Capstone Project-3

Cardiovascular Risk Prediction

Pratik M Gumble

Point for Discussion

- □ Problem Statement
- □ Introduction
- □ Dataset and Variable Information
- □ Dataset Processing
- □ Data Encoding
- □ EDA
- □ Outliers Removal
- □ Balancing Imbalanced Dataset
- □ Applying Model
- □ Predictive Model
- □ Feature Importance
- □ Conclusion

Problem Statement

The classification goal is to predict whether the patient has a 10-year risk of future coronary heart disease (CHD).

Introduction

Cardiovascular disease is a major health burden worldwide in the 21st century. The information which is gathered by data analysis of hospitals is utilizing by applying different blends of calculations and algorithms for the early-stage prediction of Cardiovascular ailments. Machine Learning is one of the slanting.

Introduction

We will use a dataset to understand how to build different classification models in python from scratch. The models that will be introduced in this project are:

Logistic Regression
Random Forest
XG Boost
K-Nearest Neighbor
SVM

After we build the models using training data, we will test the accuracy of the model with test data and determine the appropriate model for this dataset.

Dataset

The dataset is from an ongoing cardiovascular study on residents of the town of Framingham, Massachusetts. The dataset provides the patients' information. It includes over 4,000 records and 15 attributes.

Dataset Variable Information

Each attribute is a potential risk factor. There are both demographic, behavioral, and medical risk factors.

Data Description

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)

Dataset Variable Information

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.)

Dataset Variable Information

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)

Dataset Variable Information

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)

Dataset Variable Information

Predict variable (desired target)

- **10-year risk of coronary heart disease CHD**

(binary: “1”, means “Yes”, “0” means “No”) - DV

Dataset Inspection & Processing

After data inspection we found that there are 3390 rows × 17 columns available in the dataset so we drop id and education columns and keep 15 columns for analysis.

NaN Values Processing

As number of count of NaN values in cigsPerDay(22),BPMeds(44) totChol(38) and BMI(14),in are not more than 20% so we replace these values with zero instead of mean, median or mode.

Where Number of NaN values in glucose is 304 so we replace with it mean value.

Data Encoding

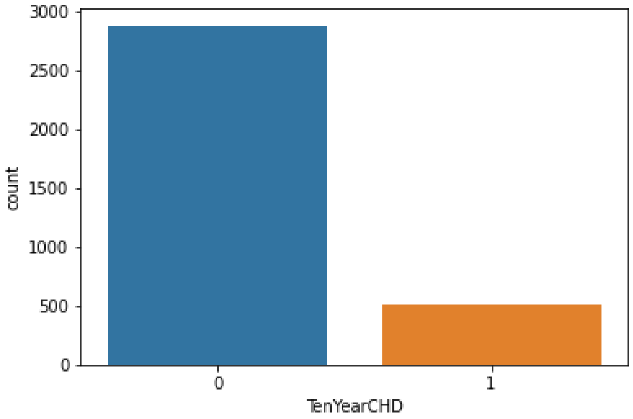
Before we start predicting, an important step to do is to convert our sex and is_smoking feature, which is a string, into integer.

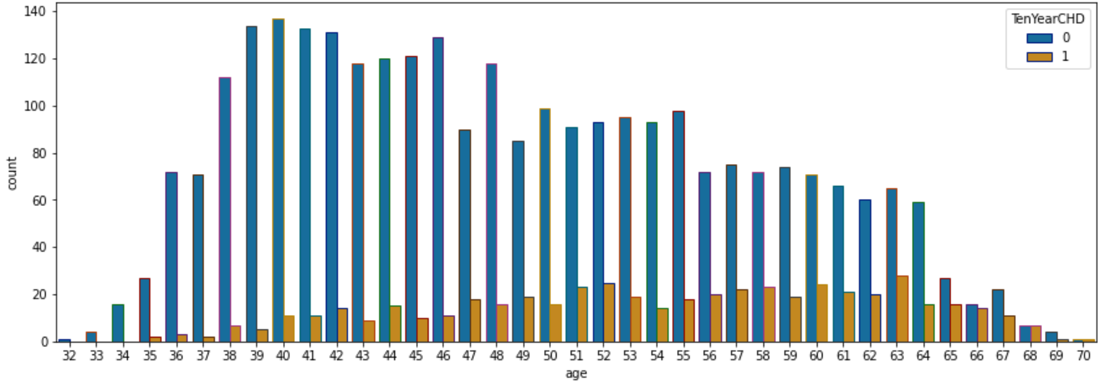
M will be converted to 1 and F will be converted to 0.

Same for YES will be converted to 1 and NO will be converted to 0.

We replaced this columns with same as a “sex" and “is_smoking".

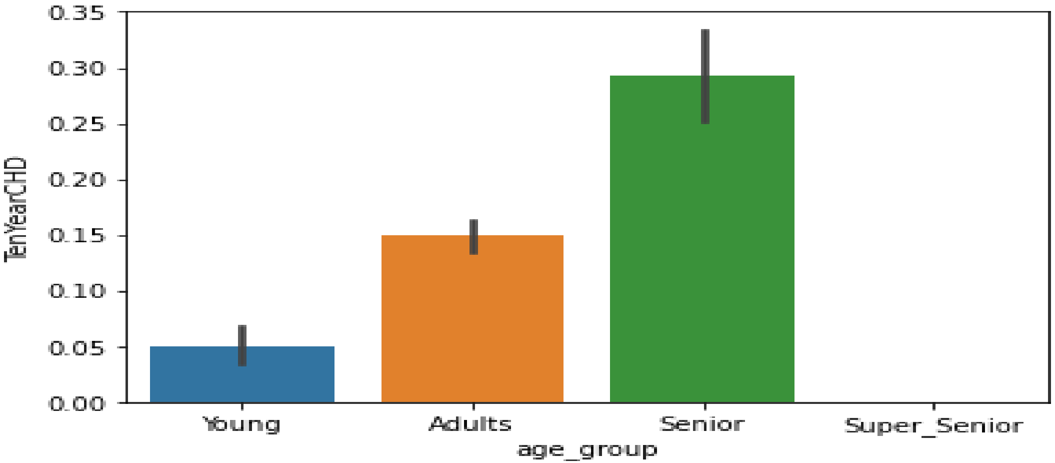
EDA



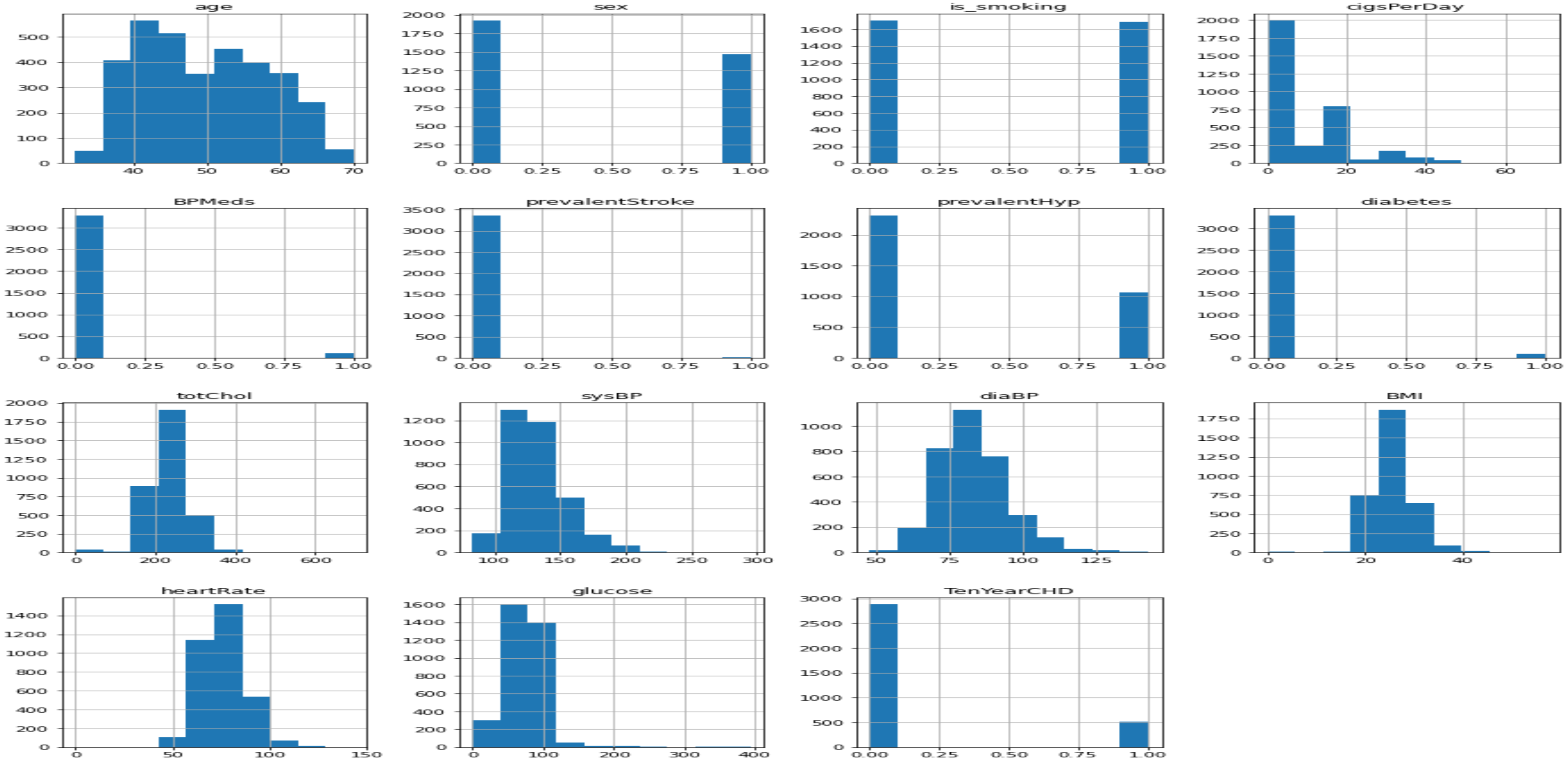


0	2879
1	511

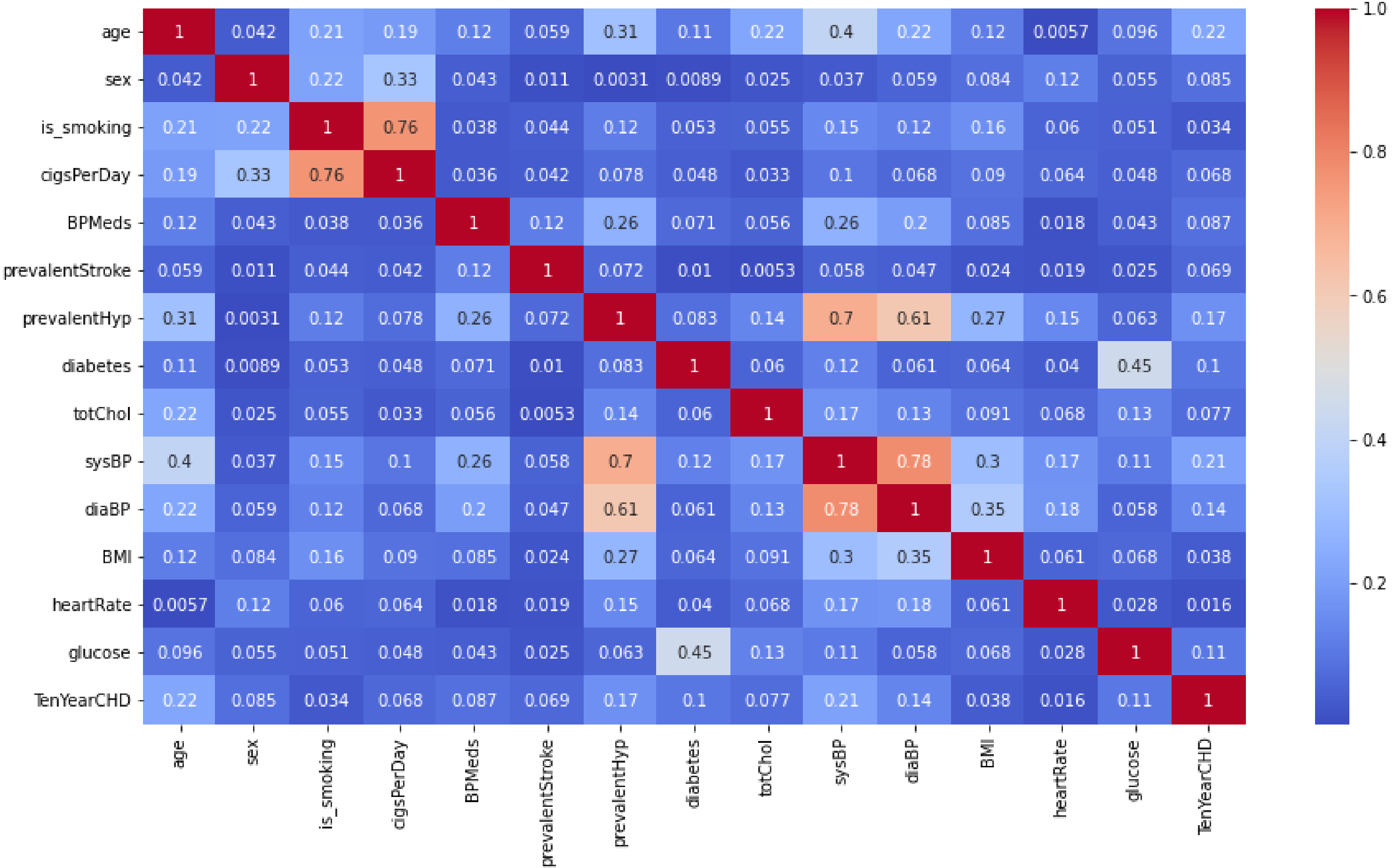
From the graph and statistical analysis it is clear that most of the people in the dataset have not 10-year risk of coronary heart disease CHD(2879 peoples) and 511 peoples have 10-year risk of coronary heart disease CHD



EDA

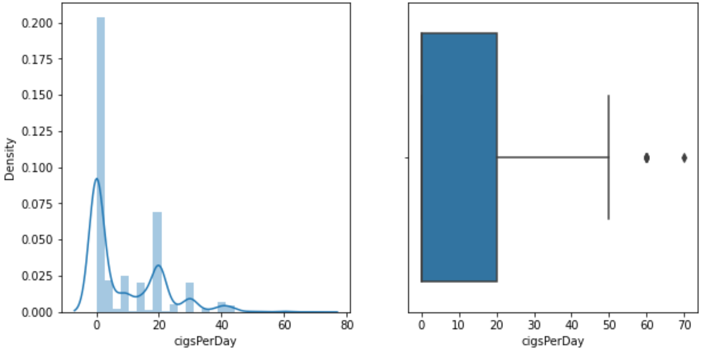

age
0
100
200
300
400
500
40
50
60
70
sex
0
250
500
750
1000
1250
1500
1750
2000
0.00
0.25
0.50
0.75
1.00
is_smoking
0
200
400
600
800
1000
1200
1400
1600
0.00
0.25
0.50
0.75
1.00
cigsPerDay
0
250
500
750
1000
1250
1500
1750
2000
0
20
40
60
BPMeds
0
500
1000
1500
2000
2500
3000
0.00
0.25
0.50
0.75
1.00
prevalentStroke
0
500
1000
1500
2000
2500
3000
3500
0.00
0.25
0.50
0.75
1.00
prevalentHyp
0
500
1000
1500
2000
0.00
0.25
0.50
0.75
1.00
diabetes
0
500
1000
1500
2000
2500
3000
0.00
0.25
0.50
0.75
1.00
totChol
0
250
500
750
1000
1250
1500
1750
2000
0
200
400
600
sysBP
0
200
400
600
800
1000
1200
100
150
200
250
300
diaBP
0
200
400
600
800
1000
50
75
100
125
BMI
0
250
500
750
1000
1250
1500
1750
0
20
40
heartRate
0
200
400
600
800
1000
1200
1400
0
50
100
150
glucose
0
200
400
600
800
1000
1200
1400
1600
0
100
200
300
400
TenYearCHD
0
500
1000
1500
2000
2500
3000
0.00
0.25
0.50
0.75
1.00

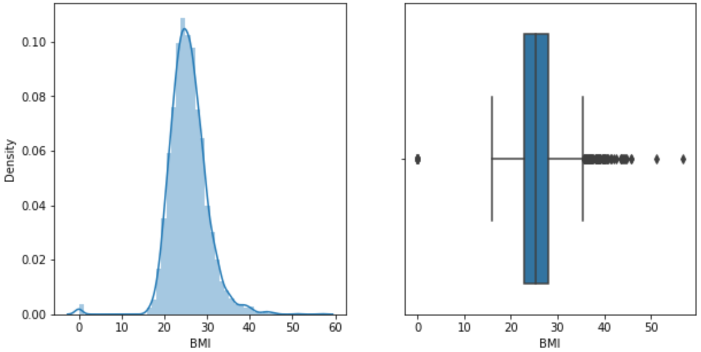
EDA

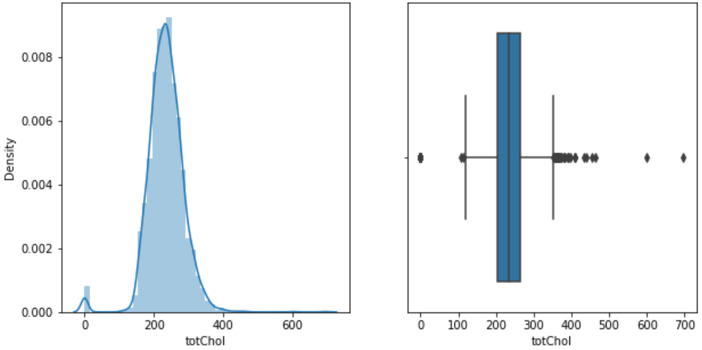


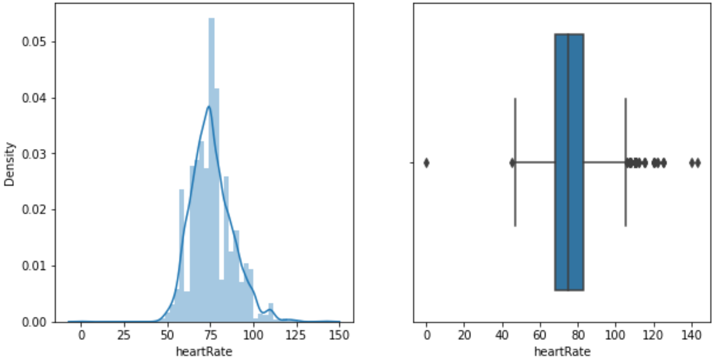
	age	sex	is_smoking	cigsPerDay	BPMeds	prevalentStroke	prevalentHyp	diabetes	totChol	sysBP	diaBP	BMI	heartRate	glucose	TenYearCHD
age	1	0.042	0.21	0.19	0.12	0.059	0.31	0.11	0.22	0.4	0.22	0.12	0.0057	0.096	0.22
sex	0.042	1	0.22	0.33	0.043	0.011	0.0031	0.0089	0.025	0.037	0.059	0.084	0.12	0.055	0.085
is_smoking	0.21	0.22	1	0.76	0.038	0.044	0.12	0.053	0.055	0.15	0.12	0.16	0.06	0.051	0.034
cigsPerDay	0.19	0.33	0.76	1	0.036	0.042	0.078	0.048	0.033	0.1	0.068	0.09	0.064	0.048	0.068
BPMeds	0.12	0.043	0.038	0.036	1	0.12	0.26	0.071	0.056	0.26	0.2	0.085	0.018	0.043	0.087
prevalentStroke	0.059	0.011	0.044	0.042	0.12	1	0.072	0.01	0.0053	0.058	0.047	0.024	0.019	0.025	0.069
prevalentHyp	0.31	0.0031	0.12	0.078	0.26	0.072	1	0.083	0.14	0.7	0.61	0.27	0.15	0.063	0.17
diabetes	0.11	0.0089	0.053	0.048	0.071	0.01	0.083	1	0.06	0.12	0.061	0.064	0.04	0.45	0.1
totChol	0.22	0.025	0.055	0.033	0.056	0.0053	0.14	0.06	1	0.17	0.13	0.091	0.068	0.13	0.077
sysBP	0.4	0.037	0.15	0.1	0.26	0.058	0.7	0.12	0.17	1	0.78	0.3	0.17	0.11	0.21
diaBP	0.22	0.059	0.12	0.068	0.2	0.047	0.61	0.061	0.13	0.78	1	0.35	0.18	0.058	0.14
BMI	0.12	0.084	0.16	0.09	0.085	0.024	0.27	0.064	0.091	0.3	0.35	1	0.061	0.068	0.038
heartRate	0.0057	0.12	0.06	0.064	0.018	0.019	0.15	0.04	0.068	0.17	0.18	0.061	1	0.028	0.016
glucose	0.096	0.055	0.051	0.048	0.043	0.025	0.063	0.45	0.13	0.11	0.058	0.068	0.028	1	0.11
TenYearCHD	0.22	0.085	0.034	0.068	0.087	0.069	0.17	0.1	0.077	0.21	0.14	0.038	0.016	0.11	1

EDA

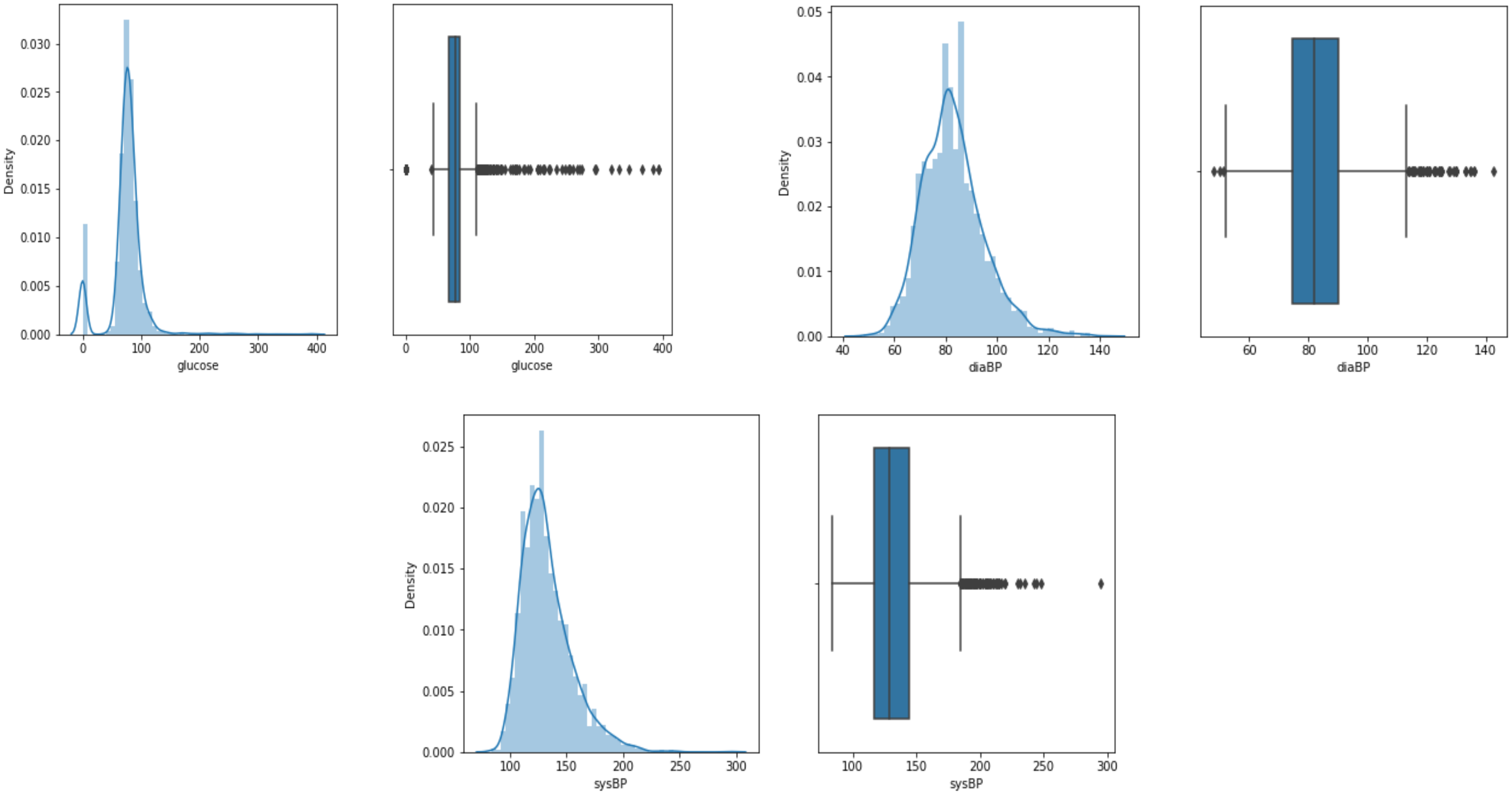

Density
0.200
0.175
0.150
0.125
0.100
0.075
0.050
0.025
0.000
0
20
40
60
80
cigsPerDay
0
10
20
30
40
50
60
70
cigsPerDay


Density
0.10
0.08
0.06
0.04
0.02
0.00
0
10
20
30
40
50
60
BMI
0
10
20
30
40
50
BMI


Density
0.008
0.006
0.004
0.002
0.000
0
200
400
600
totChol
0
100
200
300
400
500
600
700
totChol


Density
0.05
0.04
0.03
0.02
0.01
0.00
0
25
50
75
100
125
150
heartRate
0
20
40
60
80
100
120
140
heartRate

EDA


Density
0.030
0.025
0.020
0.015
0.010
0.005
0.000
0
100
200
300
400
glucose
0
100
200
300
400
glucose
Density
0.05
0.04
0.03
0.02
0.01
0.00
40
60
80
100
120
140
diaBP
60
80
100
120
140
diaBP
Density
0.025
0.020
0.015
0.010
0.005
0.000
100
150
200
250
300
sysBP
100
150
200
250
300
sysBP

Outliers Removal

Interquartile Range (IQR)

After removing outliers using IQR, the data contains (2439, 15)

Z-Score

After using Z-score to detect and remove outliers, the number of records in the dataset is (3046, 15).

As the number of records available is higher after Z-score, we will proceed with z score.

Imbalanced Dataset

Sampling technique comes to save us and deal with imbalanced data.
There are two sampling techniques available to handle the imbalanced data:

Under Sampling
Over Sampling

Here we choose Oversampling technique as Unlike under-sampling, this method leads to no information loss.

Now to overcome data imbalance we apply Oversampling.
In oversampling we use SMOTE.
It stands for Synthetic Minority Oversampling Technique.

```
Original dataset shape Counter({0: 2646, 1: 400})
Resample dataset shape Counter({0: 2646, 1: 2646})
```

Implementation

Now Lets implement 5 models on our dataset:

1. Logistic Regression
2. Random Forrest
3. XGBoost
4. K-NN
5. SVM

1. Logistic Regression

Train accuracy	Test accuracy	Train precision	Test precision	Train recall	Test recall	Train f1 score	Test f1 score	Train ROC-AUC	Test ROC-AUC
67%	68%	67%	65%	70%	71%	68%	68%	67%	68%

2. Random Forest

Train accuracy	Test accuracy	Train precision	Test precision	Train recall	Test recall	Train f1 score	Test f1 score	Train ROC-AUC	Test ROC-AUC
100%	90%	100%	87%	100%	93%	100%	90%	100%	90%

3. XG Boost

Train accuracy	Test accuracy	Train precision	Test precision	Train recall	Test recall	Train f1 score	Test f1 score	Train ROC-AUC	Test ROC-AUC
85%	80%	84%	77%	86%	83%	85%	80%	85%	80%

4. KNN Classifier

Train accuracy	Test accuracy	Train precision	Test precision	Train recall	Test recall	Train f1 score	Test f1 score	Train ROC-AUC	Test ROC-AUC
83%	77%	77%	69%	95%	93%	85%	79%	82%	78%

4. SVM Classifier

Train accuracy	Test accuracy	Train precision	Test precision	Train recall	Test recall	Train f1 score	Test f1 score	Train ROC-AUC	Test ROC-AUC
78%	74%	76%	70%	84%	93%	80%	75%	78%	75%

Comparison

Classifier	Train accuracy	Test accuracy	Train precision	Test precision	Train recall	Test recall	Train f1 score	Test f1 score	Train ROC-AUC	Test ROC-AUC
Logistic Regression	67%	68%	67%	65%	70%	71%	68%	68%	67%	68%
Random Forest	100%	90%	100%	87%	100%	93%	100%	90%	100%	90%
XGBoost	85%	80%	84%	77%	86%	83%	85%	80%	85%	80%
KNN	83%	77%	77%	69%	95%	93%	85%	79%	82%	78%
SVM	78%	74%	76%	70%	84%	93%	80%	75%	78%	75%

It is quite evident from the results Random forest is the best model that can be used for cardiovascular risk prediction dataset since all the performance metrics (accuracy, precision, recall and roc-auc score) show a higher value for the random forest model !

Predictive Model

Predictive model help us to understand possible future occurrences by analyzing our existance model.
Predictive models make assumptions based on what has happened in the past and what is happening now.
If incoming, new data shows changes in what is happening now, the impact on the likely future outcome must be recalculated, too.

In our predictive model we can see that Random forest classifier predict well as patient having **10-year risk of coronary heart disease or not.**

Classifier	Actual	Prediction	Remarks
Logistic Regression	1	1	The Person have Cardiovascular Risk for next 10 years
Random Forest	1	1	The Person have Cardiovascular Risk for next 10 years
XG Boost	1	1	The Person have Cardiovascular Risk for next 10 years
KNN Classifer	1	1	The Person have Cardiovascular Risk for next 10 years

Classifier	Actual	Prediction	Remarks
Logistic Regression	0	1	The Person have Cardiovascular Risk for next 10 years
Random Forest	0	0	The Person does not have Cardiovascular Risk for next 10 years
XG Boost	0	1	The Person have Cardiovascular Risk for next 10 years
KNN Classifer	0	1	The Person have Cardiovascular Risk for next 10 years

Feature Importance

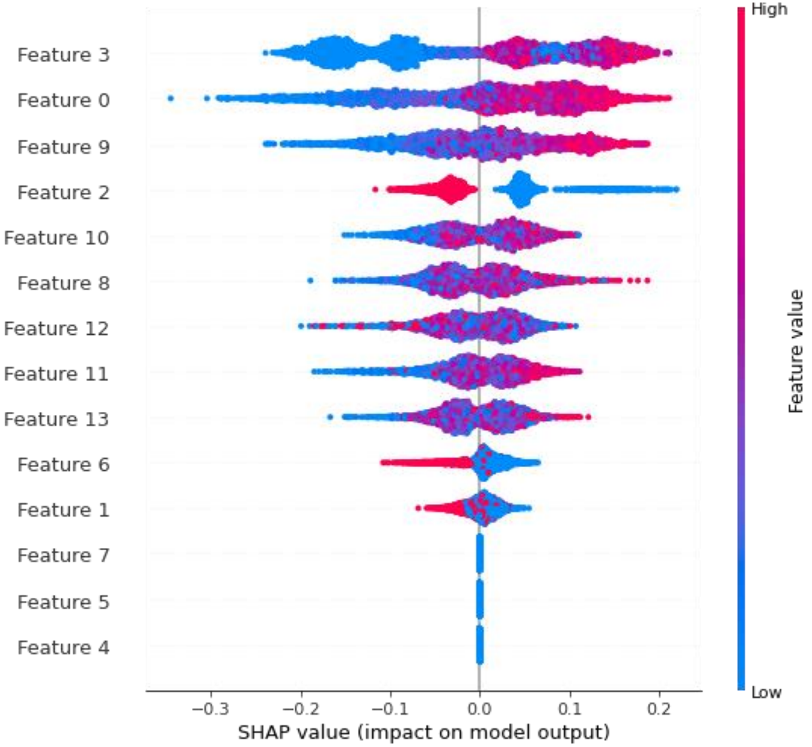
Feature (variable) importance indicates how much each feature contributes to the model prediction. Basically, it determines the degree of usefulness of a specific variable for a current model and prediction.

By analyzing variable importance scores, we would be able to find out irrelevant features and exclude them. Reducing the number of not meaningful variables in the model may speed up the model or even improve its performance.

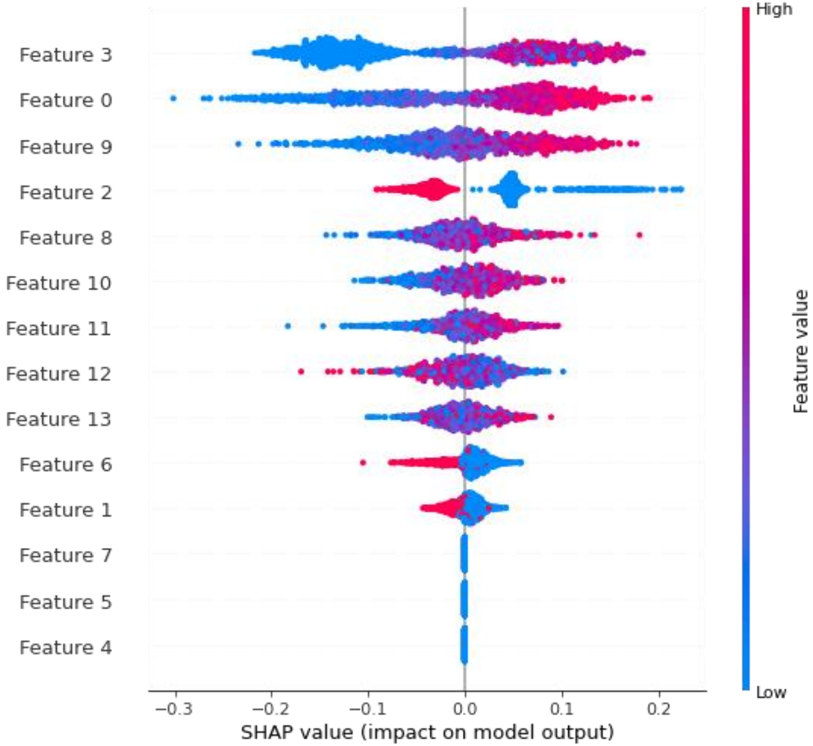
In almost all classifier CigsPerDay, age, sysBP, diaBP, tatchol, heartrate, BMI, glucose are the important feature to determine person having 10 year risk for heart diastase or not. Out of that CigsPerDay, age and BP are the most important feature which decided the person having chance of 10 year heart diseases or not.

Feature Importance

Training



Testing



Conclusion

The main aim of this project is to compare the accuracy and other parameters like precision, recall ,f1 score, auc roc of all the classification algorithms to evaluate the risk of 10-year CHD using 14 features. After implementing four classification models and comparing their accuracy and other scores, we can conclude that for this dataset Random forest Classifier is the appropriate model to be used. Also out of all features CigsPerDay, age and BP are the most important feature which decided the person having chance of 10 year heart diseases or not.

Thank You