Capstone Project

Cardiovascular Disease Risk Prediction

Individual contributer

Mohammed Javeed

Introduction

- Cardiovascular diseases (CVDs) are the leading cause of death globally. An estimated 17.9 million people died from CVDs in 2019, representing 32% of all global deaths. Of these deaths, 85% were due to heart attack and stroke.

- Out of the 17 million premature deaths (under the age of 70) due to noncommunicable diseases in 2019, 38% were caused by CVDs.

- Most cardiovascular diseases can be prevented by addressing behavioural risk factors such as tobacco use, unhealthy diet and obesity, physical inactivity and harmful use of alcohol.

- It is important to detect cardiovascular disease as early as possible so that management with counselling and medicines can begin.

Steps in the project

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.

Problem statement

- The dataset is from an ongoing cardiovascular study on resident sof the town of Framingham,Massachusetts.
  The classification goal is to predict whether the patient has a 10-year risk of future coronary heartdisease(CHD).

- •The data set provides the patients' information. It includes over 4,000 records and 15 attributes.

- Numerical columns:
  id,age,totChol,sysBP,diaBP,BMI,heartRate,glucose

- Categoricalcolumns:
  education,cigsPerDay,sex,is_smoking,BPMeds,prevalentStroke,prevalentHyp,diabetes, TenYearCHD

  Datasetshape:(3390,17)

Data Dictionary

The meanings of the various columns are as follows :-

●Demographic:

1.Sex: male or female("M" or "F")

2.Age: Age of the patient;(Continuous - Although the recorded ages have been truncated to whole numbers, the concept of age is continuous)

●Behavioral:

1.is_smoking: whether or not the patient is a current smoker ("YES" or "NO")

2.Cigs Per Day: the number of cigarettes that the person smoked on average in one day.(continuous as one can have any number of cigarettes, even half a cigarette.)

●Medical(history):

1.BP Meds: whether or not the patient was on blood pressure medication (Nominal)

2.Prevalent Stroke: whether or not the patient had previously had a stroke (Nominal)

3.Prevalent Hyp: whether or not the patient was hypertensive (Nominal)

4.Diabetes: whether or not the patient had diabetes (Nominal)

●Medical(current):

1.Tot Chol: total cholesterol level (Continuous)

2.Sys BP: systolic blood pressure (Continuous)

3.Dia BP: diastolic blood pressure (Continuous)

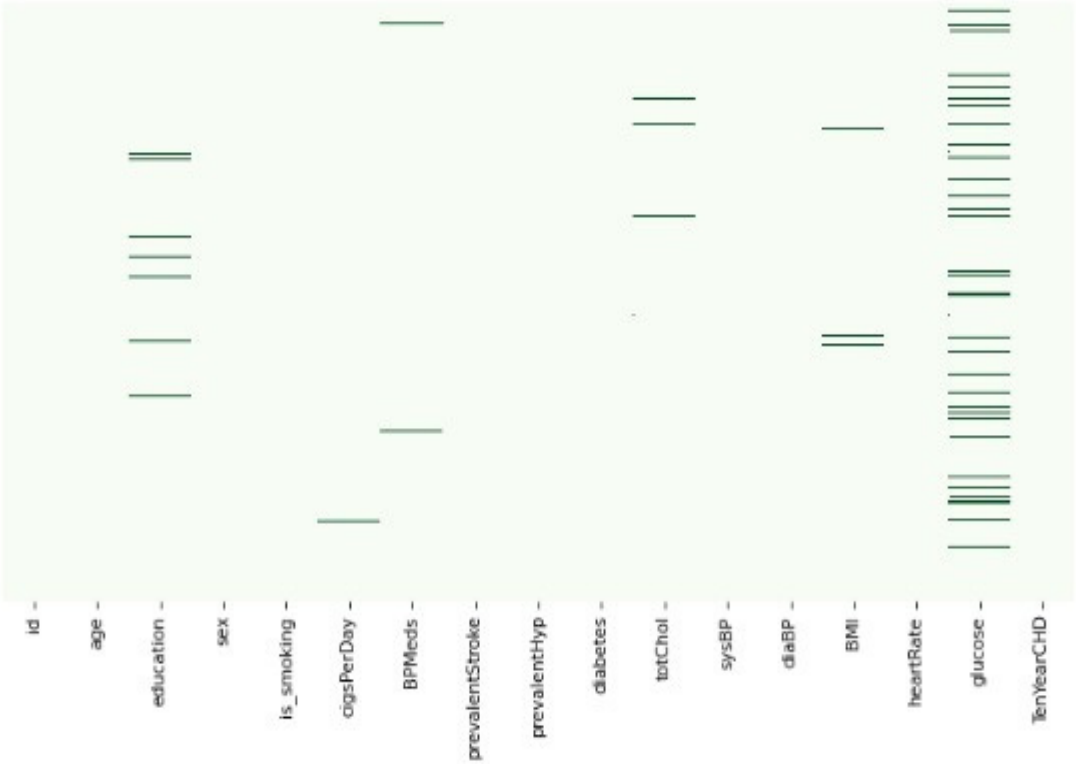
4.BMI: Body Mass Index (Continuous)

5.Heart Rate: heart rate (Continuous)

6.Glucose: glucose level (Continuous)

EDA

Visualization of Missing values



Missing values in the data

glucose	304
education	87
BPMeds	44
totChol	38
cigsPerDay	22
BMI	14
heartRate	1
age	0
sex	0
is_smoking	0
prevalentStroke	0
prevalentHyp	0
diabetes	0
sysBP	0
diaBP	0
TenYearCHD	0

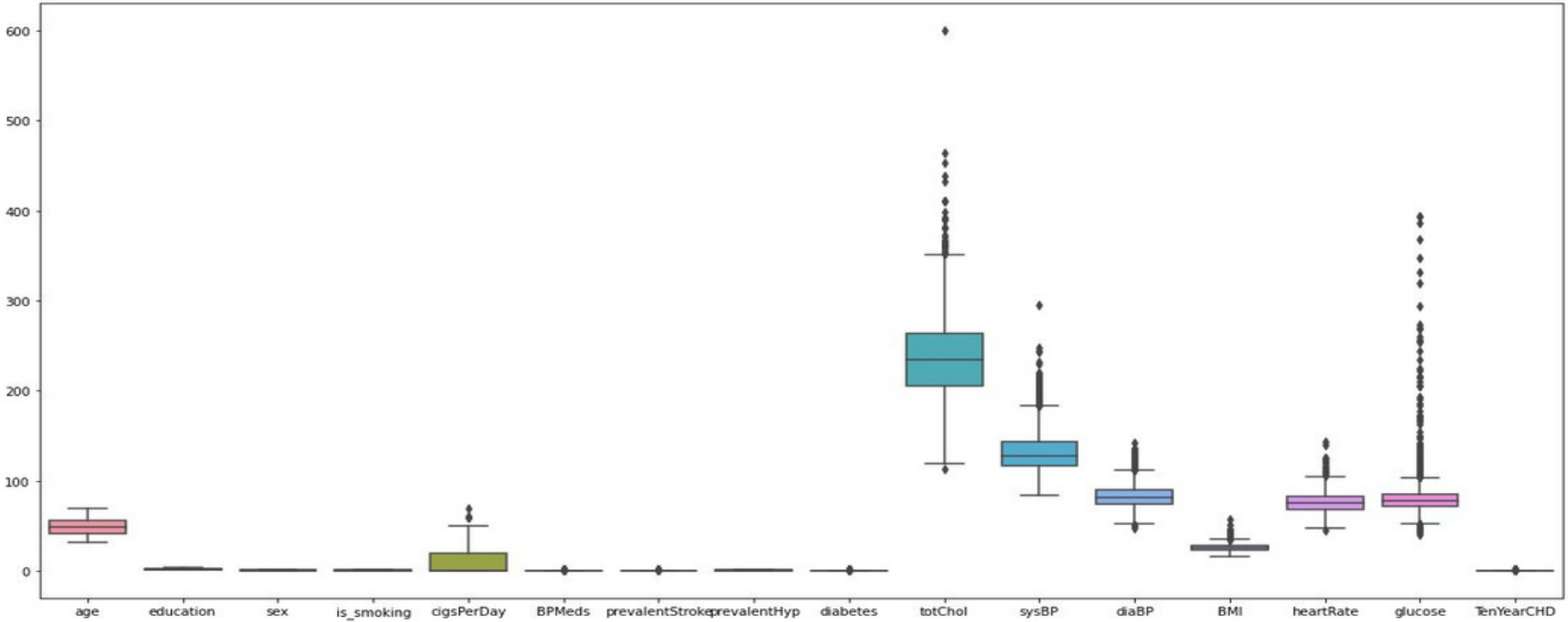
dtype: int64

Missing values %

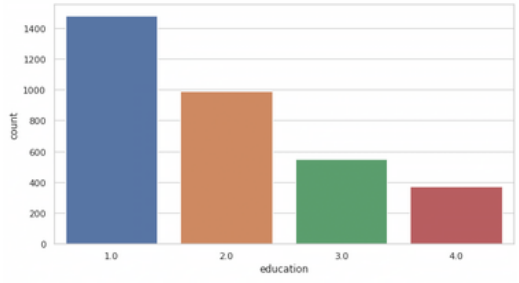
glucose	8.967552
education	2.566372
BPMeds	1.297935
totChol	1.120944
cigsPerDay	0.648968
BMI	0.412979
heartRate	0.029499
age	0.000000
sex	0.000000
is_smoking	0.000000
prevalentStroke	0.000000
prevalentHyp	0.000000
diabetes	0.000000
sysBP	0.000000
diaBP	0.000000
TenYearCHD	0.000000

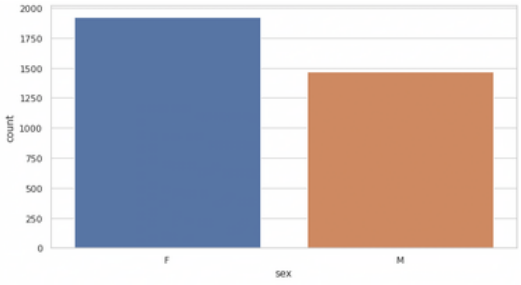
dtype: float64

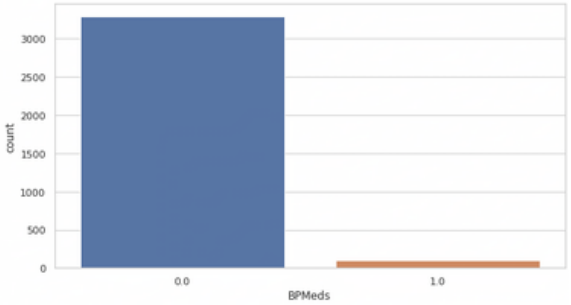
OUTLIERS DETECTION:

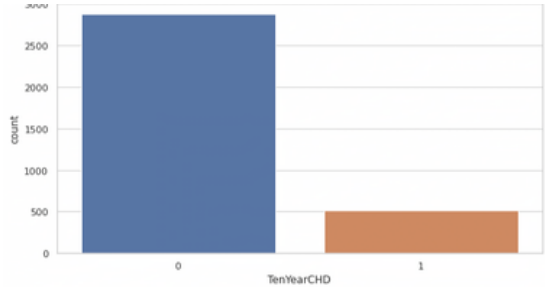


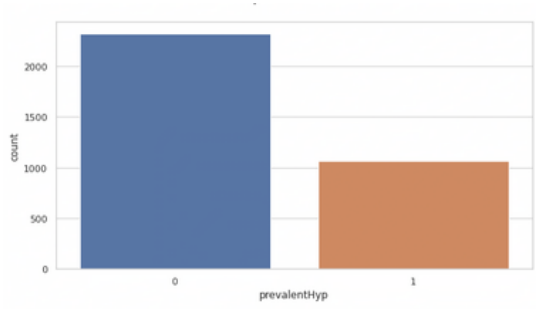
Bivariate Analysis:

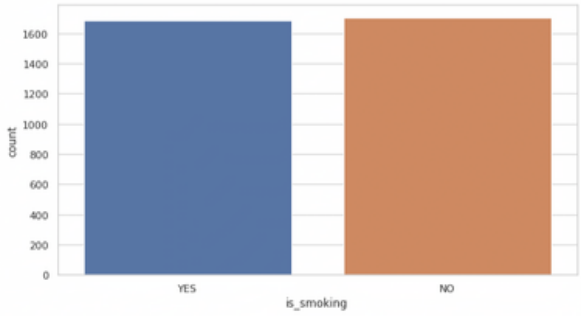






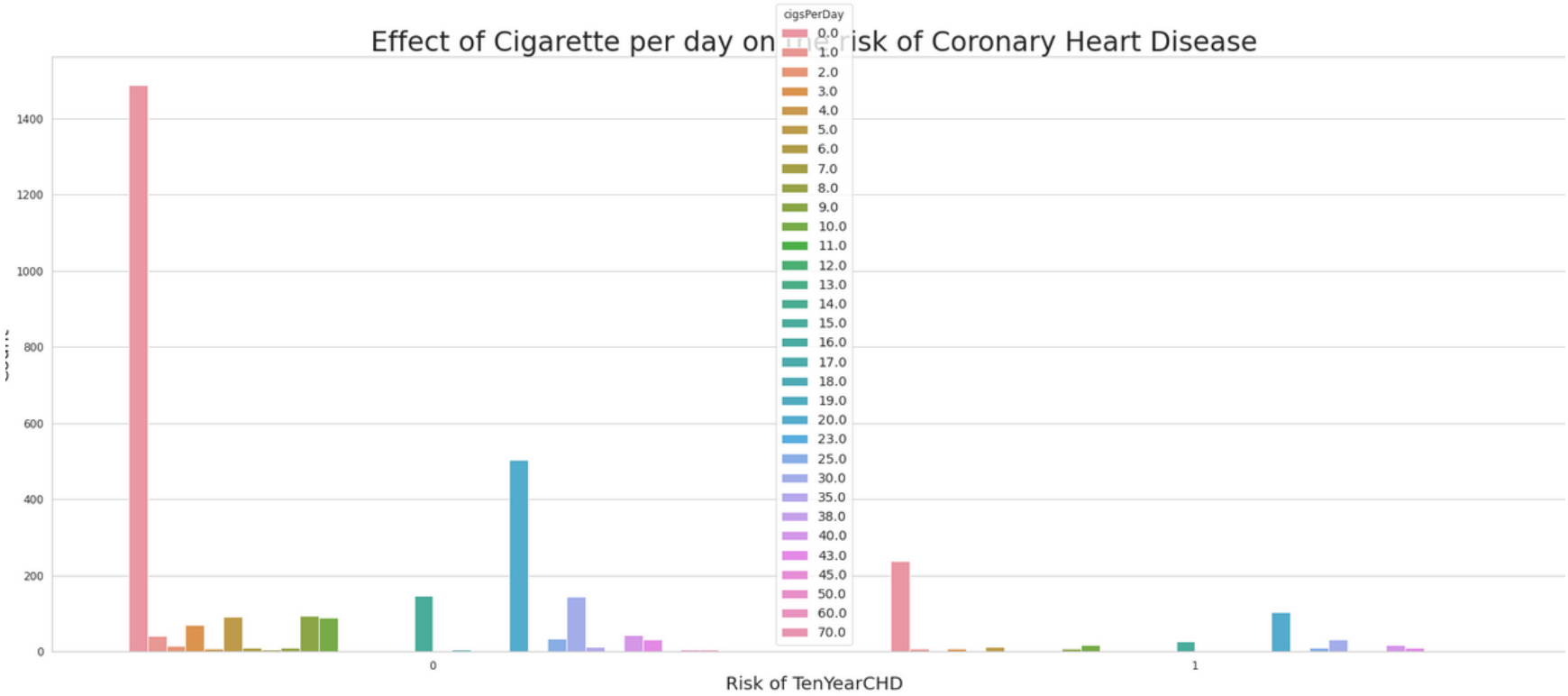






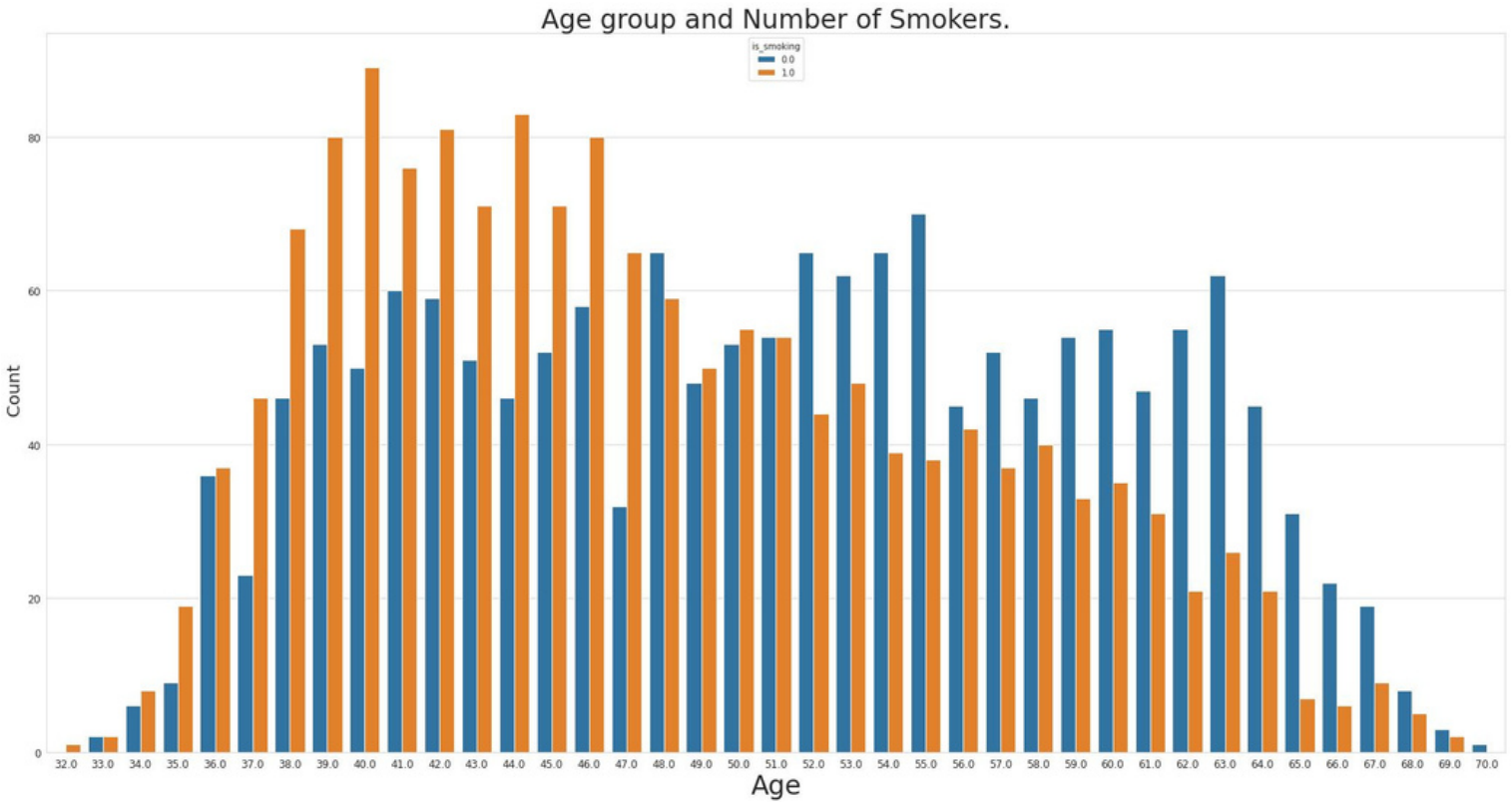
EDA

(BIVARIATE ANALYSIS)



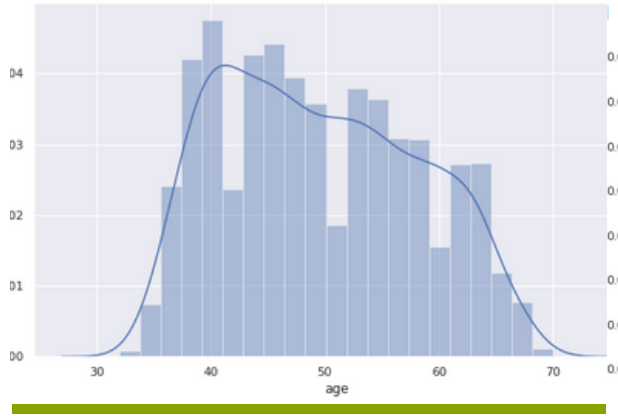
EDA

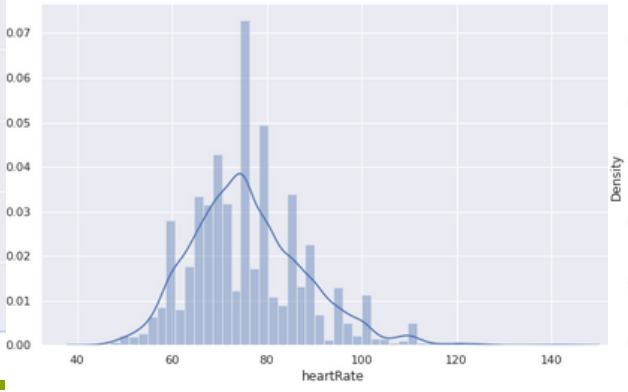
(BIVARIATE ANALYSIS)

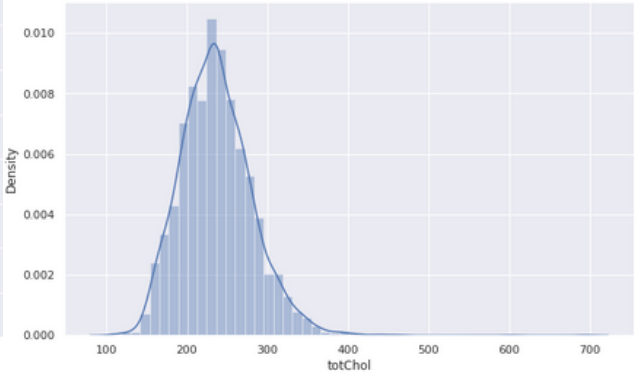


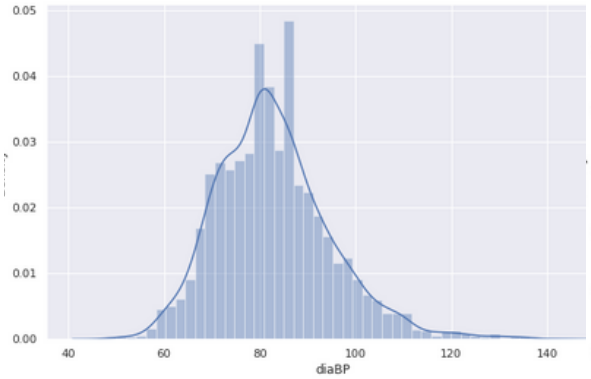
EDA

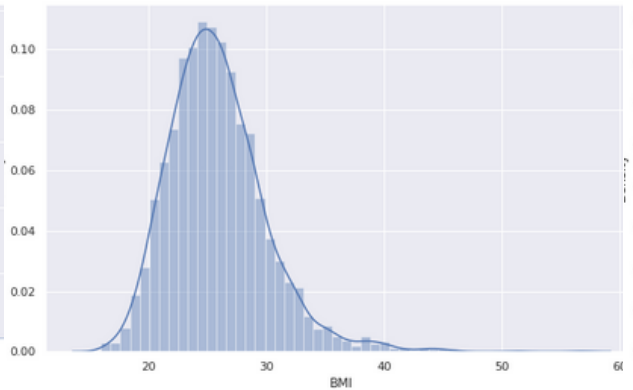
Bi-variate Analysis

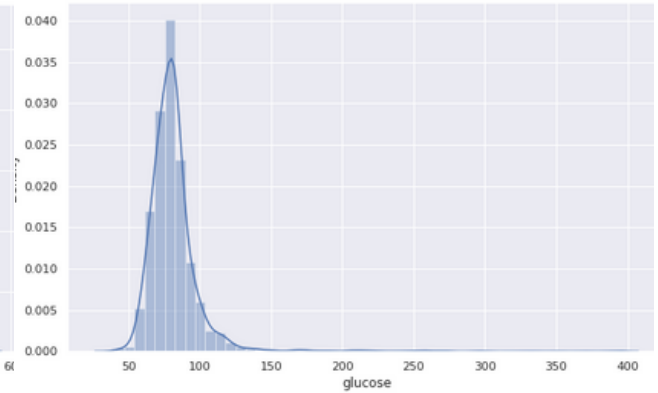










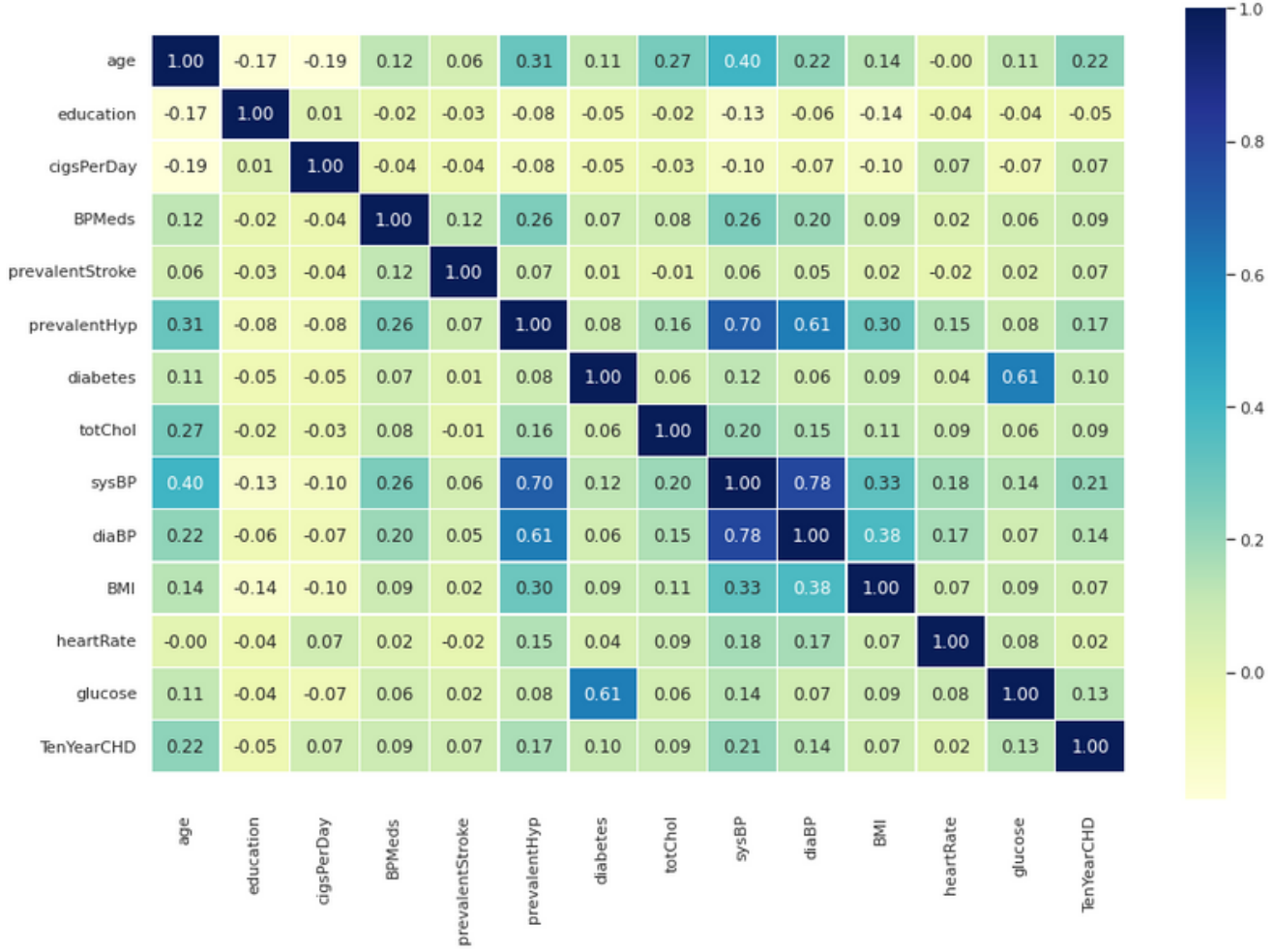


Histogram of several numerical columns

Findings of Bivariate Analysis

- In the middle age group, the number of current smokers starts to rise
- Coronary heart disease is slightly more common in females
- A very small number of people were smoking at the age of 32, or perhaps none at all.
- There is a reduction in CHD risk with fewer cigarettes per day
- It is normal for systolic blood pressure to be less than 120. As systolic blood pressure rises above this range, the risk of coronary heart disease increases.
- Higher cholesterol levels are associated with higher chances of heart disease.
- A diastolic blood pressure reading of 80mm Hg is considered normal. Coronary heart disease increases with higher diastolic blood pressure values;

Multivariate analysis



Building Model

Before building the models,we perform the traintestsplit.We have taken 70% of the data as train data and 30% of the data as test data. .
There are many classification models available in supervised machine learning.
The models which we have used are,
(1)Logisticregression
(2)RandomForest
(3)K–nearestneighbor
(4) XGB classifier
(5) SVC

CLASSIFICATION MODEL AND MODEL EVALUATION:

CLASSIFICATION METRICS COMPARISON

	Model	Train Accuracy	Test Accuracy	Train Precision	Test Precision	Train Recall	Test Recall	Train ROC AUC	Test ROC AUC
0	Logistic Regression	0.738602	0.738715	0.771608	0.736243	0.686131	0.705455	0.739201	0.737279
1	Random Forest	1.000000	0.907986	1.000000	0.909594	1.000000	0.896364	1.000000	0.907484
2	XGB Classifier	0.875380	0.865451	0.942958	0.917548	0.802061	0.789091	0.876217	0.862153
3	KNN	0.874946	0.823785	0.856736	0.793570	0.903821	0.852727	0.874616	0.825035
4	SVC	0.767043	0.763889	0.824045	0.789583	0.685702	0.689091	0.767972	0.760658

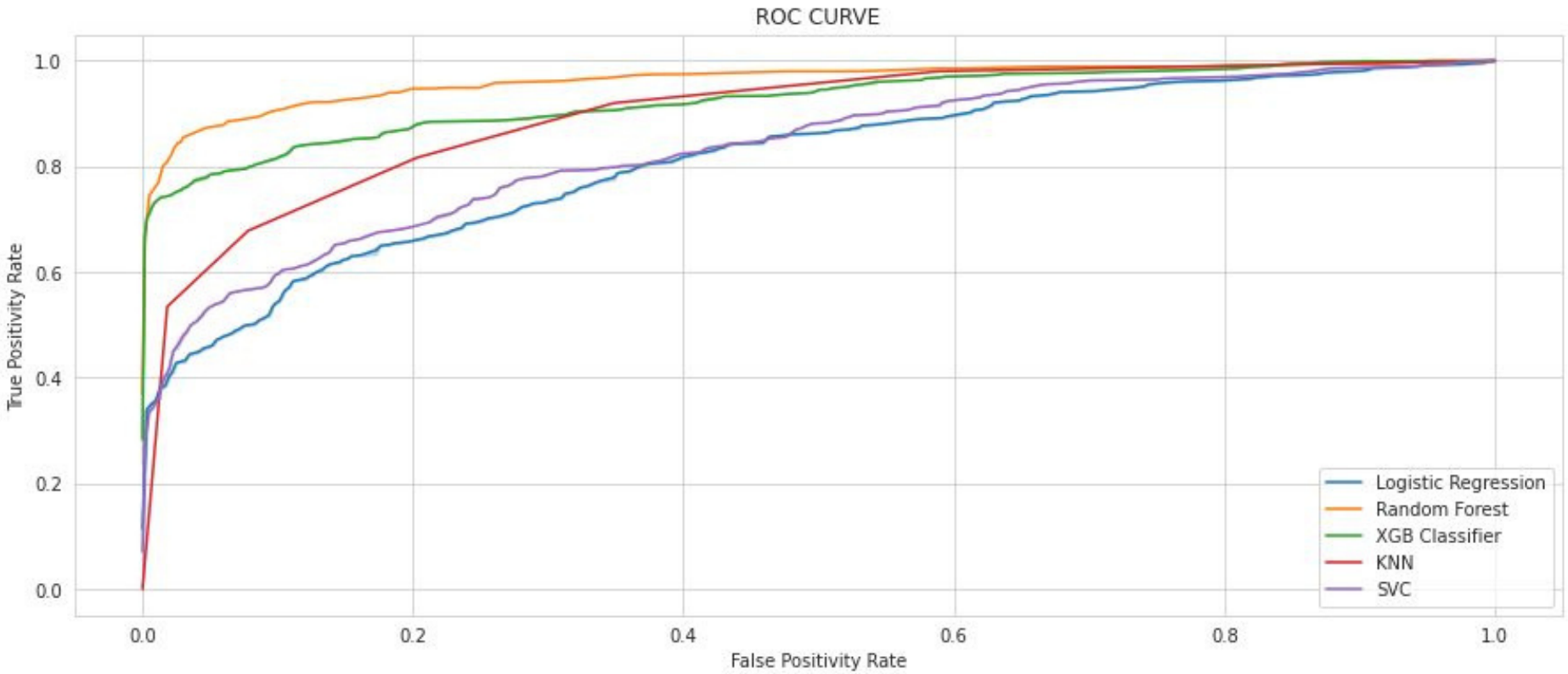
Model Building and Evaluation

- For this project, 5 models have been experimented with.

	Model	Train Accuracy	Test Accuracy	Train Precision	Test Precision	Train Recall	Test Recall	Train ROC AUC	Test ROC AUC
0	Logistic Regression	0.746201	0.727431	0.785714	0.736000	0.684843	0.669091	0.746901	0.724911
1	Random Forest	1.000000	0.912326	1.000000	0.927619	1.000000	0.885455	1.000000	0.911166
2	XGB Classifier	0.881676	0.862847	0.942022	0.901639	0.816230	0.800000	0.882423	0.860133
3	KNN	0.881025	0.806424	0.875897	0.786340	0.890940	0.816364	0.880912	0.806853
4	SVC	0.767477	0.757812	0.830705	0.801782	0.678403	0.654545	0.768494	0.753352

- Random forest is overfitting on train data, but at the same time is performing better than other models on the test data. It is closely followed by XGB Classifier.

Model Building and Evaluation



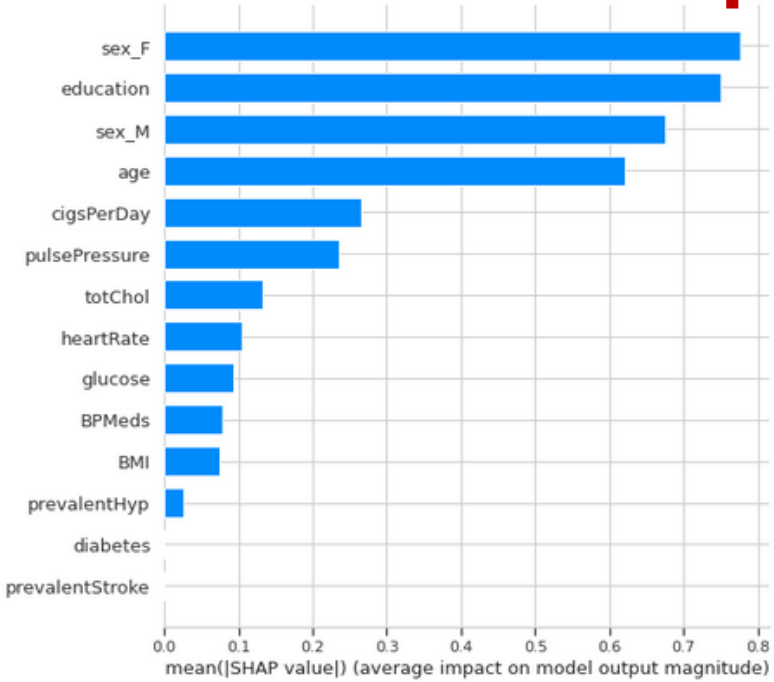
Hyperparameter Tuning

- Hyperparameter tuning was performed on the Random Forest model using Grid Search CV. An attempt was done to reduce overfitting and improve results further.

	Model	Train Accuracy	Test Accuracy	Train Precision	Test Precision	Train Recall	Test Recall	Train ROC AUC	Test ROC AUC
0	Random Forest	0.982197	0.888889	0.983226	0.893657	0.981537	0.870909	0.982205	0.888112

- The result of the hyperparameter tuning explain that, overfitting was reduced to a small extent, but results did not improve drastically, the results have remained more or less the same after hyperparameter tuning.

SHAP Feature Importance



- The above figure explains that, features such as gender, age and pulse pressure are highly influencing the possibility of Coronary Heart Disease (CHD).

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.

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.

Conclusion

- A model is developed with almost 89% accuracy, 89% precision and 87% recall on the test data.
- With more knowledge from an expert in the field of cardiovascular health, new variables could be developed to enhance the predictions further.
- Better models other than the ones used here could be used to improve predictions.
- This project has provided experience in an important field of healthcare and has clearly illustrated the application of machine learning in this field.
- Machine learning can help make lives better and with early predictions of diseases, casualties can be avoided.

THANK YOU!