Capstone Project

Cardio_Vascular_Risk_Prediction

Individual Project

Shambhuraj Desai

Contents

- Problem Statements
- Data Summary
- Data Cleaning & Manipulation
- Exploratory Data Analysis
- Feature Selection
- Feature Transformation
- Balancing Dataset
- Scaling Dataset
- Logistic Regression (classification)
- k-NN
- k-NN Hyperparameter tunning
- SVM
- Conclusion

Problem Statement

The dataset is from an outgoing cardiovascular study on residents of the town of Framingham, Massachusetts. The classification goal is to predict whether a patient has 10 year risk of future coronary heart disease (CHD). The dataset provides the patients information. It includes over 4000 records and 15 attributes. Each attribute is potential risk factor. There are both demographic and behavioral and medical risk factor.

By using the above data information and given dataset we need to predict the Cardio Vascular Risk among the patients.

Data Summary

The collected data had 4000 records/observations and 17 columns/features. For this project we will be analysing Cardio Vascular Risk among the patients. This dataset contains information about the patients and daily habits like patients age, education, their gender, whether he/she smoking the cigarettes or not, how many cigarettes per day, he/she taking BP medicines or not, whether affected by prevalent stroke or not, prevalent hypertension, diabetic patient or not, total cholesterol, systolic BP, diastolic BP, BMI, Heart Rate, glucose, Ten Year Coronary Heart Disease report.

Data Description

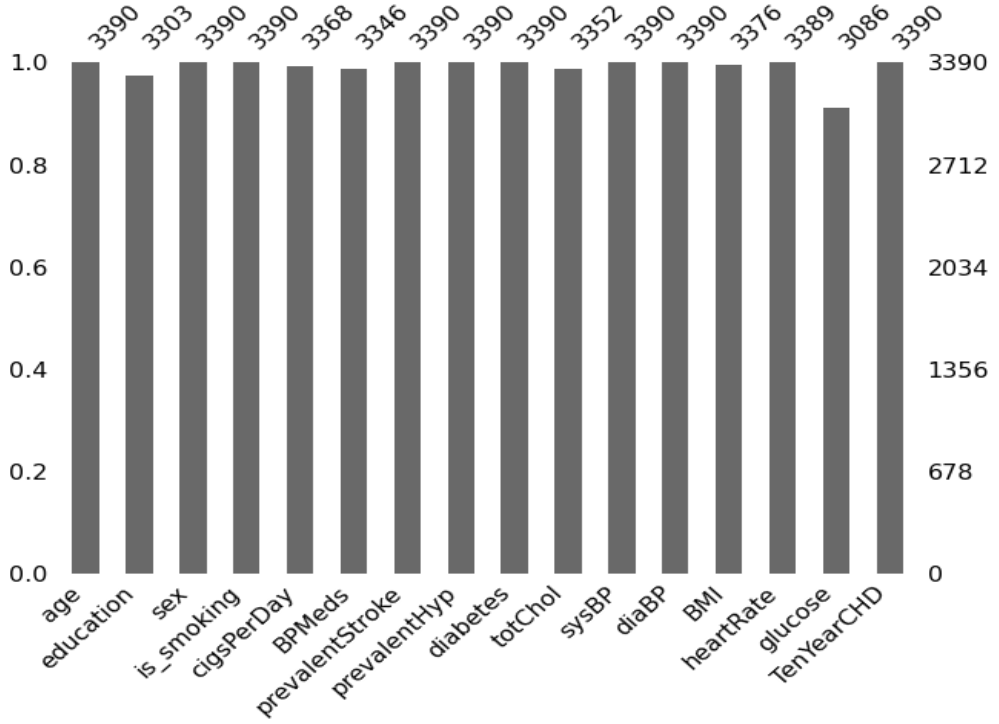
- **Sex** – Gender of patient
- **Age** – Age of patient
- i**s_smoking** – Whether patient smoking or not
- **cigs_per_day** – how many cigarettes smoking
- **BP_meds** – taking Blood Pressure medicines or not
- **Prevalent_stroke** – affected by prevalent stroke or not
- **Diabetes** – Patient has diabetes or not
- **TotChol** – Total cholesterol of patient
- **Sys_BP** – Systolic Blood Pressure
- **Dia_BP** – Diastolic Blood Pressure
- **BMI** – Body mass index
- **Heart_Rate** – heart rate measure

Data Cleaning

Missing values:

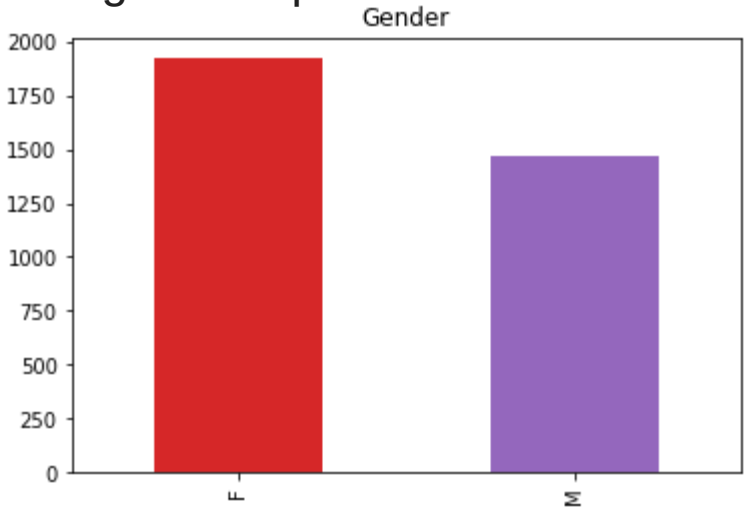
In our data of 4000 rows and 17 columns and we have missing data in 7 columns.

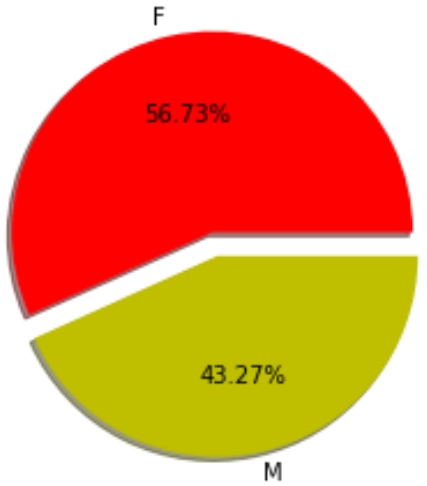
- Education have 87 missing values
- cigs_per_Day have 22 missing values
- BP meds have 44 missing values
- totChol have 38 missing values
- BMI have 14 missing values
- Heart rate have 1 missing value
- Glucose have 304 missing values



Data Wrangling

Data Cleaning & Manipulation





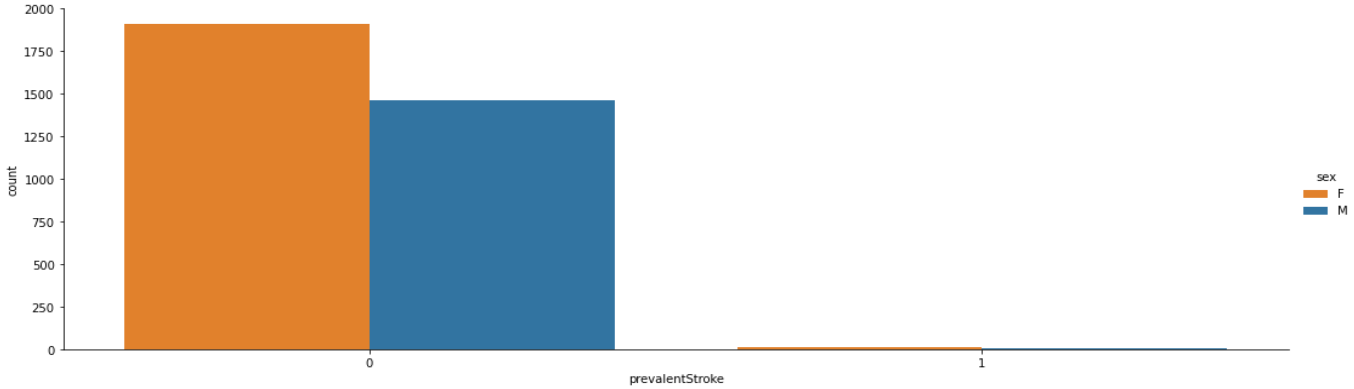
Our dataset contains more female patients than male patients and here we can conclude that Females facing more heart related issues in their daily life.
As number of males are also facing lots of health problems.

Interpretation:

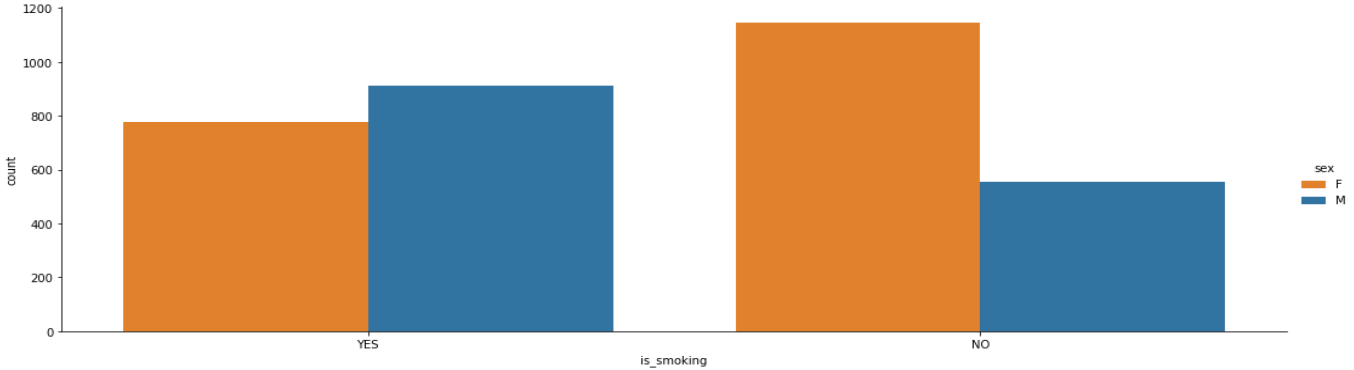
- Female patients are greater than male patients.
- Females 56.73% and Males 43.27%

Bivariate Analysis

Prevalent stroke i.e., the stroke occurs when supply of blood to part of brain is interrupted of reduced.



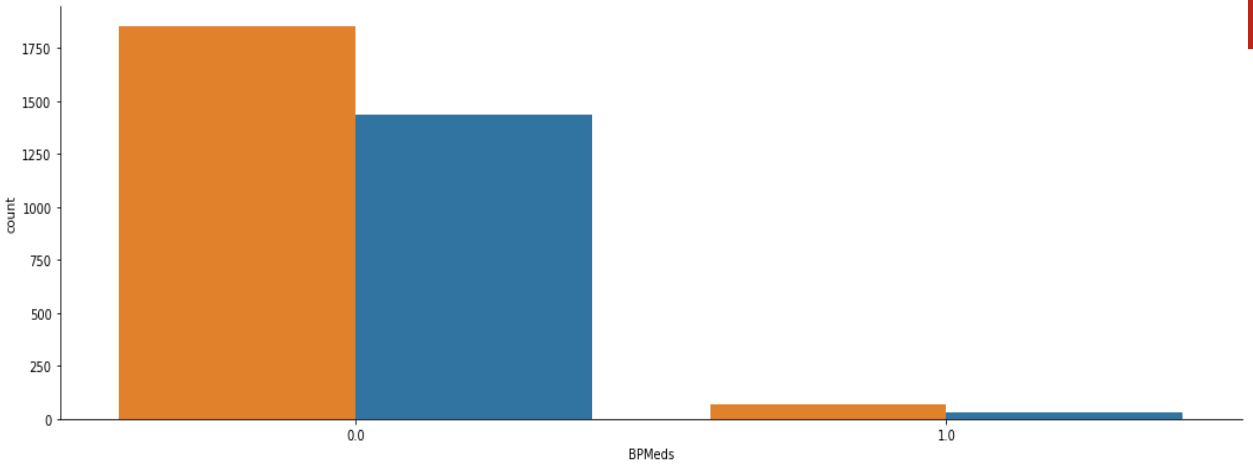
The feature ‘is_smoking’ gives us information whether the patient is smoking cigarettes or not.



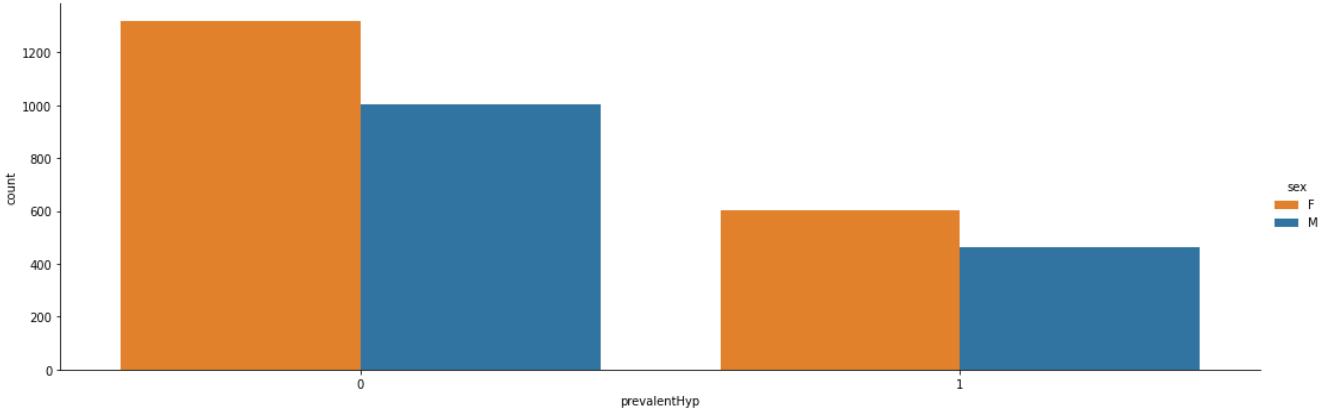
Interpretation

- Prevalent stroke in both male and females are seen very rarely
- Males count is high in smoking habit

The feature ‘BPMeds’ tells us whether the patient is consuming any blood pressure medicines or not. And we can see more number of Females eating more BPMeds than males.



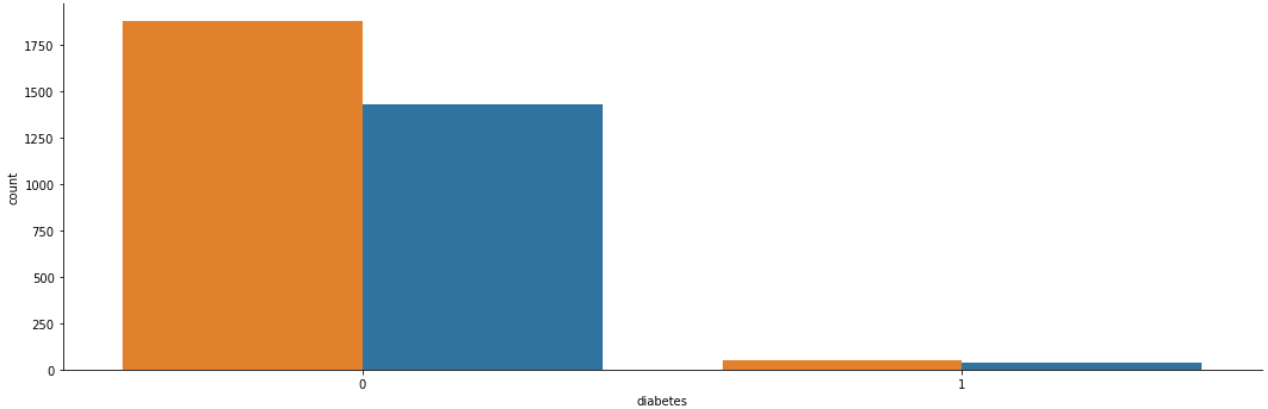
Prevalent hypertension means the blood pressure is higher than the normal blood pressure, and from our dataset female patients are affected by the prevalent hypertension than males.



Interpretation:

- The number of Females consuming the BP medicines is slightly higher than males.
- Females have high number of prevalent hyper tension than males.

Diabetes means increasing the sugar level in blood more than normal range, from the dataset we observe that fraction of female patients have more diabetes than male patients.



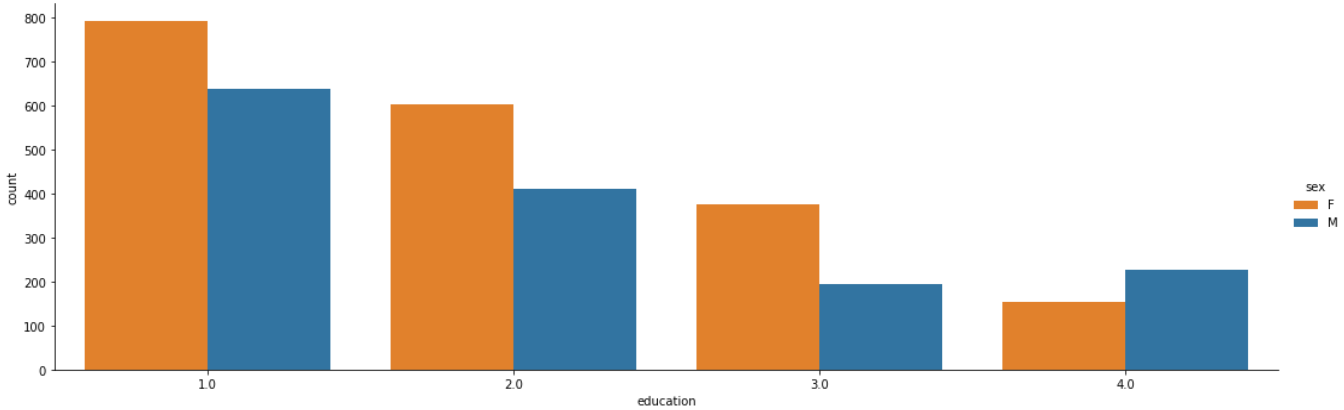
Education is matters in the health care.

1 – Primarily educated

2 – Secondarily educated

3 – Higher secondary

4 – Well educated



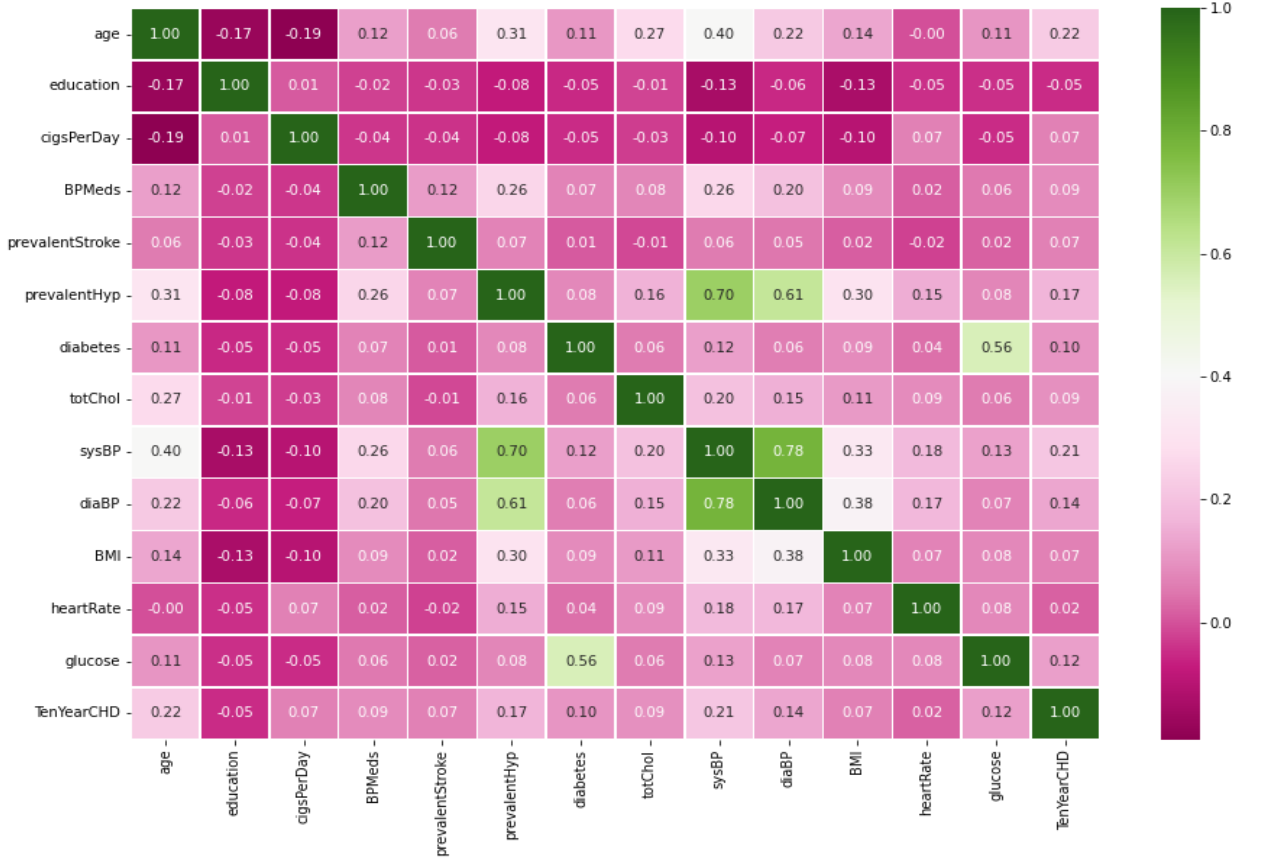
Interpretation:

- Very rarely seen the diabetes in the dataset, and Females have the greater in number
- In higher qualification males are greater in number and in other qualifications females have higher number

Exploratory Data Analysis

Correlation Heatmap-

This is correlation heatmap is correlation matrix of all the features we have in our dataset.
It tells us about the correlation between the two variables.
Higher the correlation ‘greenish’ is color and lower the correlation ‘pinkish’ it becomes

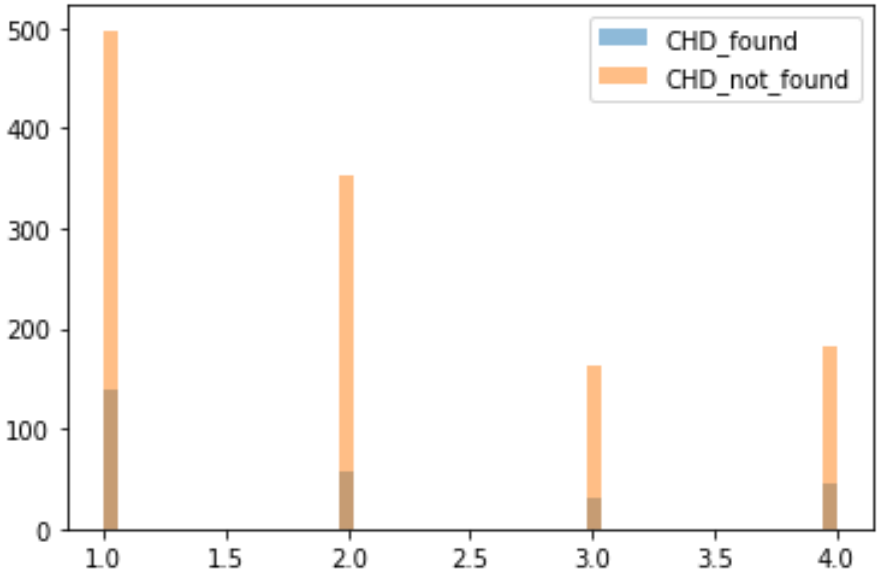


Interpretation:

- Maximum correlation between systolic BP and diastolic BP have seen.
- sysBP and diaBP have lowest correlation with target variable TenYearCHD.

Hist plot – Education (Male)

The subdivided histogram is for the male patients which are distributed by Education the patients had and the bars are divided by whether the male patients are affected by the ‘Coronary Heart Disease’ or not.



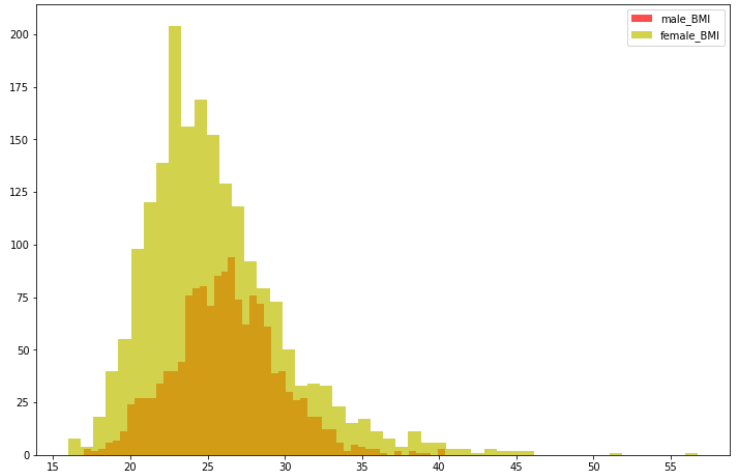
Interpretation :

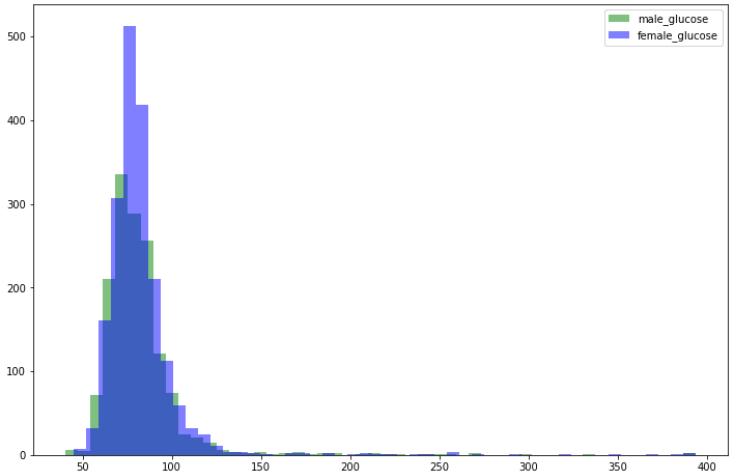
- Coronary Heart Disease is seen in the patient’s which are primarily educated.
- The well educated patient’s seen in less number of CHD disease.
- We can conclude that, education matters with respect to health management.

Data Analysis

Comparison of Male & Female patients

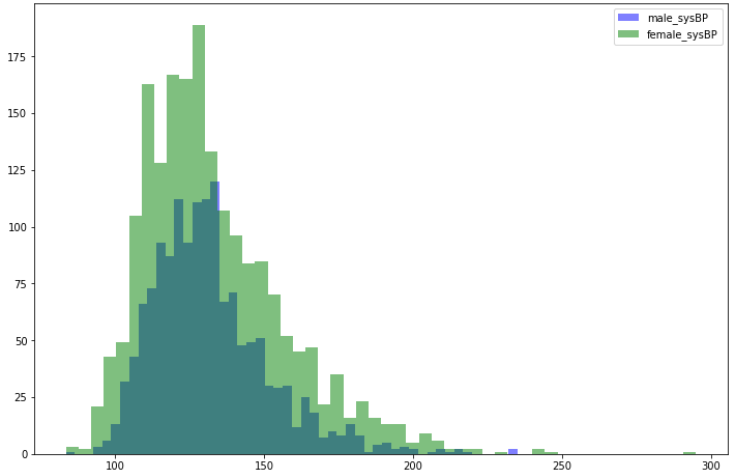
The above graphs tells us about the distribution of given dataset for male and females over various features.

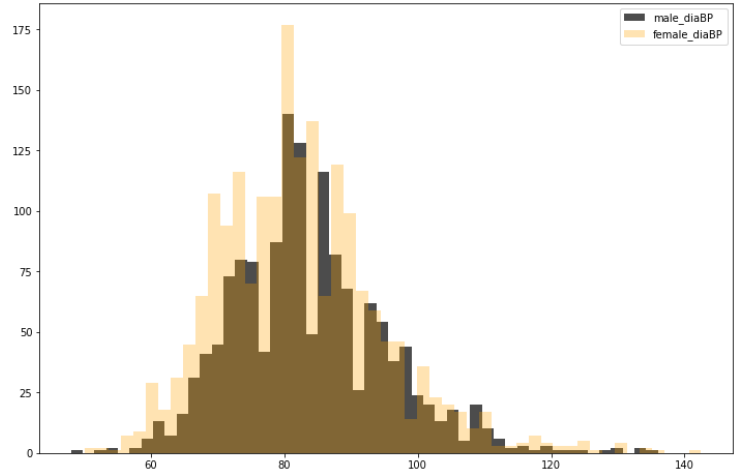




- Male ‘BMI’ is represents a normal distribution than female ‘BMI’ and we can see female ‘BMI’ data is positively skewed.
- Also female category data have higher values than the male category.

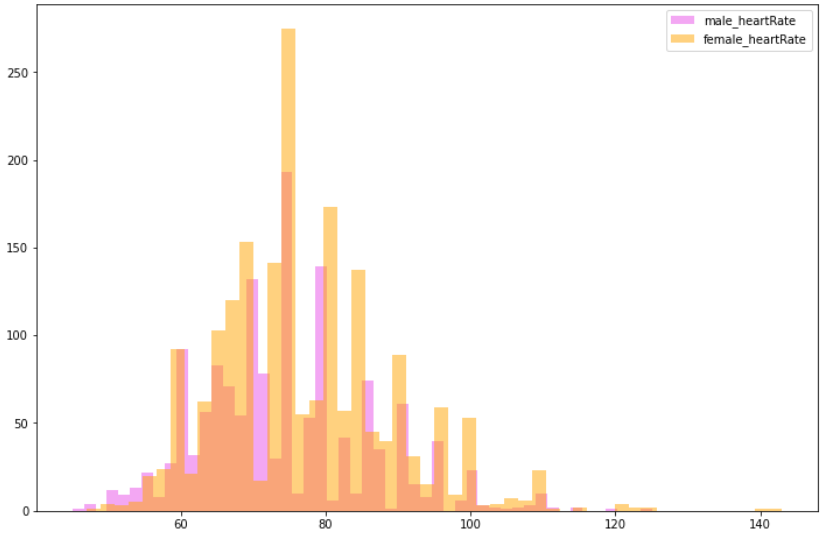
- For feature ‘glucose’ it shows us for both male and female category highly positive skewness.
- In this feature also females have higher values than males.

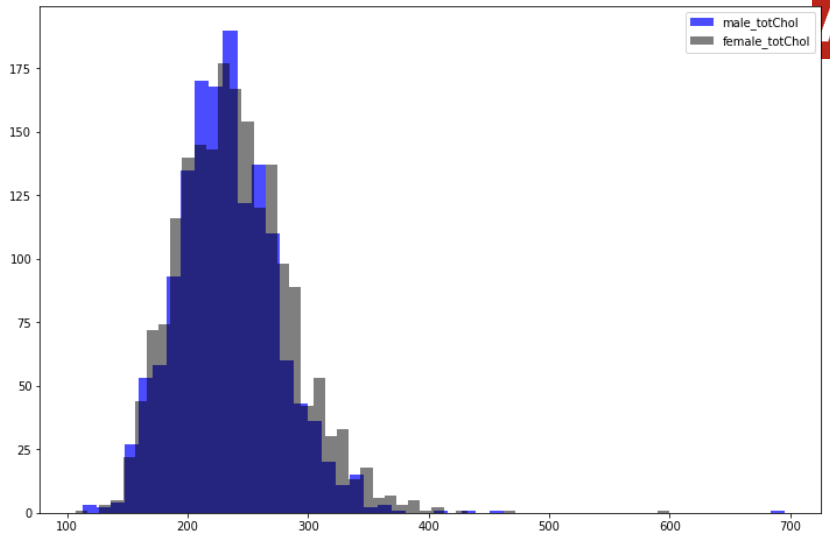




- Feature ‘systolic blood pressure’ for both the categories is positively skewed.
- In this feature females have higher peak values than male category.
- Here we can see the for female patients the blood pressure is more than normal range (90-120)

- Feature ‘diastolic blood pressure’ for both the categories shows almost normal distribution.
- We can see some high peak values in female category than male’s.
- In ‘diastolic blood pressure’ also females have some higher values than normal range.

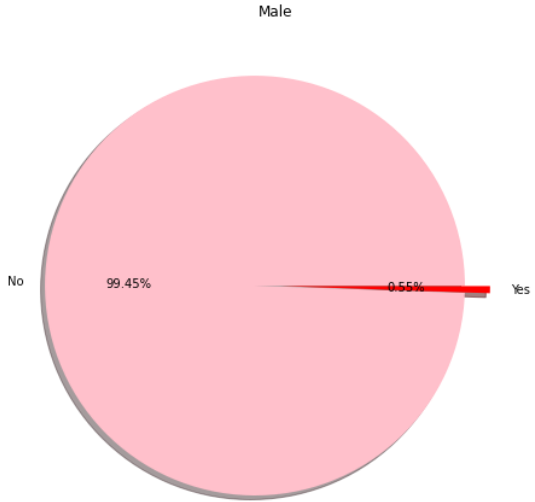


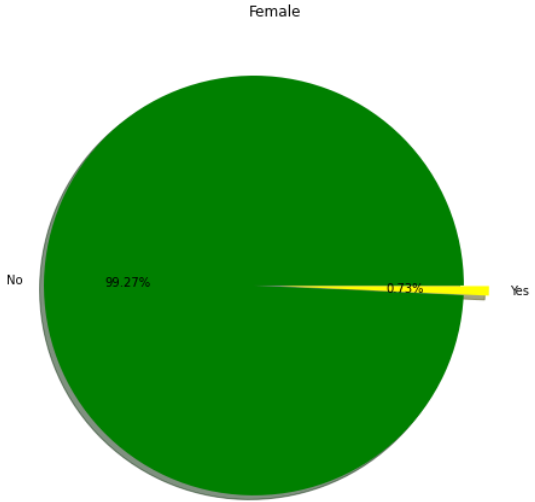


- Normal ‘heart rate’ range is 60 – 100, but we can see some high peaks and lower peaks in this distribution.
- Female category shows some high peaks which are beyond the normal range.
- Males patients also have some higher values in the heart rate feature.

- For the humans ‘total cholesterol’ should be less than 200 mg/dL
- Here the graph interprets that both the categories have well controlled ‘total cholesterol’ level which shows below 200 mg/dL.
- The data for both categories is positively skewed.

Prevalent stroke





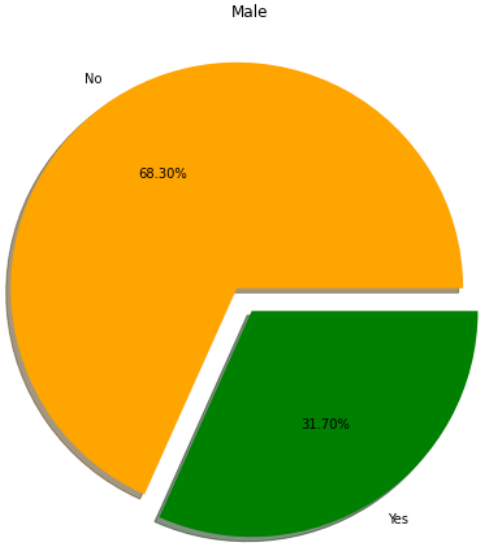
The ‘prevalent stroke’ is defined as “the proportion of patients who had a first stroke in a given population at a specified time” Considering the all male patients and finding 0.55% of them are affected is just negligible.

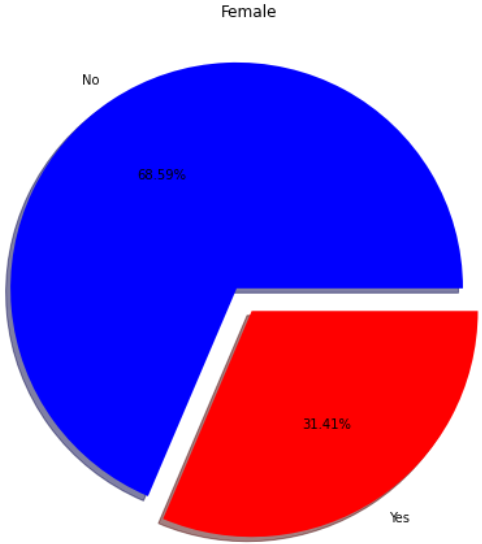
For the female patients, considering 0.73% of affected patients by ‘prevalent stroke’ is negligible.

As we see for both categories ‘prevalent stroke’ is not a big problem.

Prevalent Hypertension

Hypertension is high blood pressure which is higher than normal range (130/80 mm Hg).

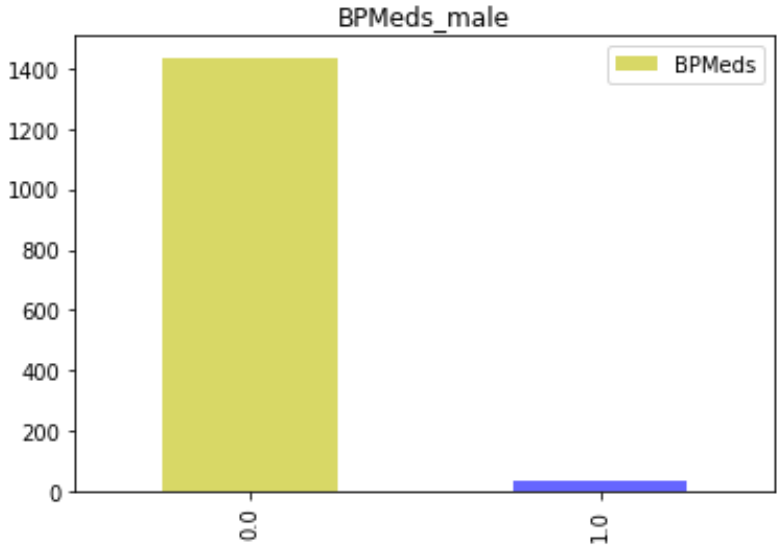


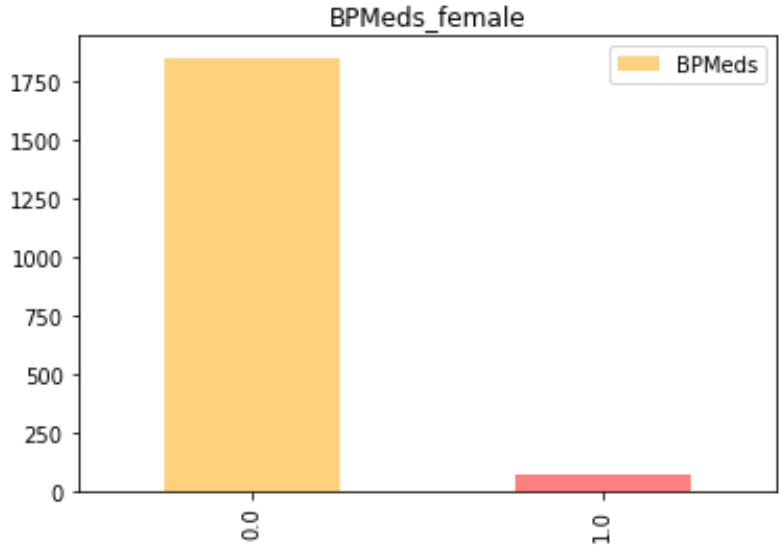


Out of total male patients 31.70% patients are affected with ‘prevalent hypertension’.
Its very big number of affected patients in male category.

31.41% of total female patients are affected by the ‘prevalent hypertension’
In this category also it very big number of patients which shows hypertension more than normal range.

BP medicine





Interpretations:

- Both categories consumes blood pressure medicines.
- Female’s consumes slightly more medicine than Male’s.

Insights of Feature Selection

Dependent variable: ‘TenYearCHD’

Independent variables: 'age', 'education', 'sex', 'is_smoking', 'cigsPerDay', 'BPMeds', 'prevalentStroke’, 'prevalentHyp', 'diabetes’, 'totChol', 'sysBP', 'diaBP', 'BMI', 'heartRate', 'glucose’

Balancing Data: ‘SMOTE’ method

Transformation: ‘Standard scaler’

Logistic Regression

Classification report of testing data

	Precision	Recall	F1-score	Support
0	0.84	1.00	0.91	563
1	0.86	0.05	0.10	115
accuracy			0.84	678
Macro avg.	0.85	0.53	0.50	678
Weighted avg.	0.84	0.84	0.77	678

Interpretations:

- Testing accuracy for predicting the Coronary Heart Disease TRUE is 0.86 .
- We have precision more than 84% for both the classes i.e., we can say that accuracy of positive predictions is good.
- Recall value for class 0 is perfectly good i.e., positive classes which correctly identified.

k-Nearest Neighbor Classifier:

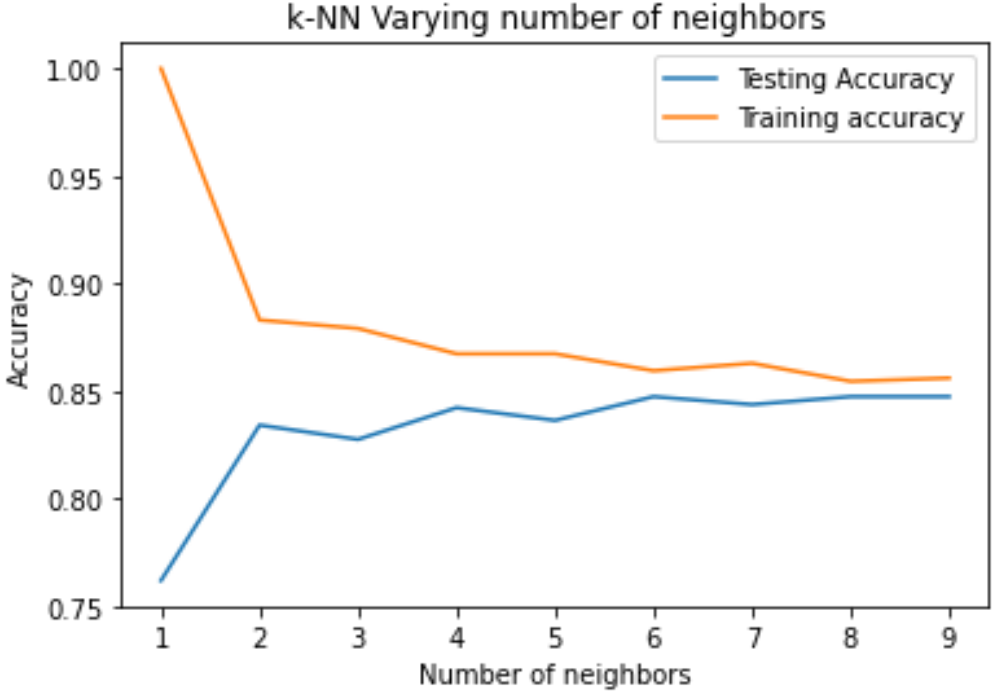
Classification report of testing data

	Precision	Recall	F1-score	Support
0	0.85	0.99	0.92	1152
1	0.43	0.05	0.09	204
accuracy			0.85	1356
Macro avg.	0.64	0.52	0.50	1356
Weighted avg.	0.79	0.85	0.79	1356

Interpretations:

- Testing accuracy for predicting the Coronary Heart Disease TRUE is 0.85.
- We have precision more than 85% for predicting the class 0 and less than 50% for predicting class 1 i.e., we can say that accuracy of positive predictions for class 0 is greater.
- Recall value for class 0 is perfectly good i.e., positive classes which correctly identified.

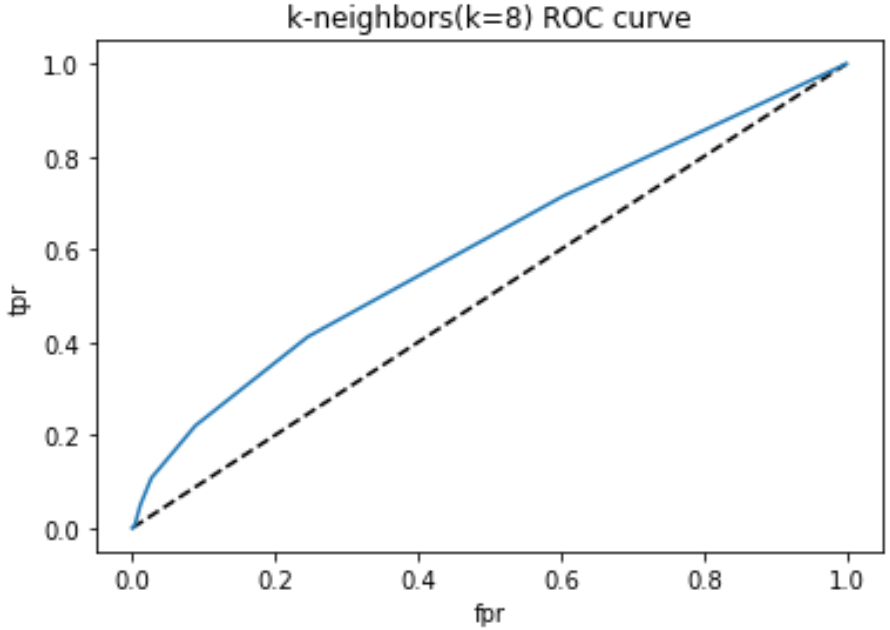
n_neighbors



Interpretations:

Here for testing and training accuracy is closer to the value 8, and therefore we take the n_neighbors value 8.

ROC-AUC curve



Interpretations:

- Here we have roc_auc score is 60% and we can say that there is high chances that classifier can distinguish between positive class values and negative class values.
- This is because classifier can detect more numbers of True Positives and Negatives instead of False Negatives and Positives.

Support Vector Machine

Classification report of testing data

	Precision	Recall	F1-score	Support
0	0.87	1.00	0.93	588
1	0.00	0.00	0.00	90
accuracy			0.87	678
Macro avg.	0.43	0.50	0.46	678
Weighted avg.	0.75	0.87	0.81	678

Interpretations:

- Testing accuracy for predicting the Coronary Heart Disease TRUE is 0.87
- We have precision more than 85% for predicting the class 0 i.e., we can say that accuracy of positive predictions for class 0 is very good.
- Recall value for class 0 is perfectly good i.e., positive classes which correctly identified.

Observations

1. In the dataset we had have some missing values and fill them.

2. In number of patients Females are comparatively more.

3. Primary educated patients are greater in number.

4. Males consumes more cigarettes per day than Females

5. No prevalent stroke history in both Male and Female cases.

6. We have seen the females have comparatively greater hypertension than males.

7. Both males and females, they were not consuming the BP medicines and the patients less than 5% consumes medicine.

Conclusion

In our analysis we initially did EDA on all the features of the dataset. Our dependet variable is 'OneYearCHD'. In our dataset we have some missing values and we fill them with different methods, we don't have duplicates in the data.

In the dataset we have two 'Object' data type and before the further operations we Encode them as integer. We use different types of visualization techniques to plot some graphs.

Further we use SMOTE(Synthetic Minority Oversampling Technique) for balancing the imbalance of the dataset. After this we scale the dataset by using Standard Scaler and then use some Machine Learning Classification techniques.

1. Among all three classifiers SVM gives the highest testing accuracy 87% and k-NN gives 85%, Logistic classifier gives 84%.
2. In hyper parameter tunning in k-NN is also gives us same accuracy.