CAPSTONE PROJECT

Cardiovascular-Risk-Prediction

TEAM MEMBERS

KIRTESH VERMA
PRAVIN BEJJO
SAHIL PARDESHI

➔ STEPS

- ❖ Problem statement
- ❖ Overview of the data
- ❖ EDA
- ❖ Data Preprocessing
- ❖ Evaluation of Models
- ❖ Conclusion

➔ PROBLEM STATEMENT

- **Heart disease is one the major cause of morbidity and mortality globally. A heart attack happens when the flow of oxygen-rich blood to a section of heart muscle suddenly becomes blocked and the heart can't get oxygen. If blood flow isn't restored quickly, the section of heart muscle begins to die.**
- **The dataset is from an ongoing cardiovascular study on residents of the town of Framingham, Massachusetts.**
- **Our goal is to predict whether the patient has a 10-year risk of future coronary heart disease (CHD) based on their present health conditions using different Machine Learning Techniques.**

➔ DATA OVERVIEW

Breakdown of Our Features:

- **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)**

DATA OVERVIEW(Contin…)

- 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”) -DV

➔ DATA OVERVIEW(Contin…)

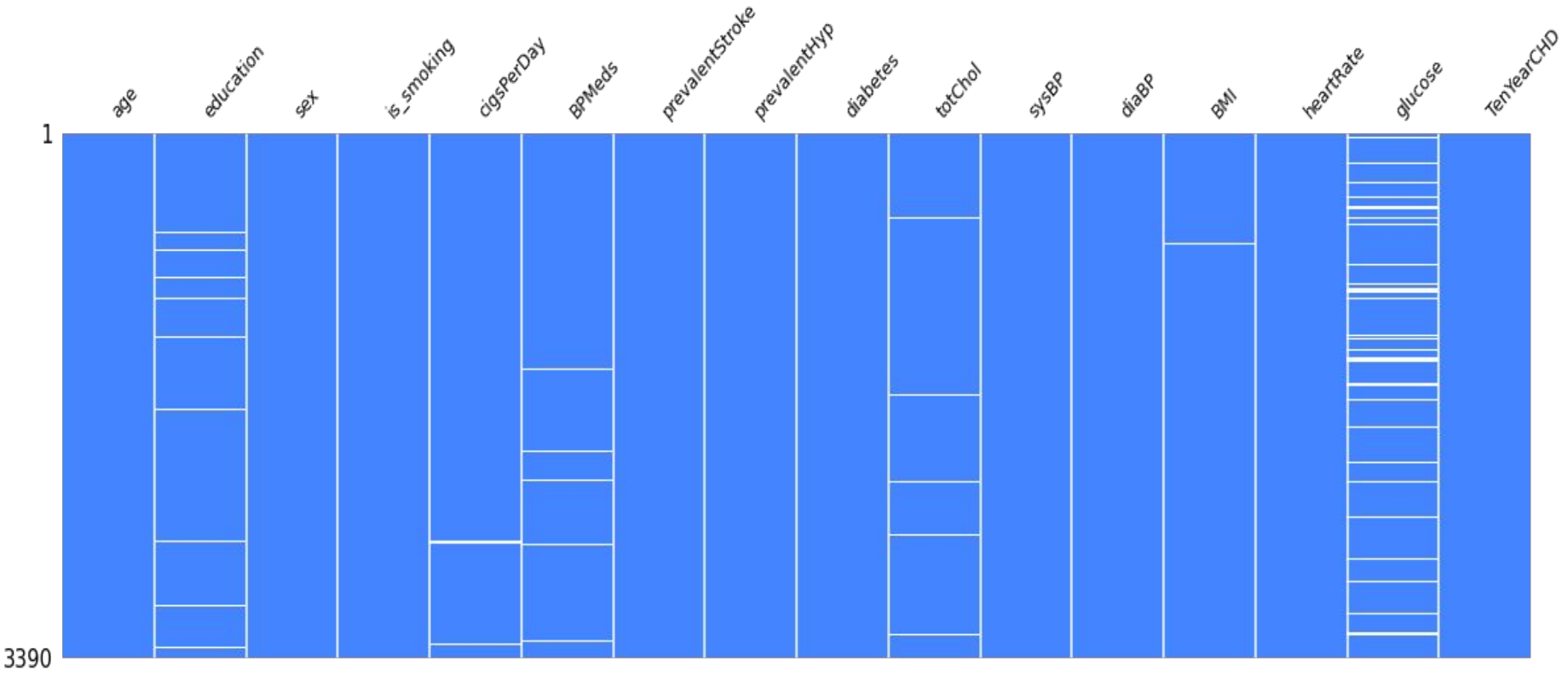
DATA INSIGHTS

- The given dataset has 17 columns and 3390 entries.
- The columns belonging to data type ‘Object’ are ‘sex’ and ‘is_smoking’.
- There are 7 columns in the dataset containing null values. They are ‘education’, ‘cigsPerDay’, ‘BPMeds’, ‘totChol’, ‘BMI’, ‘heartRate’, ‘glucose’.

```
<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
```

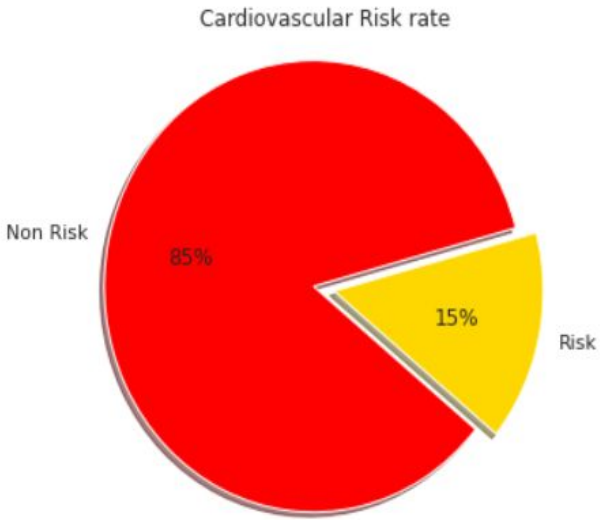
➔ EXPLORATORY DATA ANALYSIS

➔ NULL VALUES


age
education
sex
is_smoking
cigsPerDay
BPMeds
prevalentStroke
prevalentHyp
diabetes
totChol
sysBP
diaBP
BMI
heartRate
glucose
TenYearCHD
1
3390

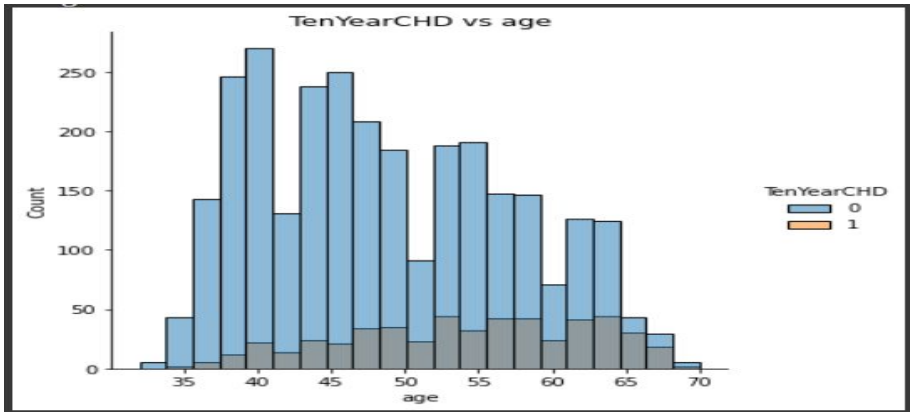
➔ NULL VALUE SUMMARY

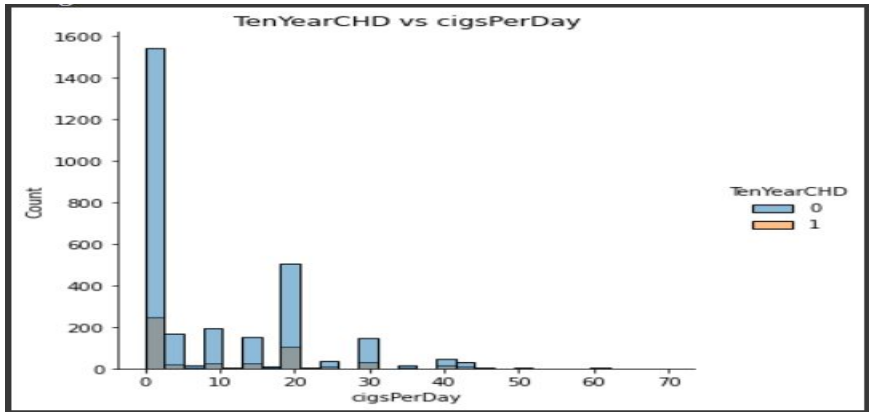
NaN Value Summary				
Feature	Total Counts	Total NaN Value Count	Minority Class having NaN Value	Percentage
Education	3390	87	13	0.38%
Cigsperday	3390	22	1	0.03%
BPMeds	3390	44	7	0.21%
totChol	3390	38	7	0.21%
BMI	3390	14	7	0.21%
heartrate	3390	1	1	0.03%
glucose	3390	304	39	1.15%
		510	75	2.21%

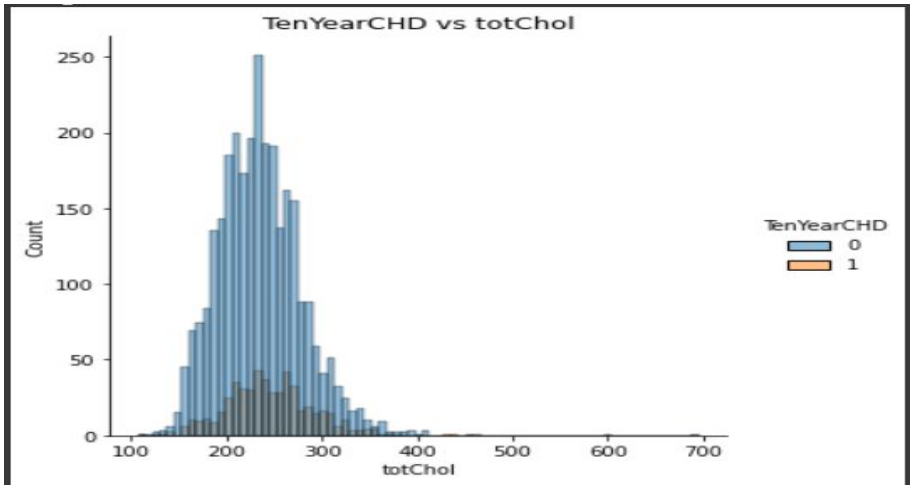


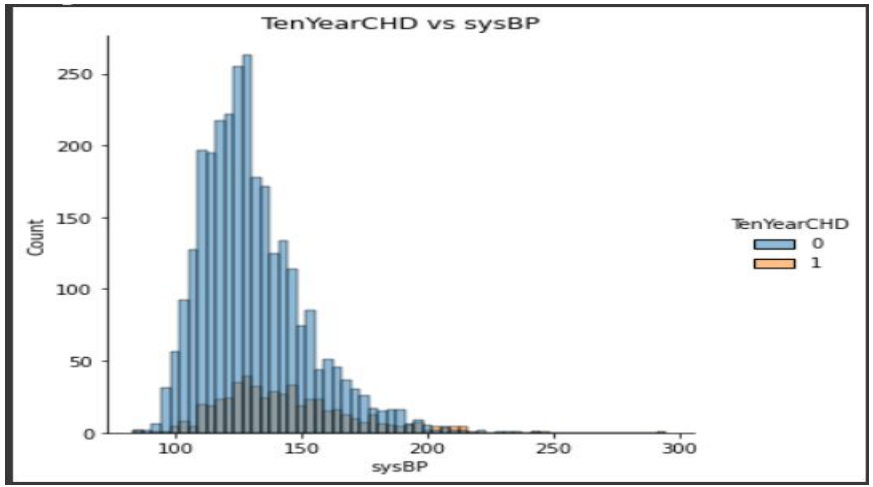
Dropping the NaN values will lead to 2.21% loss of the minority set, which is already at 15%

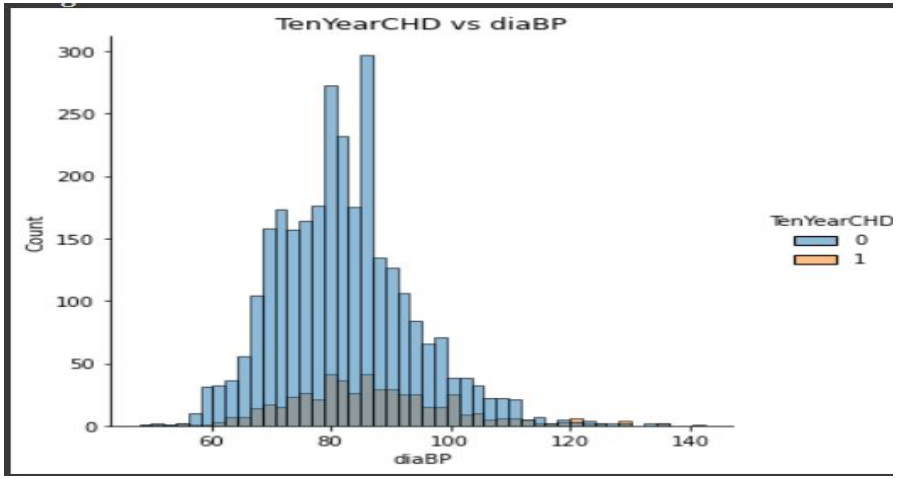
➔ DISTRIBUTION PLOT (NUMERICAL FEATURES)

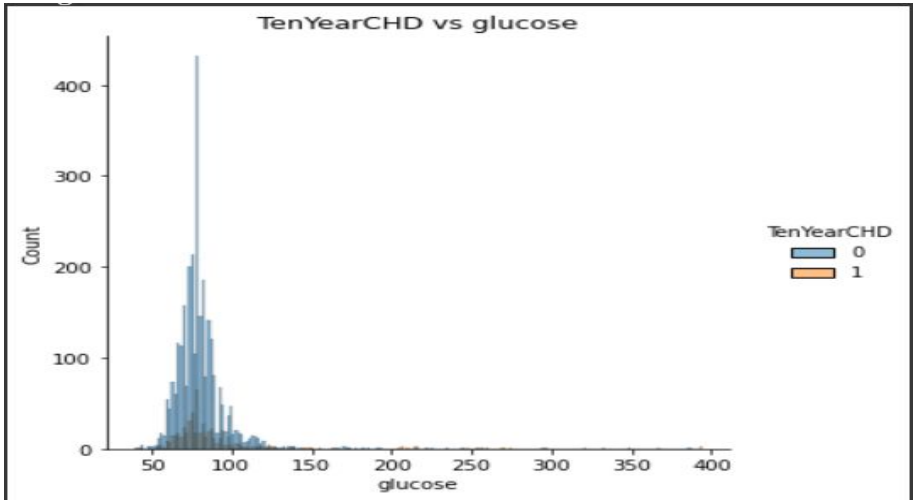


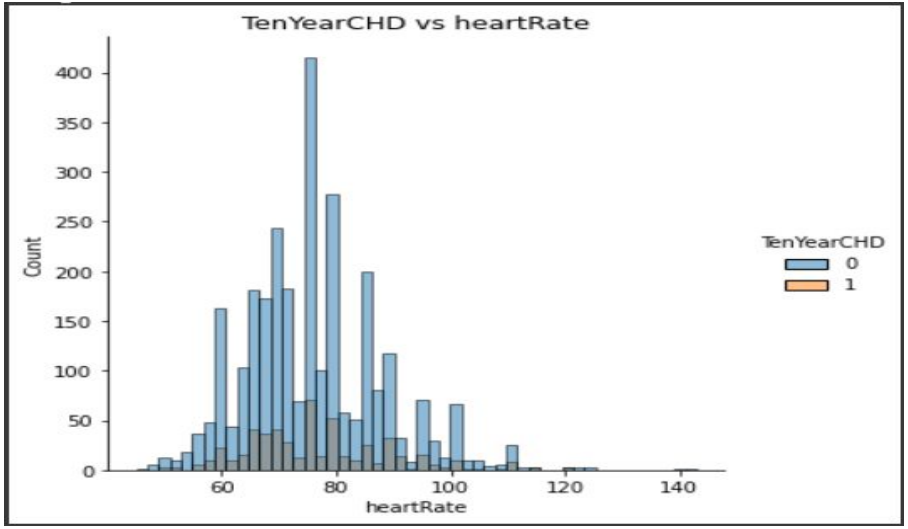


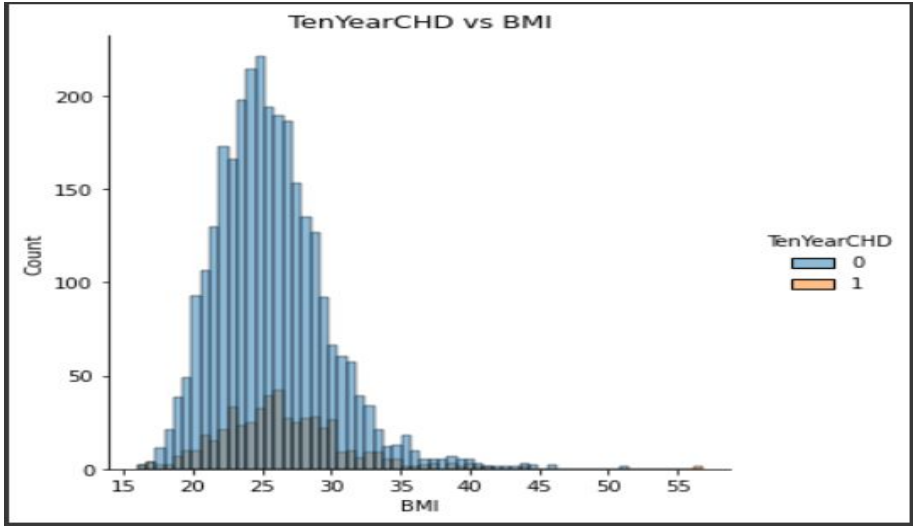





TenYearCHD vs diaBP
Count
diaBP
TenYearCHD
0
1


TenYearCHD vs glucose
Count
glucose
TenYearCHD
0
1

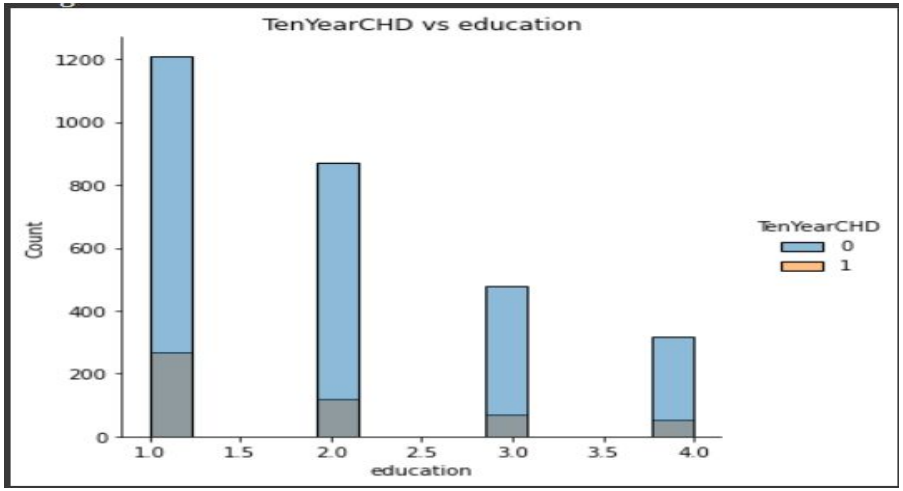

TenYearCHD vs heartRate
Count
heartRate
TenYearCHD
0
1

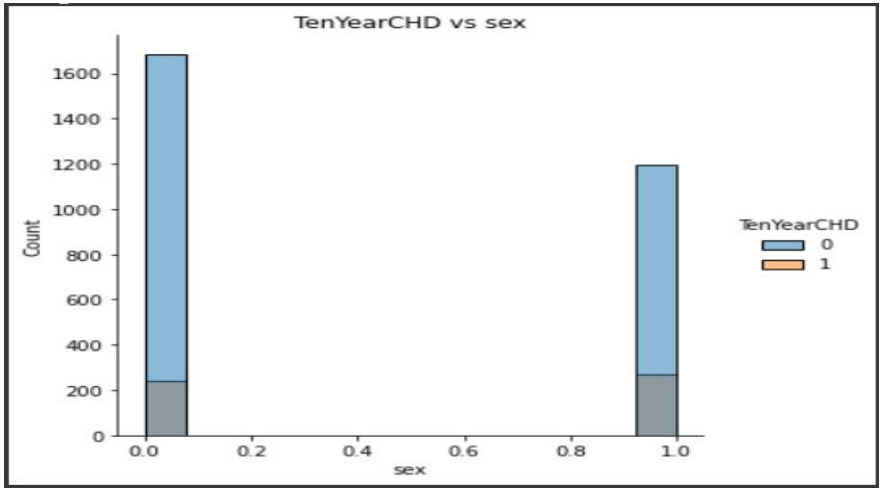

TenYearCHD vs BMI
Count
BMI
TenYearCHD
0
1

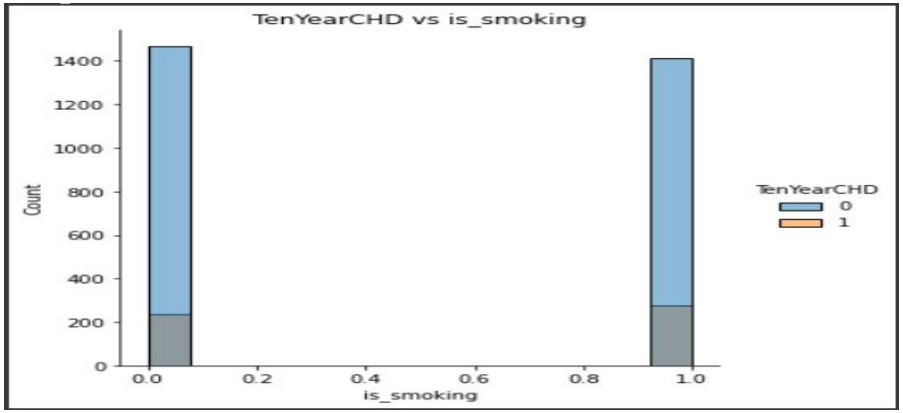
➔ INSIGHTS

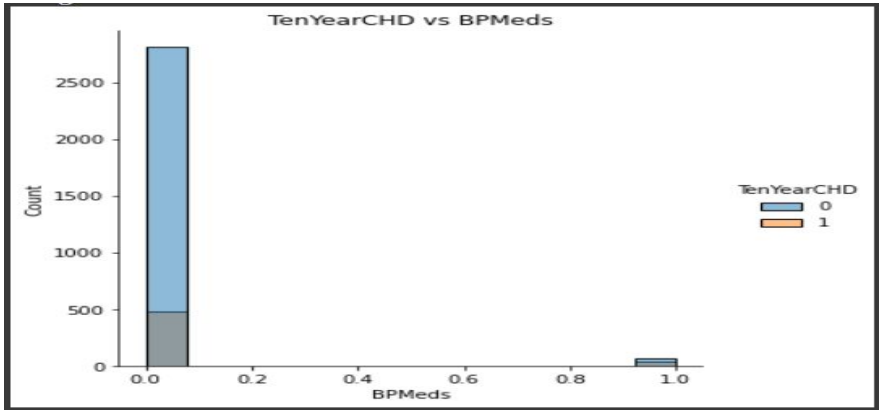
- Glucose and totalChol are highly right skewed.
- cigsPerDay, sysBP and BMI are moderately right skewed.
- Ages, diaBP and heartRate are somewhat normally distributed.

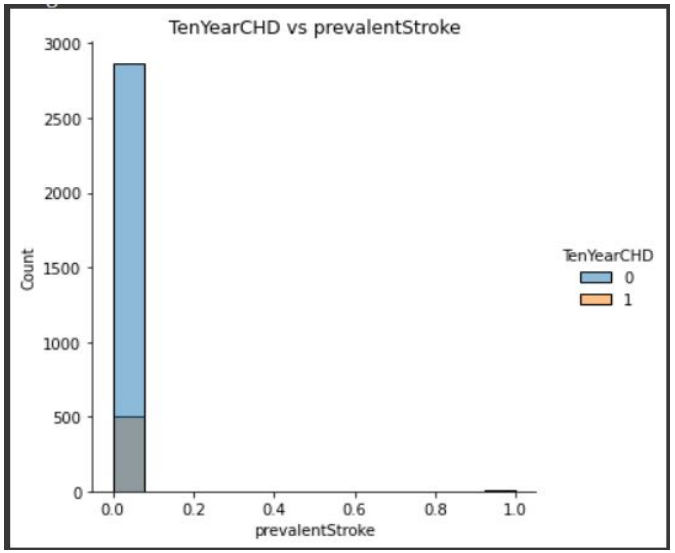
➔ DISTRIBUTION PLOT (CATEGORICAL FEATURES)

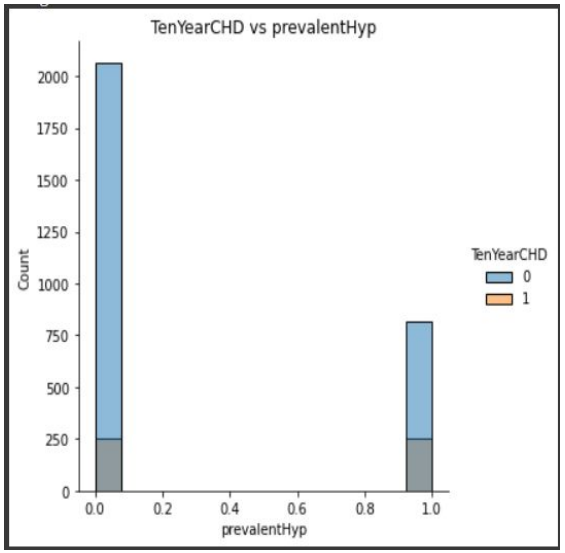


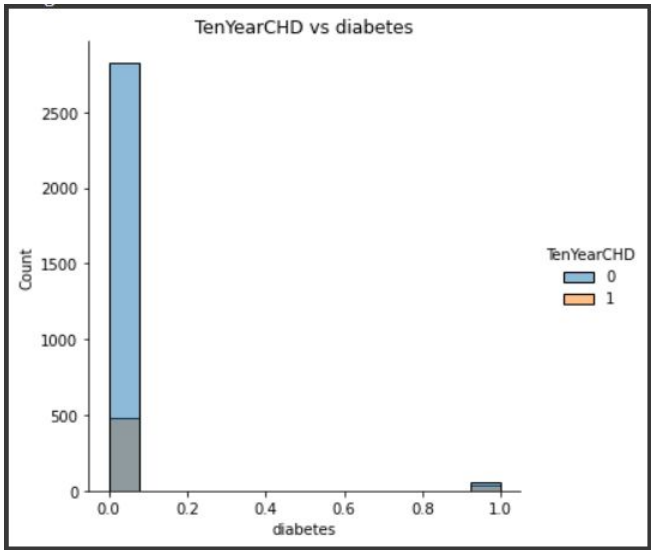






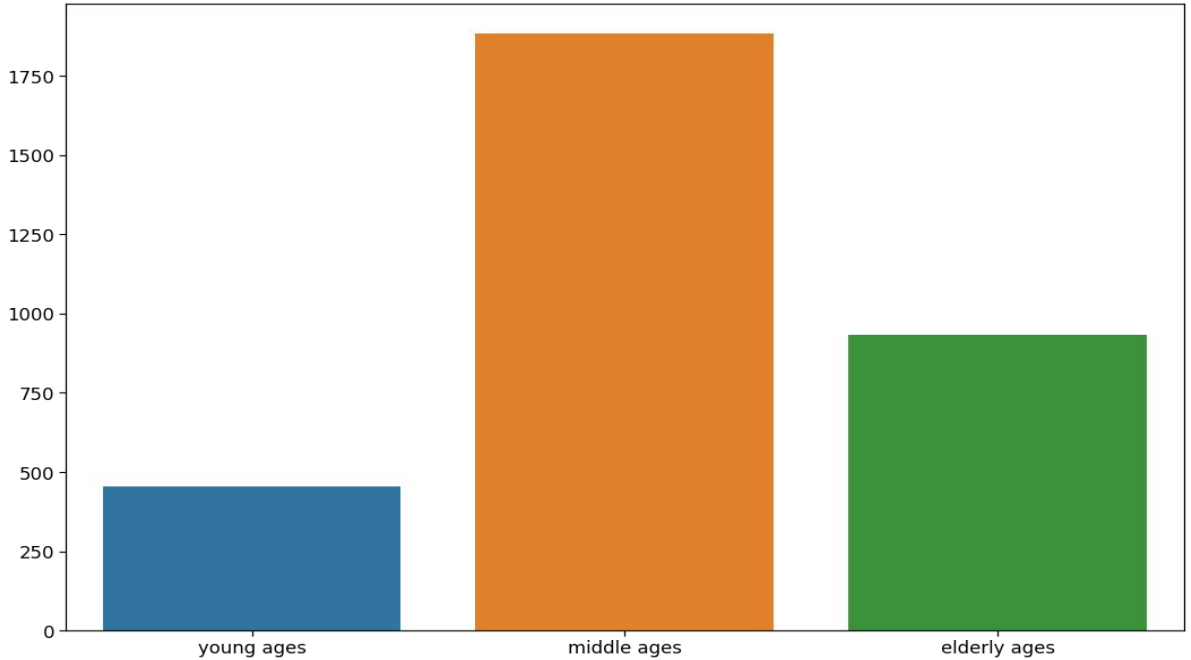






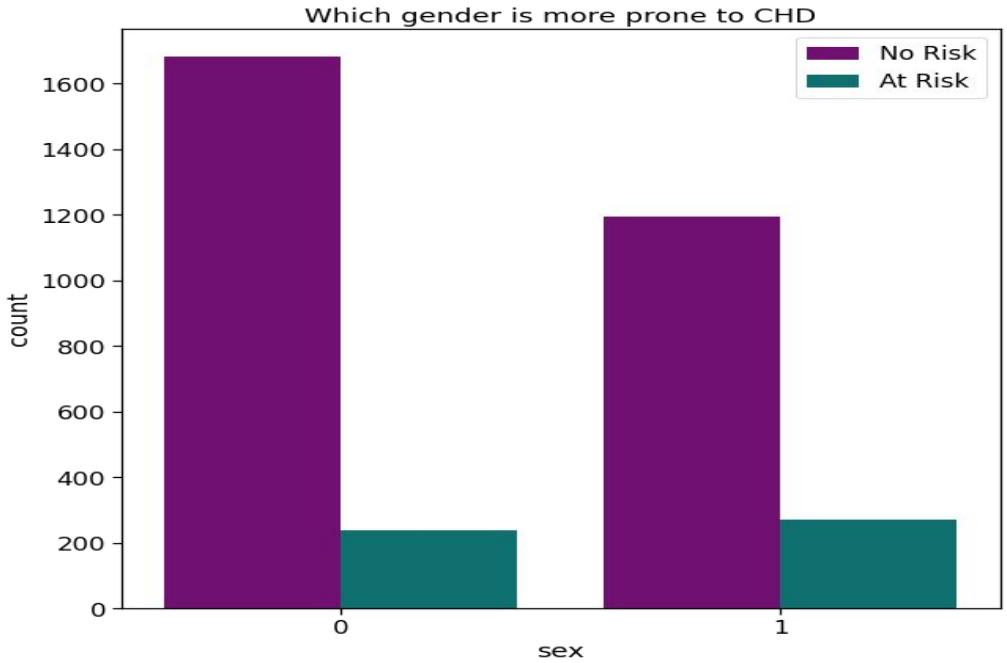
The total count of people who are on BP meds, who previously had a stroke and people who have diabetes is very less.

➔ Age wise distribution of Population



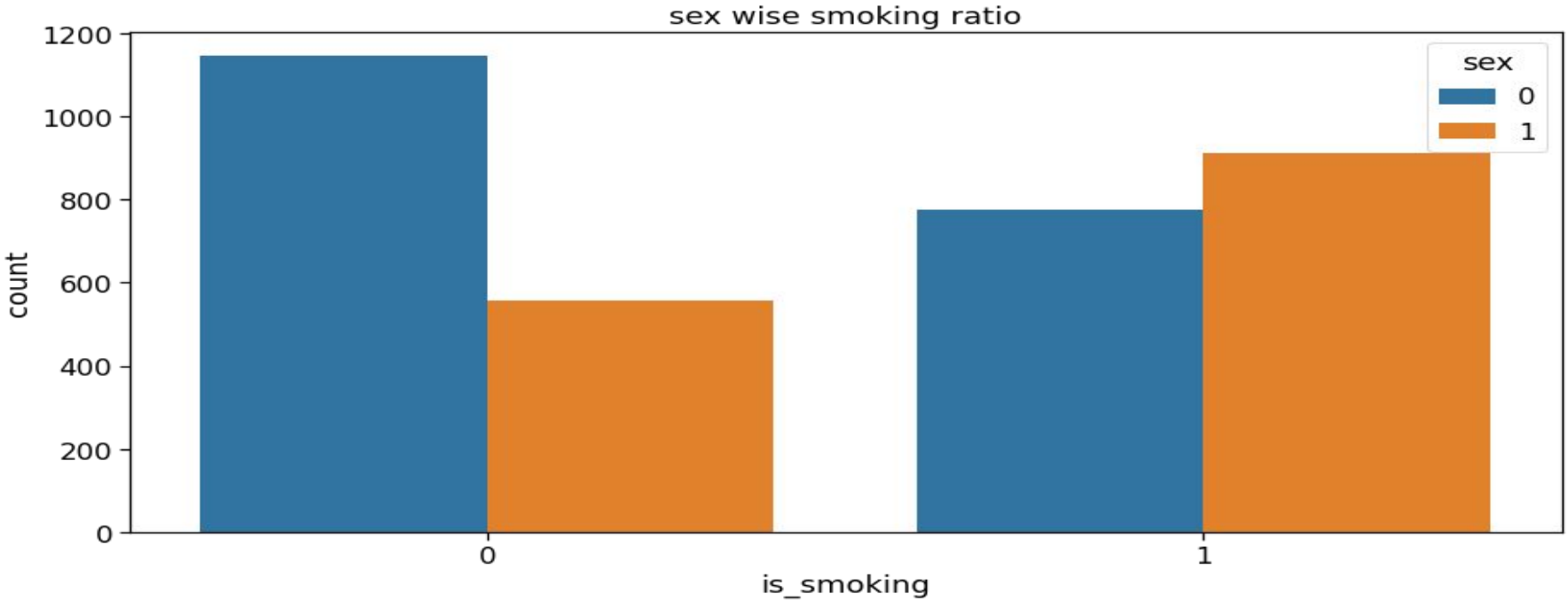
- Number of people belonging to middle age (between 40-55) group are highest in the given dataset followed by elderly age (above 55) group.
- Number of people belonging to young age group (between 29-40) are least.

➔ Gender Prone to CHD



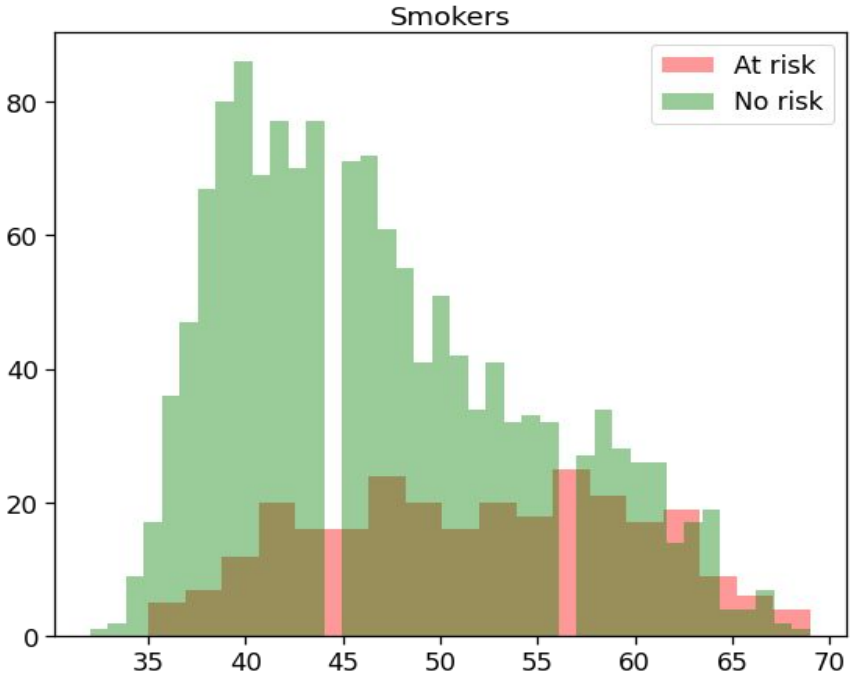
- The number of males and females which are at risk of CHD is equal.
- The number of females who are not at risk is higher than that of male.

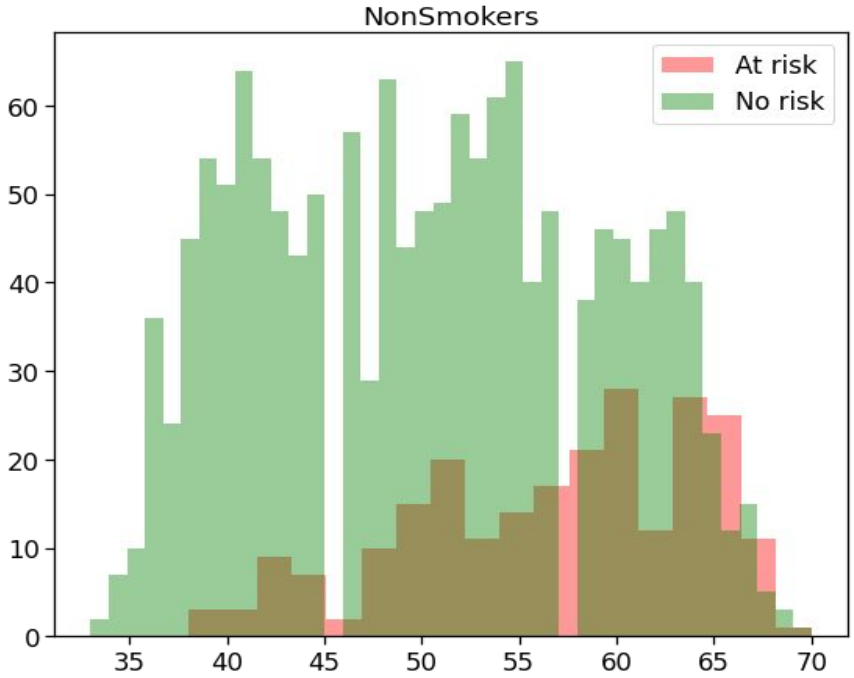
➔ Sex wise smoking ratio



- Number of male smokers is higher than female smoker even though the total count of male is less than female.
- Number of non-smoking female is higher than number of non-smoking male.

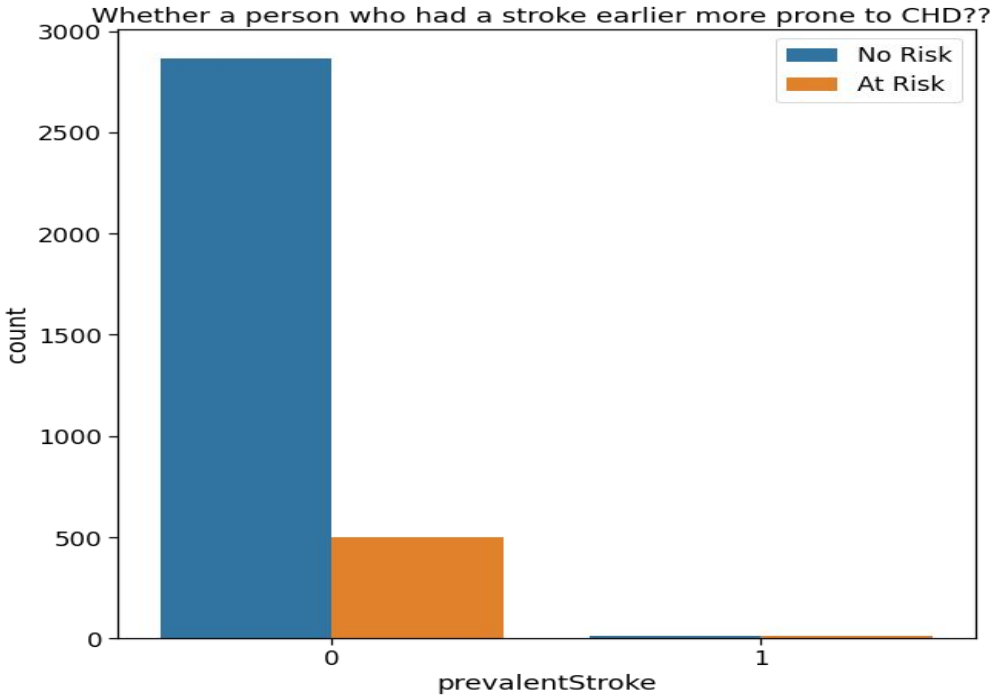
➔ Smoker and Non-Smokers





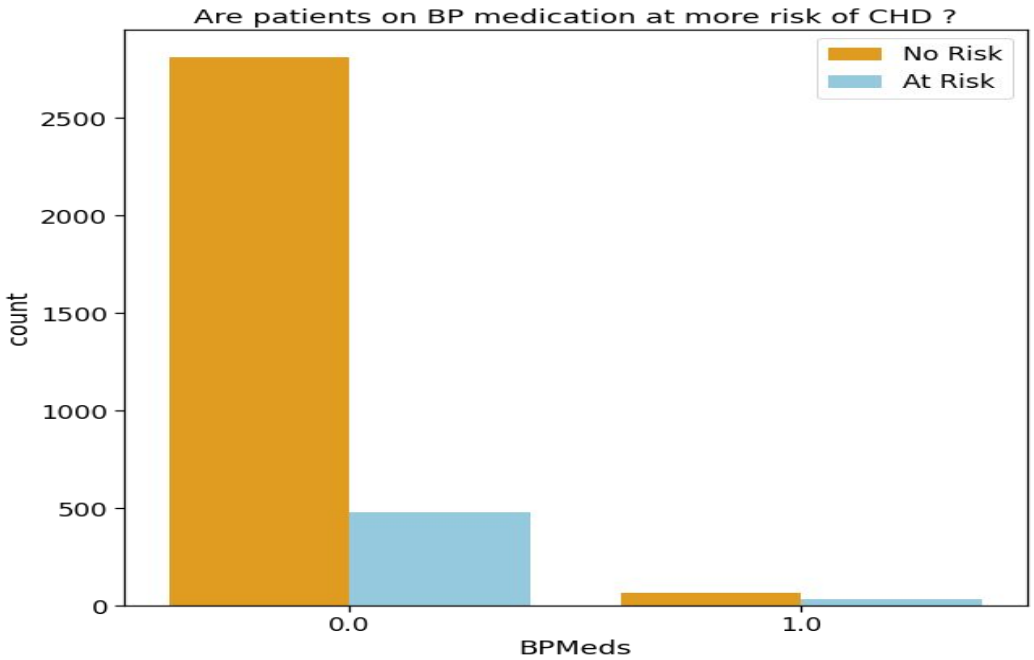
Age plays an important role in cardiovascular risk irrespective of the person smokes or not.

➔ Previous Heart Stroke



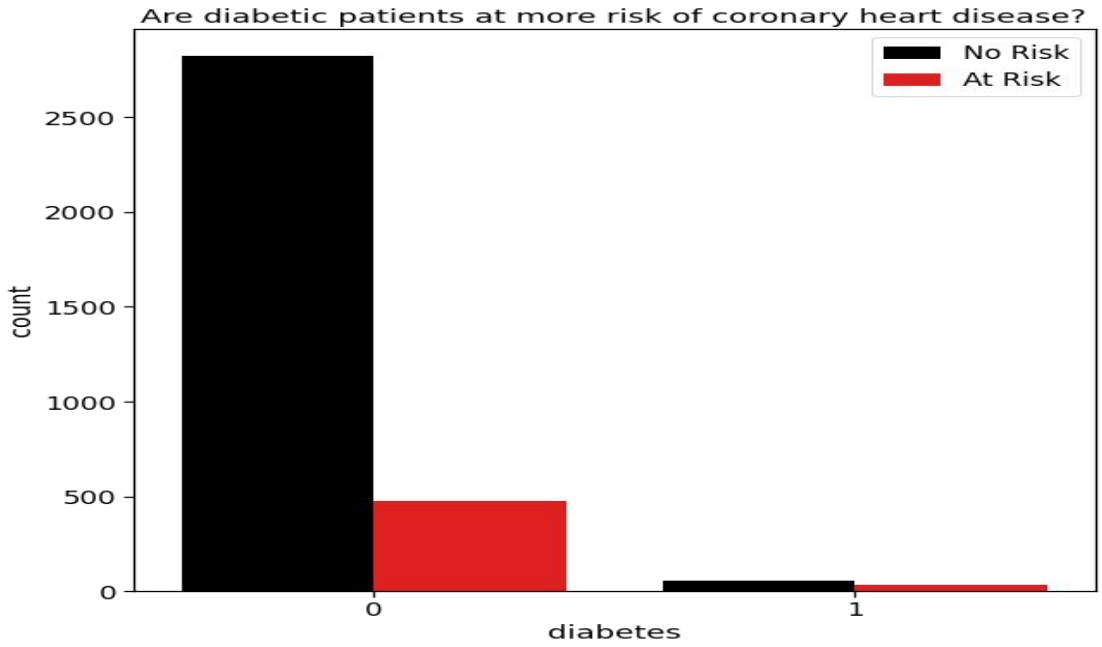
The person who previously had a heart stroke are more at risk to CHD than those who did not.

➔ Patients on BP medication



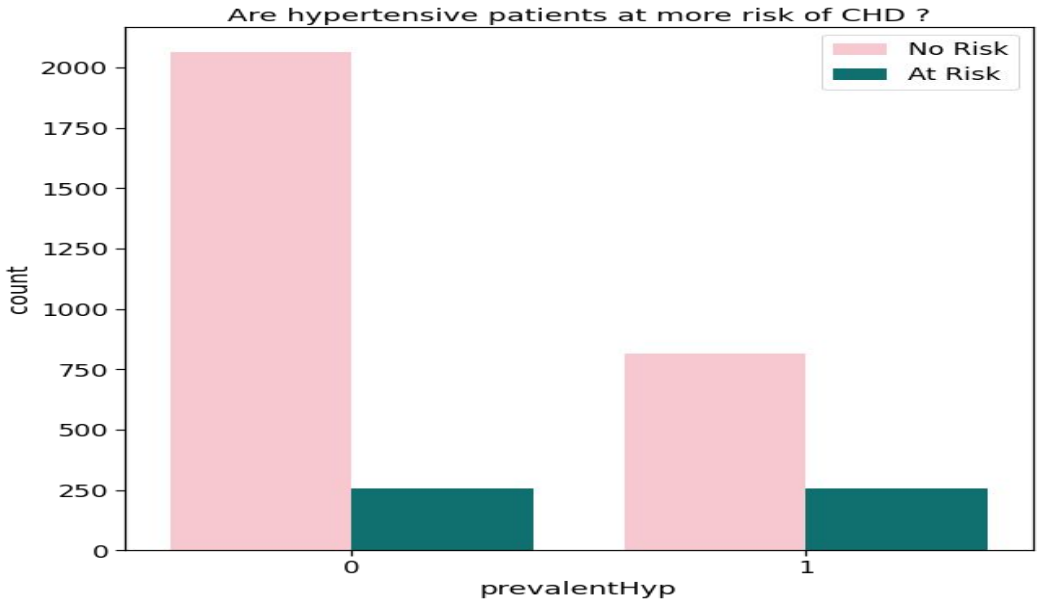
Patients on BP medication are more prone to CHD where as those who are not on BP medication has severely less chances of getting CHD.

➔ Diabetic Patients



Diabetic patients have higher risk of CHD than non-diabetic patients.

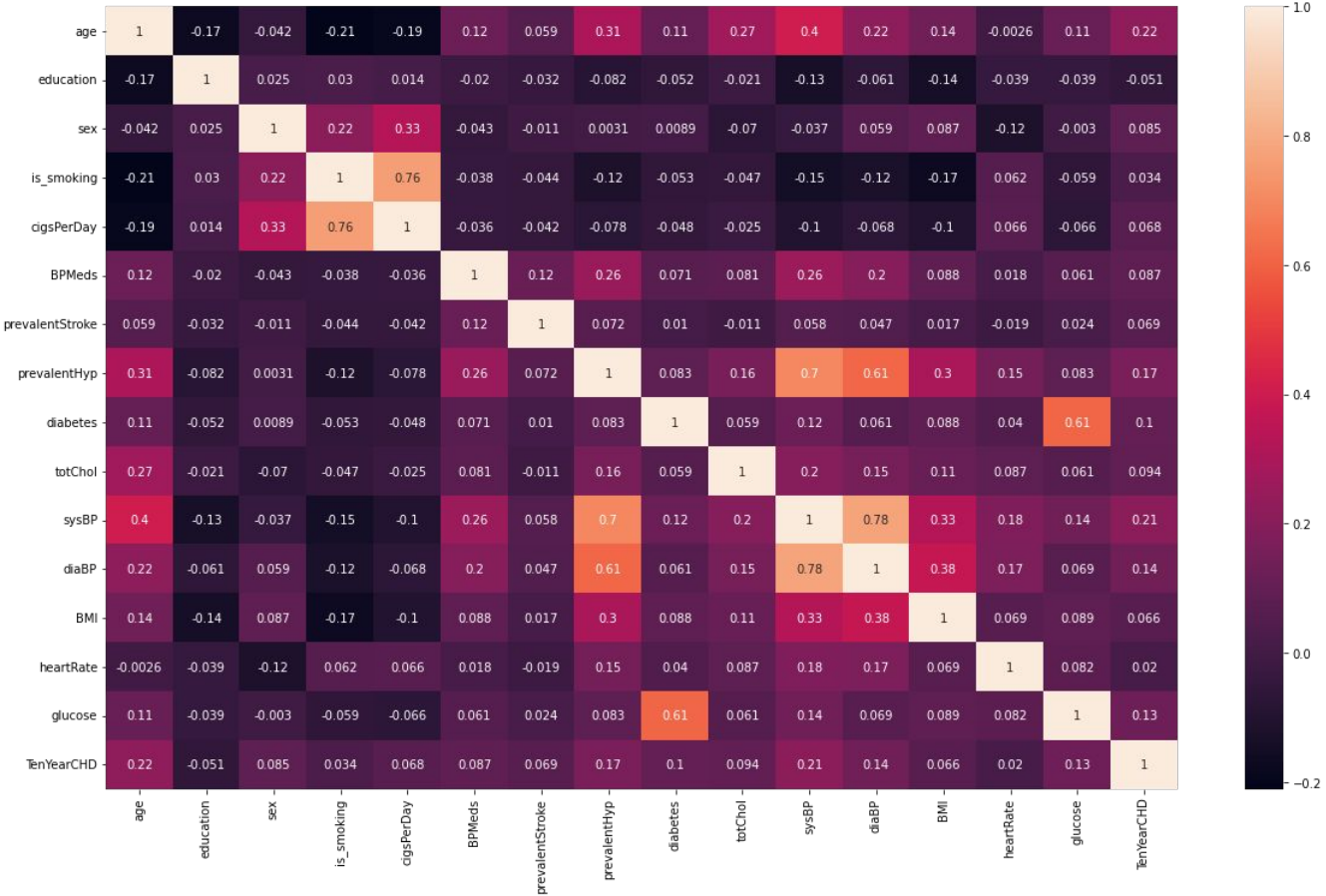
➔ Hypertensive Patients



- Out of all the people who are not Hypertensive, the number of people getting CHD is very less.
- People who are hypertensive has more chances of getting CHD.

➔ HEAT MAP

- sysBP and diaBP shows strongest positive correlation of 0.78 with one another.
- ls_smoking and cigsPerDay has a positive correlation of 0.76.



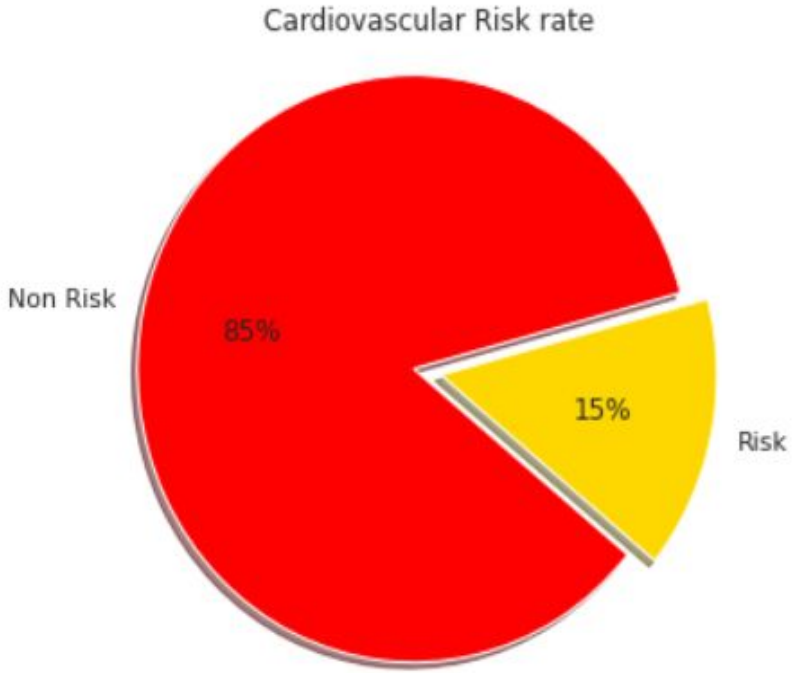
➔ DATA PREPROCESSING

- There are 7 features in the dataset containing null values. They are education, cigsPerday, BPMeds, totChol, BMI, heartRate and glucose. We filled the null values with their respective median and mode.
- We did label encoding for categorical features like ‘sex’ and ‘is_smoking’.
- As sysBP, diaBP are highly correlated with one another we feature engineered a new feature by adding both the features and dividing by 2. We named it ‘avgBP’ and dropped both the columns.

➔ DATA PREPROCESSING (Conti…)

Handling class imbalance

- Number of people who have CHD is very less .i.e. Only 15%. Even our machine learning model gives high accuracy it could be misleading.
- To balance our highly imbalanced class we use the oversampling technique called SMOTE i.e. Synthetic Minority Over-sampling Technique.”

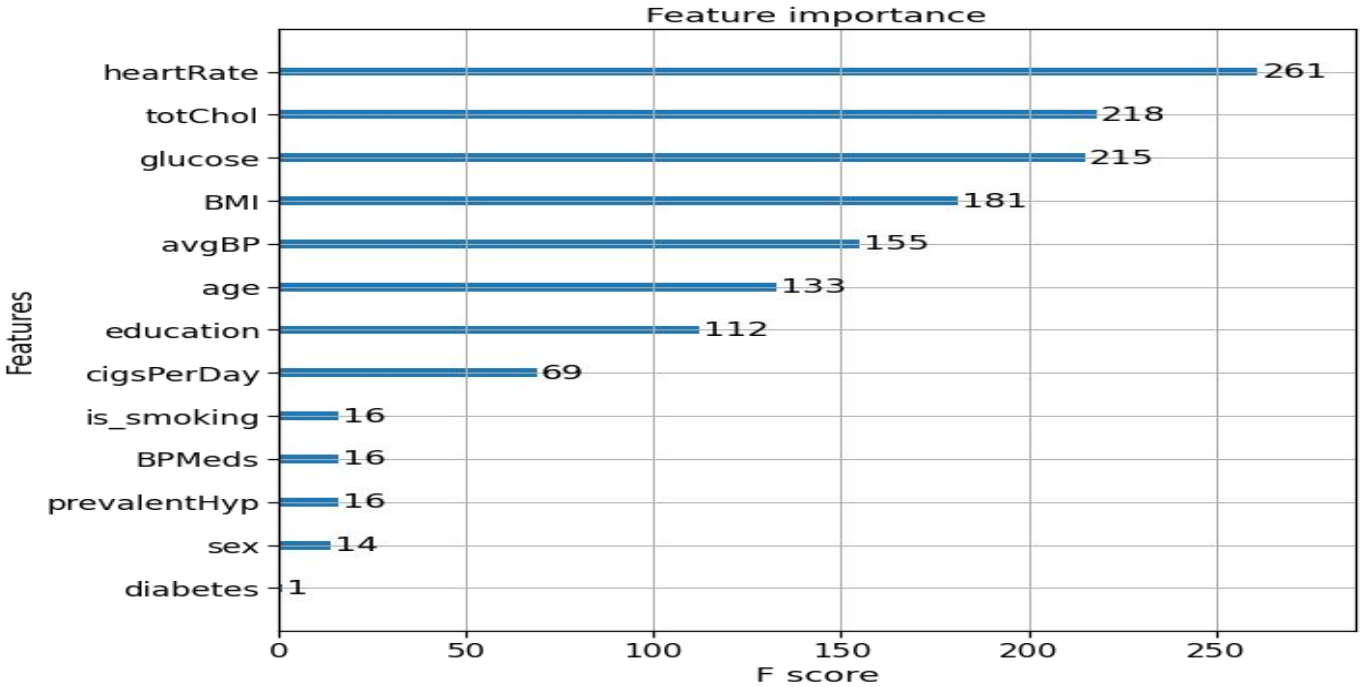


➔ EVALUATION

- The model that performed best on the given dataset is XGBoost with an accuracy and f1 score of 0.94 followed by Random Forest.
- KNN and SVM with hyperparameter tuning both have almost same level of accuracy and recall.
- Logistic Regression model performed the worst among all with 0.67 accuracy and f1 score.

	Model	Accuracy	Precision	Recall	F1 Score
7	XGBoost	0.943287	0.970516	0.914352	0.941597
6	Random Forest	0.882523	0.890201	0.872685	0.881356
4	KNN with hyperparameter tuning	0.820023	0.898148	0.776777	0.833065
2	SVM with hyperparameter tuning	0.802083	0.836806	0.782468	0.808725
0	Decision Tree	0.780671	0.821192	0.717593	0.765905
3	K Nearest Neighbour	0.755787	0.855324	0.713320	0.777895
1	Support Vector Machines	0.730324	0.775463	0.711253	0.741971
5	Logistic Regression	0.674190	0.673611	0.674392	0.674001

➔ FEATURE IMPORTANCE (XGBoost)



- heartRate is the most important feature in predicting the CHD.
- totChol and glucose have the somewhat same level of importance.
- BMI, avgBP and age are other important features.

➔ CONCLUSION

- Number of people belonging to middle age group are highest whereas number of people belonging to young age group are lowest
- Male and female both are equally prone to CHD
- Number of male smoker is higher than female smokers.
- Age is an important aspect in predicting the CHD. Middle and older age people are more prone to CHD than young people. Young people are least likely to get CHD.
- People who suffered previously from a heart stroke have severely high chances of getting CHD.
- People on BP medication, diabetic patients and hypertenisve patients have higher chance of getting a CHD than other people.
- XGBoost performed the best among all other models with highest accuracy and f1 score of 94%.
- heartRate is the most important feature in predicting the CHD followed by totChol and glucose.

THANK YOU!