Cardiovascular Risk Prediction

Vikas Chaudhary
Data Science trainee,
AlmaBetter, Bengaluru

Abstract: Like any other sector, the health sector is also going through a phase of disruption with the confluence of old practices and the power of AI/ML. Among all the health-related issues Cardiovascular diseases (CVDs) are among the biggest concern for humanity because of its sudden and irremediable impacts. Any piece of technology that can help in reducing the burden of casualties from CVDs will be a game changer in the field of Medicine.

Keywords: Python, EDA, Feature Engineering, ML, Classification, K-Nearest Neighbour, Decision Tree Classifier, Logistic Regression Classifier, Naïve Bayes Classifier, Support Vector Classifier (SVC), Random Forest Classifier, ROC_AUC, etc.

Problem Statement:

The given dataset is from an ongoing cardiovascular study on residents of the town of Framingham, Massachusetts (US). It has 3390 rows and 17 columns (16 independent and 1 dependent). The objective of the project is to predict if the patient has 10-year risk of future coronary heart disease (CHD) or cardiovascular risk using Machine Learning (ML) classification algorithms.

Below are the different features presented in the given dataset:

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)

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”)

INTRODUCTION

As per the World Health Organization (WHO), Cardiovascular diseases (CVDs) are the leading cause of death globally, taking an estimated 17.9 million lives each year. CVDs are a group of disorders of the heart and blood vessels and include coronary heart disease, cerebrovascular disease, rheumatic heart disease and other conditions. More than four out of five CVD deaths are due to heart attacks and strokes, and one-third of these deaths occur prematurely in people under 70 years of age.

With the help of medical data and ML tools, the health sector can be revolutionized especially in filling the gap which is there in the conventional health system. This project is just a microcosm of the emerging HealthTech.

HealthTech is going to play an important role in ensuring the idea of Universal Health Coverage (UHC). The new business opportunities in the field of the health sector are going to be data-driven and are largely based on the preventive health care system.

Successful deployment of such tools/services will give some respite to governments and authorities across the globe because it will make both public and coffers healthy.

Tools Used: Google Colab notebook, 'Python' programming language along with Python Libraries.

Steps Involved

Exploratory Data Analysis (EDA) and Feature Engineering

The DataFrame contains 3390 rows and 17 columns. The following steps are involved in EDA.

Null/NaN, Outliers treatment:

The given dataset has 'NaN', 'Null, or missing values is different features, all the null/NaN values are filled with different methods or approaches.

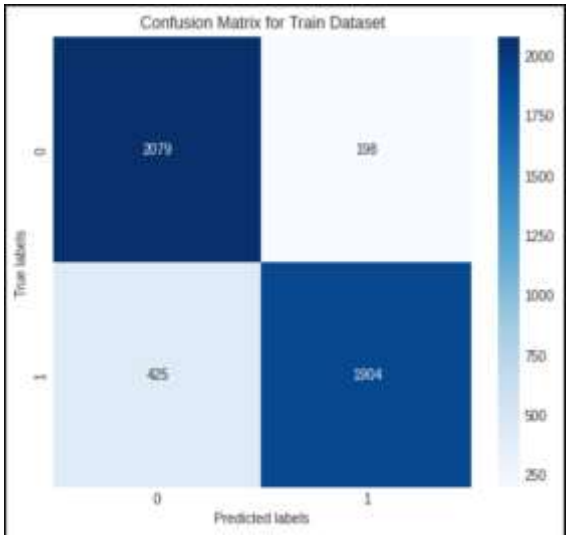
ML Classification Models Implementation

1. **K-Nearest Neighbor Classifier:**

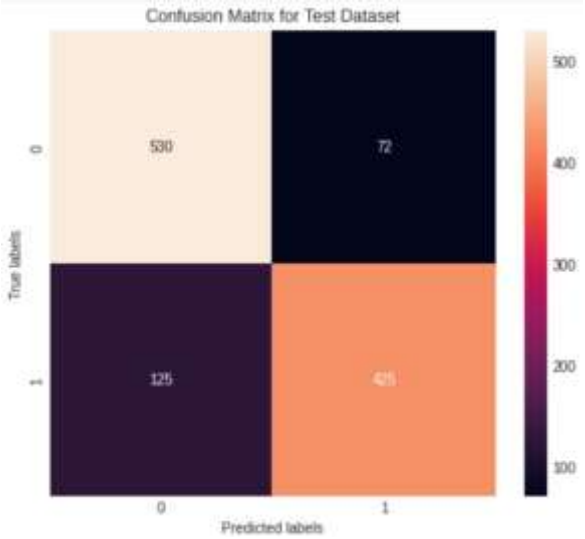
The k-nearest neighbors algorithm, also known as KNN or k-NN, is a non-parametric, supervised learning classifier, which uses proximity to make classifications or predictions about the grouping of an individual data point. While it can be used for either regression or classification problems, it is typically used as a classification algorithm, working off the assumption that similar points can be found near one another.


```
Train Accuracy = 0.8647416413373861
Test Accuracy = 0.8289930555555556
```

Train Dataset



Test Dataset



Classification Report

	precision	recall	f1-score	support
0	0.83	0.91	0.87	2277
1	0.91	0.82	0.86	2329
accuracy			0.86	4606
macro avg	0.87	0.87	0.86	4606
weighted avg	0.87	0.86	0.86	4606

Classification Report

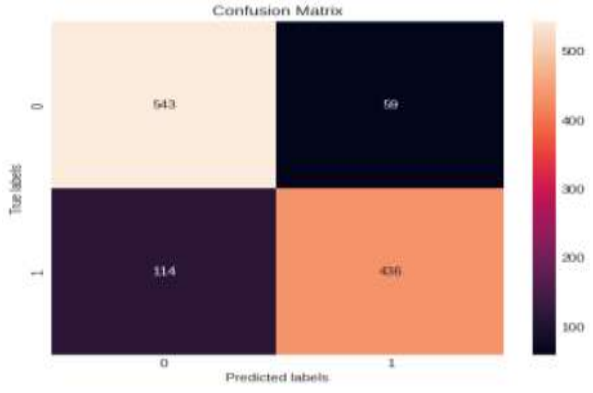
	precision	recall	f1-score	support
0	0.81	0.88	0.84	602
1	0.86	0.77	0.81	550
accuracy			0.83	1152
macro avg	0.83	0.83	0.83	1152
weighted avg	0.83	0.83	0.83	1152

Hyper-parameter Tuning (K-Nearest Neighbor Classifier) using GridSearchCV

```
Train Accuracy = 0.9450716456795484
Test Accuracy = 0.8498263888888888
```

Train Dataset

Test Dataset



Classification Report

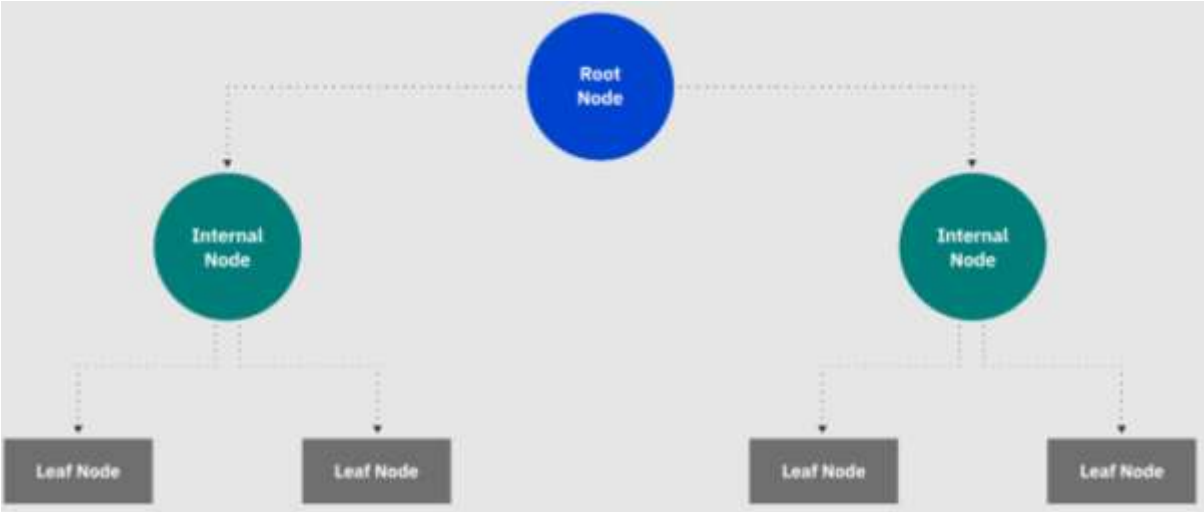
	precision	recall	f1-score	support
0	0.90	1.00	0.95	2277
1	1.00	0.89	0.94	2329
accuracy			0.95	4606
macro avg	0.95	0.95	0.94	4606
weighted avg	0.95	0.95	0.94	4606

Classification Report

	precision	recall	f1-score	support
0	0.83	0.90	0.86	602
1	0.88	0.79	0.83	550
accuracy			0.85	1152
macro avg	0.85	0.85	0.85	1152
weighted avg	0.85	0.85	0.85	1152

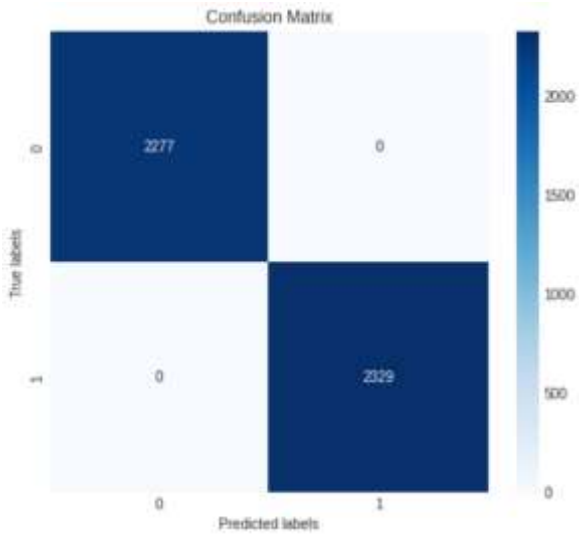
2. Decision Tree Classifier

A decision tree is a non-parametric supervised learning algorithm, which is utilized for both classification and regression tasks. It has a hierarchical, tree structure, which consists of a root node, branches, internal nodes and leaf nodes.

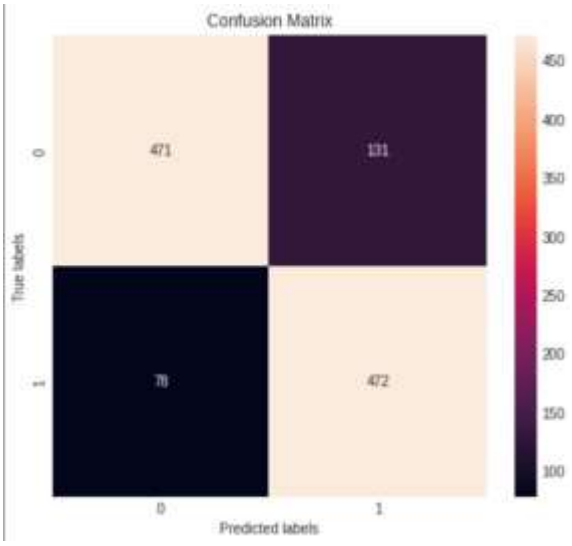



```
Train Accuracy = 1.0
Test Accuracy = 0.8185763888888888
```

Train Dataset



Test Dataset



Classification Report

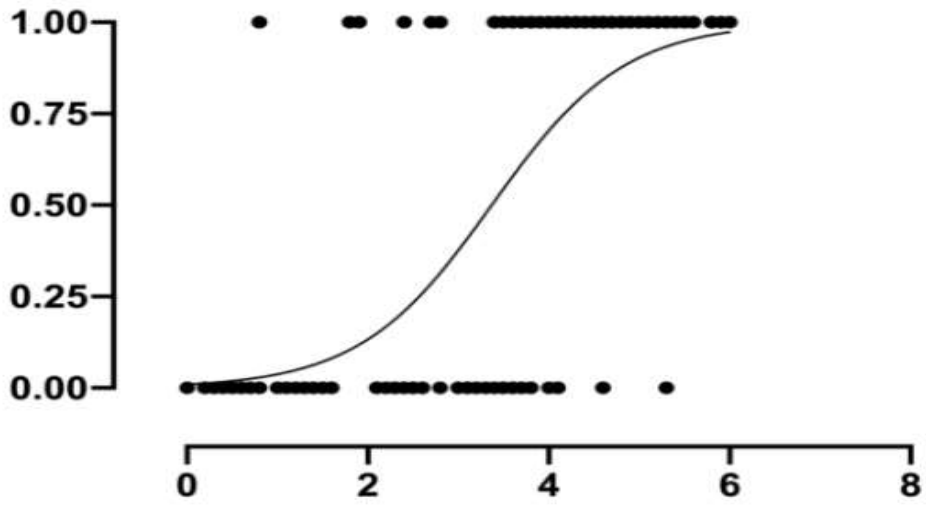
	precision	recall	f1-score	support
0	1.00	1.00	1.00	2277
1	1.00	1.00	1.00	2329
accuracy			1.00	4606
macro avg	1.00	1.00	1.00	4606
weighted avg	1.00	1.00	1.00	4606

Classification Report

	precision	recall	f1-score	support
0	0.86	0.78	0.82	602
1	0.78	0.86	0.82	550
accuracy			0.82	1152
macro avg	0.82	0.82	0.82	1152
weighted avg	0.82	0.82	0.82	1152

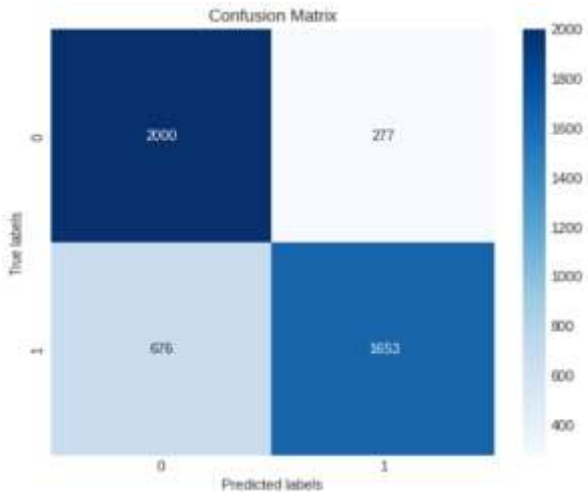
3. Logistic Regression Classifier

This type of statistical model (also known as *logit model*) is often used for classification and predictive analytics. Logistic regression estimates the probability of an event occurring, such as voted or didn't vote, based on a given dataset of independent variables. Since the outcome is a probability, the dependent variable is bounded between 0 and 1.

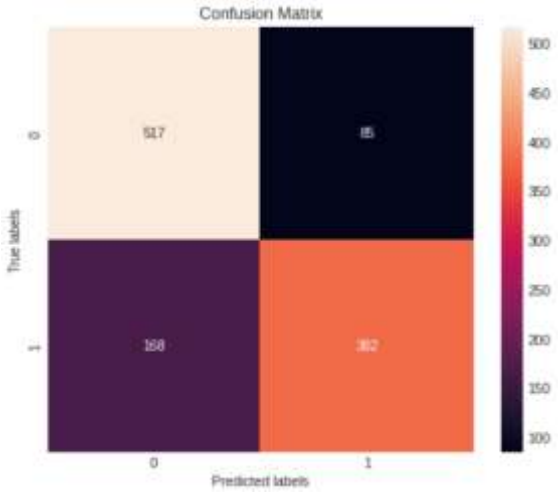



```
Train Accuracy = 0.7930959617889709
Test Accuracy = 0.7803819444444444
```

Train Dataset



Test Dataset



Classification Report

	precision	recall	f1-score	support
0	0.75	0.88	0.81	2277
1	0.86	0.71	0.78	2329
accuracy			0.79	4606
macro avg	0.80	0.79	0.79	4606
weighted avg	0.80	0.79	0.79	4606

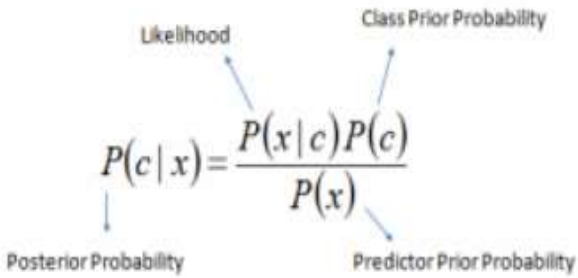
Classification Report

	precision	recall	f1-score	support
0	0.75	0.86	0.80	602
1	0.82	0.69	0.75	550
accuracy			0.78	1152
macro avg	0.79	0.78	0.78	1152
weighted avg	0.78	0.78	0.78	1152

4. Naïve Bayes Classifier

Naive Bayes classifiers are a collection of classification algorithms based on Bayes' Theorem. It is not a single algorithm but a family of algorithms where all of them share a common principle, i.e. every pair of features being classified is independent of each other.

Bayes' Theorem

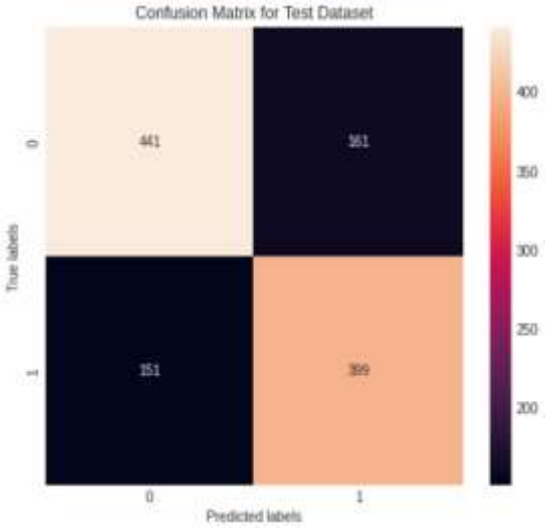


$$P(c|X) = P(x_1|c) \times P(x_2|c) \times \cdots \times P(x_n|c) \times P(c)$$

```
Train Accuracy = 0.7344767694311767
Test Accuracy = 0.7291666666666666
```

Train Dataset

Test Dataset



Classification Report

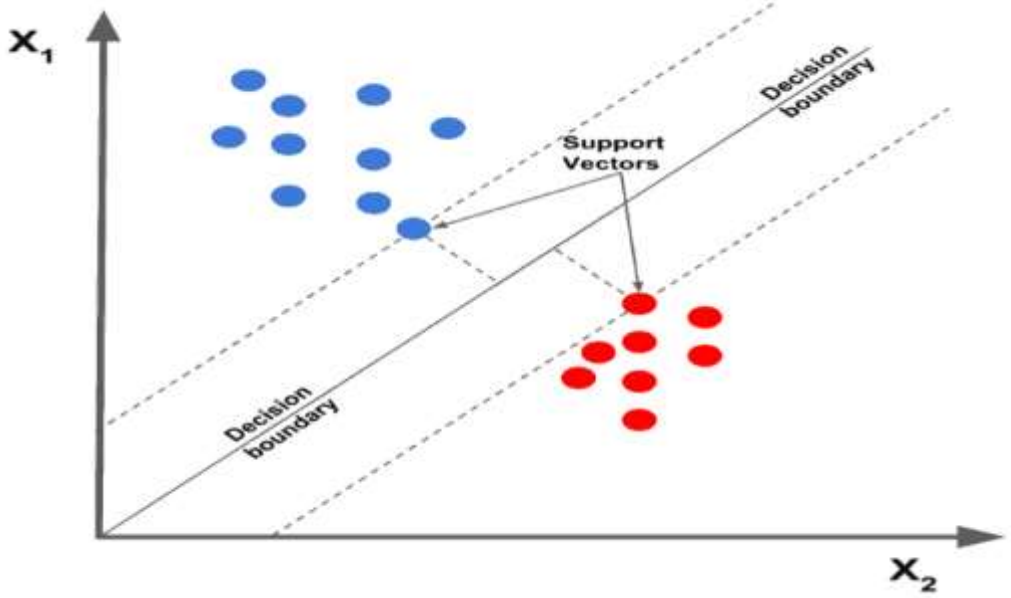
	precision	recall	f1-score	support
0	0.73	0.74	0.73	2277
1	0.74	0.73	0.73	2329
accuracy			0.73	4606
macro avg	0.73	0.73	0.73	4606
weighted avg	0.73	0.73	0.73	4606

Classification Report

	precision	recall	f1-score	support
0	0.74	0.73	0.74	602
1	0.71	0.73	0.72	550
accuracy			0.73	1152
macro avg	0.73	0.73	0.73	1152
weighted avg	0.73	0.73	0.73	1152

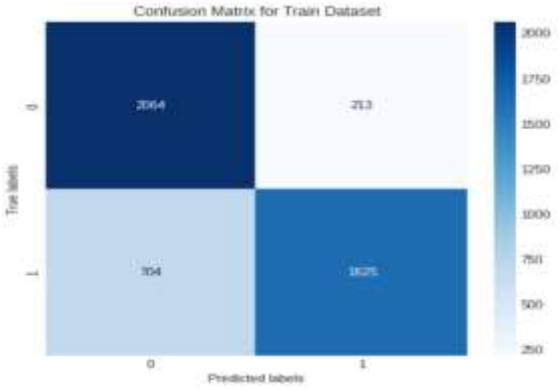
5. Support Vector Classifier (SVC)

Support Vector Machine (SVM) is a robust classification and regression technique that maximizes the predictive accuracy of a model without overfitting the training data. SVM is particularly suited to analyzing data with very large numbers (for example, thousands) of predictor fields.

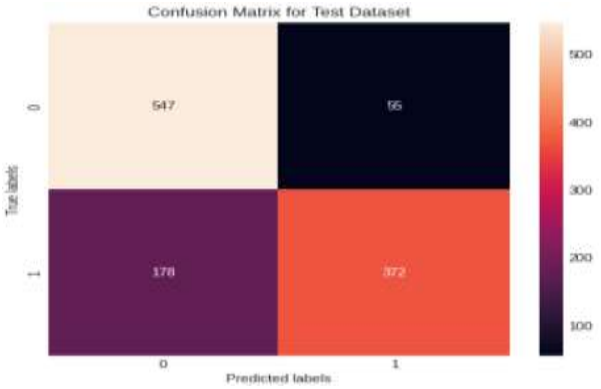


```
Train Accuracy = 0.8009118541033434
Test Accuracy = 0.7977430555555556
```

Train Dataset



Test Dataset



Classification Report

	precision	recall	f1-score	support
0	0.75	0.91	0.82	2277
1	0.88	0.70	0.78	2329
accuracy			0.80	4606
macro avg	0.81	0.80	0.80	4606
weighted avg	0.82	0.80	0.80	4606

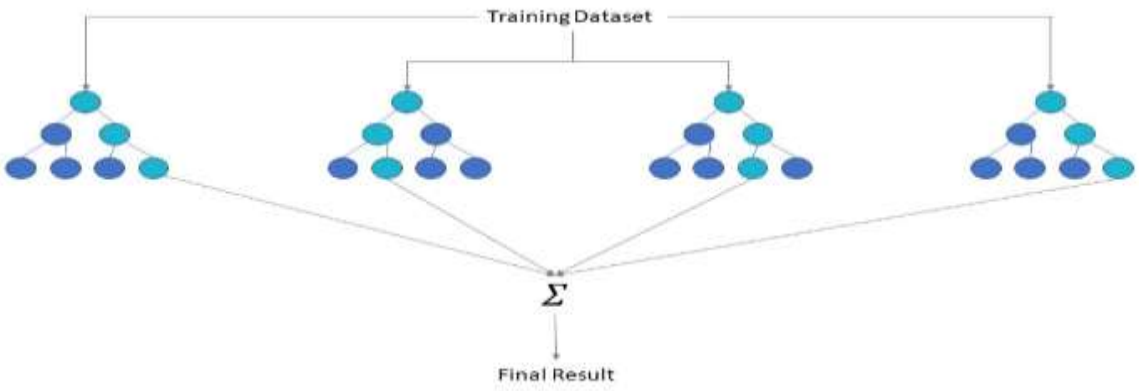
Classification Report

	precision	recall	f1-score	support
0	0.75	0.91	0.82	602
1	0.87	0.68	0.76	550
accuracy			0.80	1152
macro avg	0.81	0.79	0.79	1152
weighted avg	0.81	0.80	0.79	1152

6. Random Forest Classifier

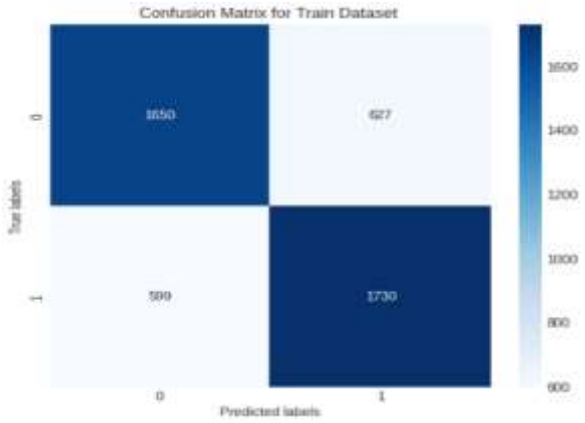
In this article, we will see how to build a Random Forest Classifier using the Scikit-Learn library of Python programming language and in order to do this, we use the IRIS dataset which is quite a common and famous dataset. The Random forest or Random Decision Forest is a supervised Machine learning algorithm used for classification, regression, and other tasks using decision trees.

The Random forest classifier creates a set of decision trees from a randomly selected subset of the training set. It is basically a set of decision trees (DT) from a randomly selected subset of the training set and then It collects the votes from different decision trees to decide the final prediction.

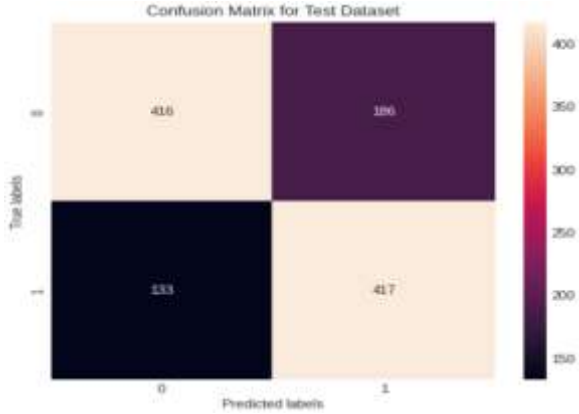


```
Train Accuracy = 0.7338254450716457
Test Accuracy = 0.7230902777777778
```

Train Dataset



Test Dataset



Classification Report

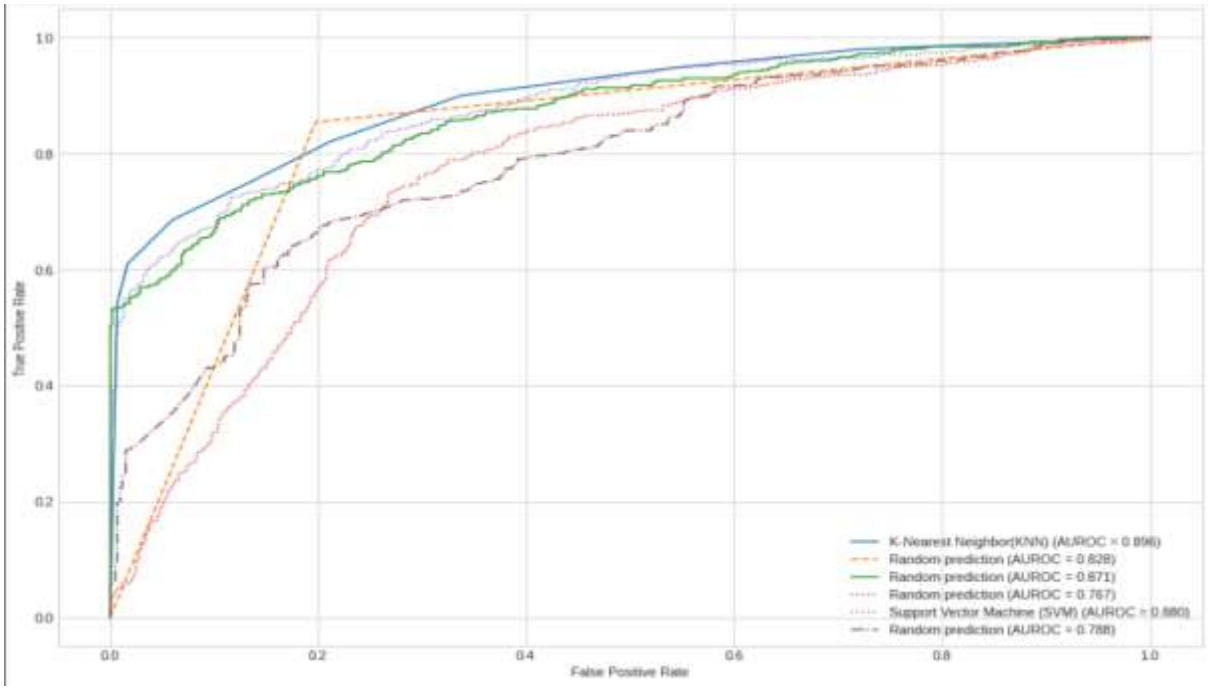
	precision	recall	f1-score	support
0	0.73	0.72	0.73	2277
1	0.73	0.74	0.74	2329
accuracy			0.73	4606
macro avg	0.73	0.73	0.73	4606
weighted avg	0.73	0.73	0.73	4606

Classification Report

	precision	recall	f1-score	support
0	0.76	0.69	0.72	602
1	0.69	0.76	0.72	550
accuracy			0.72	1152
macro avg	0.72	0.72	0.72	1152
weighted avg	0.73	0.72	0.72	1152

ROC_AUC (Curve)

AUC - ROC curve is a performance measurement for the classification problems at various threshold settings. ROC is a probability curve and AUC represents the degree or measure of separability. It tells how much the model is capable of distinguishing between classes. Higher the AUC, the better the model is at predicting 0 classes as 0 and 1 classes as 1



Conclusion

- Age is the biggest factor for the cardiovascular disorders.
- 'male', 'smokers', patients on BP Medication, with Prevalent Stroke, with Prevalent Hypertension and Diabetic are more prone to heart disease.
- For Decision Tree and Random Forest Classifiers 'age' is most important feature.
- For Logistic Regression Classifier dummy features of 'education' are the top 4 important features.
- Best performance of Models on test dataset for class 1 (Risk of TenYearCHD):
    - Recall - Decision Tree
    - Precision – SVC
    - F1 Score - Decision Tree
    - Accuracy - K-Nearest Neighbor(KNN)

Best AUROC Score: K-Nearest Neighbor(KNN)

Reference:

1. GeeksforGeeks
2. Analytics Vidhya
3. Stack Overflow
4. https://towardsdatascience.com/
5. Youtube, etc