Capstone Project-3

Cardiovascular Risk Prediction

By-Md.ImranHaji

Points to discuss:-

- Problem Statement.
- Data summary.
- Feature Engineering and Data Analysis
- Understand the relationships between variables
- Hypothesis Testing
- Variable _Selection.
- Model Implementation and evaluation.
- Conclusion.

Problem Statement.

Despite of the age, rise in number of Cardiovascular Heart Disease (CHD) is the one of the leading reason for death annually worldwide in recent past years. However, it can be prevented if caught early and simple changes are made in lifestyle . This project would explore a set of given data and known factors for heart disease ,and develop a classification machine learning model to predict risk of develo ping heart disease within the next ten years.

Along with model building ,will look in some feature behavior.

- Spread of dependent variable?
- How each independent variable is behaving with target variable?
- How is the relationship between Gender variable to type of blood pressures?
- How is the relationship between Gender variable to cholesterol and heartrate ?
- Behavior of age vs (‘totChol', 'sysBP','diaBP', 'BMI', 'heartRate', 'glucose’) with target variable?
- How is the relationship between cigsPerDay and target variable ?

Data summary.

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

Data Description

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.)

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)

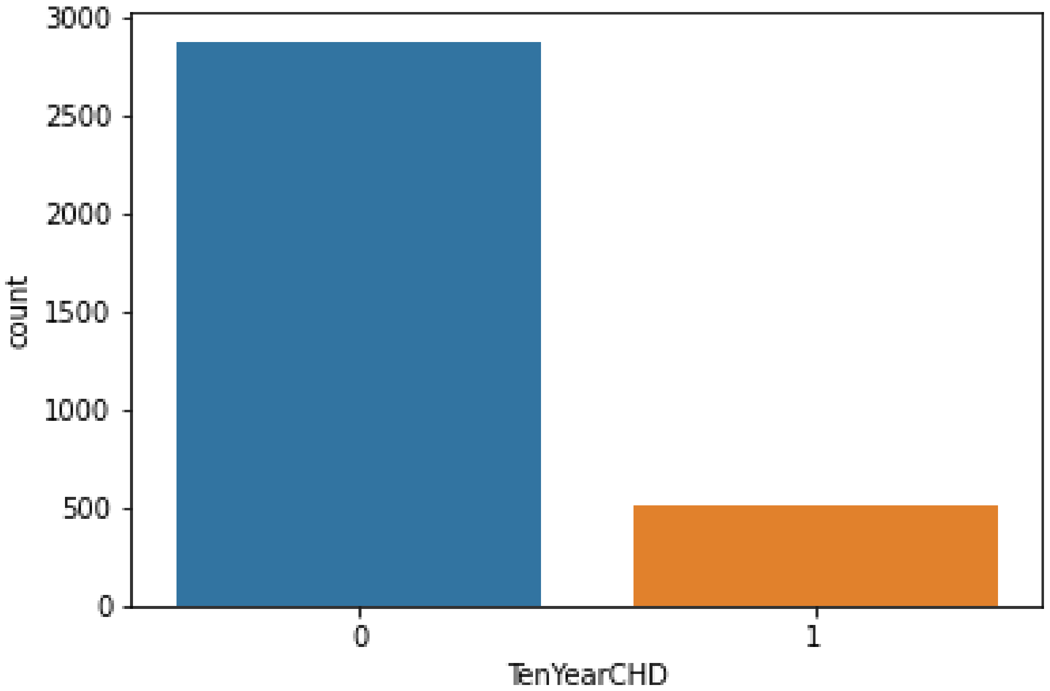
10-year risk of coronary heart disease CHD(binary: “1”, means “Yes”, “0” means “No”) - Dependent Variable

Understand the relationships between variables

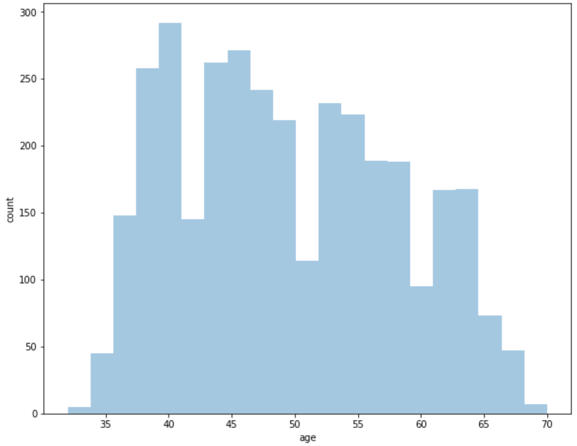
Here our dependent variable is TenYearCHD

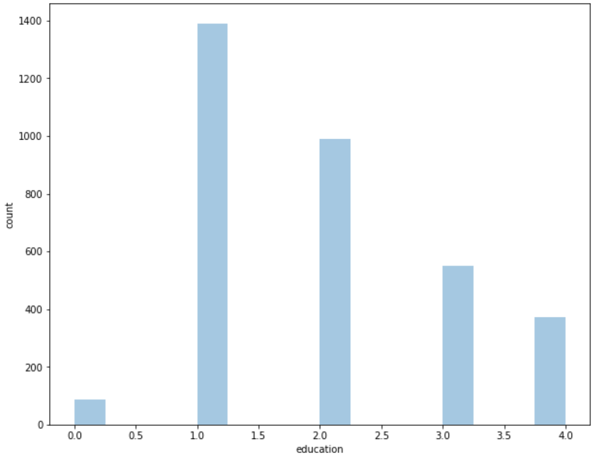
- 10-year risk of coronary heart disease CHD (binary: “1”, means “Yes”, “0” means “No”)

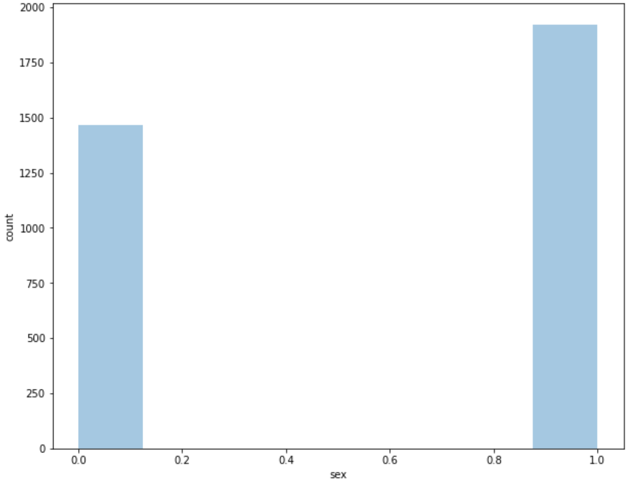
- Well, people with no heart disease are more , but as this is our target variable it leads to imbalance data set .Therefore further data imbalance manipulation will be done before using this data in model deployment

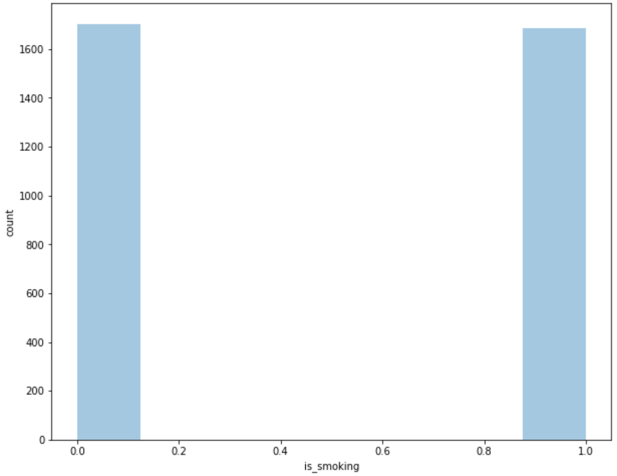


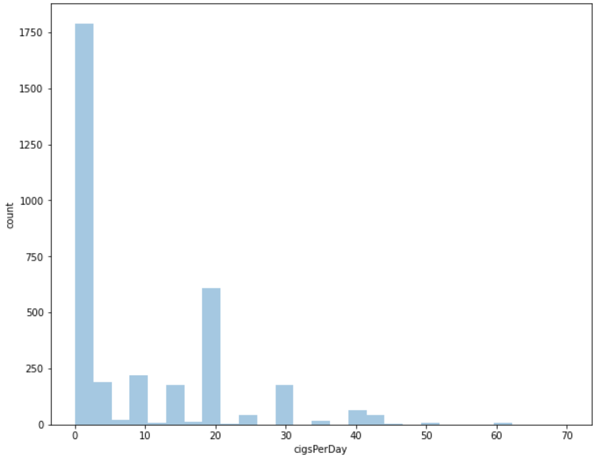
Univariant Analysis:

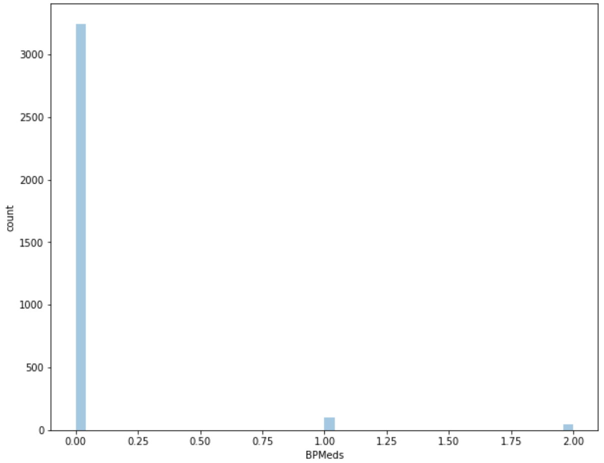


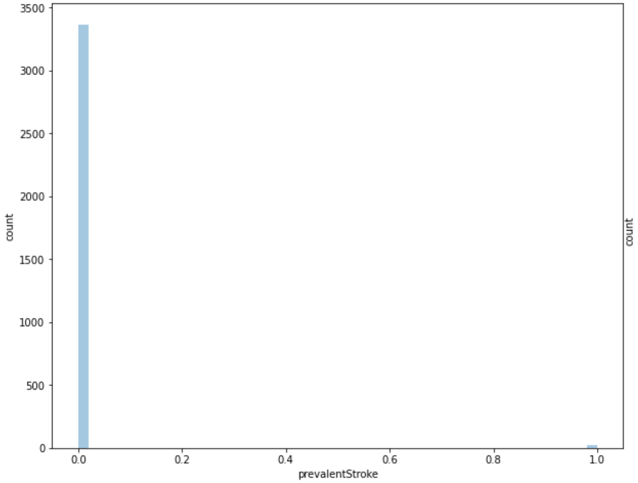


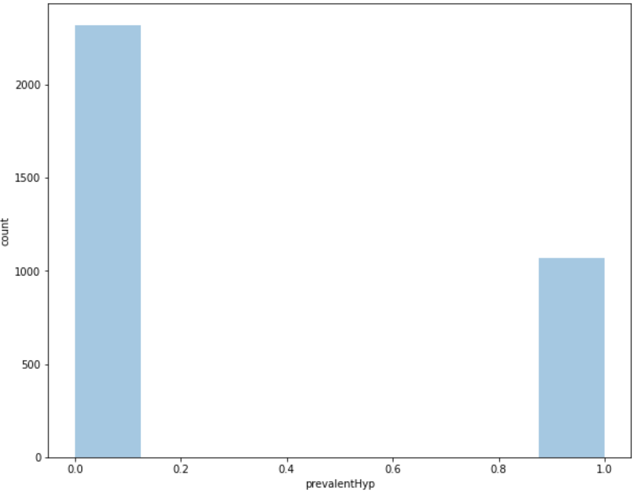


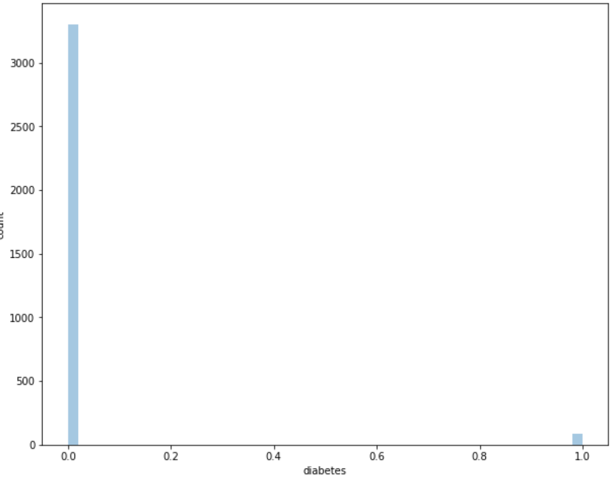


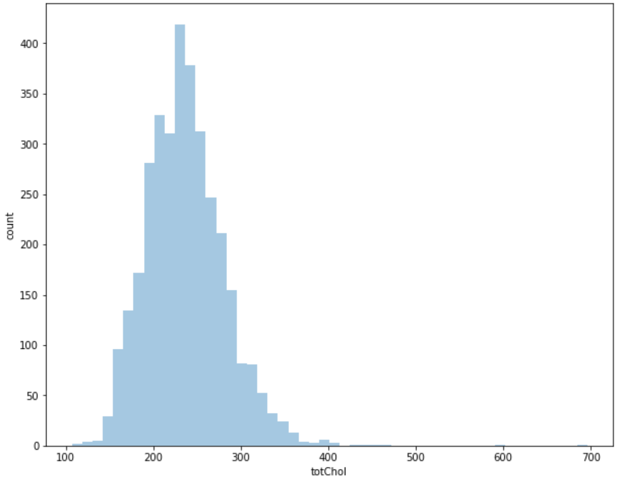


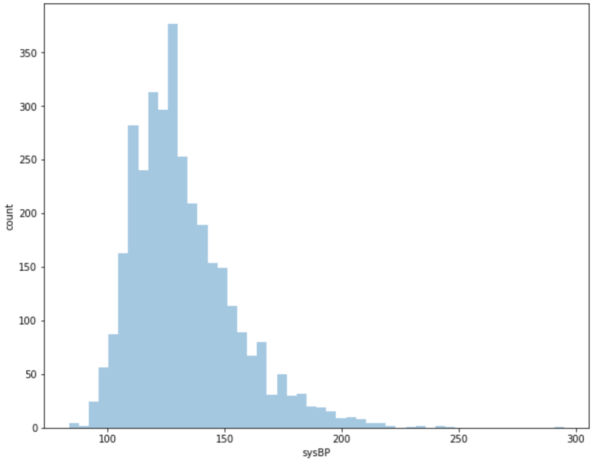


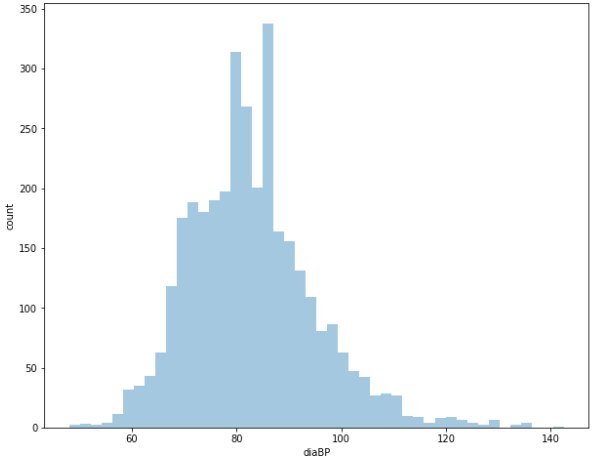


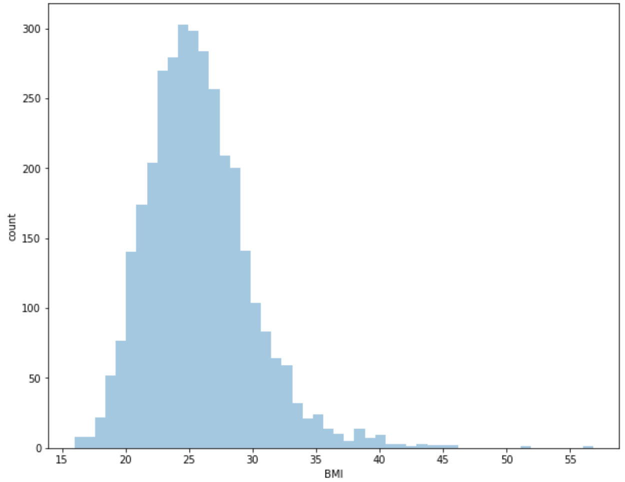


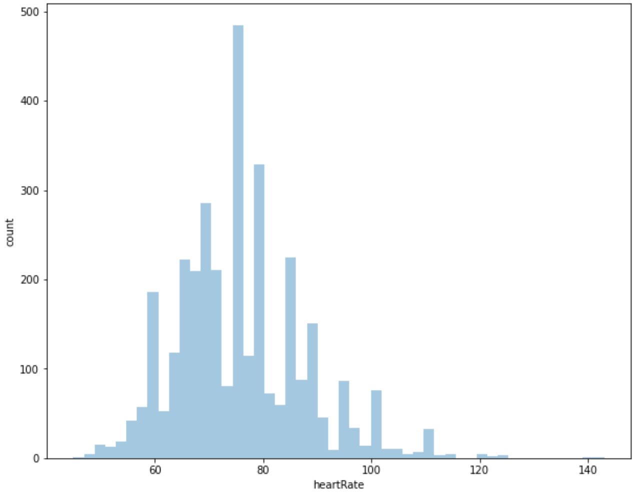


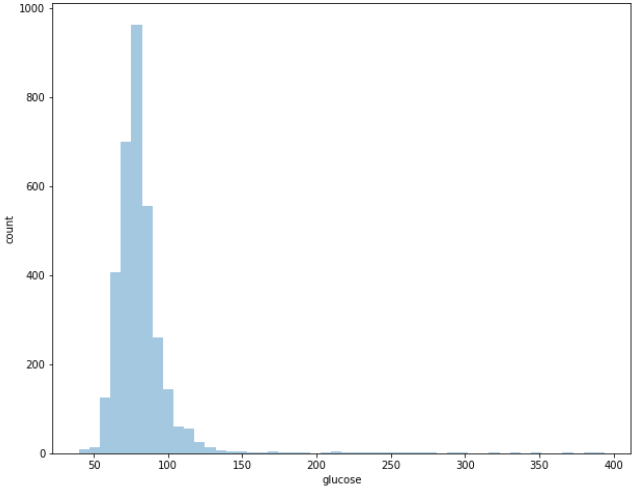












Observation:

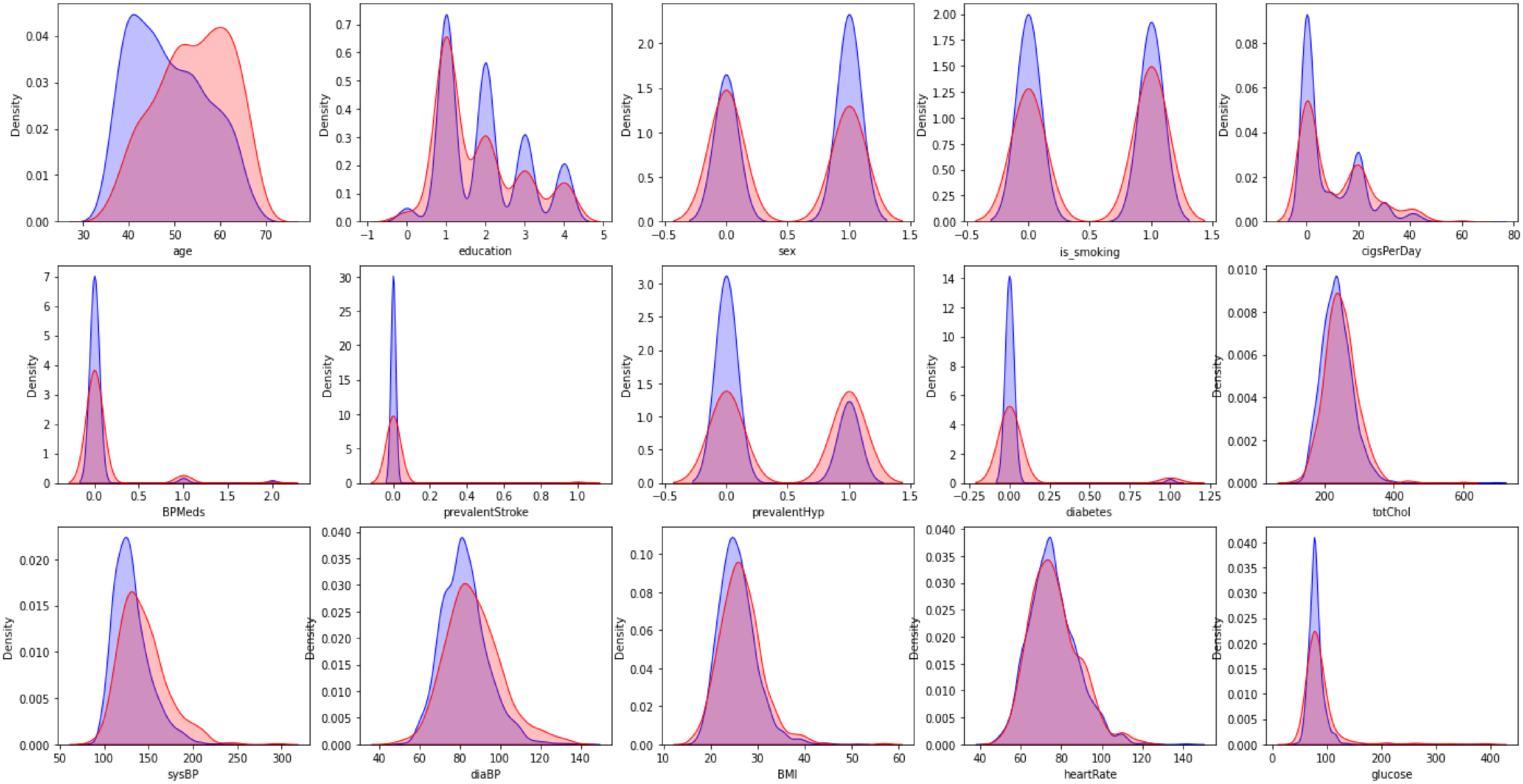
- Sample was taken in age group of people between 32 to 70
- Education Level: 1 with count of 1391, 2 with count of 990, 3 with count of 549, 4 with count of 373,0.0 with count of 87
- Gender, 0=Male with 1467 count , 1 = Female with 1923 count
- patient is a smoker 1= yes with count 1687 , 0=no with count 1703 almost balanced set

Observation:

- Number of Cigarettes smoked per day value range 0 to 70 high is as :

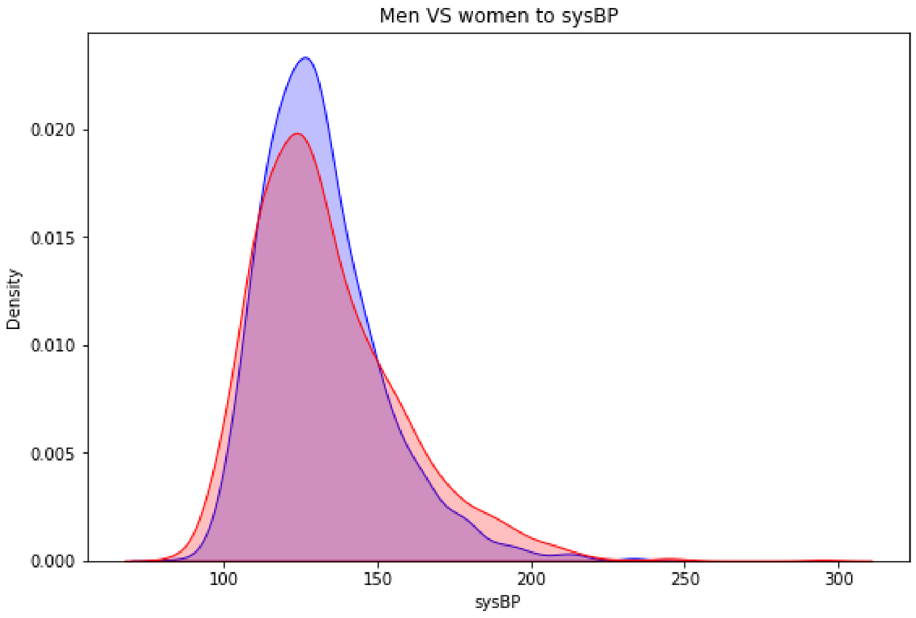
  0.0 -- 1725
  20.0 -- 606
  30.0 -- 176
  15.0 -- 172
- Blood Pressure Medications 0 = no with count of 3246, 1=yes with count of 100, 2 = unknown with count of 44
- Prevalence of stroke 0=none with count of 3368, 1 = had occurrences of stroke with count of 22
- Prevalence of hypertension 0=none with count of 2321 , 1= has prevalence hypertension with count of 1069
- patient has diabetes 0=no with count of 3303 , 1=yes with count of 87
- Total Cholesterol value range between 107-696 where as 125 to 200 is optimum range our data set is highly concentrated between 200 to 270 is right skewed
- systolic blood pressure value range between 83 - 295 where as normal 120-130 and our data is between 100 to 150 is nearly right skewed
- diastolic blood pressure A normal range for adults is 60 mmHg to 80 most of our data is between 70 to 100 is near to normally distributed.
- Body Mass Index most of our data is between 20 to 30 is nearly right skewed.
- Heart Rate a normal range 70 to 110 bpm most of our data is between 60 to 100 is near to normally distributed.
- Glucose (72 to 99 mg/dL) when fasting (140 mg/dL) 2 hours after eating most of our data is between 60 to 90 and right skewed

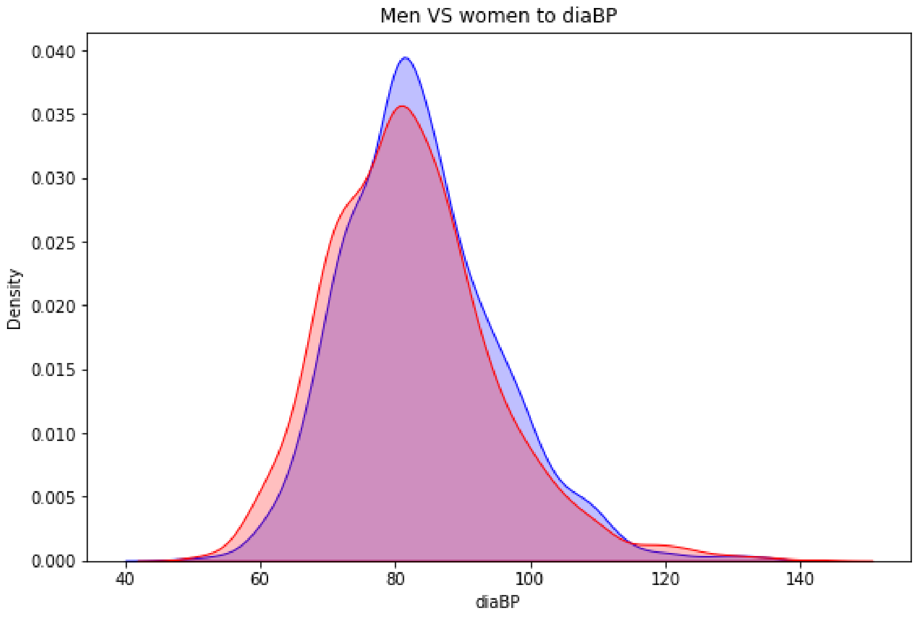
Bivariant analysis



Observation

- As age of people Increases CHD risk increased
- Education Level: of every group had both almost same proportion between having and not having risk however no risk is greater.
- Gender, 0=Male had high chances of CHD Risk , 1 = Female had more coot on no risk.
- patient is a smoker 1= yes, 0=no have same proportion to having and not having risk
- Number of Cigarettes smoked per day if increased defiantly increased the CHD risk
- Blood Pressure Medications 0 = no ,1=yes has more CHD Risk
- Prevalence of hypertension 0=none with more No Risk , 1= has prevalence hypertension had equal share of CHD risk
- patient has diabetes 0=no with more No Risk, 1=yes had equal share of CHD risk
- Every range of total Cholesterol have both NO Risk and CHD risk.
- systolic blood pressure has CHD risk if increased.
- diastolic blood pressure has CHD risk if increased.
- Body Mass Index has CHD risk if increased.
- Heart Rate has almost equal distribution of NO Risk and CHD risk
- Glucose has CHD risk if increased

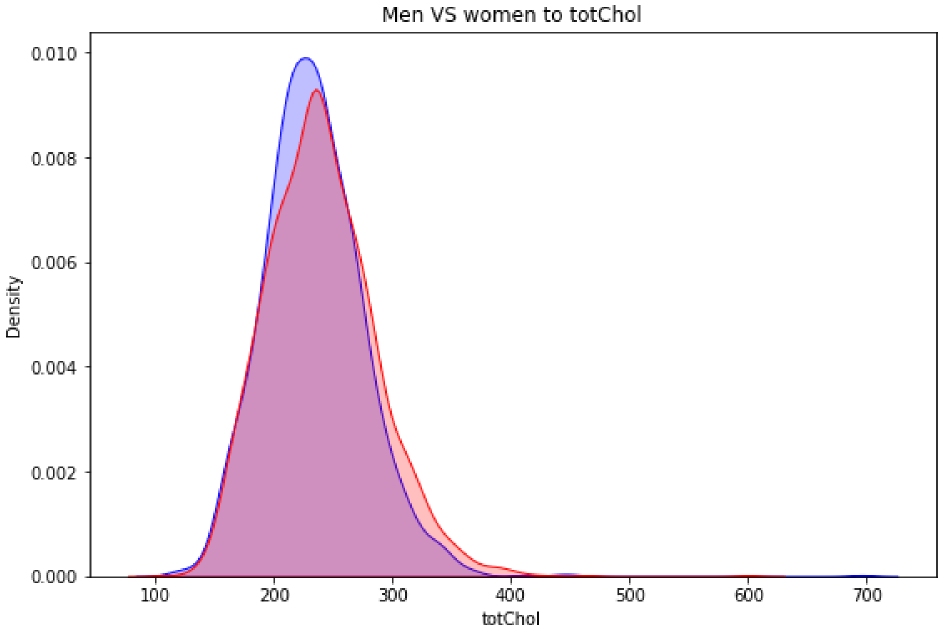

Men VS women to sysBP

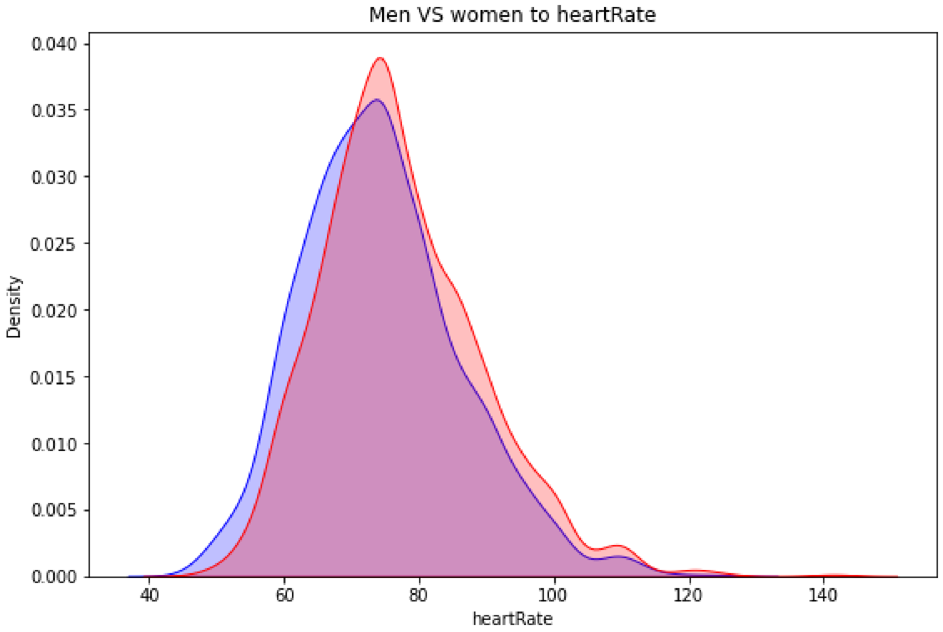

Men VS women to diaBP

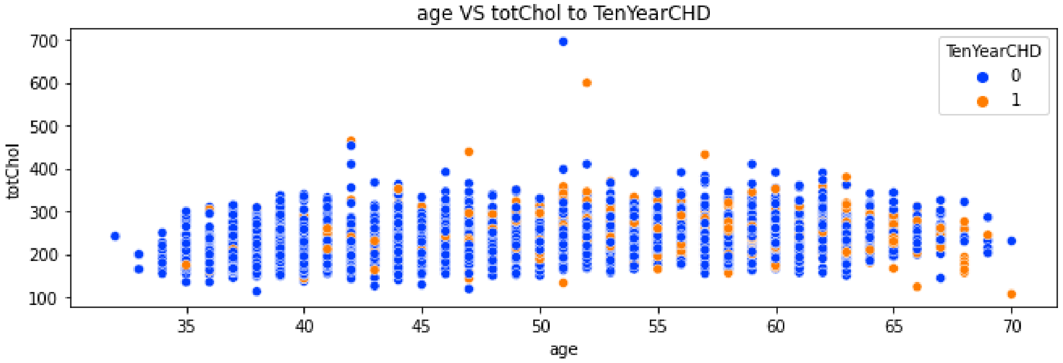
Observation

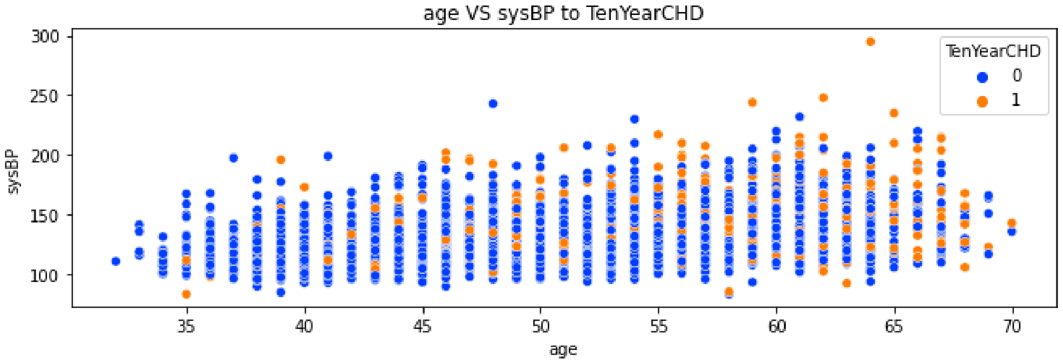
- Man and Women had same distribution over range of systolic blood pressure and diastolic blood pressure

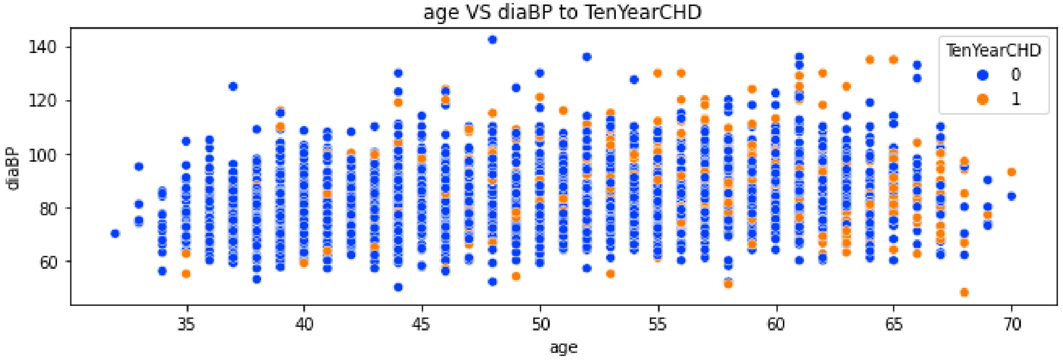
- Women had little more distribution in high range of both in total cholesterol and heartrate

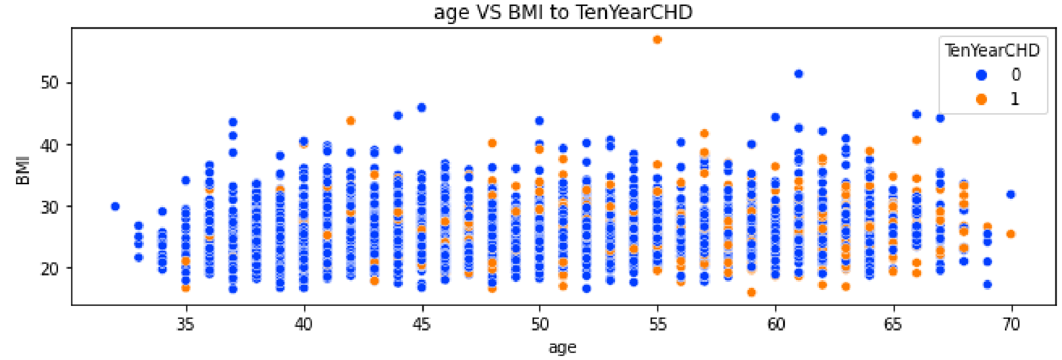

Men VS women to totChol

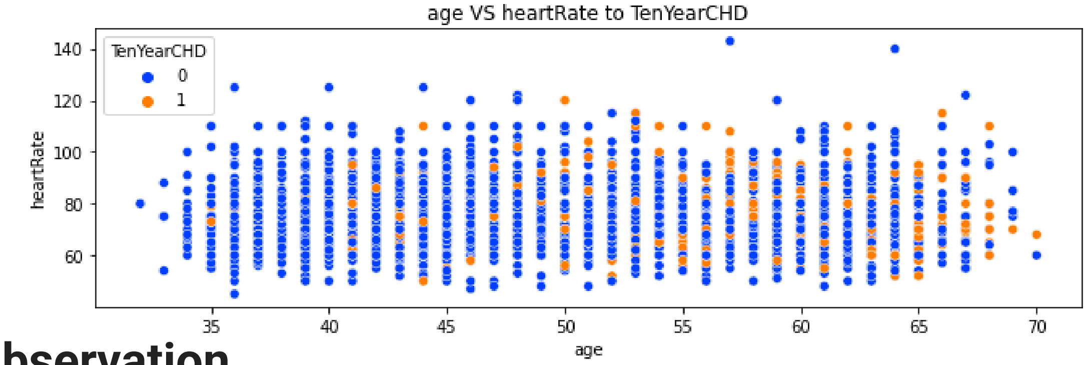

Men VS women to heartRate

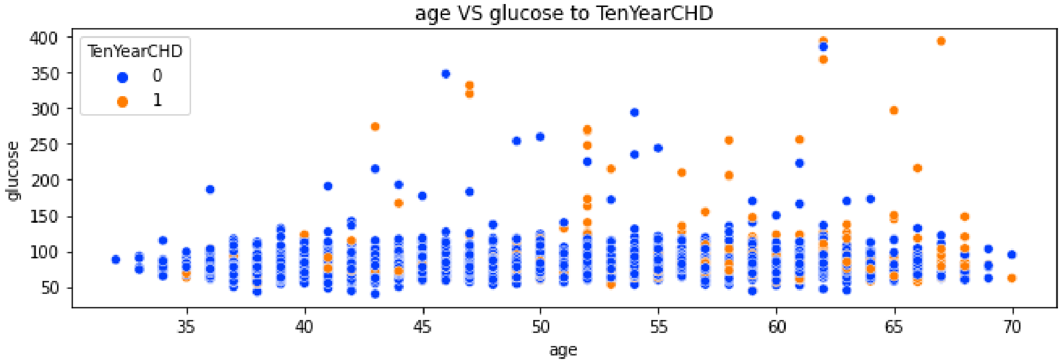








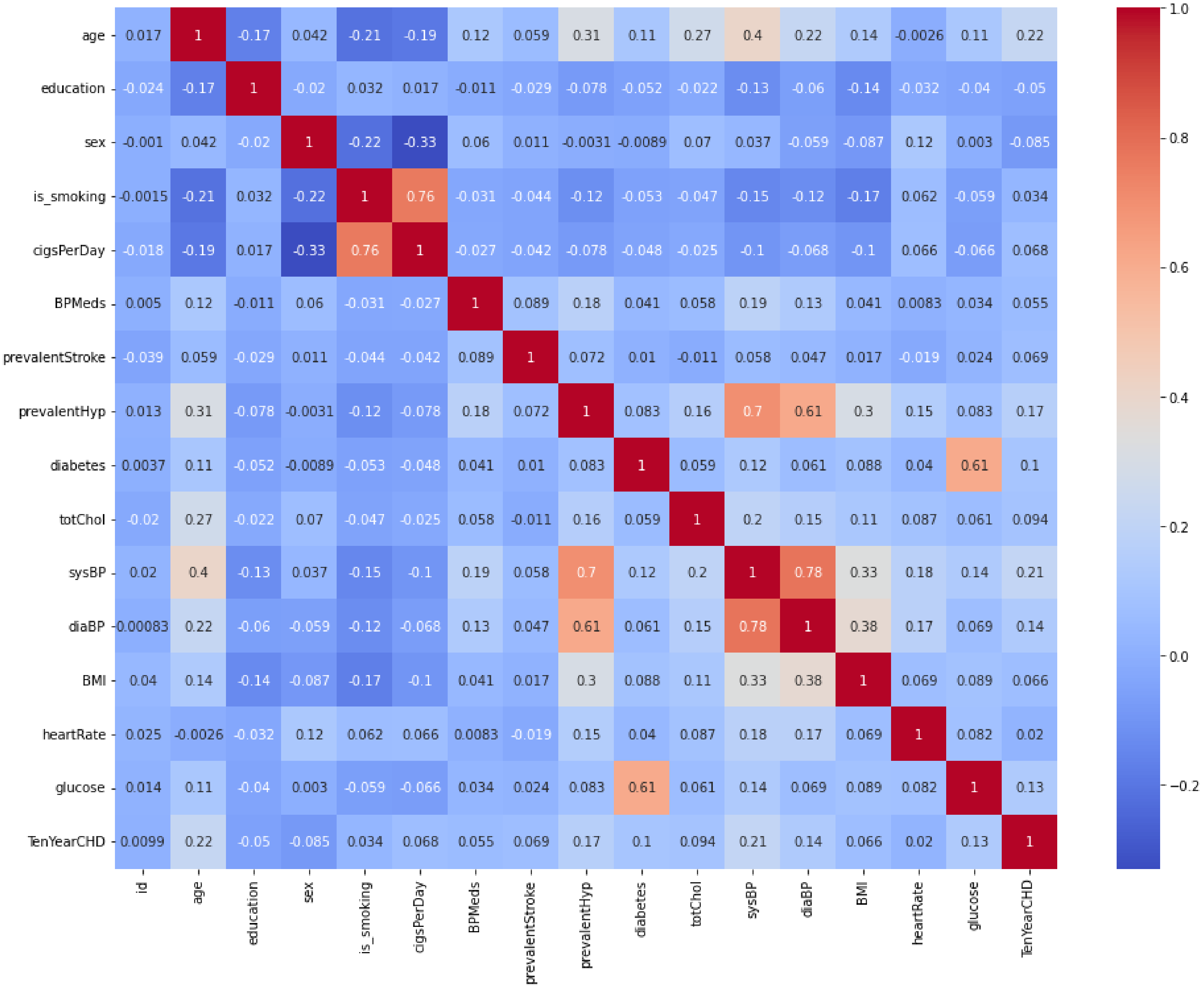




Observation

- Different age group had almost same share of totChol,BMI,heart rate ,sysBP and diaBP but as age is increased CHD risk increased
- As age increases there was little increase in count for glucose increased which also had high Risk in CHD

Correlation Heatmap

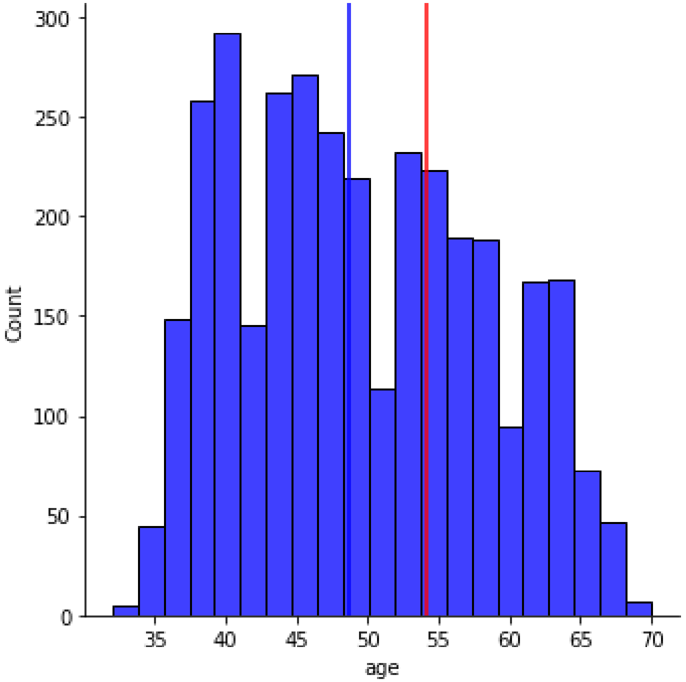


Observation

- Age has strongest weight in predicting CHD , followed by blood pressure variable,glucose and Prevalence of hypertension.
- Blood glucose and presence of diabetes are closely related, also systolic and diastolic blood pressure and is_smoking with cigerday which was expected.
- Prevalence of hypertension has strong relation with systolic blood pressure and diastolic blood pressure.

Hypothesis Testing

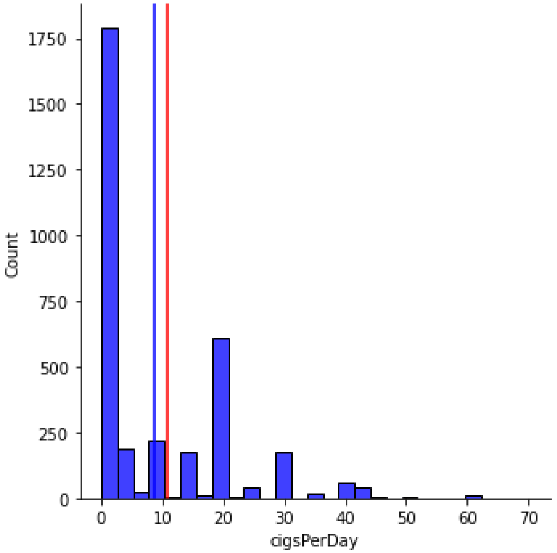
- Null Hypothesis : Increasing age donot have the same risk of developing heart disease.
- Alternate Hypothesis : Increasing age have the risk of developing heart disease.



T-testing is used to compare the means between groups of continuous data

Reject the null hypothesis. Increasing age have the risk of developing heart disease for
p = 1.8455541117753288e-38

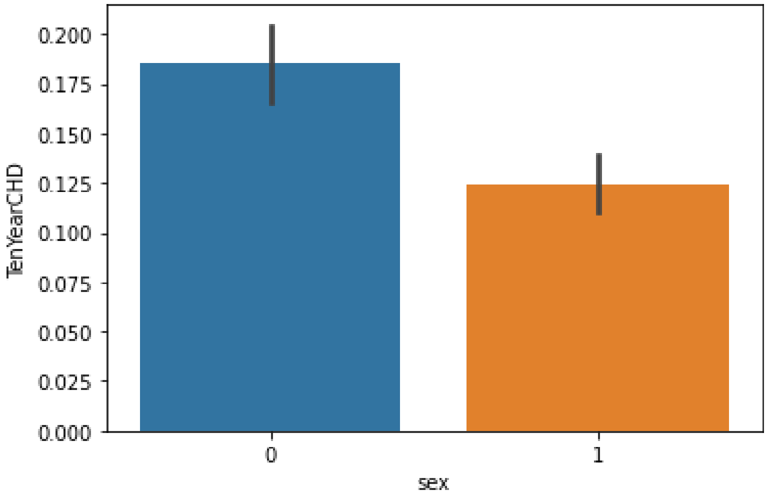
- Null Hypothesis : CigsPerDay donot have the same risk of developing heart disease.
- Alternate Hypothesis : CigsPerDay have the risk of developing heart disease.



T-testing is used to compare the means between groups of continuous data

Reject the null hypothesis.cigsPerDay have the risk of developing heart disease
p = 0.0002813024053734548

- Null Hypothesis : There is no relationship between gender and CHD risk
- Alternate Hypothesis : There is a relationship between gender and CHD risk



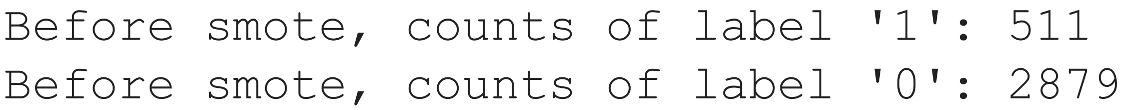
Chi-Squared testing is used when testing statistical independence or association between categorical variables.
Reject the null hypothesis, and there is a relationship between gender and CHD risk
P = 1.060878293561798e-06

Variable _Selection

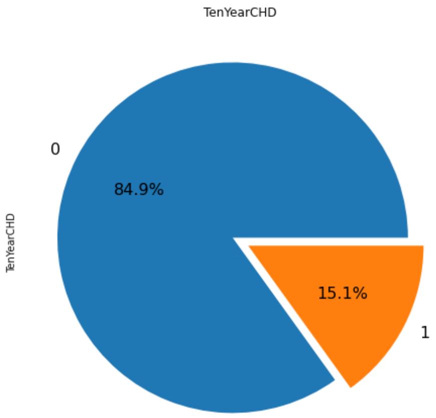
After doing VIF ,below given variables are used for modelling

variables	VIF
age	6.69
education	3.76
sex	2.48
is_smoking	4.75
cigsPerDay	4.02
BPMeds	1.09
prevalentStroke	1.02
prevalentHyp	1.65
diabetes	1.04

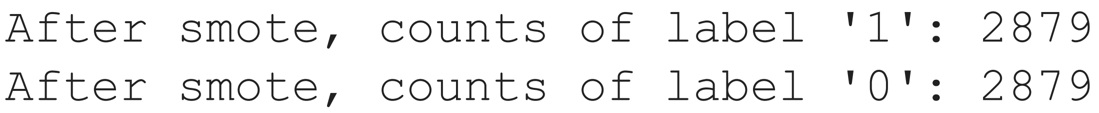
Note : Since our data is imbalanced **SMOTE** is used for balancing the data before model implementation.



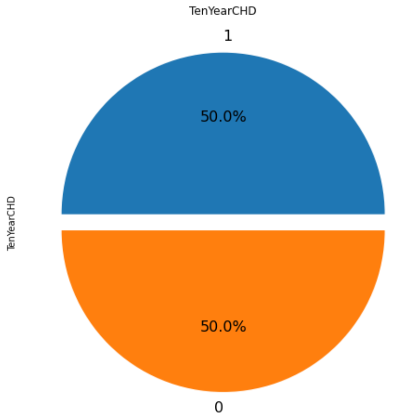
```
Before smote, counts of label '1': 511
Before smote, counts of label '0': 2879
```



Our data is highly imbalanced as 0 has 85% and 1 had only 15%



```
After smote, counts of label '1': 2879
After smote, counts of label '0': 2879
```



Model Implementation

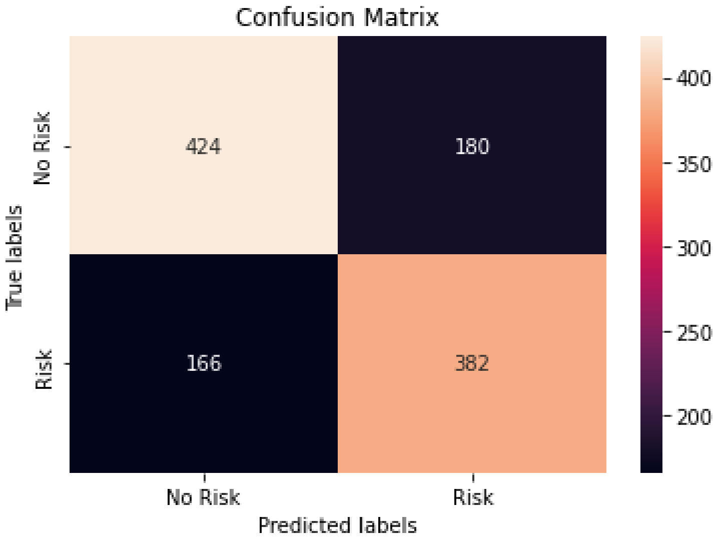
- Since our project is classification, I used Logistic ,random forest and XGBoost model with hyperparameter tuning
- Grid Search CV for hyperparameter tuning
- Stratified K_Fold for cross validation as It maintains the same class ratio throughout the K folds as the ratio in the original dataset.

Model Evaluation:

- Accuracy Score.
- Precession.
- Recall.
- F1 score.

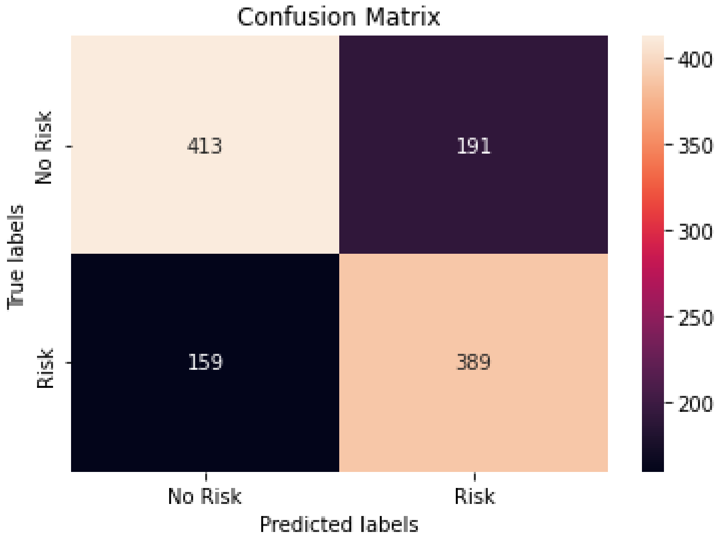
Logistic Regression:

Model Before hyperparameter tunning



```
The accuracy on train data is 0.667390360399479
The accuracy on test data is 0.6996527777777778

 precision  recall  f1-score
0 0.719     0.702   0.710
1 0.680     0.697   0.688
```



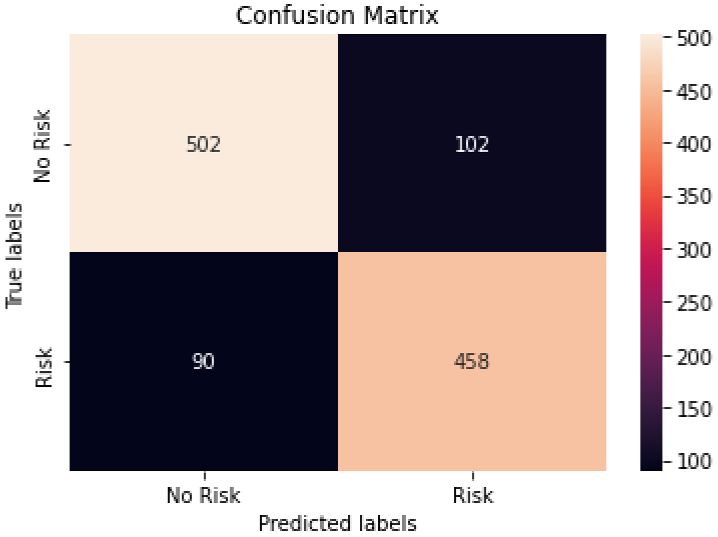
```
The accuracy on train data is 0.6682587928788537
The accuracy on test data is 0.6961805555555556

  precision recall f1-score
0 0.722     0.684  0.702
1 0.671     0.710  0.690
```

Not a big difference is seen here even after tunning.

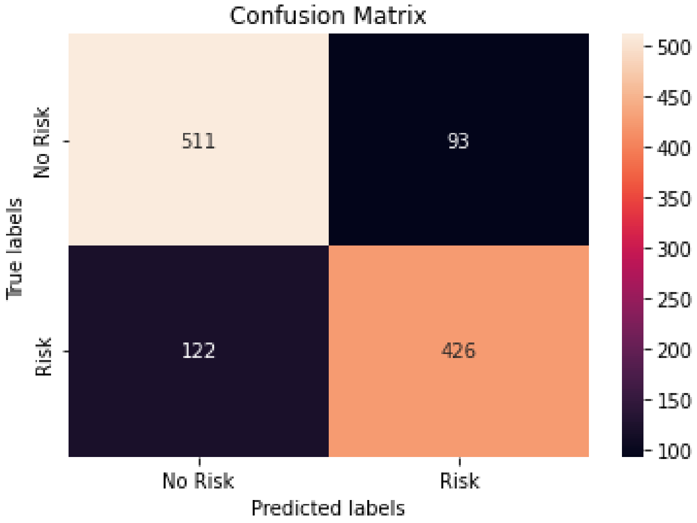
Random Forest Classifier:

Model Before hyperparameter tunning



```
The accuracy on train dataset is 0.9122883195831524
The accuracy on test dataset is 0.8333333333333334

  precision recall f1-score
0 0.848     0.831  0.839
1 0.818     0.836  0.827
```



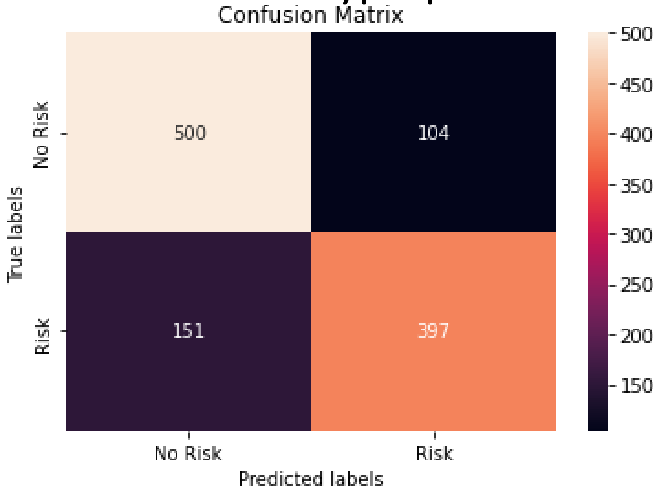
```
The accuracy on train dataset is 0.8334780720798958
The accuracy on test dataset is 0.8107638888888888

  precision recall f1-score
0 0.807     0.846   0.826
1 0.821     0.777   0.799
```

After tuning train and test had very less difference in accuracy, which helped model in overfitting.

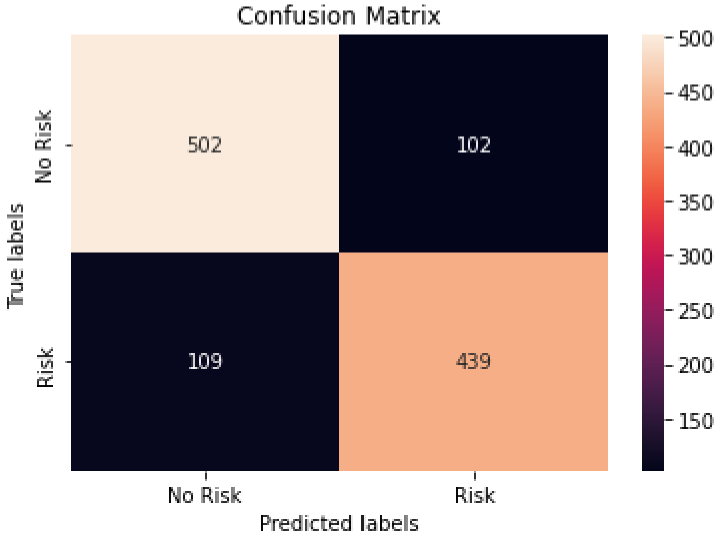
XGBClassifier:

Model Before hyperparameter tunning



```
The accuracy on train dataset is 0.7809379070777247
The accuracy on test dataset is 0.7786458333333334

   precision recall f1-score
0  0.768      0.828   0.797
1  0.792      0.724   0.757
```

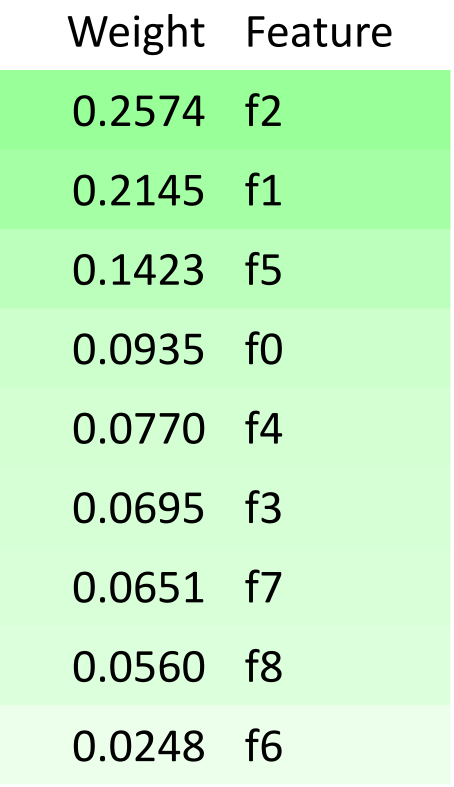


```
The accuracy on train dataset is 0.8627876682587928
The accuracy on test dataset is 0.8168402777777778
  precision recall f1-score
0 0.822      0.831   0.826
1 0.811      0.801   0.806
```

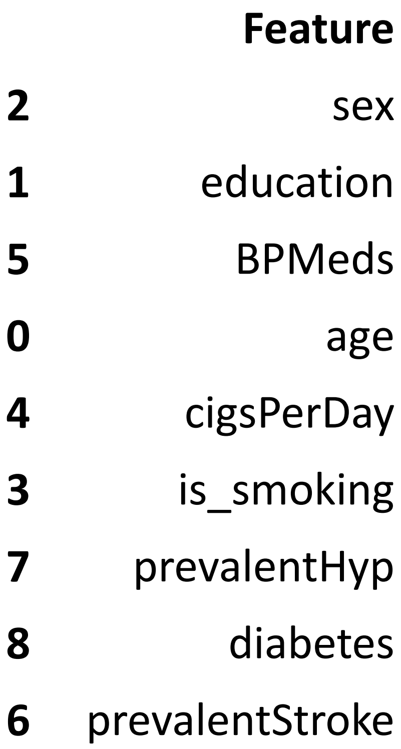
Xgboost had performed well after tunning.

Best estimator:

XGB_optimal_model

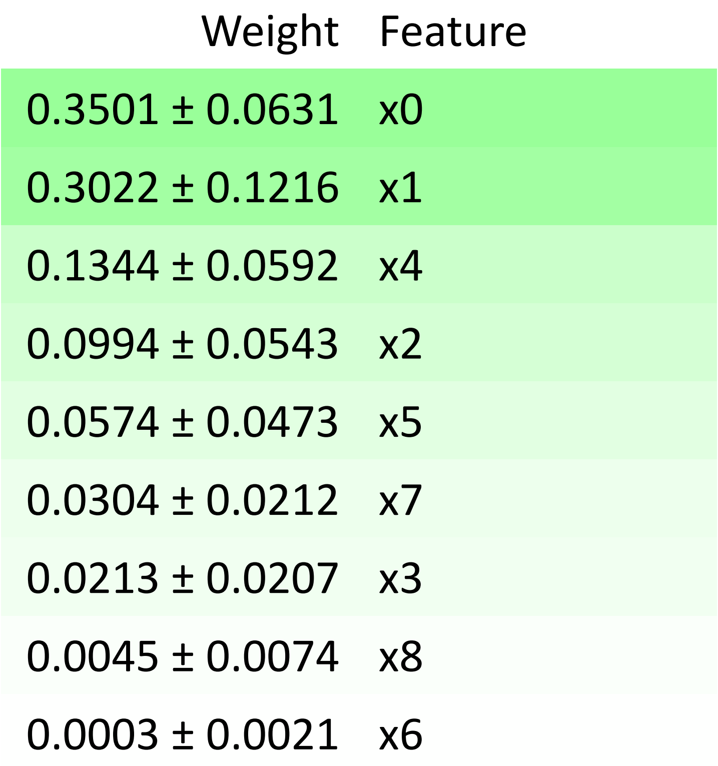


Weight	Feature
0.2574	f2
0.2145	f1
0.1423	f5
0.0935	f0
0.0770	f4
0.0695	f3
0.0651	f7
0.0560	f8
0.0248	f6

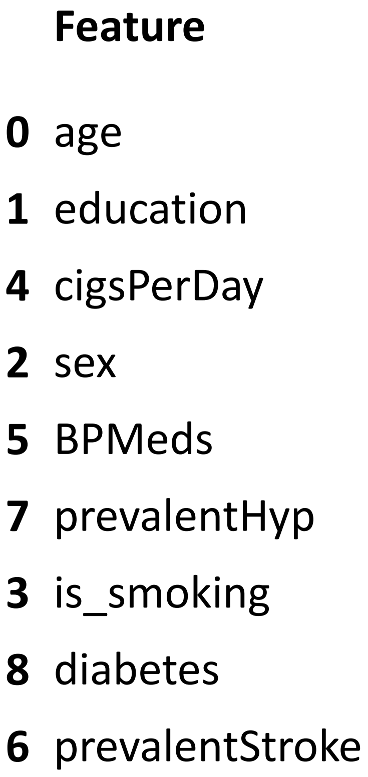


	Feature
2	sex
1	education
5	BPMeds
0	age
4	cigsPerDay
3	is_smoking
7	prevalentHyp
8	diabetes
6	prevalentStroke

rf_optimal_model



Weight	Feature
0.3501 ± 0.0631	x0
0.3022 ± 0.1216	x1
0.1344 ± 0.0592	x4
0.0994 ± 0.0543	x2
0.0574 ± 0.0473	x5
0.0304 ± 0.0212	x7
0.0213 ± 0.0207	x3
0.0045 ± 0.0074	x8
0.0003 ± 0.0021	x6



	Feature
0	age
1	education
4	cigsPerDay
2	sex
5	BPMeds
7	prevalentHyp
3	is_smoking
8	diabetes
6	prevalentStroke

Conclusion.

- Dependent variable i.e. TenYearCHD has little correlation between features and of all the variable only Age has strongest weight in predicting CHD , followed by blood pressure variable, glucose and Prevalence of hypertension.
- Gender variable has large positive correlation in predicting dependent variable for XGBmodel,where as Age for RF model. Education had equal importance in both the model
- After Hyper parameter tuning Logistic regression did not have much difference , but for Rf train and test had very less difference in accuracy, which helped model in overfitting.
- Over all XBG Xgboost had performed well after tunning

References :-

1. mygreatlearning.com
2. GeeksforGeeks
3. Analytics Vidhya
4. Almabetter Notes.