CLASSIFICATION CAPSTONE PROJECT

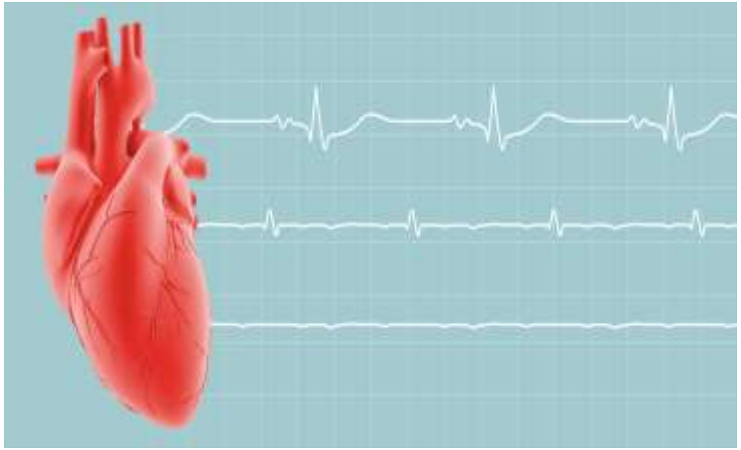
CORONARY HEART DISEASE PREDICTION

Presented By:

- **Raj Vijay Sonar (Cohort Vindhya)**

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 both demographic, behavioral, and medical risk factors.



The traditional risk factors for coronary heart disease are high LDL cholesterol, low HDL cholesterol, high blood pressure, family history, diabetes, smoking and being older than 45 for men

DATA DESCRIPTION

Demographic:

- sex: Male or Female("M" or "F")
- age: Age of the patient (Continuous)
- education: Education level of a person

Behavioral:

- is_smoking: Whether or not the patient is a current smoker ("YES" or "NO")
- cigsPerDay: The number of cigarettes that the person smoked on an average in one day (Continuous)

Medical (History):

- BPMeds: Whether or not the patient was on blood pressure medication (Nominal)
- prevalentStroke: Whether or not the patient had previously had a stroke (Nominal)
- prevalentHyp: Whether or not the patient was hypertensive (Nominal)
- diabetes: Whether or not the patient had diabetes (Nominal)

Medical (Current):

- totChol: Total Cholesterol Level (Continuous)
- sysBP: Systolic Blood Pressure (Continuous)
- diaBP: Diastolic Blood Pressure (Continuous)
- BMI: Body Mass Index (Continuous)
- heartRate: Heart Rate (Continuous)
- glucose: Glucose Level (Continuous)

Predict Variable (Desired Target):

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

The dataset consists of 3390 rows and 16 columns out of which 8 columns are continuous and 7 are categorical and 1 column is the target column which is also binary categorical

DATA PRE-PROCESSING

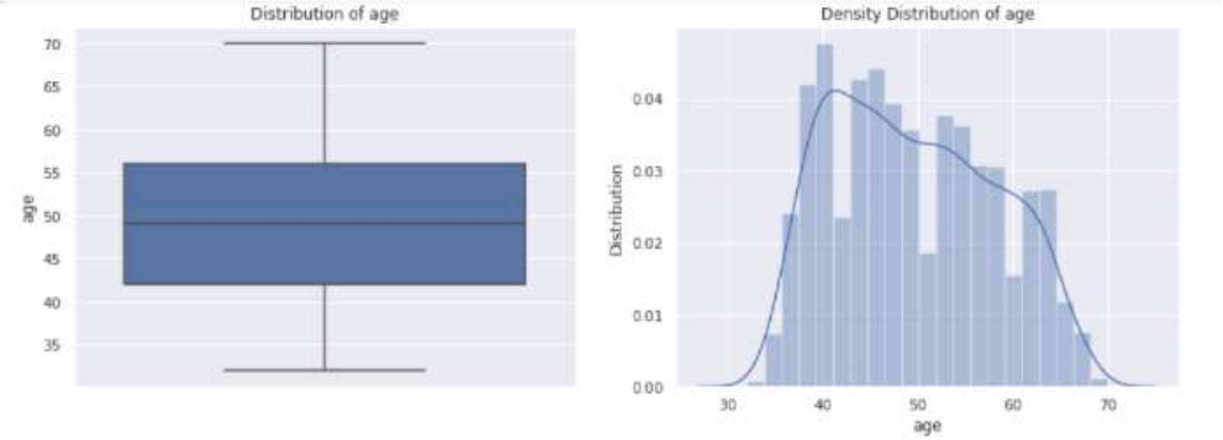
- **Handling Missing Values**

1. The columns ‘cigsPerDay’, ‘education’, ‘BPMeds’, ‘totChol’, ‘BMI’, ‘heartRate’ and ‘glucose’ had missing (Nan) values but not much and were manageable.
2. Imputed the missing values with the median of the respective columns for all the variables except ‘cigsPerDay’.
3. The median of ‘cigsPerDay’ column was 0 but after checking the ‘is_smoking’ column found that the people has smoking habits, so imputed those Nan values with the mean of the column which was around 9.

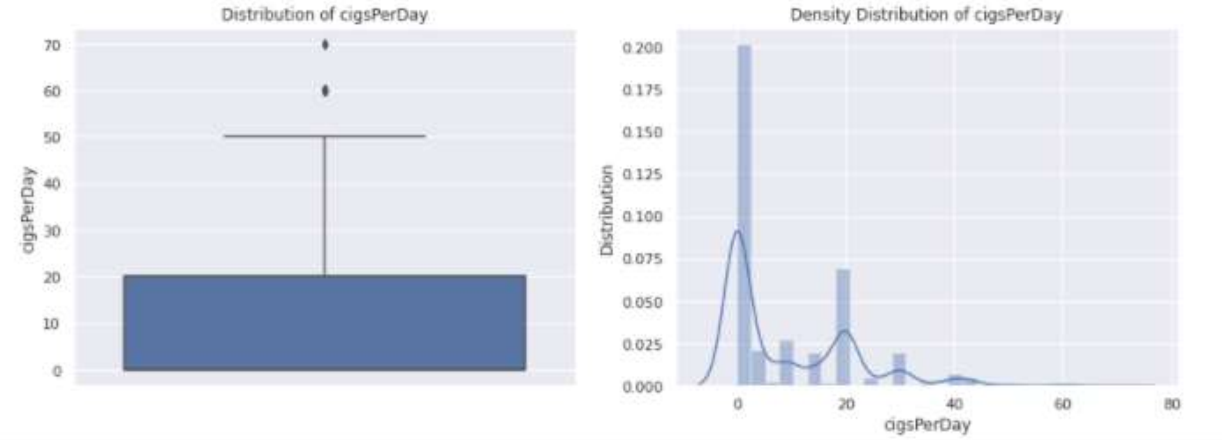
- **Converting Data-Type**

Converting the data-type of the columns ‘education’ and ‘BPMeds’ from float to integer for uniformity.

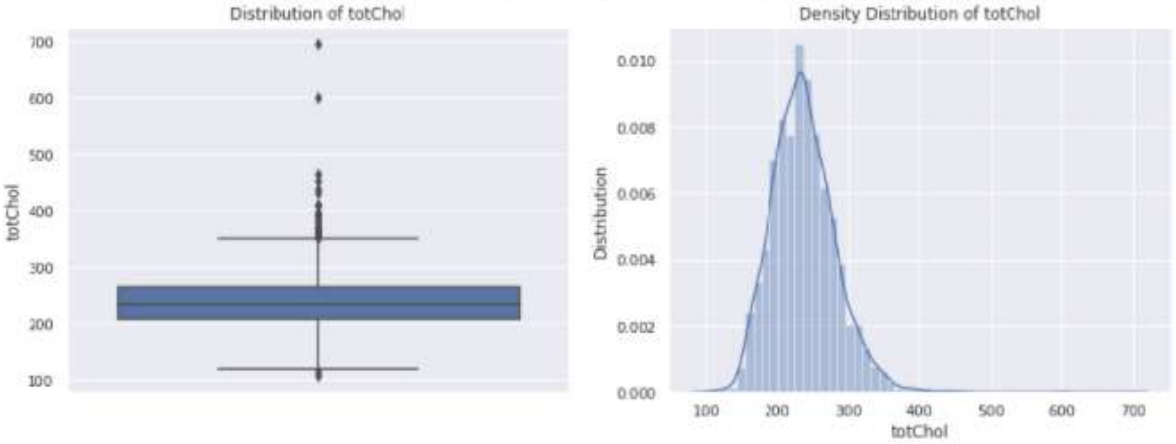
EXPLORATORY DATA ANALYSIS
UNIVARIATE ANALYSIS



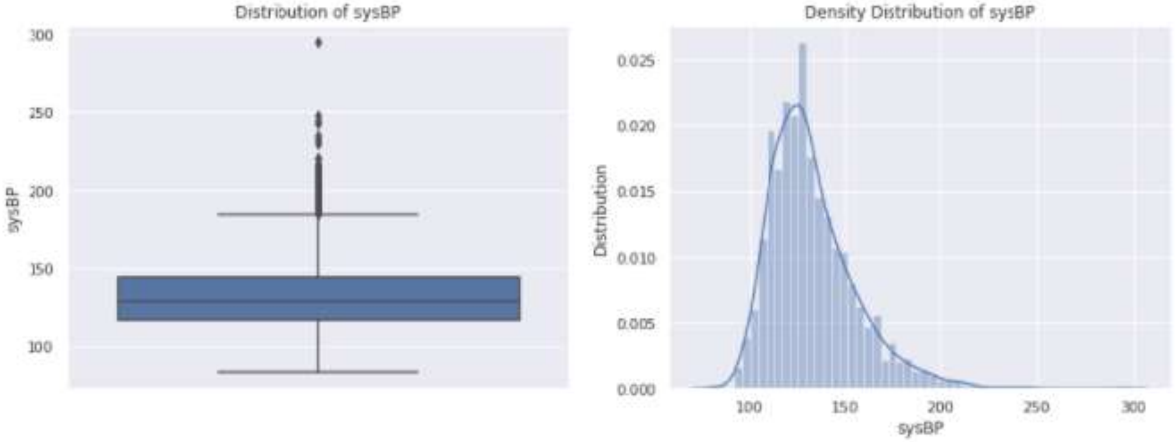
Boxplot and density distribution plot of the column ‘age’



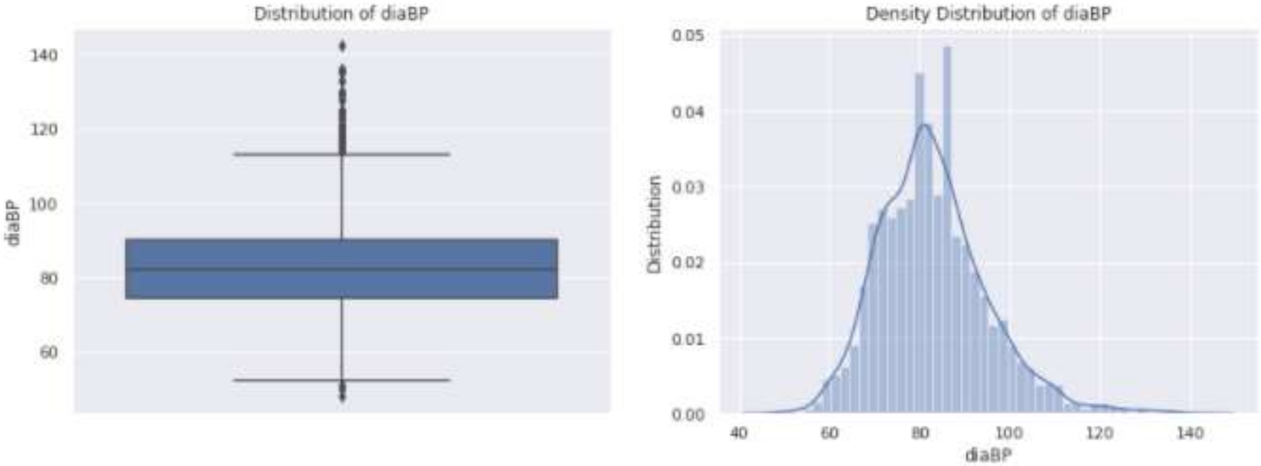
Boxplot and density distribution plot of the column ‘cigsPerDay’



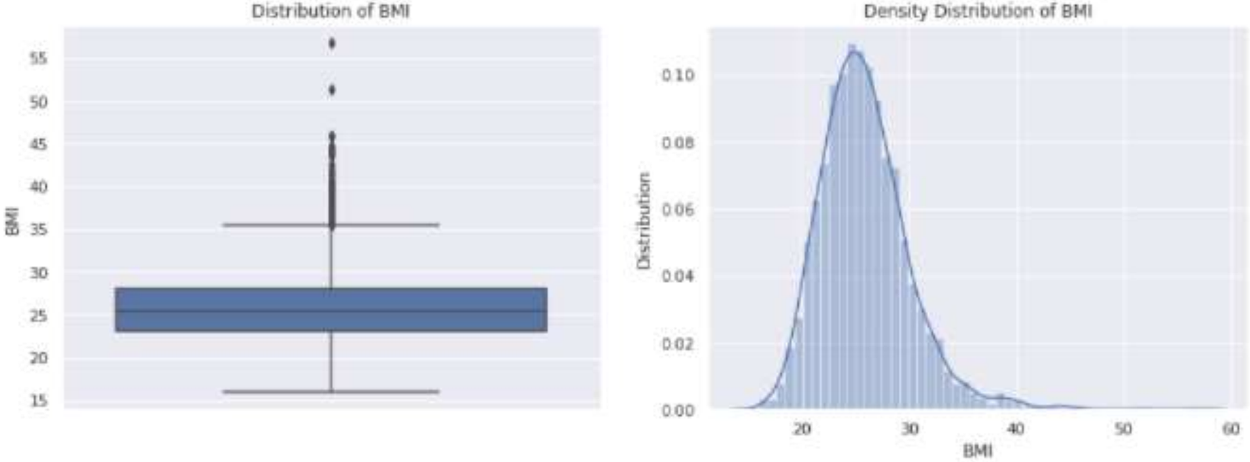
Boxplot and density distribution plot of the column ‘totChol’



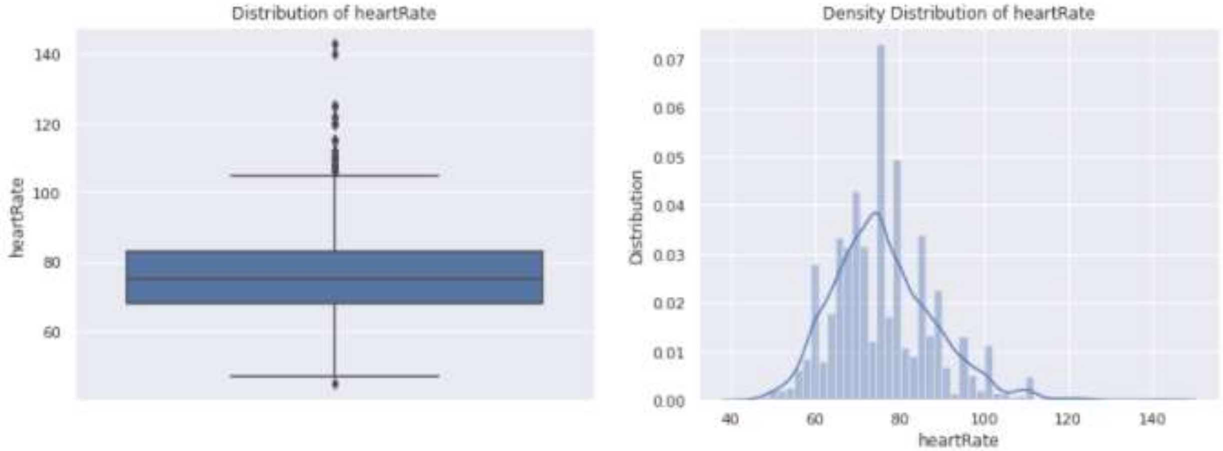
Boxplot and density distribution plot of the column ‘sysBP’



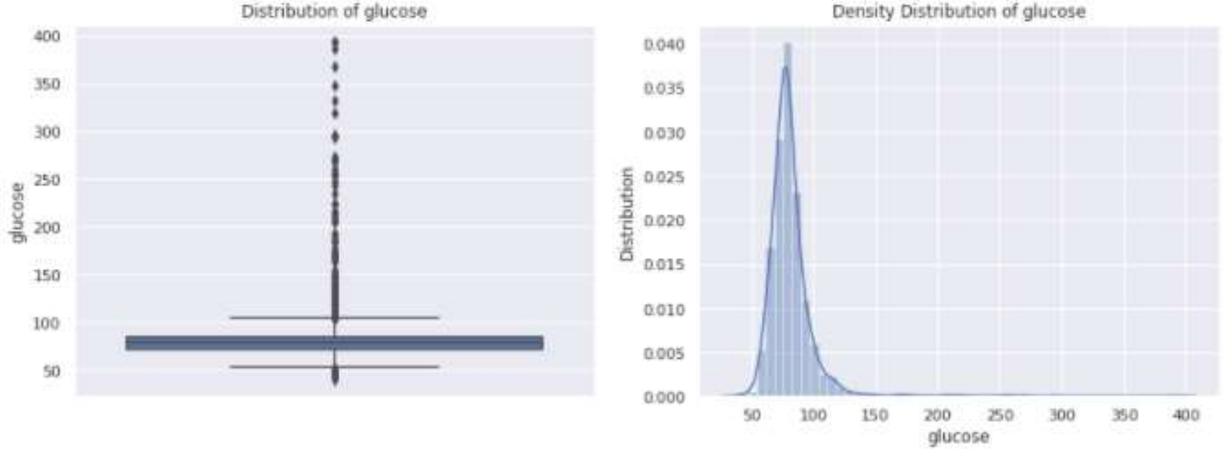
Boxplot and density distribution plot of the column ‘diaBP’



Boxplot and density distribution plot of the column ‘BMI’

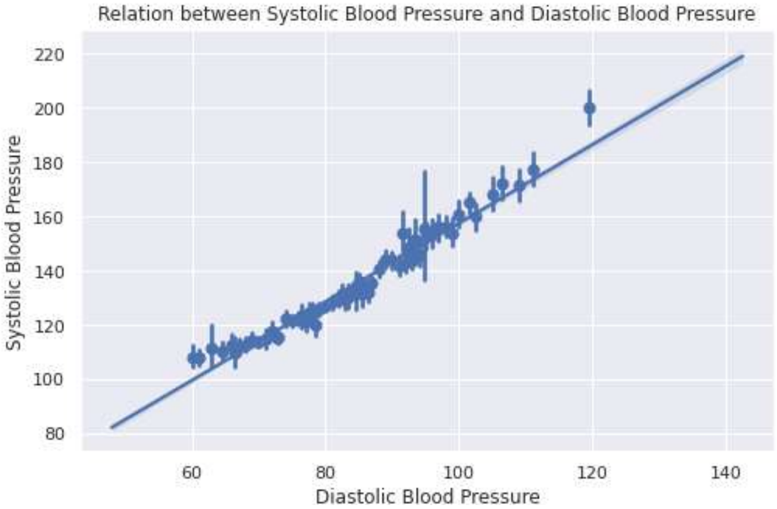


Boxplot and density distribution plot of the column ‘heartRate’

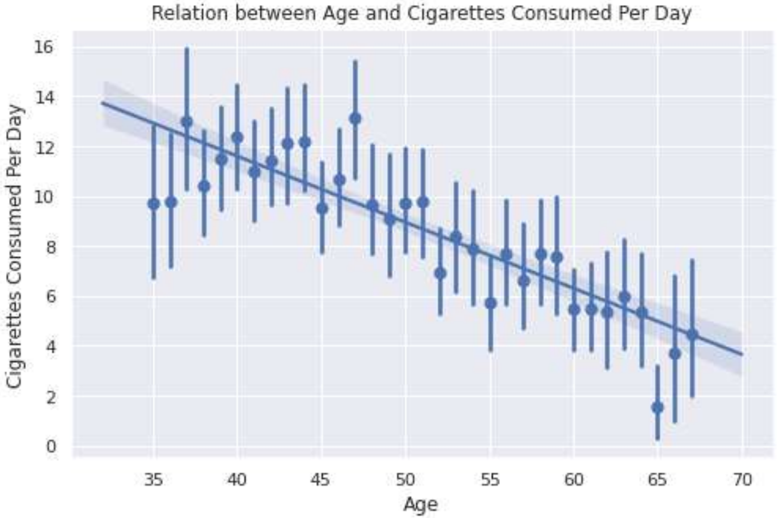


Boxplot and density distribution plot of the column ‘glucose’

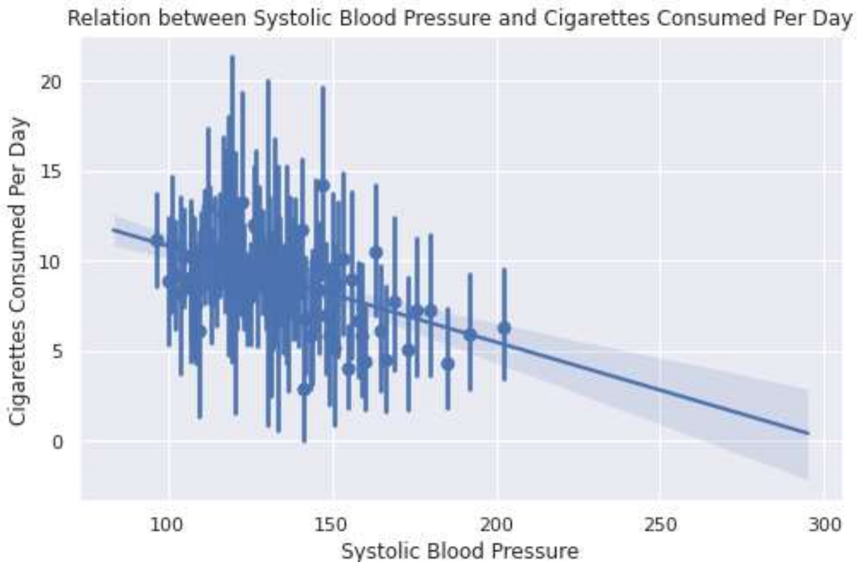
EXPLORATORY DATA ANALYSIS
BIVARIATE ANALYSIS



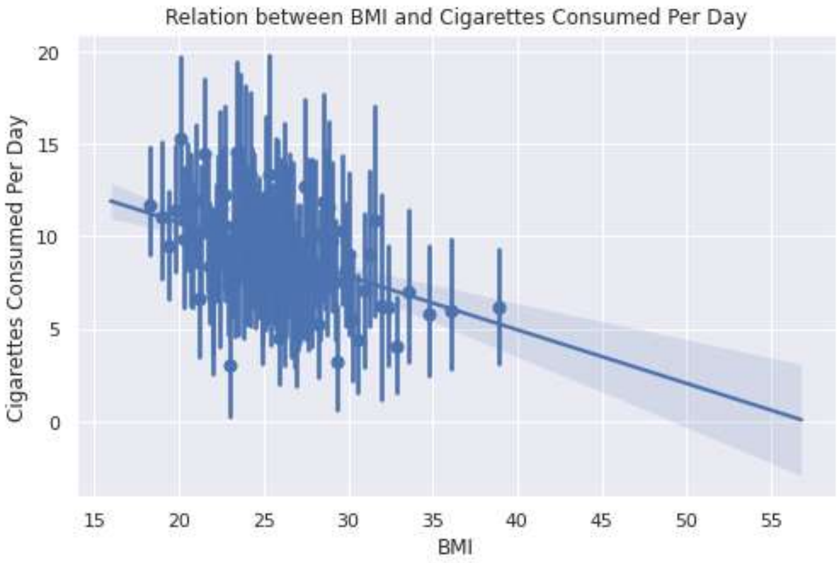
It can be observed that the relation between Systolic Blood Pressure and Diastolic Blood Pressure is always linear with each other, so both of them give the same correlation and behaviour with the other variables



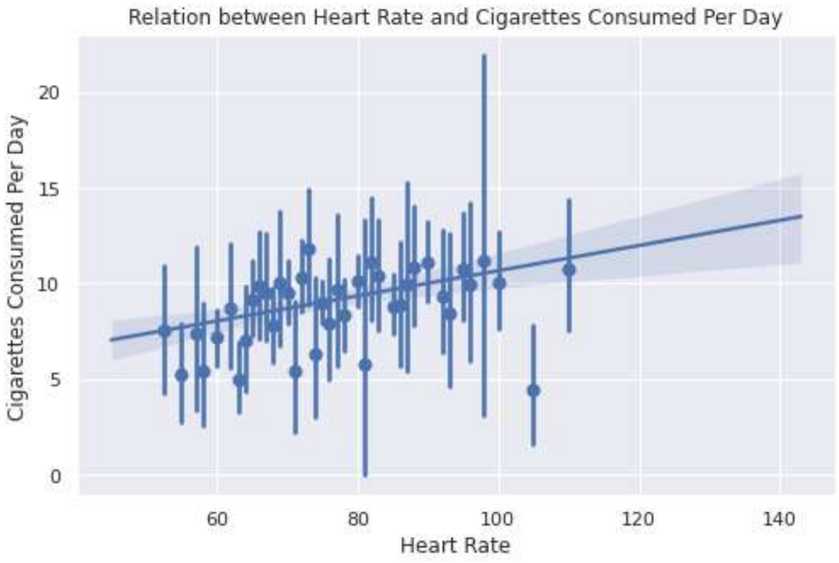
It can be seen that the consumption of cigarettes per day decreases with the increase in the age

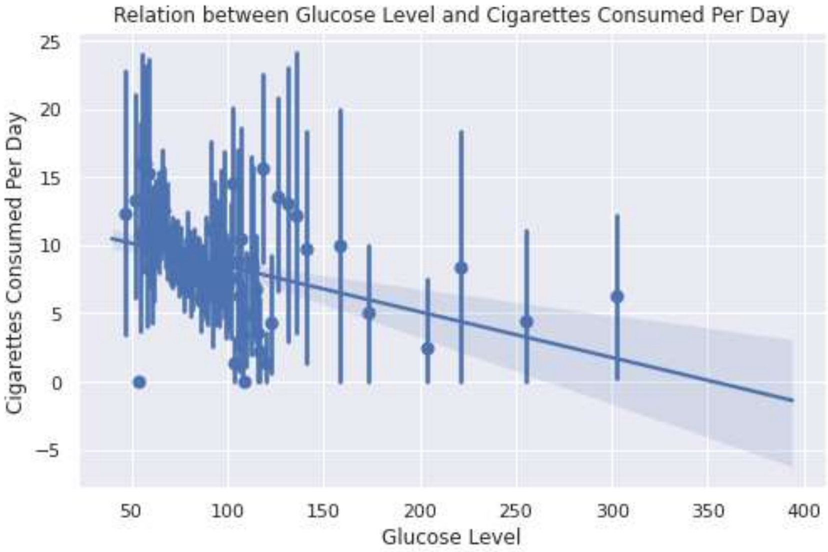


As the consumption of the cigarettes per day increases the Systolic Blood Pressure decreases



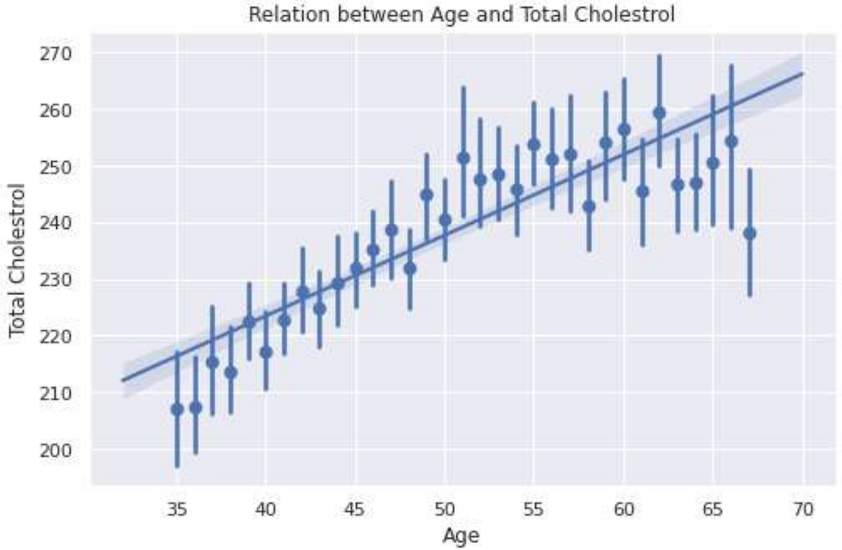
As the consumption of cigarettes per day increases the BMI decreases



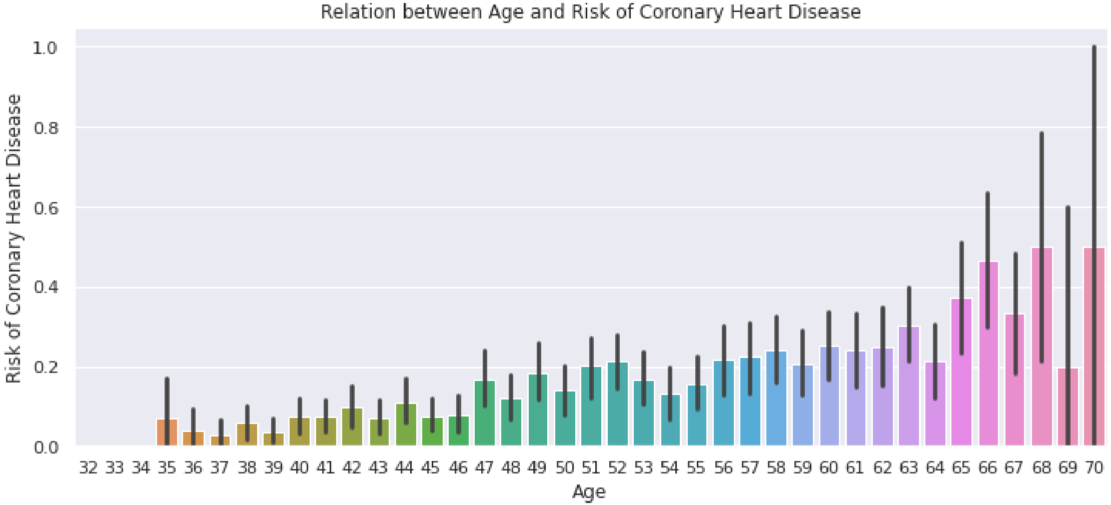


As the consumption of the cigarettes per day increases the Heart Rate of a person also increases

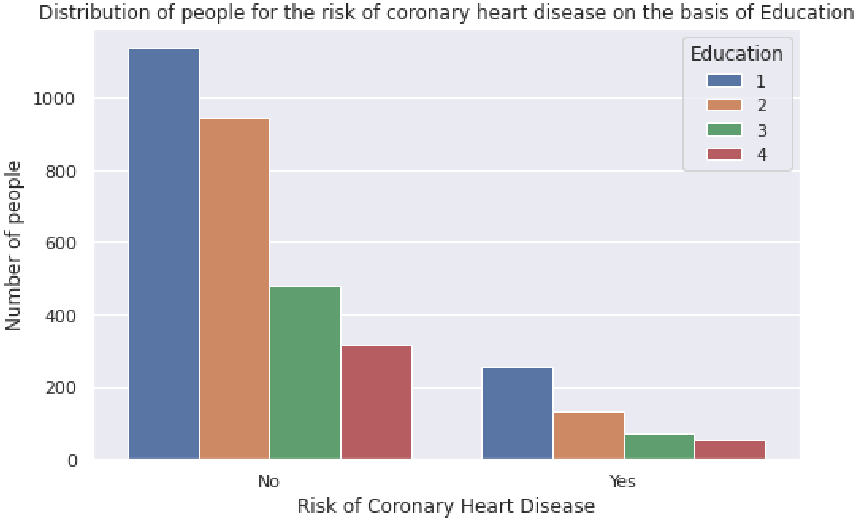
As the consumption of the cigarettes per day increases the glucose level decreases



As the age of a person increases the total cholesterol of the body also increases

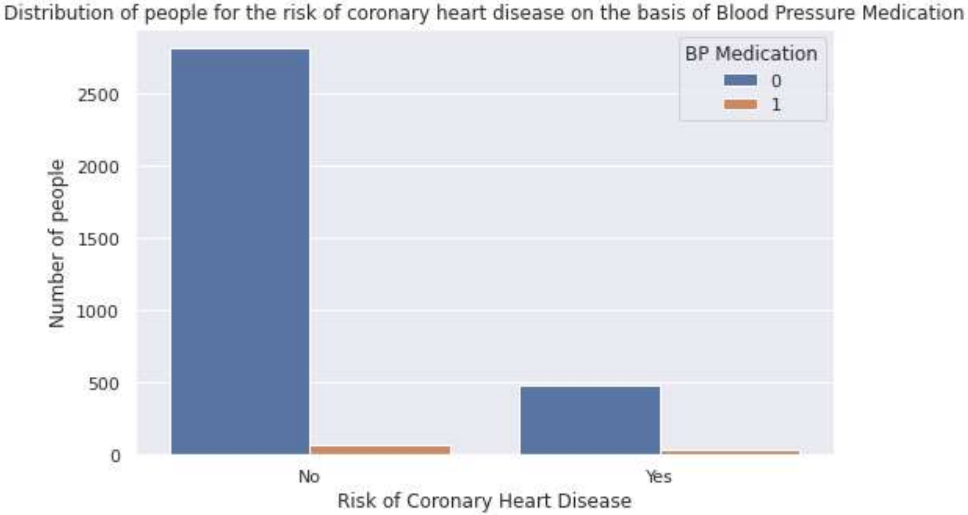


It can be observed that as the age of a person increases the probability of getting the risk of Coronary Heart Disease also increases, not that significantly but by a small proportion



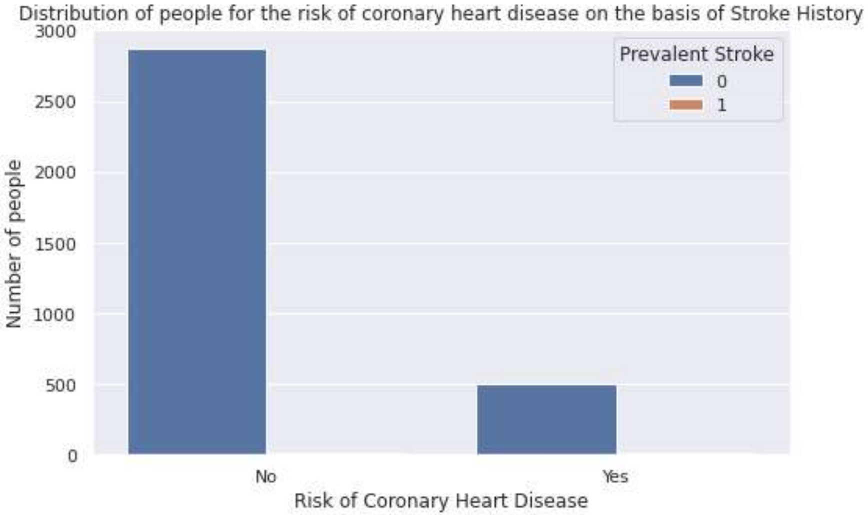
Distribution of people for the risk of coronary heart disease on the basis of Education

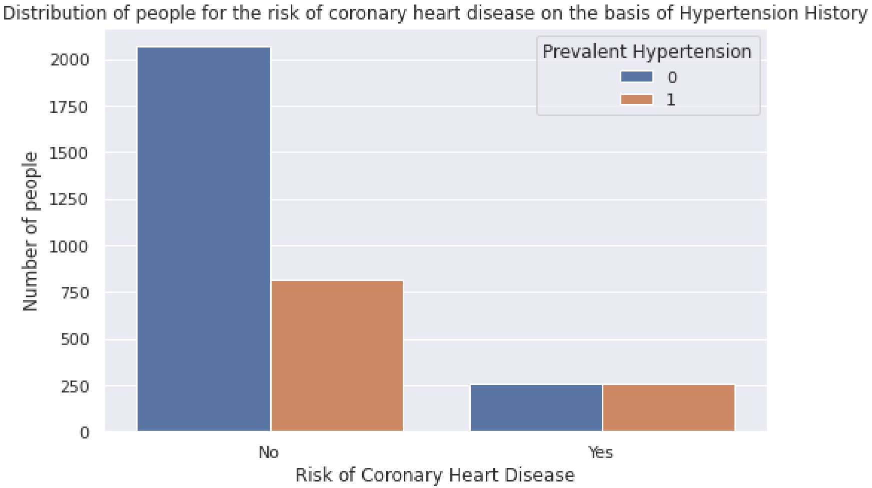
It can be studied that as the level of education is increasing the risk of getting the risk of Coronary Heart Disease is decreasing, it can be because people getting awareness through education



Distribution of people for the risk of coronary heart disease on the basis of Blood Pressure Medication

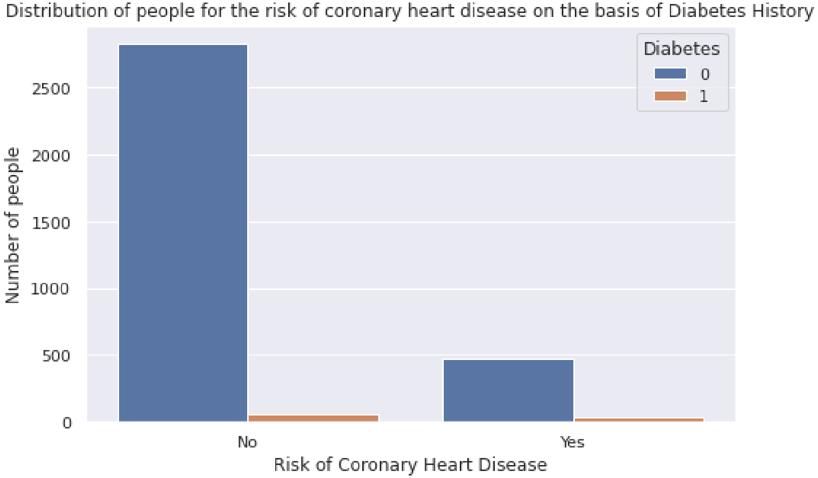
It can be seen that very few people are on the Blood Pressure Medication and so Blood Pressure Medication is not causing a significant value towards the risk of Coronary Heart Disease

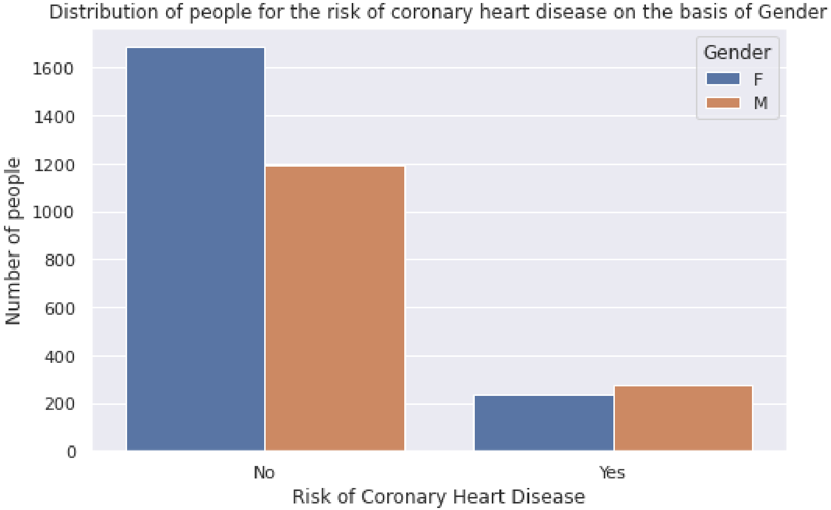




It can be seen that very few people are having the history of Heart Stroke, almost negligible and so the variable is not playing a crucial role in deciding the risk of Coronary Heart Disease

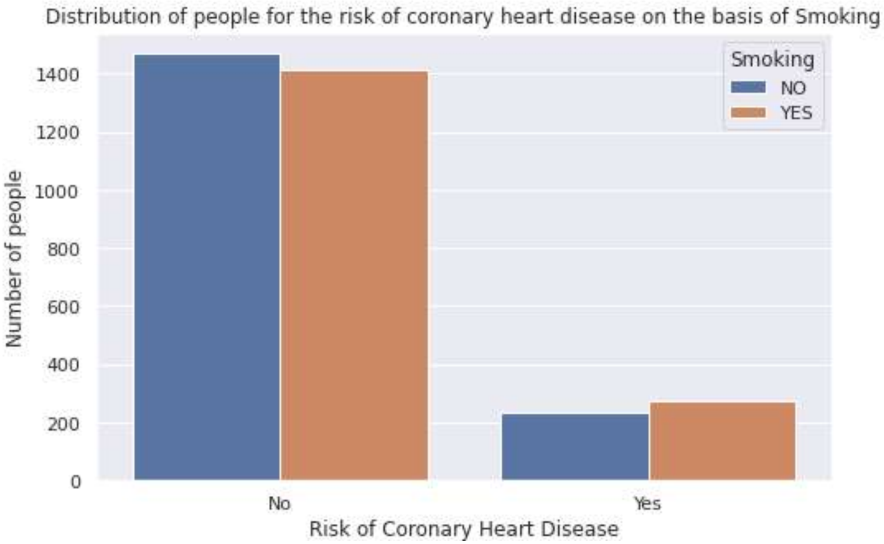
People having Hypertension history or not are almost the same towards the risk of Coronary Heart Disease





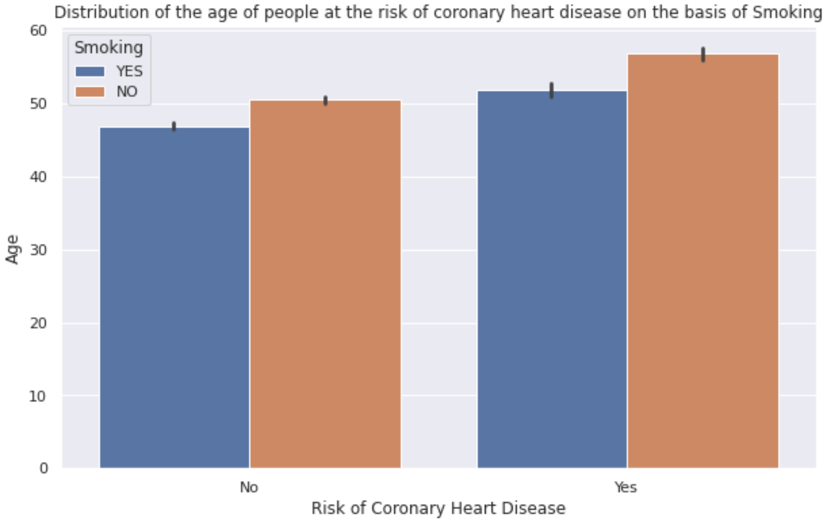
Having or not having diabetes is not adding any value towards the risk of Coronary Heart Disease as very few number of people are suffering from it

It can be seen that Males are a bit more vulnerable to the risk of Coronary Heart Disease as compared to Females

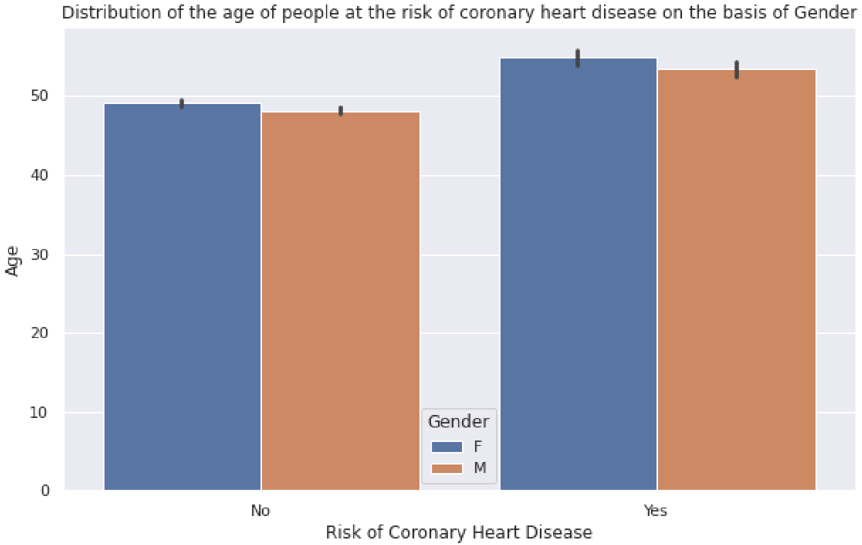


People who have smoking habits are more inclined towards the risk of Coronary Heart Disease compared to the people who do not have smoking habits

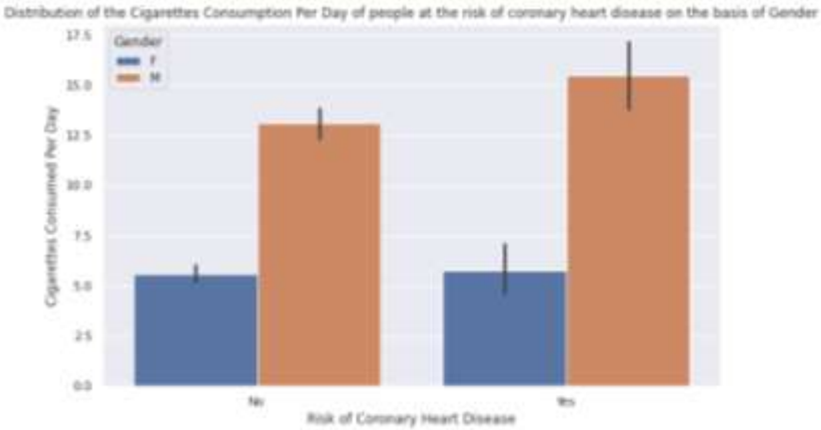
EXPLORATORY DATA ANALYSIS
MULTIVARIATE ANALYSIS

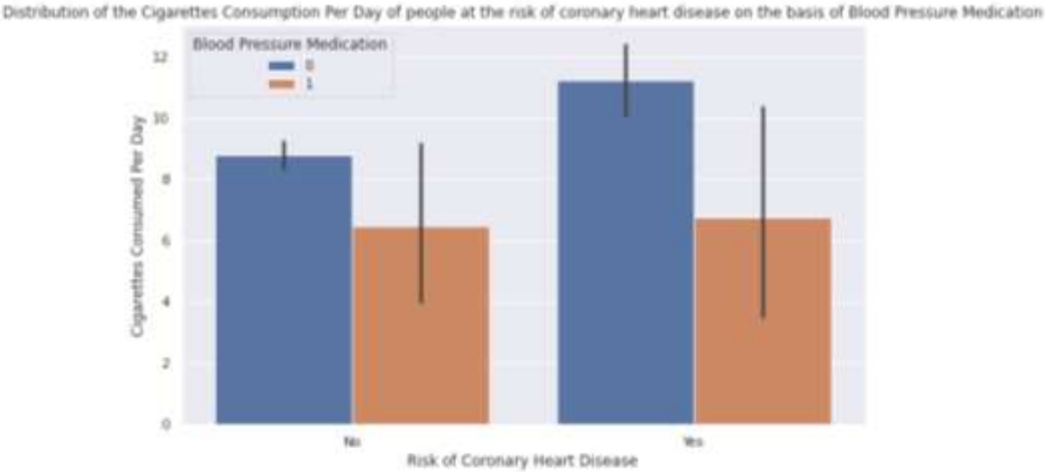


When compared between smoking habit and age, age plays a vital role towards the risk of Coronary Heart Disease



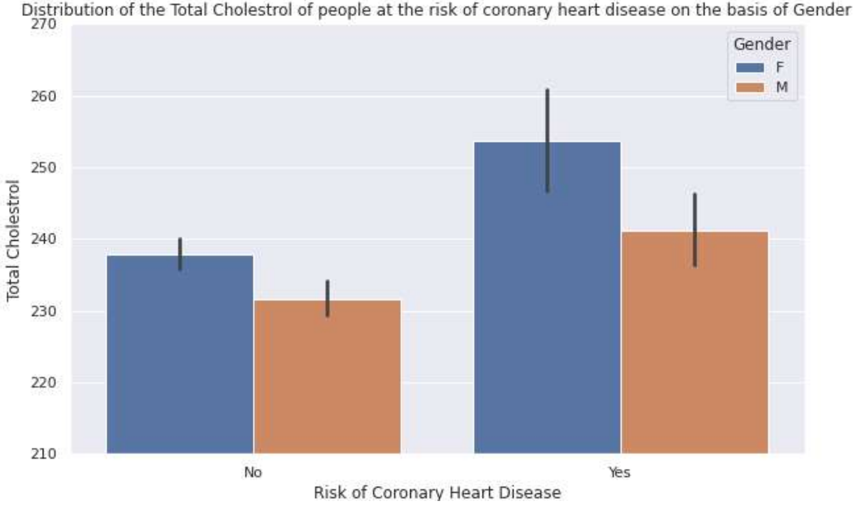
When compared between age and gender, there can not be seen a remarkable deciding factor towards the risk of Coronary Heart Disease





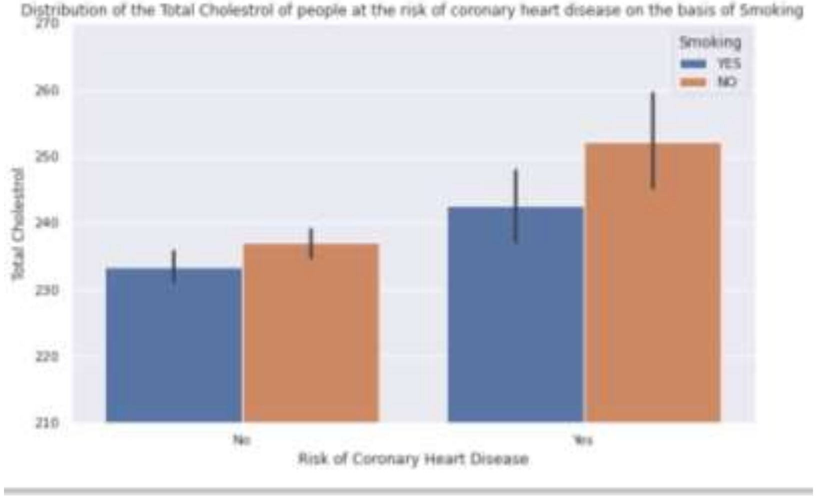
It can be concluded from the graph that more number of males consume cigarettes as compared to females and thus males are more likely to get the risk of Coronary Heart Disease

Cigarettes consumed per day plays an important role in deciding the risk of Coronary Heart Disease when it comes to the comparison with Blood Pressure Medication. Even a person who is not medicated towards Blood Pressure still can get risky towards Coronary Heart Disease if cigarettes consumption per day is high



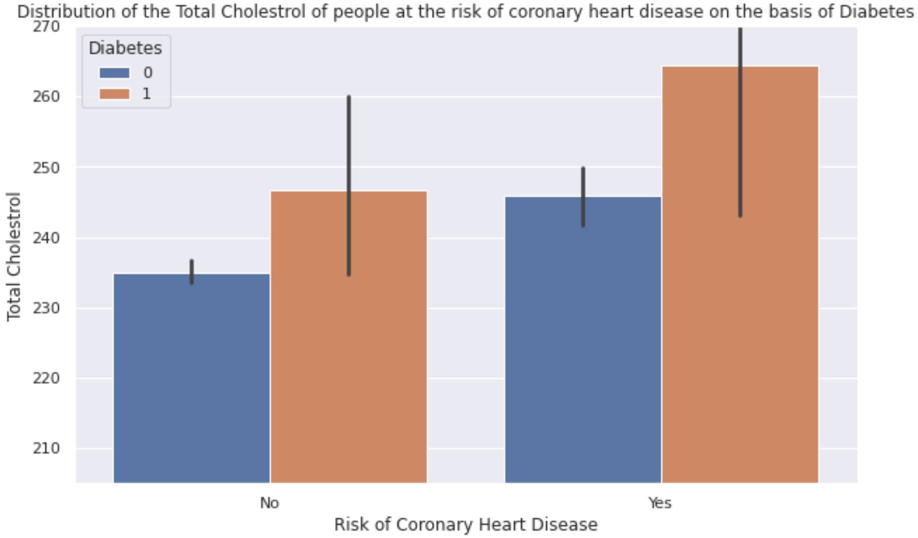
Distribution of the Total Cholestrol of people at the risk of coronary heart disease on the basis of Gender

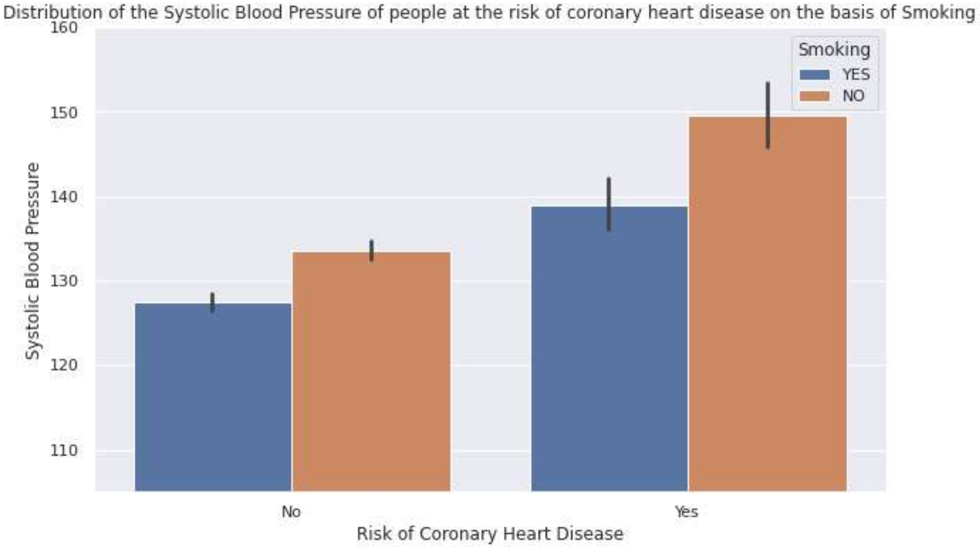
It can be concluded that females have higher amount of total cholesterol as compared to males and thus it leads to the risk of Coronary Heart Disease in females more than males



Distribution of the Total Cholestrol of people at the risk of coronary heart disease on the basis of Smoking

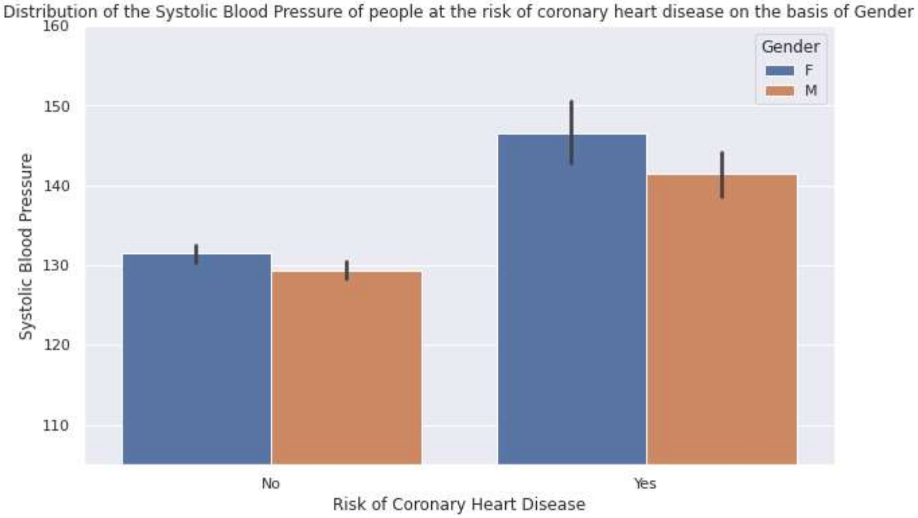
Total cholesterol leads towards the risk of Coronary Heart Disease when compared with smoking habit, even if the person is not having smoking habit can be at a risk of Coronary Heart Disease if the cholesterol levels are high

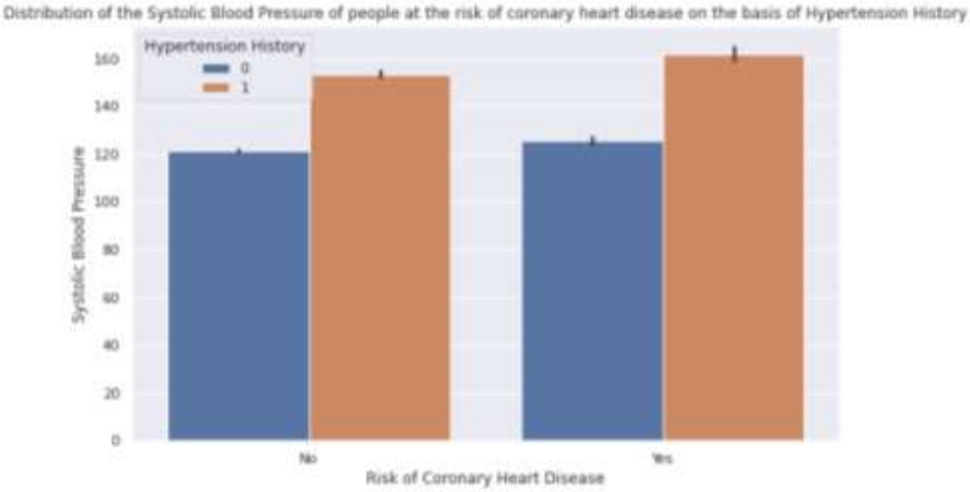




People having diabetes tend to have more probability of having the risk of Coronary Heart Disease irrespective of the total cholesterol levels

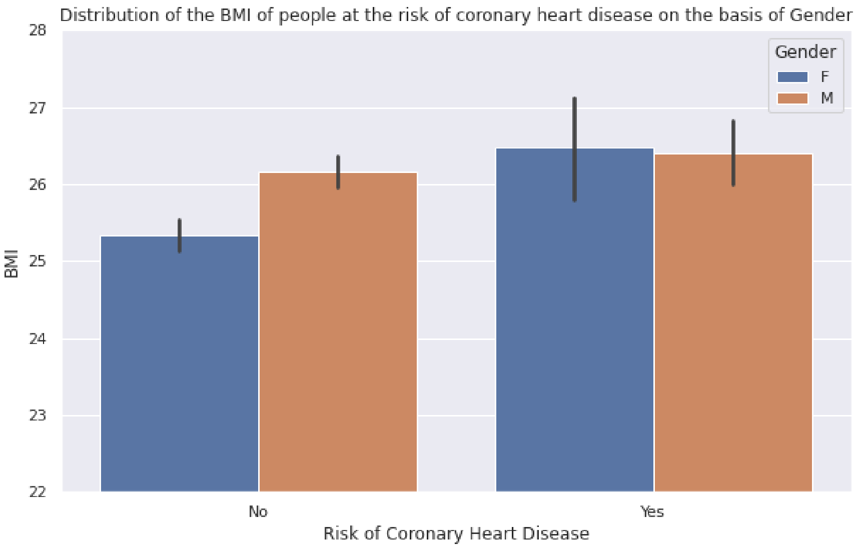
It shows that even people do not have smoking habits but can develop the risk of Coronary Heart Disease if the Systolic Blood Pressure is high

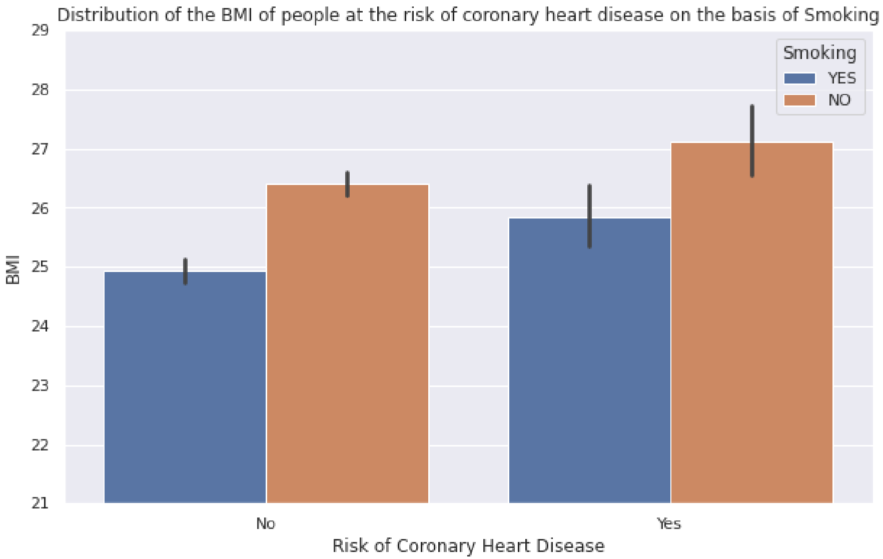




Females can be seen having higher Systolic Blood Pressure and have a higher tendency of getting the risk of Coronary Heart Disease

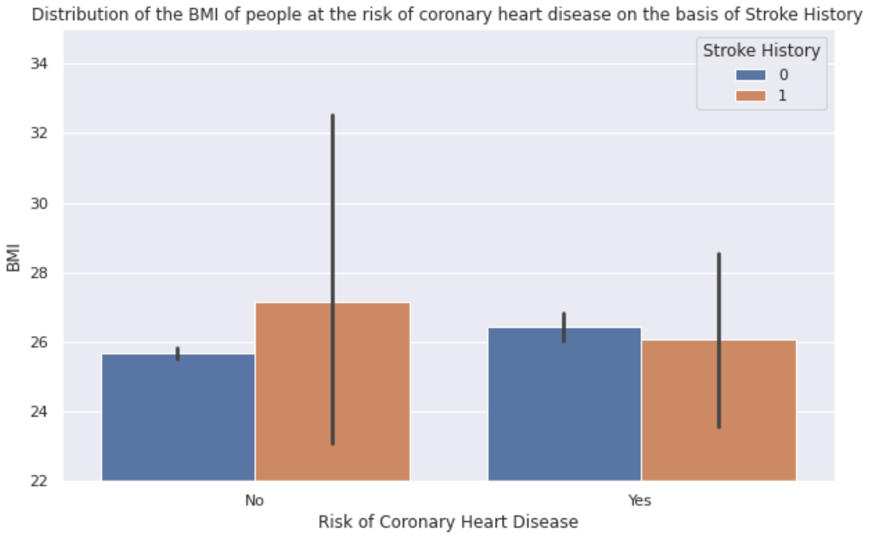
Hypertension history of people helps to determine the risk of Coronary Heart Disease when on the same systolic blood pressure level.





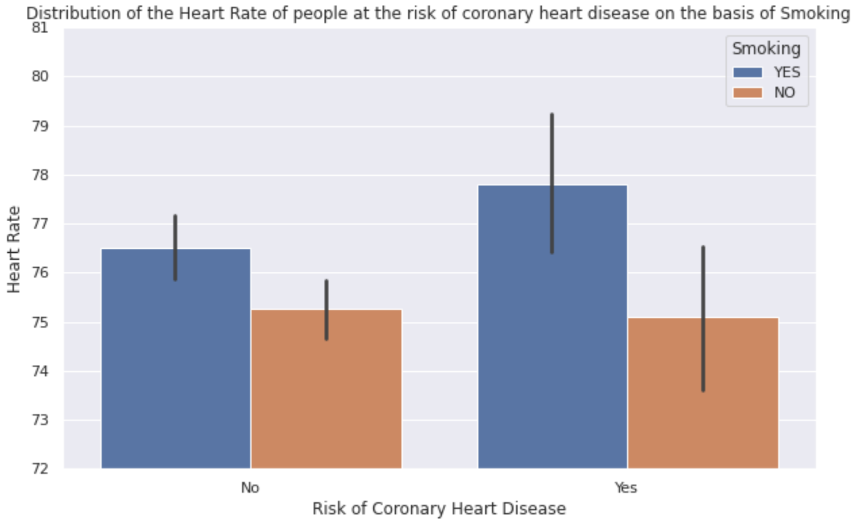
Gender does not play a vital role in deciding the risk of Coronary Heart Disease when it comes to the comparison with BMI

Higher BMI leads to the risk of Coronary Heart Disease irrespective of the smoking habit of people



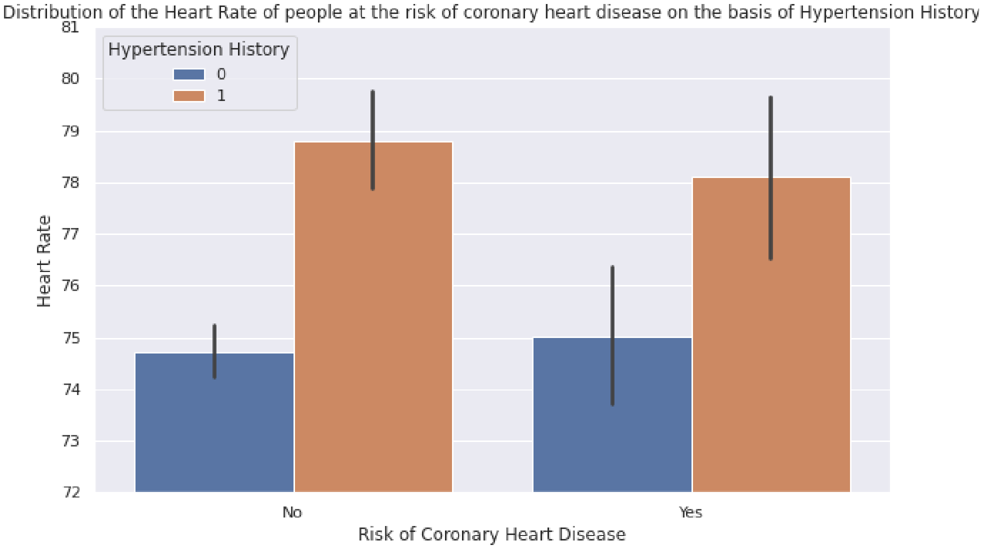
Distribution of the BMI of people at the risk of coronary heart disease on the basis of Stroke History

Stroke History is not helping to decide the risk of Coronary Heart Disease when compared with the BMI level



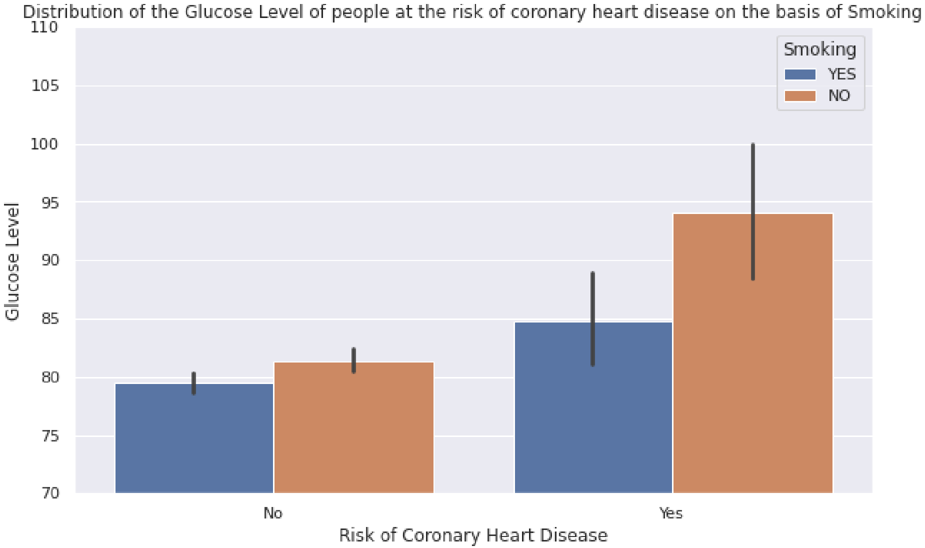
Distribution of the Heart Rate of people at the risk of coronary heart disease on the basis of Smoking

It helps to conclude that having smoking habit increases the heart rate and smoking habit forms significant evidence in deciding the risk of Coronary Heart Disease



Distribution of the Heart Rate of people at the risk of coronary heart disease on the basis of Hypertension History

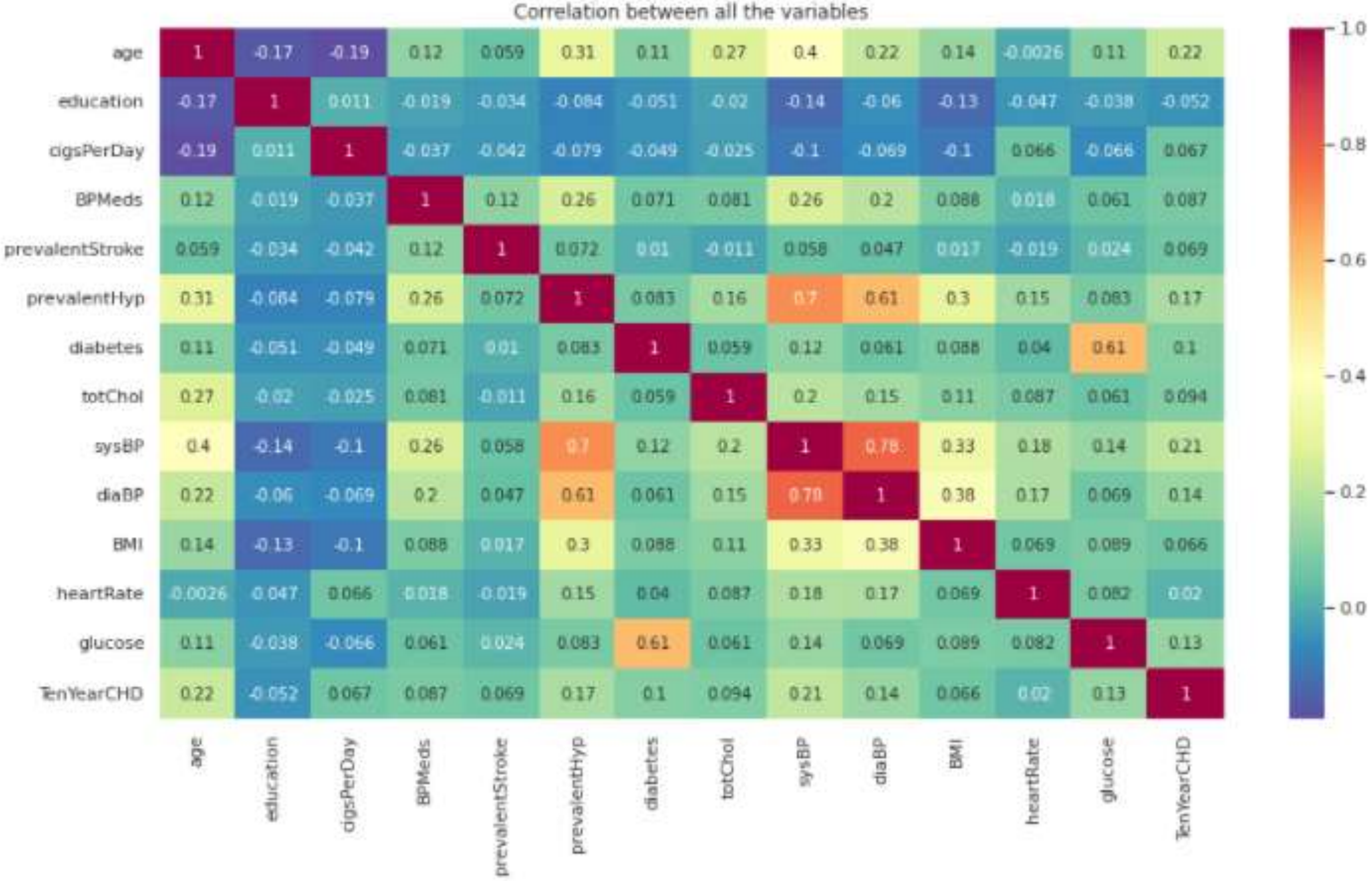
Hypertension History signifies the risk of Coronary Heart Disease when compared with the heart rate. People with hypertension history can be seen having a risk of Coronary Heart Disease



Distribution of the Glucose Level of people at the risk of coronary heart disease on the basis of Smoking

Higher glucose level helps to conclude the higher risk of Coronary Heart Disease irrespective of the smoking habits

CORRELATION MATRIX OF ALL THE VARIABLES



Correlation between all the variables

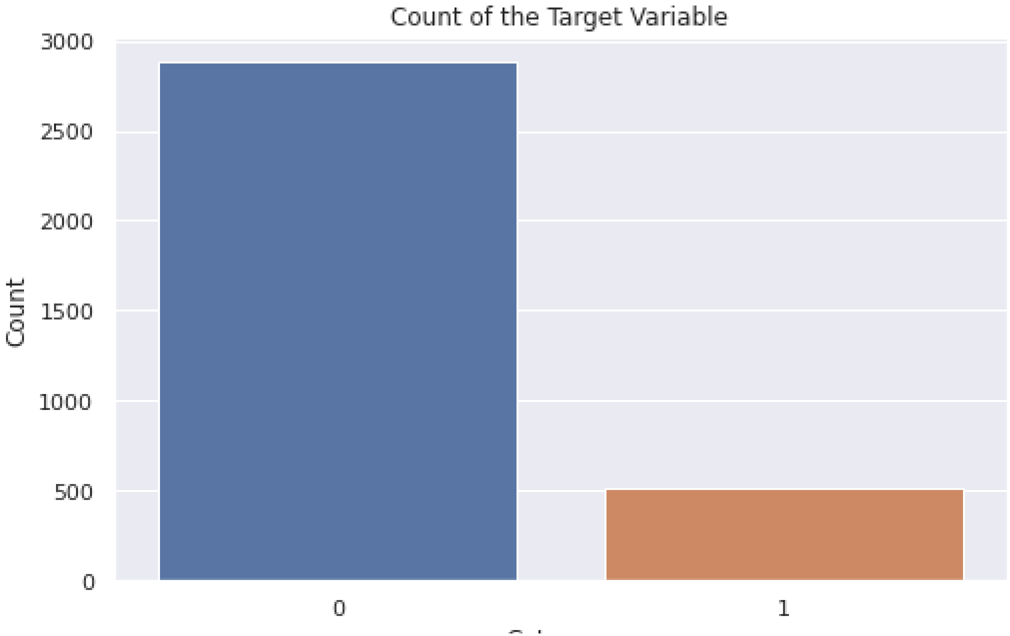
	age	education	cigsPerDay	BPMeds	prevalentStroke	prevalentHyp	diabetes	totChol	sysBP	diaBP	BMI	heartRate	glucose	TenYearCHD
age	1	-0.17	-0.19	0.12	0.059	0.31	0.11	0.27	0.4	0.22	0.14	-0.0026	0.11	0.22
education	-0.17	1	0.011	-0.019	-0.034	-0.084	-0.051	-0.02	-0.14	-0.06	-0.13	-0.047	-0.038	-0.052
cigsPerDay	-0.19	0.011	1	-0.037	-0.042	-0.079	-0.049	-0.025	-0.1	-0.069	-0.1	0.066	-0.066	0.067
BPMeds	0.12	-0.019	-0.037	1	0.12	0.26	0.071	0.081	0.26	0.2	0.088	0.018	0.061	0.087
prevalentStroke	0.059	-0.034	-0.042	0.12	1	0.072	0.01	-0.011	0.058	0.047	0.017	-0.019	0.024	0.069
prevalentHyp	0.31	-0.084	-0.079	0.26	0.072	1	0.083	0.16	0.7	0.61	0.3	0.15	0.083	0.17
diabetes	0.11	-0.051	-0.049	0.071	0.01	0.083	1	0.059	0.12	0.061	0.088	0.04	0.61	0.1
totChol	0.27	-0.02	-0.025	0.081	-0.011	0.16	0.059	1	0.2	0.15	0.11	0.087	0.061	0.094
sysBP	0.4	-0.14	-0.1	0.26	0.058	0.7	0.12	0.2	1	0.78	0.33	0.18	0.14	0.21
diaBP	0.22	-0.06	-0.069	0.2	0.047	0.61	0.061	0.15	0.78	1	0.38	0.17	0.069	0.14
BMI	0.14	-0.13	-0.1	0.088	0.017	0.3	0.088	0.11	0.33	0.38	1	0.069	0.089	0.066
heartRate	-0.0026	-0.047	0.066	0.018	-0.019	0.15	0.04	0.087	0.18	0.17	0.069	1	0.082	0.02
glucose	0.11	-0.038	-0.066	0.061	0.024	0.083	0.61	0.061	0.14	0.069	0.089	0.082	1	0.13
TenYearCHD	0.22	-0.052	0.067	0.087	0.069	0.17	0.1	0.094	0.21	0.14	0.066	0.02	0.13	1

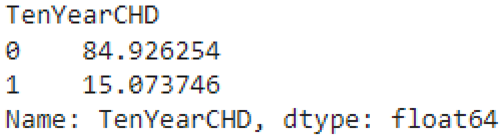
DATA PREPARATION FOR MODELLING

1. Converting the columns ‘sex’ and ‘is_smoking’ from string values to numerical values.
2. Splitting the data into X and Y as independent and dependent variable respectively.
3. Dropping the columns ‘diabetes’, ‘prevalentStroke’, ‘BPMeds’ and ‘diaBP’ from the independent variables because the number of people having diabetes, stroke history and on Blood Pressure Medication is very minimal.

   Diastolic Blood Pressure makes collinear relationship with Systolic Blood Pressure and thus using Systolic Blood Pressure is enough to train the model and it also helps in removing multi-collinearity.
4. Splitting the data into training and testing dataset with 25% data as testing dataset and remaining 75% data as training dataset.
5. Scaling the data using Standard Scaler function to make all the numerical values lie in the same range for the better accuracy of model.

HANDLING IMBALANCED DATASET





```
TenYearCHD
0    84.926254
1    15.073746
Name: TenYearCHD, dtype: float64
```

It can be observed that the target variable of the dataset is imbalanced. The value of getting the risk of Coronary Heart Disease is 15.07% whereas the value of not getting the risk of Coronary Heart Disease is 84.92%.

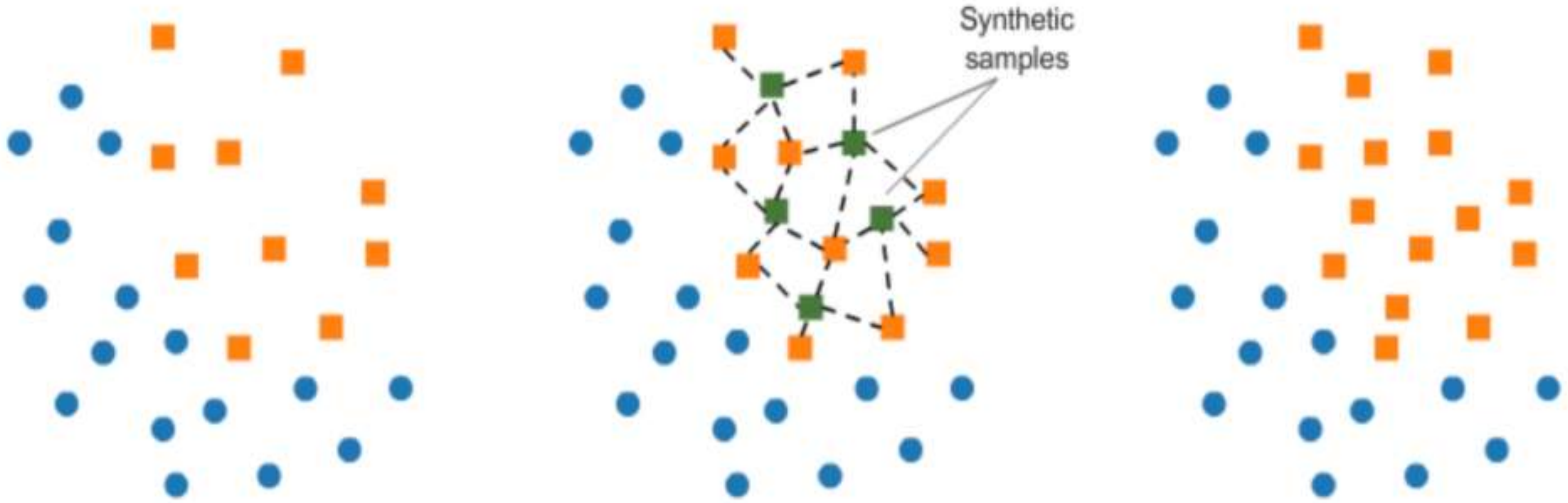
To deal with the imbalanced dataset an oversampling technique known as SMOTE (Synthetic Minority Oversampling Technique) was used.

SMOTE generates synthetic data for the minority class. It works by randomly picking a point from the minority class and computing the k-nearest neighbors for this point. The synthetic points are added between the chosen point and its neighbors.

WORKING OF SMOTE

SMOTE algorithm works in 4 simple steps:

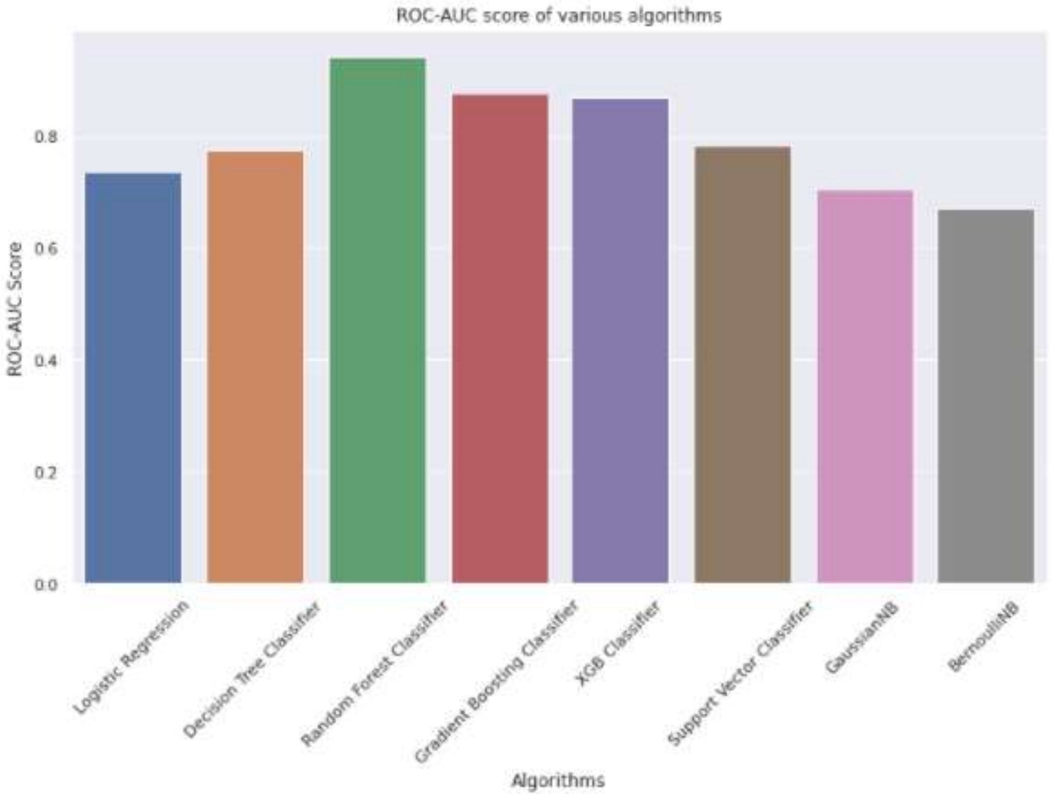
1. Choose a minority class as the input vector
2. Find its k nearest neighbors (k-neighbors is specified as an argument in the SMOTE() function)
3. Choose one of these neighbors and place a synthetic point anywhere on the line joining the point under consideration and its chosen neighbor
4. Repeat the steps until data is balanced



TRAINING THE MODEL

The classification algorithms are made to run over the training dataset with the use of cross-validation and then the algorithms showing the best results are selected for further hyper-parameter tuning.
Classification algorithms used:

1. Logistic Regression
2. Decision Tree Classifier
3. Random Forest Classifier
4. Gradient Boosting Classifier
5. XG Boost Classifier
6. Support Vector Classifier
7. Gaussian Naive Bayes Classifier
8. Bernoulli Naive Bayes Classifier

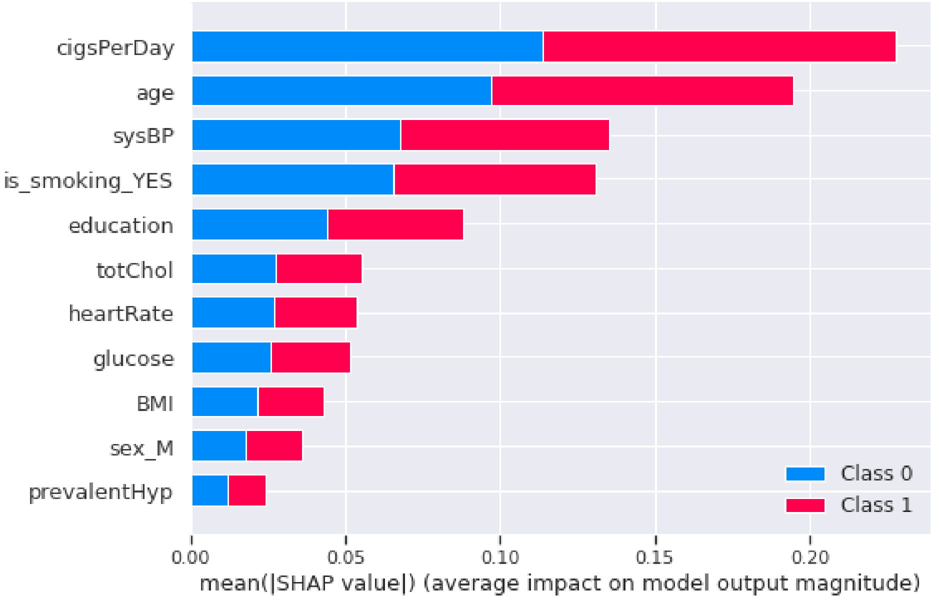


RANDOM FOREST CLASSIFIER

Random forest consists of a large number of individual decision trees that operate as an ensemble. Each individual tree in the random forest spits out a class prediction and the class with the most votes becomes our model's prediction.

Hyper-parameters used for tuning:

- n_estimators: Number of trees in the forest
- max_depth: Maximum number of levels in each decision tree
- min_samples_split: Minimum number of data points placed in a node before the node is split
- min_samples_leaf: Minimum number of data points allowed in a leaf node
- bootstrap: Method for sampling data points (with or without replacement)

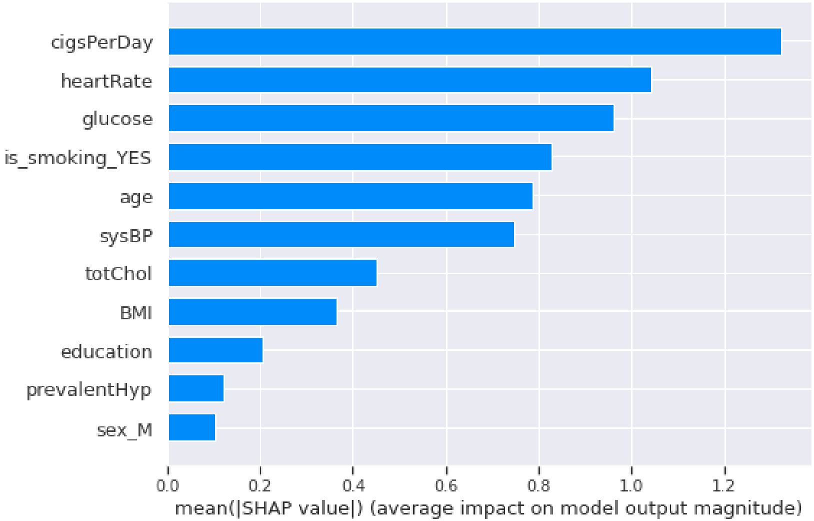


GRADIENT BOOSTING CLASSIFIER

In Gradient Boosting, each predictor tries to improve on its predecessor by reducing the errors. But the fascinating idea behind Gradient Boosting is that instead of fitting a predictor on the data at each iteration, it actually fits a new predictor to the residual errors made by the previous predictor.

Hyper-parameters used for tuning:

- n_estimators: Number of trees in the forest
- max_depth: Maximum number of levels in each decision tree
- min_samples_split: Minimum number of data points placed in a node before the node is split
- min_samples_leaf: Minimum number of data points allowed in a leaf node
- max_features: The number of features to consider while searching for a best split
- max_leaf_nodes: The maximum number of terminal nodes or leaves in a tree.
- learning_rate: The learning parameter controls the magnitude of this change in the estimates.

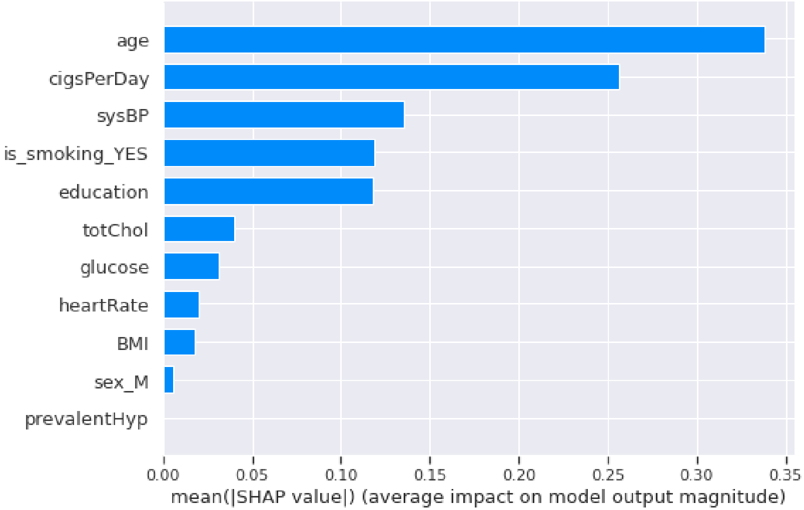


XG BOOST CLASSIFIER

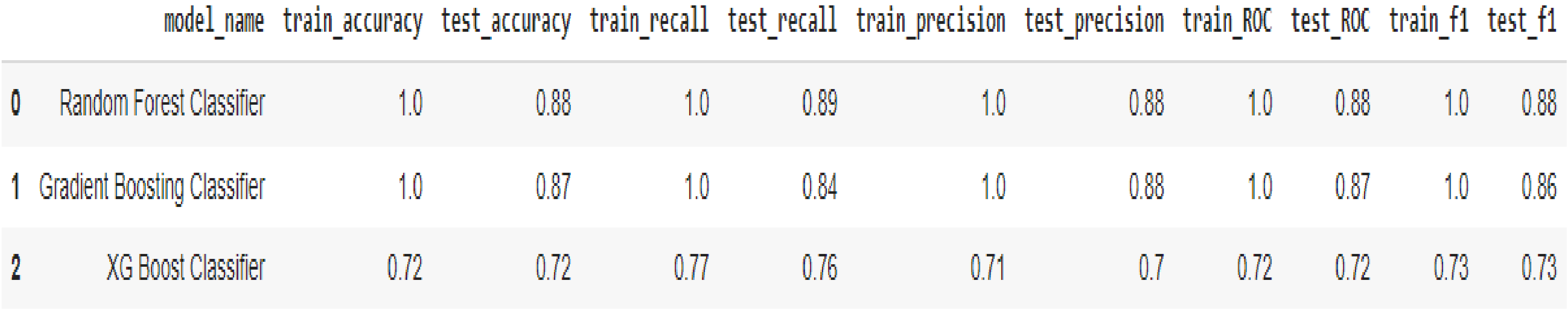
Decision trees are created in sequential form. Weights play an important role in XGBoost. Weights are assigned to all the independent variables which are then fed into the decision tree which predicts results. The weight of variables predicted wrong by the tree is increased and these variables are then fed to the second decision tree. These individual classifiers/predictors then ensemble to give a strong and more precise model.

Hyper-parameters used for tuning:

- n_estimators: Number of trees in the forest
- max_depth: Maximum number of levels in each decision tree
- gamma: It specifies the minimum loss reduction required to make a split.
- max_delta_step: It allow each tree's weight estimation to be considered
- min_child_weight: Defines the minimum sum of weights of all observations required in a child.
- reg_alpha: L1 regularization term on weights
- reg_lambda: L2 regularization term on weights
- learning_rate: The learning parameter controls the magnitude of this change in the estimates.



FINAL RESULT TABLE



	model_name	train_accuracy	test_accuracy	train_recall	test_recall	train_precision	test_precision	train_ROC	test_ROC	train_f1	test_f1
0	Random Forest Classifier	1.0	0.88	1.0	0.89	1.0	0.88	1.0	0.88	1.0	0.88
1	Gradient Boosting Classifier	1.0	0.87	1.0	0.84	1.0	0.88	1.0	0.87	1.0	0.86
2	XG Boost Classifier	0.72	0.72	0.77	0.76	0.71	0.7	0.72	0.72	0.73	0.73