Capstone Project

Cardio Vascular Risk Prediction

Aim of Project

- In this project we were supposed to make a classifier models where 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, behavioural, and medical risk factors.

Methodology

- 1. Data Cleaning (Outlier detection, Checking Null Values)
- 2. Exploratory Data Analysis
- 3. Data Processing (Scaling and Feature Selection)
- 4. Data Splitting
- 5. Model Training - Used various Models
- 6. Evaluation Metrics

Data Cleaning

IQR

id	1694.50
age	14.00
education	2.00
cigsPerDay	20.00
BPMeds	0.00
prevalentStroke	0.00
prevalentHyp	1.00
diabetes	0.00
totChol	58.00
sysBP	27.00
diaBP	15.50
BMI	5.02
heartRate	15.00
glucose	16.00
TenYearCHD	0.00

Null Values before cleaning

id	0
age	415
education	87
sex	0
is_smoking	0
cigsPerDay	433
BPMeds	44
prevalentStroke	0
prevalentHyp	0
diabetes	0
totChol	448
sysBP	415
diaBP	415
BMI	426
heartRate	416
glucose	693
TenYearCHD	0
sex_F	0
sex_M	0
is_smoking_NO	0
is_smoking_YES	0

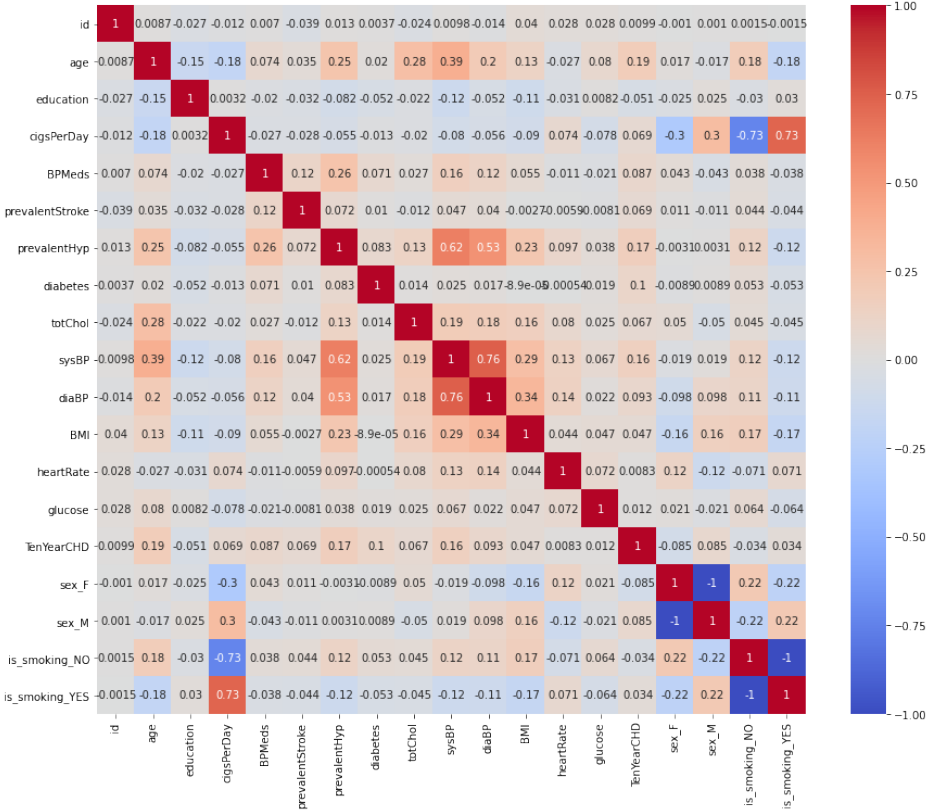
Data Cleaning

```
id                0
age               0
education         0
sex               0
is_smoking        0
cigsPerDay        0
BPMeds            0
prevalentStroke   0
prevalentHyp      0
diabetes          0
totChol           0
sysBP             0
diaBP             0
BMI               0
heartRate         0
glucose           0
TenYearCHD        0
sex_F             0
sex_M             0
is_smoking_NO     0
is_smoking_YES    0
```

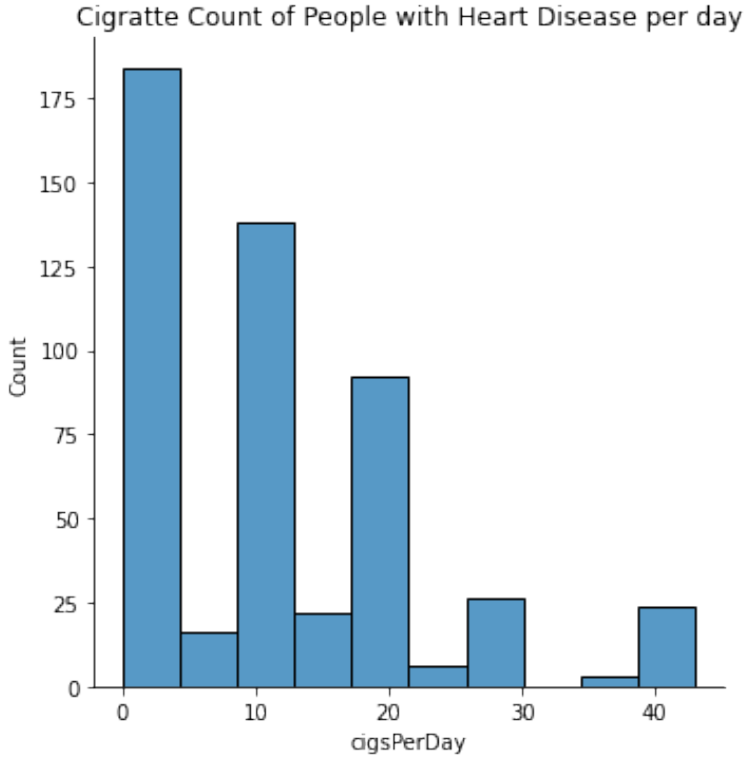
- Missing values can be either replaced or that respective column is dropped.
- Dropping of column only takes place when null values are more than 50%.
- Null values can be replaced with mean, median and mode.
- Algorithms like KNN are sometimes used to fill null values.

Exploratory Data Analysis

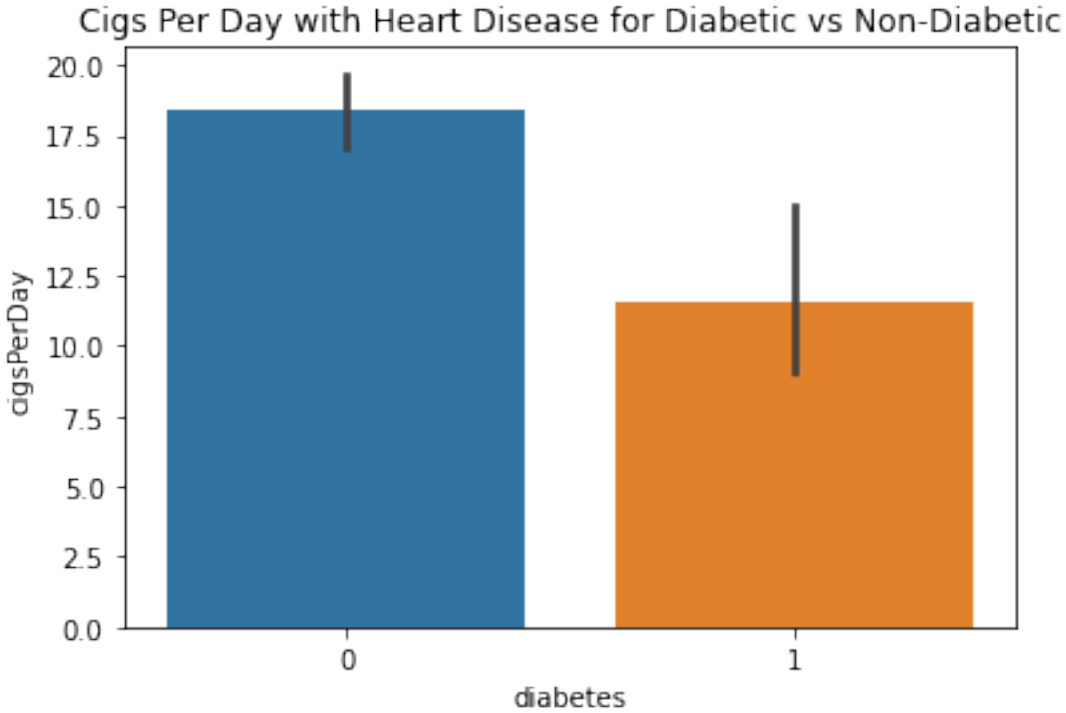
Correlation



Cigarette count of people suffering from Heart Disease.

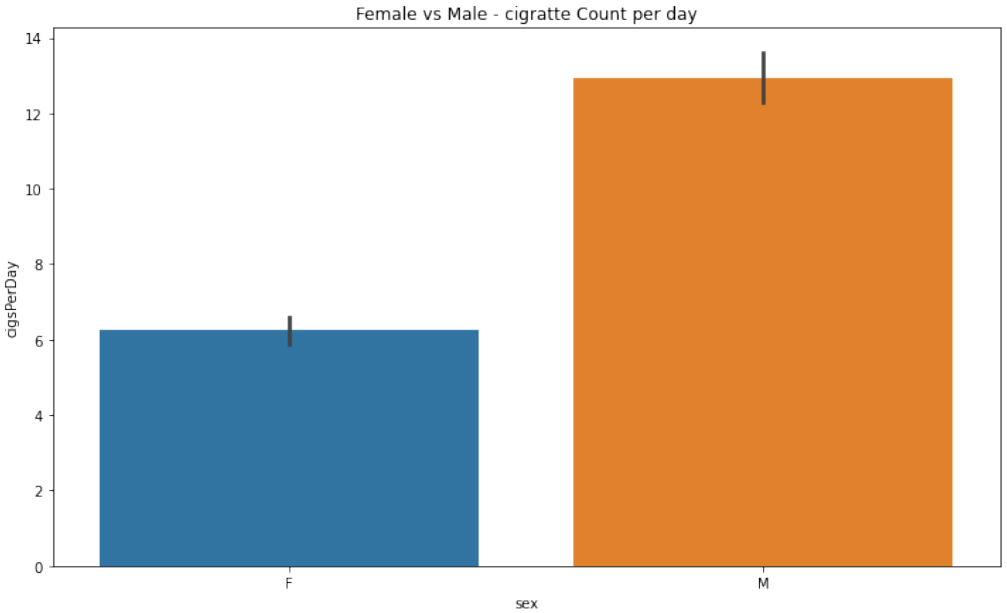


Diabetic vs Non-Diabetic



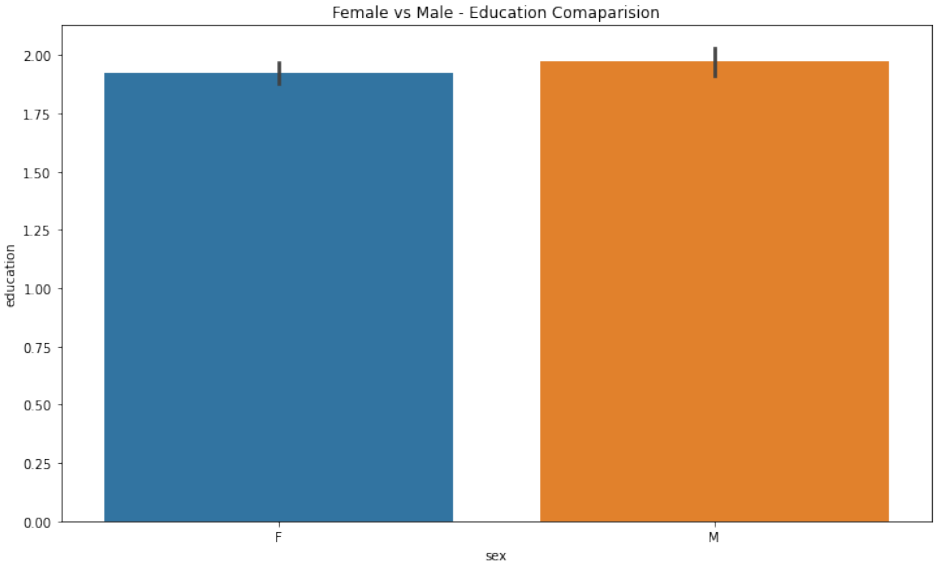
- Those who are suffering from heart disease include more non-diabetic people.

Gender Comparison of Patients



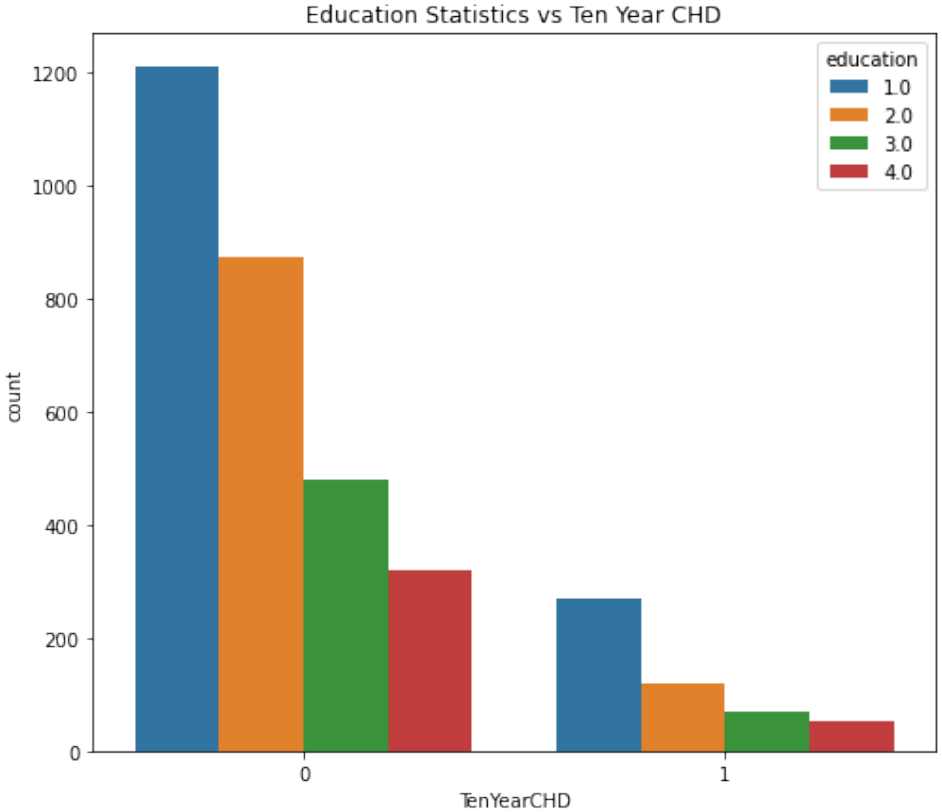
- Males suffer more than female from heart diseases.

Average Education Comparison.

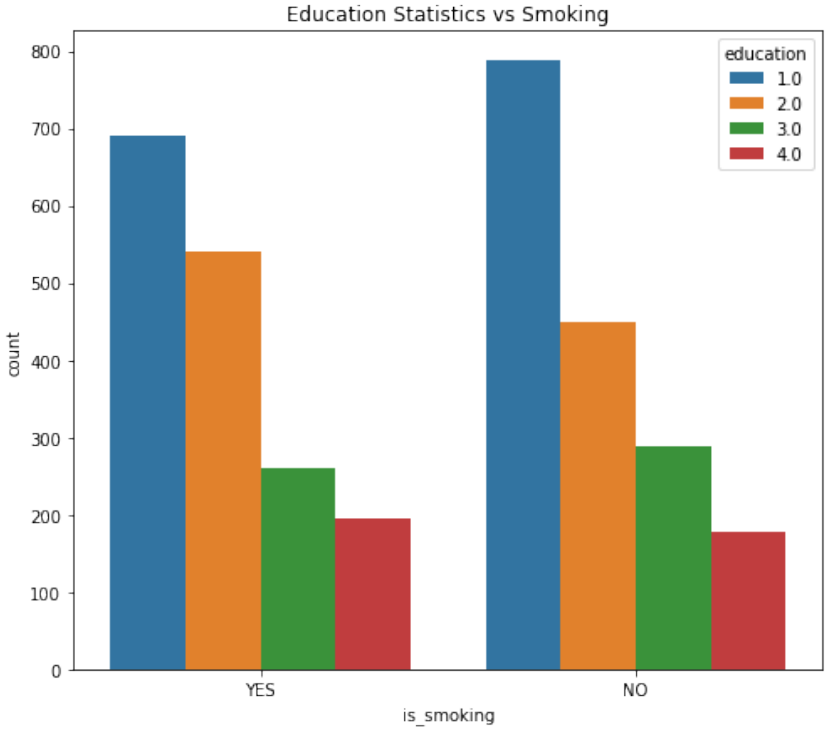


- On comparing all the people who took part in survey, it is seen that males are more educated than females (on average).

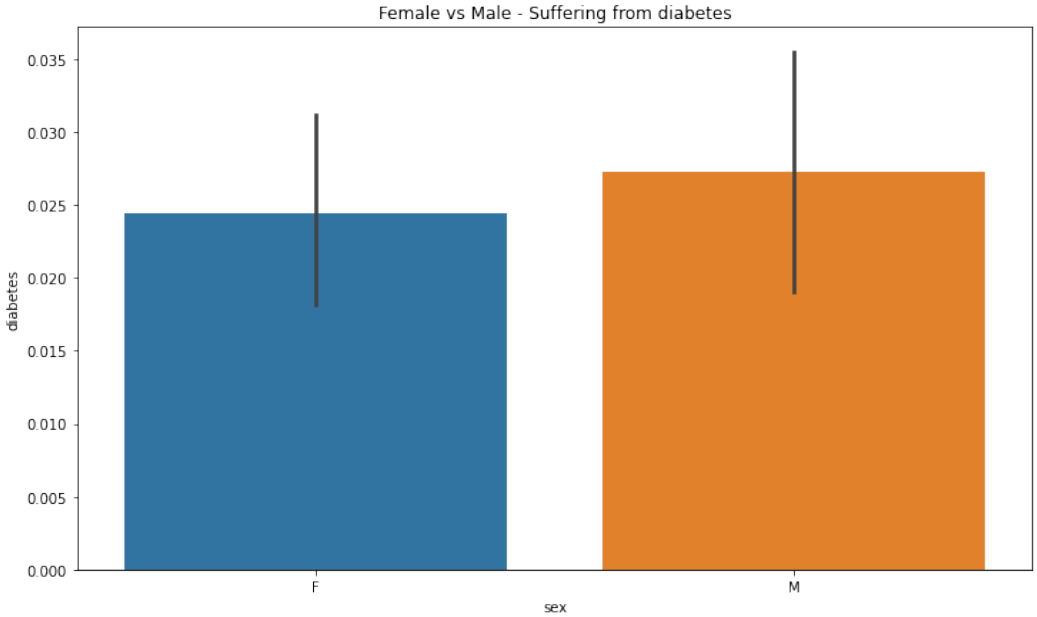
Education Statistic vs TenYearCHD



Education Statistics vs Smoking

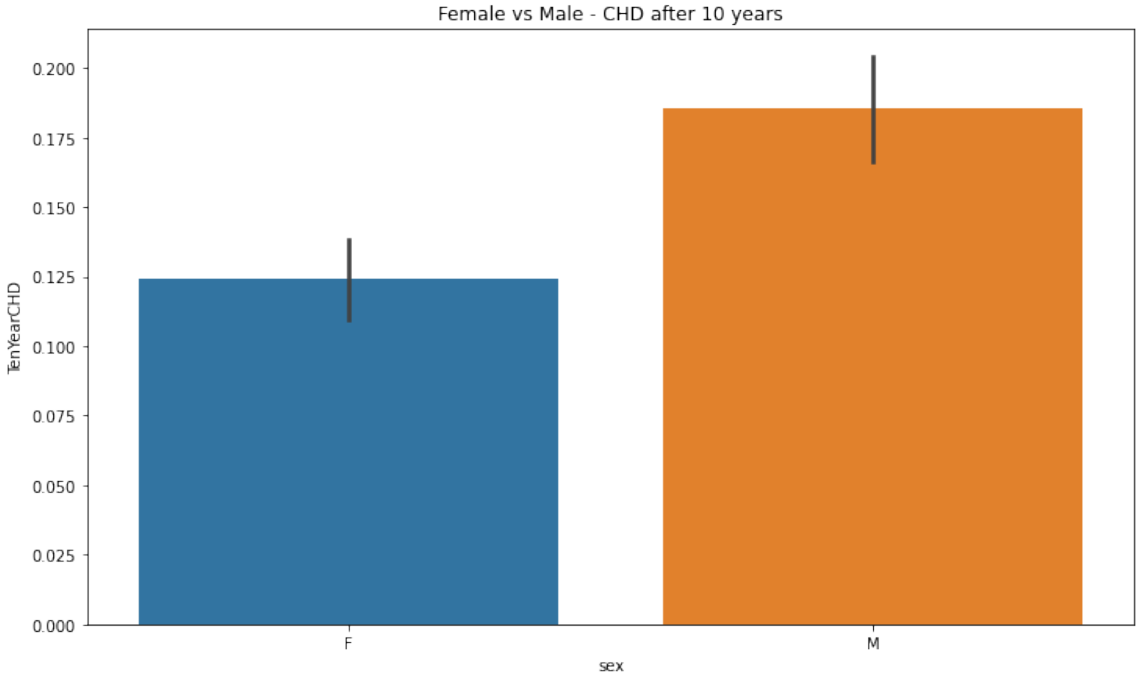


Diabetes Comparison

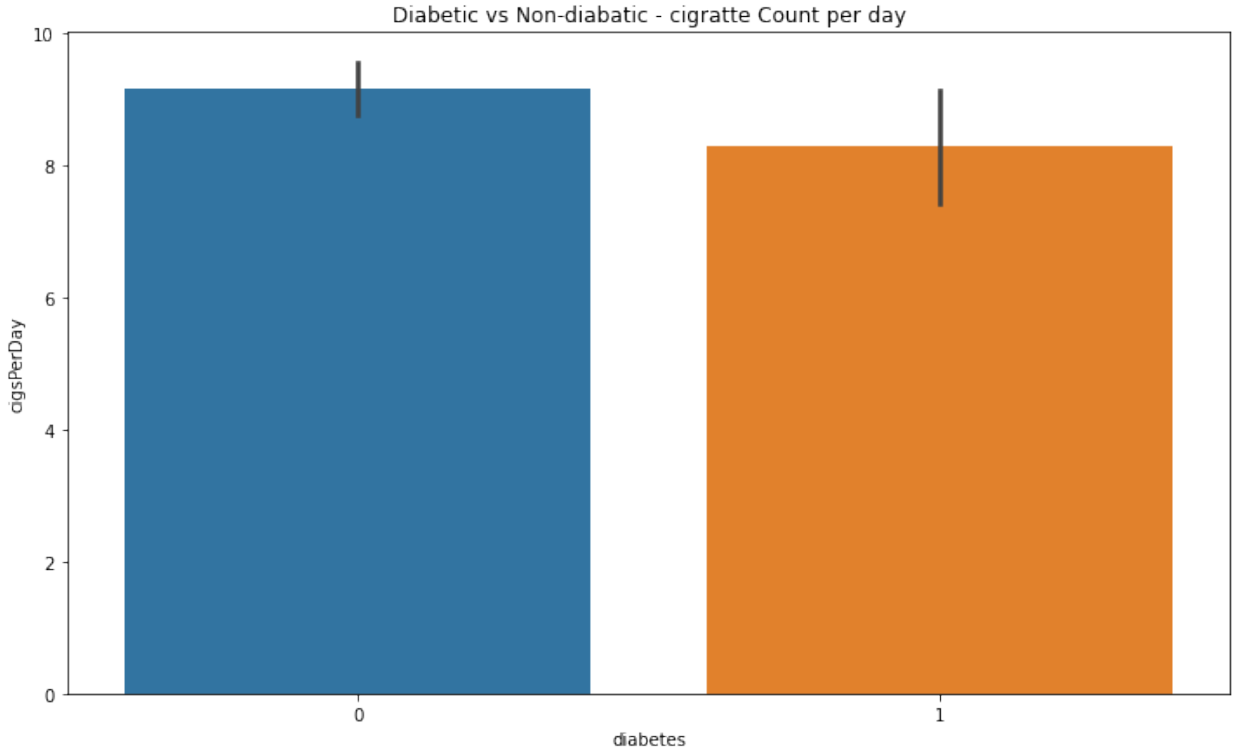


- The Dataset has more males suffering from diabetes than female.

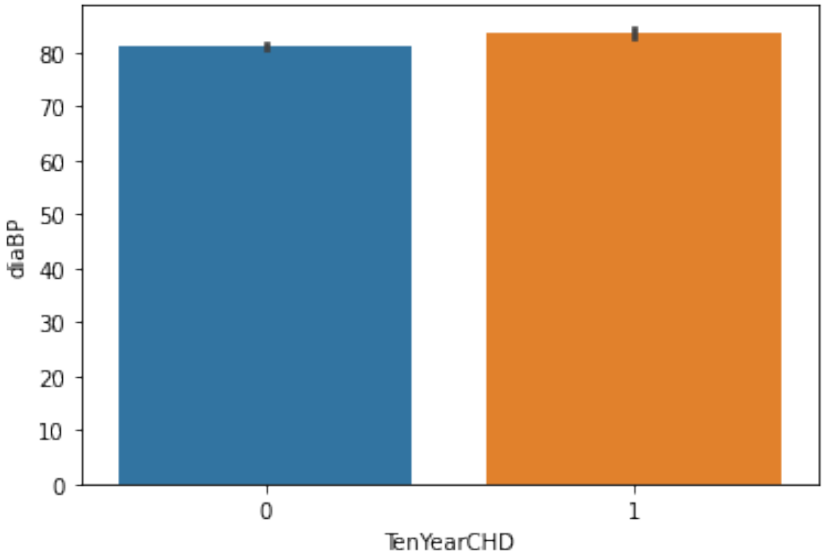
TenYearCHD Comparison.

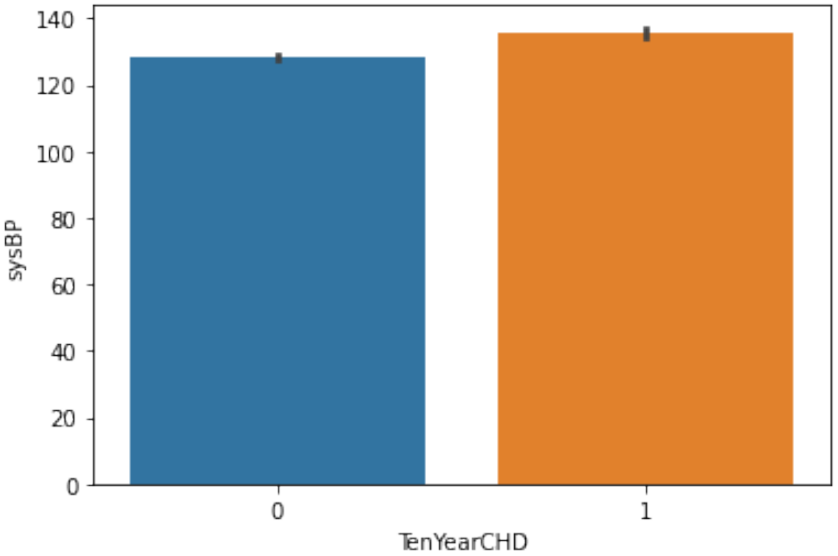


Cigs Per Day Count for Diabetic vs Non-Diabetic

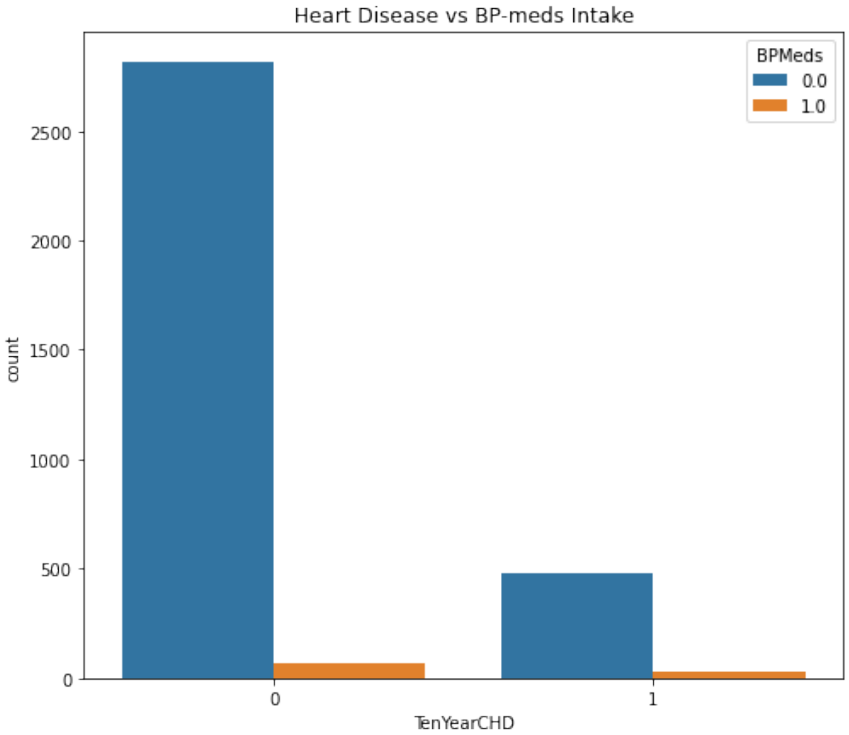


TenYearCHD vs Blood Pressure

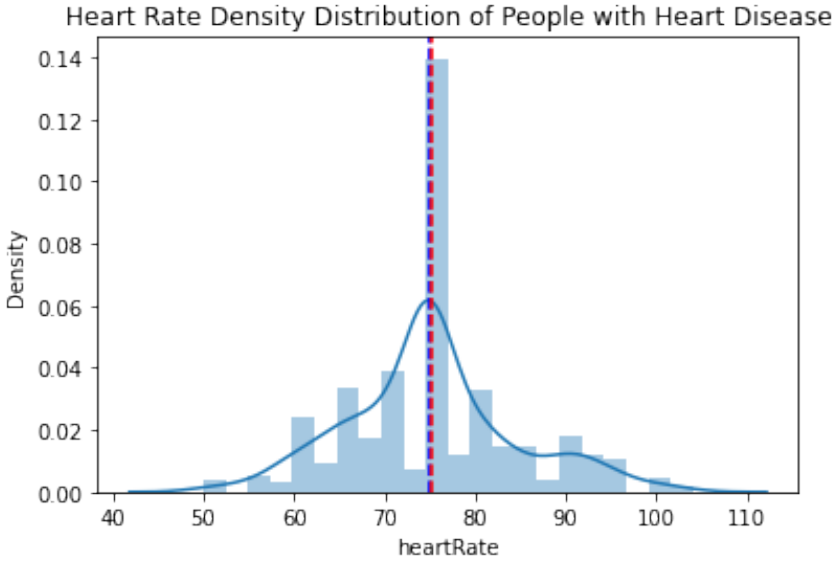


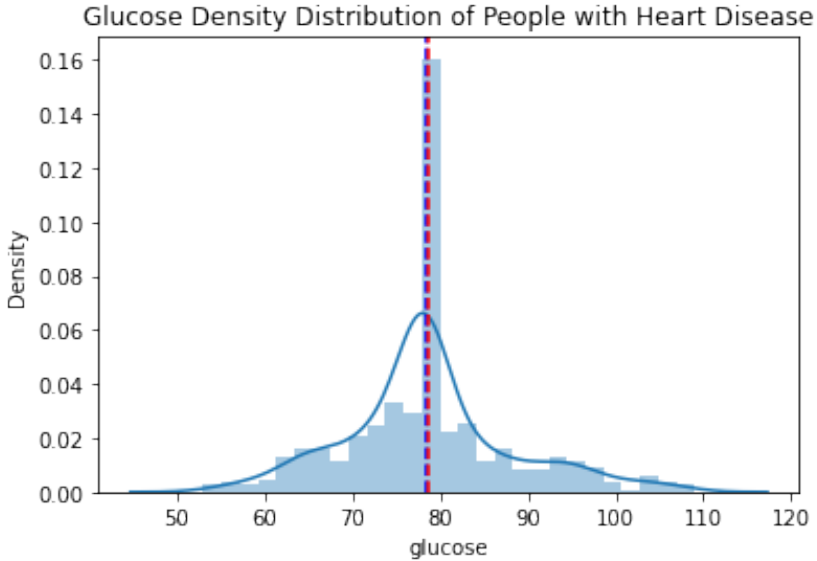


Effect of BP Medicines.

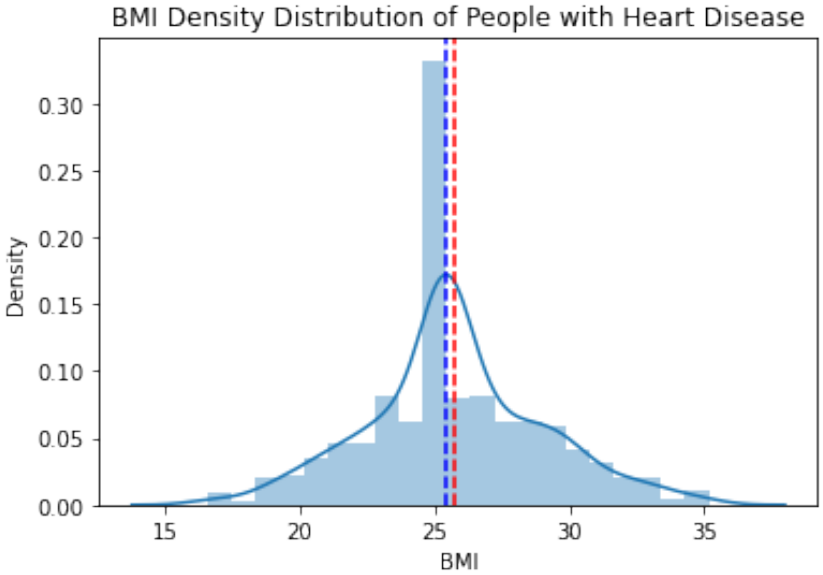


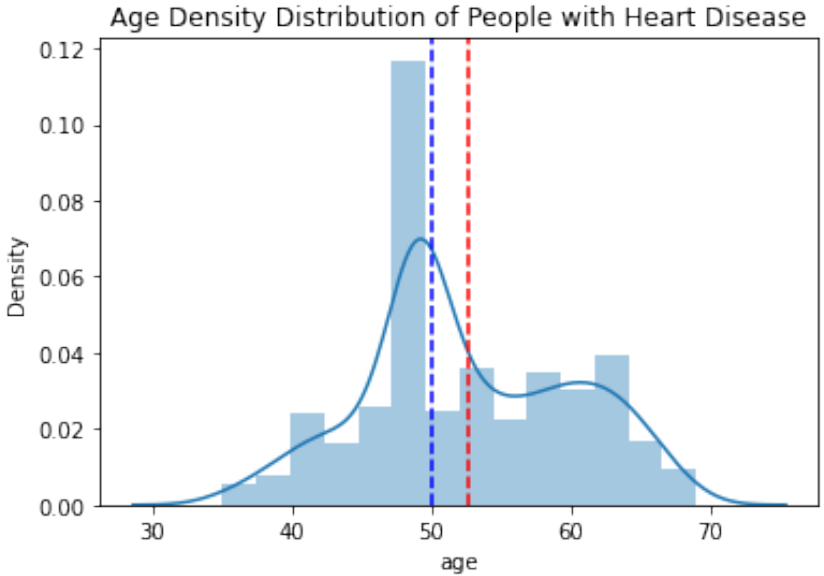
Numerical Parameters Density Distribution





Continued.


BMI Density Distribution of People with Heart Disease
Density
0.00
0.05
0.10
0.15
0.20
0.25
0.30
15
20
25
30
35
BMI


Age Density Distribution of People with Heart Disease
Density
0.00
0.02
0.04
0.06
0.08
0.10
0.12
30
40
50
60
70
age

Data Preprocessing

Scaling

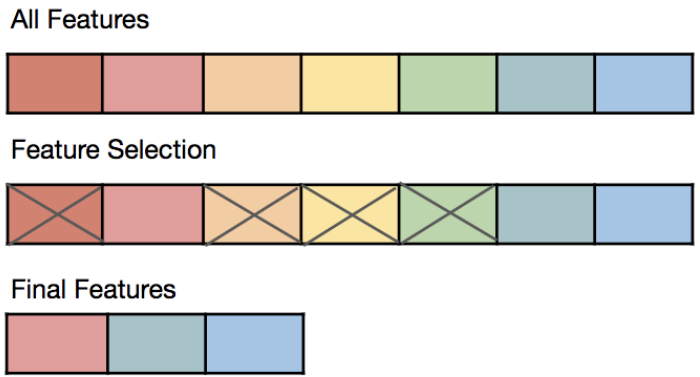
- Feature scaling in machine learning is one of the most critical steps during the pre-processing of data before creating a machine learning model. Scaling can make a difference between a weak machine learning model and a better one.
- The most common techniques of feature scaling are Normalization and Standardization.
- Normalization is used when we want to bound our values between two numbers, typically, between [0,1] or [-1,1]. While Standardization transforms the data to have zero mean and a variance of 1, they make our data **unitless**. Refer to the below diagram, which shows how data looks after scaling in the X-Y plane

Standard Scaling

- The Standard Scaler assumes data is normally distributed within each feature and scales them such that the distribution centered around 0, with a standard deviation of 1.
- Centering and scaling happen independently on each feature by computing the relevant statistics on the samples in the training set. If data is not normally distributed, this is not the best Scaler to use.

Feature Selection Using RFE

- Recursive feature elimination (RFE) is a feature selection method that fits a model and removes the weakest feature (or features) until the specified number of features is reached.
- Features are ranked by the model's coef_ or feature_importances_ attributes, and by recursively eliminating a small number of features per loop, RFE attempts to eliminate dependencies and collinearity that may exist in the model.

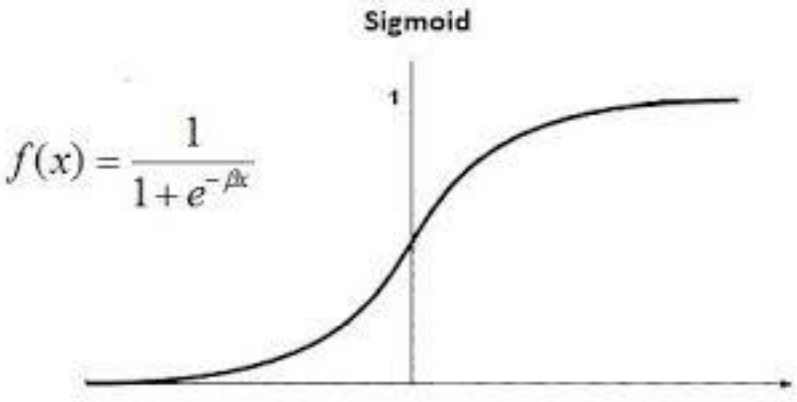


Data Splitting

- Education, CigsPerDay, ‘is_smoking_NO' , ‘sex_M', ‘prevalentHyp', ‘BPMeds', 'diabetes' are the features which were selected and data is distributed with 80% train and 20% test data.

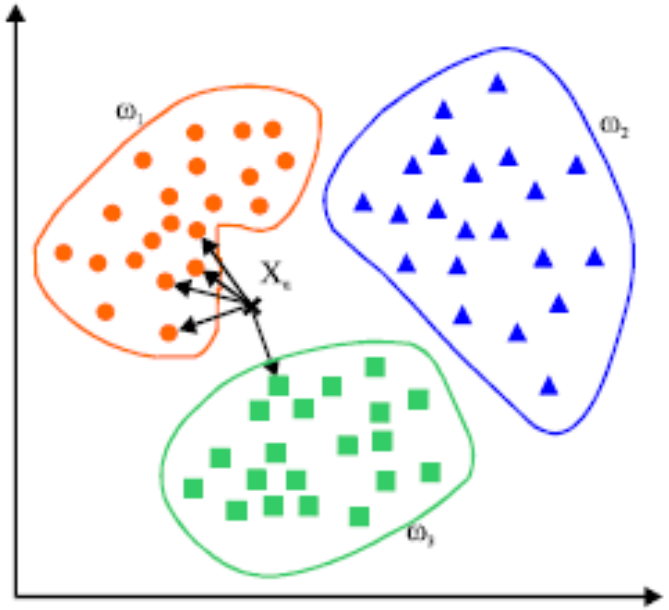
Model - Training : Logistic Regression

- Logistic regression is one of the most popular Machine Learning algorithms, which comes under the Supervised Learning technique.
- It is used for predicting the categorical dependent variable using a given set of independent variables. It predicts the output of a categorical dependent variable.



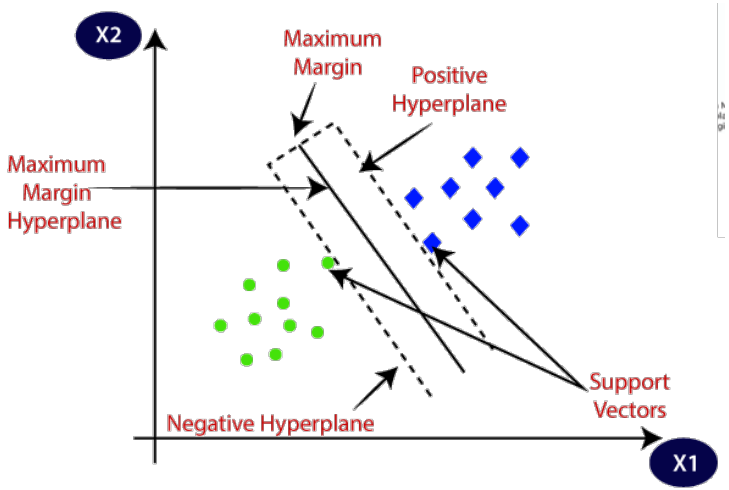
KNN Classifier

- K-Nearest Neighbour is one of the simplest Machine Learning algorithms based on Supervised Learning technique.
- It assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories.



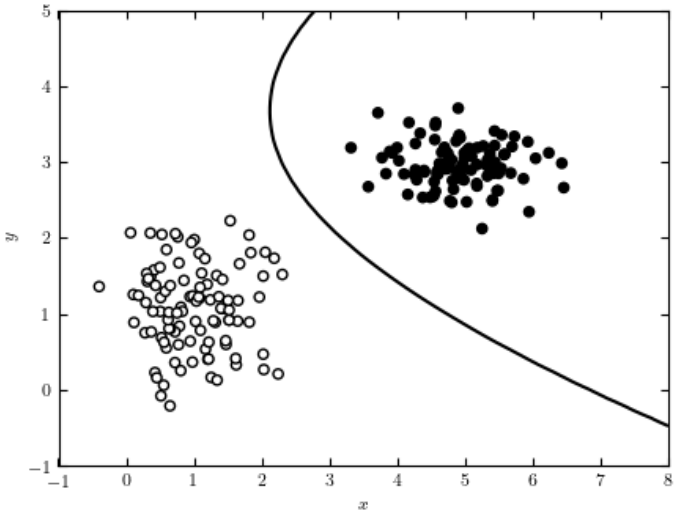
Support Vector Machine

- Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. However, primarily, it is used for Classification problems in Machine Learning. The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called hyperplane. It chooses the extreme points/vectors that help in creating the hyperplane.
- These extreme cases are called as support vectors, and hence algorithm is termed as Support Vector Machine. Consider the above diagram in which there are two different
- categories that are classified using a decision boundary or hyperplane.



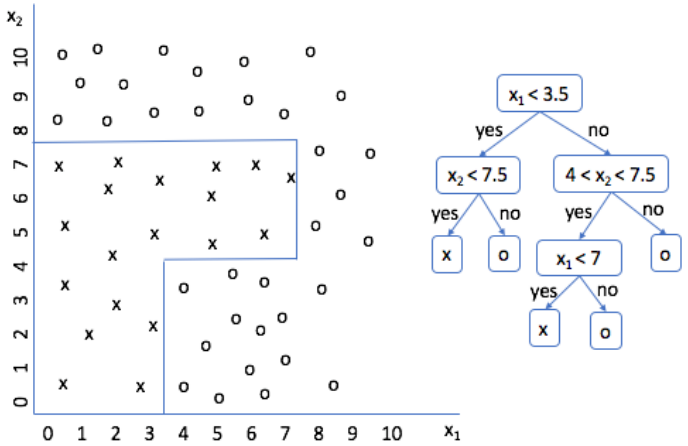
Naive Bayes Classifier

- A Naive Bayes classifier is a probabilistic machine learning model that's used for classification task. The crux of the classifier is based on the Bayes theorem.
- Using Bayes theorem, we can find the probability of A happening, given that B has occurred. Here, B is the evidence and A is the hypothesis.
- The assumption made here is that the predictors/ features are independent. That is presence of one particular feature does not affect the other. Hence it is called naive.



Decision Tree Classifier

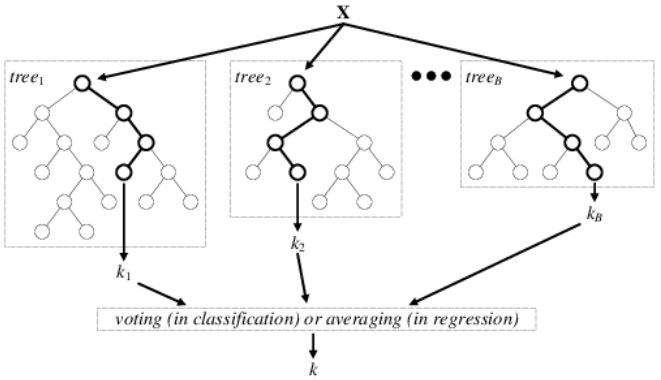
- Decision Tree Classifier is a simple and widely used classification technique. It applies a straight forward idea to solve the classification problem. Decision Tree Classifier poses a series of carefully crafted questions about the attributes of the test record.
- Each time time it receive an answer, a follow-up question is asked until a conclusion about the class label of the record is reached. The decision tree classifiers organized a series of test questions and conditions in a tree structure.



Random Forest Classifier

- Random forest classifiers fall under the broad umbrella of ensemble-based learning methods. They are simple to implement, fast in operation, and have proven to be extremely successful in a variety of domains.
- The key principle underlying the random forest approach comprises the construction of many “simple” decision trees in the training stage and the majority vote (mode) across them in the classification stage.

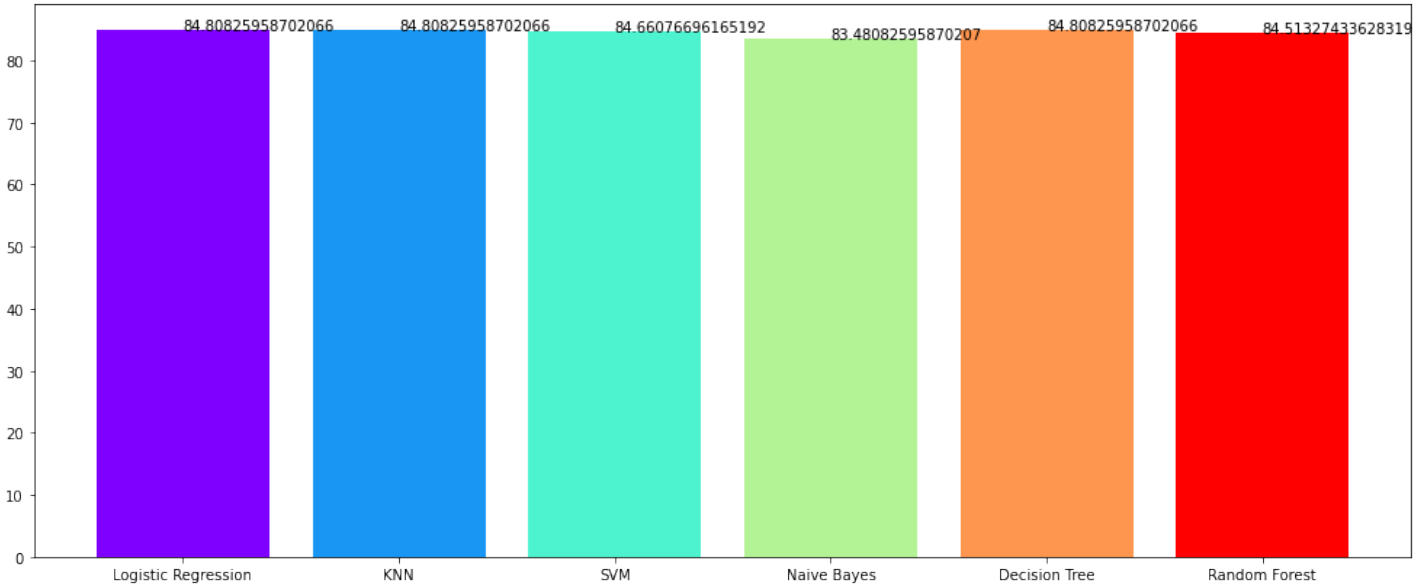
- Among other benefits, this voting strategy has the effect of correcting for the undesirable property of decision trees to overfit training data.



K-fold Cross Validation

- Cross-Validation is just a method that simply reserves a part of data from the dataset and uses it for testing the model(Validation set), and the remaining data other than the reserved one is used to train the model.
- In k-fold CV, the dataset is split into ‘k’ number of subsets, k-1 subsets then are used to train the model and the last subset is kept as a validation set to test the model.
- Then the score of the model on each fold is averaged to evaluate the performance of the model.

Model Accuracies



Conclusions

- Logistic Regression got better result than any model.
- Highest Number of cigarette smoked in a day is 50.
- Males consume more cigarettes than females in a day.
- People with less education are more prone to have heart disease after 10
- years.
- People with less education are more prone to get addicted to smoking.
- More males are suffering from diabetes than female.
- Those who have high BP are more prone to heart disease.
- Those who have low BP are less prone to heart disease.
- Non- diabetic people smokes more

Measure that can be taken

1. No smoking
2. Maintain Healthy daily life.
3. BMI should be checked regularly inorder to have a note of ourselves.
4. As we get older, selection of food must be done properly, so that one can control cholesterol, glucose etc.
5. Having proper cardio routine should be must. Yoga, walking or jogging are good enough.