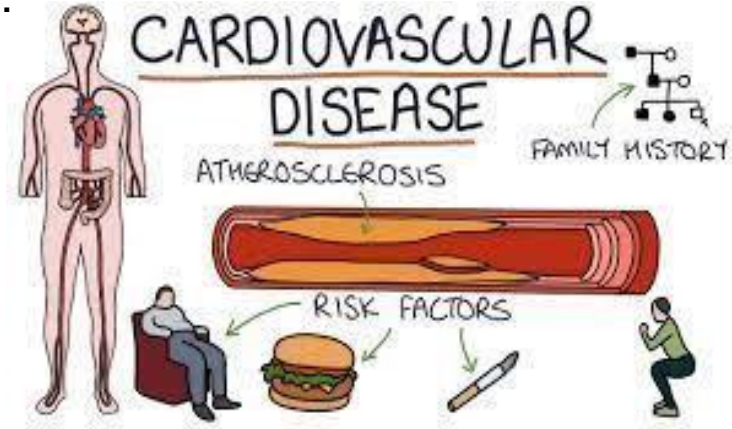
Capstone Project

Cardiovascular Risk Prediction

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.



Key Steps:

- Defining the problem statement
- Data Cleaning
- EDA and data visualization
- Data preprocessing
- Feature selection
- Preparing Dataset for model
- Applying model
- Model validation and selection

Why is predictive analytics useful for Cardiovascular risk ?

- To know which patients are in risk:
- To know which disease lead to Cardiovascular risk:
- To know what habit lead to Cardiovascular risk:
- To know what should be BMI, BP, Diabetes and Cholesterol level:

Dataset:

Rows : 3390

Columns : 17

	id	age	education	sex	is_smoking	cigsPerDay	BPMeds	prevalentStroke	prevalentHyp	diabetes	totChol	sysBP	diaBP	BMI	heartRate	glucose	TenYearCHD
0	0	64	2.0	F	YES	3.0	0.0	0	0	0	221.0	148.0	85.0	NaN	90.0	80.0	1
1	1	36	4.0	M	NO	0.0	0.0	0	1	0	212.0	168.0	98.0	29.77	72.0	75.0	0
2	2	46	1.0	F	YES	10.0	0.0	0	0	0	250.0	116.0	71.0	20.35	88.0	94.0	0
3	3	50	1.0	M	YES	20.0	0.0	0	1	0	233.0	158.0	88.0	28.26	68.0	94.0	1
4	4	64	1.0	F	YES	30.0	0.0	0	0	0	241.0	136.5	85.0	26.42	70.0	77.0	0

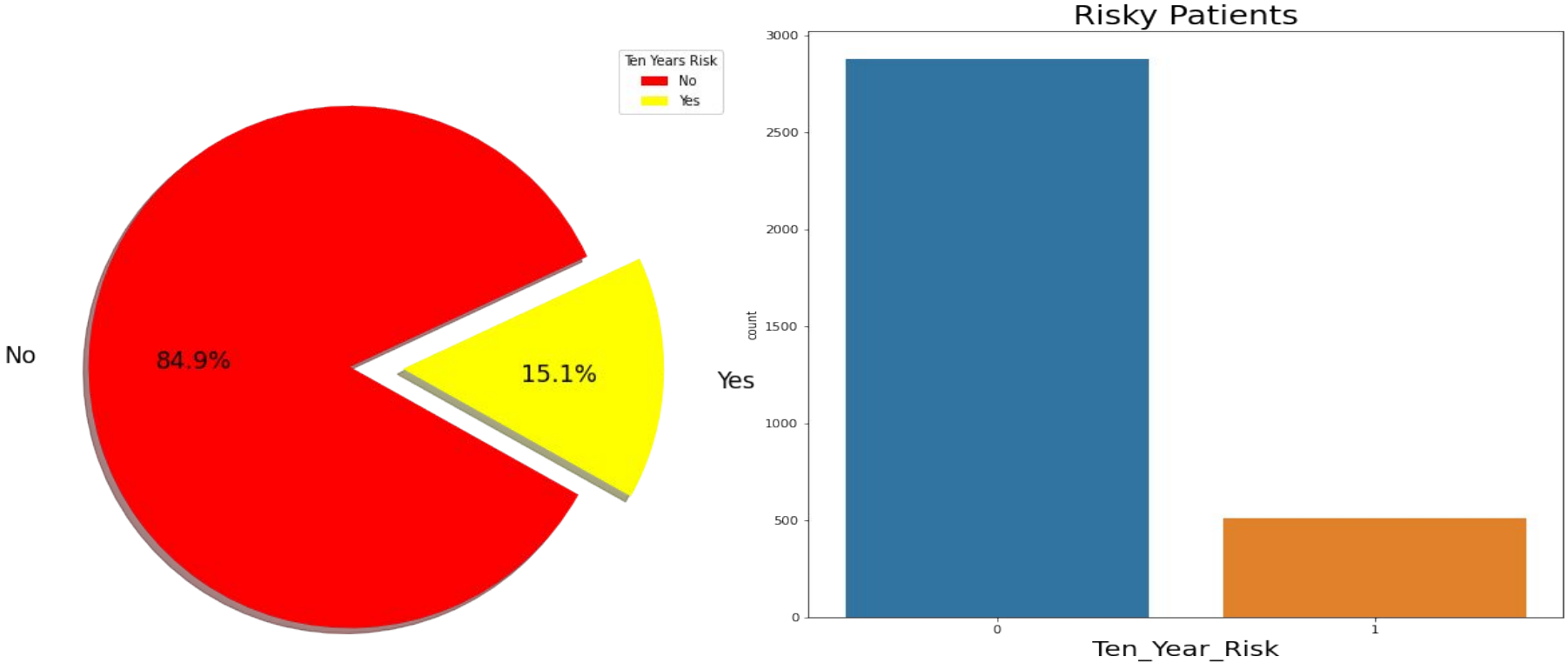
Variable Names:

- Sex:
- Age:
- is_smoking:
- Cigs Per Day:
- BP Meds:
- Prevalent Stroke:
- Prevalent Hyp:
- Diabetes:
- Tot Chol:
- Sys BP:
- Dia BP:
- BMI:
- Heart Rate:
- Glucose:
- 10-year risk

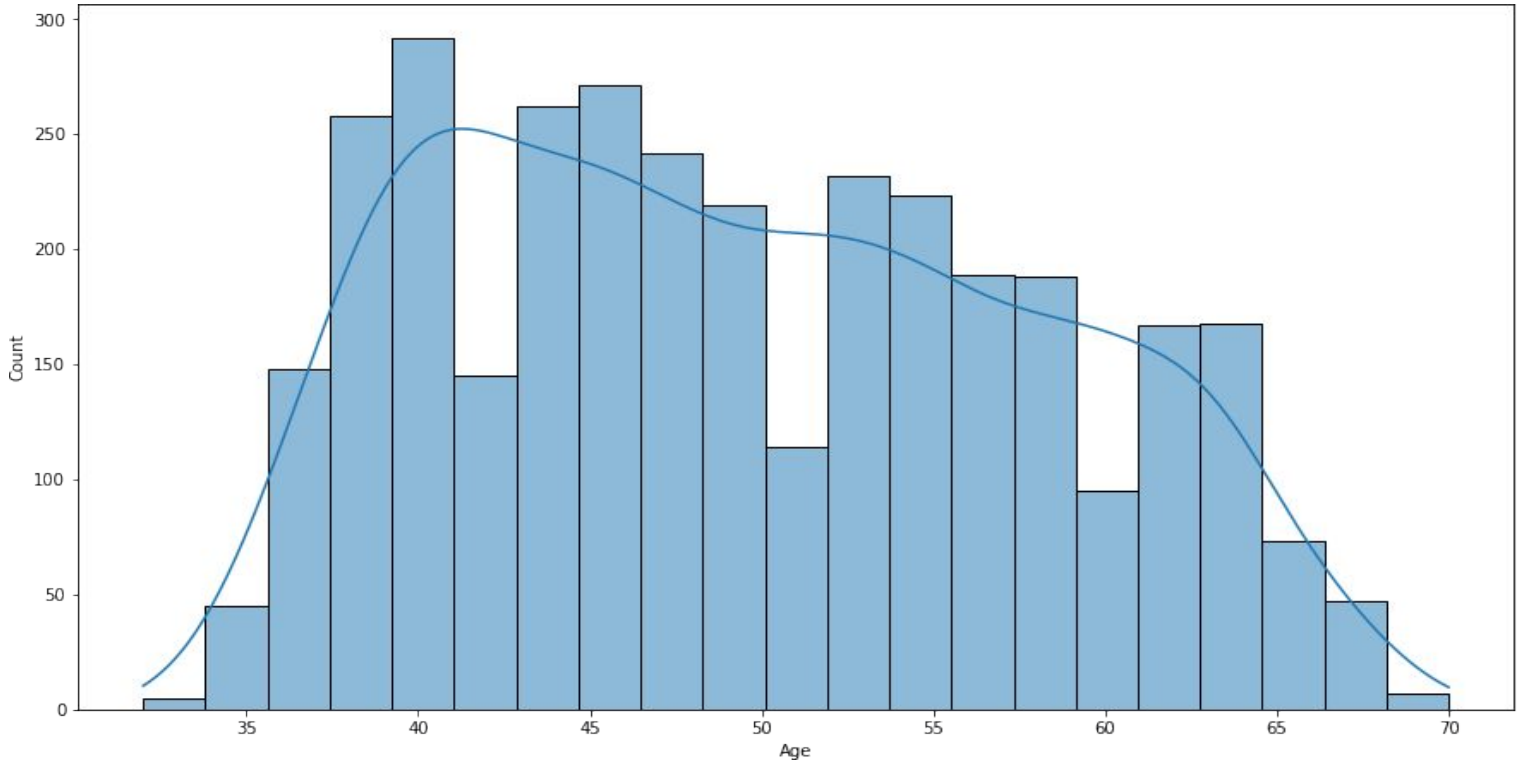
Exploratory Data Analysis:

EDA is used for analyzing what the data can tell us before the modeling or by applying any set of instructions/code.

Risky Patients:

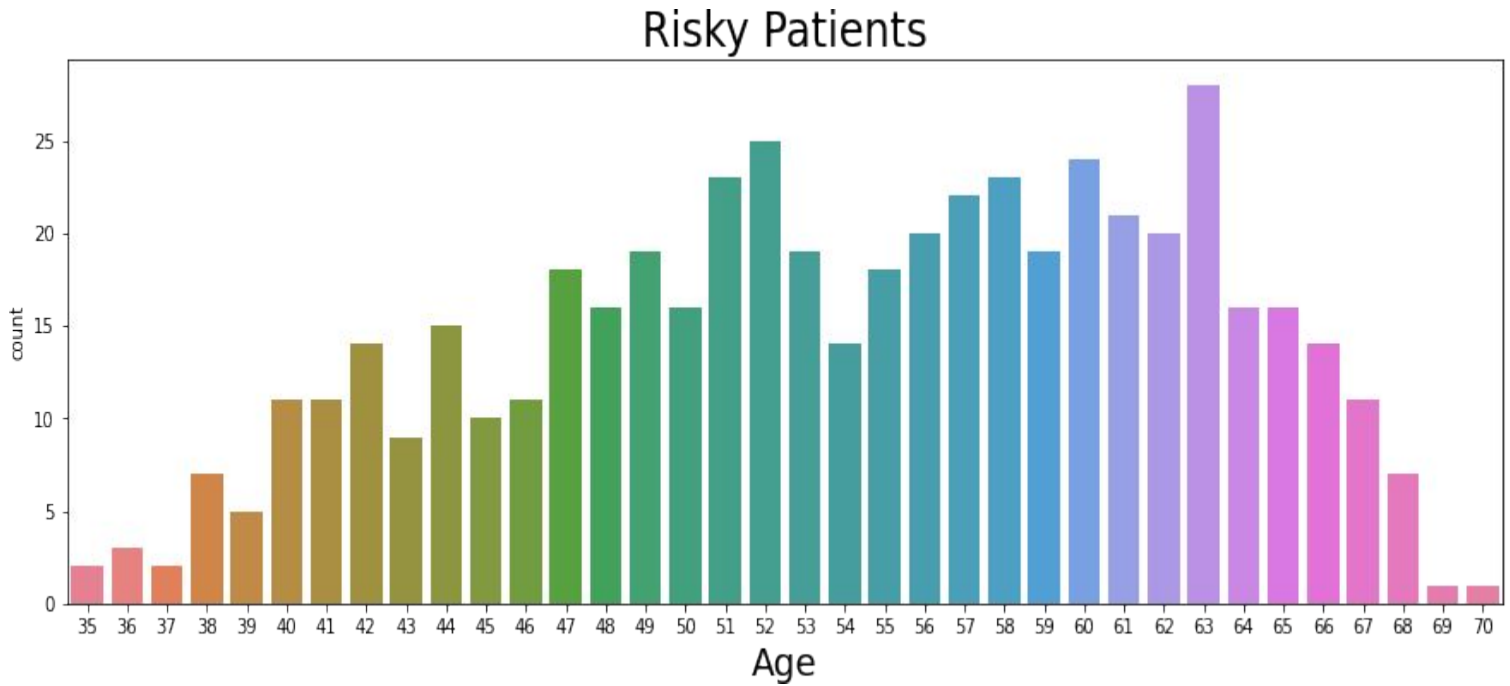


Distribution of age:



- Here we can see our data set contain the patients age between 32 to 70

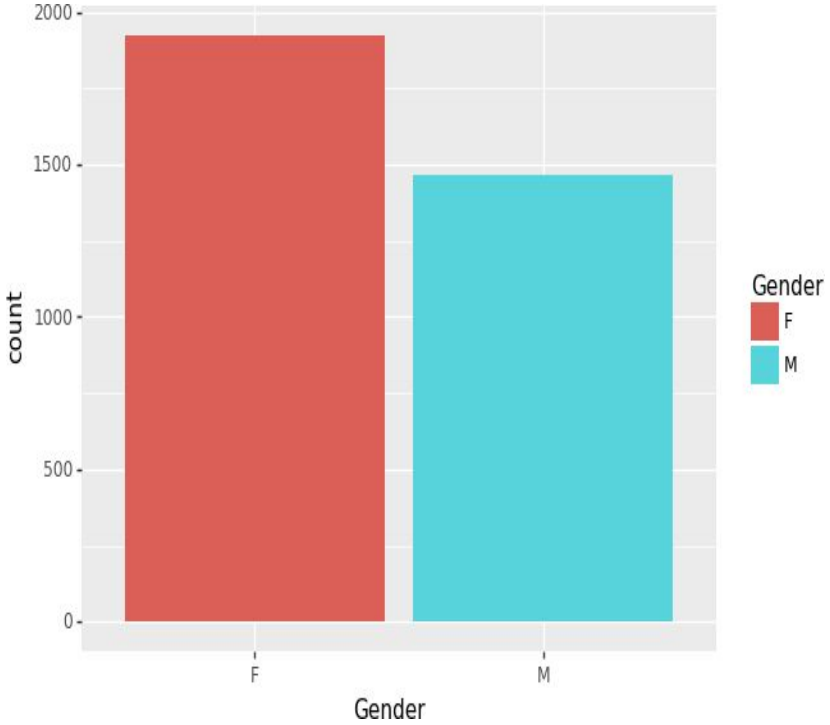
Risky Patients With Respect To Age:

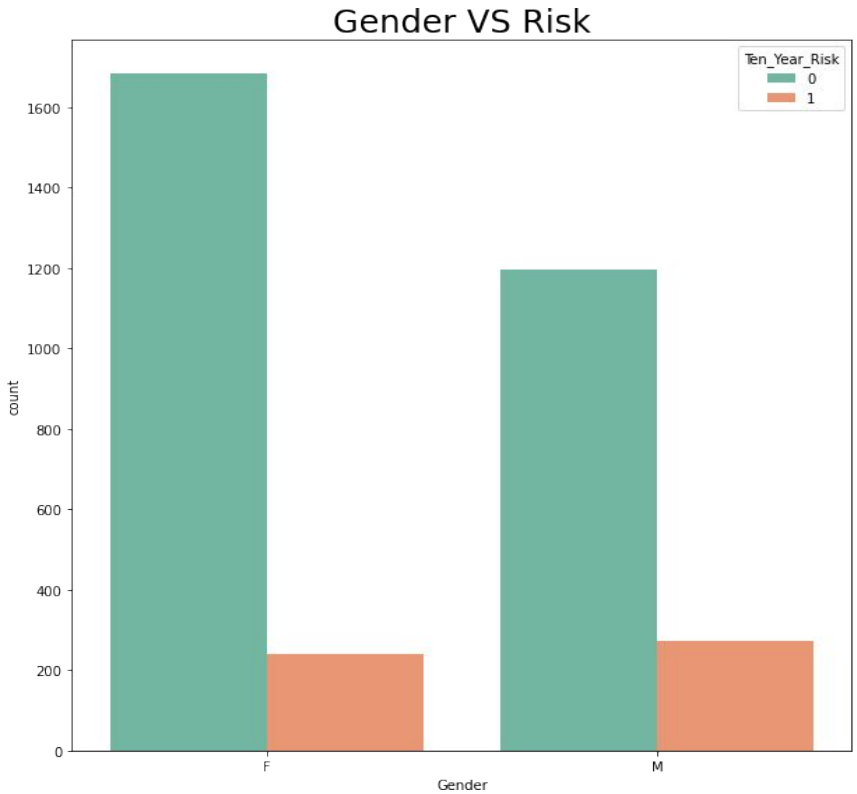


OBSERVATION:

Here we can see there is more risk of cardiovascular disease in patients of age between 51 to 63.

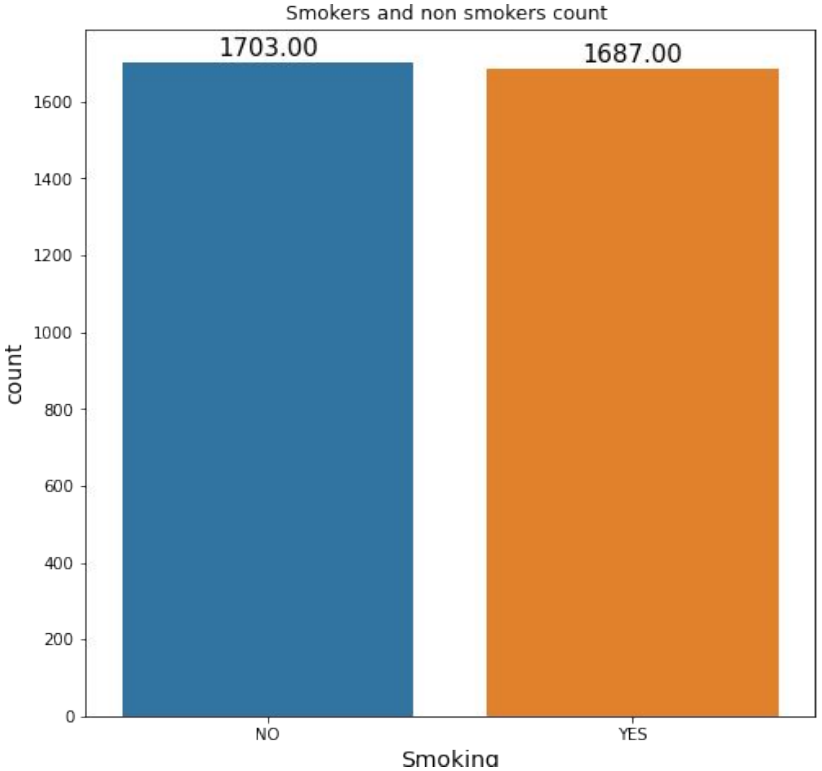
Risk with respect to gender:

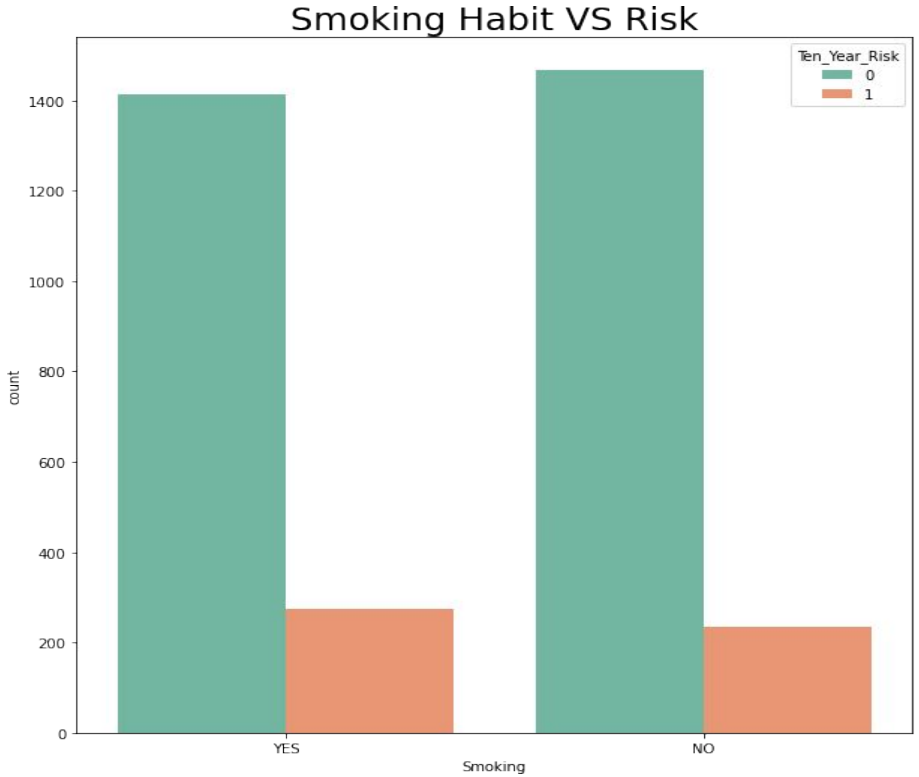




we can see the count of male and female are same in risk which is around 200 , Although females are more than males in our dataset.

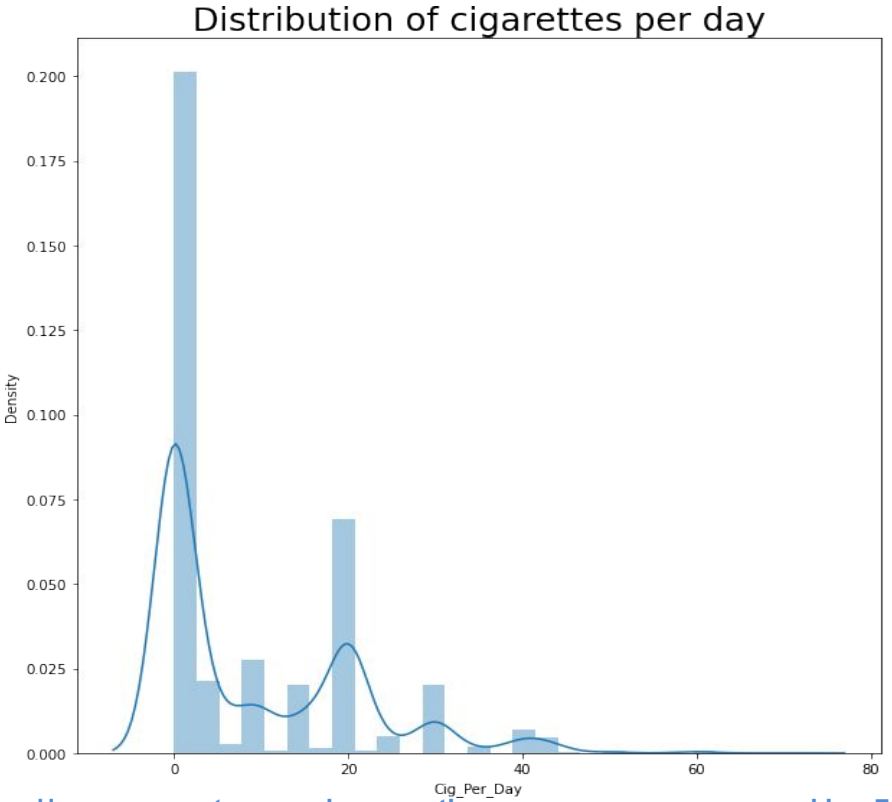
Risk With Respect To Smoking Habit:

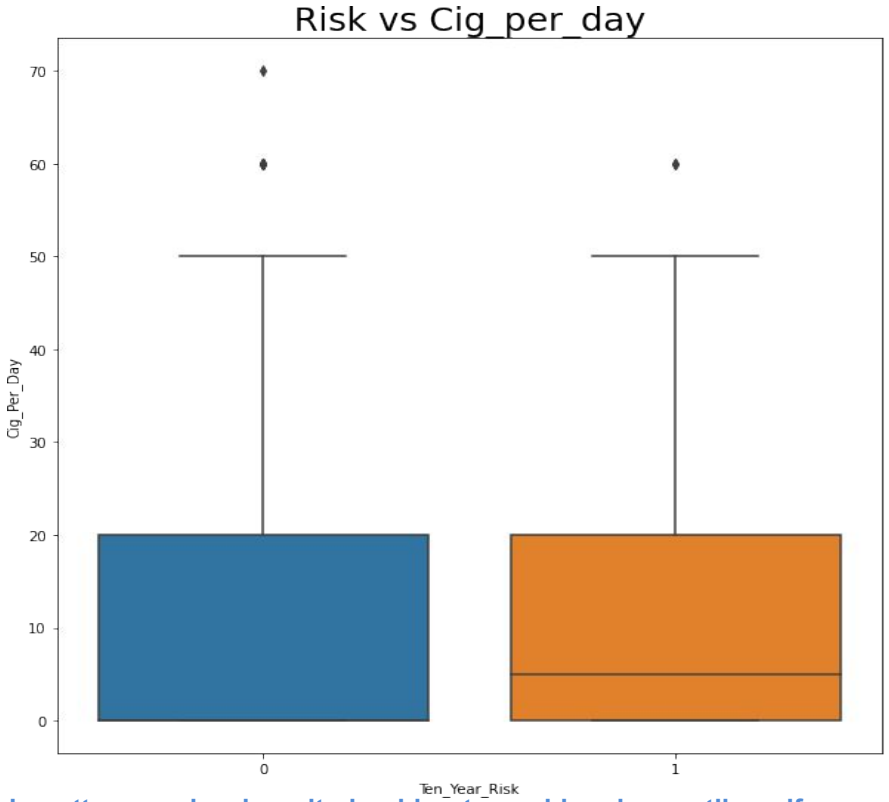




Here we can see around 250 smokers are in risk and around 210 non-smokers are is risk for cardiovascular disease, after seeing this trend we cannot directly say thay only smokers are at risk.

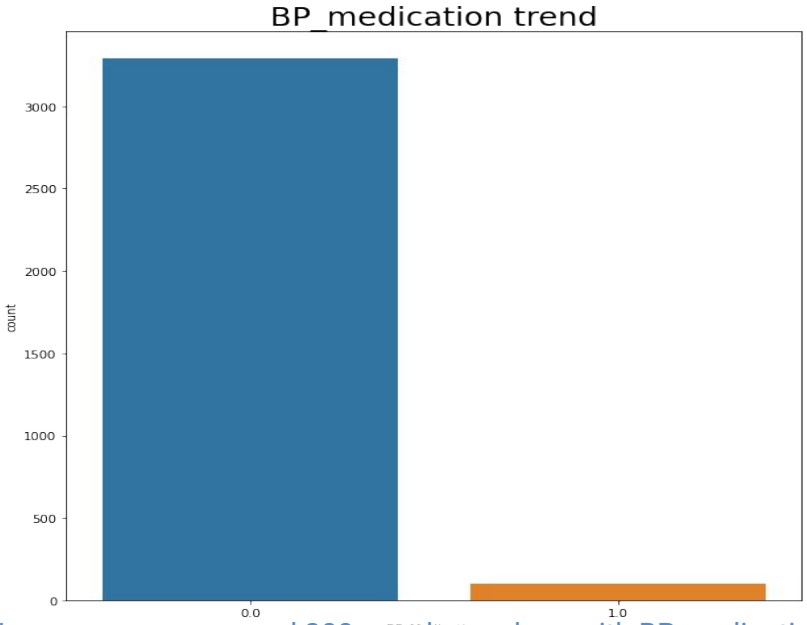
Risk With Respect To Cigarettes Per Day:

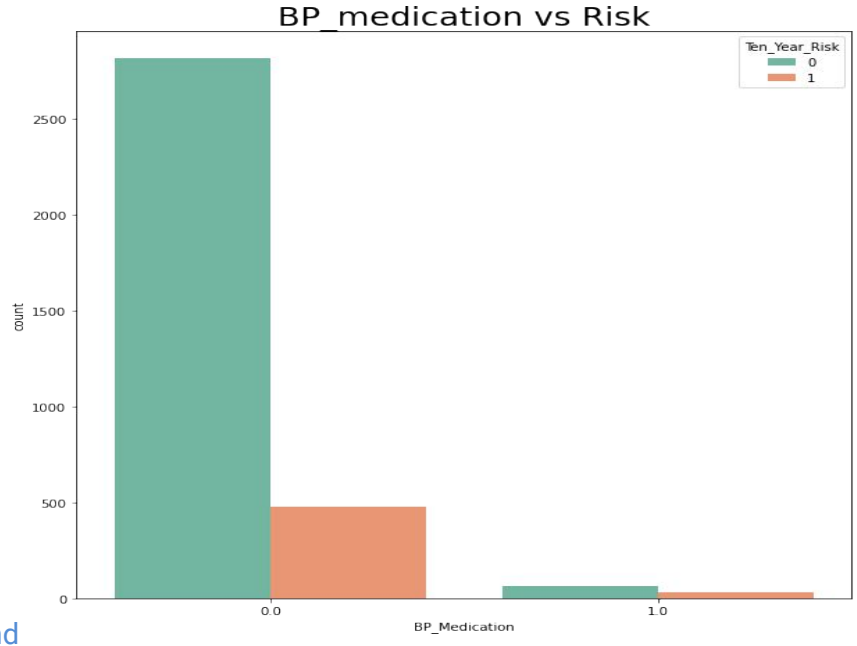




Here some extreme values are there some persons are smoking 70 cigarettes per day, here it should not considered as outliers, if done it may reduces the power of the test. Here i proceed forward with these higher values. Since its considered as continous points, but the distribution its look alike discrete distribution.

Risk With Respect To BP Medication:

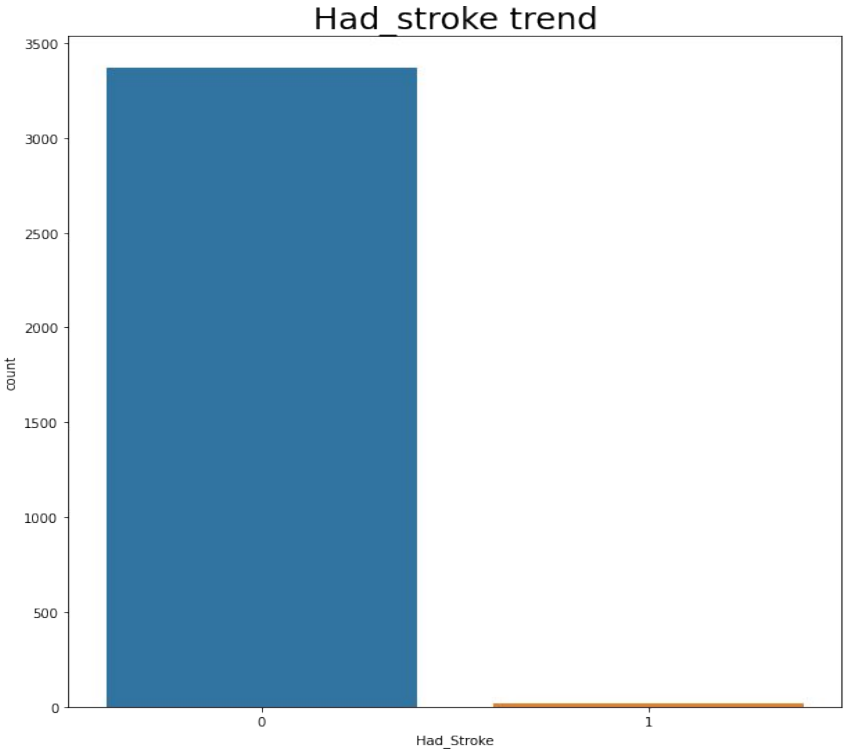




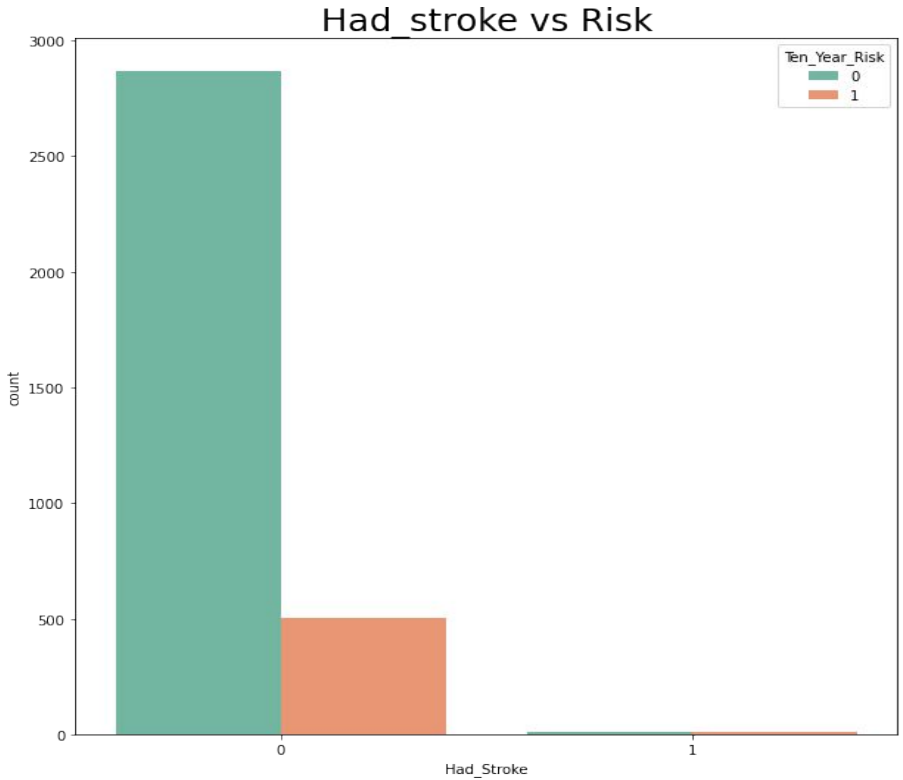
Here we can see around 200 people are done with BP medication and around 3200 people have not done any medication

Here we can see around 200 people are done with BP medication and around 3200 people have not done any medication. Here we can see there are very few people who are done with BP medication which are around 200 but many people have not taken any BP medication and they are around 3200. after seeing this trend we cannot say that after taking medication there is no risk.

Risk With Respect To Stroke:

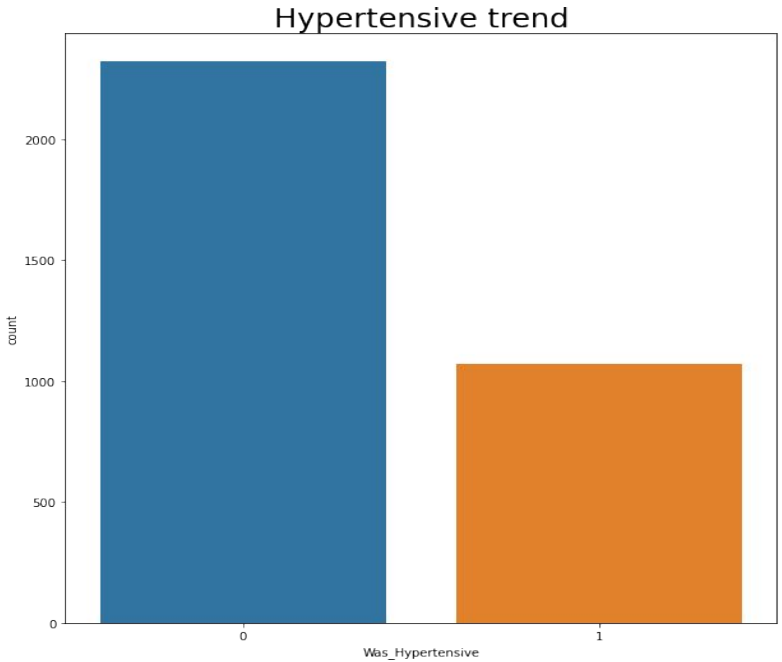


Here we can see very few people had stroke and around 3350 people yet not had stroke and are healthy enough.

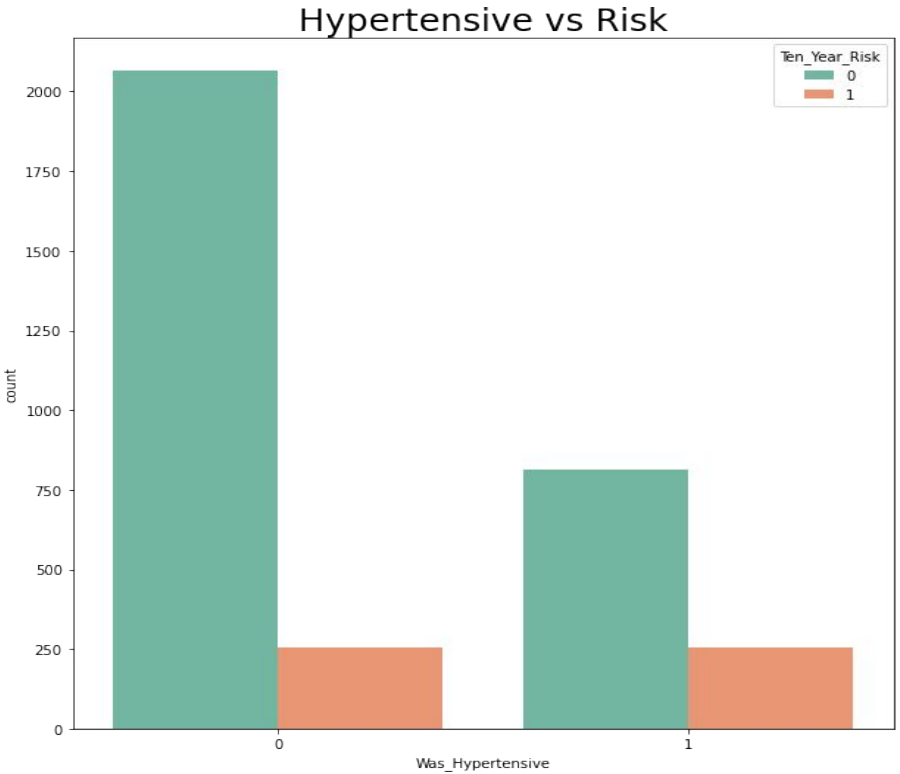


There are very few people that had stroke.
Here around 500 patients who did not had stroke yet and are a risk and around 2800 patients are safe.

Risk With Respect To Hypertensive:

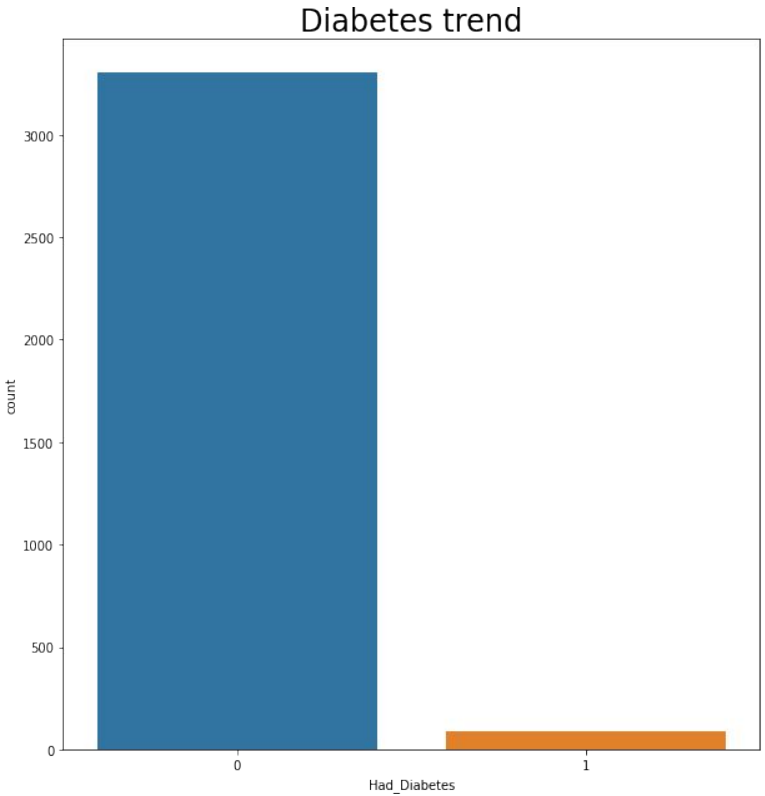


Here we can see 1000 people facing hypertensive and around 2400 people are not facing any hypertensive.

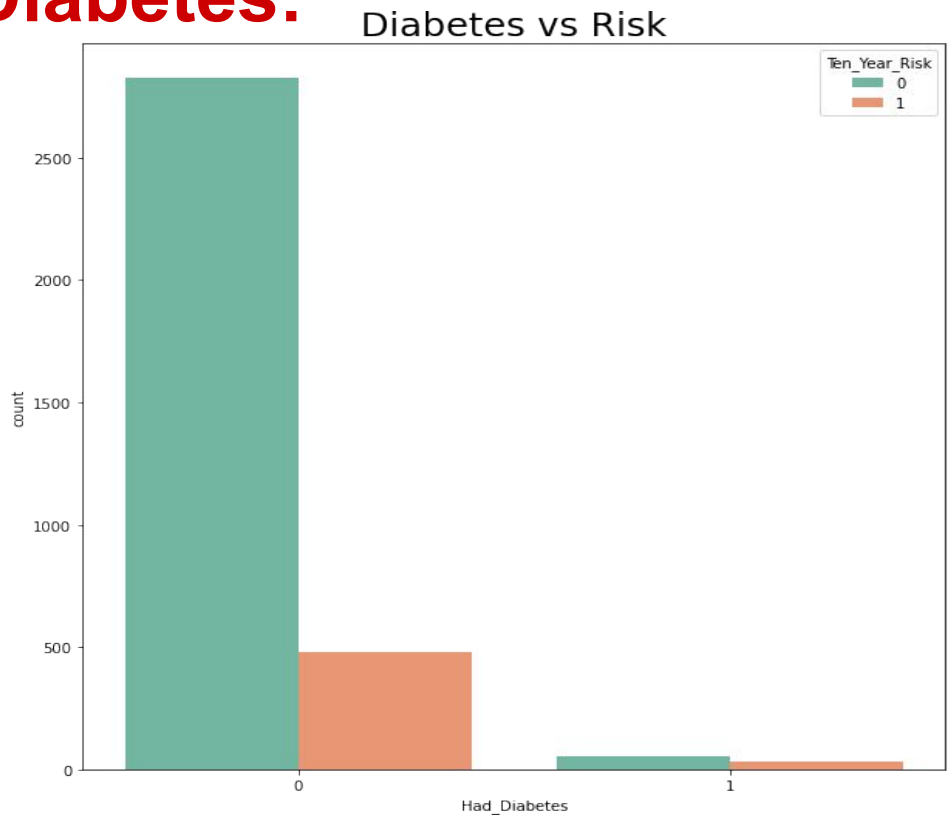


Here around 250 people with hypertensive are in risk and around 255 people with no hypertensive are at risk.

Risk With Respect To Diabetes:

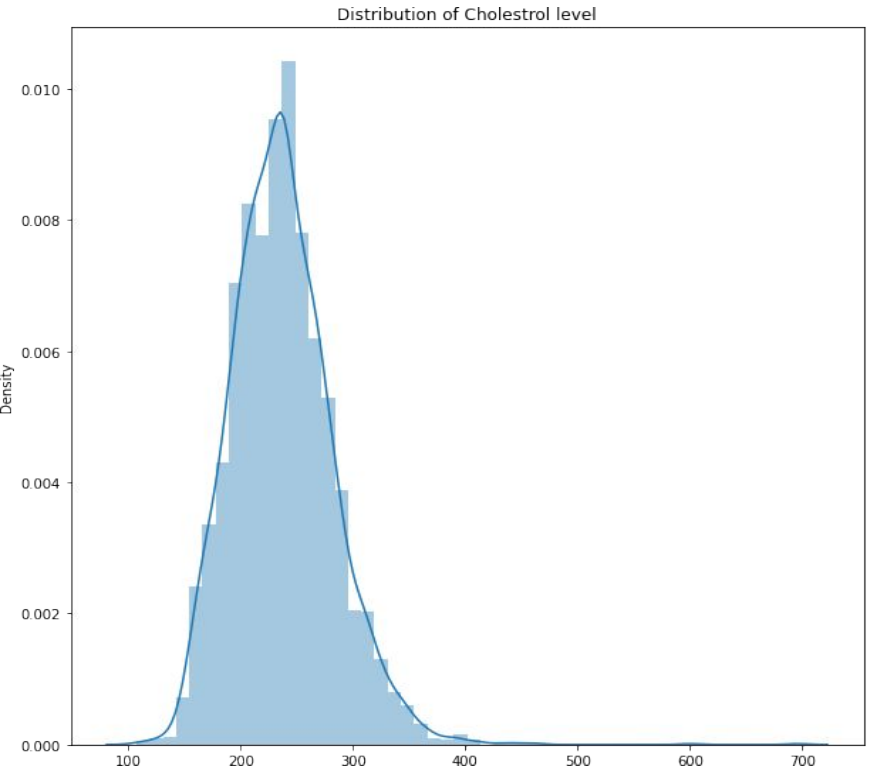


Here we can see around 100 people had diabetes and around 3300 people yet not had diabetes.

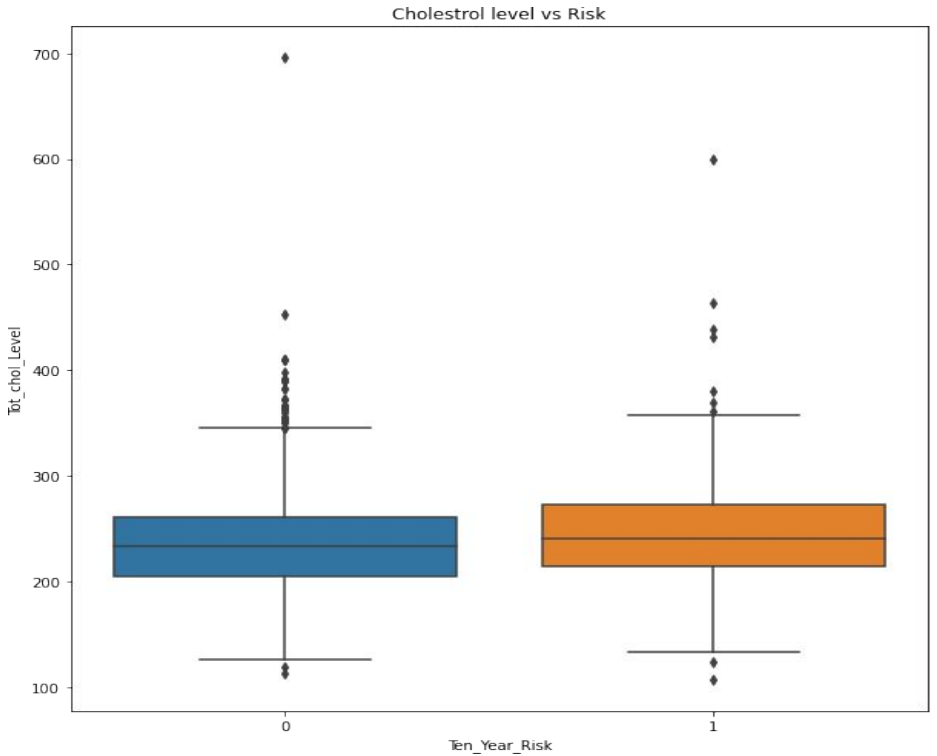


Here we can see people who did not had diabetes are more and around 500 people who did not had diabetes are at risk.
And there are very few people who had diabetes are at risk.
So diabetes feature is not helping that much in ten years risk.

Risk With Respect To Cholesterol level:

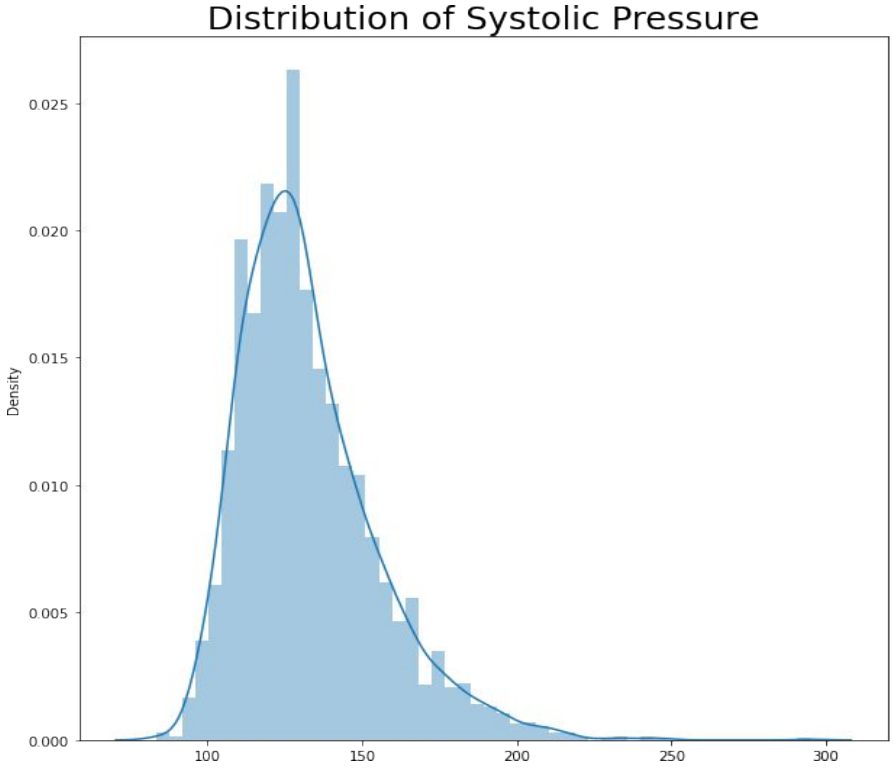


Here we can see most people cholestrol level lies between 200 to 280.

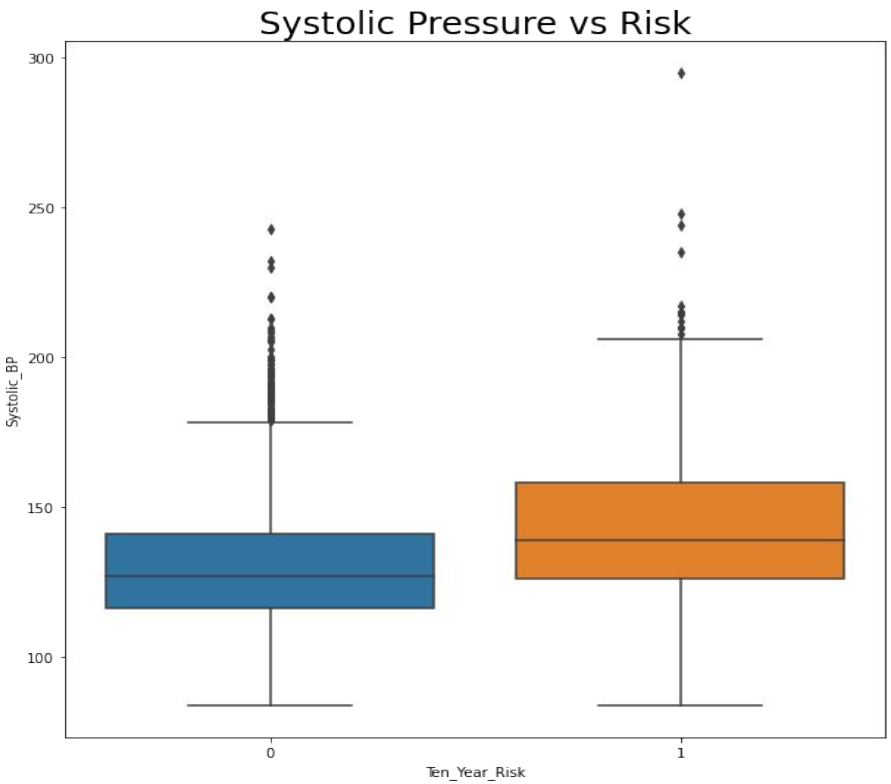


we can see most of the people who are not in risk their Cholestrol level lies between 210 to 280 and people who are in risk their cholestrol level lies between 215 to 285 there in not huge difference it is quite normal

Risk With Respect To Systolic Pressure:

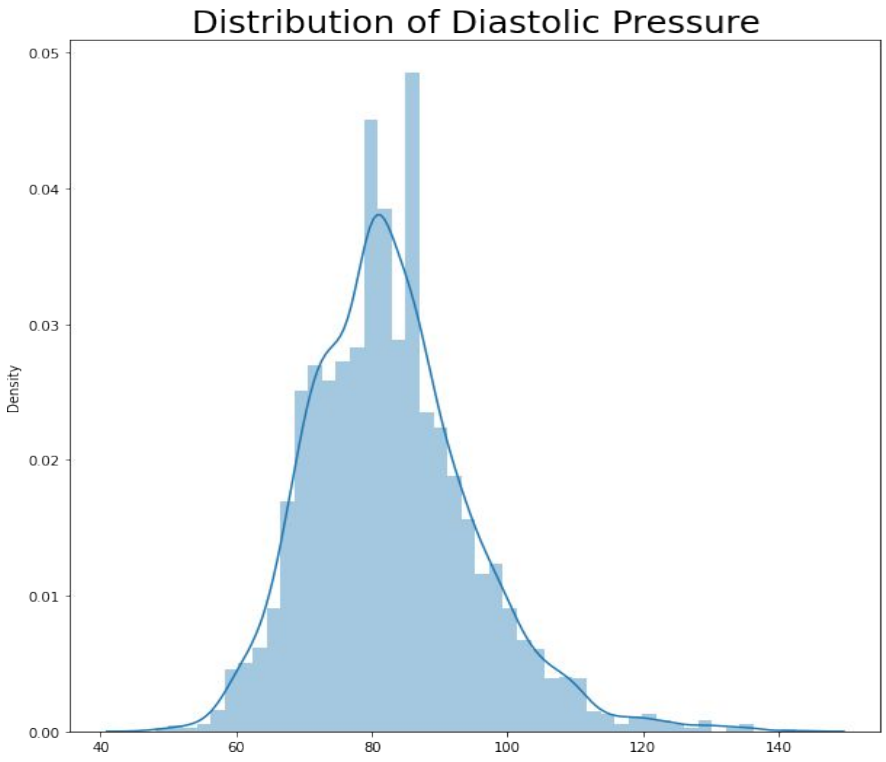


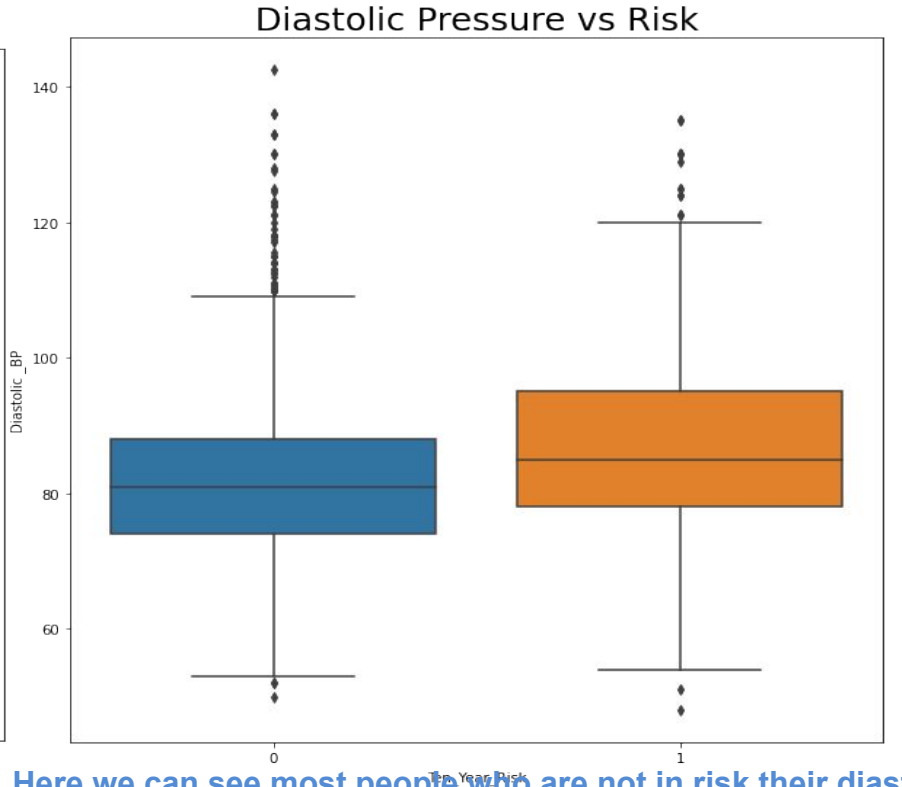
Here we can see most people systolic pressure lies between 110 to 140.



Here we can see most people who are not in risk their systolic BP lies between 110 to 140 and people who are at risk their systolic BP lies between 125 to 160 here we can say people with high systolic BP are at risk.

Risk With Respect To Diastolic Pressure:

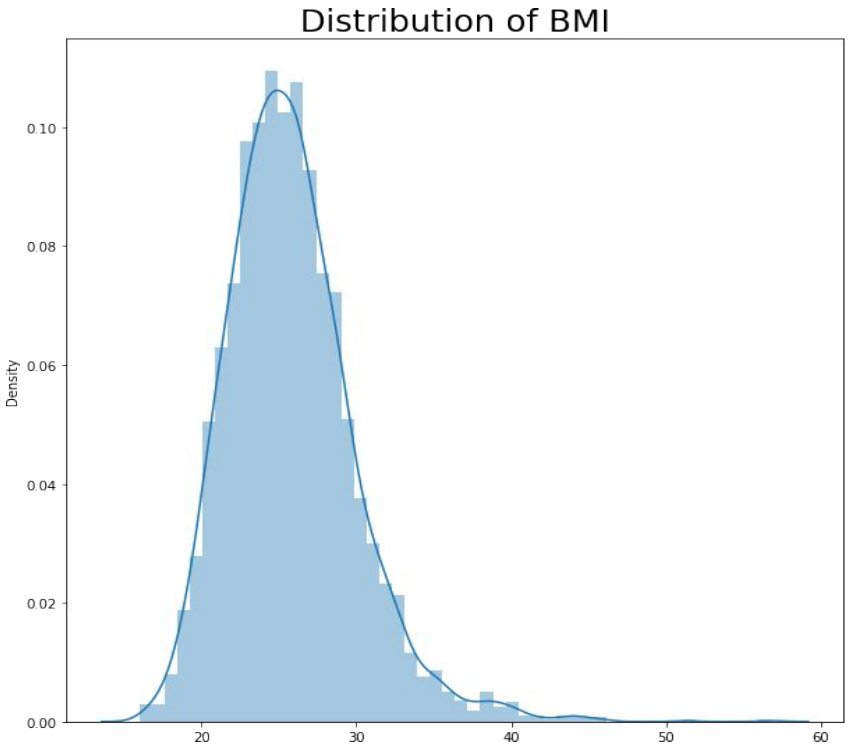




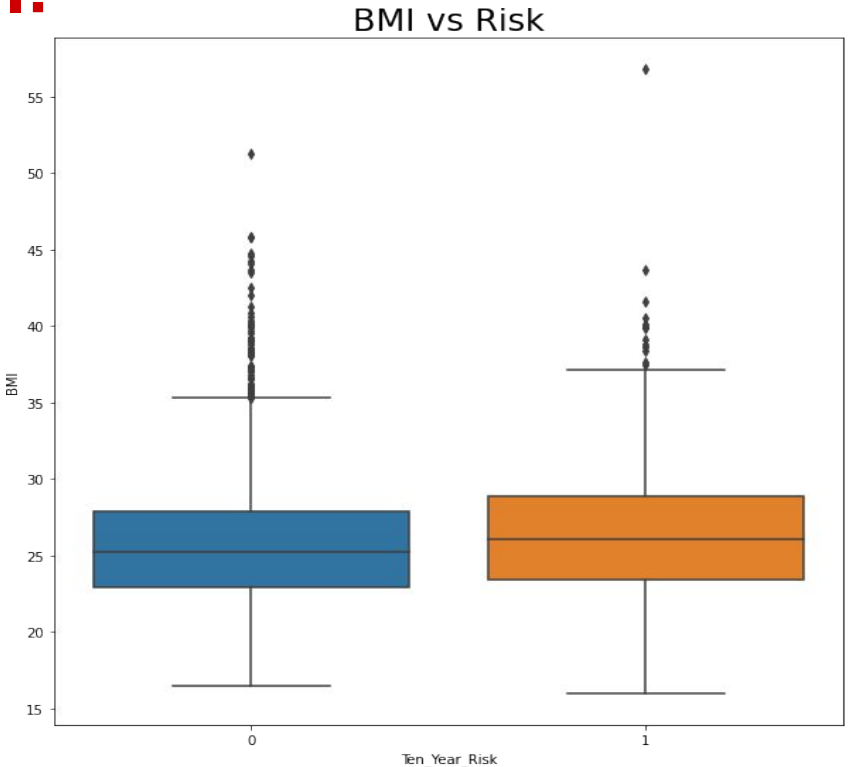
we can see most people diastolic BP lies between 75 to 85 approx.

Here we can see most people who are not in risk their diastolic BP lies between 75 to 85 and people who are at risk their diastolic BP lies between 89 to 90. here we can say there is a slight increase in diastolic BP of people who are in risk

Risk With Respect To BMI:

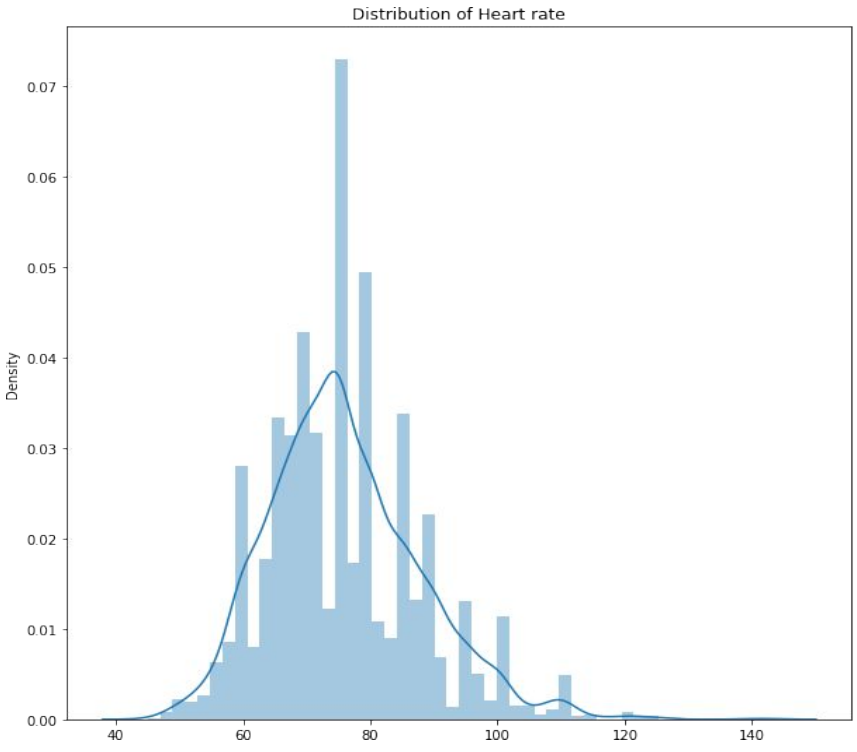


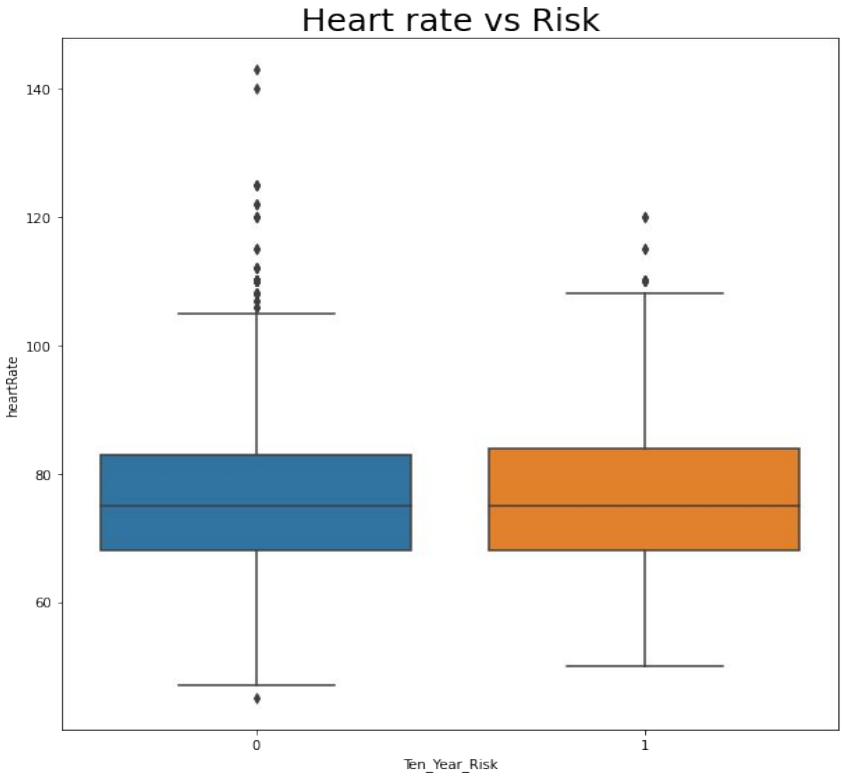
we can see most people BMI lies between 22 to 28.



Here we can see most people who are not in risk their BMI lies between 22to 28 and people who are at risk their BMI lies between 23 to 29 approx. WE cannot see any difference BMI is approx. same of risky and not risky people.

Risk With Respect To Heart Rate:

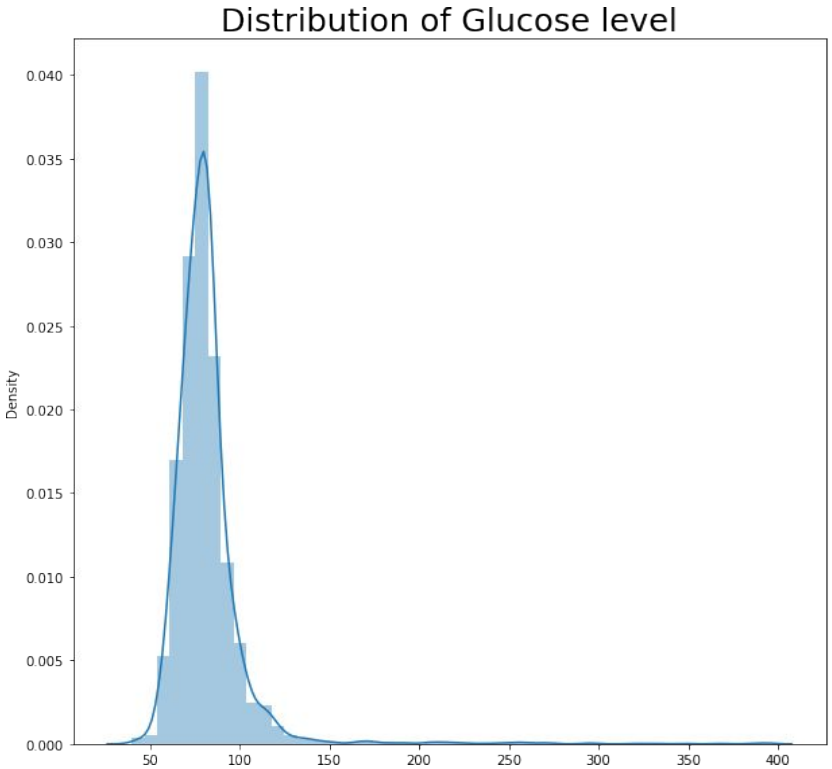


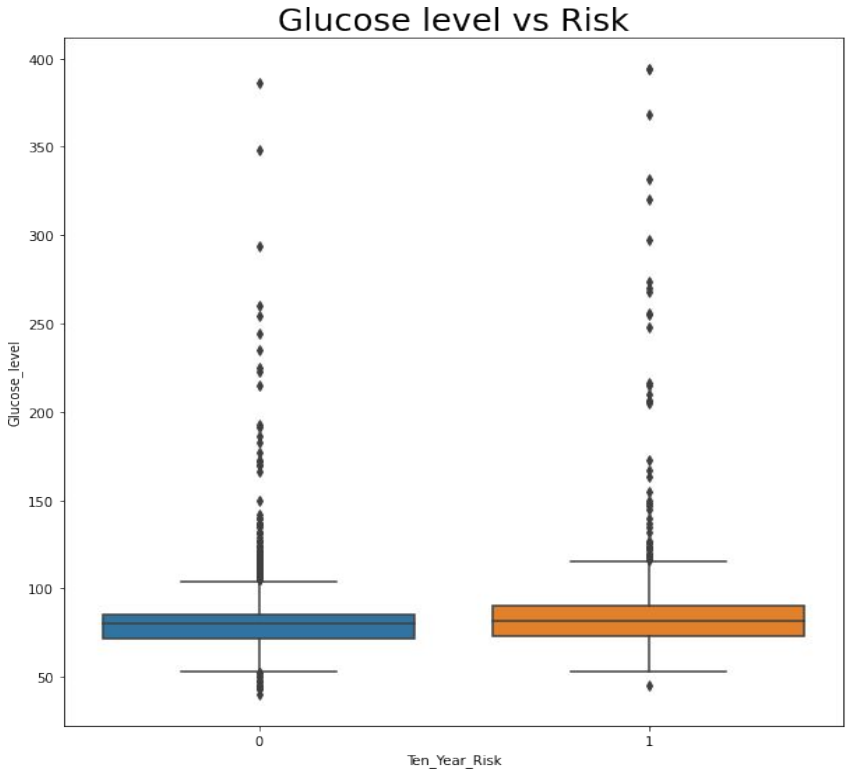


Here we can see most people Heart rate lies between 65 to 80.

Here we can see most people who are not in risk their heart rate lies between 68 to 83 and people who are at risk their heart rate lies between 68 to 84. which is same for risky and not risky people

Risk With Respect To Glucose Level:

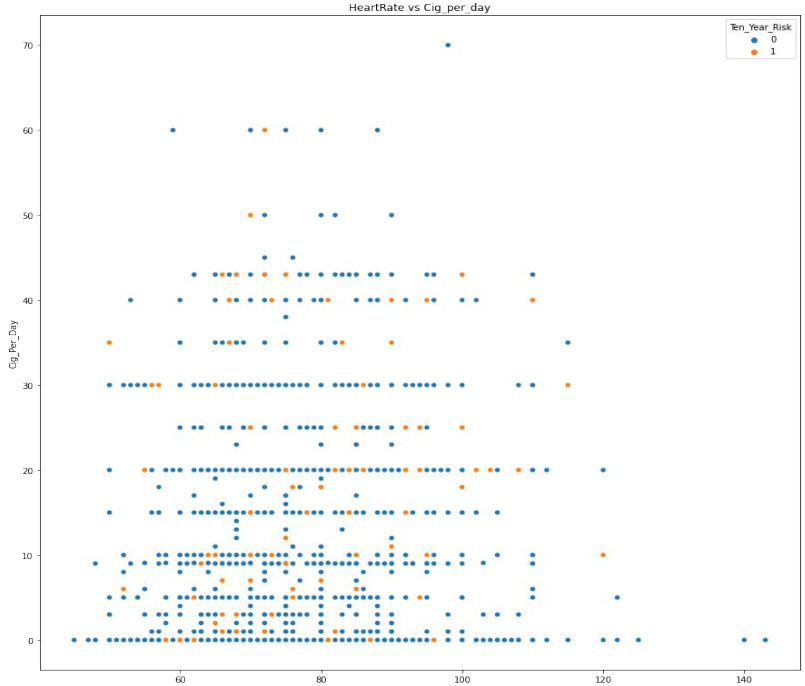




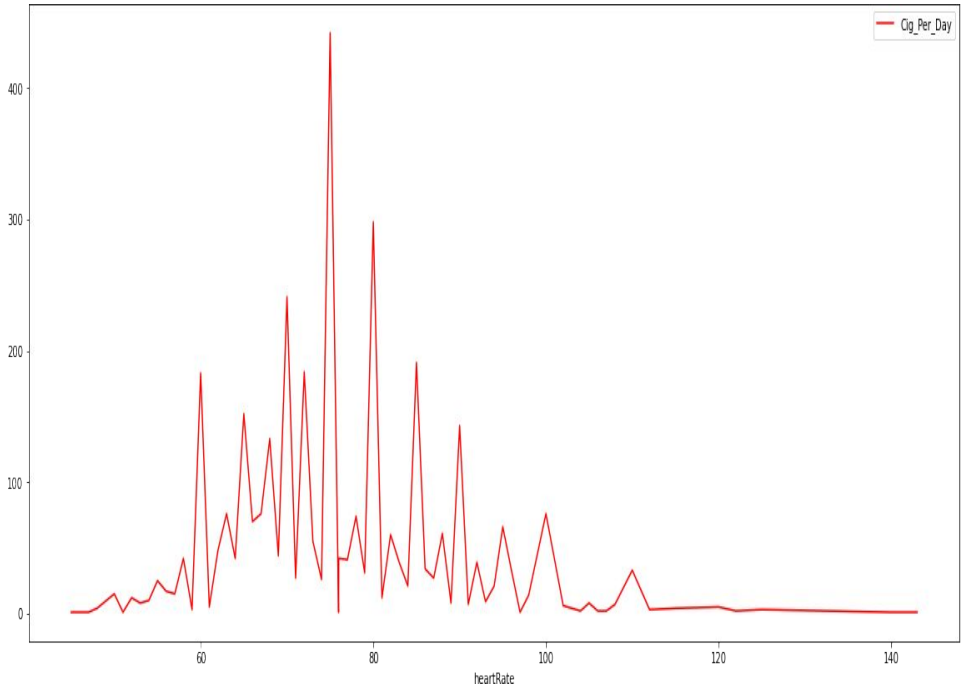
we can see most people glucose level lies between 60 to 80.

Here we can see there is not that difference between the glucose level of risky and non risky patients. glucose level lies between 70 to 80 for both risky and non risky pattients.

Heart Rate With Respect To Cigarettes Per Day:

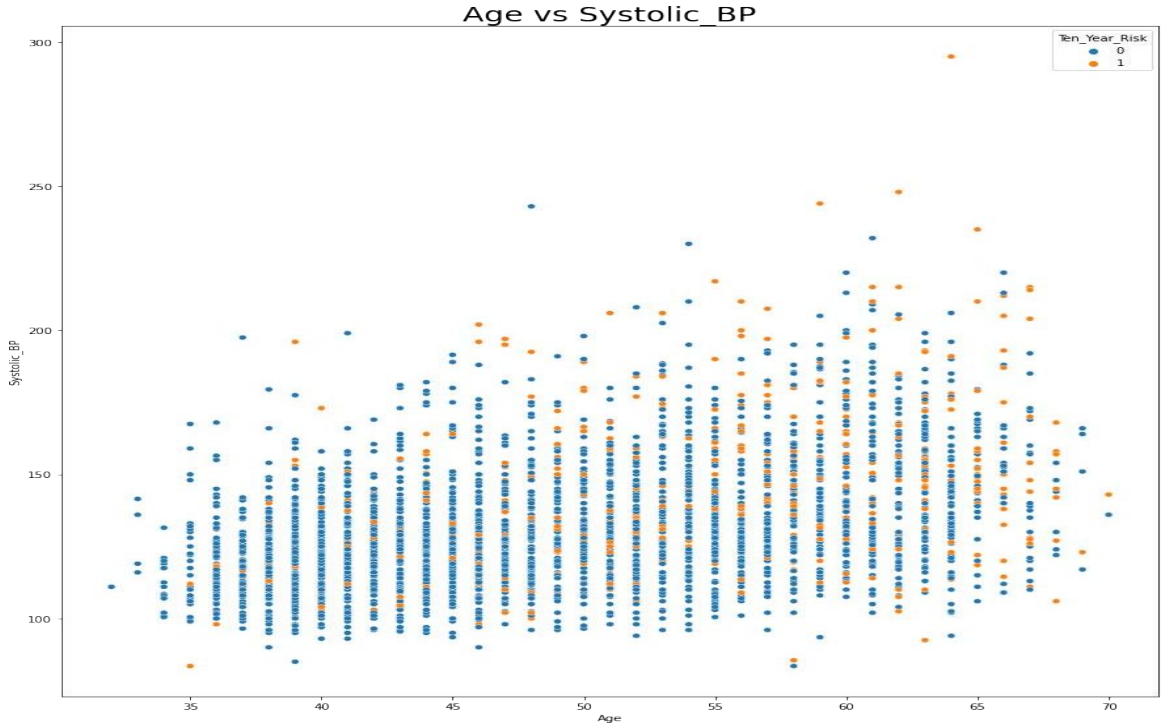


Though Observations are scatter around the 2 dimension space, where it is diificult to find the proper decision boundary that separates the two classes. Here we can see most people smoke cigarettes between 1 to 10 approx. and there heart rate lies between 60 to 100.



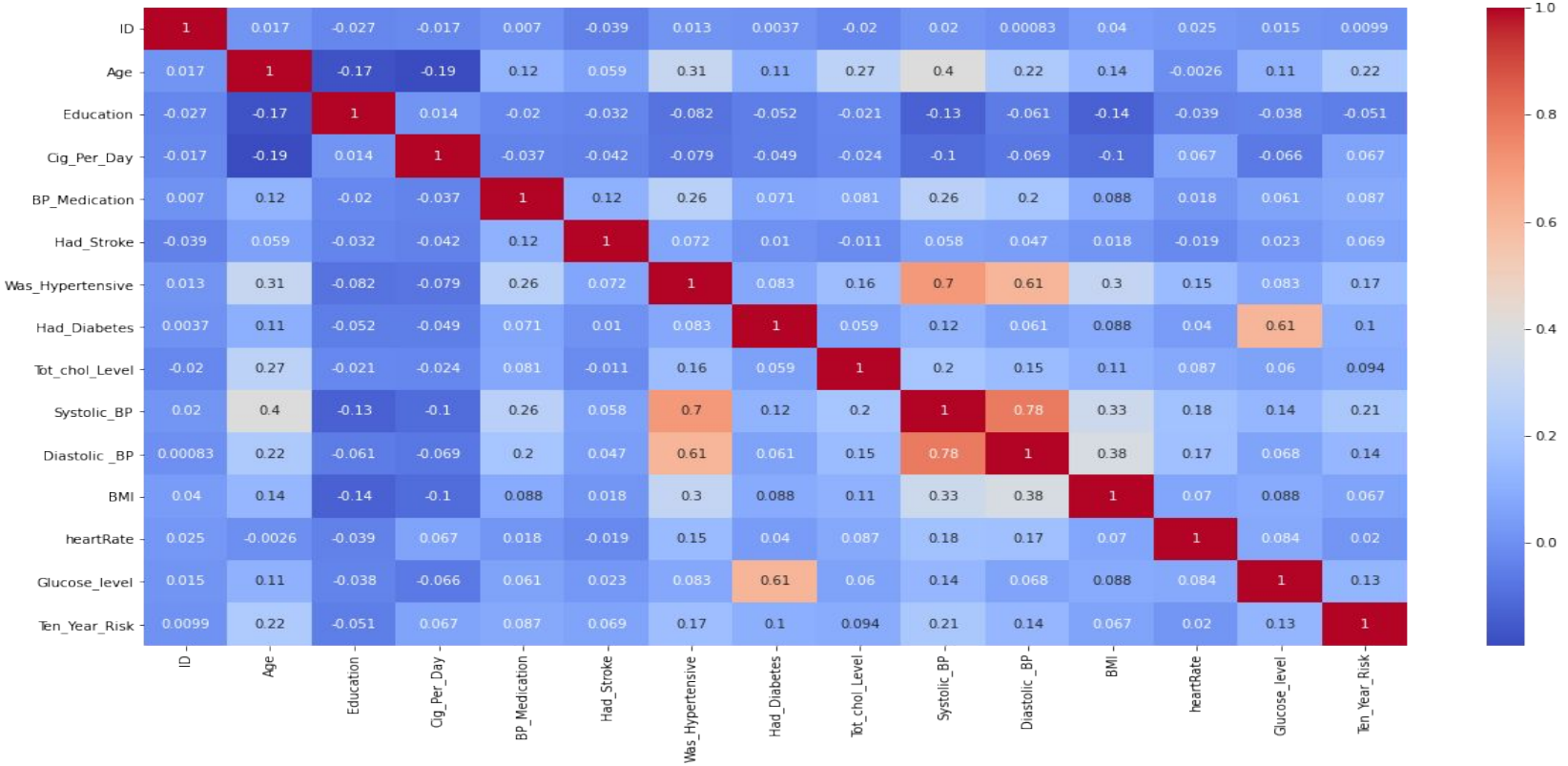
Here we can see most people smoke cigarettes between 1 to 10 approx. and there heart rate lies between 60 to 100.

Systolic BP With Respect To Age:



Here we can see age lies between 32 to 70 and most of the people's systolic BP lies between 90 to 200.

Correlation Matrix:



Here we can see systolic BP and was hypertensive variable are correlated, diastolic BP and was hypertensive variable are correlated, Glucose level and had diabetes variable are correlated diastolic BP and systolic BP are highly Correlated

Data Preprocessing:

Data After Converting to Numerical :

Rows: 3390

Columns: 18

Which columns we have converted ?

- Smoking
- Gender

Models Used:

1. Logistic Regression
2. Support Vector Machine (SVM)
3. SVM with Linear Kernel
4. SVM with Polynomial Kernel
5. SVM with Gaussian Radial Base Kernel
6. Neural Networks
7. Random Forest Classifier
8. XGBoost

Model Validation and Selection:

Observation 1:

At first I've tried Logistic Regression but scores were not that satisfying.

Observation 2:

Then I've tried using Support Vector Machine, i got good scores as compared to logistic regression. Then I used SVM with Kernel Tricks I have used Linear kernel, Polynomial Kernel and Gaussian Radial Base Kernel I got good scores with polynomial kernel and radial base kernel.

Observation 3:

Then I've used Neural networks with one hidden layer but we did not got good scores.

Observation 4:

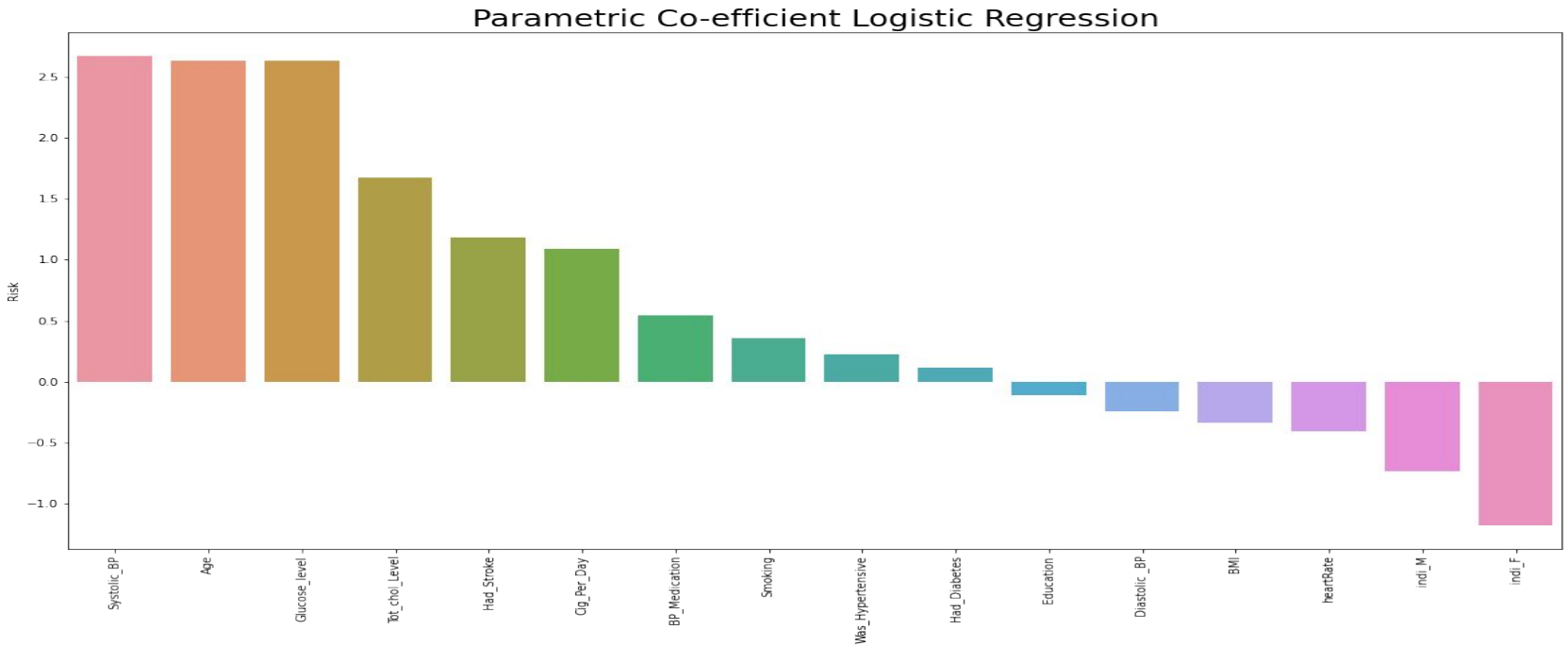
And lastly I've used Ensemble Learning models like Random Forest Classifier and XGBoost Classifier and both algorithms performed really well as compared to other models and with XGBoost I got the best Scores, so my optimal model is XGBoost Classifier.

Model Validation and Selection:

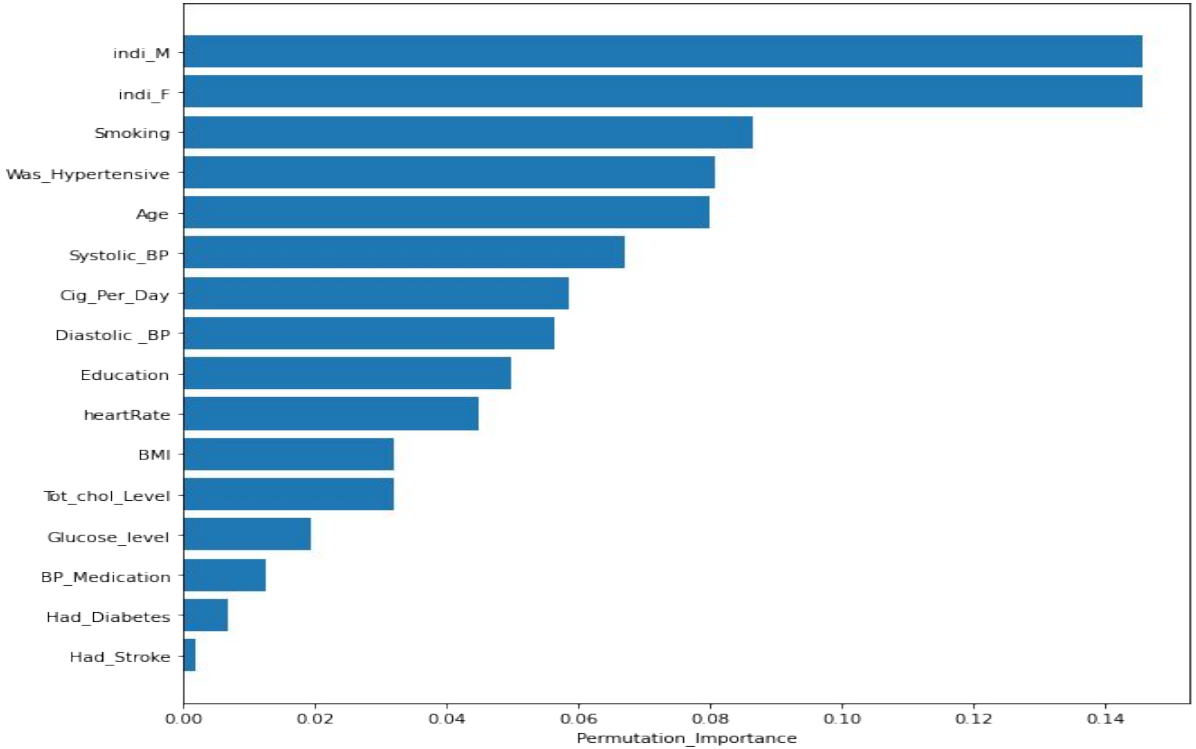
As XGBoost had performed really well so the best hyperparameters we got are:

```
'colsample_bytree': 0.7
'learning_rate': 0.03
'max_depth': 7
'min_child_weight': 4
'n_estimators': 500
'nthread': 4
'objective':
'reg:linear' 'silent': 1
'subsample': 0.7
```

Feature Importance for Logistic Regression:

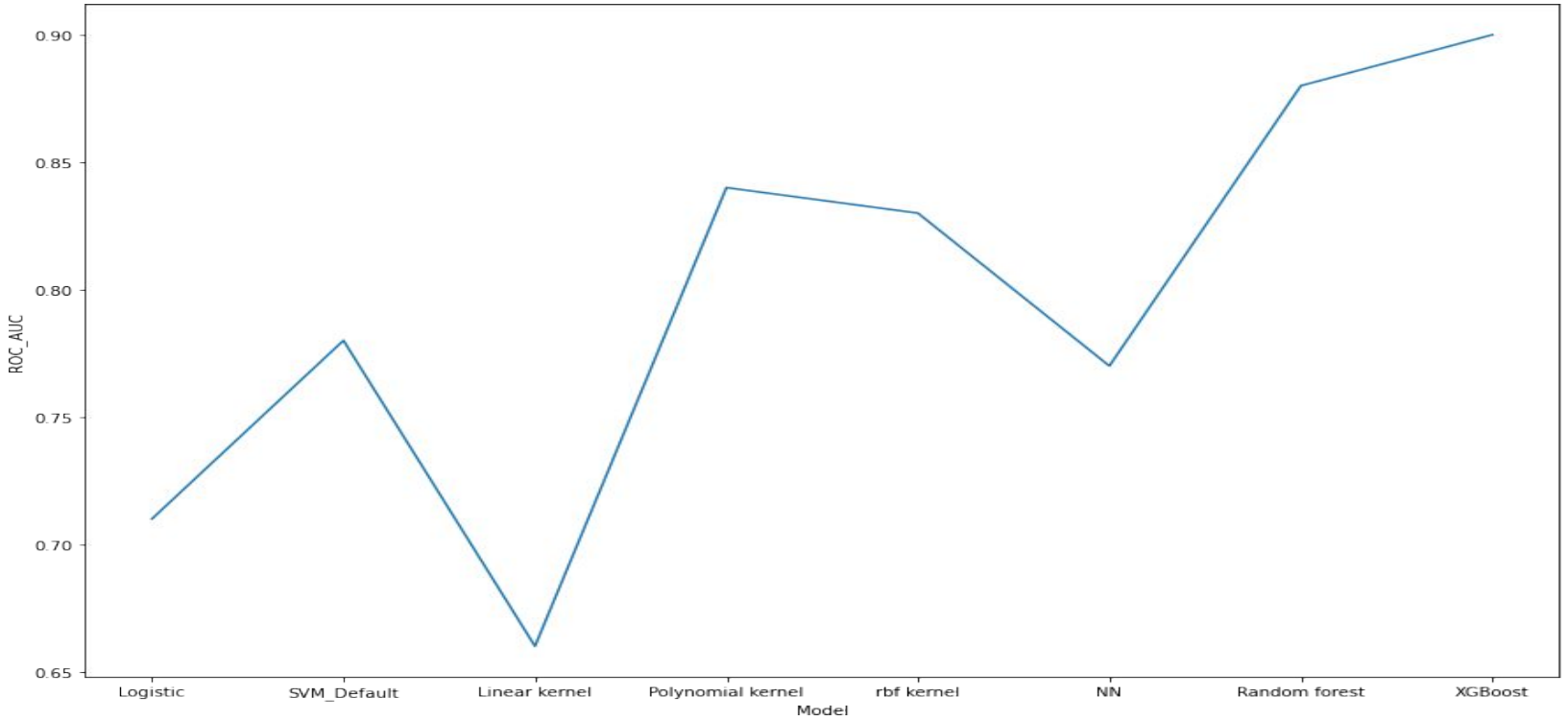


Feature Importance for Support Vector Machine:



Here we can see the important features which are helping model to get the good accuracy.

ROC AUC Comparison:



Challenges:

- Execution takes time.
- As there were many null values present in data set it took time to clean the dataset.
- Difficulty in selecting the appropriate graph for trend.

Conclusion:

- There are 15.1 % people in our dataset are is risk for cardiovascular disease and 84.9 % people are safe (ten year risk).
- There is more risk of cardiovascular disease in patients of age between 51 to 63.
- The count of male and female are same in risk which is around 200, though females are more than males in our dataset
- .Around 250 smokers are in risk and around 210 non-smokers are is risk for cardiovascular disease.
- We can't evidentially state smoking will lead to heart disease, as we seen from count plot there is no huge difference between these to commune and also our extreme smoker who smokes 70 cigarettes per day is not having ten year risk
- There are very few people who are done with BP medication which are around 200 but many people have not taken any BP medication and they are around 3200. We cannot say that after taking medication person is safe.
- Around 500 patients who did not had stroke yet and are at risk and around 2800 patients are safe.
- around 250 people with hypertensive are in risk and around 255 people with no hypertensive are at risk.
- Here we can see people who did not had diabetes are more and around 500 people who did not had diabetes are at risk.And there are very few people who had diabetes are at risk.
- most of the people who are not in risk their Cholesterol level lies between 210 to 280 and people who are in risk their cholestrol level lies between 215 to 285 there in not huge difference it is quite normal.
- most people who are not in risk their systolic BP lies between 110 to 140 and people who are at risk their systolic BP lies between 125 to 160. we can say people with high systolic BP are at risk.
- most people who are not in risk their diastolic BP lies between 75 to 85 and people who are at risk their diastolic BP lies between 89 to 90. we can say there is a slight increase in diastolic BP of people who are in risk.
- most people who are not in risk their BMI lies between 22 to 28 and people who are at risk their BMI lies between 23 to 29 approx. WE cannot see any difference BMI is approx. same of risky and not risky people

Conclusion:

- most people who are not in risk their heart rate lies between 68 to 83 and people who are at risk their heart rate lies between 68 to 84. which is same for risky and not risky people.
- there is not that difference between the glucose level of risky and non risky patients. glucose level lies between 70 to 80 for both risky and non risky patients.
- most people smoke cigarettes between 1 to 10 approx. and there heart rate lies between 60 to 100.
- age of people lies between 32 to 70 and most of the people's systolic BP lies between 90 to 200.
- With logistic regression we got the accuracy score of 0.68 on train data and 0.66 on test data.
- With Support vector machine we got the train accuracy score of 0.79 and test accuracy score of 0.78.
- With support vector linear kernel we got the train accuracy score of 0.67 and test accuracy score of 0.67.
- With support vector polynomial kernel we got the train accuracy score of 0.95 and test accuracy score of 0.84.
- With Gaussian radial base kernel we got the train accuracy score of 0.97 and test accuracy score of 0.83.
- With Neural Networks we got the Train Accuracy score of 0.81 and test accuracy score of 0.77.
- With Random forest classifier we got the train accuracy of 0.97 and test accuracy of 0.89.
- With XGBoost classifier we got the train accuracy score of 0.99 and test accuracy of 0.90.


Thank
you!