Capstone Project - 3

Cardiovascular Risk Prediction

Swapnil Patil

Problem Statement

Cardiovascular disease(CVD) is the leading cause of death worldwide and a major public health concern.

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

Providing a valid model for cardiovascular disease risk classification of each population has become a high priority for scientists and organizations working in this field.

Introduction

Cardiovascular disease (CVD) is a series of diseases involving the circulatory system, including coronary heart disease, heart failure, arrhythmia and others, which is generally related to atherosclerosis.

Heart disease is the leading cause of death in the world. The most common type is coronary heart disease, which can cause a heart attack.

Day by day the cases of Cardiovascular diseases(CVD) are increasing at a rapid rate and it's very important and concerning to predict any such diseases beforehand.

Several risk prediction models of cardiovascular disease have been developed for different populations in the past decade.

In this study, we have provided a prediction model for 10-year risk assessment of Coronary Heart Disease(CHD) to predict whether the patient is likely to be diagnosed with a disease.

Data Overview

- ➢ For the project, the dataset we used is from cardiovascular study on **residents of the town of Framingham, Massachusetts**. It includes over 4,000 records and **15 attributes**.
- ➢ Each attribute is a potential risk factor. There are **demographic, behavioral and medical risk factors**.
- ➢ Target Variable:- **TenYearCHD:** 10-year risk of coronary heart disease CHD

Attributes Information

Demographic-

- **Sex**: male or female.
- **Age**: Age of the patient
- **Education**: education of patient

Medical(history)-

- **BPMeds**: the patient was on blood pressure medication or not
- **PrevalentStroke**: the patient previously had a stroke or not
- **PrevalentHyp**: patient was hypertensive or not
- **Diabetes**: the patient had diabetes or not

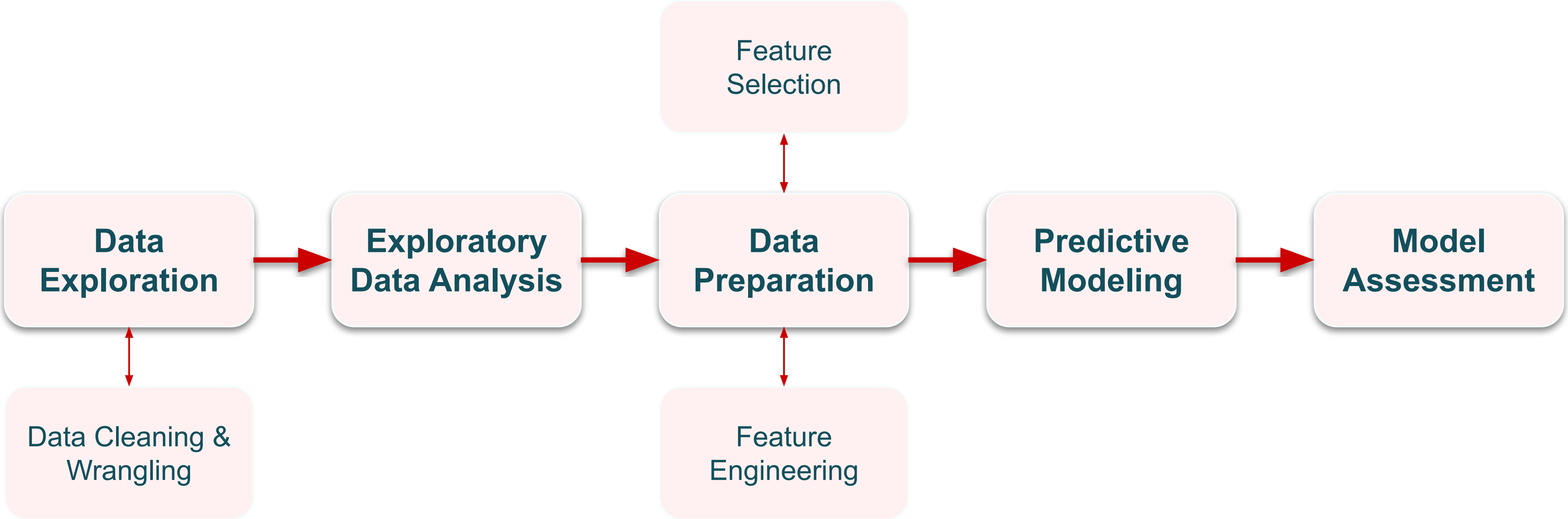
Behavioral-

- **is_smoking**: patient is a current smoker or not
- **CigsPerDay**: the number of cigarettes that the person smoked on average in one day.

Medical(current)-

- **TotChol**: total cholesterol level
- **SysBP**: systolic blood pressure
- **DiaBP**: diastolic blood pressure
- **BMI**: Body Mass Index
- **HeartRate**: heart rate
- **Glucose**: glucose level

Steps Involved



Feature Selection

Data Exploration

Exploratory Data Analysis

Data Preparation

Predictive Modeling

Model Assessment

Data Cleaning & Wrangling

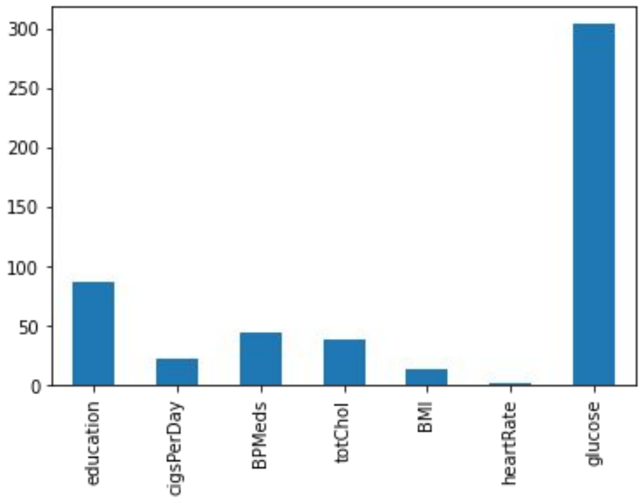
Feature Engineering

Data Exploration

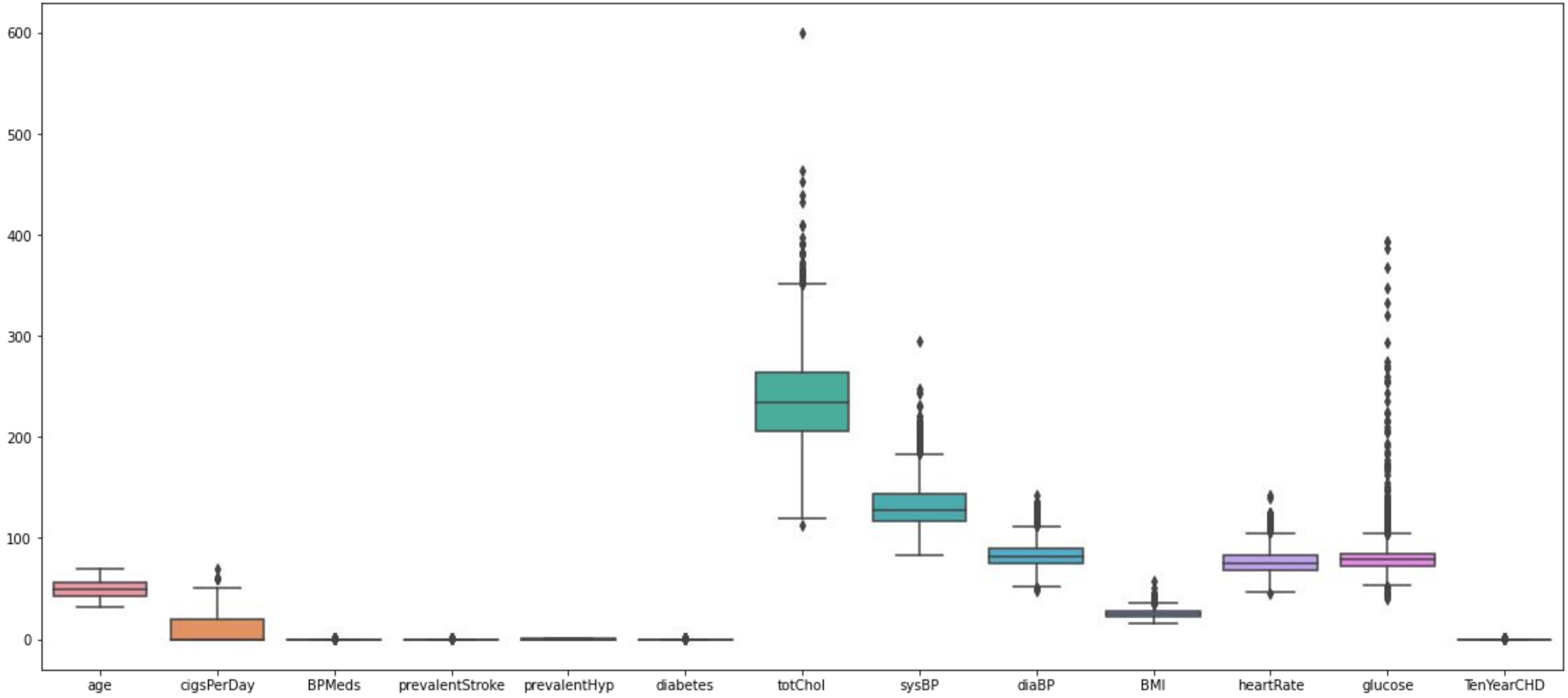
- It includes 3390 patient records and 15 attributes.
- Data consisted of both continuous variables as well as categorical variables.
- Some data features are binary (binary: “1”, means “Yes”, “0” means “No”) such as BPMeds, PrevalentStroke, PrevalentHyp and Diabetes.

Data Cleaning & Wrangling

- There were no duplicate values in the dataset
- 206 missing values in 7 different columns
- glucose had max null values so imputed nan values based on the diabetes column
- Dropped remaining missing values

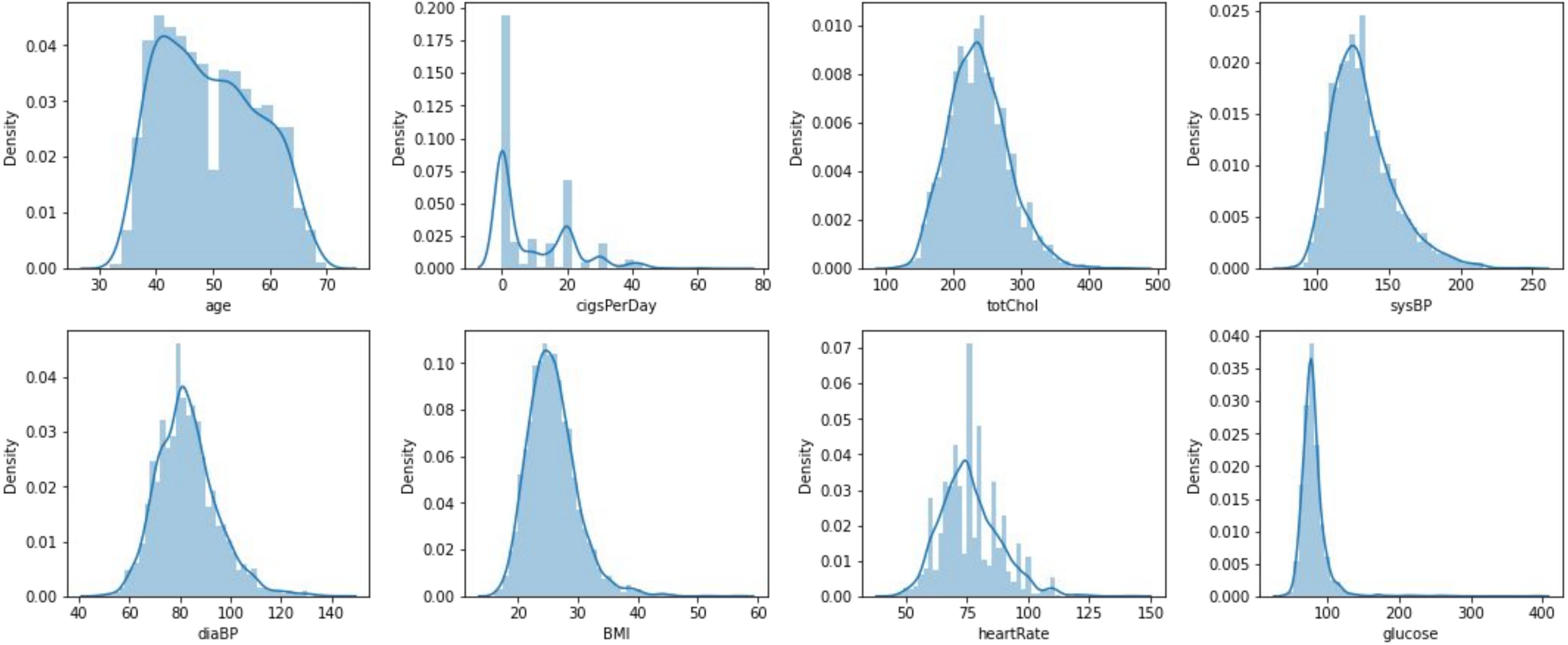


Outliers Treatment



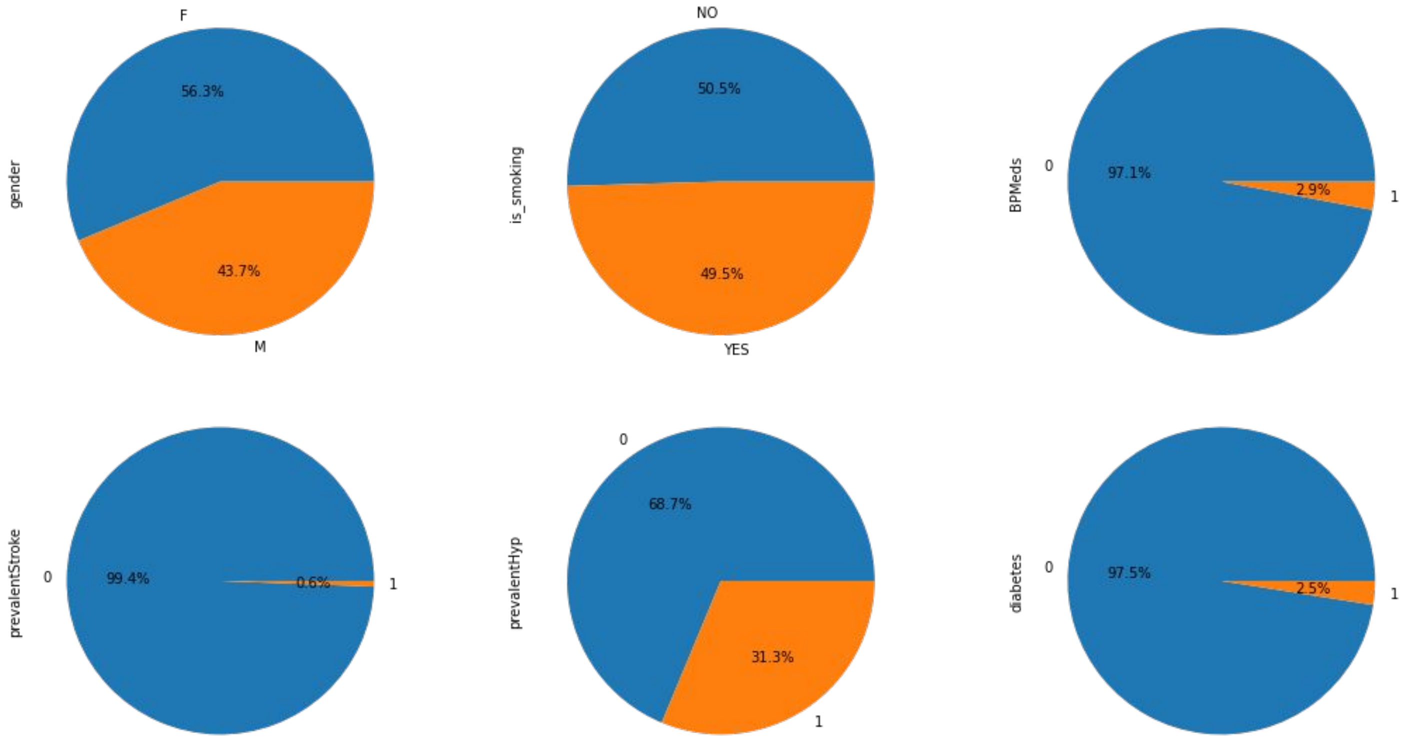
Exploratory Data Analysis - Univariate Analysis

Distribution of Continuous Variables



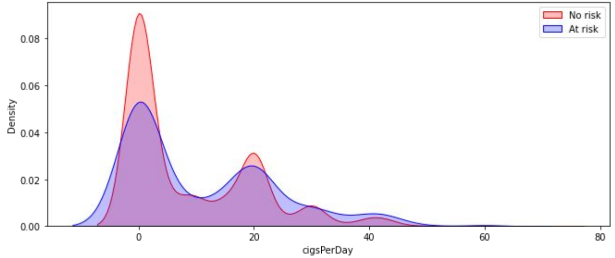
Univariate Analysis contd.

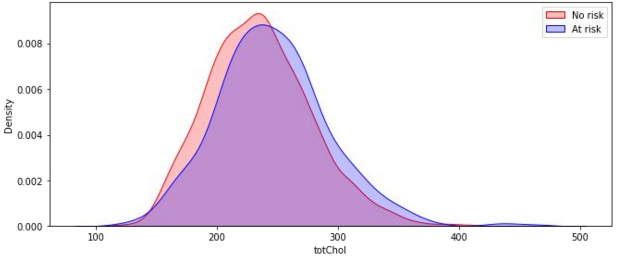
Distribution of Categorical Variables

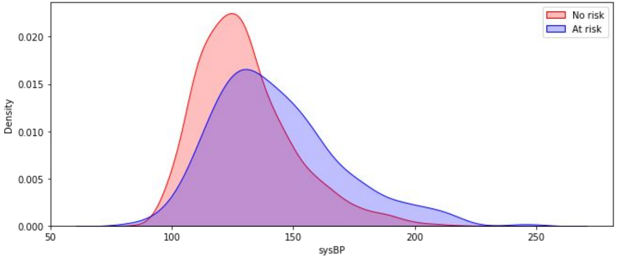


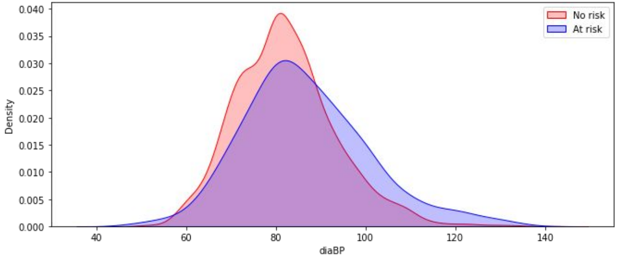
Bivariate Analysis - Continuous Variables

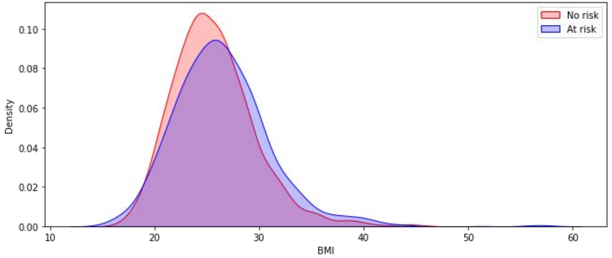
Which continuous variables are higher risk factors

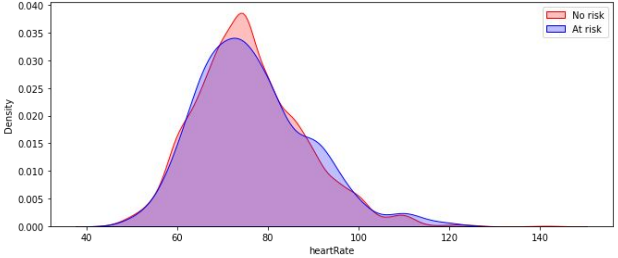


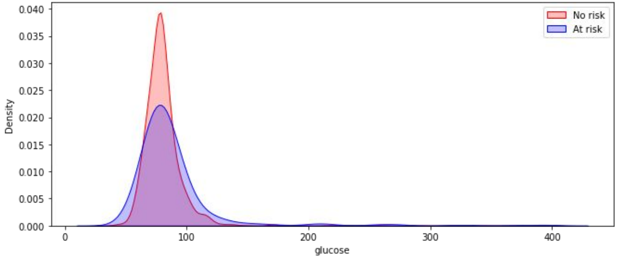






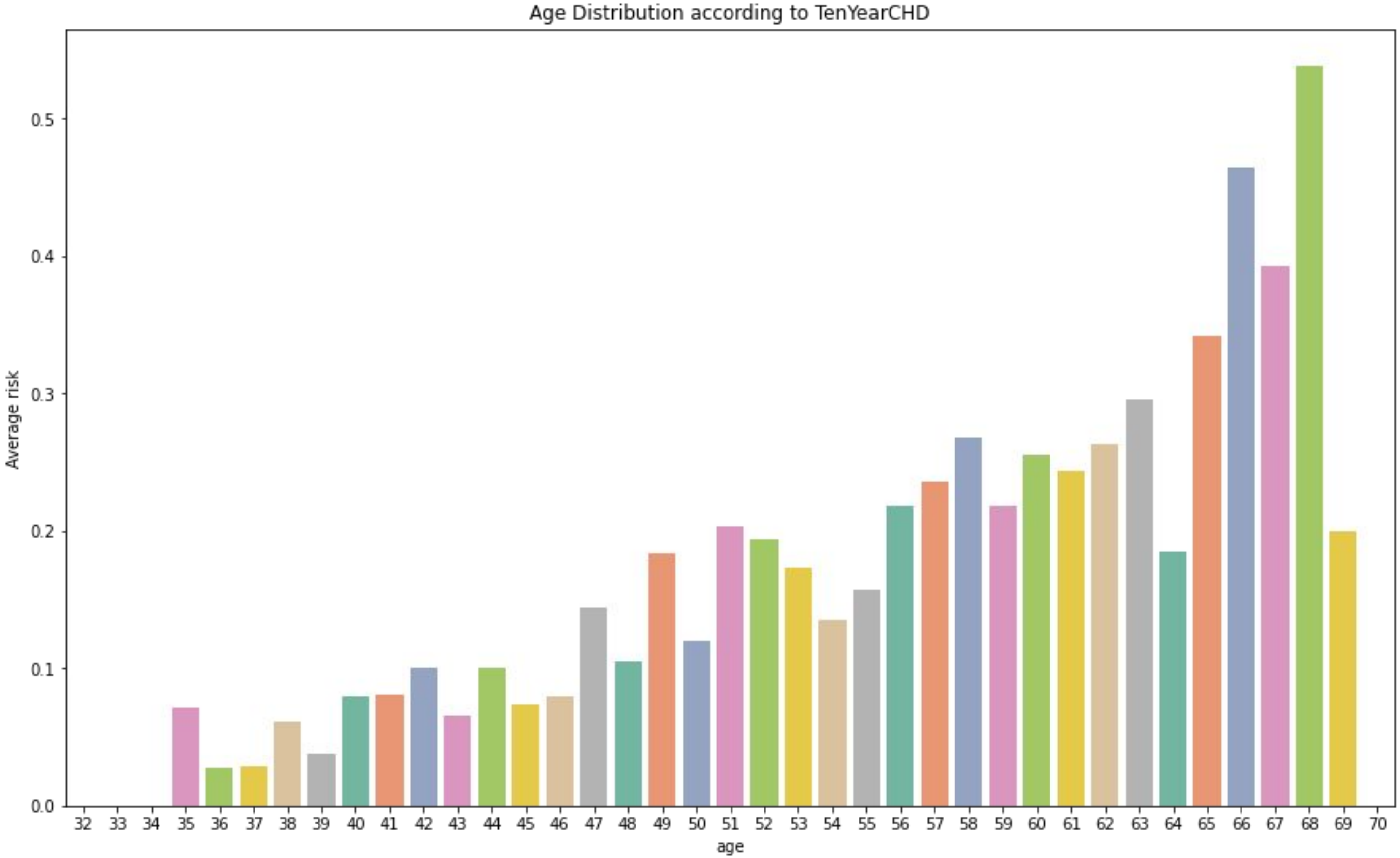






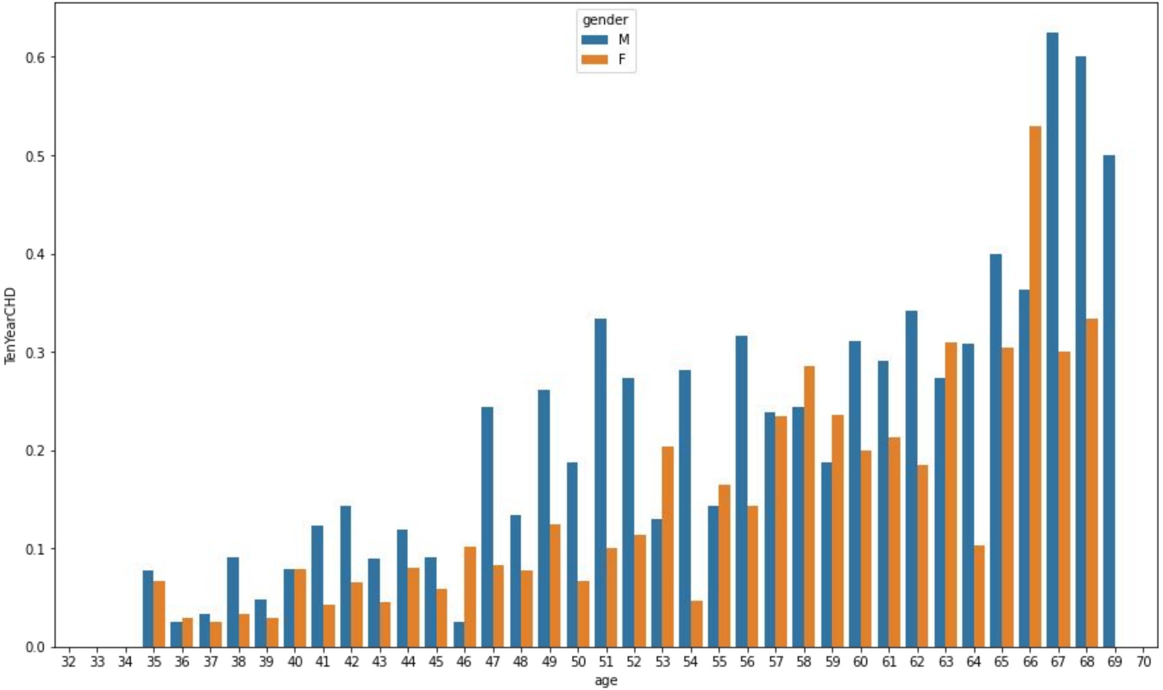
- Patients at risk have a slight increase in cholesterol
- Patients having higher systolic BP tends to have high risk of CHD
- DiaBP is also high for CHD risk patients

At what age the risk of CHD is more

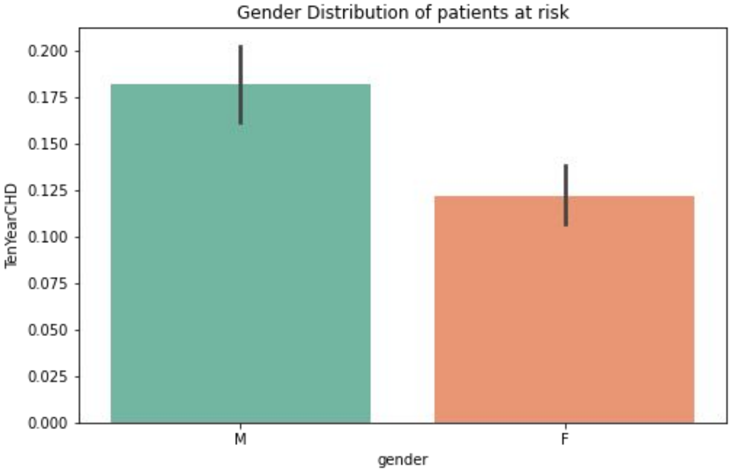


Bivariate Analysis - Categorical Variables

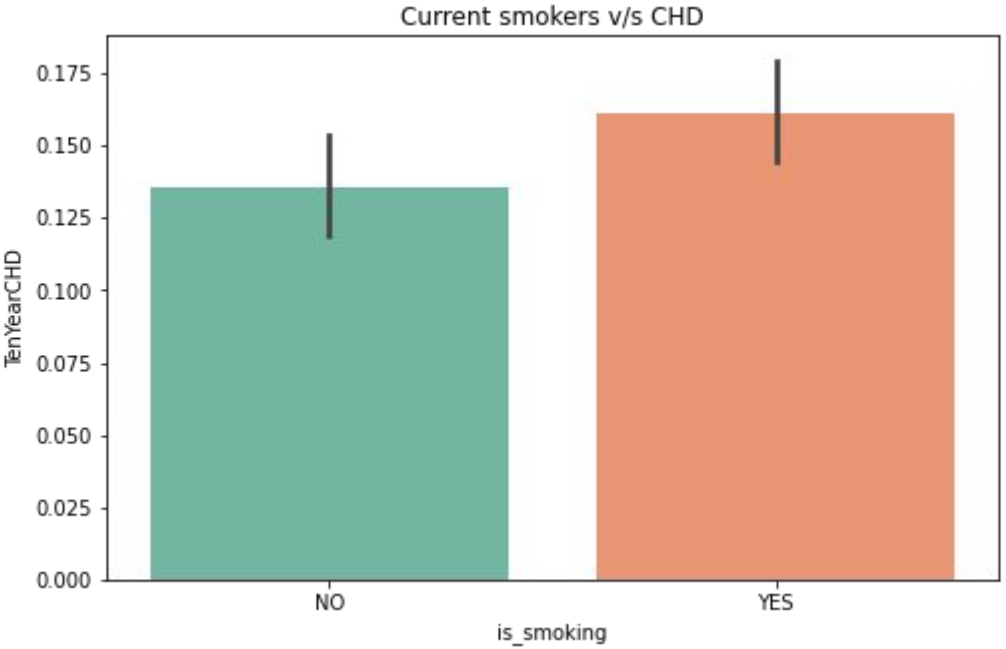
Distribution of age according to gender over TenYearCHD

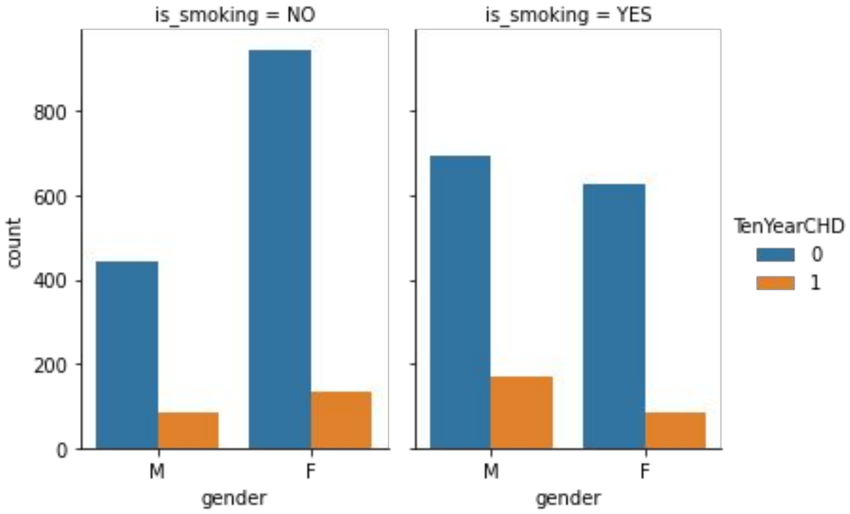


Which gender has most risk of CHD



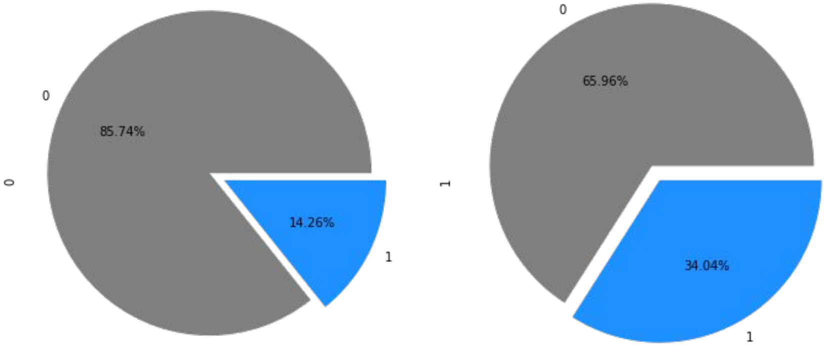
Does smokers have risk of CHD



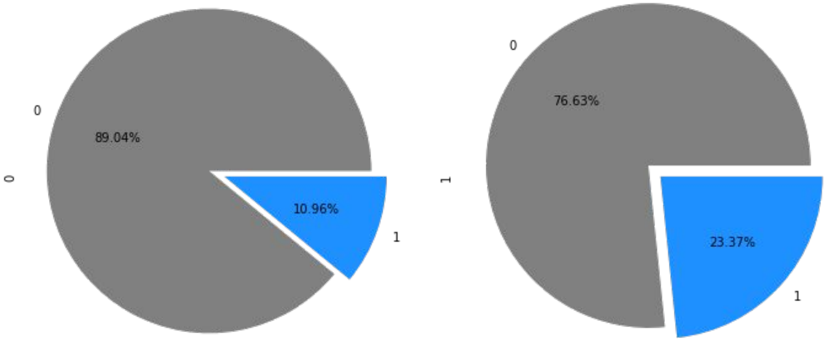


Bivariate Analysis - Binary Variables

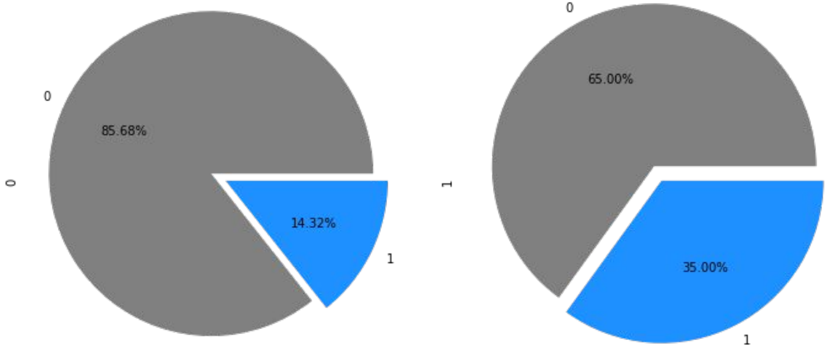
Risk of CHD to a patient on BP medication



Risk of CHD to a patient having hypertension



Risk of CHD to a diabetic patient



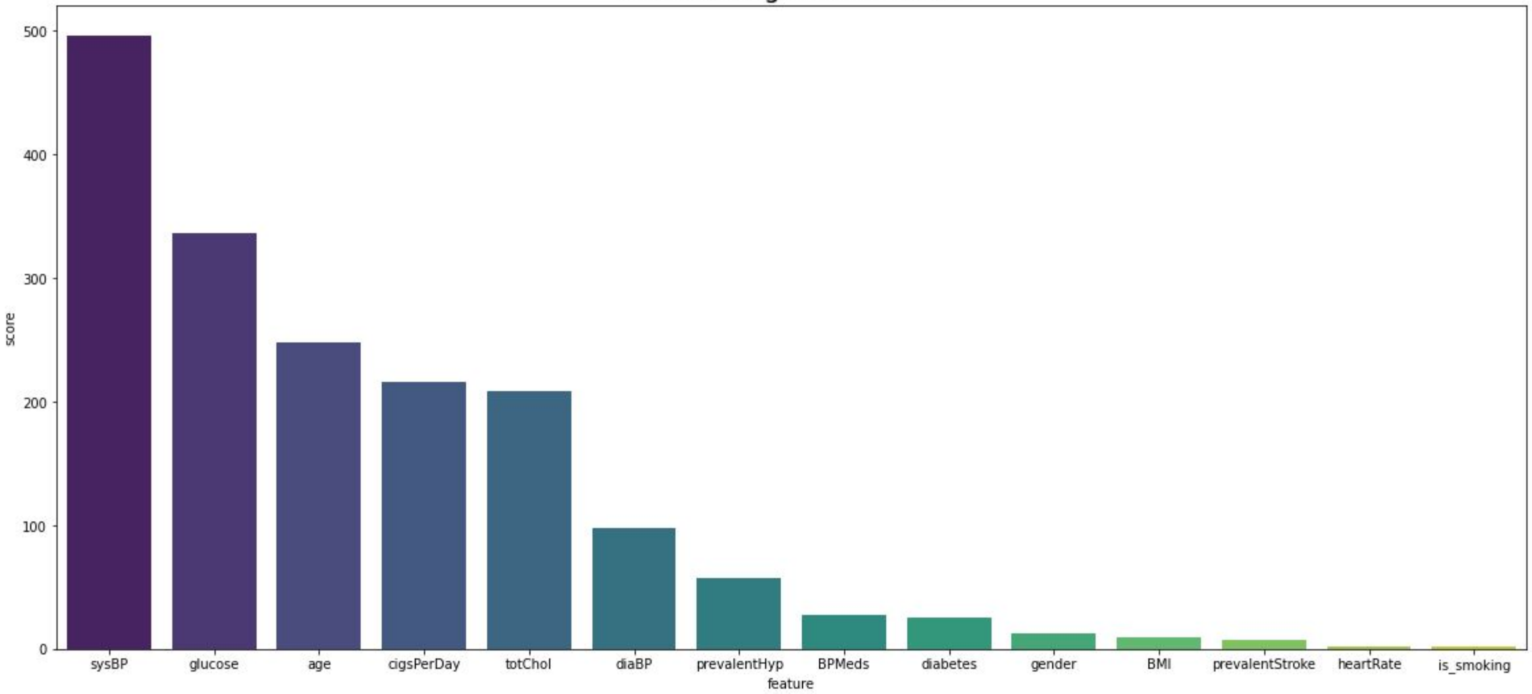
Data Preparation

Feature Engineering - Converting categorical features - gender and is_smoking into binary interpretation.

Feature Selection - SelectKBest method to select top most important features

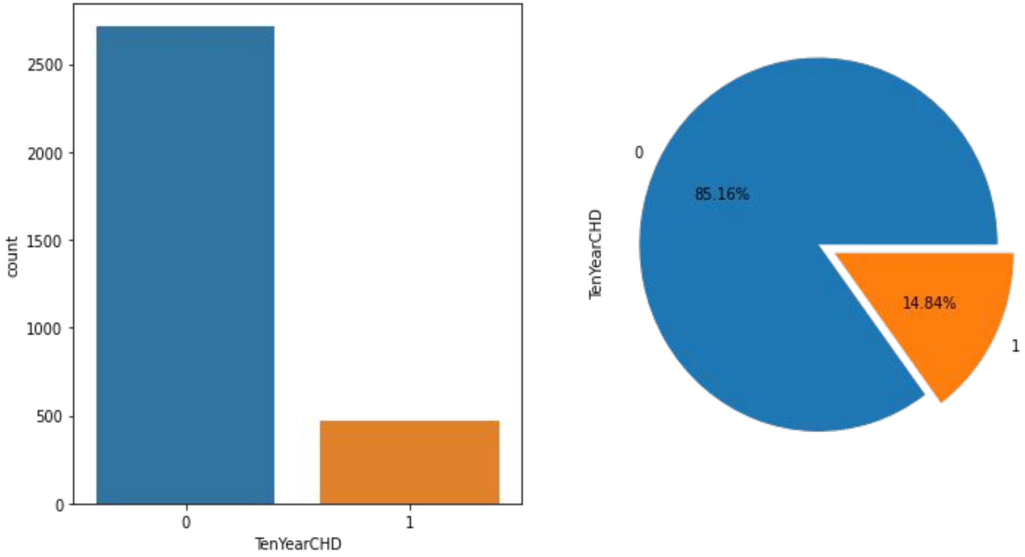
Selecting best 8 features - **sysBP, glucose, age, cigsPerDay, totChol, diaBP, prevalentHyp, gender**

Plot showing Best Features



Analyzing Dependent Variable - TenYearCHD

Distribution of TenYearCHD



Balancing the target variable

Original Dataset

Shape:- (3187, 9)

Class 0: 2714

Class 1: 473

Proportion: 5.74 : 1

After Random-over sampling

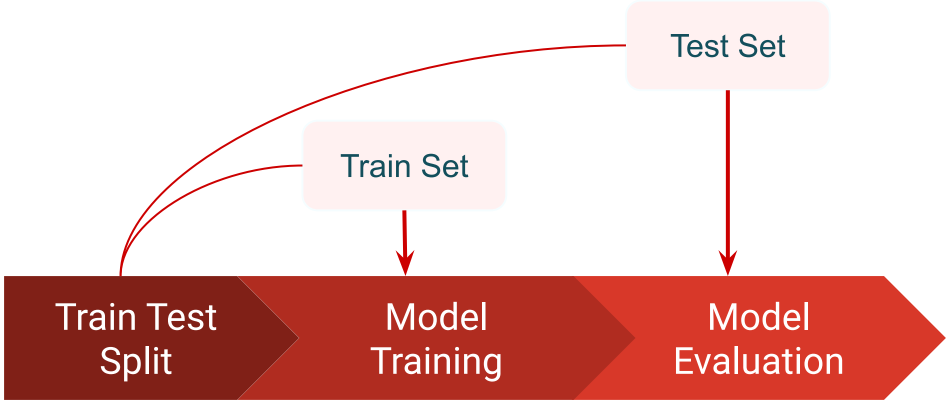
Shape:- (5428, 9)

Class 0: 2714

Class 1: 2714

Proportion: 1 : 1

Predictive Modeling



Classification Models used -

1)Logistic Regression

2)K-Nearest Neighbors

3)Support Vector Classifier

4)Decision Tree Classifier

5)Random Forest Classifier

6)Gradient Boosting Classifier

Predictive modeling includes -

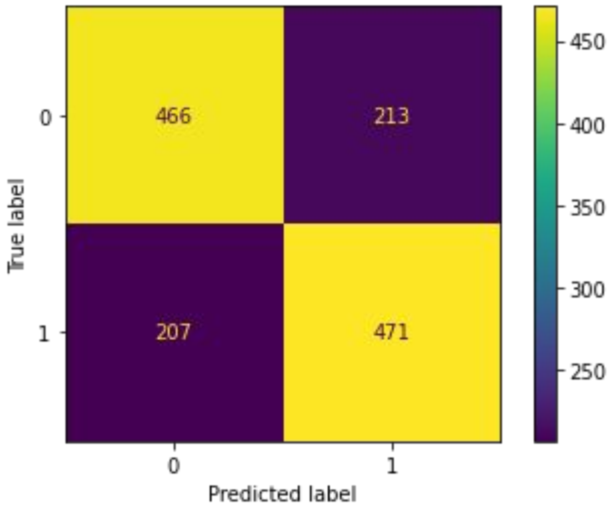
- Building and training the models
- Tuning the hyperparameters to get better performance
- Model Evaluation and Selection

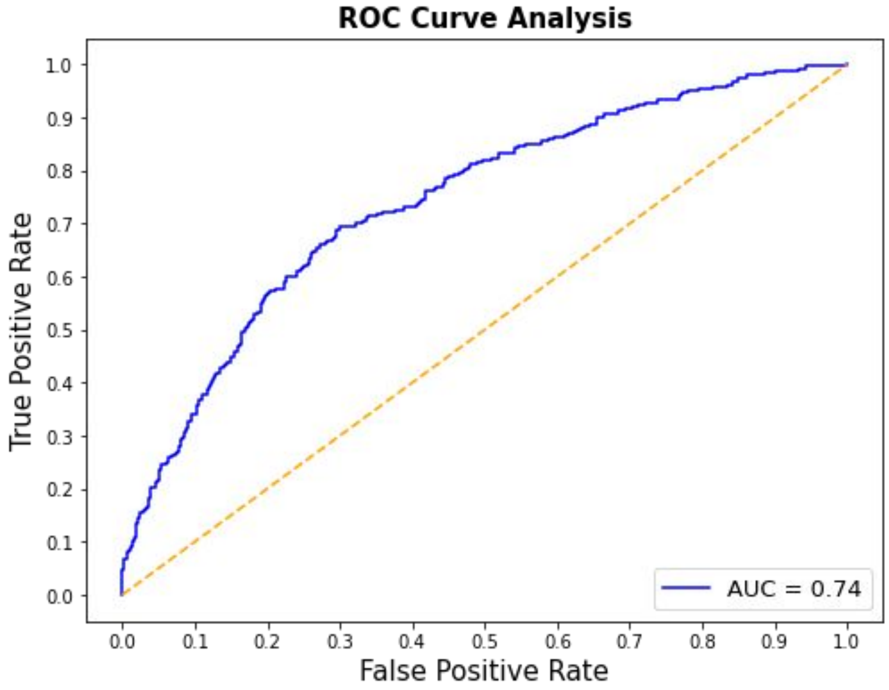
Logistic Regression

Test set Metrics

```
Accuracy  : 0.6904937361827561

Precision : 0.6885964912280702

Recall    : 0.6946902654867256

f1 score  : 0.6916299559471366
```

Confusion Matrix





K-Nearest Neighbors

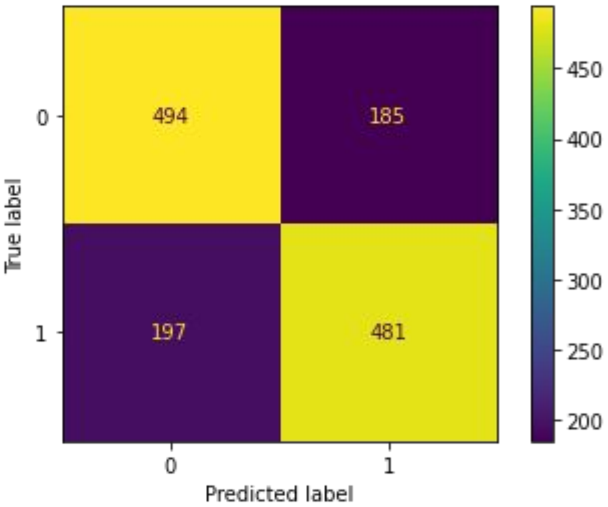
Test set Metrics

```
Accuracy   : 0.7184966838614592
Precision  : 0.7222222222222222
Recall     : 0.7094395280235988
f1 score   : 0.7157738095238095
```

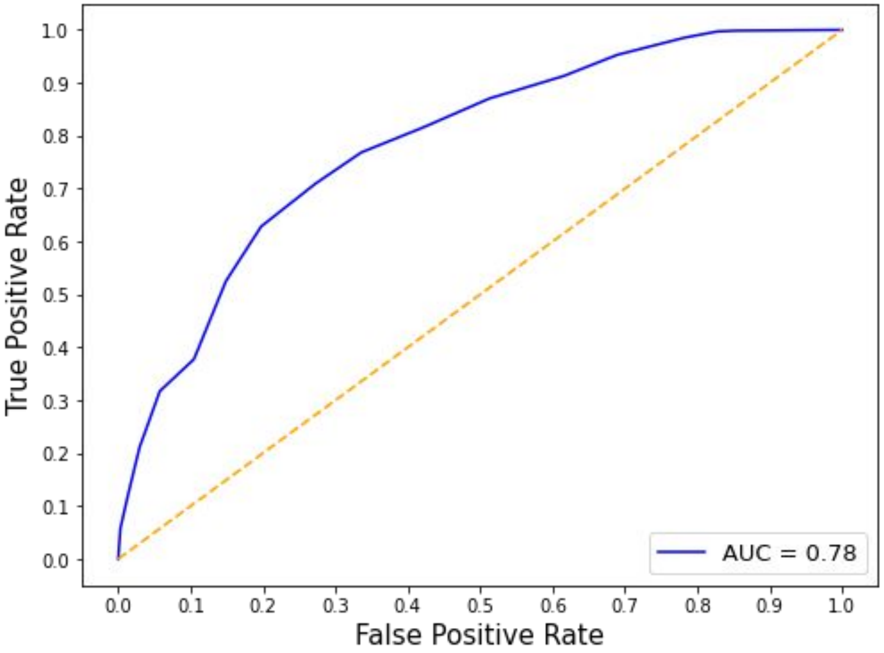
Model

GridSearchCV(cv=5, estimator=KNeighborsClassifier(),
 param_grid={'n_neighbors': array([**16**, 17, 18, 19, 20, 21, 22, 23,
24, 25, 26, 27, 28, 29, 30])})

Confusion Matrix



ROC Curve Analysis

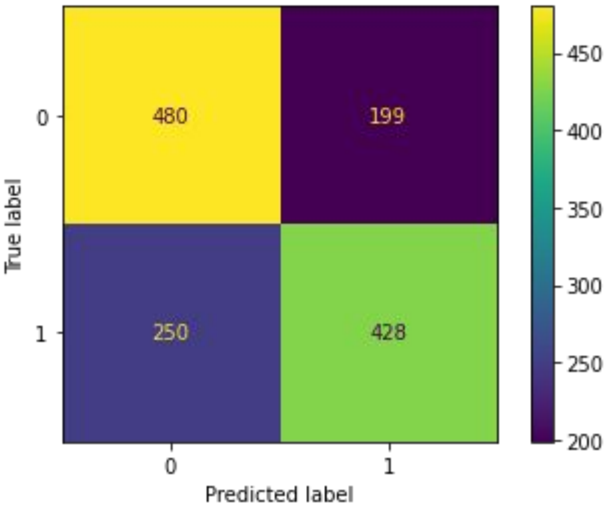


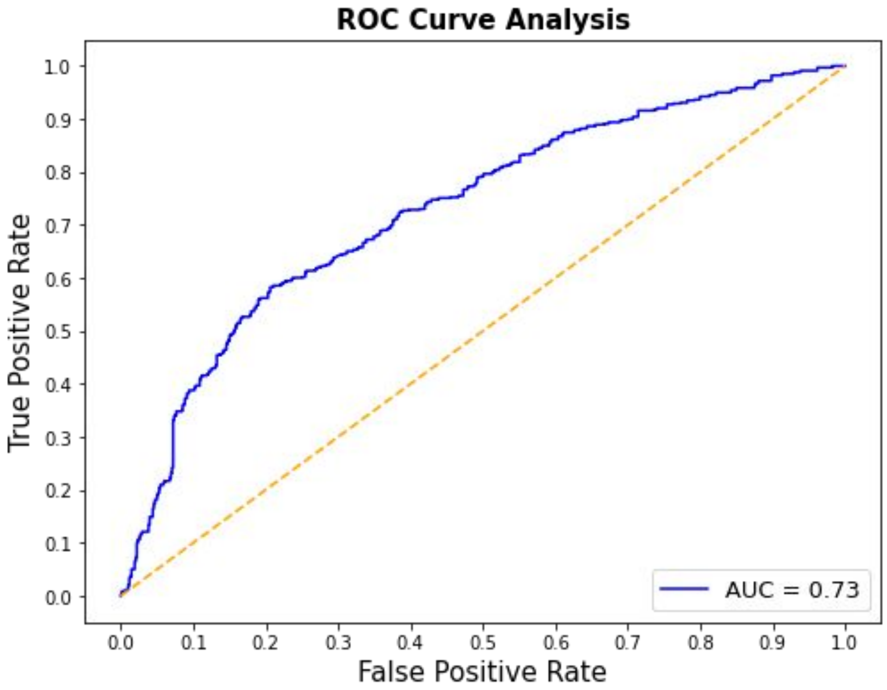
Support Vectors Classifier

Test set Metrics

```
Accuracy  : 0.6691230655858511

Precision : 0.682615629984051

Recall    : 0.6312684365781711

f1 score  : 0.6559386973180076
```

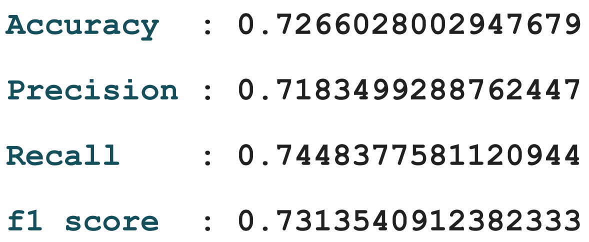
Confusion Matrix





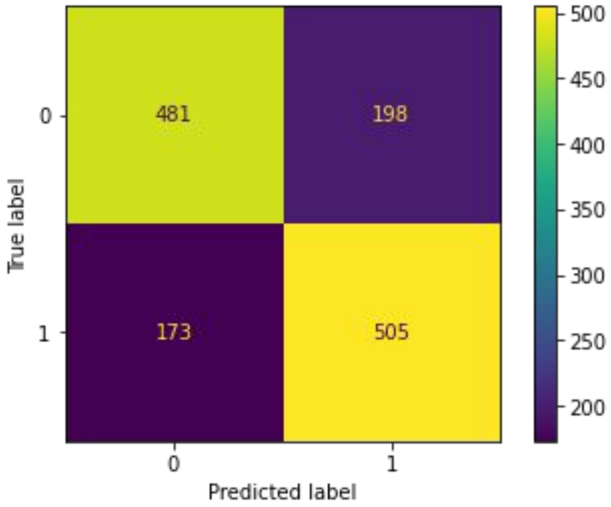
Decision Tree Classifier

Test set Metrics



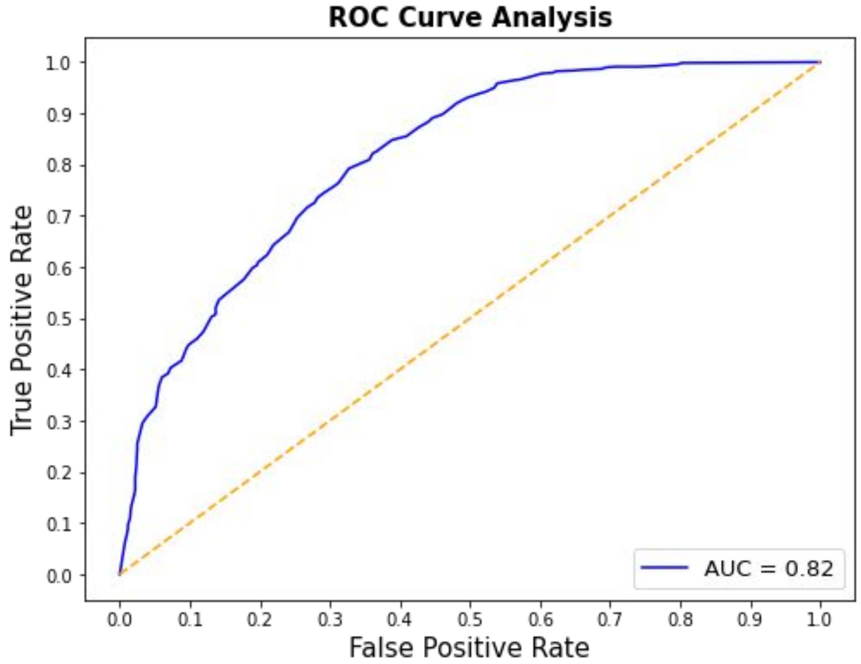
```
Accuracy  : 0.7266028002947679
Precision : 0.7183499288762447
Recall    : 0.7448377581120944
f1 score  : 0.7313540912382333
```

Confusion Matrix



Model

```
GridSearchCV(cv=5, estimator=DecisionTreeClassifier(random_state=0),
          param_grid={'criterion': ['gini', 'entropy'], 'max_depth': [10, 15, 20, 25],
                          'min_samples_leaf': [30, 40, 50]},
          scoring='accuracy', verbose=3)
```



Random Forest Classifier

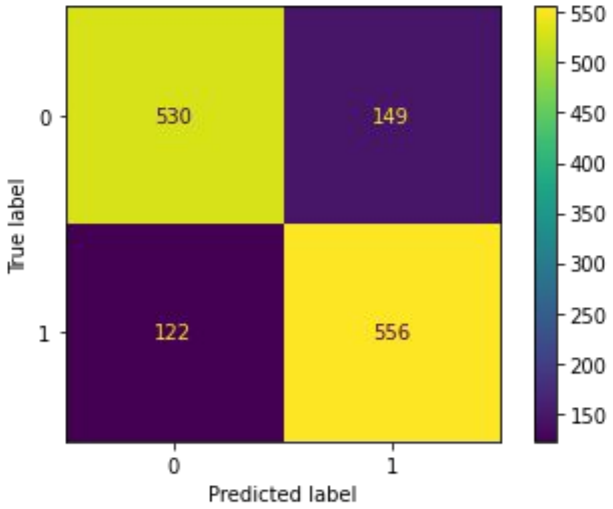
Test set Metrics

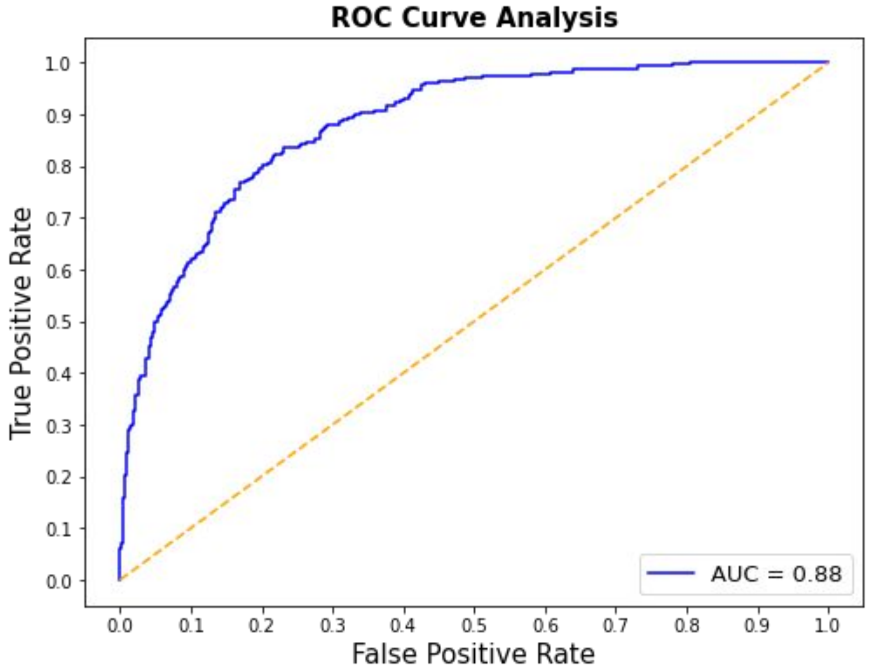
```
Accuracy  : 0.8002947678703022
Precision : 0.7886524822695036
Recall    : 0.8200589970501475
f1 score  : 0.8040491684743312
```

Model

```
RandomizedSearchCV(cv=3, estimator=RandomForestClassifier(random_state=0),
                   param_distributions={'max_depth': [3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15],
                                        'min_samples_leaf': [10, 20, 30],
                                        'n_estimators': [100, 150, 200]}, verbose=2)
```

Confusion Matrix





Gradient Boosting Classifier

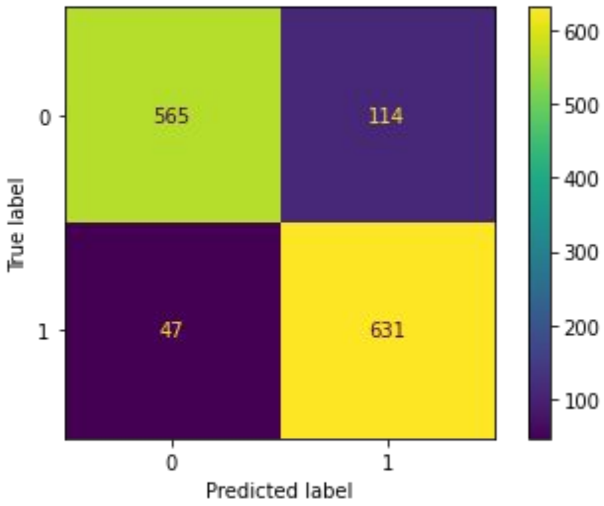
Test set Metrics

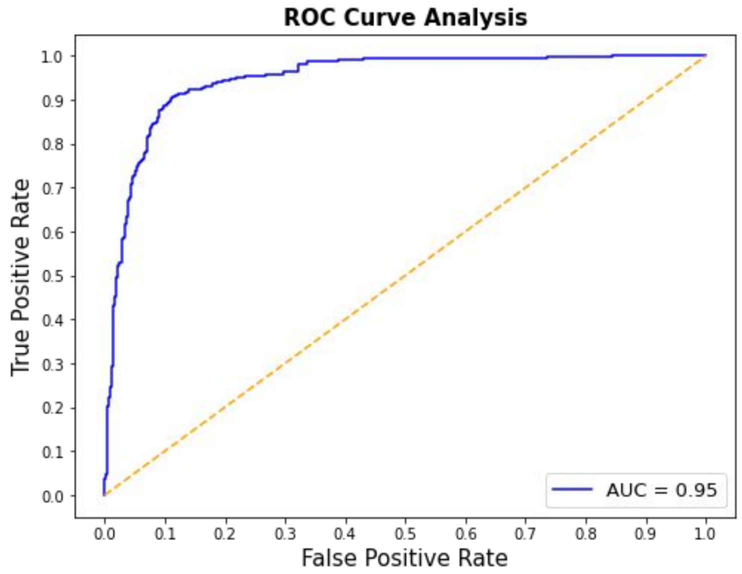
```
Accuracy  : 0.8813559322033898
Precision : 0.8469798657718121
Recall    : 0.9306784660766961
f1 score  : 0.886858749121574
```

Model

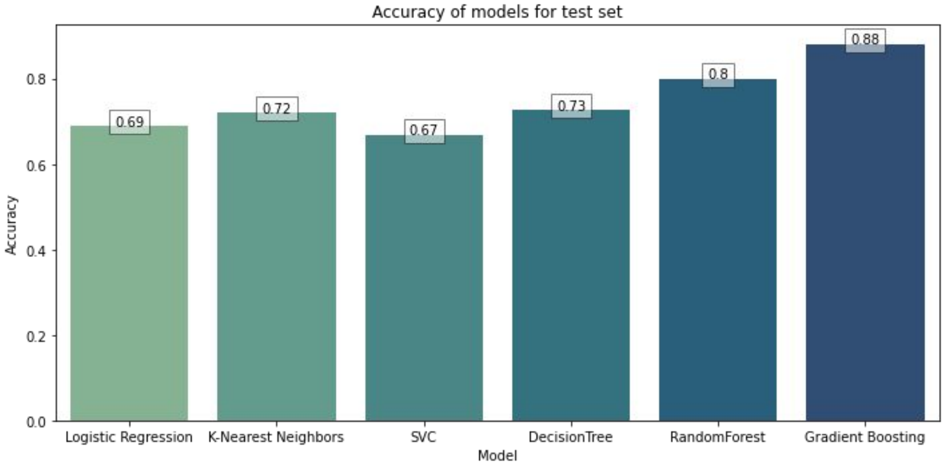
```
RandomizedSearchCV(cv=5, estimator=GradientBoostingClassifier(learning_rate=0.2,
                                                              max_depth=5, n_estimators=150, random_state=0,
                                                              subsample=0.5),
                   param_distributions={'max_features': range(5, 9),
                                        'min_samples_leaf': range(20, 41, 10),
                                        'min_samples_split': range(30, 51, 10)},
                   scoring='roc_auc', verbose=2)
```

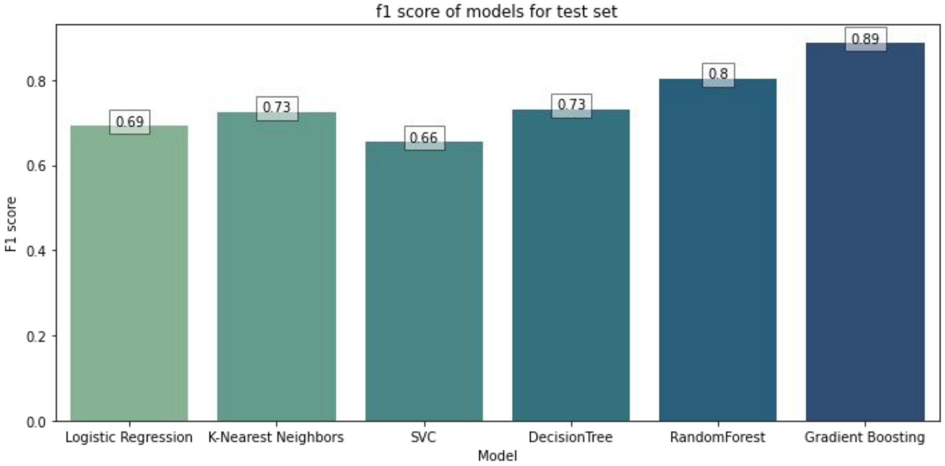
Confusion Matrix



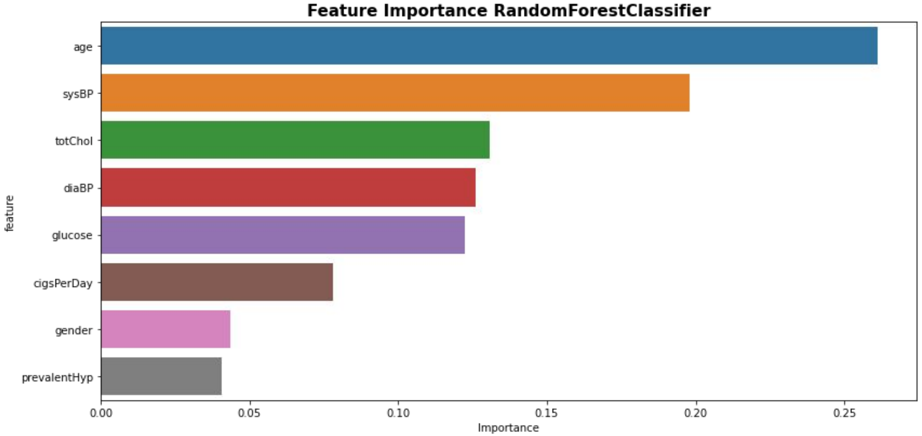


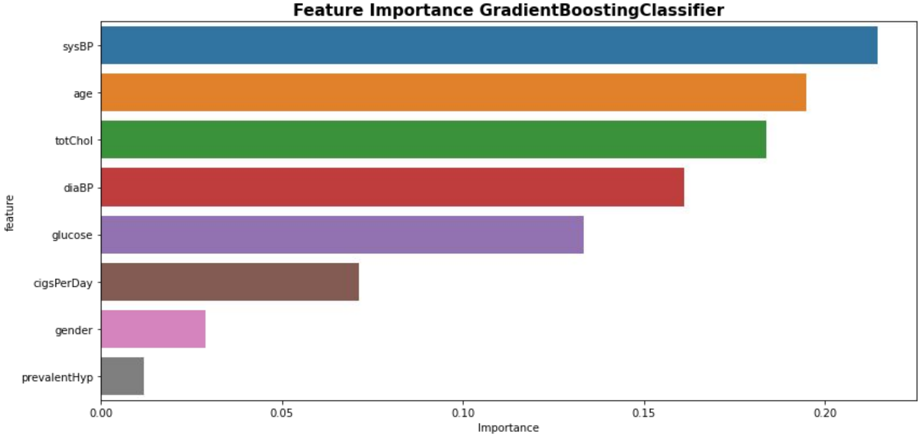
Result



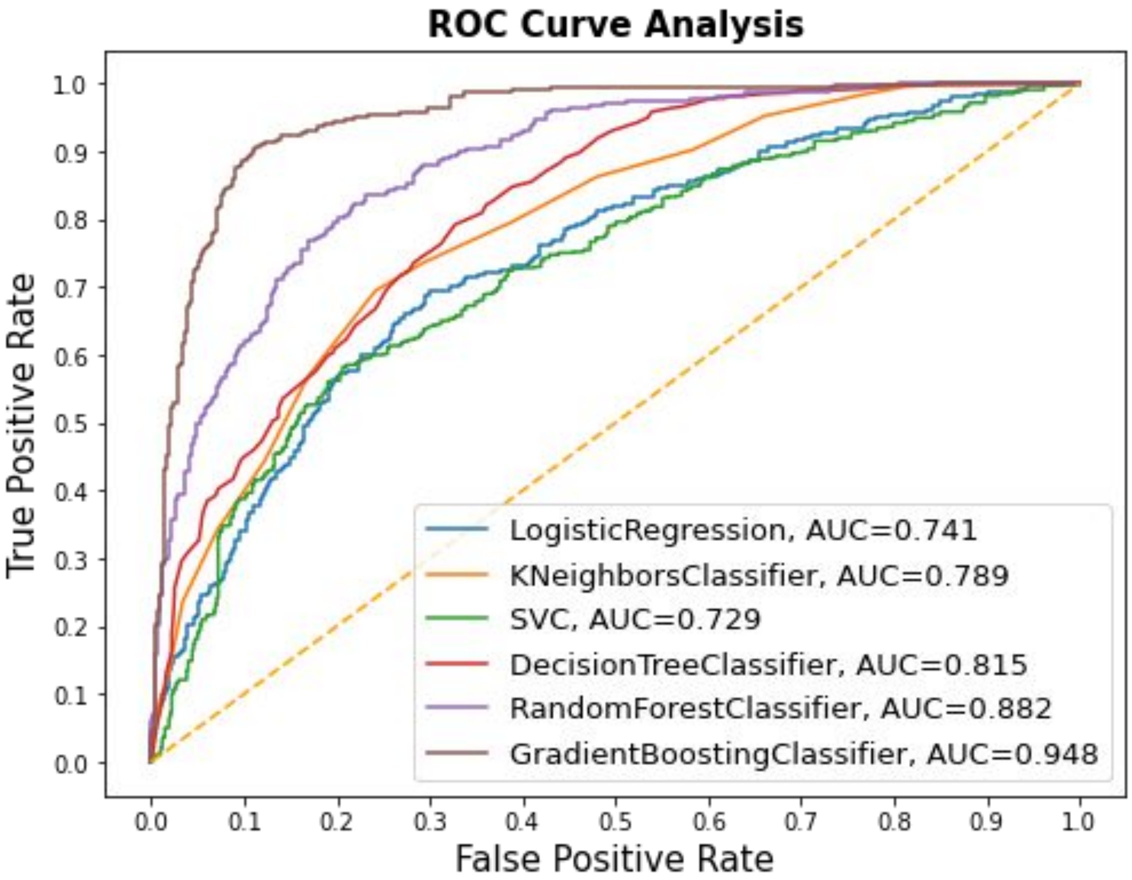


Feature Importance





ROC Curve Analysis



Challenges

- It was health care related dataset so handling the data was so sensitive.
- Imputing the missing values
- Outliers treatment was so challenging as higher values were also important
- Target variable was imbalance which could have affected fitting our models so balancing was important
- Some of the features were not so important
- Tuning the hyperparameters of models and fitting models

Conclusion

- Patients having higher sysBP and DiaBP tends to have high risk of Coronary heart disease.
- The peoples of age above 55 have high risk of contracting disease and the risk increases with age. Men have more risk than females.
- Age, systolic BP and total cholesterol are the most influential features.
- Gradient Boosting model is found to be the best model.
- Therefore, Gradient Boosting model can be used to predict whether the patient will contract Coronary Heart Disease in next 10-years.

Q & A