CAPSTONE PROJECT -III

Supervised ML - Classification

Cardiovascular_Risk_Prediction

Team Members:

Shubham Joshi

Kanika Kakra

Akshay Fasale

Rishikesh Damale

✓Acknowledgement

- We would like to express our gratitude towards the entire team of "*Almabetter*" for acknowledging us with such important domain and providing us an opportunity to work on real life problems through Capstone Project.

✓Problem Statement

- 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. Each attribute is a potential risk factor. There are both demographic, behavioral, and medical risk factors.

✓Algorithm

1.Importing the necessary packages and libraries.
2.Mounting the drive for importing the data.
3.Checking for missing, Nan values, Null values.
4.Observing the datatypes and features.
5.Observing the correlation among independent variables.
6. Exploring the categorical values, numerical features from data set.
7.Exploring the target variable.
8.Handling imbalanced dataset with SMOTE.
9.Splitting the data and training the data.
10.Observing the results.

✓Data Description

Target variable (desired target)

- • 10-year risk of coronary heart disease CHD(binary: “1”, means “Yes”, “0” means “No”) – DV

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 the patient is a current smoker ("YES" or "NO")
- • **CigsPerDay:** 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.)

Medical(history)

- • **BP Meds:** whether the patient was on blood pressure medication (Nominal).
- • **Prevalent Stroke:** whether the patient had previously had a stroke (Nominal).
- • **Prevalent Hyp:** whether the patient was hypertensive (Nominal).
- • **Diabetes:** whether the patient had diabetes (Nominal).

Medical(current)

- • **TotChol:** total cholesterol level (Continuous).
- • **SysBP:** systolic blood pressure (Continuous).
- • **DiaBP:** 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)

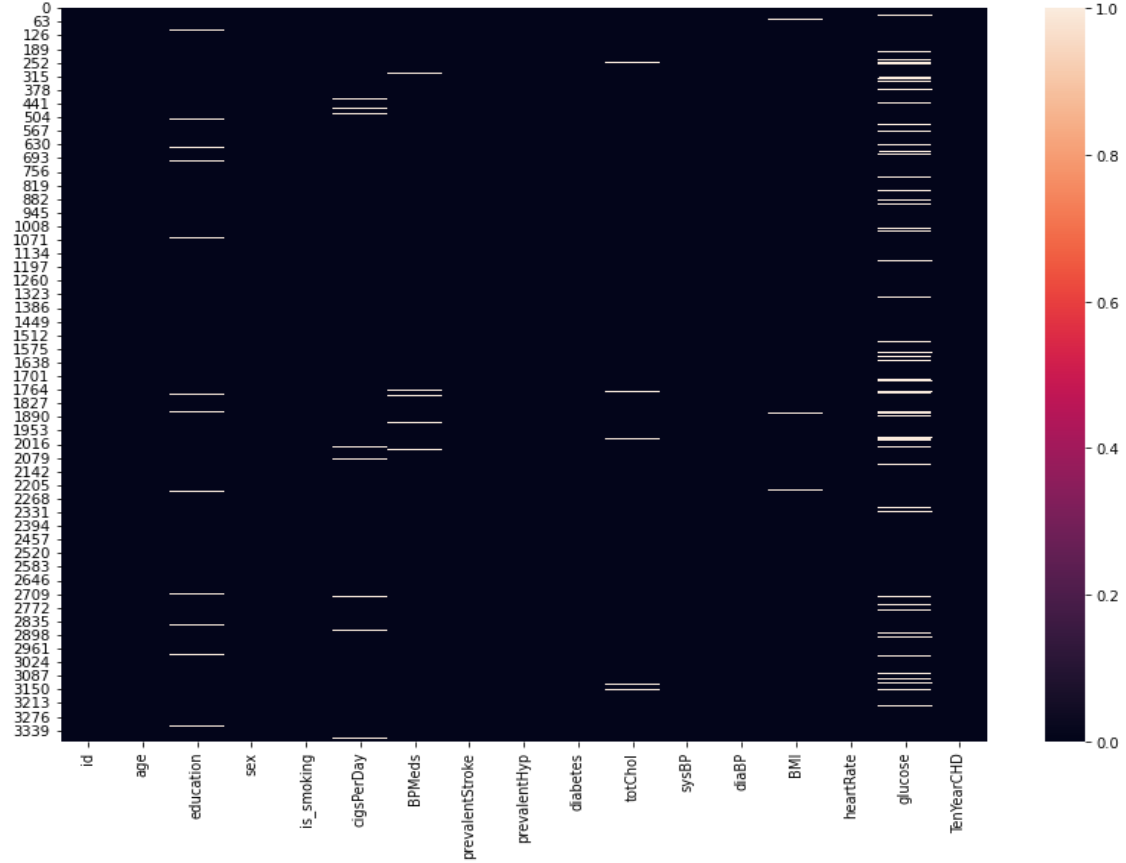
✓EDA and Feature Engineering

- Exploratory data analysis or commonly known as EDA helps to explore data, and possibly formulate hypotheses that might cause new data collection and experiments. It's a scientific approach to get the story of the data.

- Feature engineering mainly have two goals:

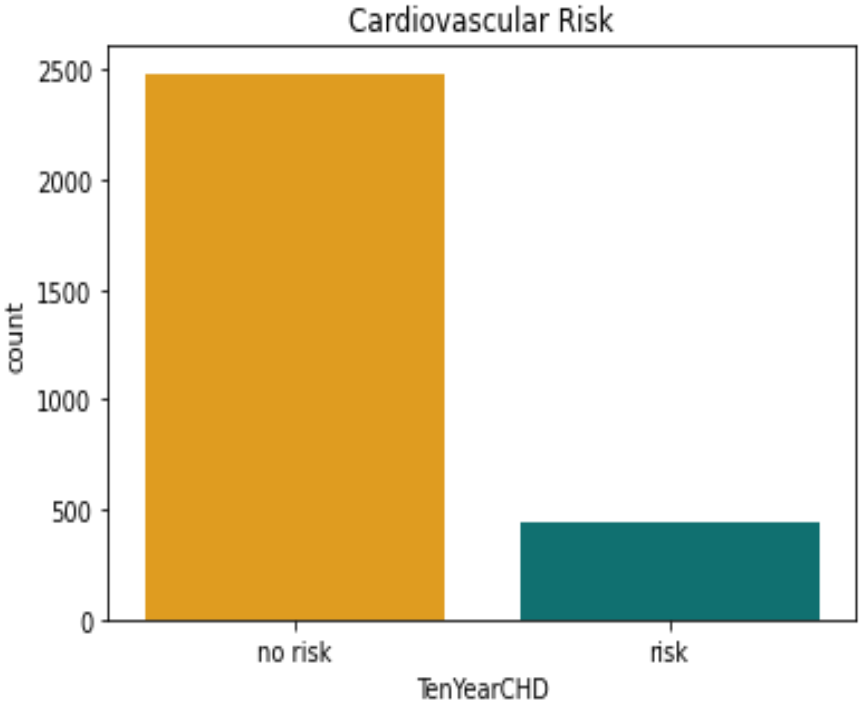
  - Preparing the proper input dataset, compatible with the machine learning algorithm requirements.
  - Improving the performance of machine learning models.

✓Heat Map for missing Values



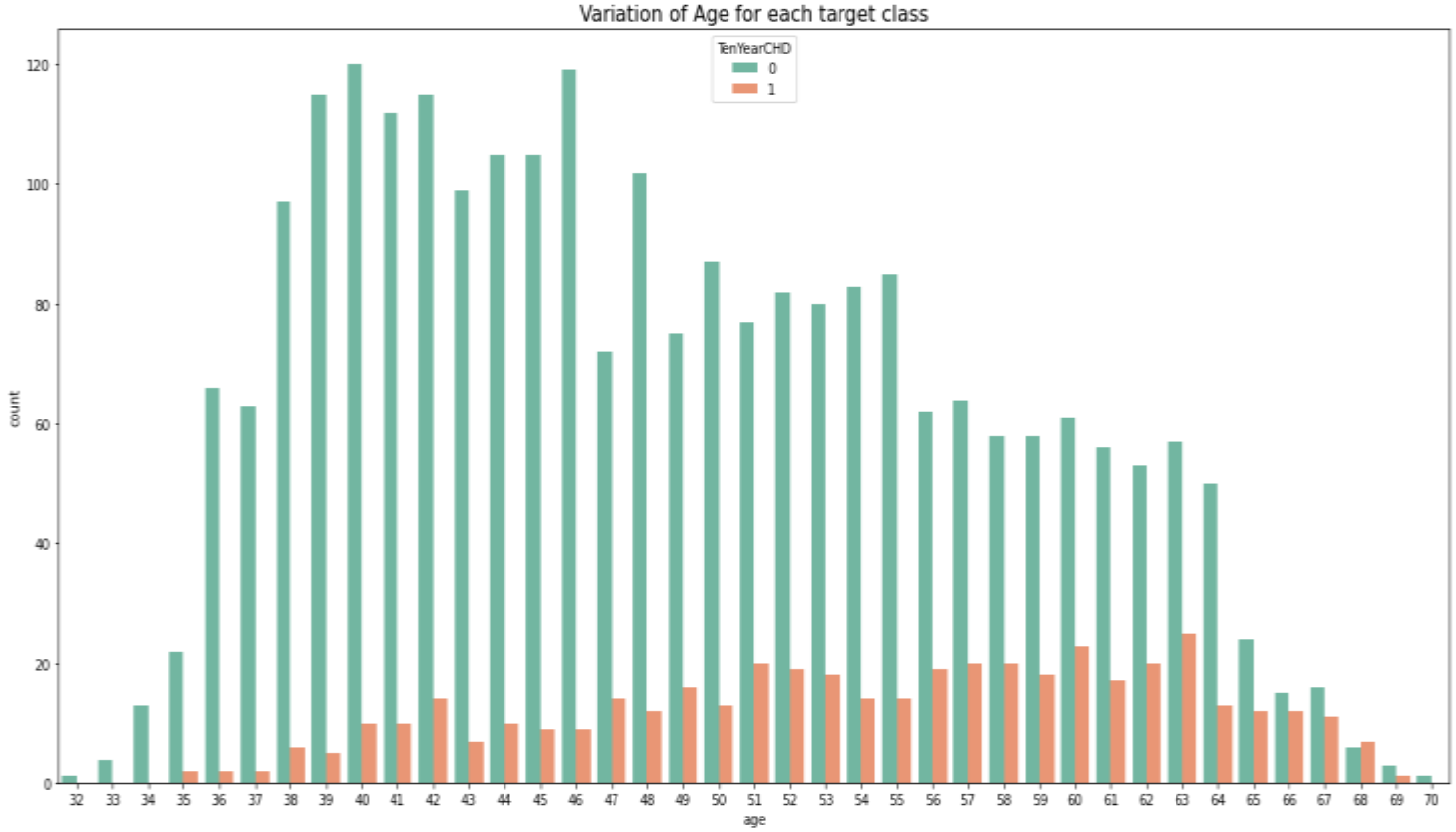
- Usually, we replace this null values with the help of other records.
- But, in this case the dataset is related to medical domain, that said, the entries in this dataset are person specific and the values vary among different individuals.
- Hence the most logical option that we should deal with such values is removing the rows with any null value.

✓Data Visualization



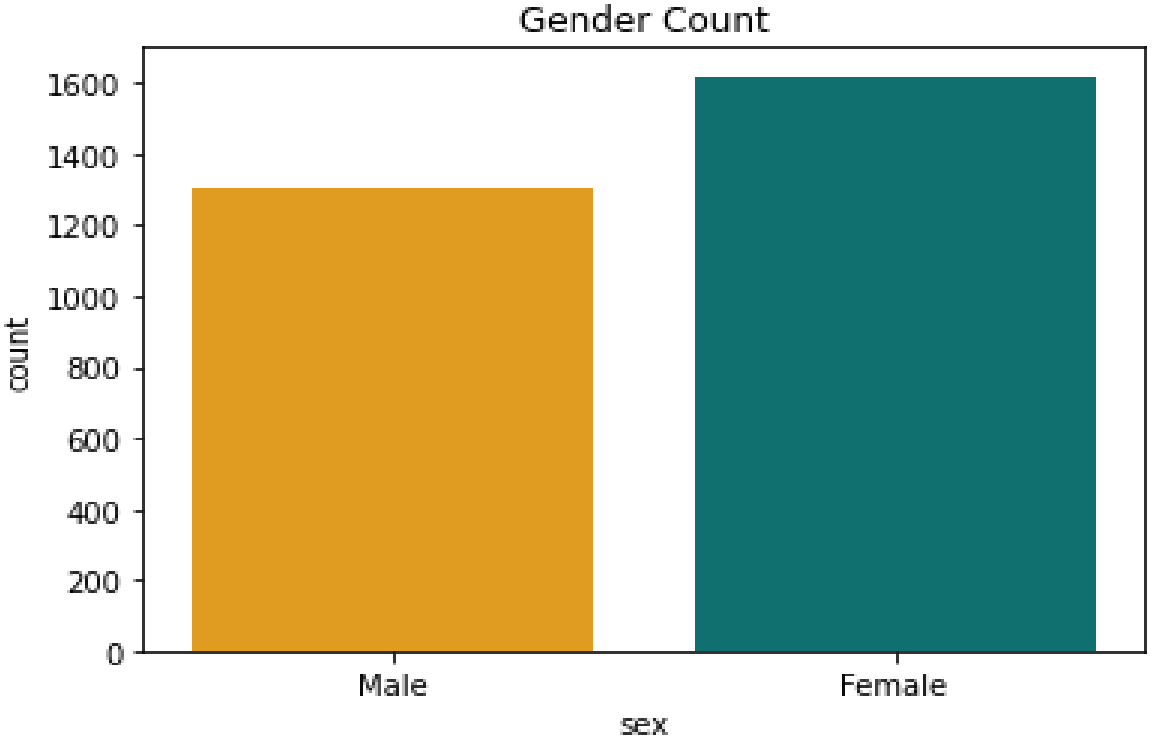
- It is clear the percentage/count of people with normal results are high and this creates a problem of class imbalance.
- It could create problems for model to perform better in such case because the model will be biased towards the normal result predictions.

CHD vs AGE



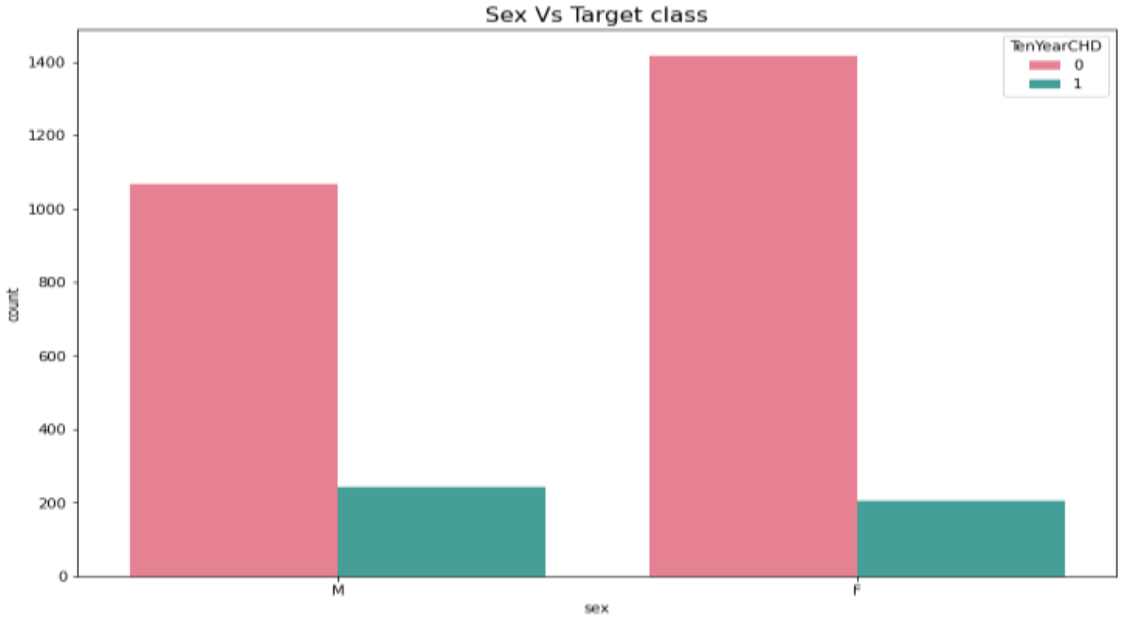
Here we see that Chances of CHD increases from age 51 to 65.

Gender



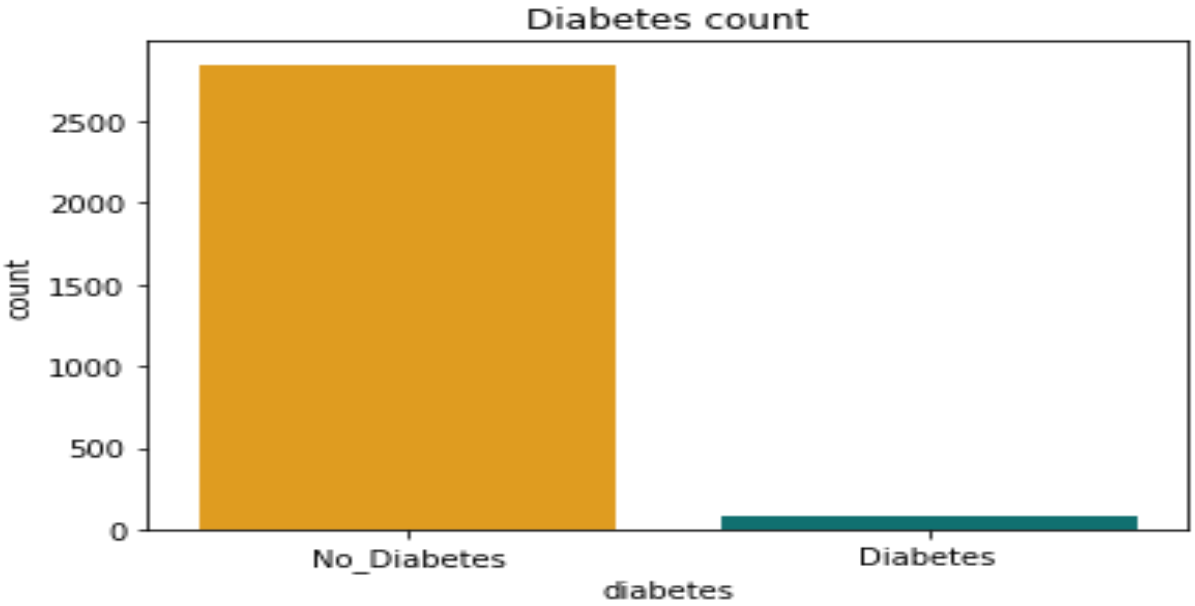
3.We observe that number of Female records are more than Male records in the dataset.

Gender vs Target class



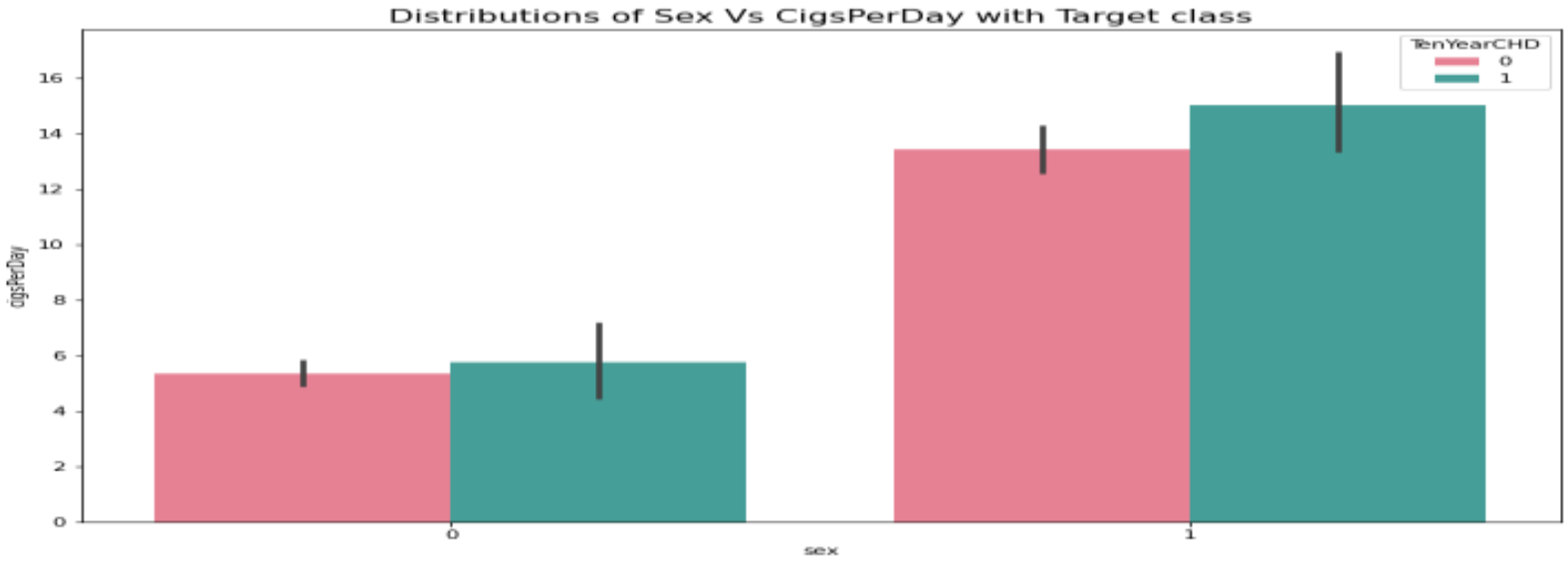
- 4.In the above bar chart we can say that no.of CHD female is less than male.

Diabetes



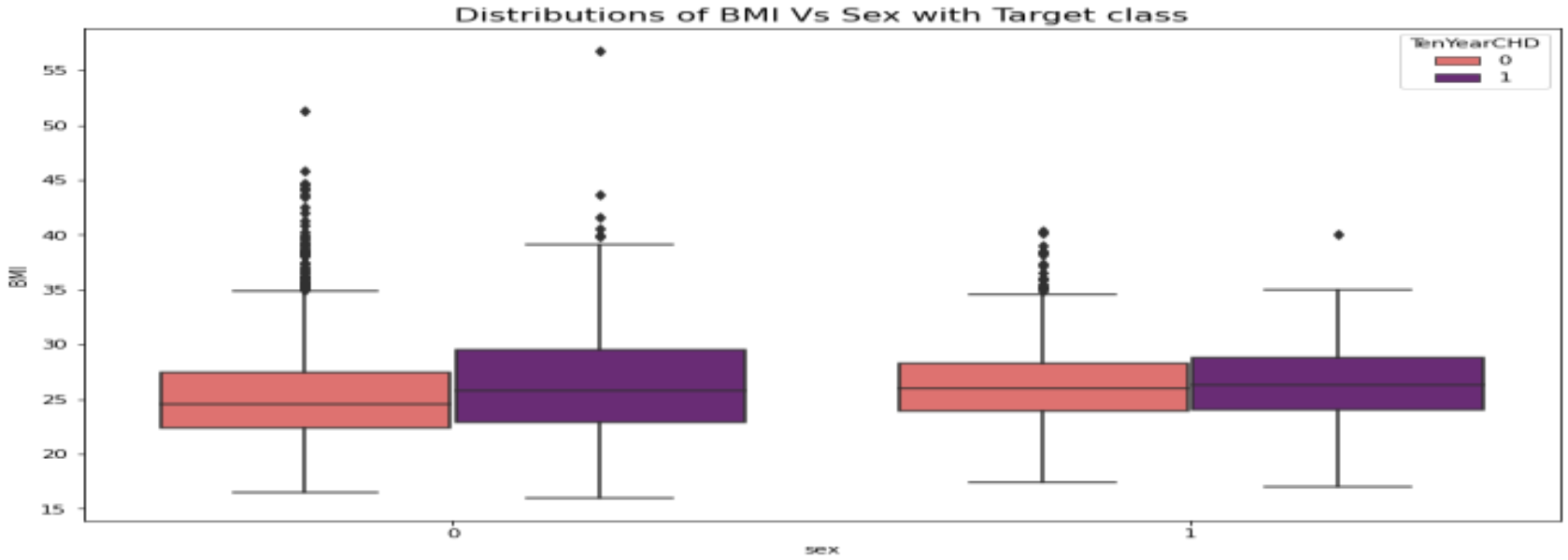
- Only 79 people are diabetic in the data ..There is a large difference in the data ...(bias towards nondiabetic).

CigsPerDay



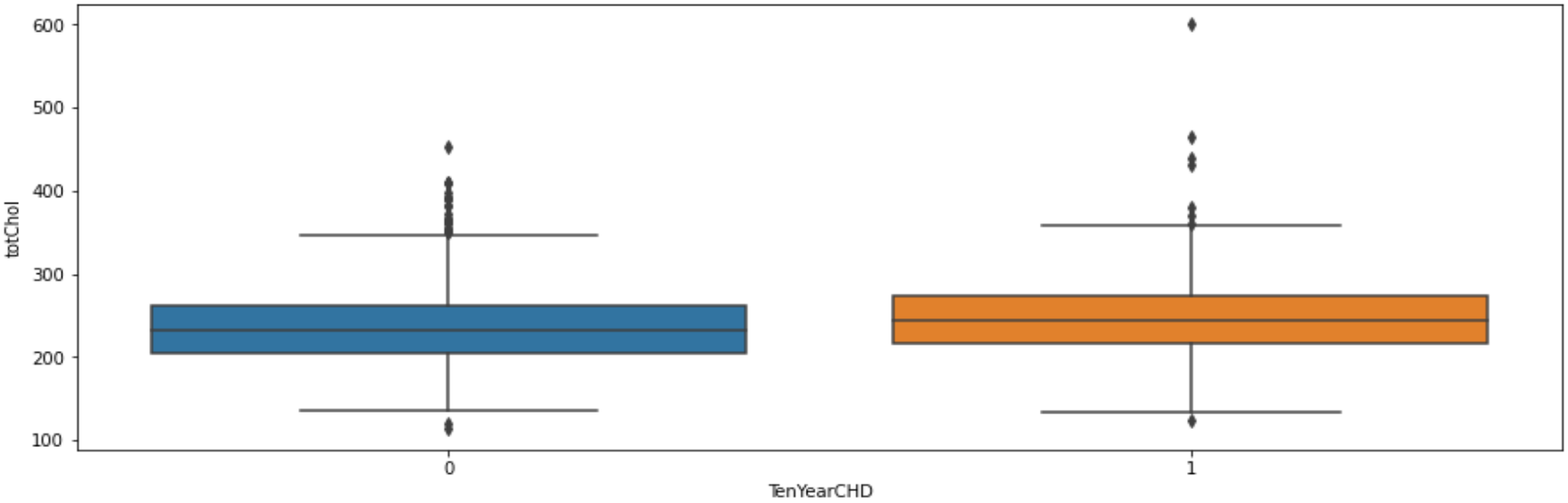
- Its clear that number of males who smokes are significantly higher than the female smokers, even though the entries for female patients were more.
- This shows smoking is more common in males and its a major cause of CHDs for male patients. 1=M , 0= F

BMI



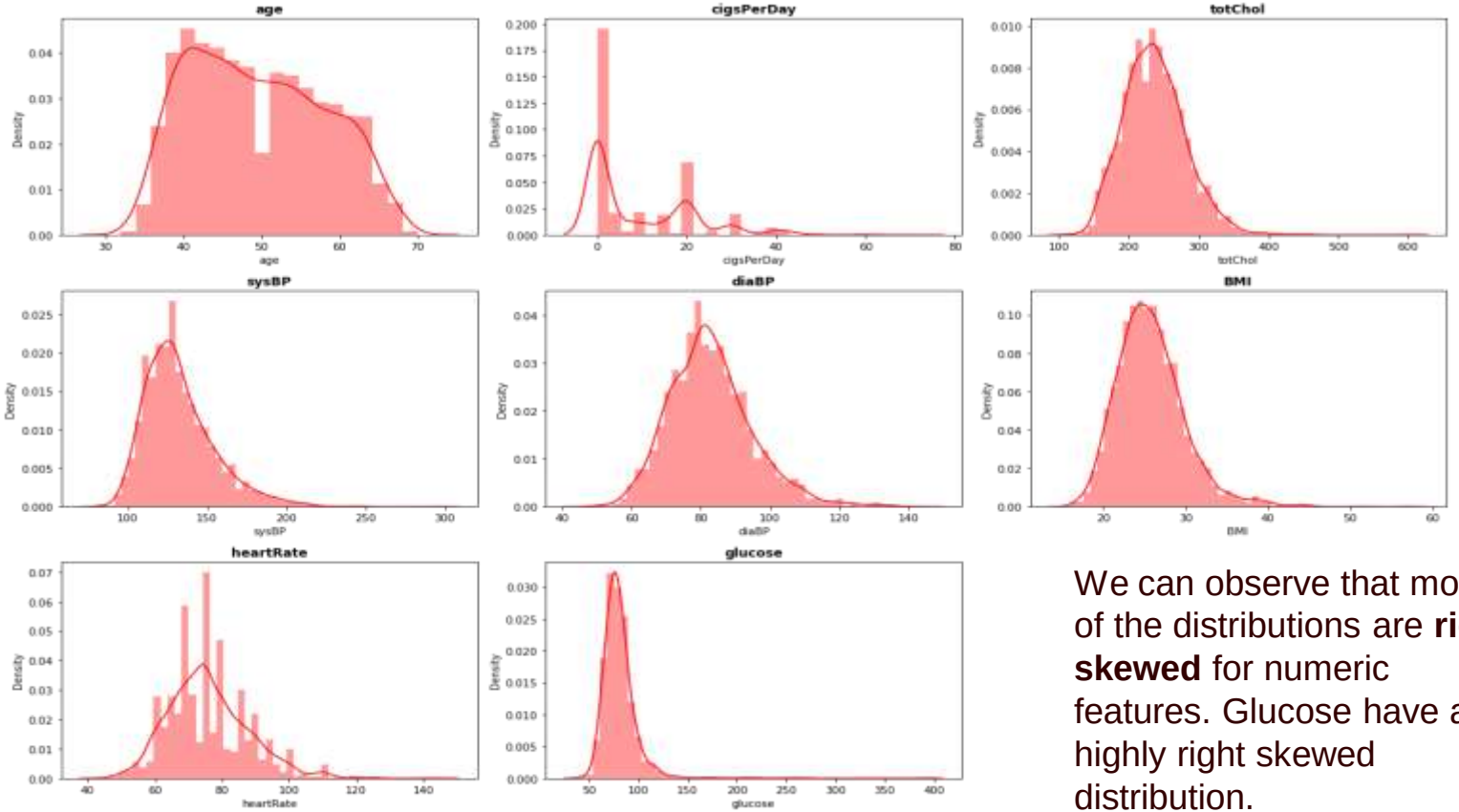
- In above plot we can say that Female BMI is more than male BMI that's leads to overweight.
- so, Female CHD is more than male CHD.

tot Chol



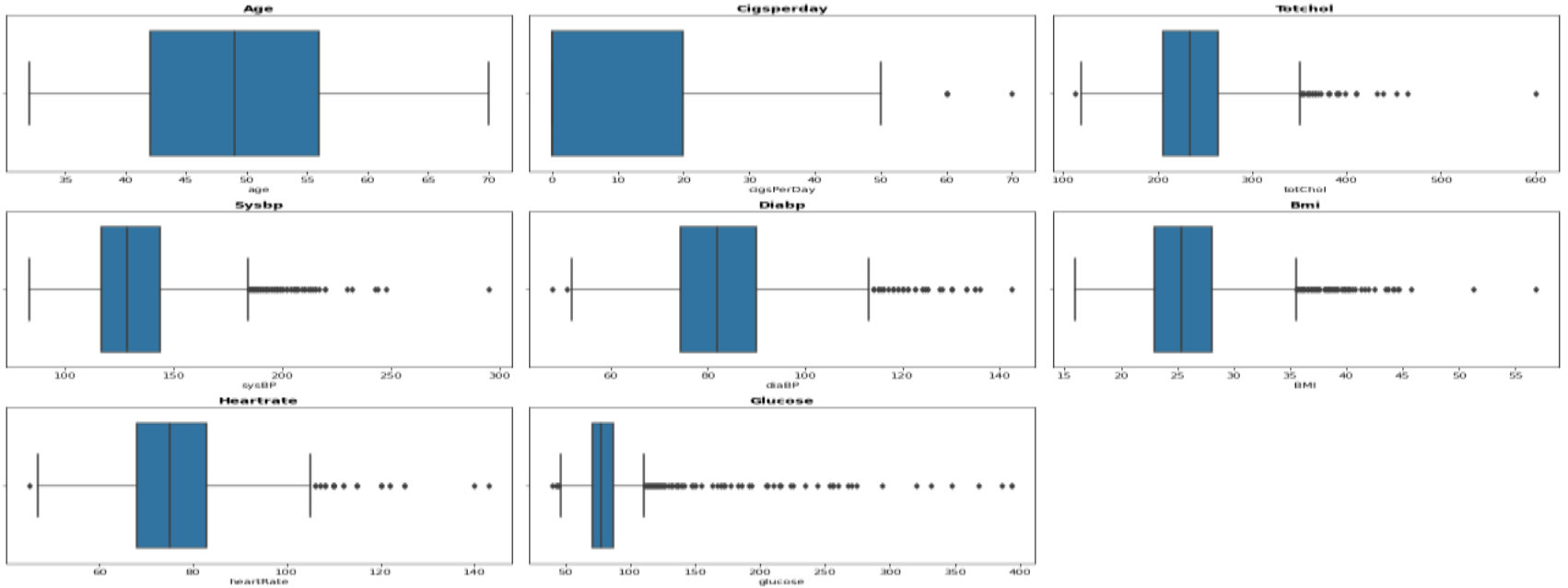
8.We observe that male Have little more cholesterol level than female

Distributions



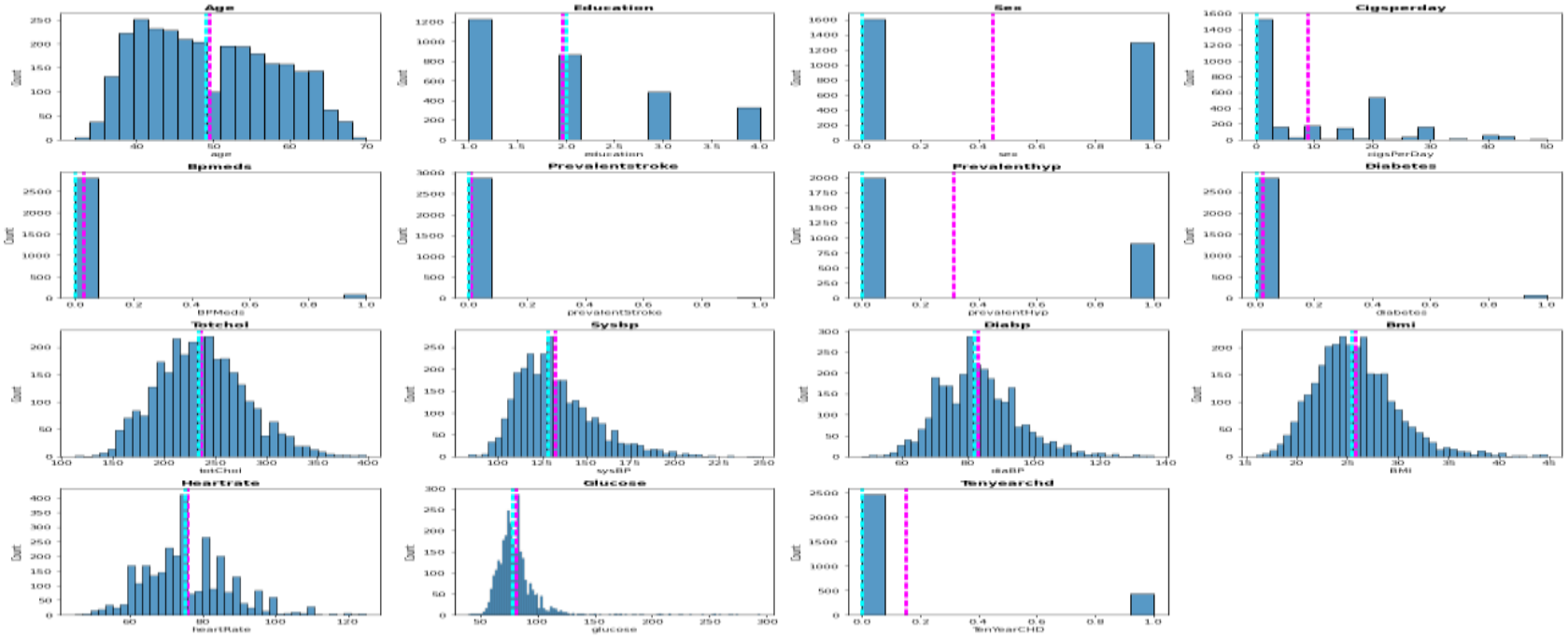
We can observe that most of the distributions are **right skewed** for numeric features. Glucose have a highly right skewed distribution.

✓Box plot distributions for numeric features.



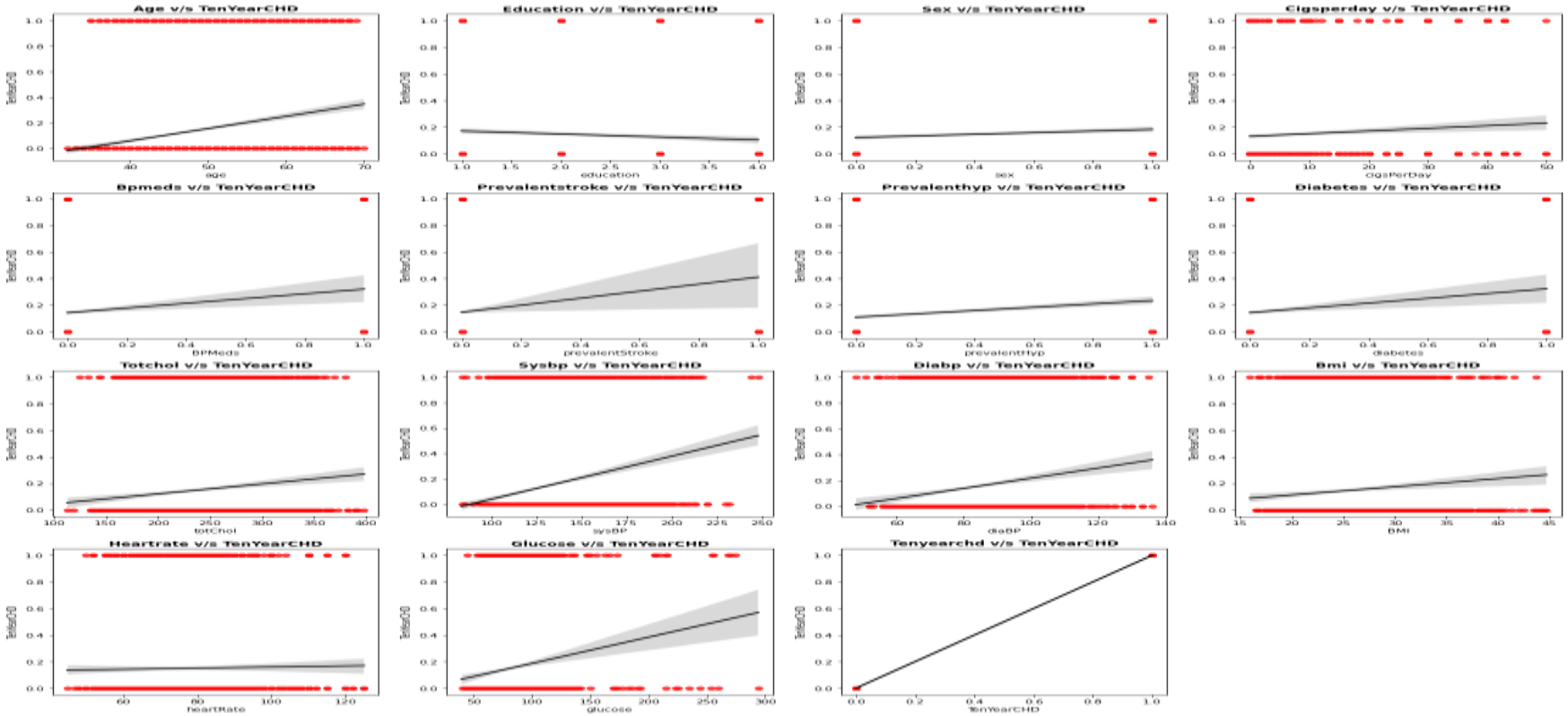
As we can see a lot of outliers in **totchol, sysbp, diabp, BMI , Glucose**. As the data come from medical survey, we can't manipulate the data. If we drop the rows with this outliers, we will lose a huge amount of important data so we can't drop them..We can only try to remove borderline outliers(with minimal data loss) which are unlikely to occur (doesn't make sense).

✓Univariate Analysis



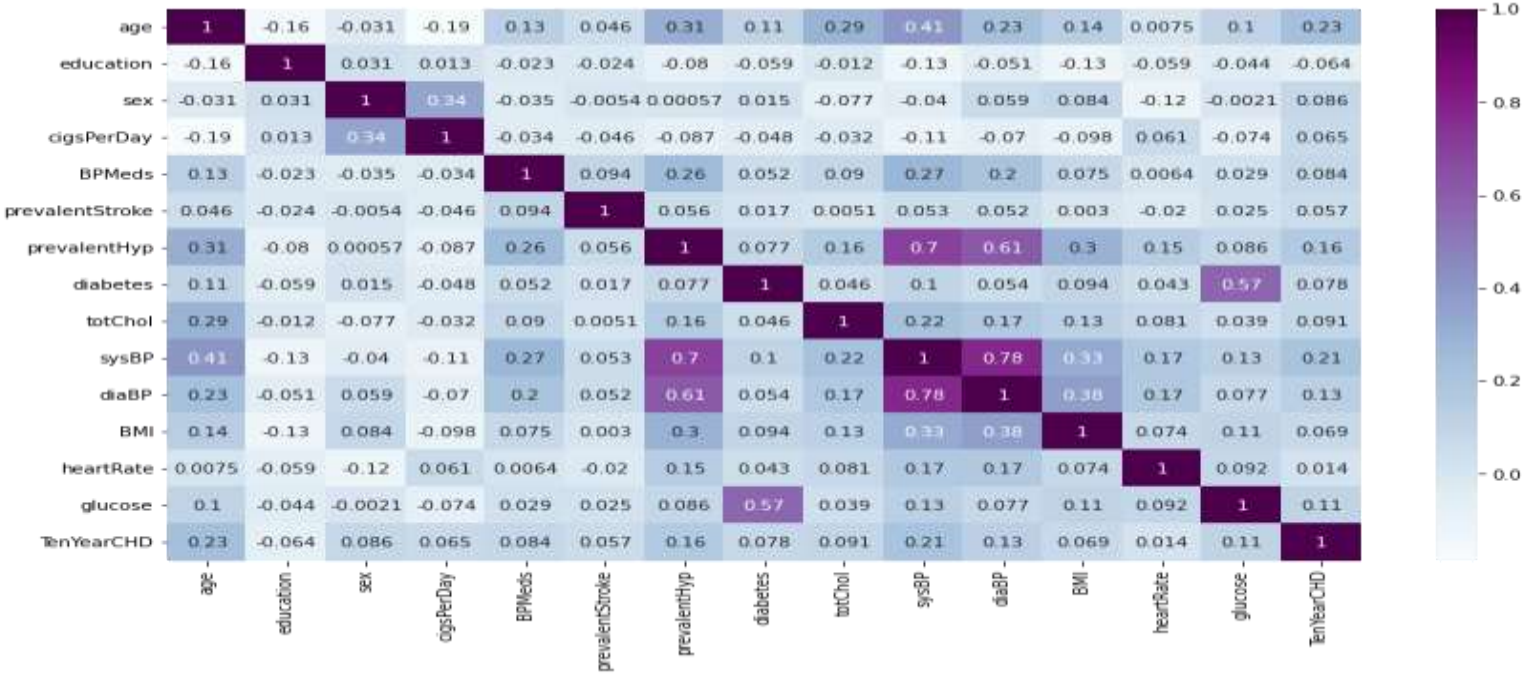
- 11.From above histograms we can see how mean, and median are positioned in distributions.

✓Bivariate Analysis



- As we can see a lot of Independent variables show relation with our Target variable.

✓Multicollinearity

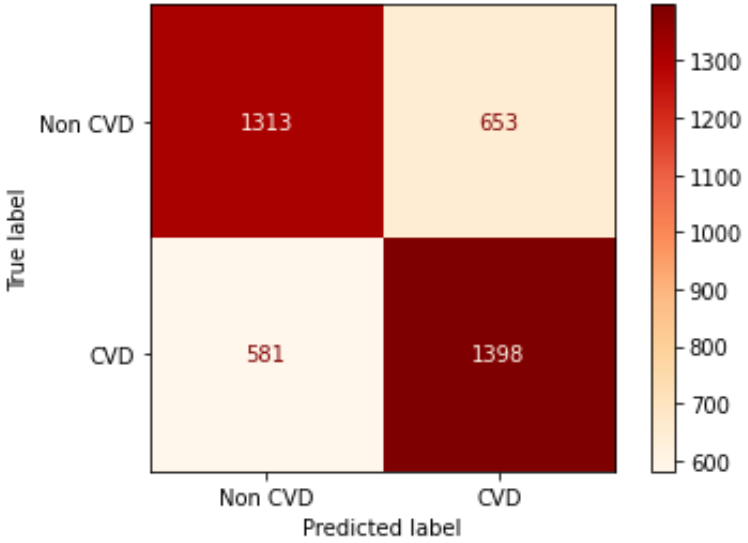


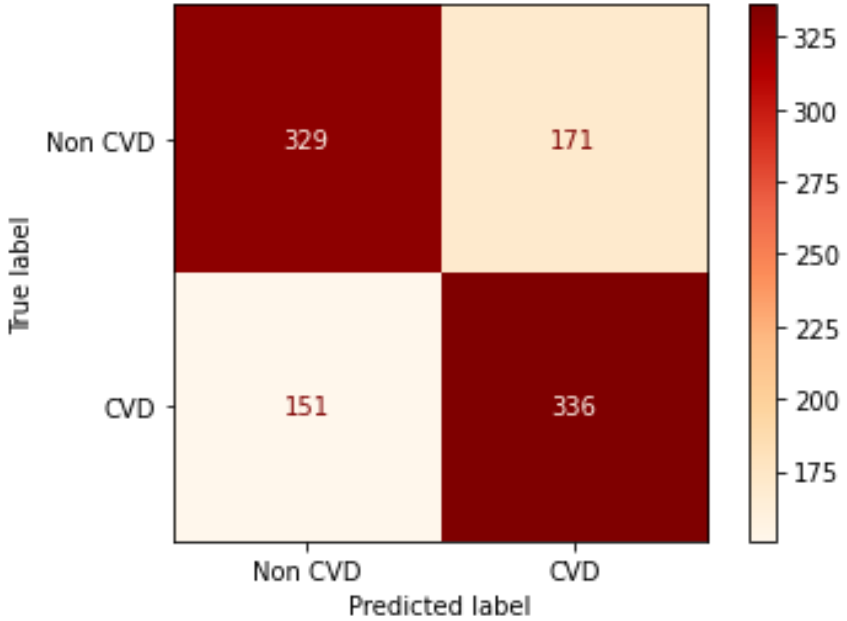
From the above Heatmap, We can see both columns(DiaBP and SysBP) are heavily correlated, there's some relationship we can establish with these two features further. And created **Mean arterial Pressure** ..which reduced multicollinearity.

Model Development

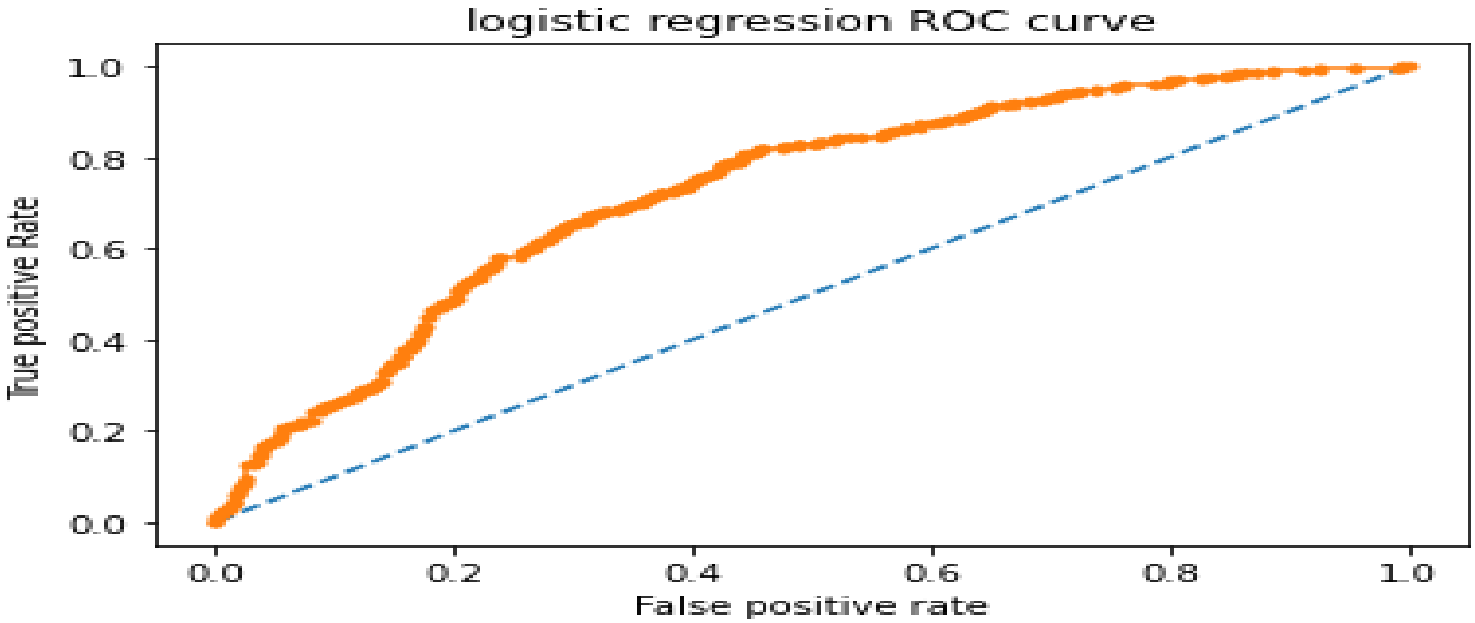
- Now that the Dataset is cleaned, and we have added all the necessary features along with some conversions of categorical features.
- And we also took care of data imbalance problem with the help of **SMOTE** (Synthetic Minority Oversampling Technique).
- Its time to split the data into training and testing sets and observe how the models are performing..
- These training and testing data are going to be same for all the model we'll build such that all the models are evaluated on a same set of parameters.

✓Logistic Regression



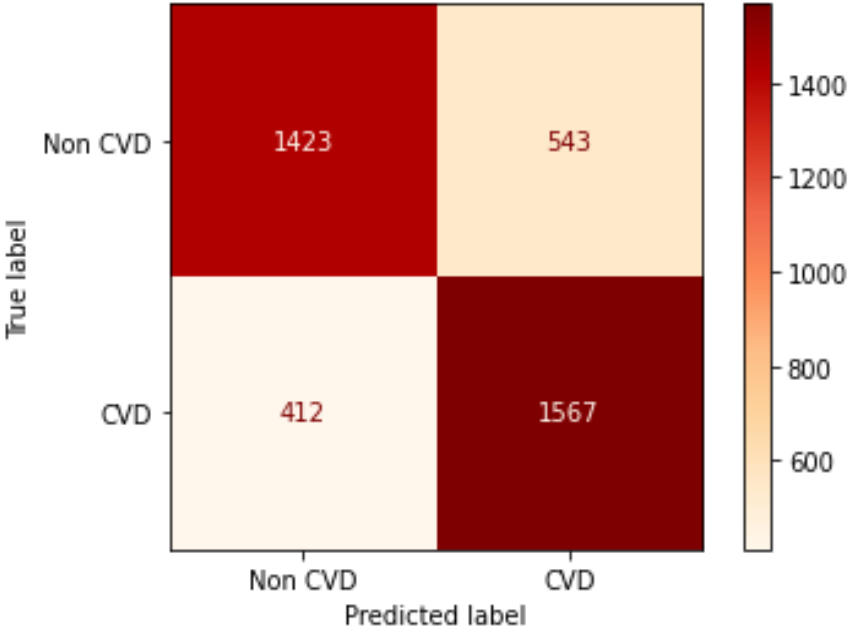


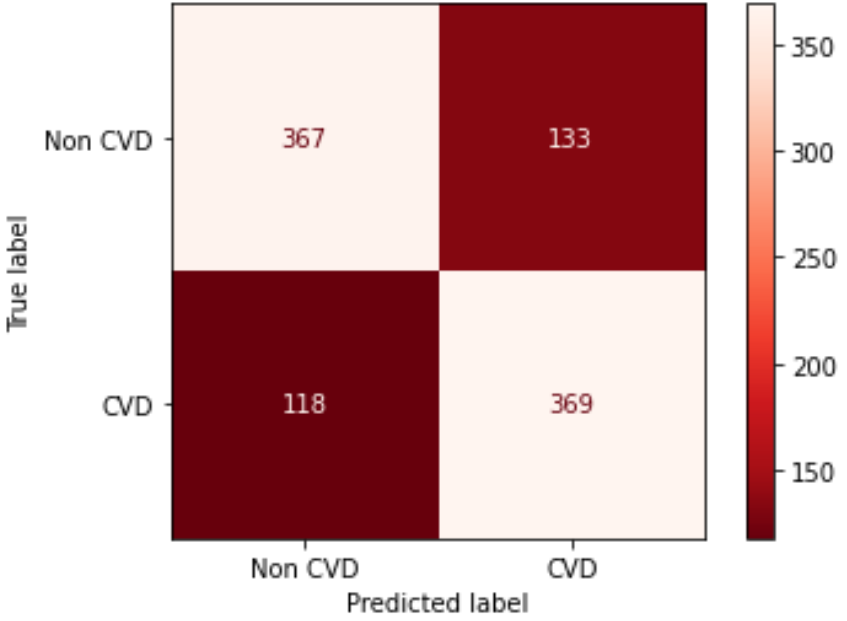
a)Visualizing the confusion matrix to evaluate the performance of the model on training set and testing set, respectively.



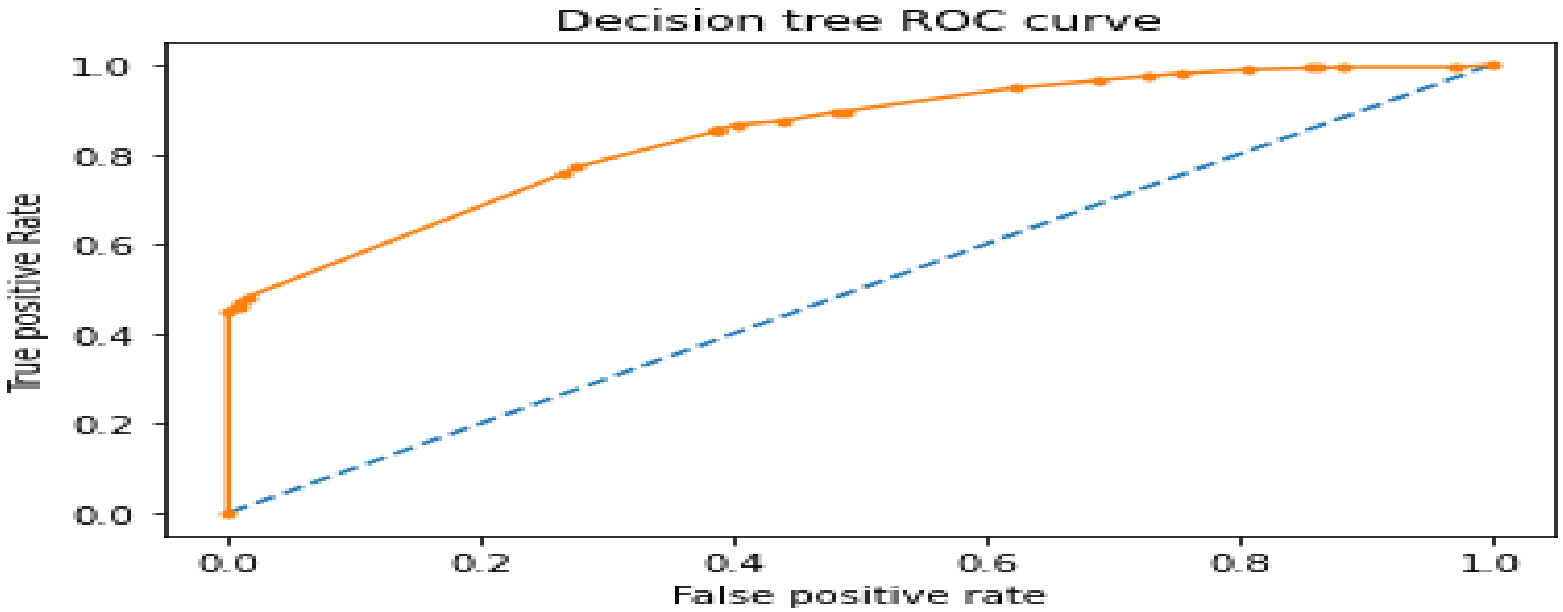
- **Precision : 0.6560**
- **Recall : 0.6633**
- **F1-Score : 0.6594**
- **Accuracy : 0.6761**
- **ROC_AUC : 0.7361**

✓Decision Tree



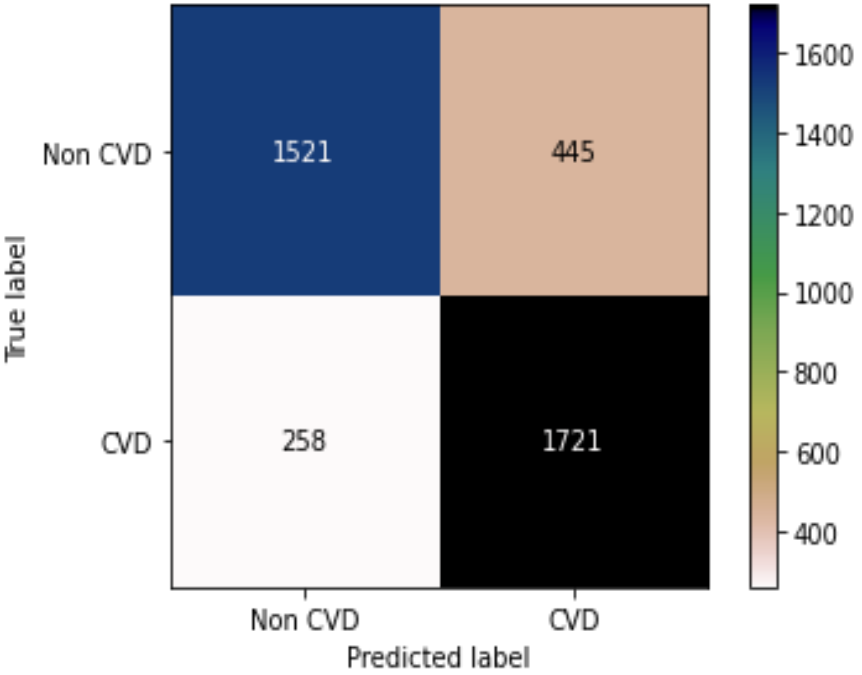


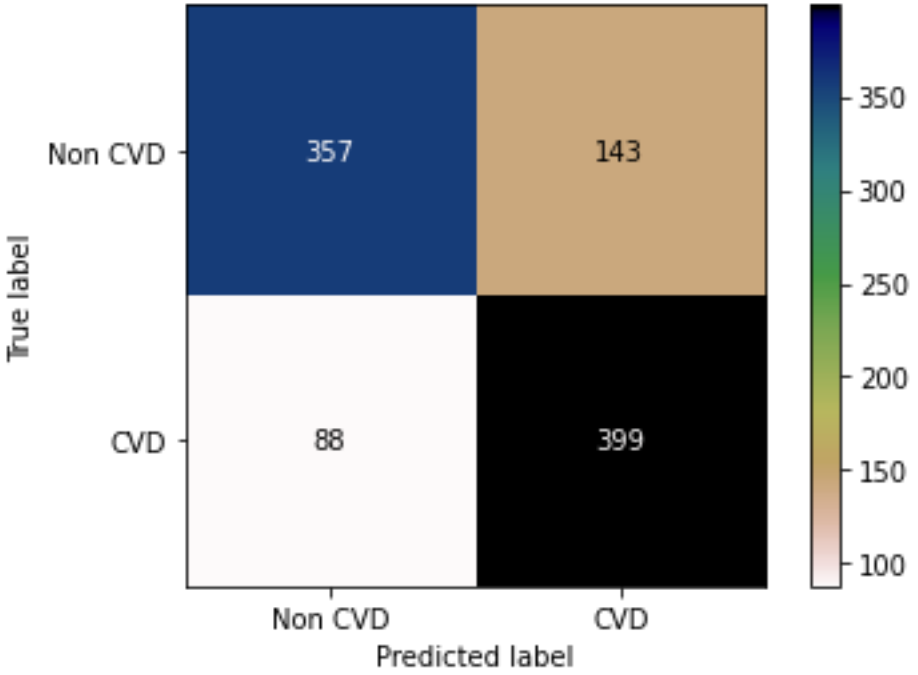
a)Visualizing the confusion matrix to evaluate the performance of the model o
n training set and testing set respectively.



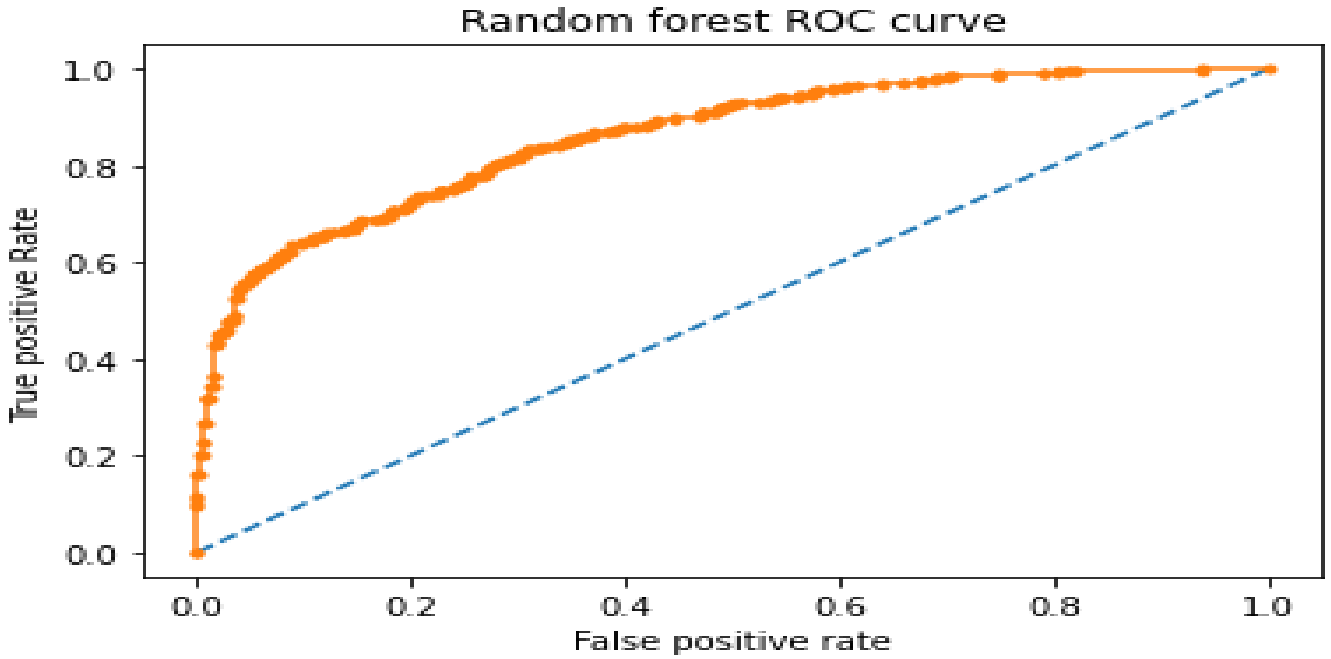
- **Precision : 0.7094**
- **Recall : 0.6789**
- **F1-Score : 0.6938**
- **Accuracy : 0.7034**
- **ROC_AUC : 0.7719**

✓Random Forest Classifier



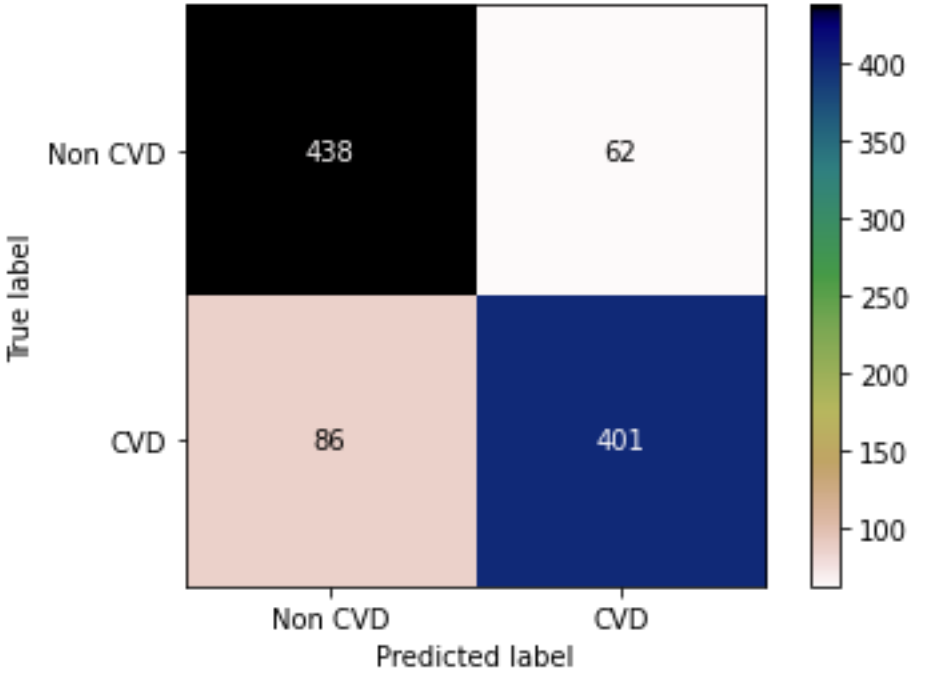


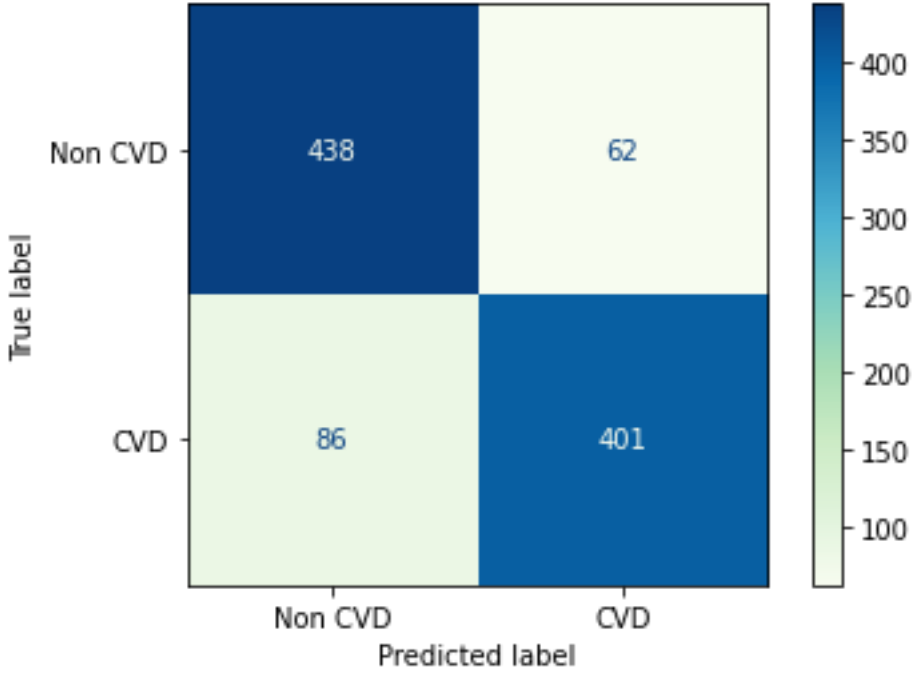
a)Visualizing the confusion matrix to evaluate the performance of the model on trainin
g set and testing set, respectively.



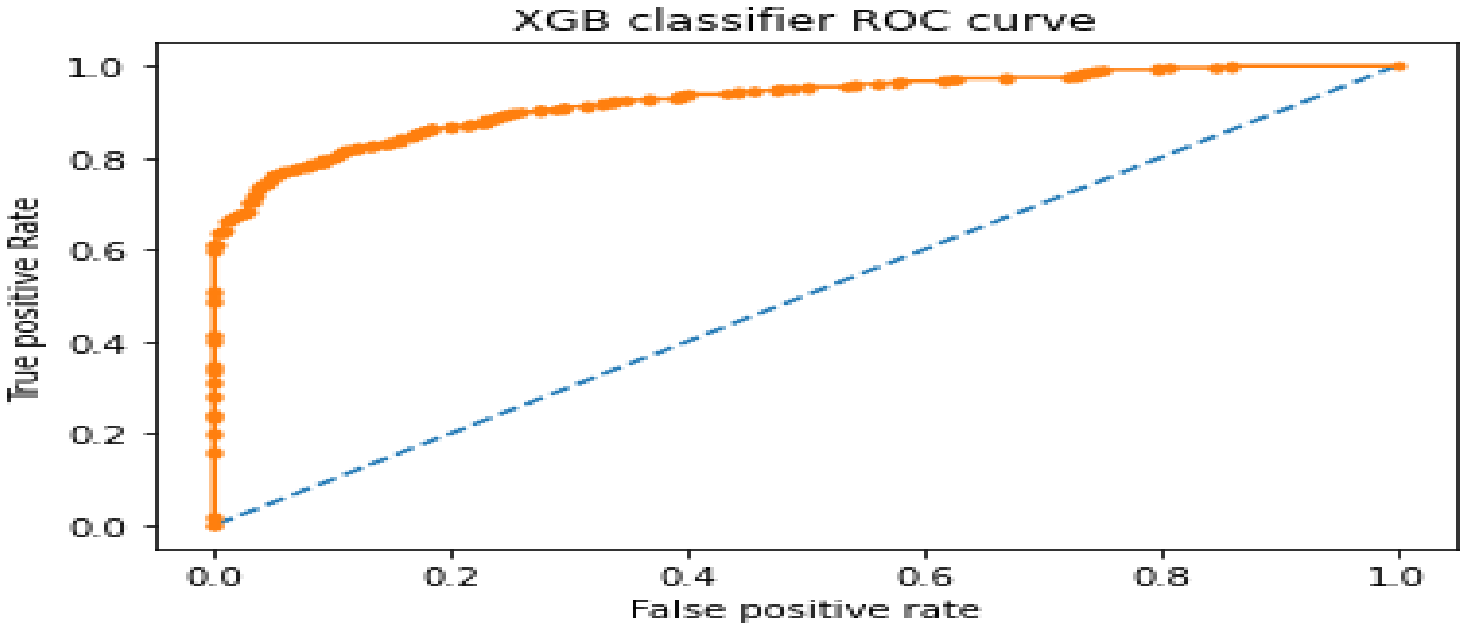
Precision : 0.7294
Recall : 0.8160
F1-Score : 0.7703
Accuracy : 0.7591
ROC_AUC : 0.8611

✓XGB Classifier



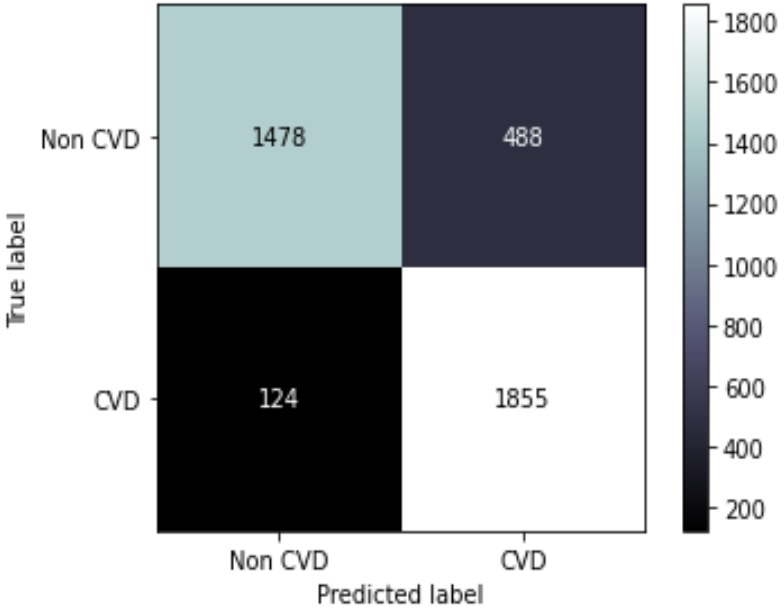


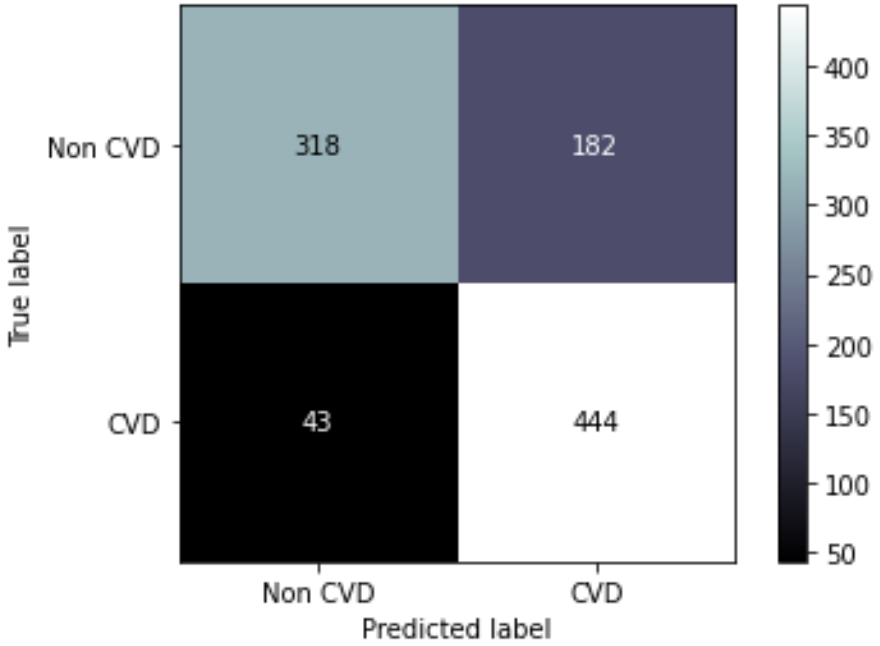
Visualizing the confusion matrix to evaluate the performance of the model on training set and testing set, respectively.



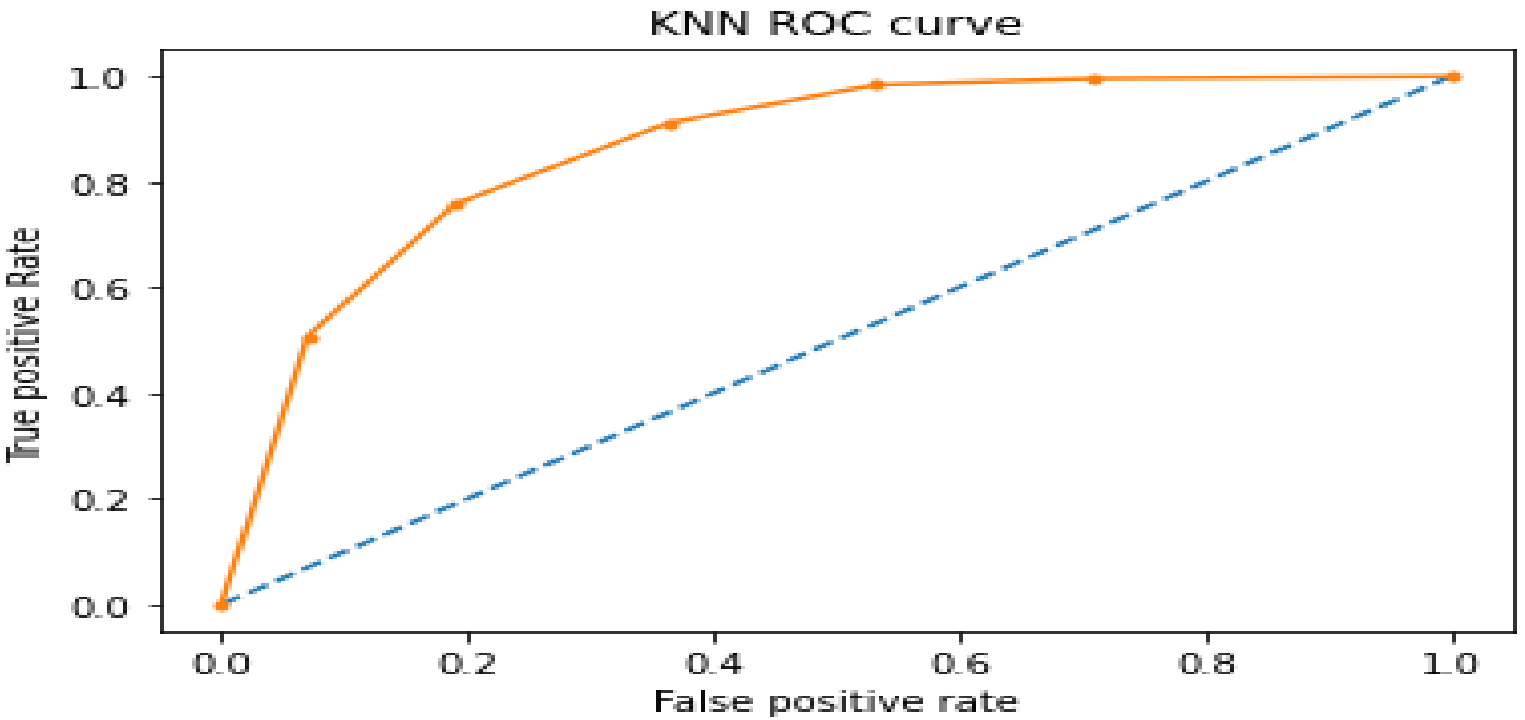
- **Precision : 0.8792**
- **Recall : 0.8037**
- **F1-Score : 0.8397**
- **Accuracy : 0.8482**
- **ROC_AUC : 0.9145**

✓Knn



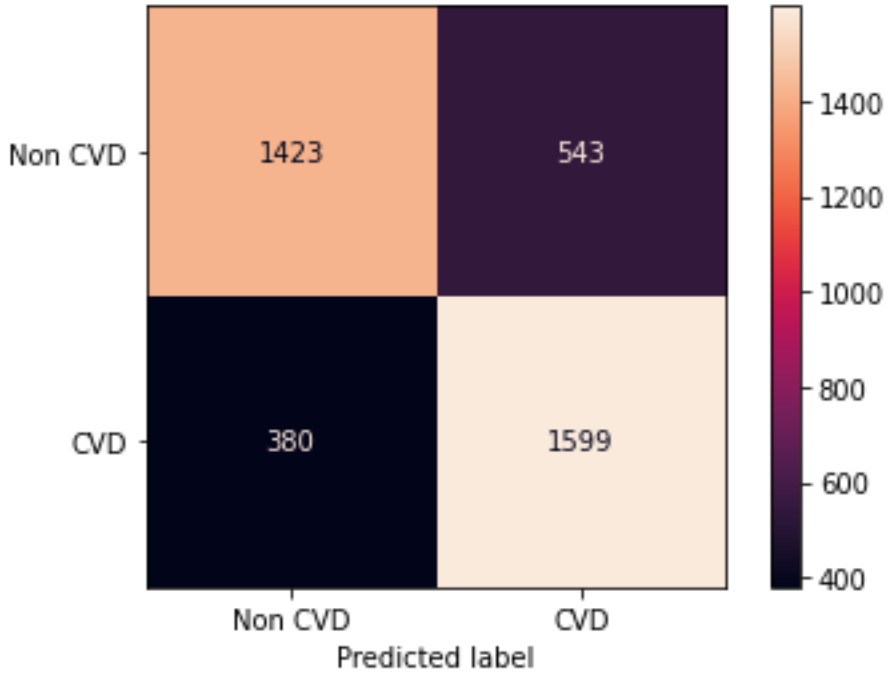


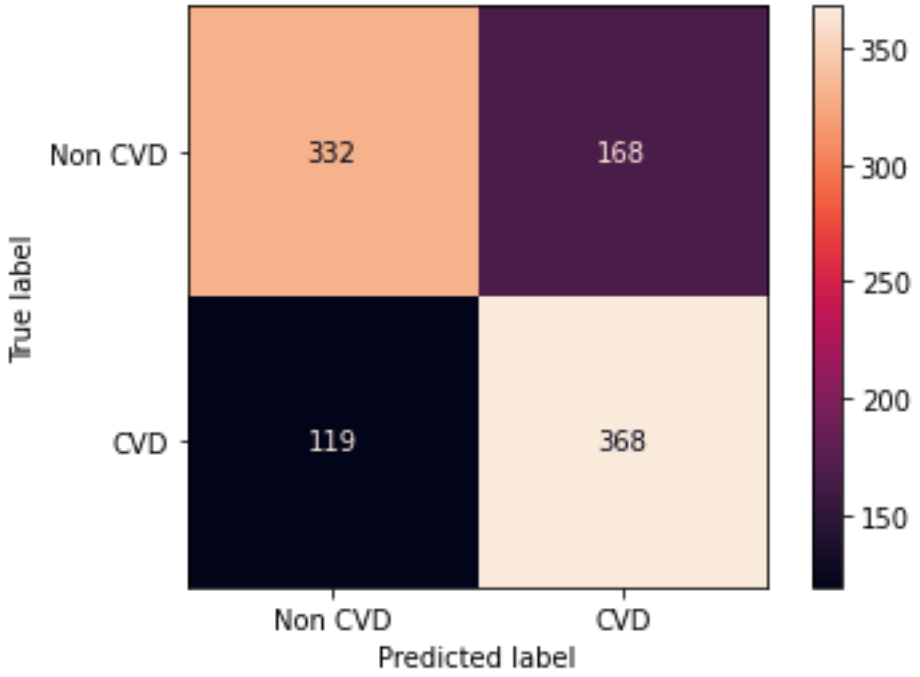
a)Visualizing the confusion matrix to evalua te the performance of the model on trai ning set and testing set, respectively.



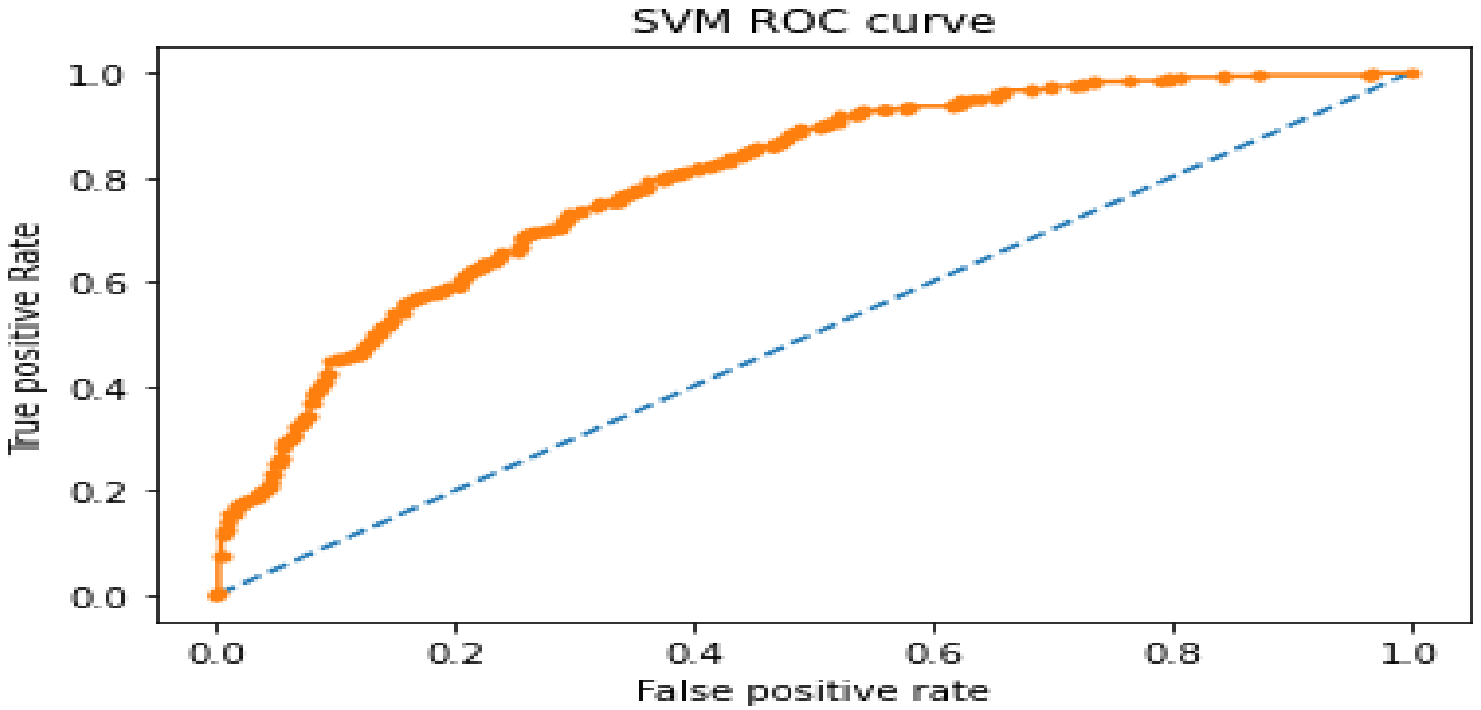
- **Precision : 0.6913**
- **Recall : 0.8609**
- **F1-Score : 0.7668**
- **Accuracy : 0.7409**
- **ROC_AUC : 0.8523**

✓Support Vector Machine





Visualizing the confusion matrix to evaluate the performance of the model on trai ning set and testing set, respectively.



precision : 0.6901
Recall : 0.8016
F1-Score : 0.7417
Accuracy : 0.7237
ROC_AUC : 0.7916

✓Conclusion

- We've noticed that **XBG Classifier** is the standout performer among all models with an f1-score of **0.8397**. it's safe to say that XGB Classifier provides an optimal solution to our problem.
- **KNN** was able to achieve f1-score of **0.7688**. It gave us the best **recall** of **0.8609** out of all the models.
- In case of **Logistic regression**, We were able to see the maximum f1-score of **0.658**.
- The **Random Forest Classifier** was able to achieve an f1-score of **0.7703** and we also noticed that in the case of **Decision-tree Classifier**, we were able to achieve an f1-score of **0.7034** for the test split.
- For **SVM(Support Vector Machines)** Classifier, the f1-score lies around **0.7417**.
- Finally , As in the medical domain (False negative values have importance we don't want to mis predict a person safe when he has the risk) recall has the most importance. KNN, XGB , Random Forest gave the best recall 0.86 ,0.80 ,0.81.
- Considering our data (medical domain) we can conclude that **XGB, KNN and Random Forest this algorithms performed well .**

✓References

https://towardsdatascience.com/
https://www.analyticsvidhya.com/
https://www.geeksforgeeks.org/python-data-visualization-tutorial