Capstone Project

Cardiovascular diseases Risk Prediction

By,
Rochak P V
rochakr4@gmail.com
cohort Hardeol

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

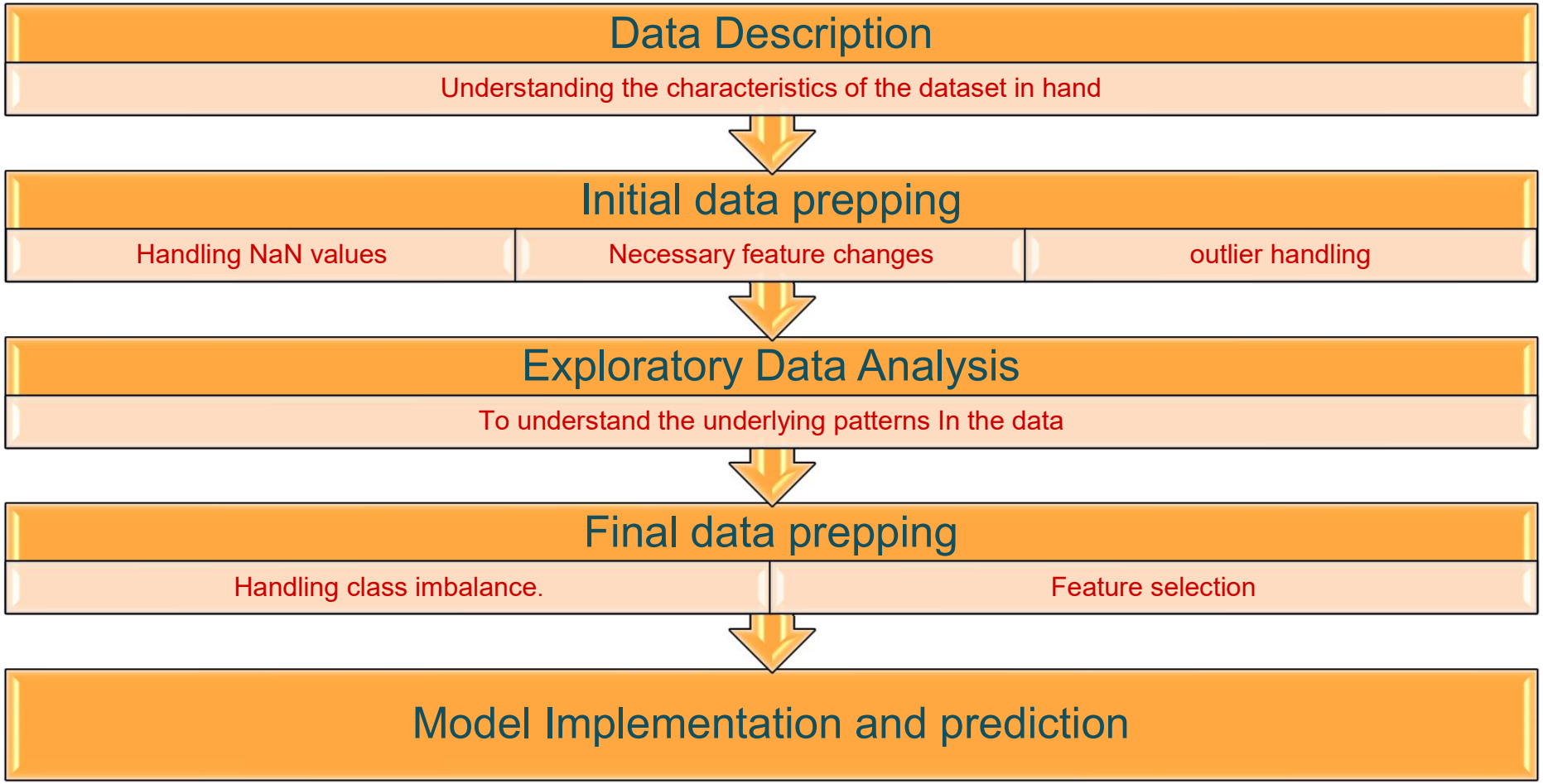
Objective

As mentioned in the problem statement, the objective of this project is to develop a supervised machine learning classification model that can predict the target variable (10-year risk of coronary heart disease CHD), with the help of given feature variables

Methodology

The Classification algorithm is a Supervised Learning technique that is used to identify the category of new observations on the basis of training data. In Classification, a program learns from the given dataset or observations and then classifies new observations into a number of classes or groups. Such as, Yes or No, 0 or 1, Spam or Not Spam, cat or dog, etc. Classes can be called targets/labels or categories

Steps involved



Data Description: Understanding the characteristics of the dataset in hand

The dataset contains information about demographic aspects i.e., sex and age, behavioral aspects i.e., whether a smoker, if so how many cigarettes does he/she smokes a day, medical history i.e., BP medication, prevalent stroke, prevalent hypertension, diabetes and current medical status i.e., total cholesterol level, systolic blood pressure, diastolic blood pressure, BMI, heart rate, glucose level

Column Information:

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)

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).

BP Meds : whether or not the patient was on blood pressure medication.

Prevalent Stroke : whether or not the patient had previously had a stroke.

Prevalent Hyp : whether or not the patient was hypertensive.

Diabetes : whether or not the patient had diabetes.

Tot Chol : total cholesterol level.

Sys BP : systolic blood pressure

Dia BP : diastolic blood pressure

BMI : Body Mass Index

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

Glucose : glucose level

TenYearCHD :(binary: “1”, means “Yes”, “0” means “No”)-Target feature

Libraries used:

NumPy, Pandas, Seaborn, Plotly, Matplot and Scikit Learn:

	age	totChol	cigsPerDay	sysBP	diaBP	BMI	heartRate	glucose
count	3390.000000	3352.000000	3368.000000	3390.00000	3390.000000	3376.000000	3389.000000	3086.000000
mean	49.542183	237.074284	9.069477	132.60118	82.883038	25.794964	75.977279	82.086520
std	8.592878	45.247430	11.879078	22.29203	12.023581	4.115449	11.971868	24.244753
min	32.000000	107.000000	0.000000	83.50000	48.000000	15.960000	45.000000	40.000000
25%	42.000000	206.000000	0.000000	117.00000	74.500000	23.020000	68.000000	71.000000
50%	49.000000	234.000000	0.000000	128.50000	82.000000	25.380000	75.000000	78.000000
75%	56.000000	264.000000	20.000000	144.00000	90.000000	28.040000	83.000000	87.000000
max	70.000000	696.000000	70.000000	295.00000	142.500000	56.800000	143.000000	394.000000

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3390 entries, 0 to 3389
Data columns (total 16 columns):
 #   Column           Non-Null Count  Dtype  
---  ------           --------------  -----  
 0   age              3390 non-null   int64  
 1   education        3303 non-null   float64
 2   sex              3390 non-null   object 
 3   is_smoking       3390 non-null   object 
 4   cigsPerDay       3368 non-null   float64
 5   BPMeds           3346 non-null   float64
 6   prevalentStroke  3390 non-null   int64  
 7   prevalentHyp     3390 non-null   int64  
 8   diabetes         3390 non-null   int64  
 9   totChol          3352 non-null   float64
 10  sysBP            3390 non-null   float64
 11  diaBP            3390 non-null   float64
 12  BMI              3376 non-null   float64
 13  heartRate        3389 non-null   float64
 14  glucose          3086 non-null   float64
 15  TenYearCHD       3390 non-null   int64  
dtypes: float64(9), int64(5), object(2)
memory usage: 423.9+ KB
```

```
id                 0
age                0
education         87
sex                0
is_smoking         0
cigsPerDay        22
BPMeds            44
prevalentStroke    0
prevalentHyp       0
diabetes           0
totChol           38
sysBP              0
diaBP              0
BMI               14
heartRate          1
glucose          304
TenYearCHD         0
```

Initial data prepping

Handling NaN values

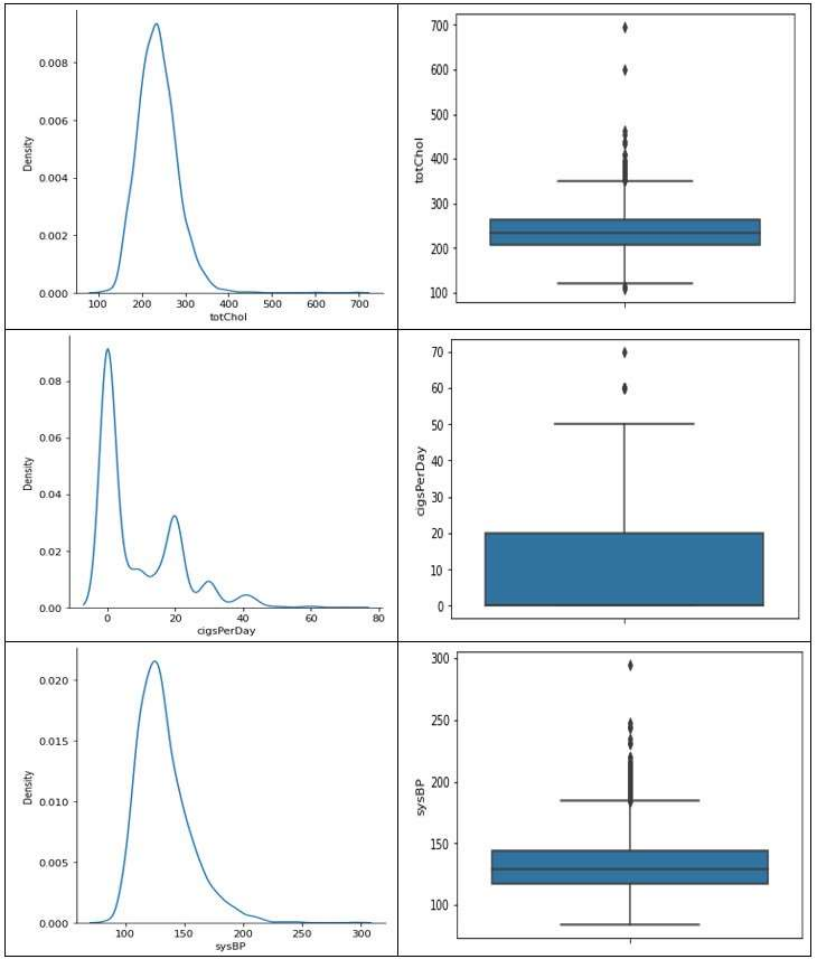
- For Nan value handling, simple imputer("most_frequent") for categorical features and KNN imputer for numerical features is implemented.

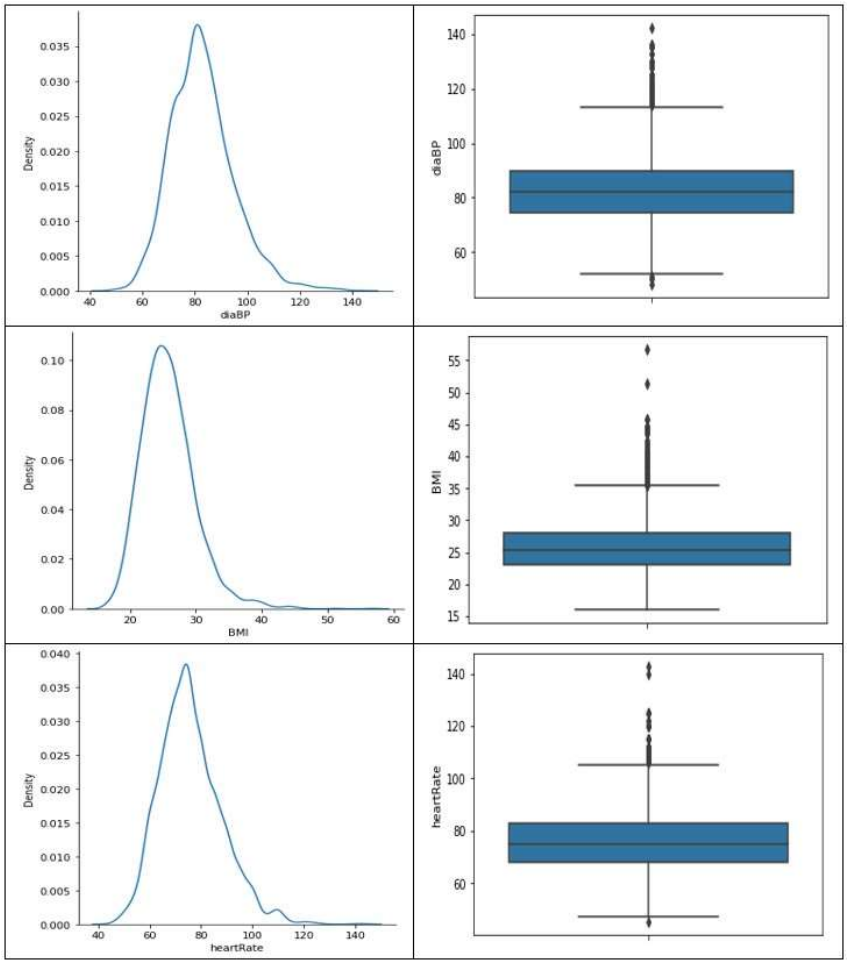
Feature alteration

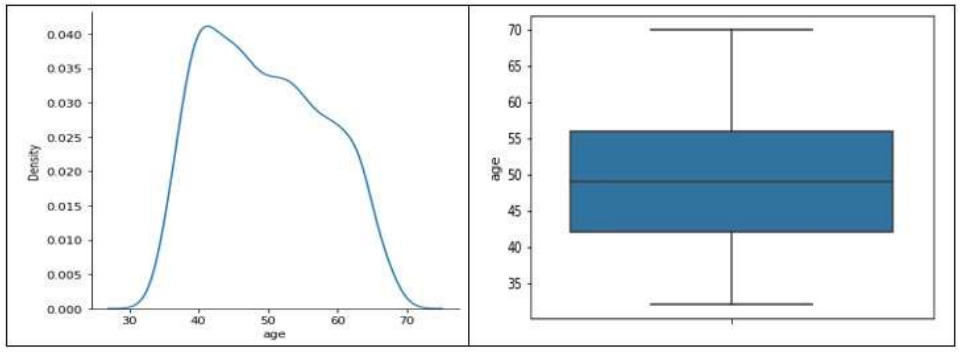
- Label encoding was done on two categorical features i.e., ‘sex’ and ‘is_smoking’
- Irrelevant features i.e., ‘id’ is removed

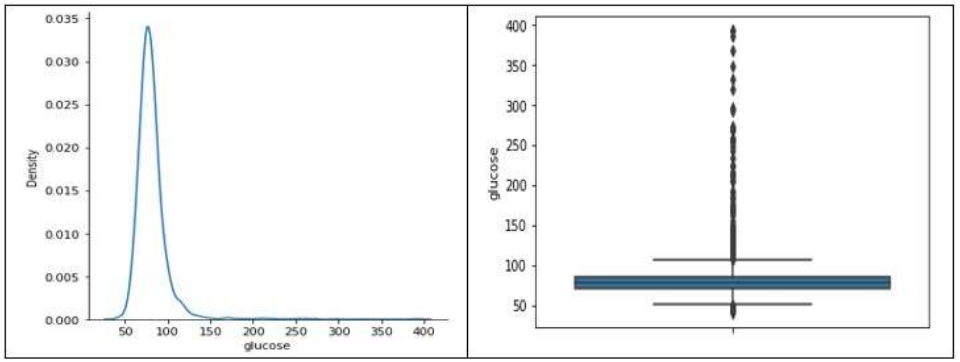
Initial data prepping

Distribution and outlier analysis:









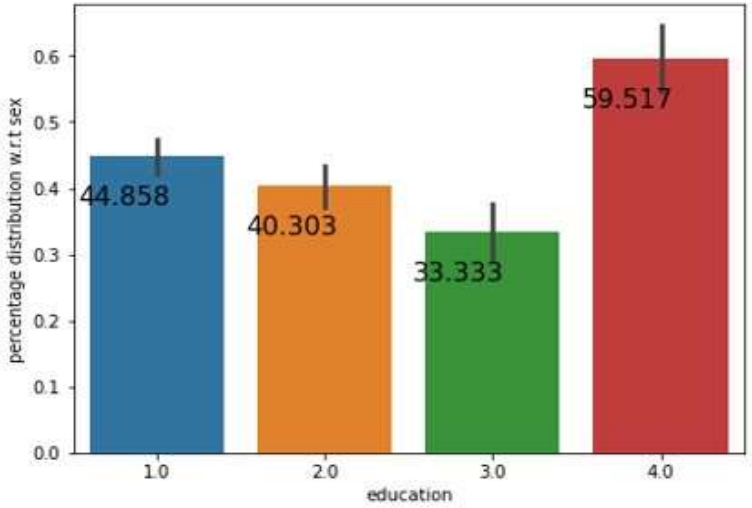
- **No outliers are observed in the age feature.**

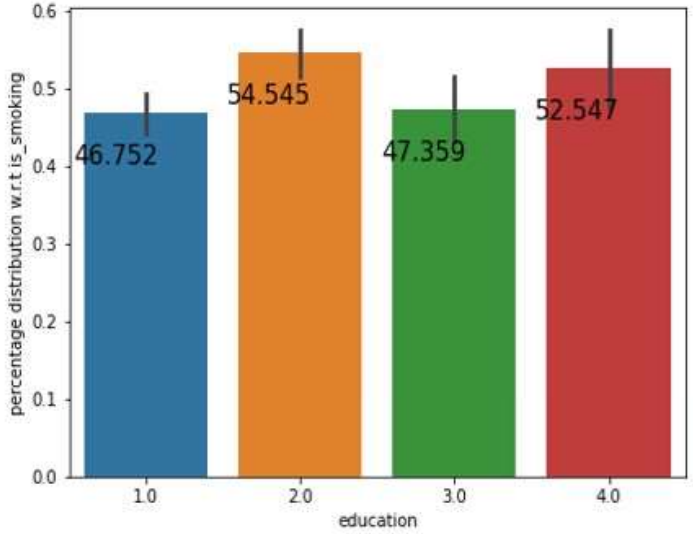
- **Outliers are observed in 'totChol', 'cigsPerDay', 'sysBP', 'diaBP', 'BMI', 'heartRate' and 'glucose' features.**

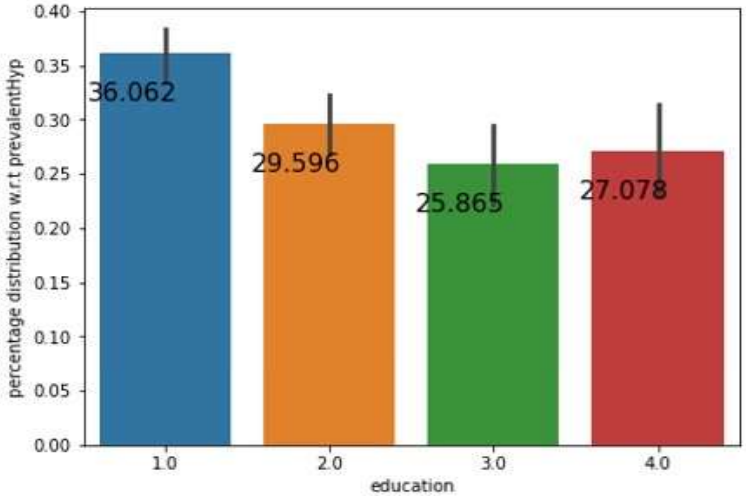
- **To handle those outliers IQR method was implemented.**

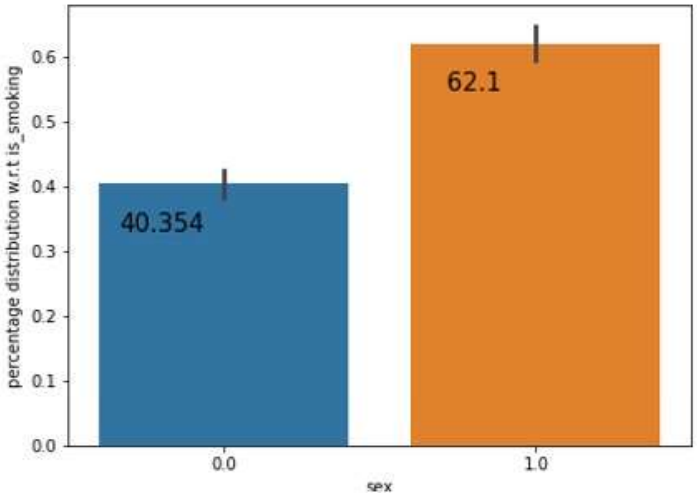
- **In the case of the ‘glucose’ feature many outliers are observed, in order to handle those outliers IQR method is implemented but the max limit is set as 145 because in medical terms glucose level above 145 is considered to be high.**

Exploratory Data Analysis



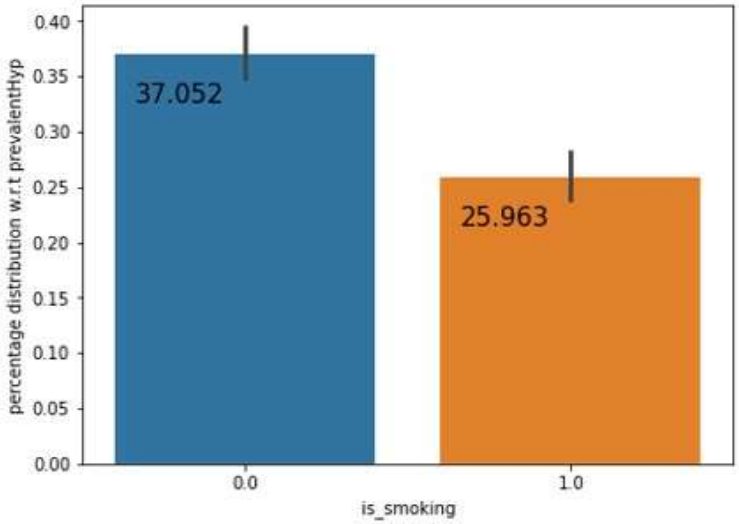


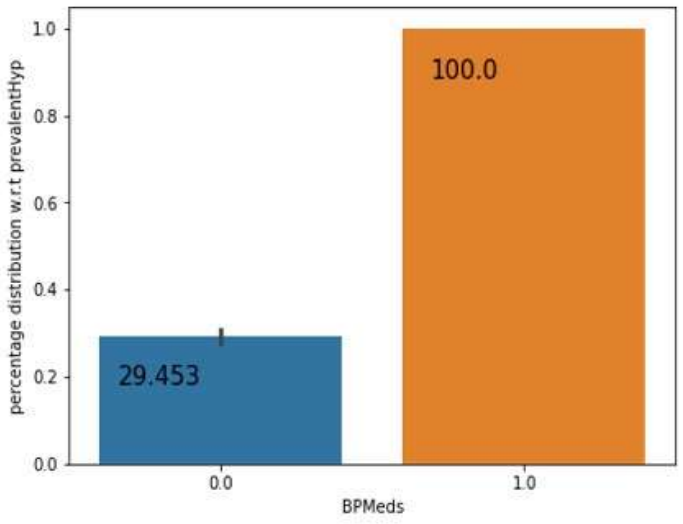


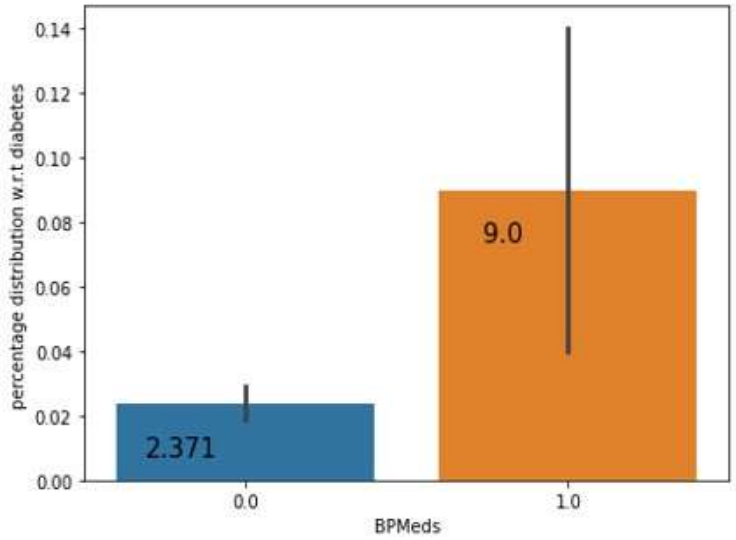


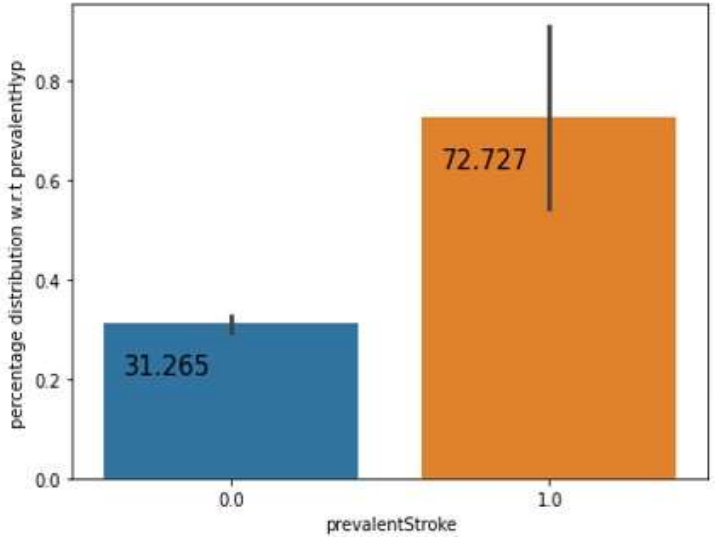
- **People attaining education 1 is high followed by 2,3 and 4.**
- **Males in education 1, 2 and 3 are comparatively less, but in the case of education-4 males are comparatively high.**
- **Around 50% of people in all education types tend to be smokers.**
- **Around 30% of people in all education types tend to have hypertension issues**
- **Males have a higher chance to be a smoker.**

Exploratory Data Analysis



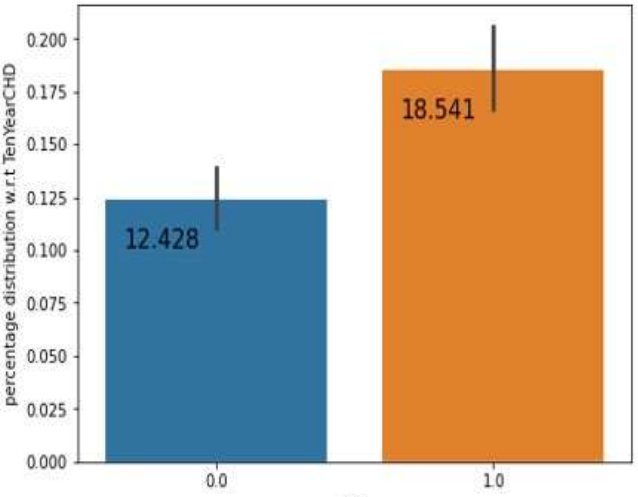


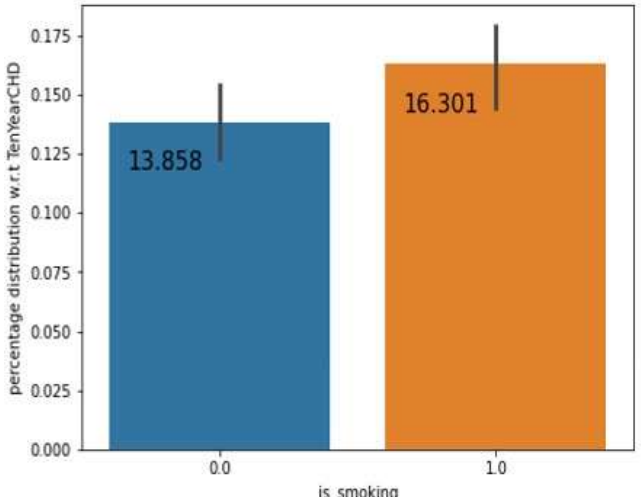


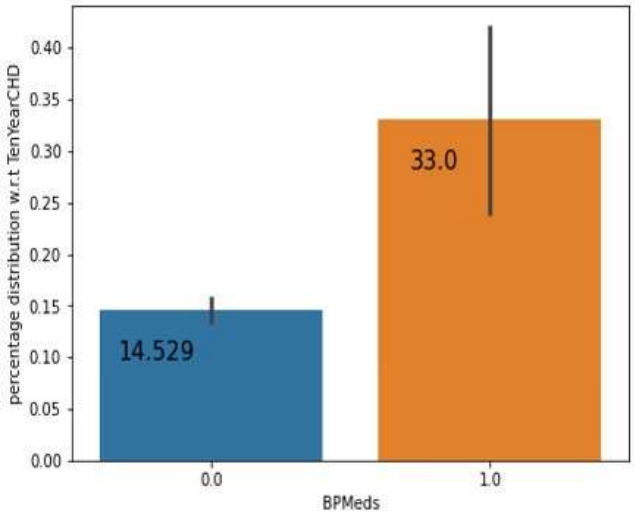


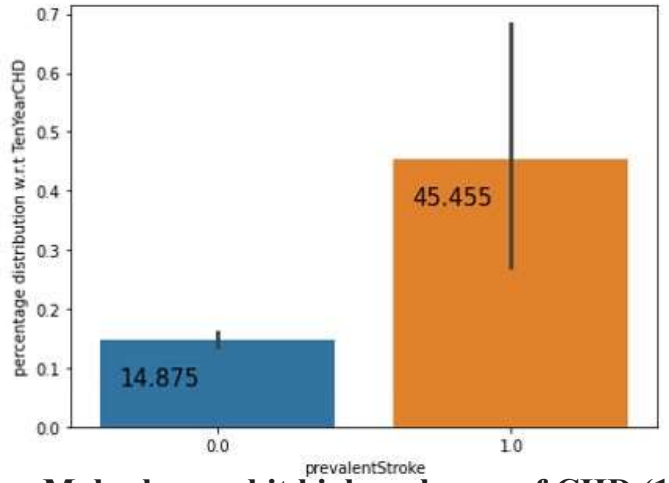
- **People on bp medication tend to have hypertension issues as well.**
- **People on bp medication tend to have a bit higher chance to be diabetic by 9%.**
- **People with prevalent stroke tend to have a higher chance (72%) of having hypertension issues**

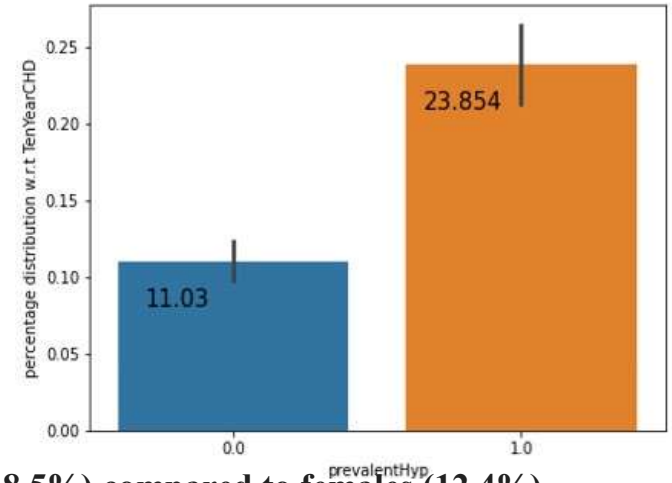
Exploratory Data Analysis

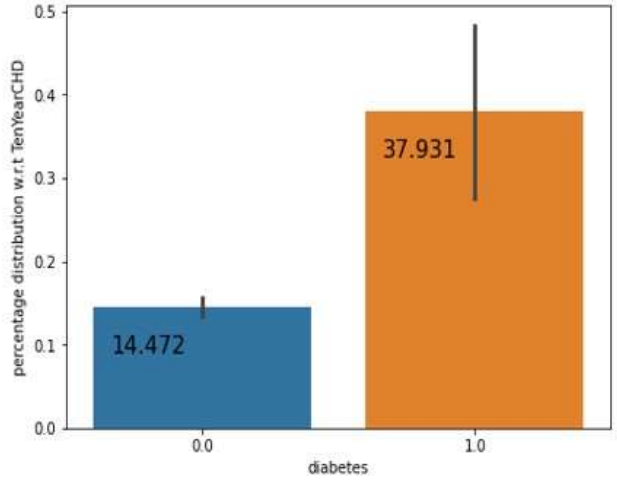






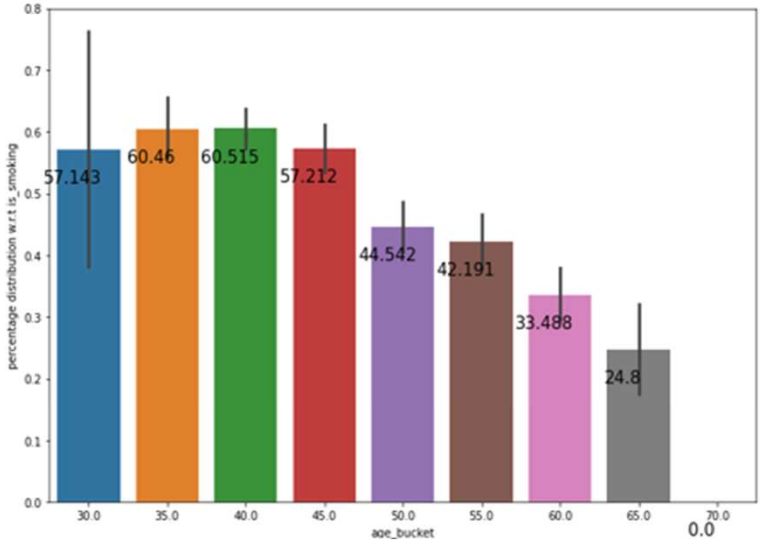


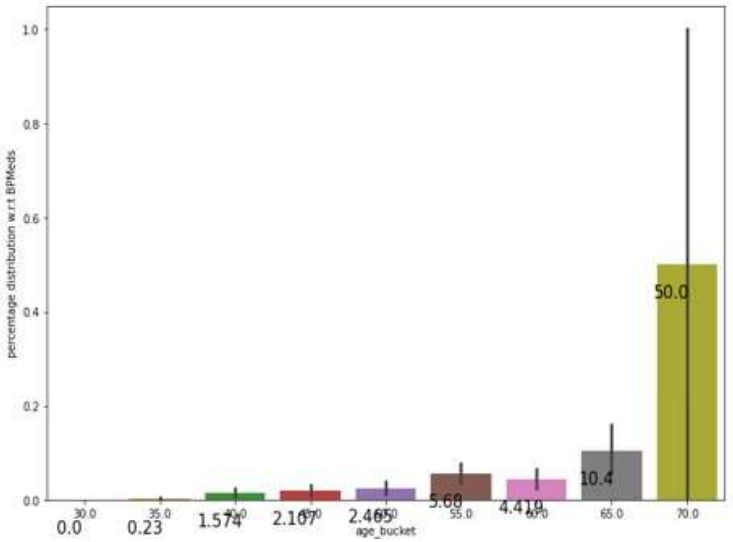


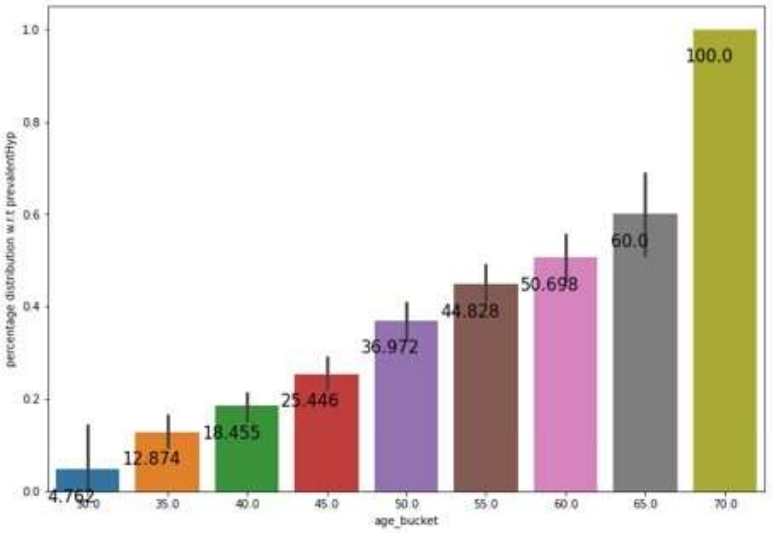


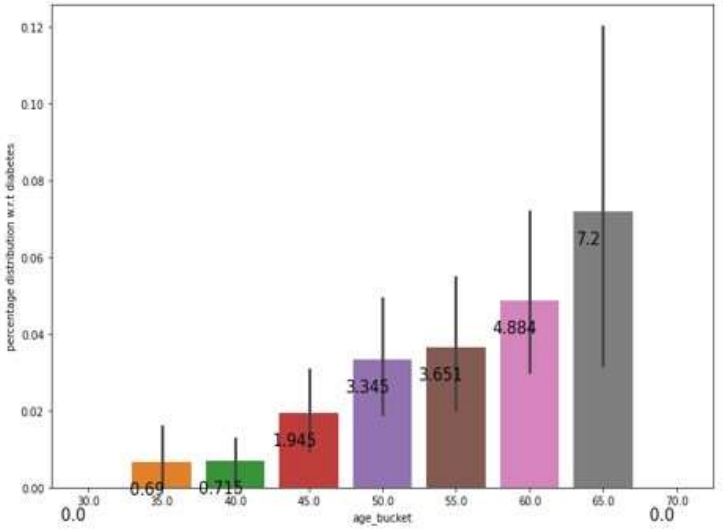
- **Males have a bit higher chance of CHD (18.5%) compared to females (12.4%).**
- **Smokers have a bit higher chance of CHD (16.3%) compared to non-smokers (13.8%).**
- **People with BP medication, prevalent stroke, prevalent hypertension and/or diabetes have a higher chance of CHD .**

Exploratory Data Analysis





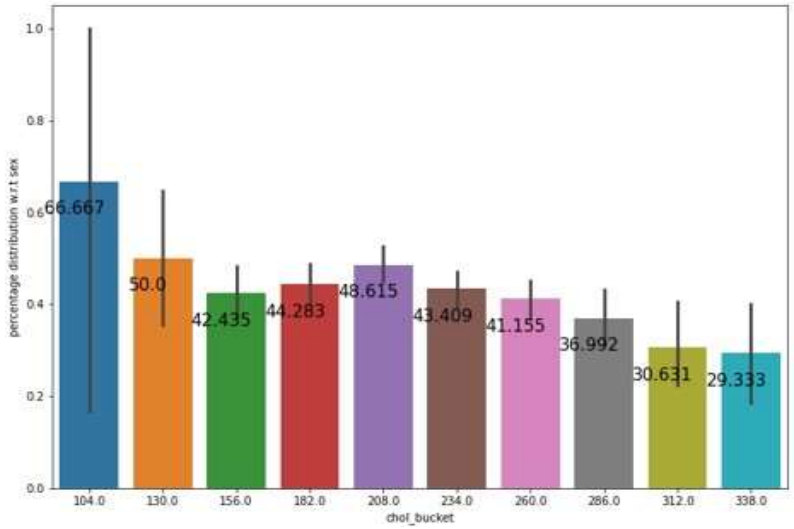


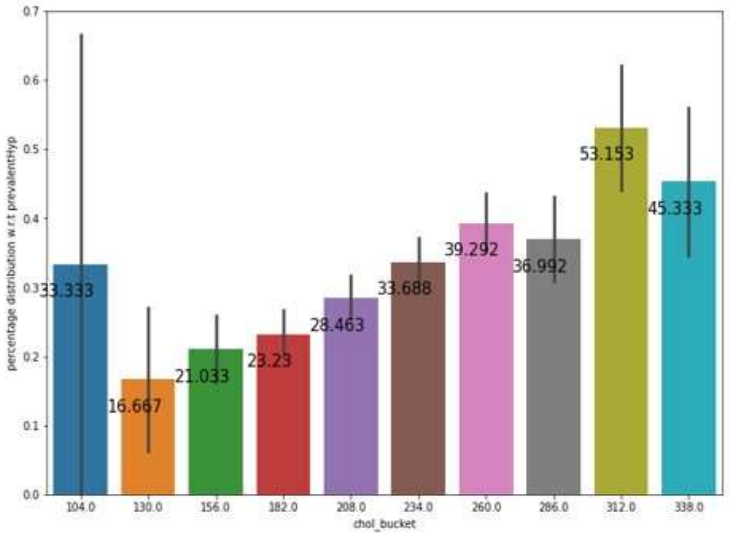


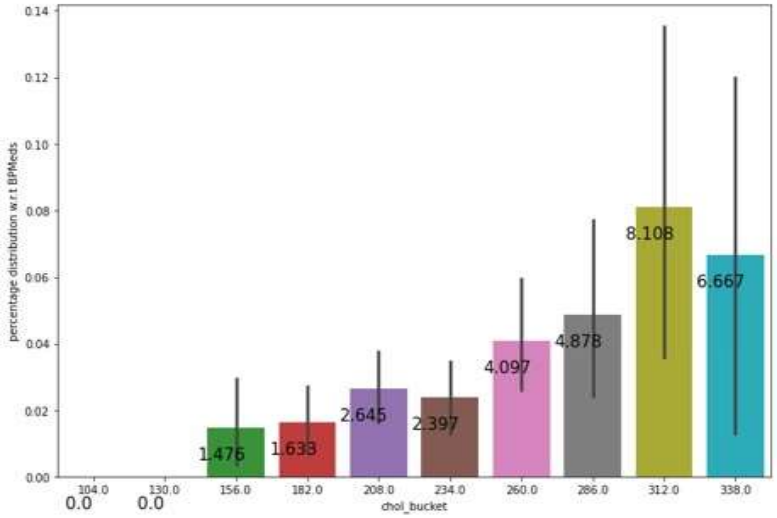
- **Smoking habits tend to decrease with an increase in age.**

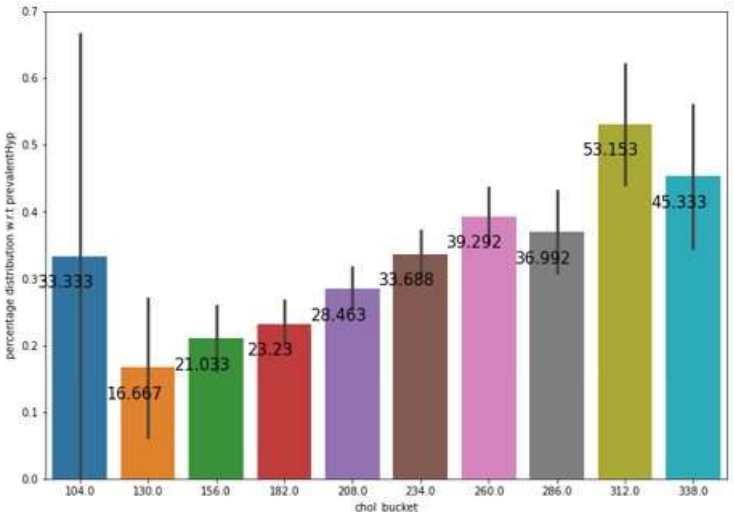
- **Bp, hypertension and diabetes issues tend to increase with increases in age.**

Exploratory Data Analysis





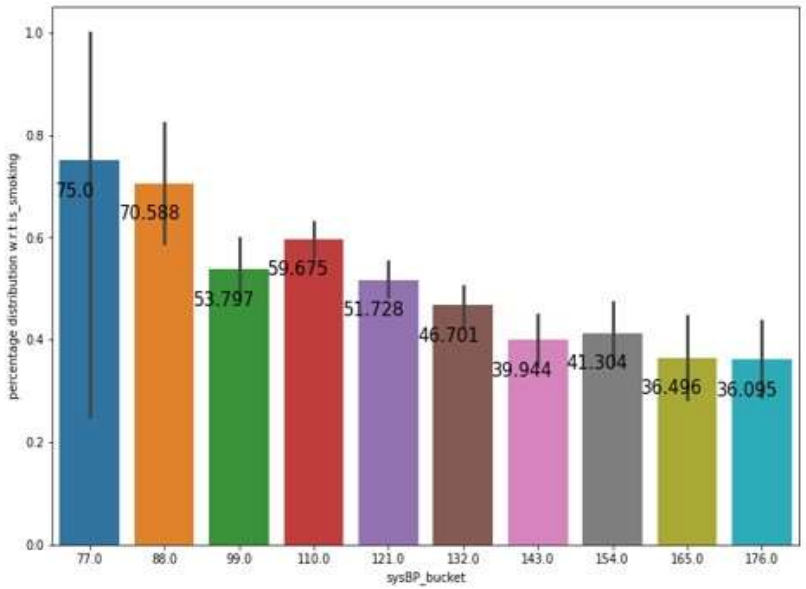


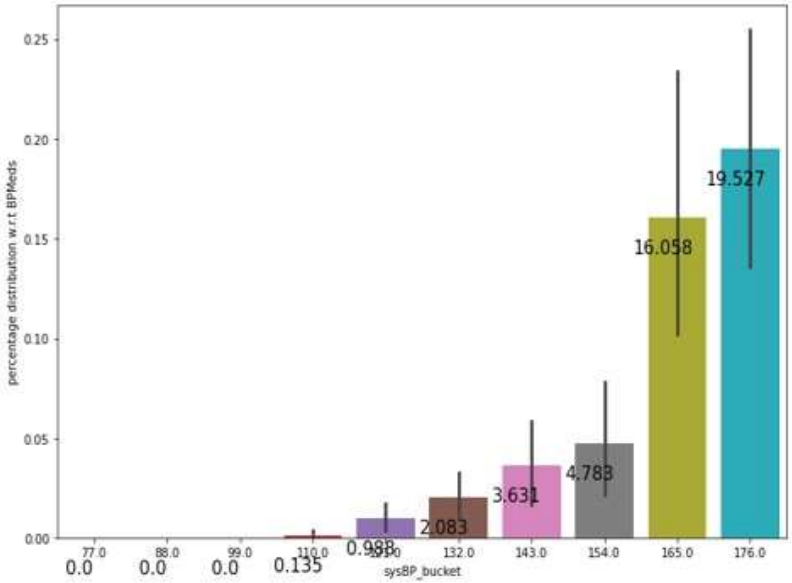


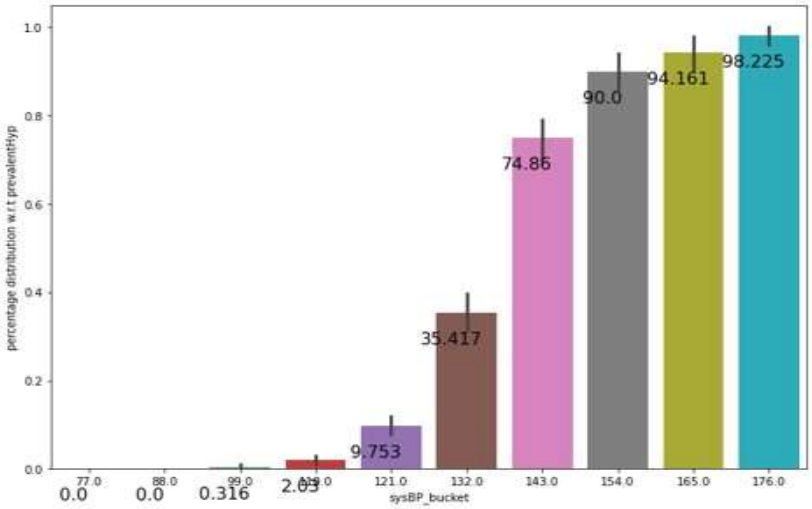
- **Females tend to have higher cholesterol levels compared to males.**

- **Bp and hypertension issues tend to increase with an increase in cholesterol levels**

Exploratory Data Analysis

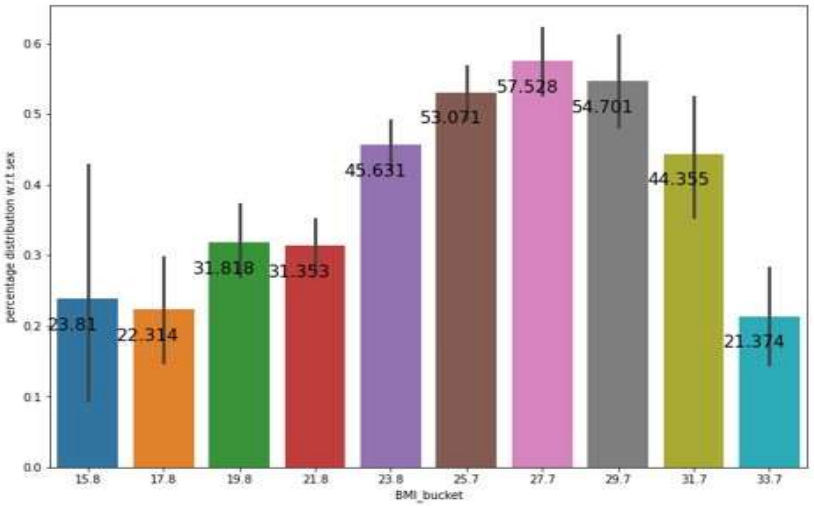


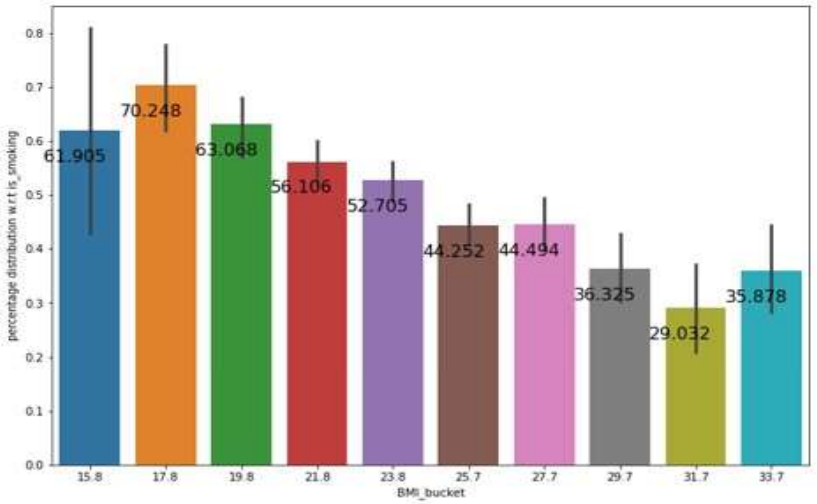


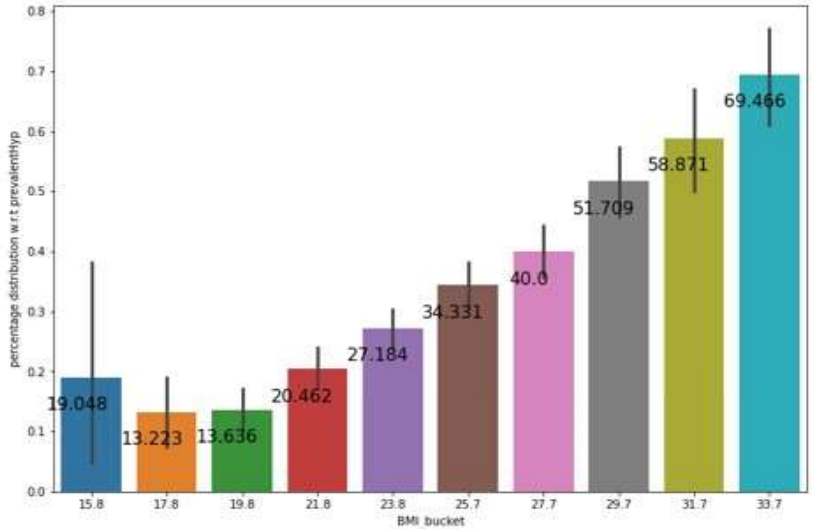


- **SysBp tends to be low for smokers.**
- **SysBp tends to be high for people with BP medication and hypertension issues.**

Exploratory Data Analysis

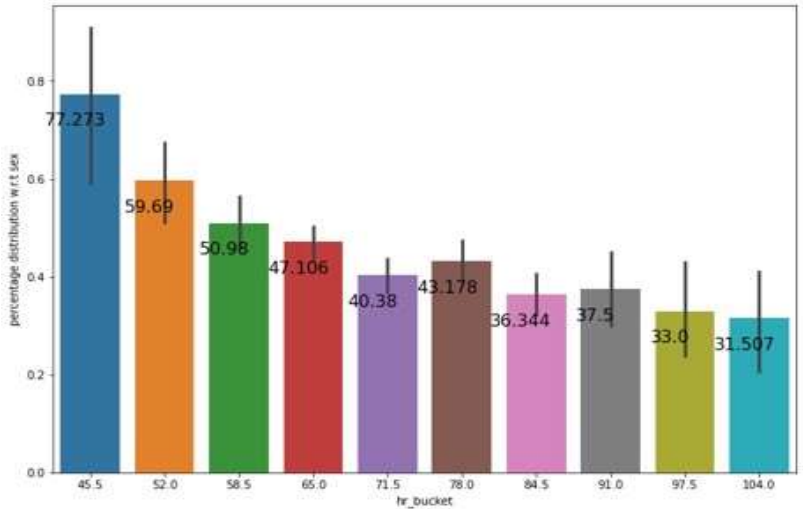


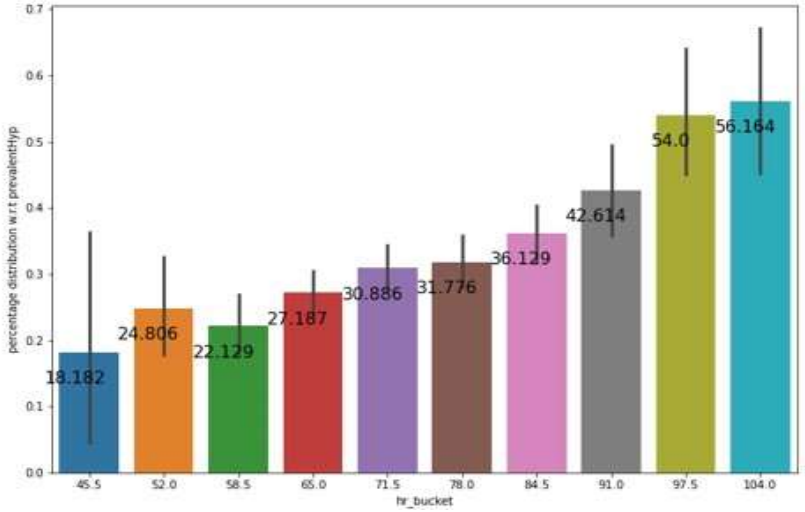


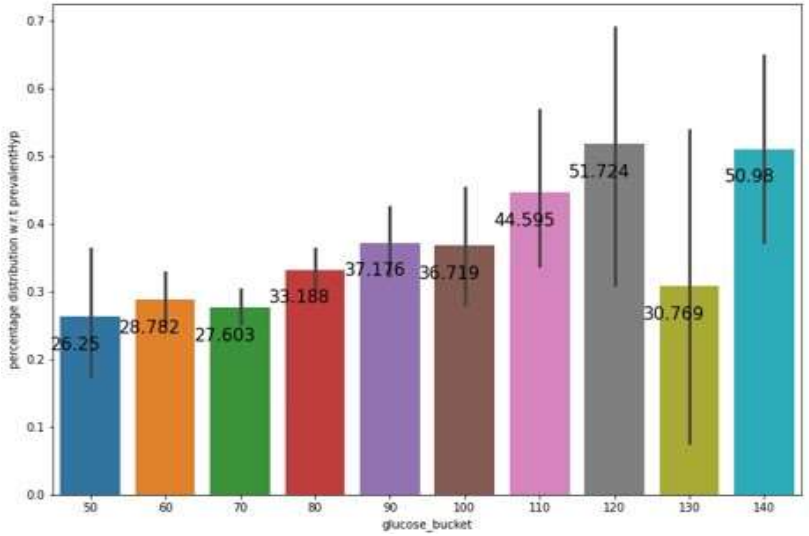


- **BMI levels from 15 to 22 and 30 to 34 are more prevalent in females.**
- **Smokers tend to have lower BMI.**
- **People with prevalent hypertension tend to have higher BMI.**

Exploratory Data Analysis



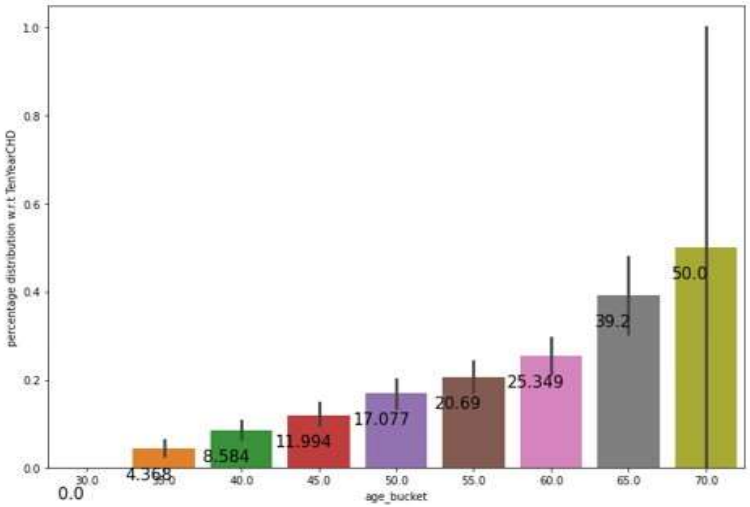


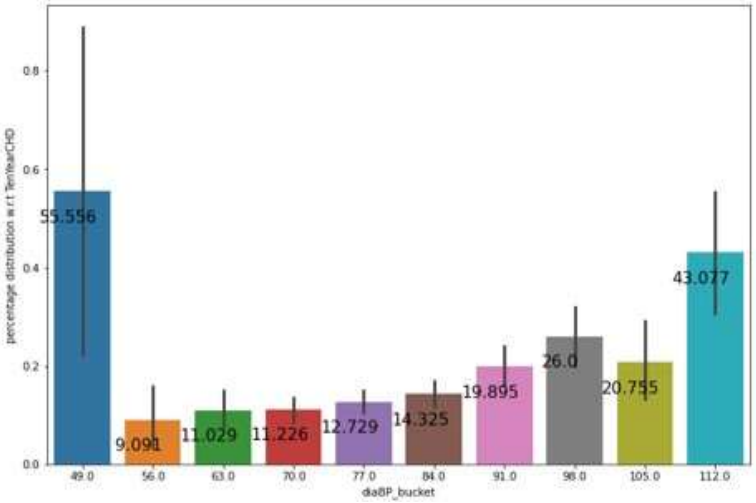


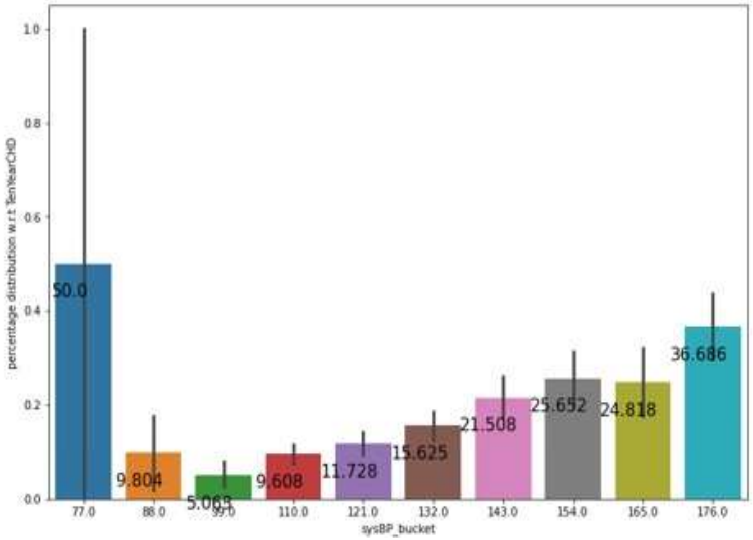
- **Females and people with prevalent hypertension tend to have higher heart rates.**

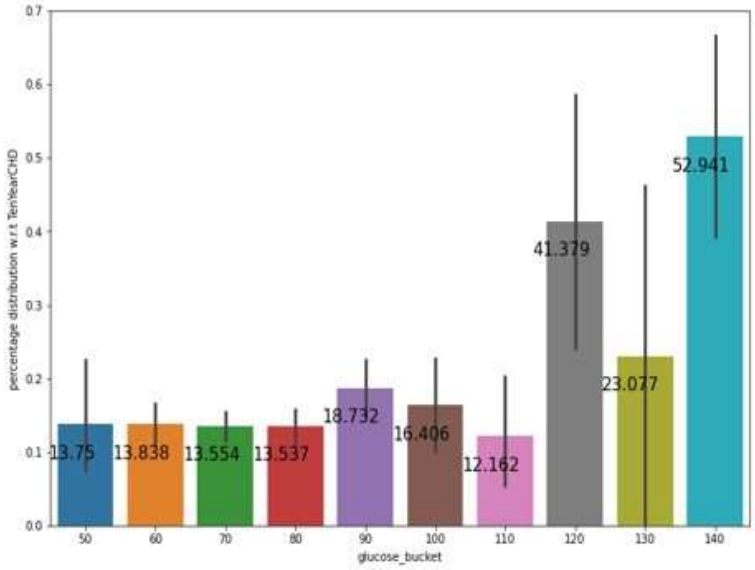
- **People with prevalent hypertension tend to have higher glucose levels.**

Exploratory Data Analysis









- **An increase in glucose levels, SysBp, DiaBp and age tend to have a higher chance for CHD.**

Feature selection and final data processing

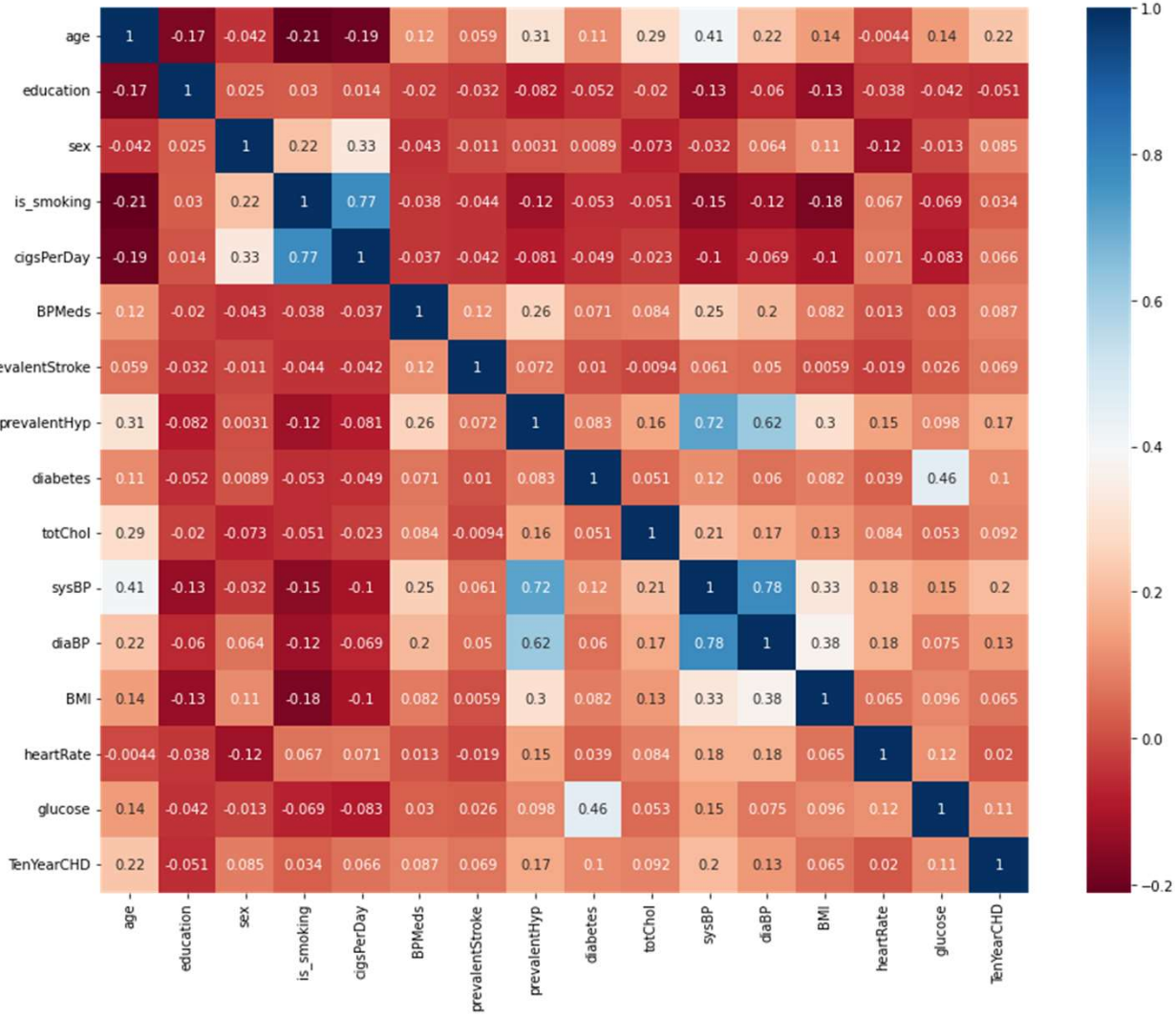
Handling class imbalance

- In this dataset class imbalance was observed with respect to the target variable i.e., 2879 ‘0s’ and 511 ‘1s’, in order to handle this imbalance SMOTE(Synthetic Minority Over-sampling Technique) is used.

Feature selection

- Correlation analysis.
- Analysing feature importance based RandomForestClassifier, DecisionTreeClassifier and XGBClassifier.
- Analysing features based on information gain.

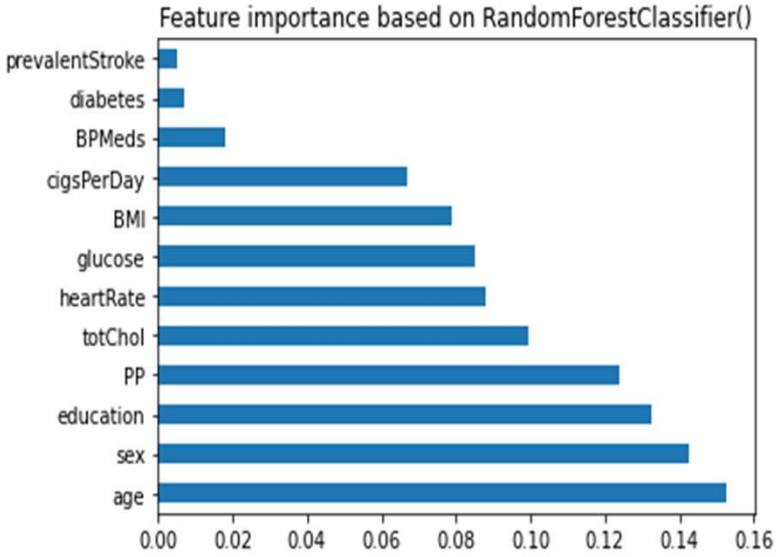
Feature selection and final data processing

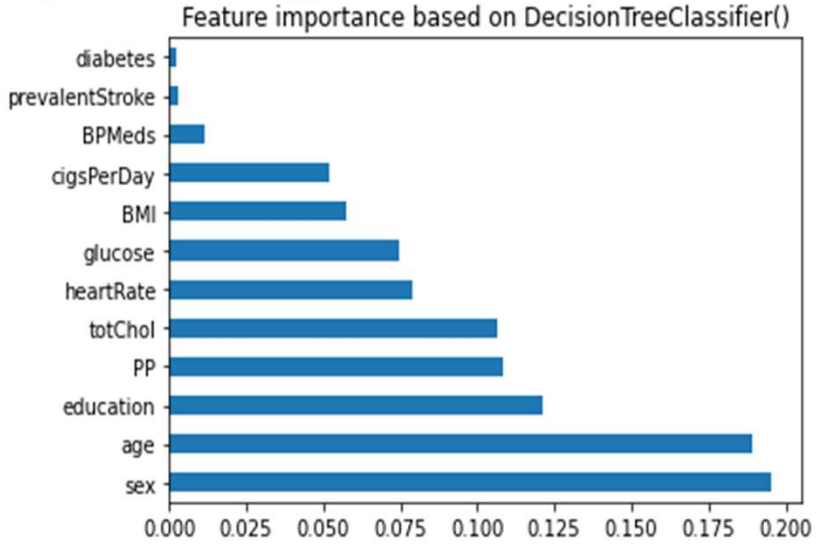


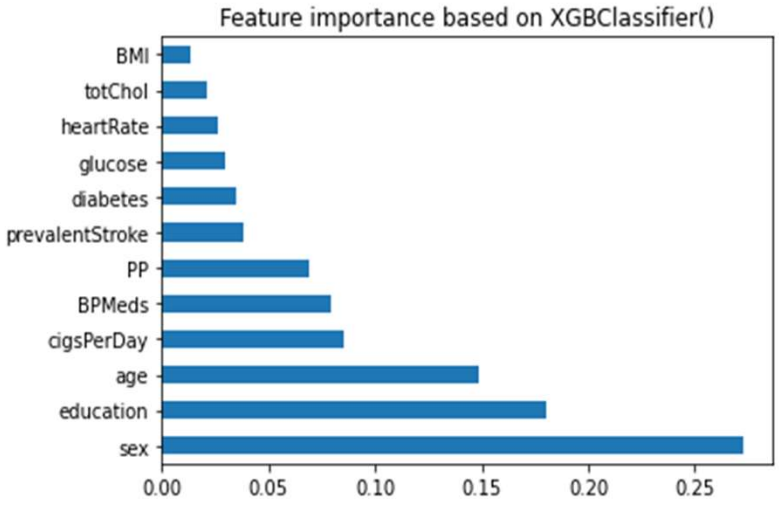
- **'SysBP' and 'DiaBp' is having high correlation of 0.78 followed by 'cigsPerDay' and 'is_smoking' with 0.77, 'SysBP' and 'prevalentHyp' with 0.72 and 'DiaBp' and 'prevalentHyp' with 0.62.**

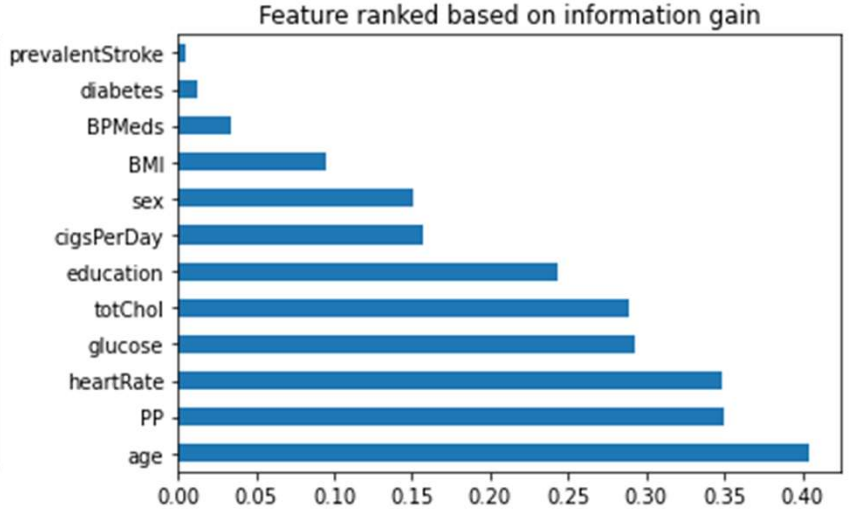
- **In order to resolve the correlation, 'prevalentHyp' , 'prevalentHyp' and 'is_smoking' features are dropped from the dataset and a new feature 'PP'(Pulse pressure) is introduced which is the difference between 'SysBP' and 'DiaBp' .**

Feature selection and final data processing









After analyzing feature importance and information gain, features such as 'prevalentStroke', 'diabetes', 'BPMeds' and 'BMI' are having less preference or significance.
So, these features are dropped from the data set.

Model Implementation and prediction

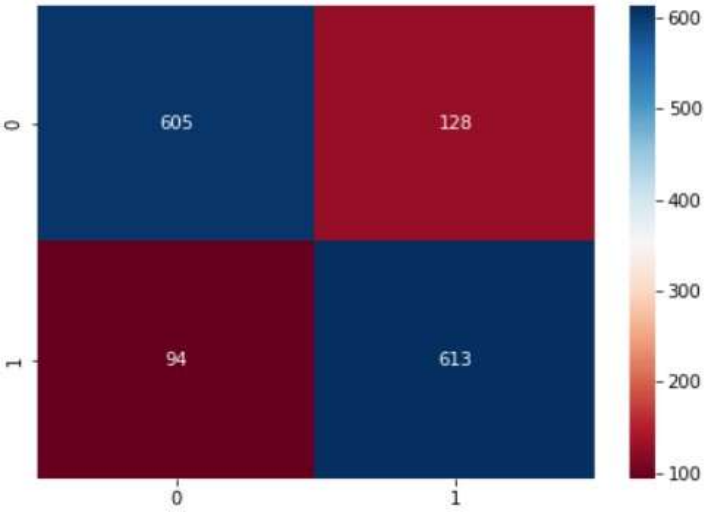
- For model implementation the independent variables selected are,
- ['education', 'sex', 'diabetes', 'age', 'totChol', 'cigsPerDay','heartRate', 'glucose']
- and the dependent variable is [‘TenYearCHD’].
- The dataset is split into test and train with a test size of 25%.
- X train shape (4318, 8)
- Y train shape (4318,1)
- X test shape (1440, 8)
- Y test shape (1440,1)
- In this project six classification techniques were used, those are,
    1. Decision tree classifier
    2. logistic regression classifier
    3. KNN classifier
    4. SVC
    5. XGB
    6. Random forest classifier

Model Implementation and prediction

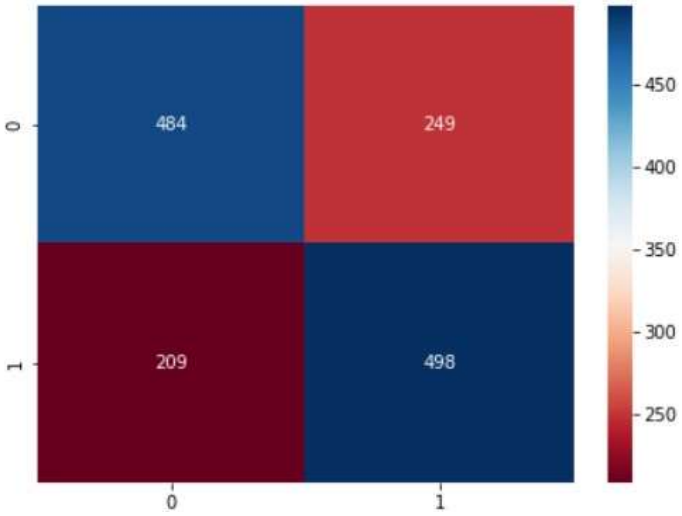
Sl. No	Model	Train Accuracy	Test Accuracy	Train Precision	Test Precision	Train Recall	Test Recall	Train ROC AUC	Test ROC AUC	Train F1	Test F1
1	Decision Tree Classifier	1.00	0.82	1.00	0.80	1.00	0.85	1.00	0.82	1.00	0.82
2	Decision Tree Classifier tunned	1.00	0.83	1.00	0.81	1.00	0.87	1.00	0.84	1.00	0.84
3	Logistic Regression	0.66	0.67	0.66	0.66	0.67	0.69	0.66	0.67	0.67	0.67
4	Logistic Regression tunned	0.67	0.68	0.67	0.67	0.68	0.70	0.67	0.68	0.67	0.68
5	KNN	0.88	0.82	0.82	0.75	0.97	0.94	0.88	0.82	0.89	0.83
6	KNN tunned	1.00	0.88	1.00	0.82	1.00	0.97	1.00	0.88	1.00	0.89
7	SVC	0.79	0.77	0.82	0.78	0.75	0.75	0.79	0.77	0.78	0.76
8	SVC tunned	0.98	0.85	0.97	0.81	0.98	0.90	0.98	0.85	0.98	0.85
9	XGB Classifier	0.88	0.86	0.95	0.92	0.81	0.79	0.88	0.86	0.87	0.85
10	XGB Classifier tunned	0.99	0.89	1.00	0.93	0.99	0.85	0.99	0.89	0.99	0.89
11	Random Forest	1.00	0.90	1.00	0.93	1.00	0.88	1.00	0.90	1.00	0.90
12	Random Forest tunned	1.00	0.91	1.00	0.92	1.00	0.89	1.00	0.91	1.00	0.91

Model Implementation and prediction

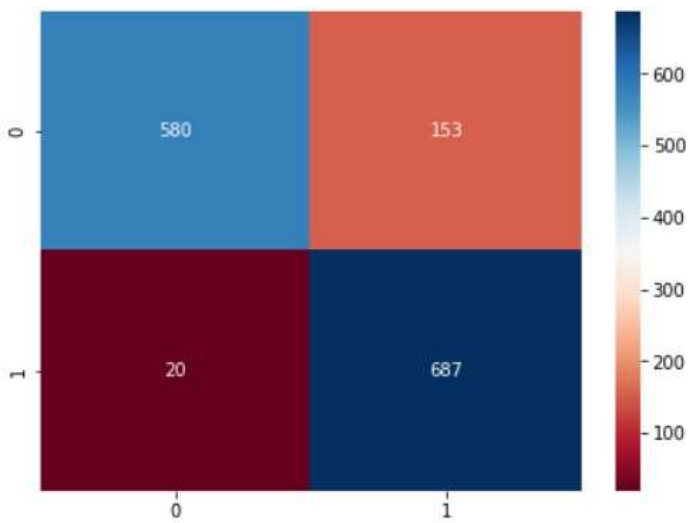
Decision tree classifier



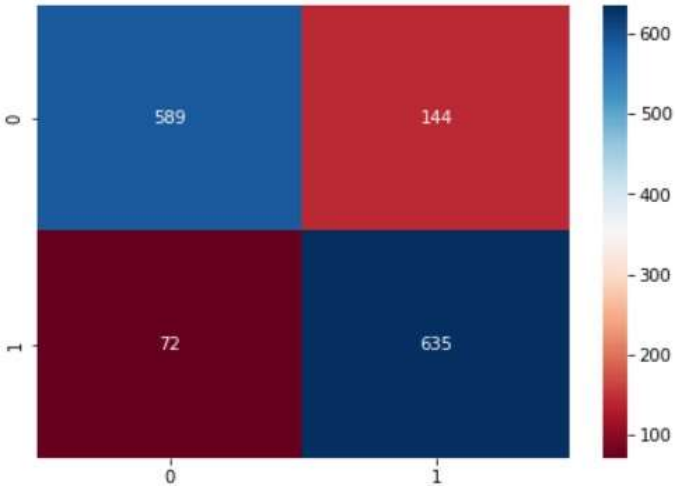
logistic regression classifier



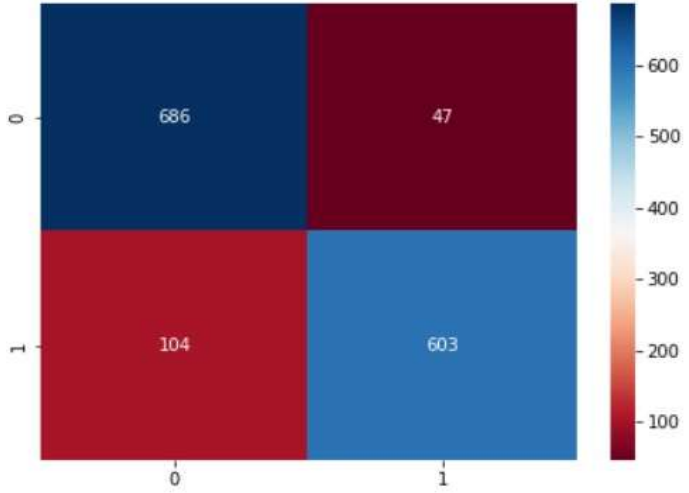
KNN classifier



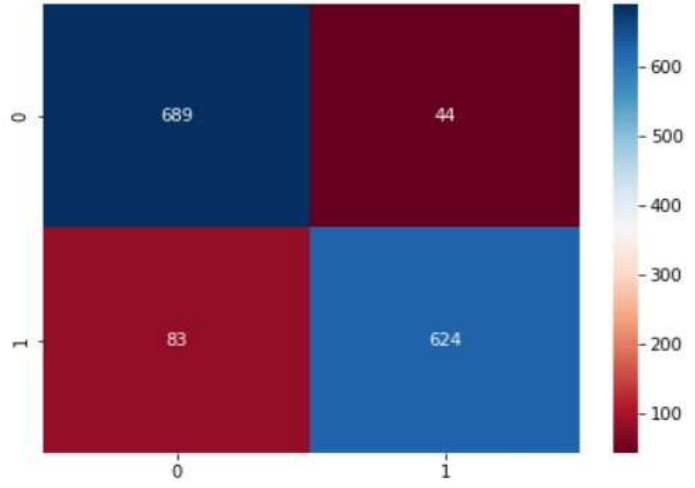
SVC



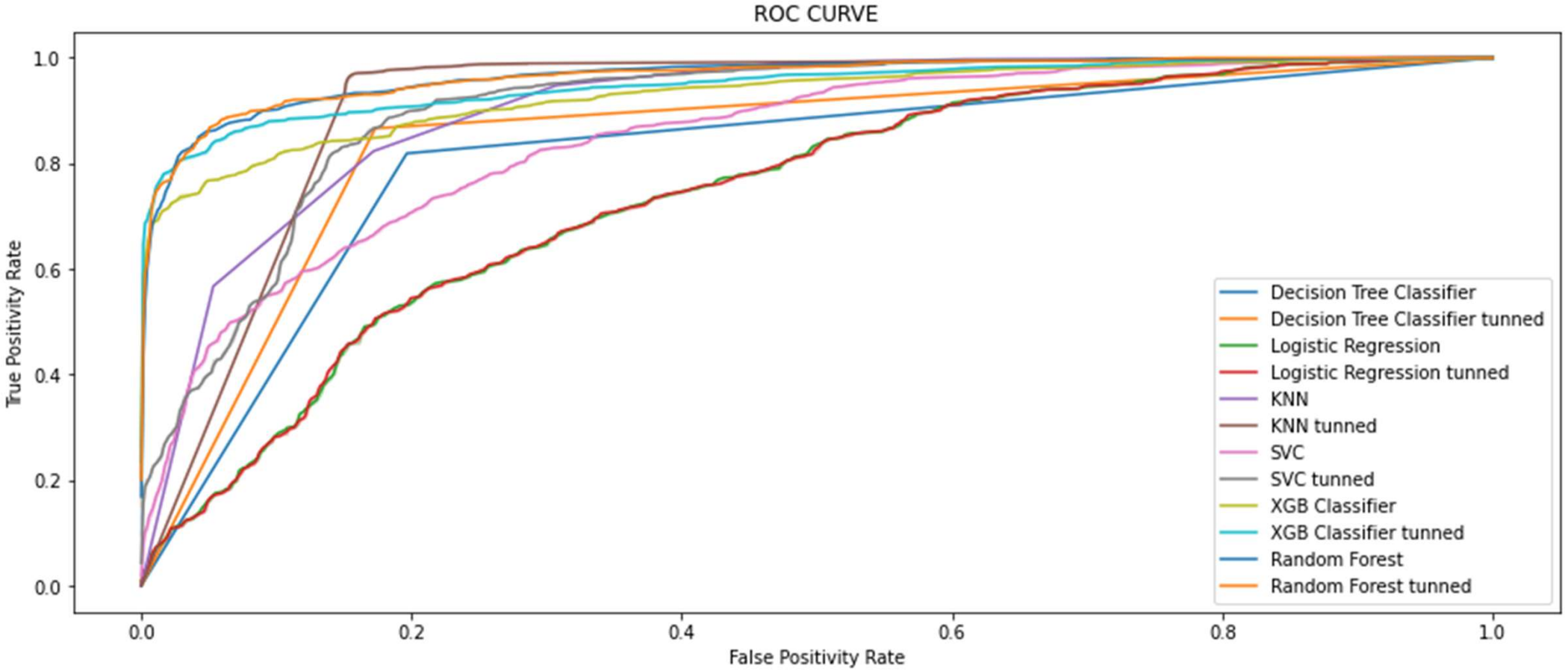
XGB classifier



Random forest classifier



Model Implementation and prediction



Conclusion:

- I started off with data cleaning and necessary feature changes, then to comprehend the data an EDA was performed which yielded many meaningful insights about the data, then the features were trimmed down based on various feature selection techniques and finally 6 classification models were implemented namely decision tree classifier, logistic regression classifier, KNN classifier, SVC, XGB and random forest classifier. All of the models were hyperparameter tunned and evaluated based on different evaluation techniques, and the main intention of hyperparameter tunning was to improve overall model performance with stress on reducing false negatives.

- Out of the models implemented Random Forest, XGB and KNN showed good results and while considering overall performance Random Forest showed great results.