Capstone Project

Supervised ML
(Classification)
On
Cardiovascular Risk Prediction

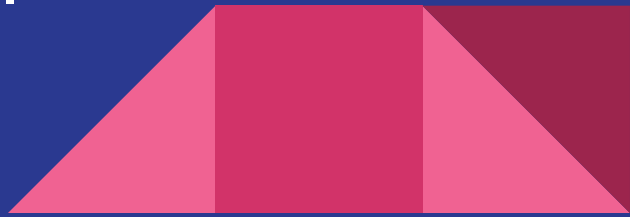
Team Members

Arvind Krishna

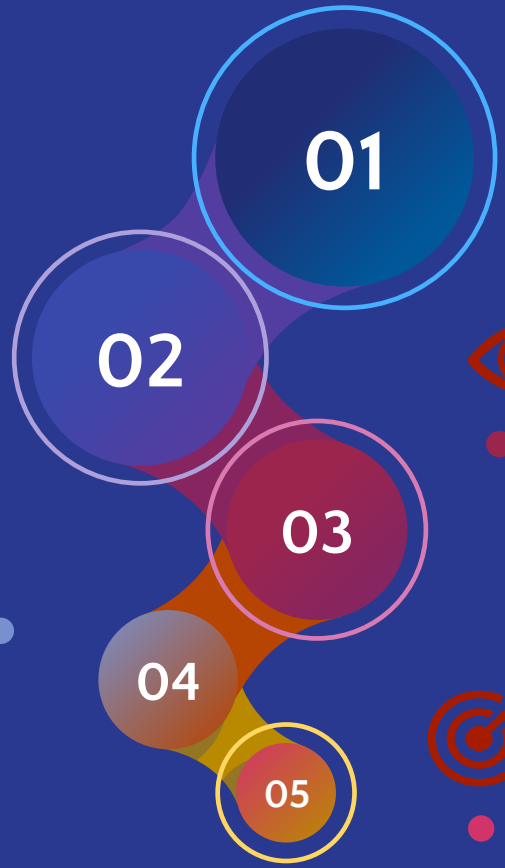
Keshav Sharma

Sahil Ahuja

Jayesh Panchal



ROADMAP TO PRESENTATION



Introduction

A brief note about the project.

Data Cleaning

Fixing the given data set and making it ready for the next process.

EDA and Feature Engineering

Analysing and manipulating data into features for supervised learning.

Model Development.

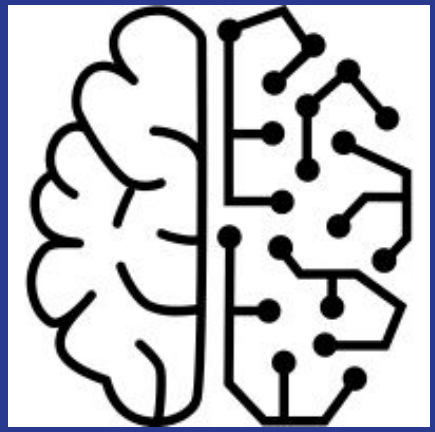
Checking the accuracy with different algorithms.

Conclusion and References

Introduction

What is machine learning ?

Machine learning (ML) is a subset of artificial intelligence (AI) which allows applications to become more accurate in predicting outcomes without being explicitly programmed to do so.



Machine learning algorithms use historical data as input to predict new output values.

Supervised ML (Classification)

Classification is the supervised machine learning technique which is used to predict the discrete values.

Using this technique we are going to predict the Cardiovascular Risk based on the available data.

About the Data

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

Objective

The classification goal is to predict whether the patient has a 10-year risk of future Coronary Heart Disease (CHD).

Data Cleaning

Data cleaning

Data cleaning is an important early step in the data analytics process in which you either remove or update information that is incomplete, incorrert, improperly formatted, duplicated, or irrelevant .

data_cardiovascular_risk.csv

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 smokes ("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)

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)

Predict variable (Desired Target)

10-year risk of coronary heart disease CHD(binary: “1”, means “Yes”, “0” means “No”) -

EDA
&
Feature Engineering

Analysis is done using systematic methods to look for trends, groupings, or other relationships between different types of data.

We then use ‘Data visualization’, to create charts, graphs, or other forms of visualization, which makes information easier to analyse and interpret.

Feature Engineering

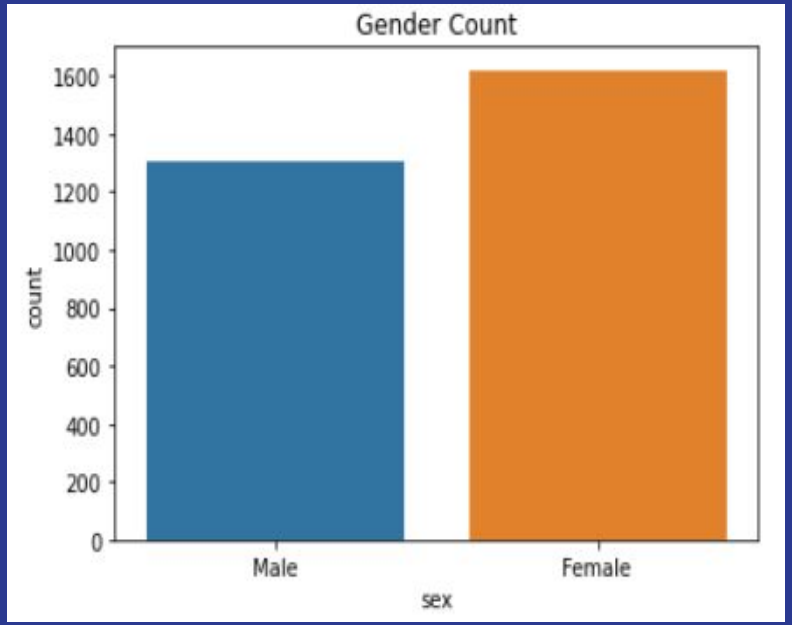
Feature engineering refers to manipulation — addition, deletion, combination, mutation — of the data set to improve machine learning model training, leading to better performance and greater accuracy.

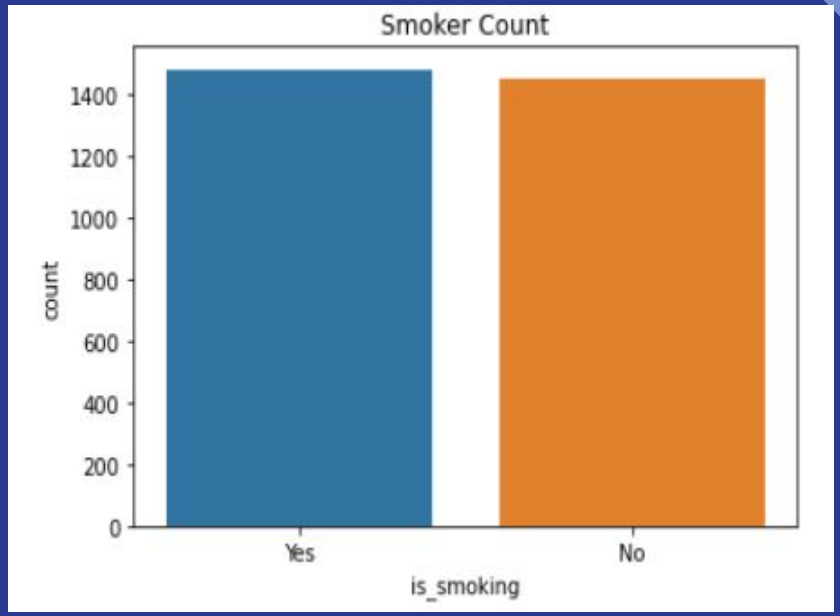
Feature Engineering consists of :

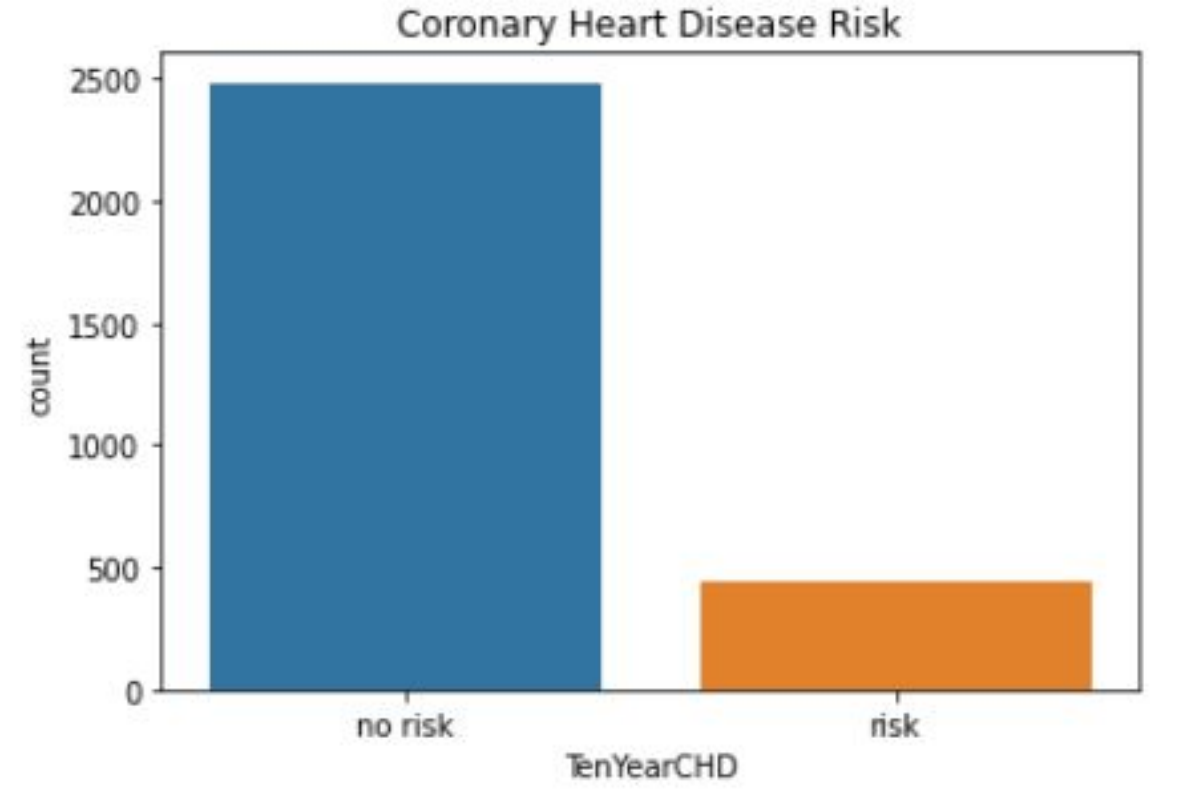
- Feature creation - creating features involves creating new variables.
- Transformation - transformation of features from one representation to another.
- Feature extraction - extracting feature from dataset to identify useful information.
- Benchmark Model - for comparing the performances between different machine learning models.

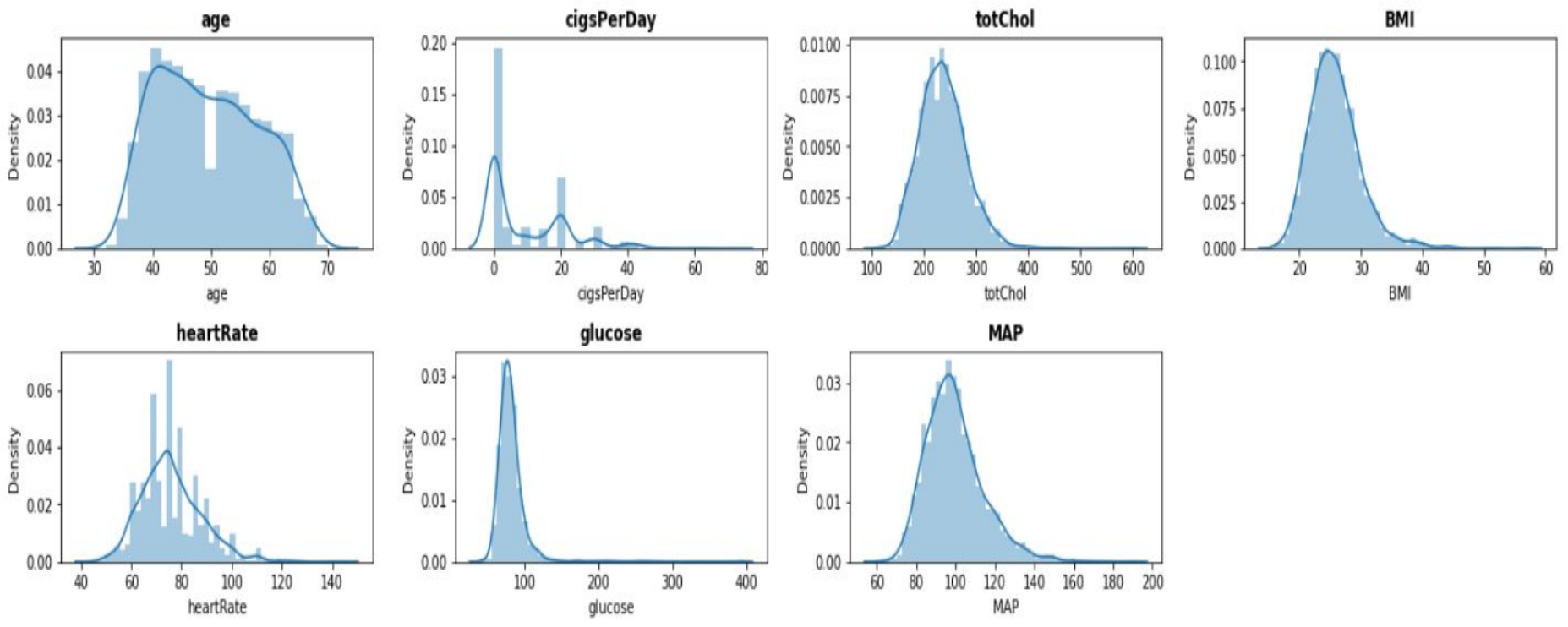
Some new derived features :

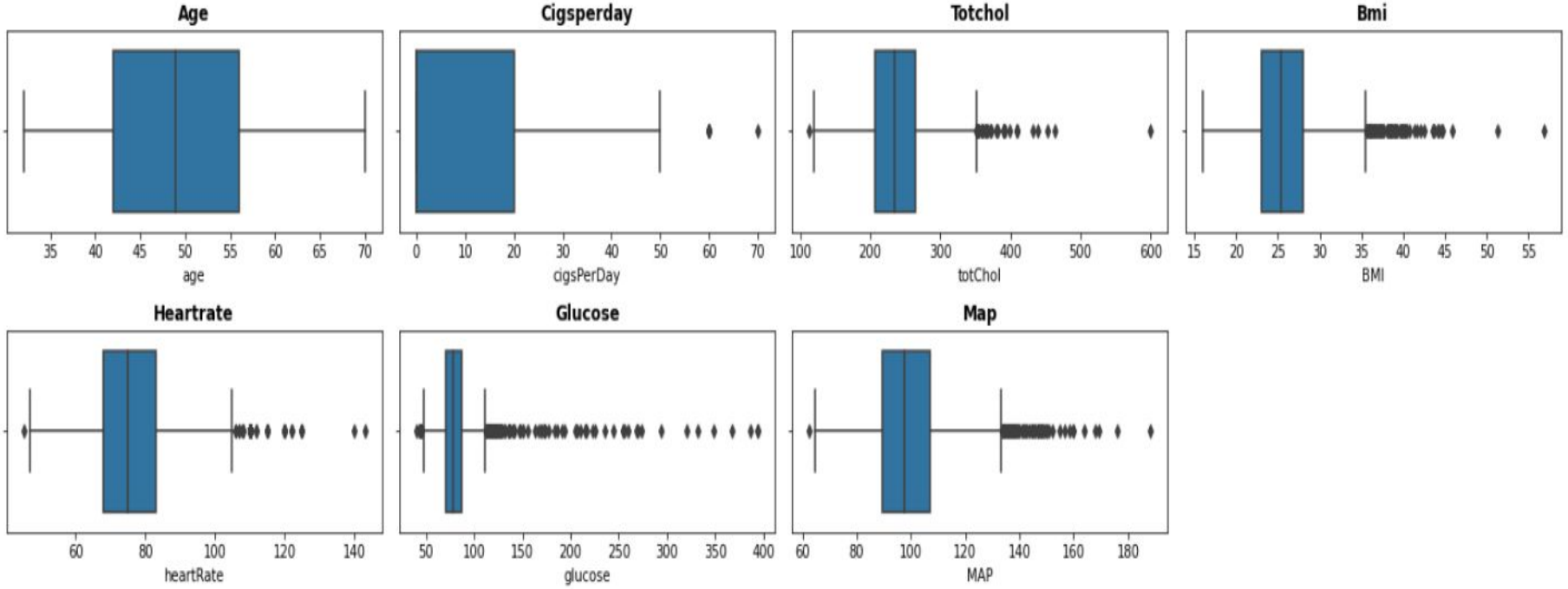
- MAP : Mean Arterial Pressure
  [sysBP + (2*diaBP)] / 3

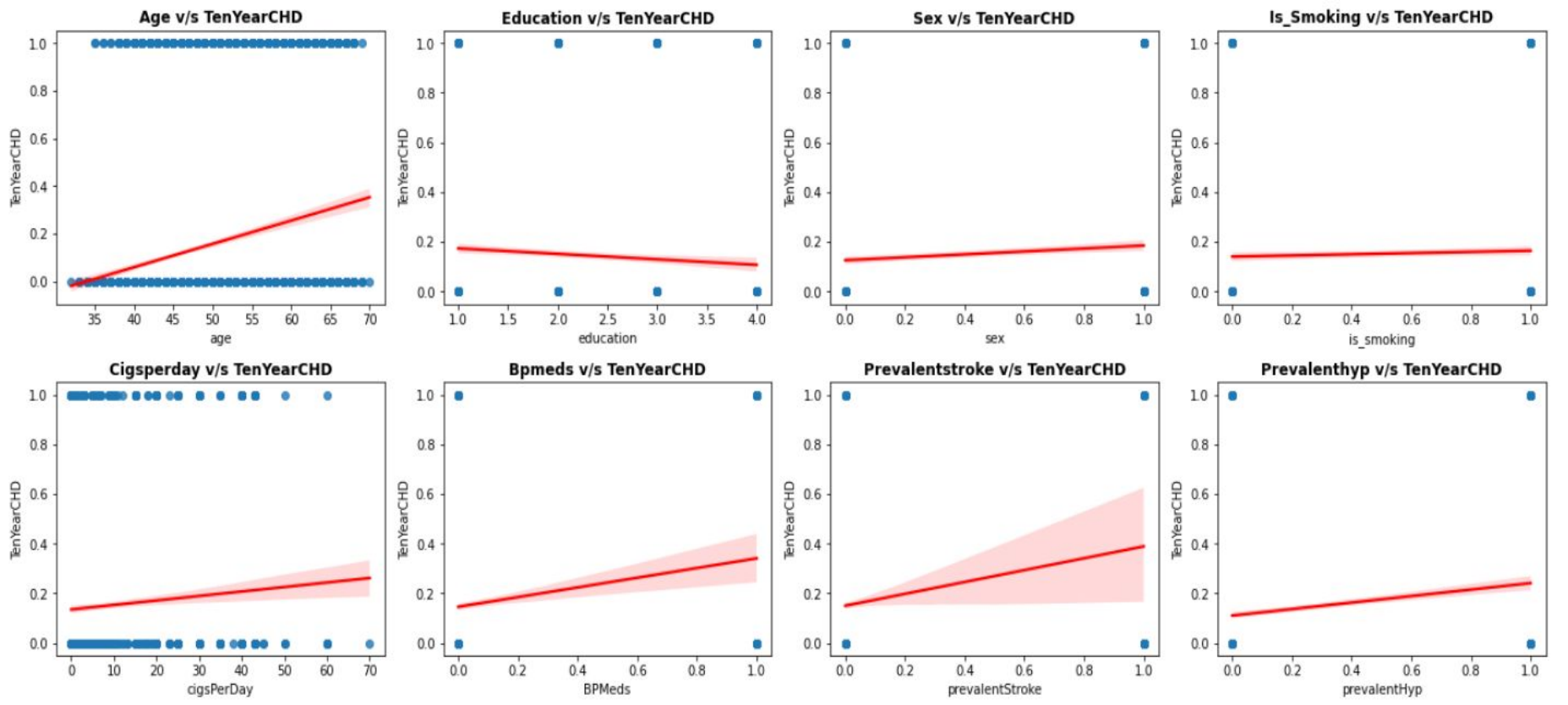

Gender Count
1600
1400
1200
1000
800
600
400
200
0
count
Male
Female
sex

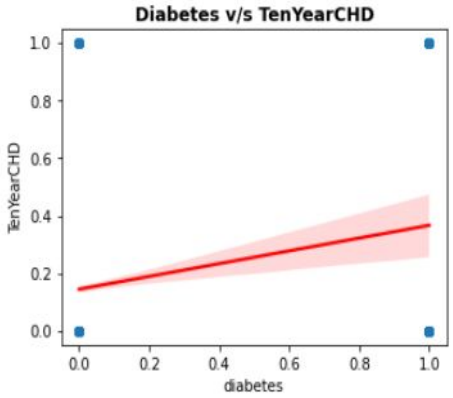

Smoker Count
1400
1200
1000
800
600
400
200
0
count
Yes
No
is_smoking

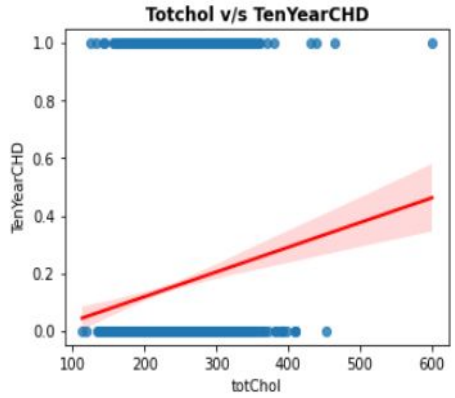

Coronary Heart Disease Risk
2500
2000
1500
1000
500
0
count
no risk
risk
TenYearCHD

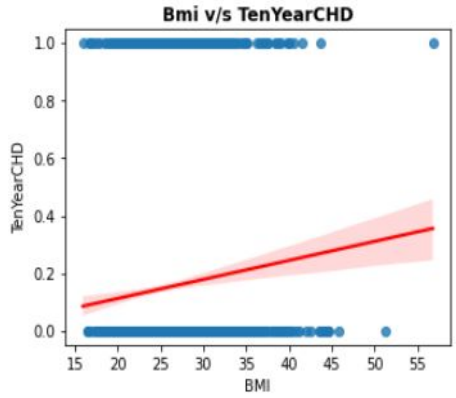

age
Density
0.04
0.03
0.02
0.01
0.00
30
40
50
60
70
age
cigsPerDay
Density
0.20
0.15
0.10
0.05
0.00
0
20
40
60
80
cigsPerDay
totChol
Density
0.0100
0.0075
0.0050
0.0025
0.0000
100
200
300
400
500
600
totChol
BMI
Density
0.100
0.075
0.050
0.025
0.000
20
30
40
50
60
BMI
heartRate
Density
0.06
0.04
0.02
0.00
40
60
80
100
120
140
heartRate
glucose
Density
0.03
0.02
0.01
0.00
100
200
300
400
glucose
MAP
Density
0.03
0.02
0.01
0.00
60
80
100
120
140
160
180
200
MAP

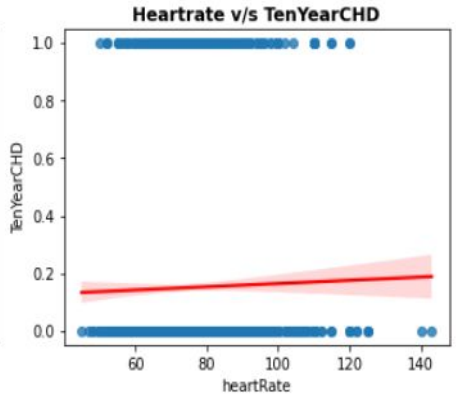

Age
35 40 45 50 55 60 65 70
age
Cigsperday
0 10 20 30 40 50 60 70
cigsPerDay
Totchol
100 200 300 400 500 600
totChol
Bmi
15 20 25 30 35 40 45 50 55
BMI
Heartrate
60 80 100 120 140
heartRate
Glucose
50 100 150 200 250 300 350 400
glucose
Map
60 80 100 120 140 160 180
MAP

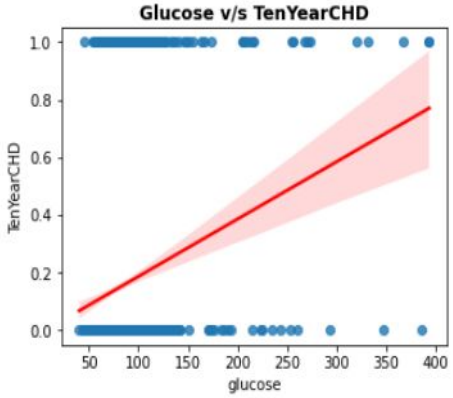

Age v/s TenYearCHD
TenYearCHD
age
Education v/s TenYearCHD
TenYearCHD
education
Sex v/s TenYearCHD
TenYearCHD
sex
Is_Smoking v/s TenYearCHD
TenYearCHD
is_smoking
Cigsperday v/s TenYearCHD
TenYearCHD
cigsPerDay
Bpmeds v/s TenYearCHD
TenYearCHD
BPMeds
Prevalentstroke v/s TenYearCHD
TenYearCHD
prevalentStroke
Prevalenthyp v/s TenYearCHD
TenYearCHD
prevalentHyp

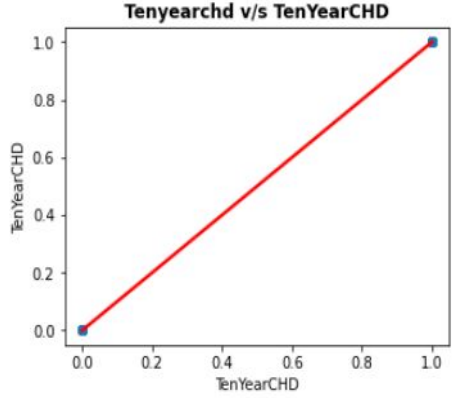

Diabetes v/s TenYearCHD
TenYearCHD
diabetes

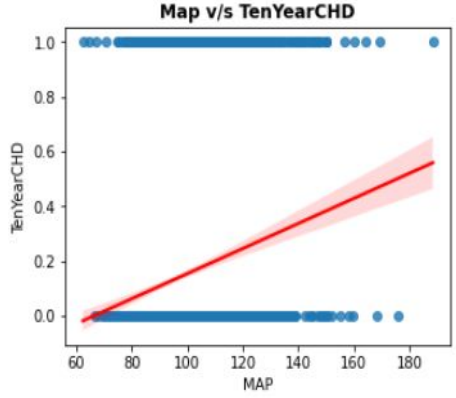

Totchol v/s TenYearCHD
TenYearCHD
totChol


Bmi v/s TenYearCHD
TenYearCHD
BMI

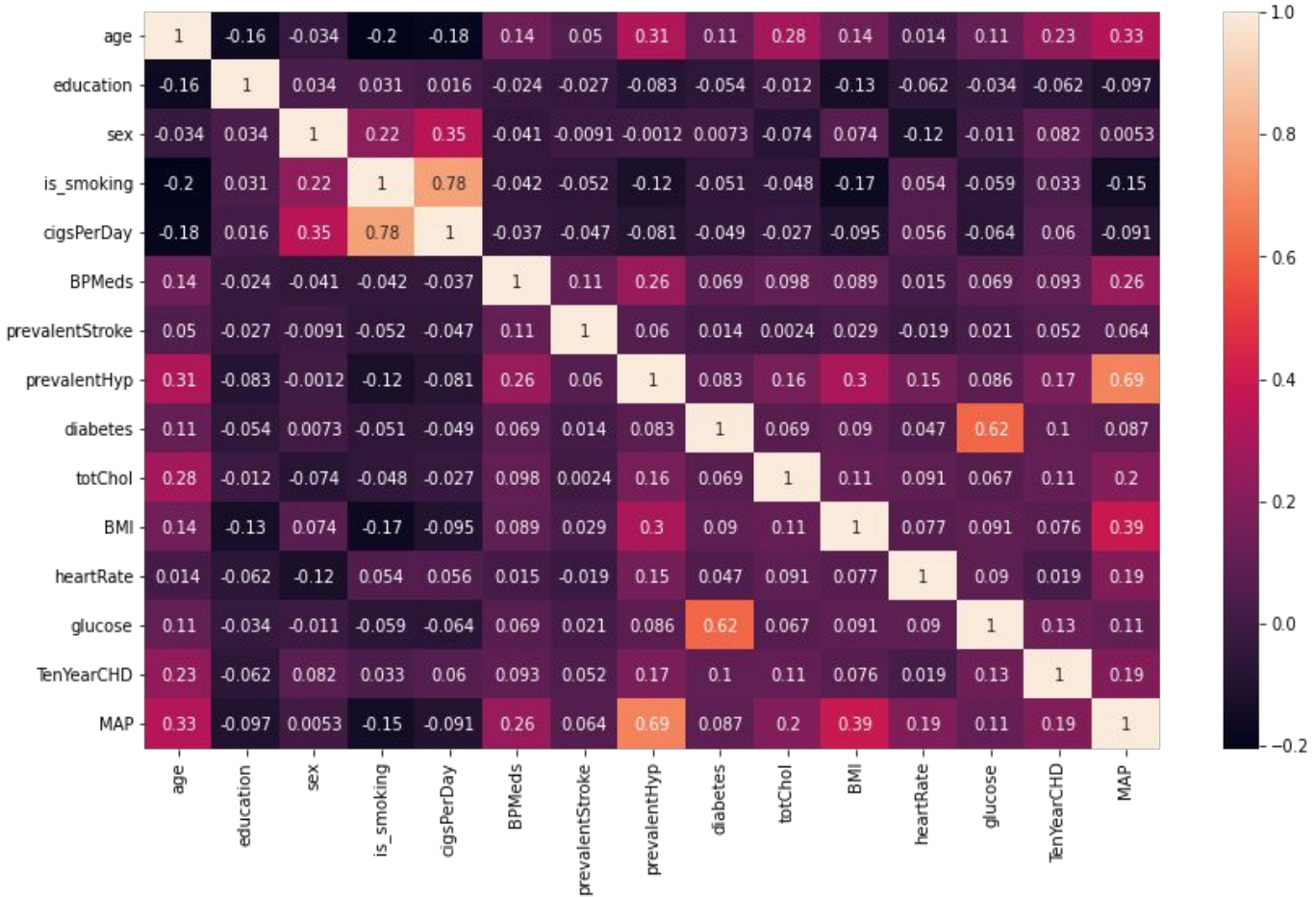

Heartrate v/s TenYearCHD
TenYearCHD
heartRate


Glucose v/s TenYearCHD
TenYearCHD
glucose


Tenyearchd v/s TenYearCHD
TenYearCHD
TenYearCHD


Map v/s TenYearCHD
TenYearCHD
MAP

Correlation Matrix



Model Development

Algorithm ?

A machine learning algorithm is the method by which the AI system conducts its task, generally predicting output values from given input data.
The two main processes of machine learning algorithms are classification and regression.

Algorithm for Classification

- K-Nearest Neighbours
- Gaussian Naive Bayes
- Logistic Regression
- Scalar Vector Machine
- XGB Classifier
- Decision Tree
- Random Forest
- Artificial Neural Network

Results Observed

	precision	recall	f1-score	support	Model
accuracy	0.656357	0.656357	0.656357	0.656357	Logistic Classifier
accuracy	0.742268	0.742268	0.742268	0.742268	K Nearest Neighbours
accuracy	0.781787	0.781787	0.781787	0.781787	Gaussian Naive Bayes Classifier
accuracy	0.640893	0.640893	0.640893	0.640893	Support Vector Machine
accuracy	0.828179	0.828179	0.828179	0.828179	XGB CLassifier
accuracy	0.826460	0.826460	0.826460	0.826460	Random Forest Classifier
accuracy	0.768041	0.768041	0.768041	0.768041	Decision Tree Classifier
accuracy	0.785223	0.785223	0.785223	0.785223	Neural network(Experimental)

Conclusion

Since the dataset we worked with, had a class imbalance, we noticed the models being biased towards the majority class by default. The biased models did deliver good accuracies……..

However, since the learning process was biased, we implemented **SMOTE (Synthetic Minority Oversampling Technique)** to handle the class imbalance, and noticed that our models performed with a significantly low bias after implementing **SMOTE.** Our results are, therefore, based on balanced train data via the use of **SMOTE**.

We've noticed that the **XGB Classifier** is the stand-out performer amongst all implemented models, with an f1-score of **0.828**.

In case of **Logistic regression**, We were able to see the maximum f1-score of **0.656**, also in case of **K-Nearest Neighbors**, the f1-score extends upto **0.742**.

The **Naive Bayes Classifier** showed a balanced result amongst the models we have implemented. It has a f1-score of **0.781**, which is neutral with regards to our observations across various models. But in case of **SVM(Support Vector Machines) Classifier**, the f1-score lies around **0.64**, which also happens to be the lowest score among all models we've implemented.

Out of the tree-based algorithms, the **Random Forest Classifier** was providing an optimal solution, achieving an f1-score of **0.826**. We also noticed that in the case of **Decision-tree Classifier**, we were able to achieve an f1-score of **0.768**.

We have also implemented a experimental **Neural network** model, and the results look promising, albeit inconclusive.

References

- Python Pandas Documentation
https://pandas.pydata.org/pandas-docs/stable

- Python MatPlotLib Documentation
https://matplotlib.org/stable/index.html

- Python Seaborn Documentation
https://seaborn.pydata.org

- Python SkLearn Documentation
https://scikit-learn.org/stable

- Keras API Official Documentation:
https://keras.io/api/

Thank You