Capstone Project

Cardiovascular Risk Prediction

Sanjay Jaiswal

Content:

- Problem Statement
- Data Inspection
- Data Analysis
- 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 patient's information. It includes over 3990 records and 16 attributes.

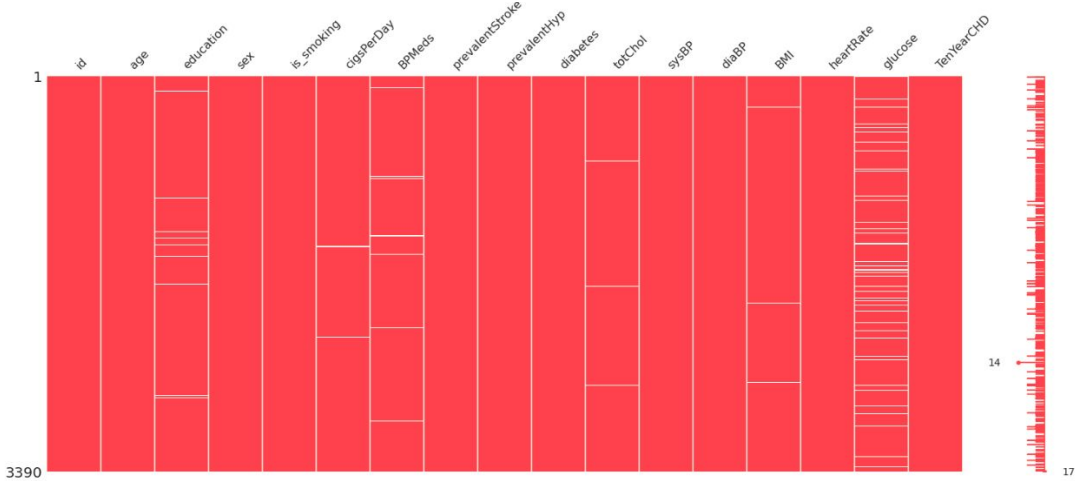
Data Inspection:

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

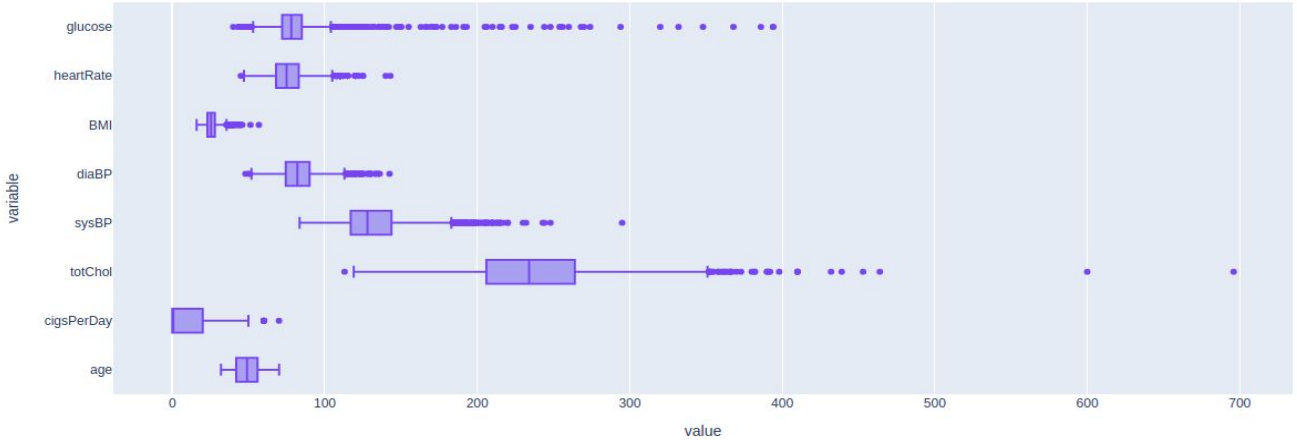
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3390 entries, 0 to 3389
Data columns (total 17 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   id               3390 non-null   int64
 1   age              3390 non-null   int64
 2   education        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
```

Data Inspection(Missing Values and Percentage):

	column_name	no.of_missing	missing_percentage
0	id	0	0.00
1	age	0	0.00
2	education	87	2.57
3	sex	0	0.00
4	is_smoking	0	0.00
5	cigsPerDay	22	0.65
6	BPMeds	44	1.30
7	prevalentStroke	0	0.00
8	prevalentHyp	0	0.00
9	diabetes	0	0.00
10	totChol	38	1.12
11	sysBP	0	0.00
12	diaBP	0	0.00
13	BMI	14	0.41
14	heartRate	1	0.03
15	glucose	304	8.97
16	TenYearCHD	0	0.00

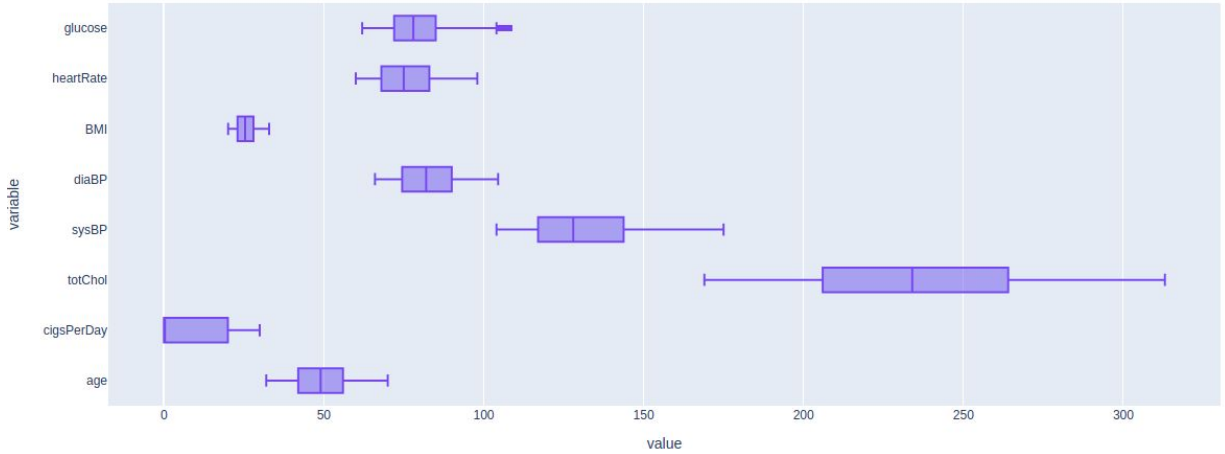


Analysis Of Outliers:

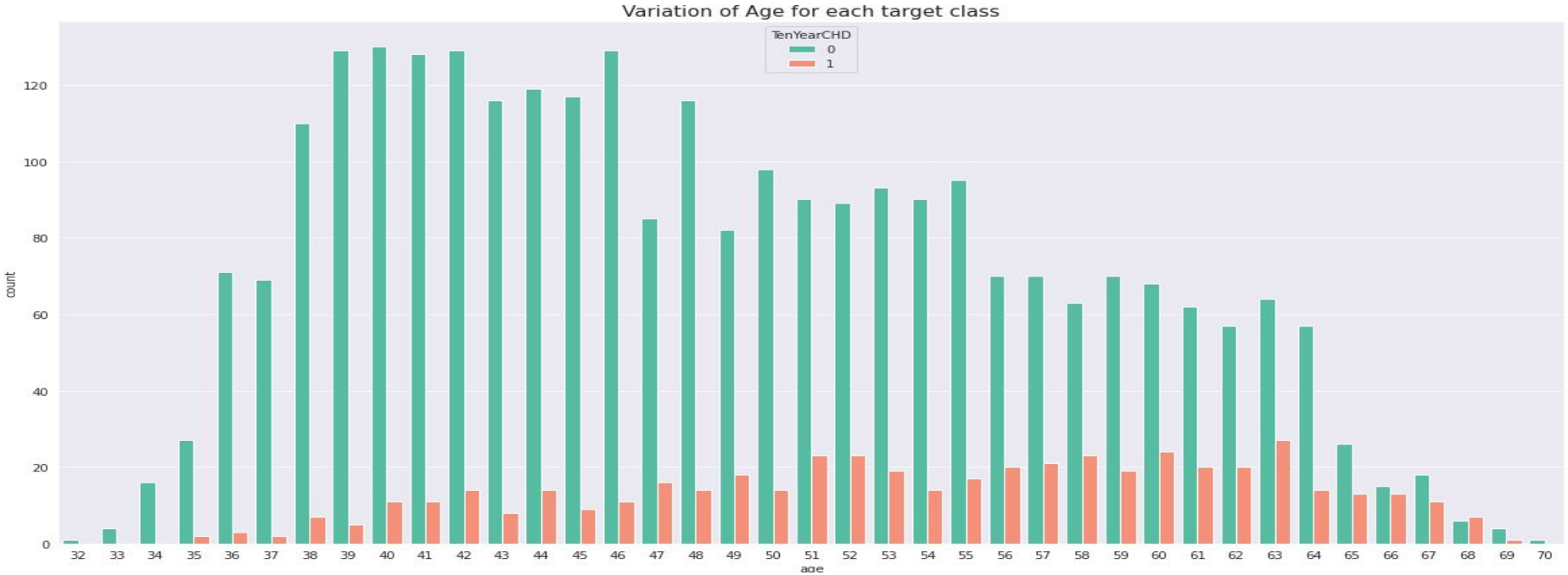


- Age has no outlier.

- Capping the outlier rows with Percentile.



Analysis Of Age for each Target class:

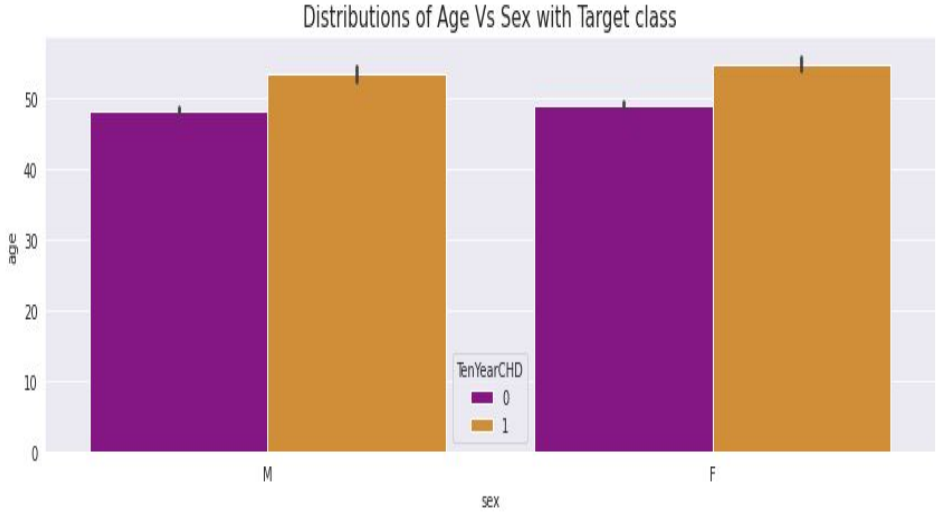


- Coronary heart disease(CHD) increases after age 51.
- Age group (34 < Age < 51) are at lower risk of cardiovascular disease.

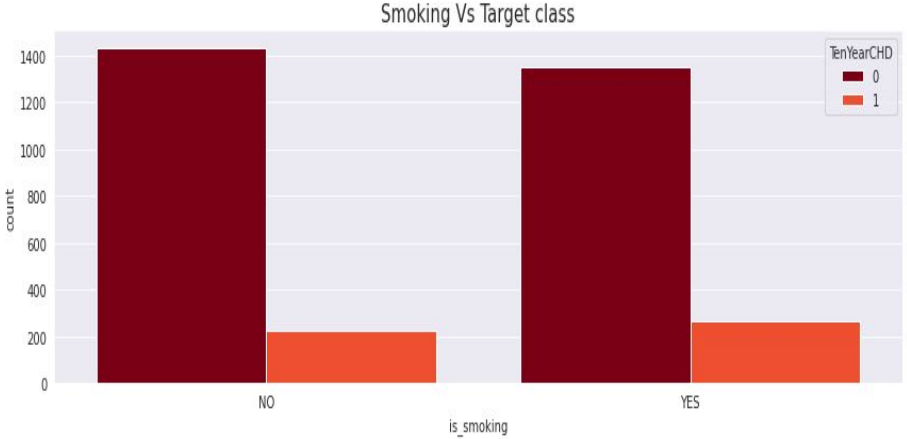
Analysis Of Age vs Sex with Target

- As we can see the countplot we can say that no. of male heart patient more than female.

- As we can see the barplot we can say that male get early CHD as compared to female.

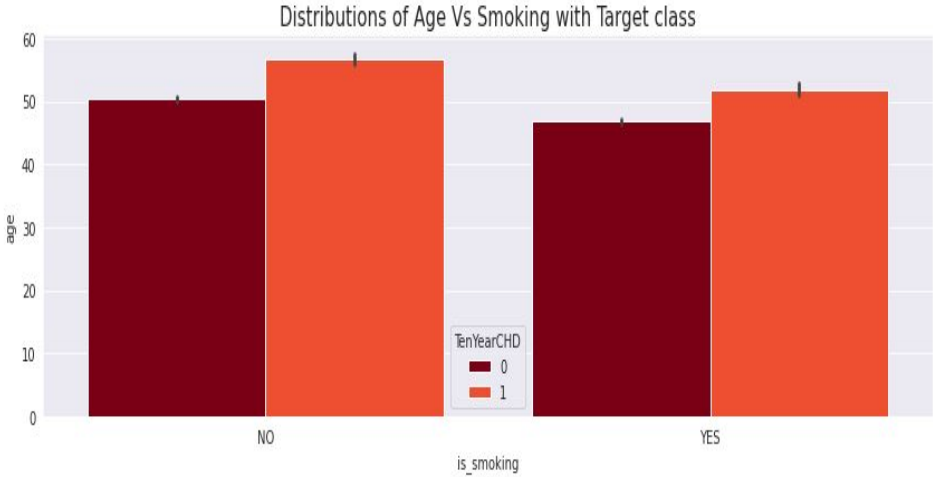


Analysis Of Age vs Smoking with



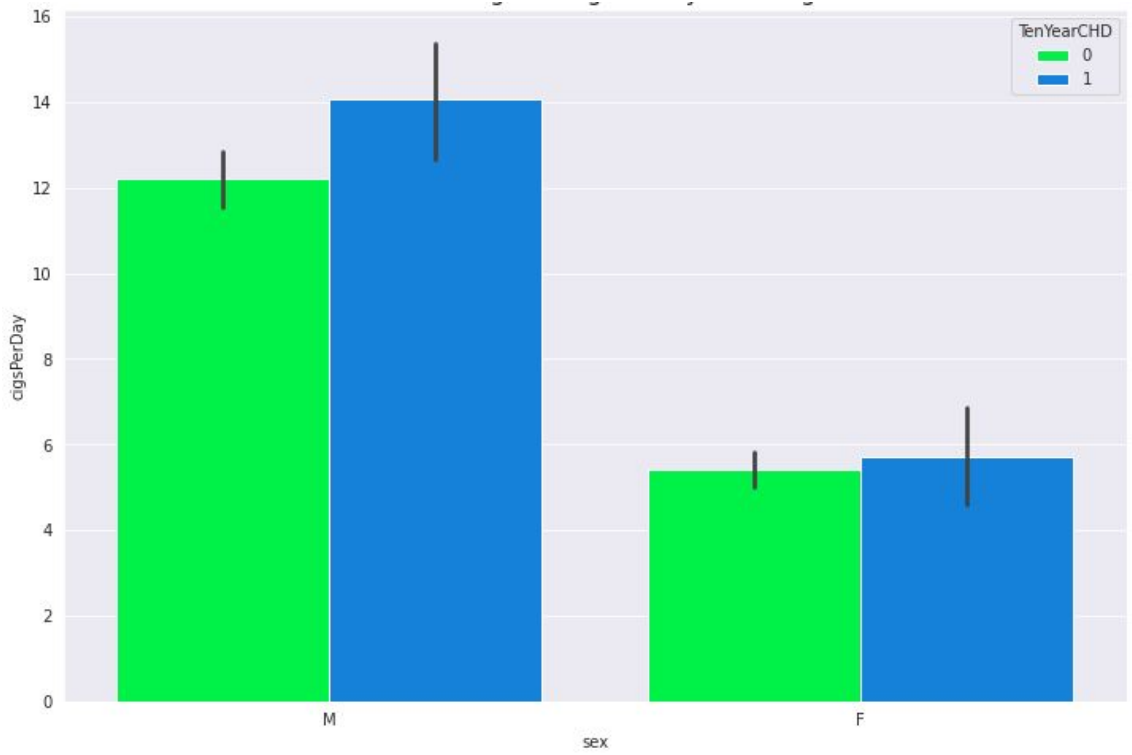
- As we can see the countplot we can say that no. of patient those who smoke more than as compared to those who won't.

- As we can see the barplot we can say that those who smoke get early heart disease as compared to those who won't.



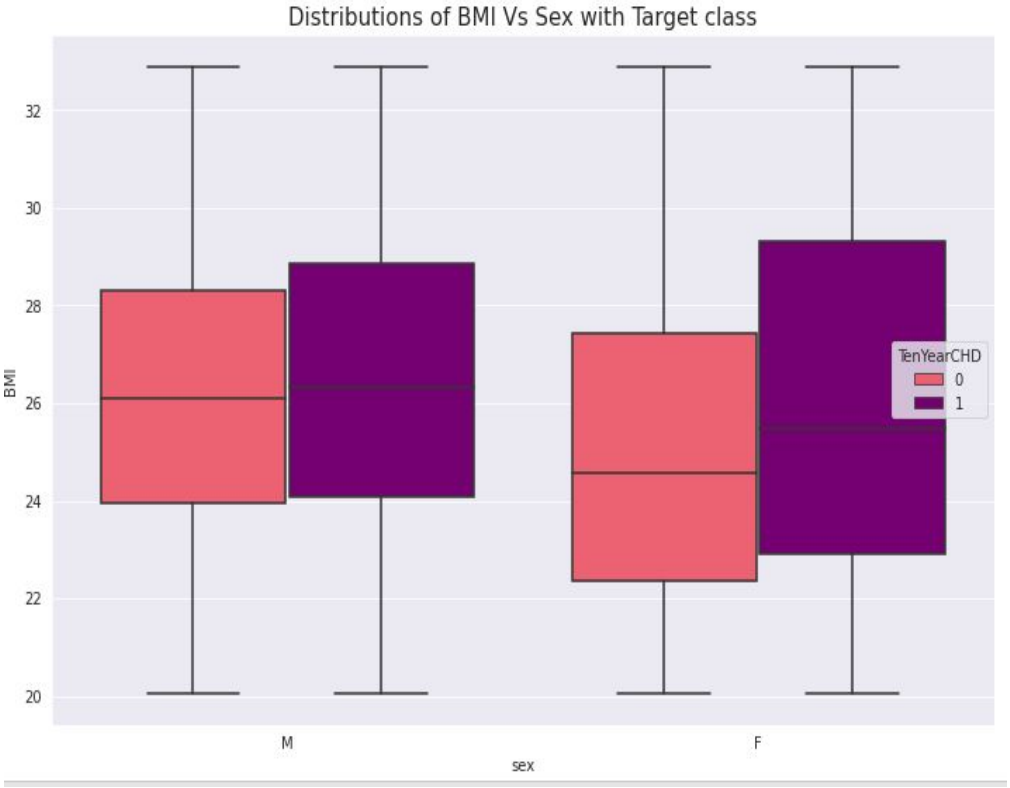
Analysis Of Cigsperday vs Sex with Target class :

- As we can see the barplot we can 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 :

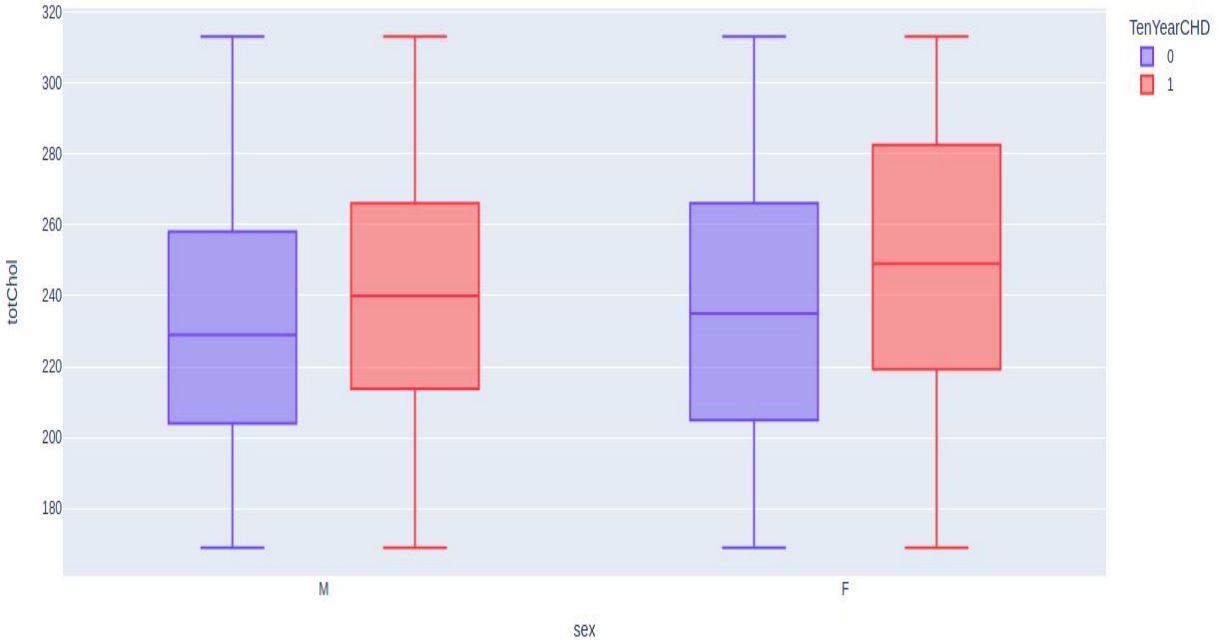
- As we can see the boxplot we can say that female BMI is more than male BMI. that's leads to OVERWEIGHT.
- So,female CHD patient more than male CHD patient.
- If your BMI is:
  - below 18.5 – you're in the underweight range
  - between 18.5 and 24.9 – you're in the healthy weight range
  - between 25 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 with Target class :

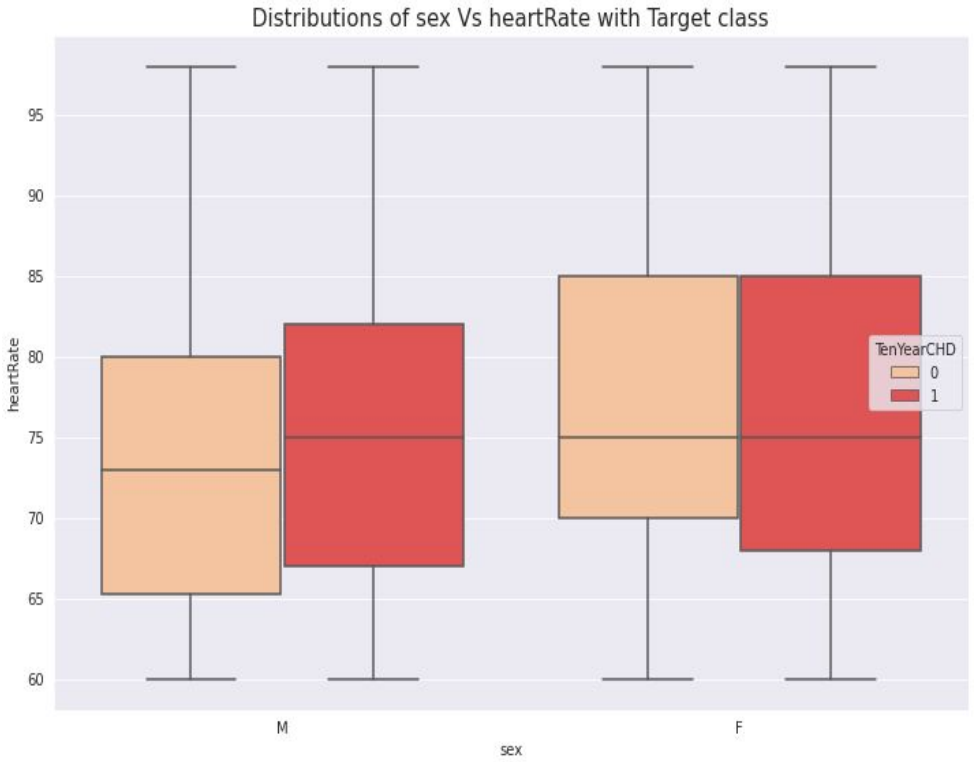
- As we can see the boxplot we can say that female cholesterol is more than male cholesterol that's leads to OVERWEIGHT.

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



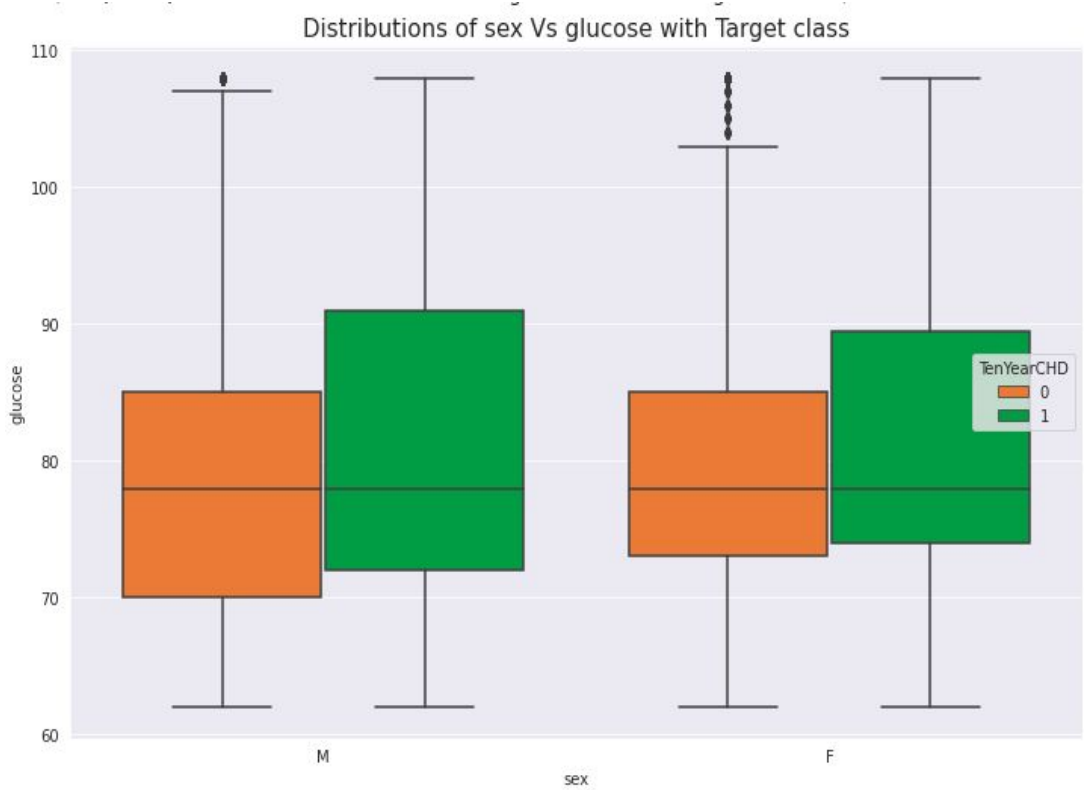
Analysis Of HeartRate vs Sex with Target class :

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

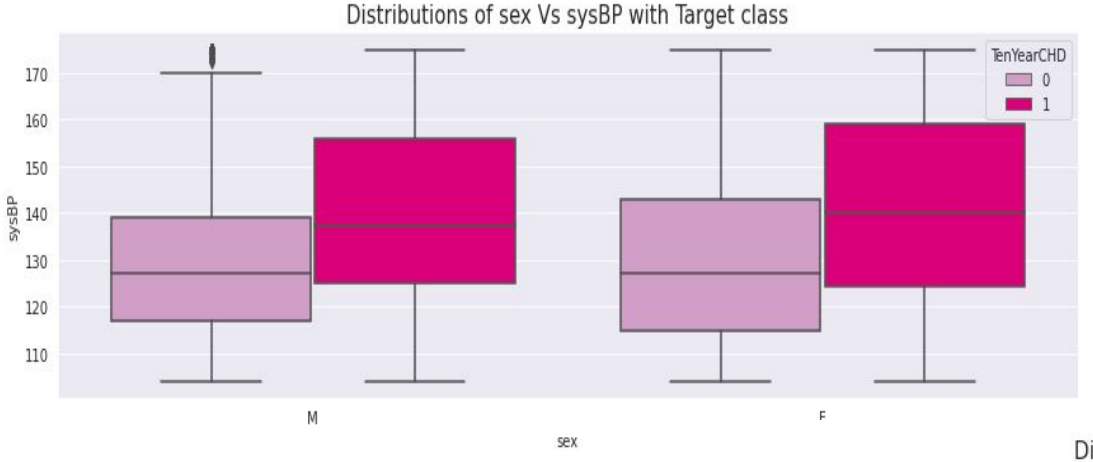


Analysis Of Glucose vs Sex with Target class :

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

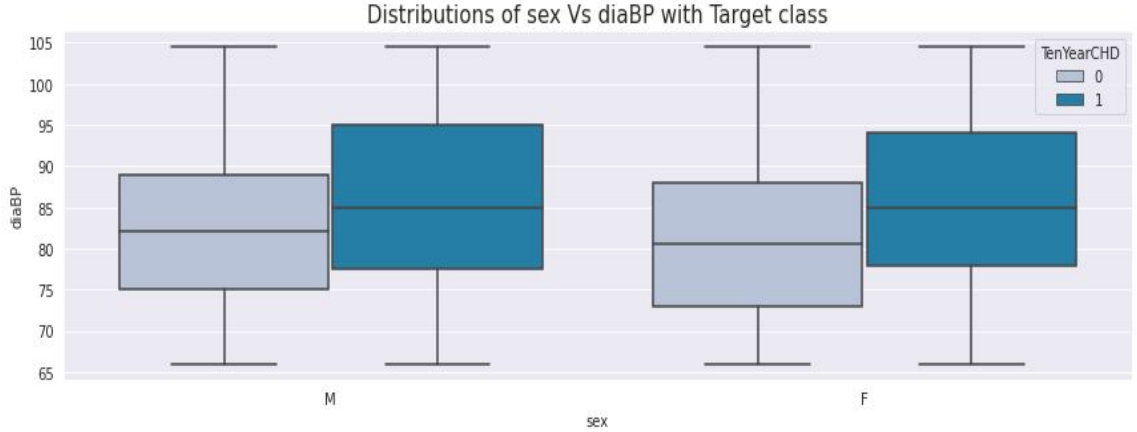


Analysis Of Systolic and Diastolic vs Sex with Target class :



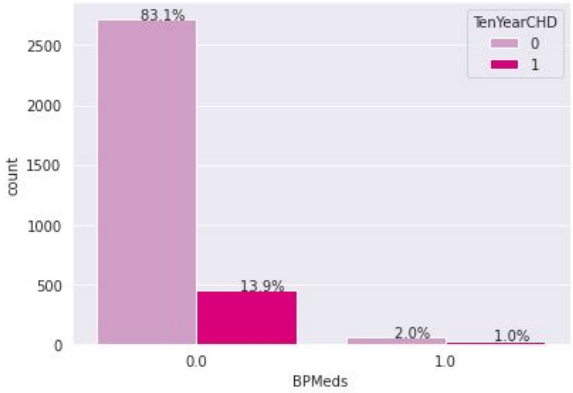
- As we can see the box plot we can say that for female heart disease patients has more Systolic BP level as compared to male heart disease patients.
- Normal < 120 mmHg.

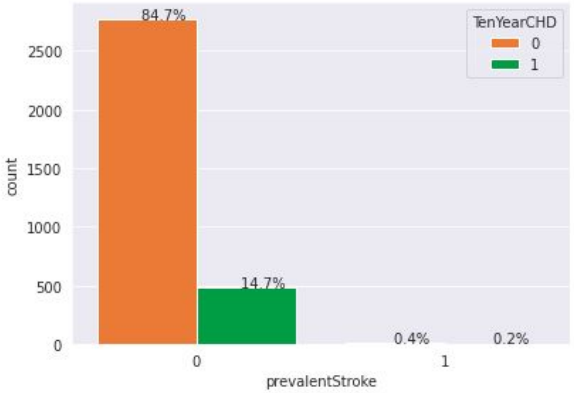
- As we can see the box plot we can say that for male heart disease patients has more Diastolic BP level as compared to female heart disease patients.
- Normal < 80 mmHg.

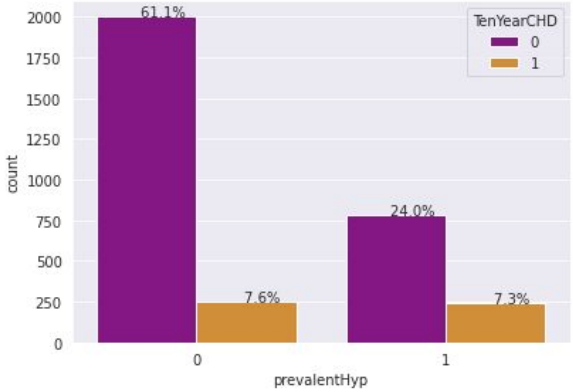


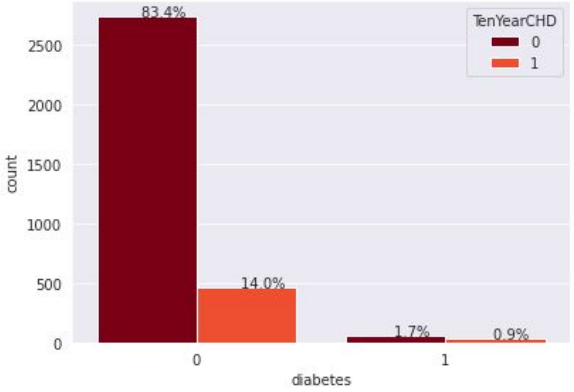
Analysis Of BPMeds | PrevalentStroke | PrevalentHyp | Diabetes vs Sex with Target class :

- 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.



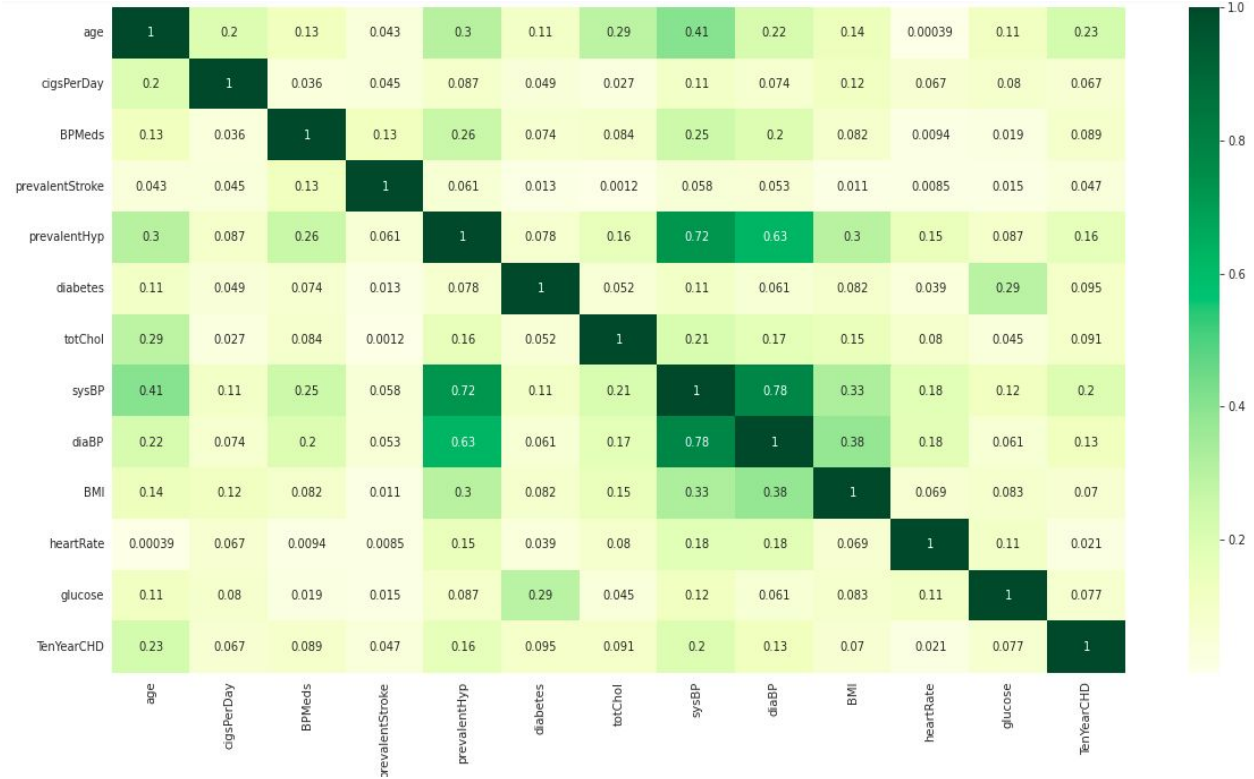






Correlation matrix:

- **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:

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

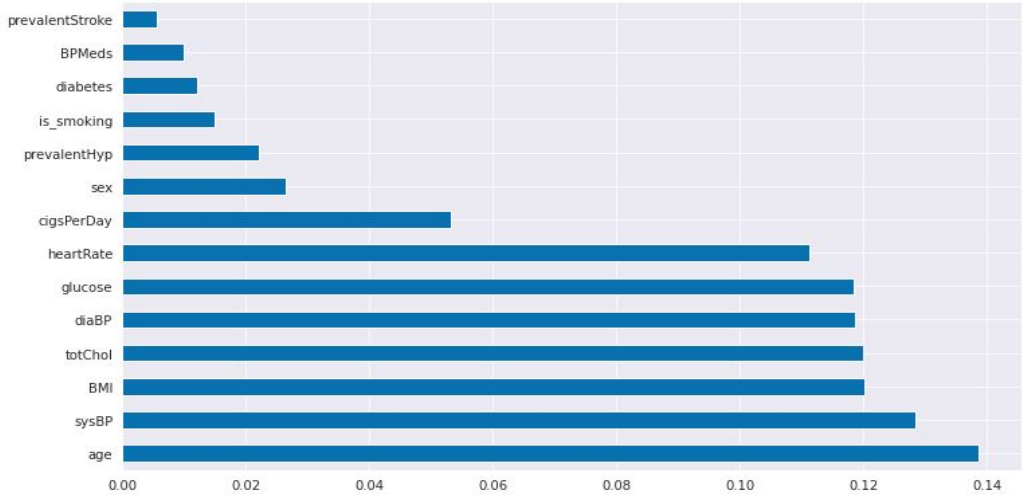
sex	is_smoking
M	NO
F	YES
M	YES
F	YES
F	NO

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

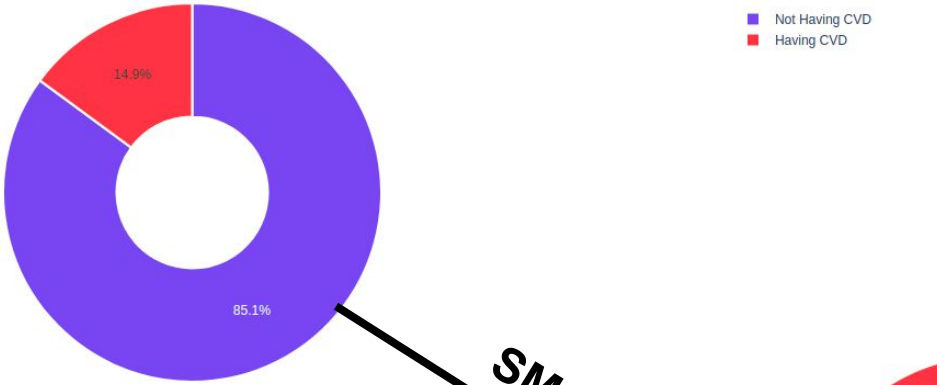
sex	is_smoking
1	0
0	1
1	1
0	1
0	0

Feature Selection:

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

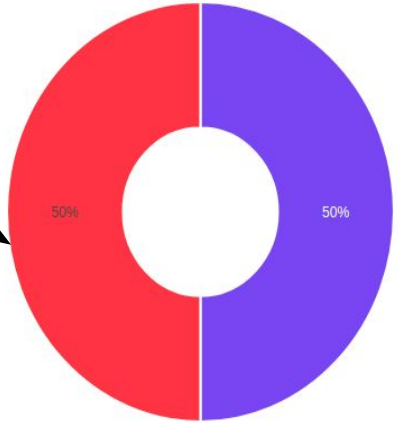


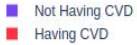
Handling Imbalanced Data:




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

SMOTE



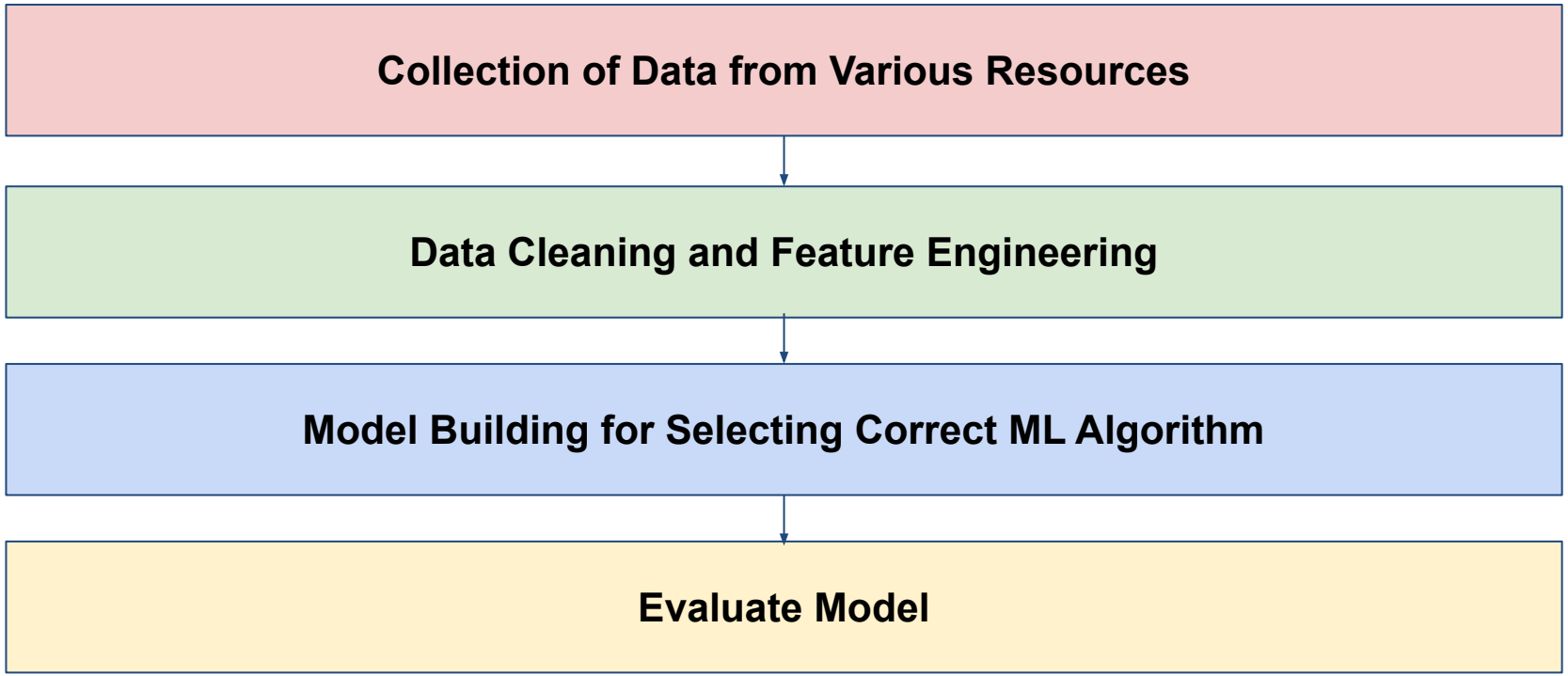


SMOTE (Synthetic Minority Oversampling Technique) works by randomly picking a point from the minority class and computing the k-nearest neighbors for this point. The synthetic points are added between the chosen point and its neighbors.

Model Building:

- Logistic Regression
- Decision Tree Classifier
- Random Forest Classifier
- XGB Classifier
- KNeighborsClassifier
- Support Vector Machine

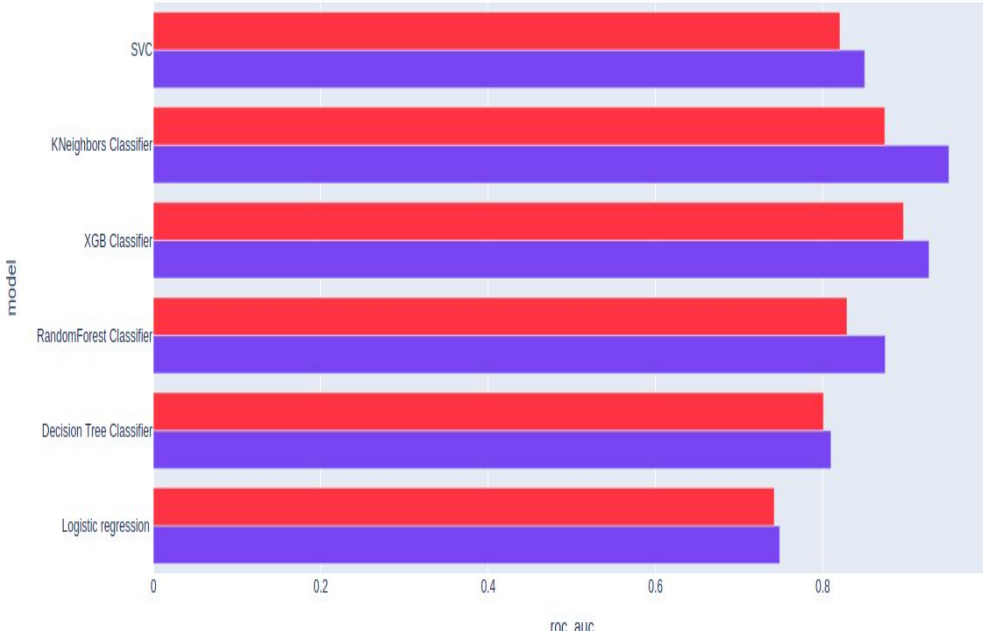
Machine Learning Process Flow:

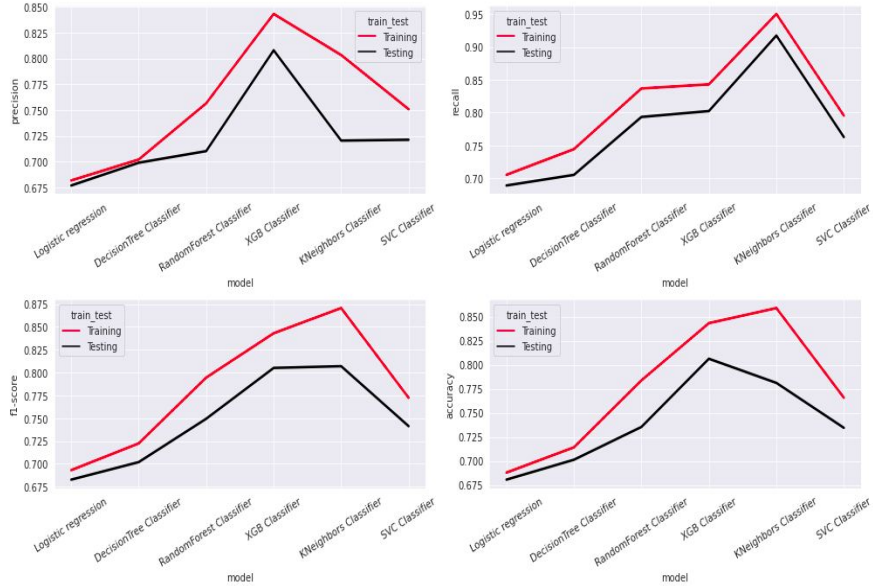


Evaluating models:

		Model	Precision	Recall	F1-Score	Accuracy	ROC_AUC
Training set	0	Logistic regression	0.6816	0.7051	0.6931	0.6877	0.7496
	1	DecisionTree Classifier	0.7019	0.7442	0.7224	0.7140	0.8106
	2	RandomForest Classifier	0.7565	0.8366	0.7945	0.7836	0.8727
	3	XGB Classifier	0.8433	0.8429	0.8431	0.8431	0.9284
	4	KNeighbors Classifier	0.8034	0.9502	0.8707	0.8588	0.9553
	5	SVC Classifier	0.7508	0.7953	0.7724	0.7656	0.8530
Testing set	0	Logistic regression	0.6767	0.6888	0.6827	0.6804	0.7391
	1	DecisionTree Classifier	0.6988	0.7050	0.7019	0.7011	0.7898
	2	RandomForest Classifier	0.7101	0.7932	0.7494	0.7352	0.8232
	3	XGB Classifier	0.8080	0.8022	0.8051	0.8061	0.8957
	4	KNeighbors Classifier	0.7203	0.9173	0.8070	0.7810	0.8678
	5	SVC Classifier	0.7211	0.7626	0.7413	0.7343	0.8169

Comparing different ML Models:





- In the above 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.
- XGB Classifier are performing well which has 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!