Capstone Project-3

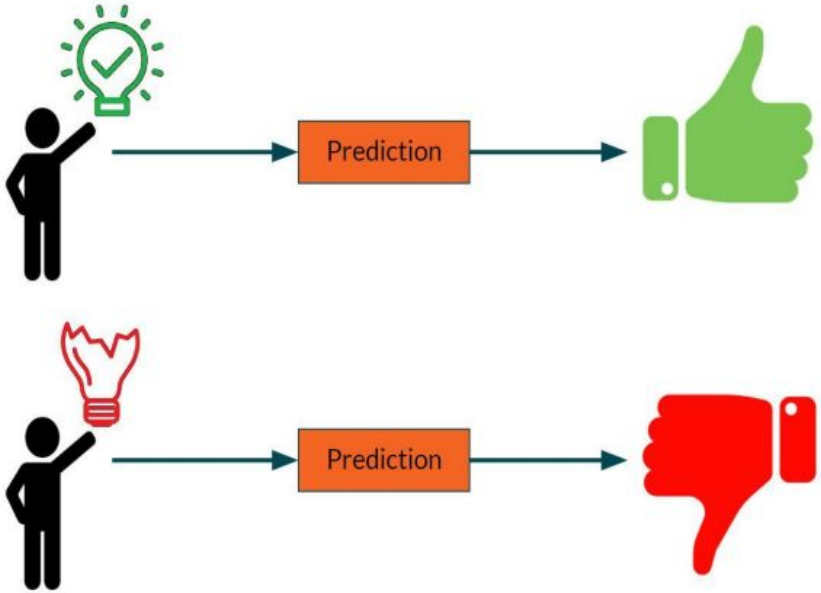
Cardiovascular Risk Prediction

By- Parth Sharma

- **STEPS INVOLVED:**

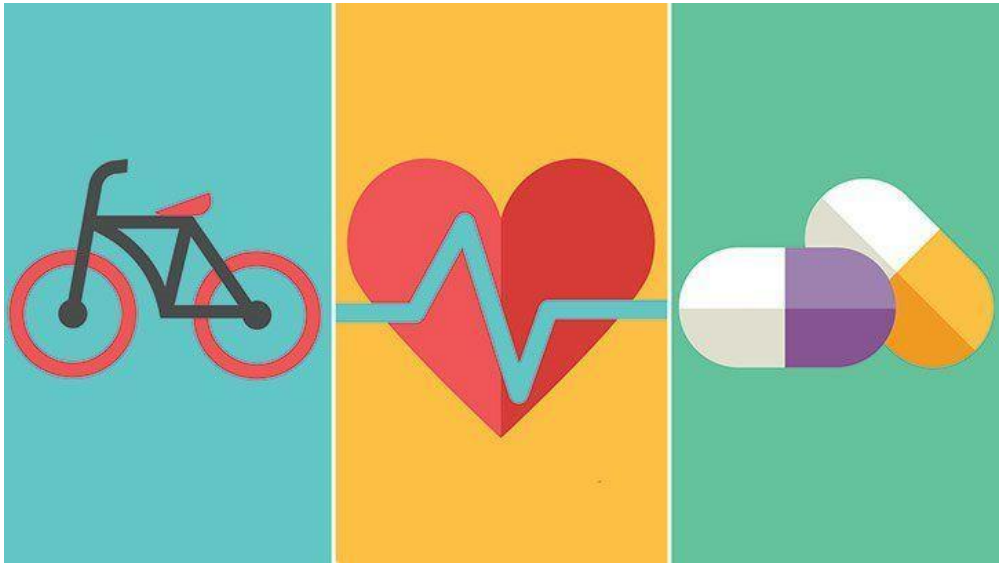
Steps that significantly contribute towards achieving the final results are listed below:

1. Defining The Problem Statement.
2. Applying The Data Pre-Processing Steps.
3. Exploratory Data Analysis.
4. Feature Selection And Transformation.
5. Classification Model Fitting.
6. Comparing The Metrics.
7. Selecting The Best Model.



WHY DO WE NEED CARDIOVASCULAR RISK PREDICTION:

- Predicting and diagnosing heart disease is the biggest challenge in the medical industry. There are many factors which influence heart diseases.
- Heart Disease is even highlighted as a silent killer which leads to the death of the person without obvious symptoms.
- The early diagnosis of heart disease plays a vital role in making decisions on lifestyle changes in high-risk patients and in turn reduces the complications.
- Machine learning can play a vital and accurate role in predicting chances of heart disease in coming potential years based upon the current way of living



- **DATA PIPELINE**:
  - **Data Processing-1**: In this initial step we went to look for different features available and tried to uncover their relevance with the target variable and we have directly dropped id column as it the one that is most irrelevant from our analysis perspective.
  - **Data Processing-2:** During this stage, then, we looked for the data types of each feature and corrected them. After that comes the null value and outliers detection and treatment. For the null values we have used iterative imputation technique and for the outliers capping is done to eliminate the outliers without any loss to the data.
  - **EDA:** EDA or Exploratory Data Analysis is the critical process of performing the initial investigation on the data. So, through this we have observed certain trends and dependencies and also drawn certain conclusions from the dataset that will be useful for further processing
  - **Feature Selection and Transformation:** During this stage, we went on to select the most relevant features using the chi-square score and next comes the feature scaling in order to bring down all the values in similar range. After that comes the treatment of class imbalance in the target variable that is done using random oversampling.

- **DATA PIPELINE**:
  - **Model Fitting and Metric Evaluation**: Since the data is transformed to an appropriate form therefore we pass it to different classification models and calculate the metrics on the basis of which we select a model that could give us better prediction.

ABOUT THE DATASET

- The dataset is from an ongoing cardiovascular study on residents of the town of Framingham, Massachusetts.
- The dataset provides the patients information.
- The data set consists of record of almost 3390 recordes and 17 features.
- Some of the features are Categorical in nature while other holds numeric data type.
- The target variable namely 'TenYearCHD' refers to whether the patient suffers from Coronary heart disease depending upon the values of current medical parameters.
- The dependent variable consists of the binary values where,1-Risk of Coronary Heart Disease and 0-No Risk of Coronary Heart Disease.

- **INDEPENDENT VARIABLES:**

These are divided in certain categories. All the dependent variables are listed below:

- **Demographic:**
  - 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).

- **INDEPENDENT VARIABLES:**
  - **Medical(current):**
    - Tot Chol: Total cholesterol level (Continuous).
    - Sys BP: Systolic blood pressure (Continuous).
    - Dia BP: Diastolic blood pressure (Continuous).
    - 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).

	id	age	education	sex	is_smoking	cigsPerDay	BPMeds	prevalentStroke	prevalentHyp	diabetes	totChol	sysBP	diaBP	BMI	heartRate	glucose
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	36	4.0	M	NO	0.0	0.0	0	1	0	212.0	168.0	98.0	29.77	72.0	75.0
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
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
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
5	5	61	3.0	F	NO	0.0	0.0	0	1	0	272.0	182.0	121.0	32.80	85.0	65.0
6	6	61	1.0	M	NO	0.0	0.0	0	1	0	238.0	232.0	136.0	24.83	75.0	79.0
7	7	36	4.0	M	YES	35.0	0.0	0	0	0	295.0	102.0	68.0	28.15	60.0	63.0
8	8	41	2.0	F	YES	20.0	NaN	0	0	0	220.0	126.0	78.0	20.70	86.0	79.0
9	9	55	2.0	F	NO	0.0	0.0	0	1	0	326.0	144.0	81.0	25.71	85.0	NaN

• DEPENDENT VARIABLES:

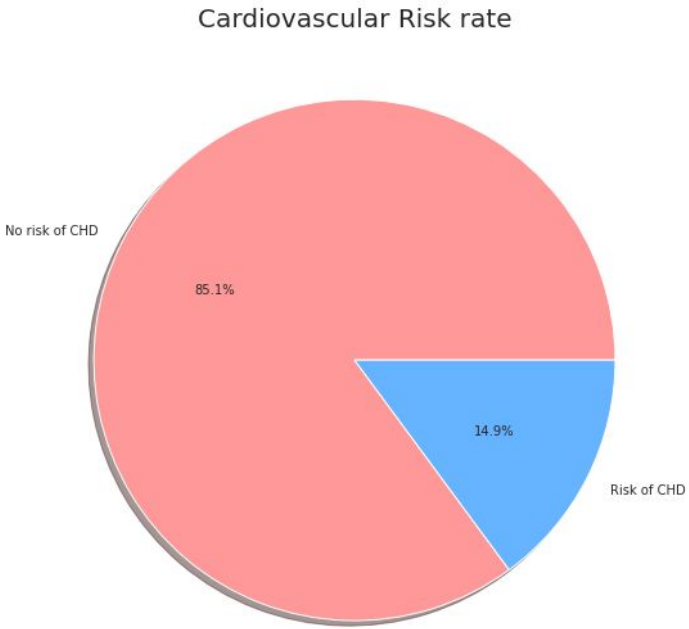
- Our dependent variable i.e TenYearCHD that refers to the Risk of Coronary Heart Disease in coming 10 years.

- 10-year risk of coronary heart disease CHD is binary i.e. it only hold two discrete values:

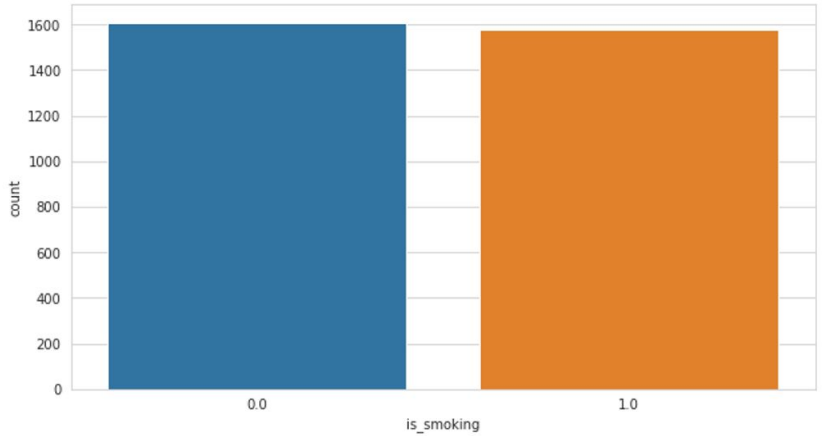
  “1”-“Yes: There is a risk of CHD”

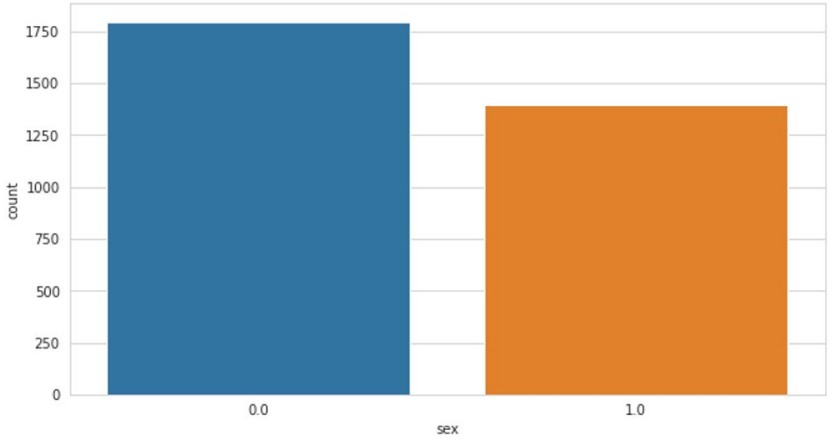
  “0” -“No: There is no risk of CHD”

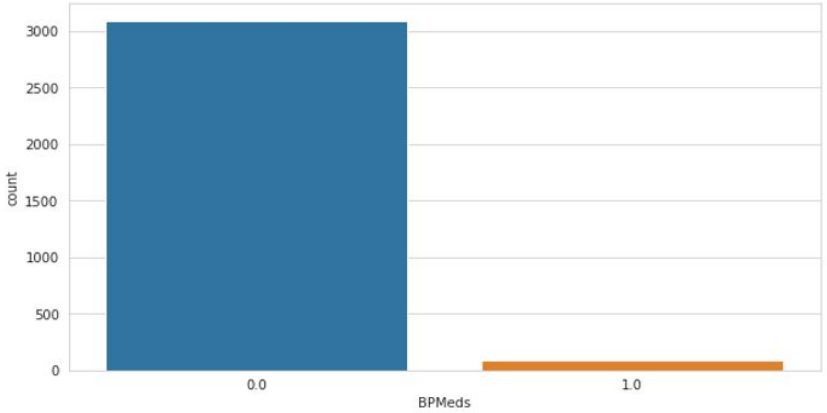
TenYearCHD
0.0
0.0
1.0
0.0
1.0
0.0
0.0
0.0
0.0
0.0

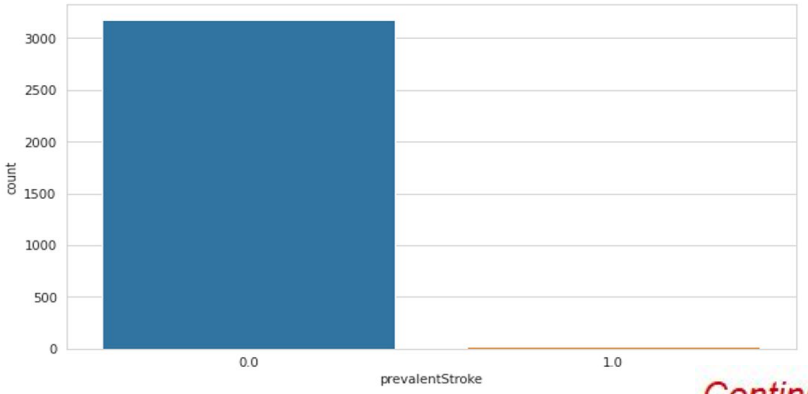


- **EXPLORATORY DATA ANALYSIS:(UNIVARIATE ANALYSIS)**



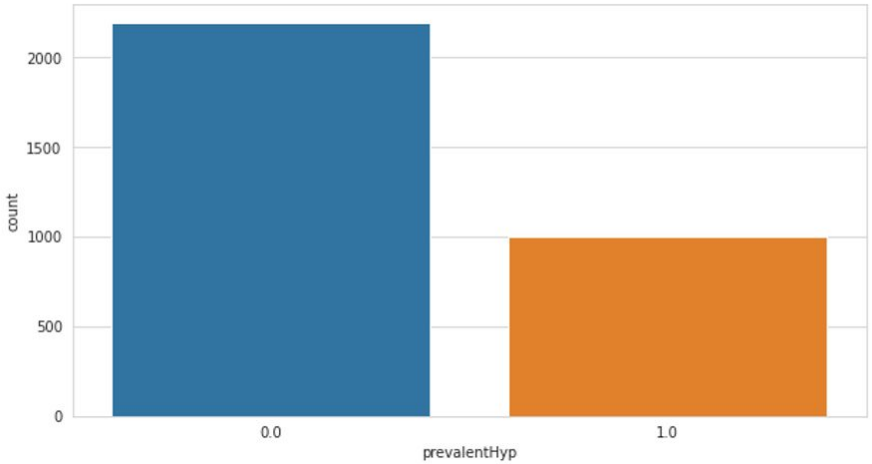


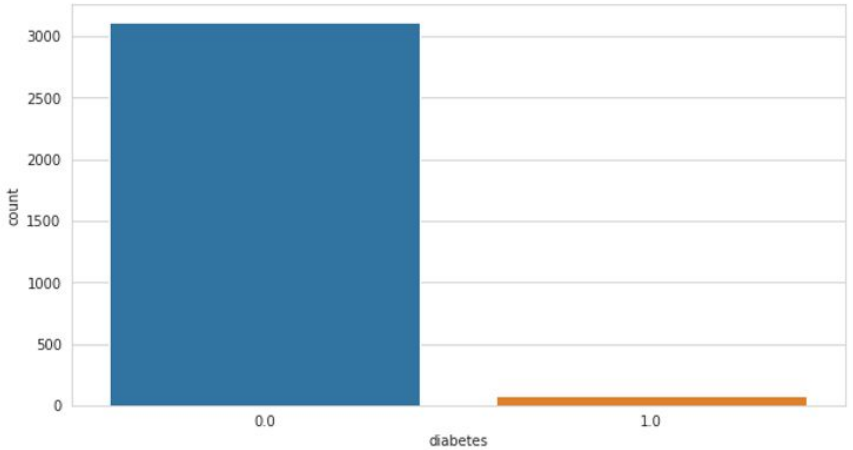




Continued…

• EXPLORATORY DATA ANALYSIS:(UNIVARIATE ANALYSIS)



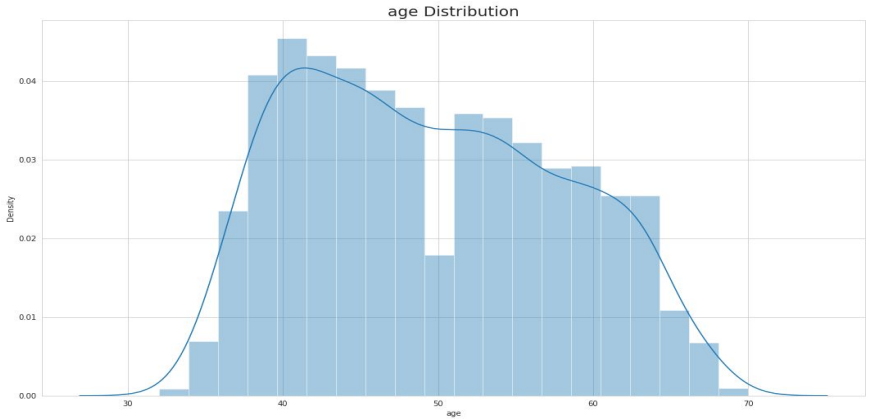


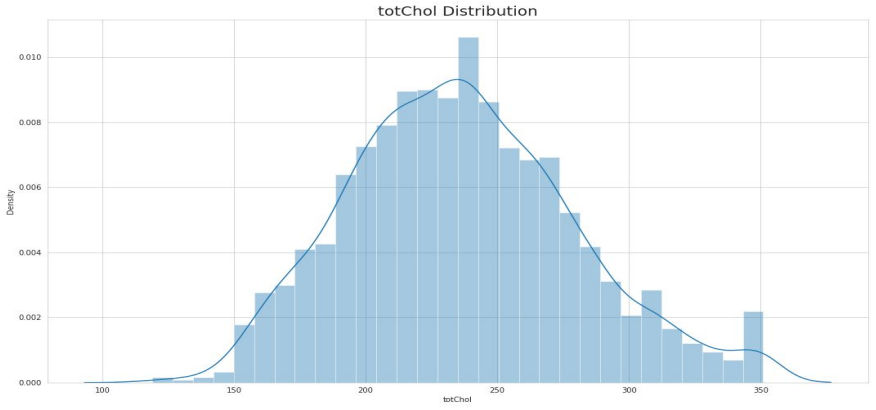
• Some Observations:

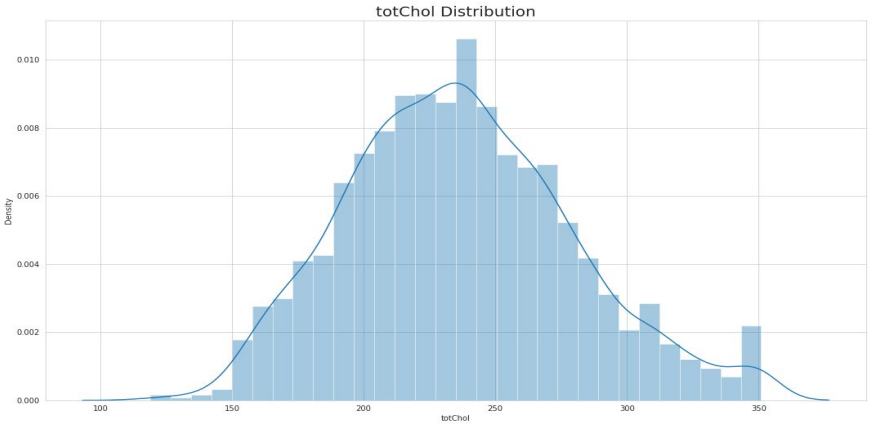
- Females are more in proportion to men by a small margin.
- There are more non-smokers than smokers, but the count for both the class is comparable.
- More than 3000 people are not on BP medication
- Only a small number of people have suffered a stroke previously.
- Around 1000 people were hypertensive.
- A large number (> 3000) of the people do not have diabetes.

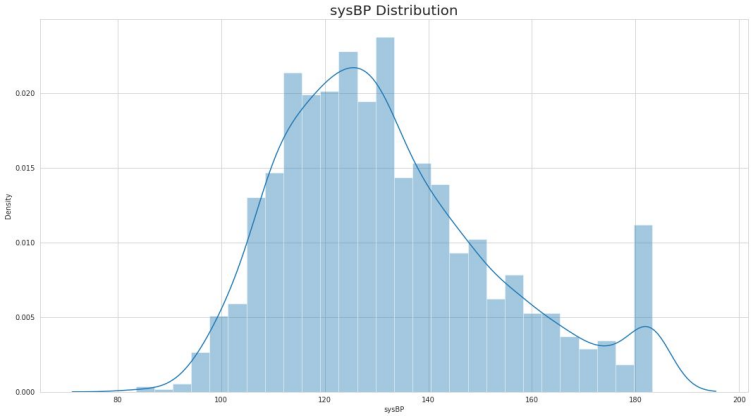
Continued…

- **EXPLORATORY DATA ANALYSIS:(UNIVARIATE ANALYSIS)**



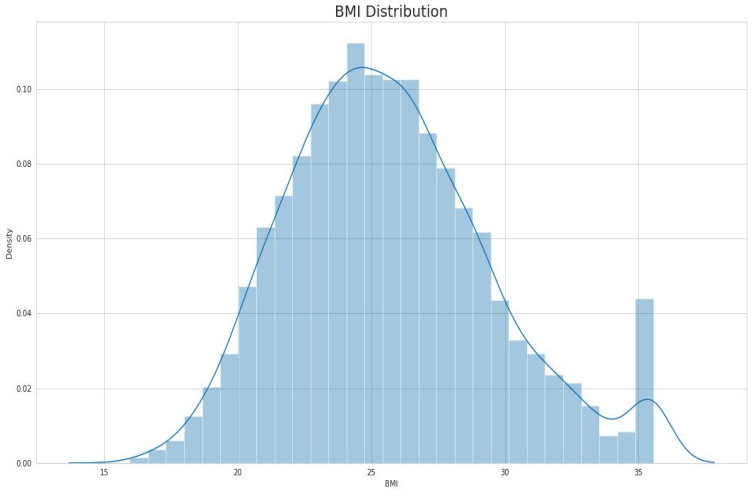






Continued…

• EXPLORATORY DATA ANALYSIS:(UNIVARIATE ANALYSIS)

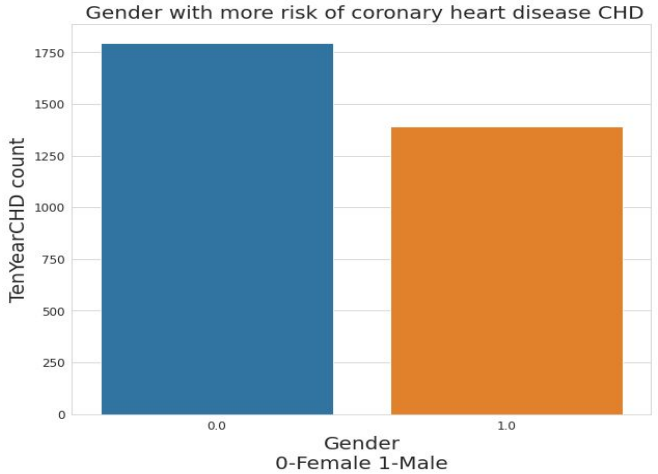


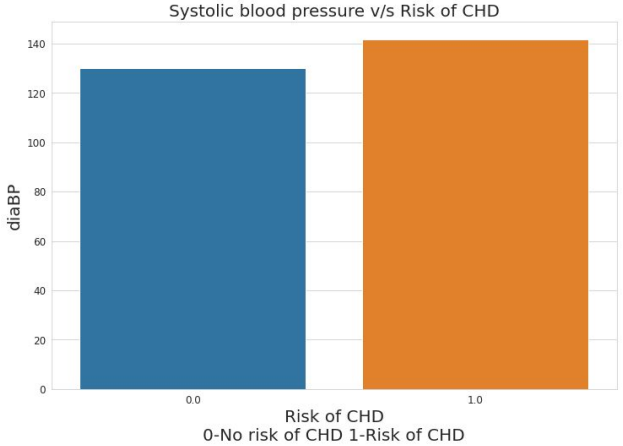
• Some Observations:

- Age ranges from 35 years to 70 years and is almost normally distributed, with most people belonging to an age group of 40.
- Cigarettes smoked per day on an average are majorly 0, but 20 cigarettes a day are also prevalent.
- Cholesterol ranges form 100 to 700, with most belonging to 150 to 350 and this range is normal according to medical documents.
- Systolic BP ranges from 100 to 180 and Diastolic BP ranges mainly from 60 to 120.
- BMI ranges mainly from 15 to 40.
- Heart rate ranges from 50 to 105 and most occurrences are around 75.
- Glucose ranges mainly from 50 to 105.

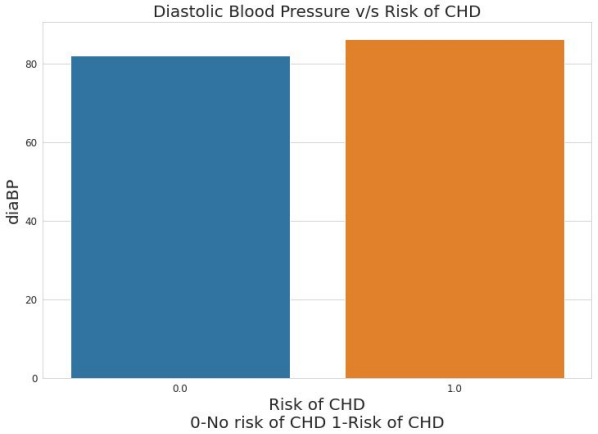
Continued…

• EXPLORATORY DATA ANALYSIS:(BIVARIATE ANALYSIS)



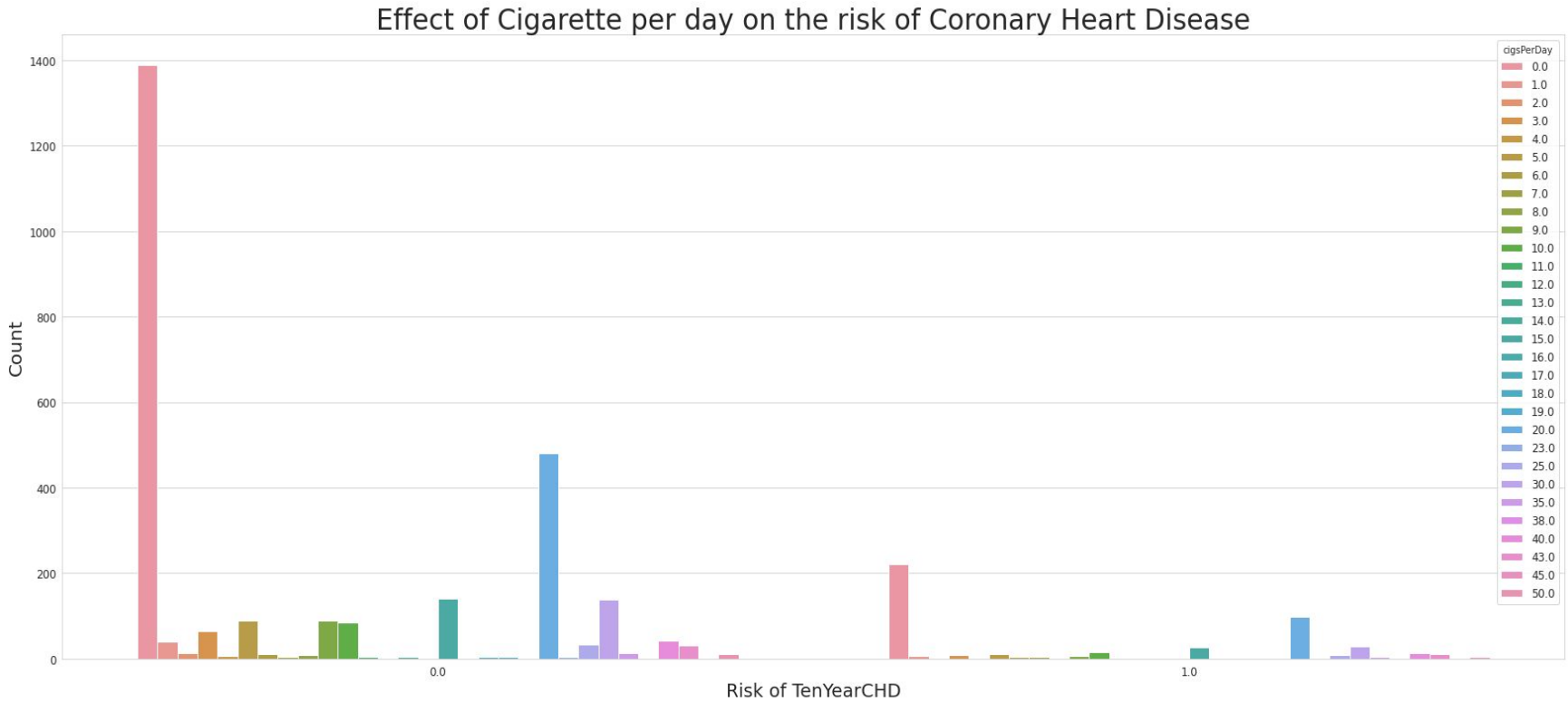


- Females have shown a slightly higher risk of coronary heart disease.
- As the value of diastolic and systolic blood pressure shoots above the normal the chances of coronary heart disease increases.

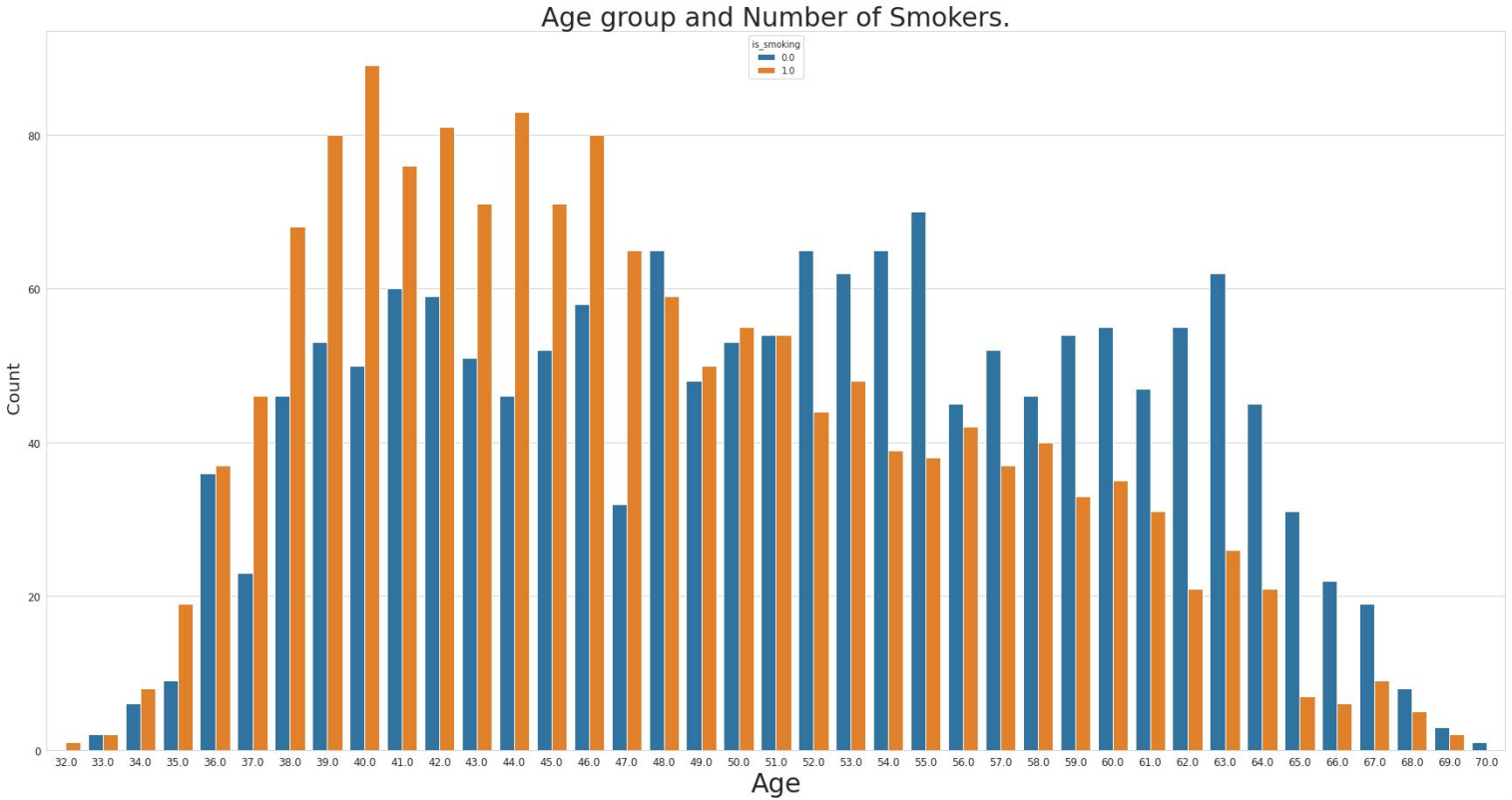


Continued…

- **EXPLORATORY DATA ANALYSIS:(BIVARIATE ANALYSIS)**

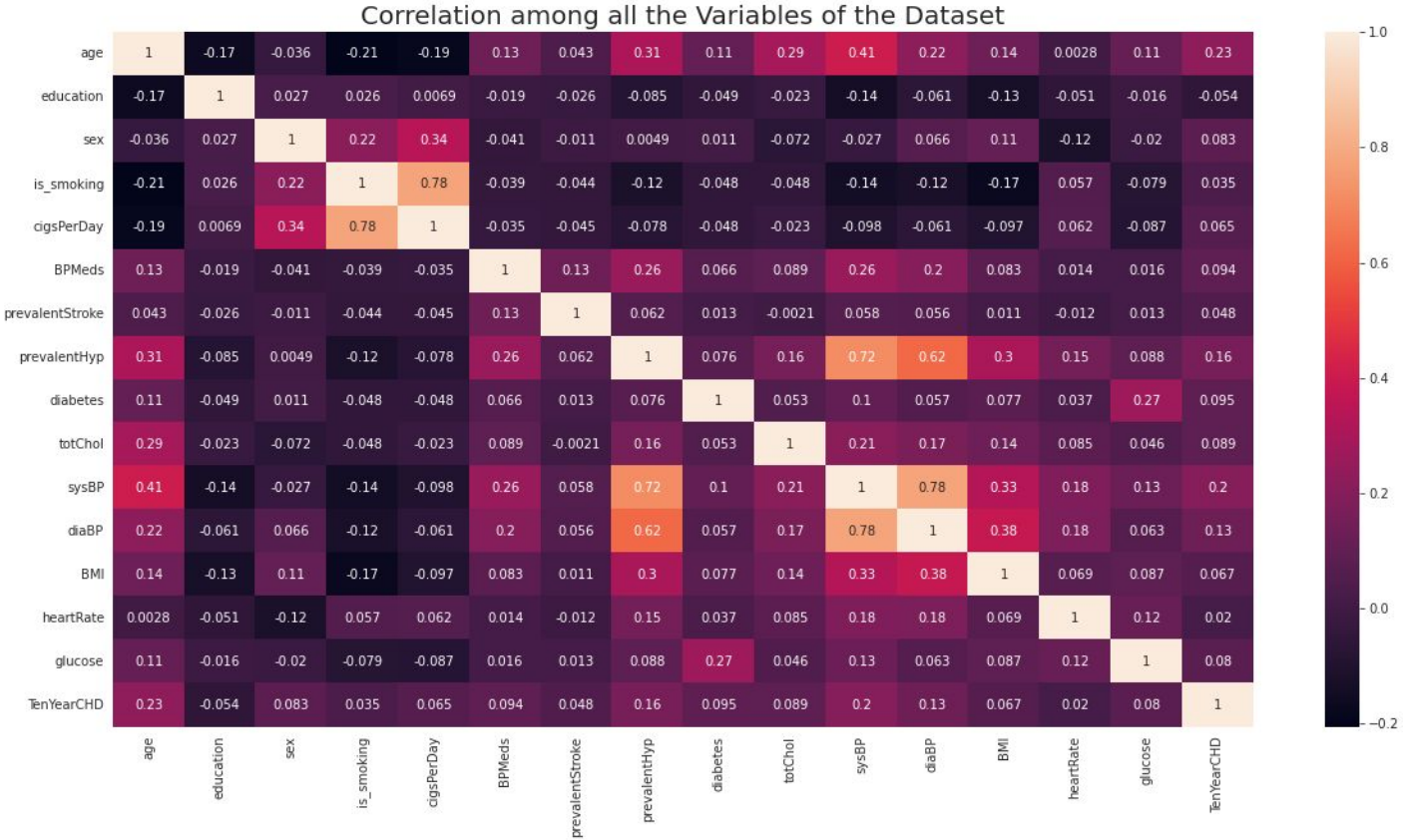


- **EXPLORATORY DATA ANALYSIS:(BIVARIATE ANALYSIS)**

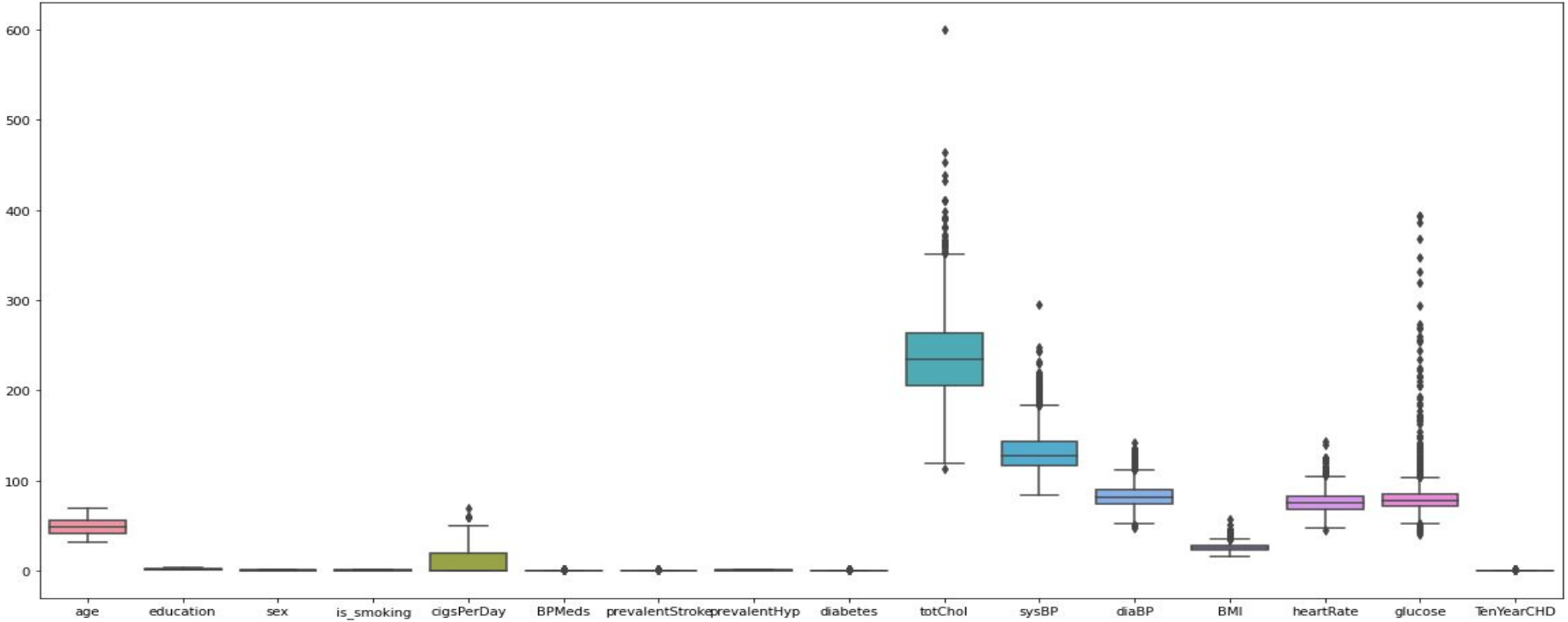


Continued…

- **CORRELATION MATRIX:**

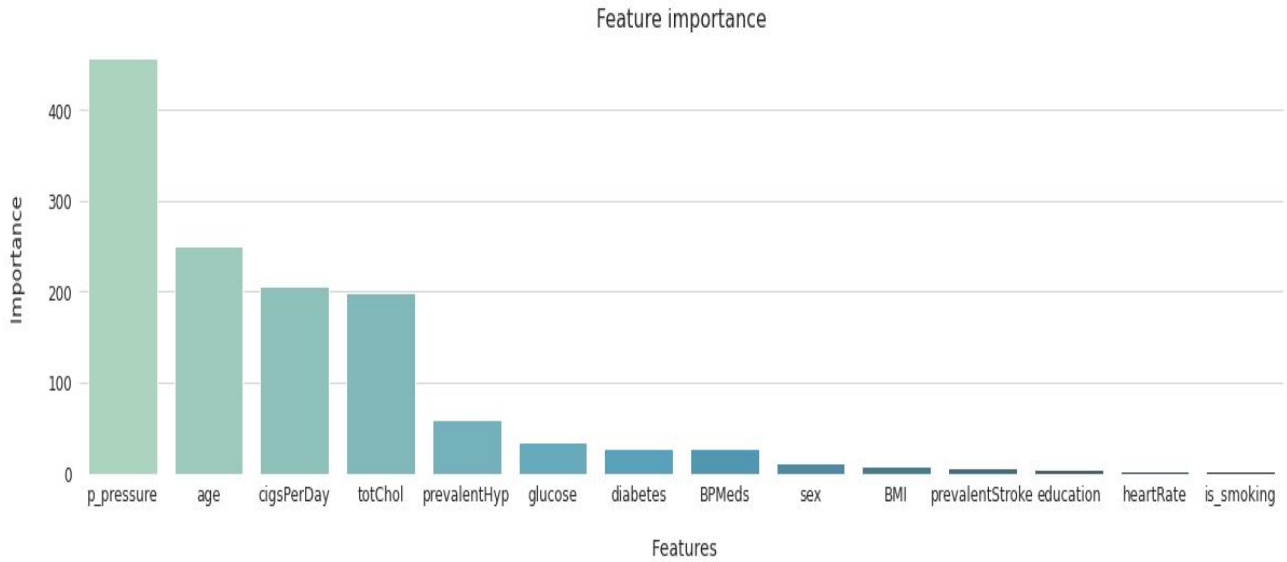


- **OUTLIERS DETECTION:**



• FEATURE SELECTION:

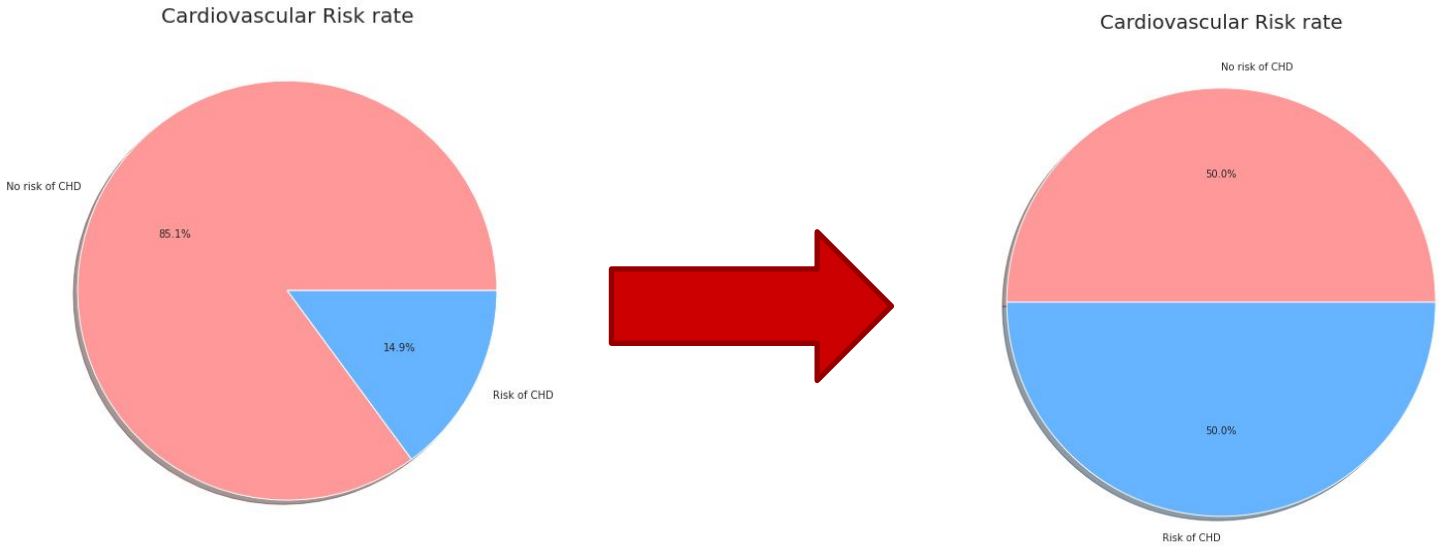
- Pulse Pressure has a Chi-score of 457.2 i.e. it has the maximum influence on the target variable followed by the age.
- prevalentStroke, education, heartRate, is_smoking accounts the least effect on the target variable as the chi-score is very low.



	Independent Feature	Chi_Score
13	p_pressure	457.250519
0	age	249.931510
4	cigsPerDay	205.859597
9	totChol	199.719662
7	prevalentHyp	58.674438
12	glucose	34.243363
8	diabetes	27.931583
5	BPMeds	27.187058
2	sex	12.252065
10	BMI	8.294096
6	prevalentStroke	7.219641
1	education	4.826685
11	heartRate	2.207399
3	is_smoking	1.960362

• DEALING WITH CLASS IMBALANCE:

- We have used the SMOTE for oversampling the minority class.
- Synthetic Minority Oversampling Technique or SMOTE is another technique to oversample the minority class. Simply adding duplicate records of minority class often don't add any new information to the model. In SMOTE new instances are synthesized from the existing data.



- **CLASSIFICATION MODEL AND MODEL EVALUATION:**

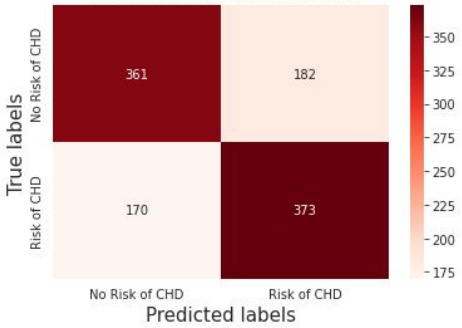
CLASSIFICATION METRICS COMPARISON

	Model	Train Accuracy	Test Accuracy	Train F1-score	Test F1-score	Train Precision	Test Precision	Train Recall	Test Recall	Train ROC AUC	Test ROC AUC
0	Logistic Regresssion	0.669968	0.675875	0.672607	0.679417	0.667271	0.672072	0.678029	0.686924	0.669968	0.675875
1	KNN	1.000000	0.883057	1.000000	0.888694	1.000000	0.847826	1.000000	0.933702	1.000000	0.883057
2	Decision Tree	1.000000	0.818600	1.000000	0.818433	1.000000	0.819188	1.000000	0.817680	1.000000	0.818600
3	Random Forest Classifier	1.000000	0.876611	1.000000	0.870406	1.000000	0.916497	1.000000	0.828729	1.000000	0.876611
4	SVM	0.791340	0.787293	0.775964	0.767839	0.837694	0.845133	0.722708	0.703499	0.791340	0.787293
5	LGBM With Grid Search CV	0.752418	0.766114	0.765437	0.775618	0.727197	0.745331	0.807923	0.808471	0.752418	0.766114
6	RFC Using Randomised Search CV	1.000000	0.894107	1.000000	0.889741	1.000000	0.928000	1.000000	0.854512	1.000000	0.894107

Continued…

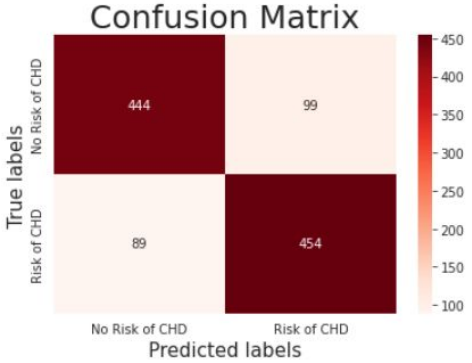
• CLASSIFICATION MODEL AND MODEL EVALUATION:

LOGISTIC REGRESSION

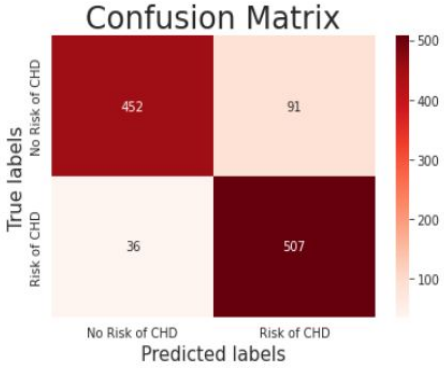


	Predicted 0	Predicted 1
Actual 0	TN	FP
Actual 1	FN	TP

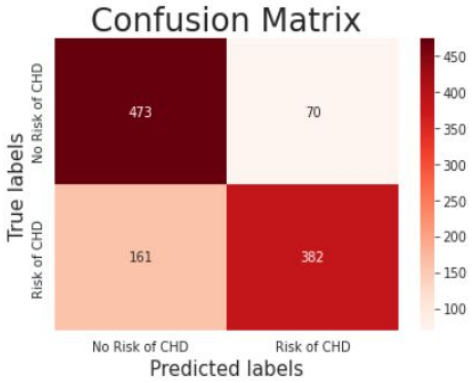
DECISION TREE



KNN



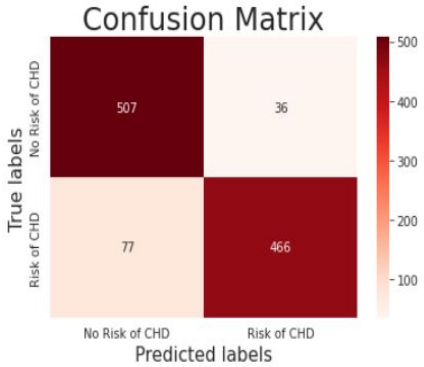
SVM



Continued…

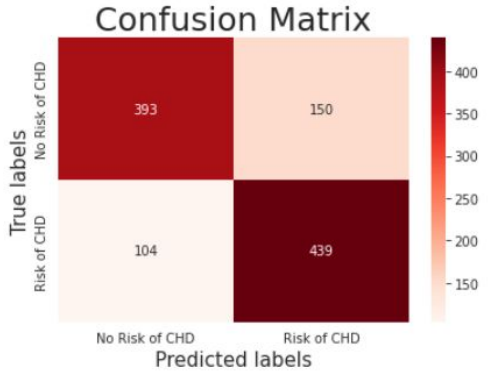
• CLASSIFICATION MODEL AND MODEL EVALUATION:

RFC WITH RANDOMIZED SEARCH CV

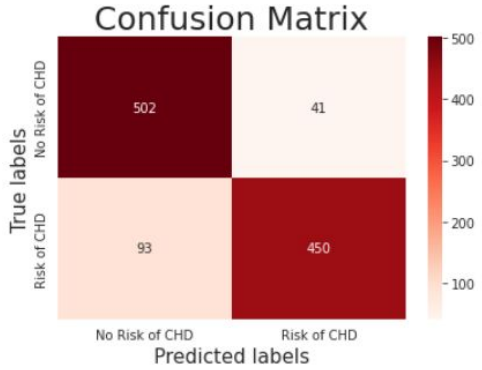


	Predicted 0	Predicted 1
Actual 0	TN	FP
Actual 1	FN	TP

LGBM



RANDOM FOREST CLASSIFIER



Continued…

- **CLASSIFICATION MODEL AND MODEL EVALUATION:**

KNN

	precision	recall	f1-score	support
0.0	0.93	0.83	0.88	543
1.0	0.85	0.93	0.89	543
accuracy			0.88	1086
macro avg	0.89	0.88	0.88	1086
weighted avg	0.89	0.88	0.88	1086

LGBM

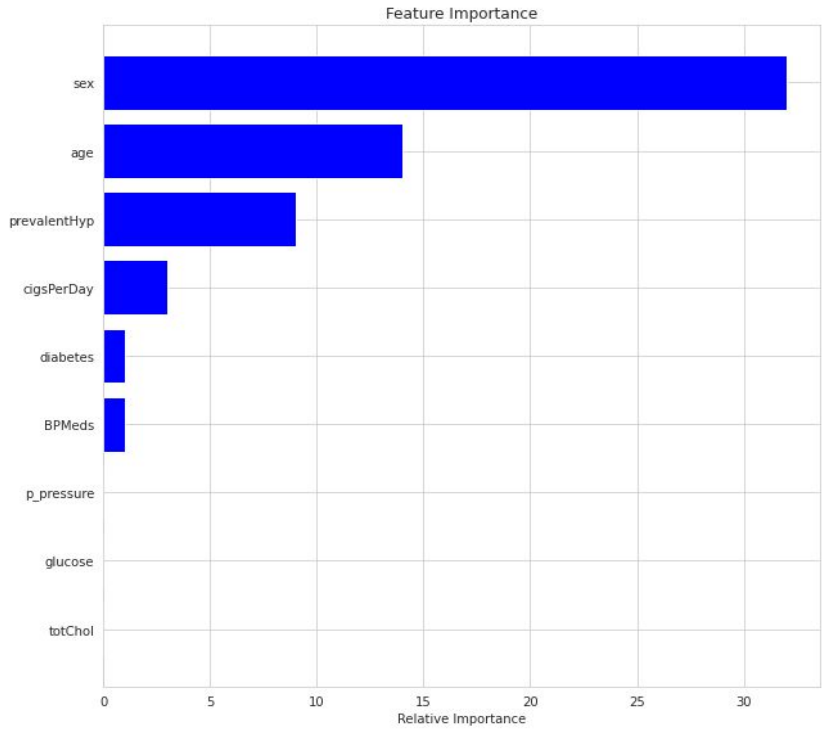
	precision	recall	f1-score	support
0.0	0.79	0.72	0.76	543
1.0	0.75	0.81	0.78	543
accuracy			0.77	1086
macro avg	0.77	0.77	0.77	1086
weighted avg	0.77	0.77	0.77	1086

RFC WITH RANDOMIZED SEARCH CV

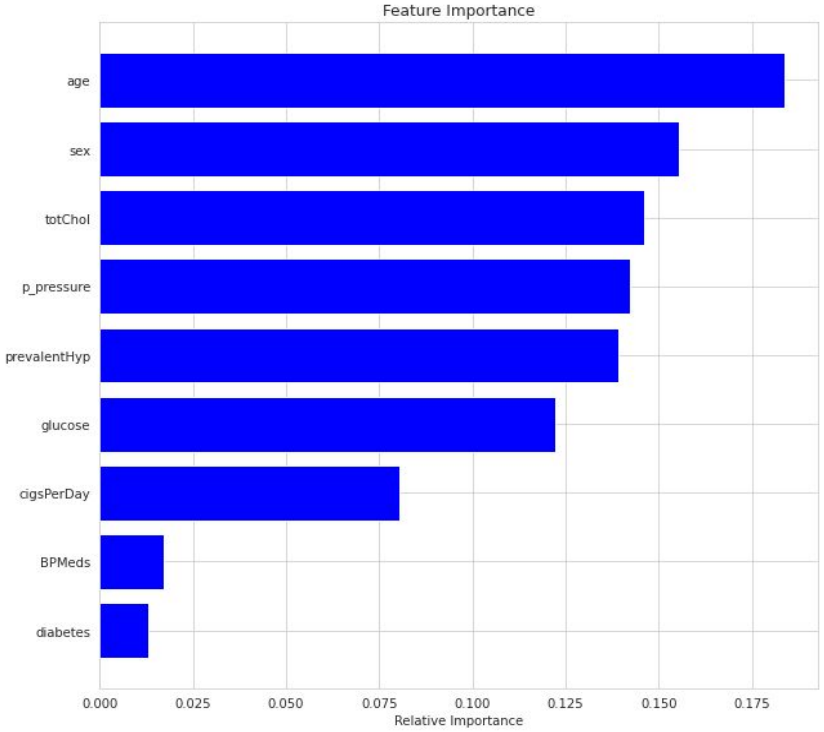
	precision	recall	f1-score	support
0.0	0.87	0.93	0.90	543
1.0	0.93	0.85	0.89	543
accuracy			0.89	1086
macro avg	0.90	0.89	0.89	1086
weighted avg	0.90	0.89	0.89	1086

- **HYPER TUNED MODEL:**
  - **Feature Impotance:**

LGBM



RFC WITH RANDOMIZED SEARCH CV



- **OBSERVATIONS:**

1. As we can observe from the metrics comparison table, Logistic Regression is not giving us good results as accuracy as well as recall is least of all the models.

2. Talking of Decision tree it was gave us good results than Logistic regression. If we compare it with others on recall even than it outperformed SVM, Random Forest Classifier.

3. K-Nearest Neighbours showed the best results as the best accuracy if we don't consider hyper tuned models but as far recall is considered it is best of all the models i.e.0.933.

Continued…

- **OBSERVATIONS:**

4. Out of our tuned models that is Light Gradient Boosting Machine With Grid Search CV and Random Forest Classifier Using Randomised Search CV, Random Forest Classifier Using Randomised Search CV performed better.

5. The hyper-tuned random forest classifier performed well in comparison with the base random forest classifier model.

6. Looking at the business problem recall is utmost important to us as there should be no case where a person having risk of CHD left unattended. Therefore, we will choose KNN as it give the highest value of recall i.e.0.933.

- **CONCLUSIONS:**

1. The models that could can be deployed according to our study is KNN with Accuracy-0.88,Precision-0.84,Recall-0.93.
2. Better model can be developed that can predict the risk of coronary heart disease.
3. With the help of the experts we can engineer an extensive amount of variable that could make our prediction more accurate.
4. Hence, we can say that Machine Learning can help save lives of many and help them to switch over to a healthy lifestyle to chop off any health related issue.

Thank You