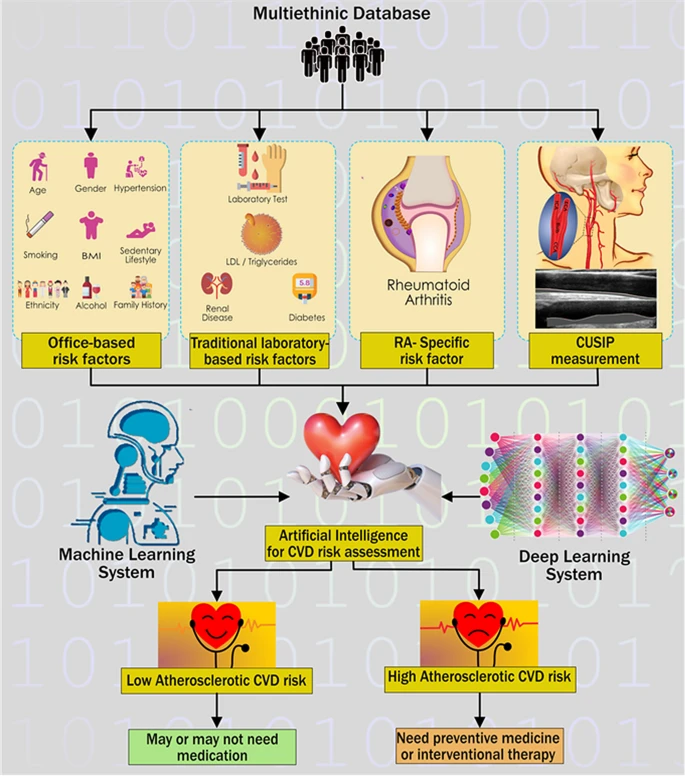
**Classification Analysis on Cardiovascular Risk Prediction**



**-** **Sanjay Verma**

**Data science trainee at**

**Almabetter**

**ABSTRACT**

Cardiovascular Diseases are the number one source of explanation for death globally, more people die annually from CVDs than from the other cause. Roughly estimated 17.9 million people died from CVDs in 2016, representing 31% of all global deaths. Early detection of cardiac diseases and continuous supervision of clinicians can reduce the death rate. Coming to CVD, the silver lining is that heart attacks are highly preventable and simple lifestyle modifications. It is, however, difficult to identify high risk patients because of the multifactorial nature of several contributory risk factors such as diabetes, high blood pressure, high cholesterol, et cetera. This is where machine learning and data mining comes to the rescue. Our aim is to develop a machine learning model capable of identifying if a person has a cardiovascular disease or not. We can do this by extracting the features of a dataset and training a machine learning model with the highest accuracy. We trained four machine learning algorithms and compared their accuracy: Logistic Regression, K-nearest Neighbors ,Decision Tree Classifier and Support Vector Machine to identify people with high risk of getting cardiovascular disease.

**PROBLEM STATEMENT**

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

**INTRODUCTION**

Coronary artery disease (CAD) is the most common heart disease seen in today’s population worldwide. Recent studies of the American Heart Association have shown that coronary artery diseases recorded 13% death in the USA in 2018 and worldwide in 2015. Amongst other diseases CAD found to be one of the most common causes of death, with the record of 15.6% of all results across the globe. Because this disease is associated with modifiable risk factors which indirectly related with lifestyle and intervention, timing of detection and diagnostic accuracy are especially relevant in clinical management of patients with CAD.

Over the past years, approaches making a significant impact in the detection and diagnosis of diseases that include machine learning (ML). In general, ‘training’ an algorithm with a control dataset for which the disease status (disease or no disease) is known, and then applying this trained algorithm to a variable dataset in order to predict the disease status in patients for whom it is not yet determined. The ML algorithm will be better trained as a predictor for disease status as larger cohorts of data are introduced. prediction with ML would empower clinicians with improved detection, diagnosis, classification, risk stratification and ultimately, management of patients, all while potentially minimizing required clinical intervention are More accurate.

**DATASET PREPROCESSING**

According to the size of the dataset, extensive cleaning of the dataset is not needed.Column names in the dataset are short as we get a meaningful explanation of these features. Dataset contains NaN values and it is removed by replacing the NAN values by unknown values.

**DATA DESCRIPTION**

**Attribute Description :**

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

**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 a 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”) - Dv

**CHALLENGES FACED**

* As data is imbalanced , to balance the data in an appropriate way was a bit tricky.
* Using the appropriate metrics for comparison of the implemented machine learning algorithms.

**APPROACH**

We checked the Outliers and correlation matrix to overcome the noise in the dataset.Also,data was balanced using the SMOTE method and scaled by Standard Scaler transformation. As the Coronary Heart Diseases dataset defines the classification problem. We decided to train the models such as Logistic regression , K-nearest Neighbors , Decision Tree Classifier & Support Vector Machine .Also, we used Hyperparameter Tuning for improvement in the model fitting to understand the better results of the model as well as the metrics.

**TOOLS USED**

Following libraries were used for analyzing the data and visualizing:

● Pandas: Extensively used to load and wrangle with the dataset.

● Matplotlib: Used for visualization.

● Seaborn: Used for visualization.

● Warnings: For filtering and ignoring warnings.

● Numpy: For some math operations in predictions.

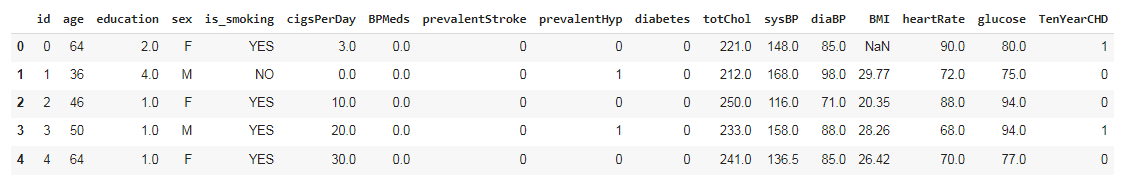
● Sklearn: For analysis and prediction.

**●** Stats models: For outliers influence.

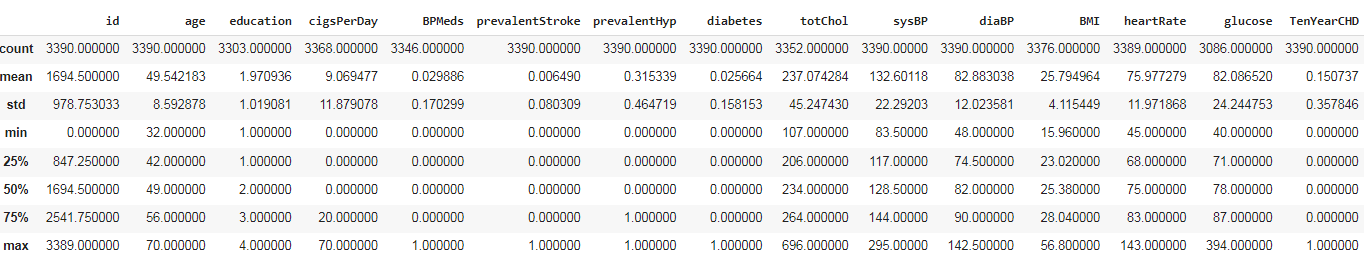
* train\_test\_split: To split the dataset into training and testing data
* StandardScaler: To scale all the features, so that the Machine Learning model better adapts to the dataset

**DATA INSIGHTS**

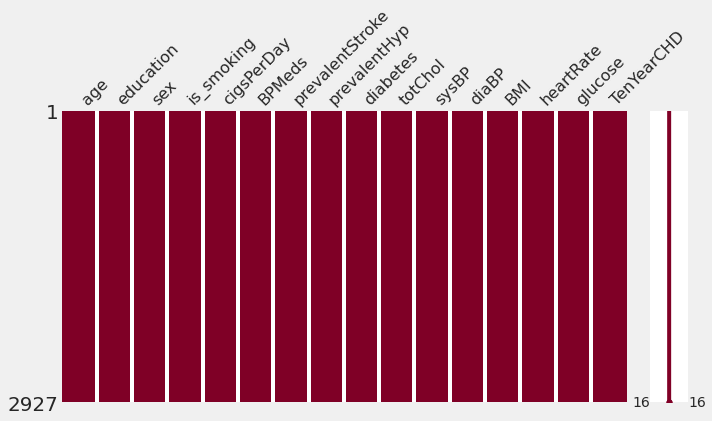
**Pandas dataframe**

****

**Describe**

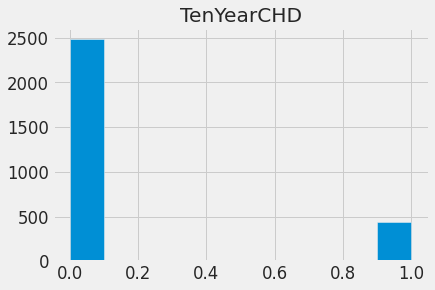
****

**Visualize the presence of NAN values-**



NAN values are removed by replacing it with unknown values with missingno matrix.The above figure shows that presence of NAN values will not reflect in the dataset.

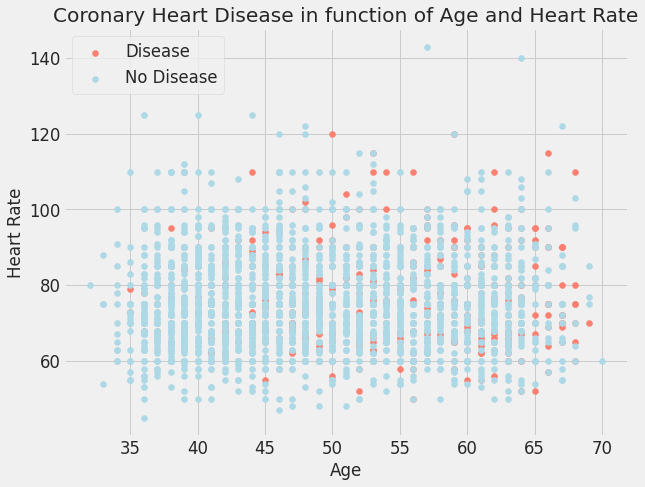
**Histogram Plot for Dependent & Independent Features-**





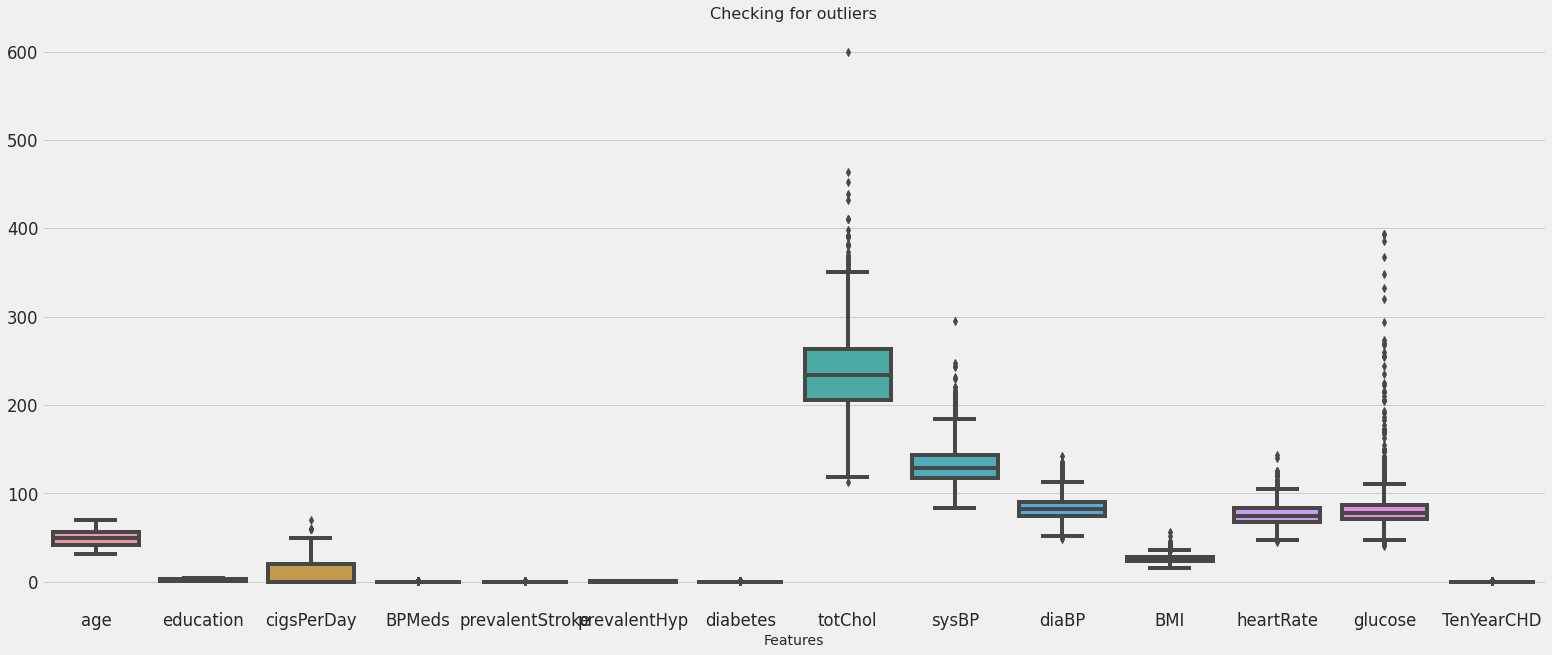
### It shows how each feature and label is distributed along different ranges, which further confirms the need for scaling. Next, wherever you see discrete bars, it basically means that each of these is actually a categorical variable. We will need to handle these categorical variables before applying Machine Learning.

**Age vs.Heart Rate for Coronary Heart Disease**



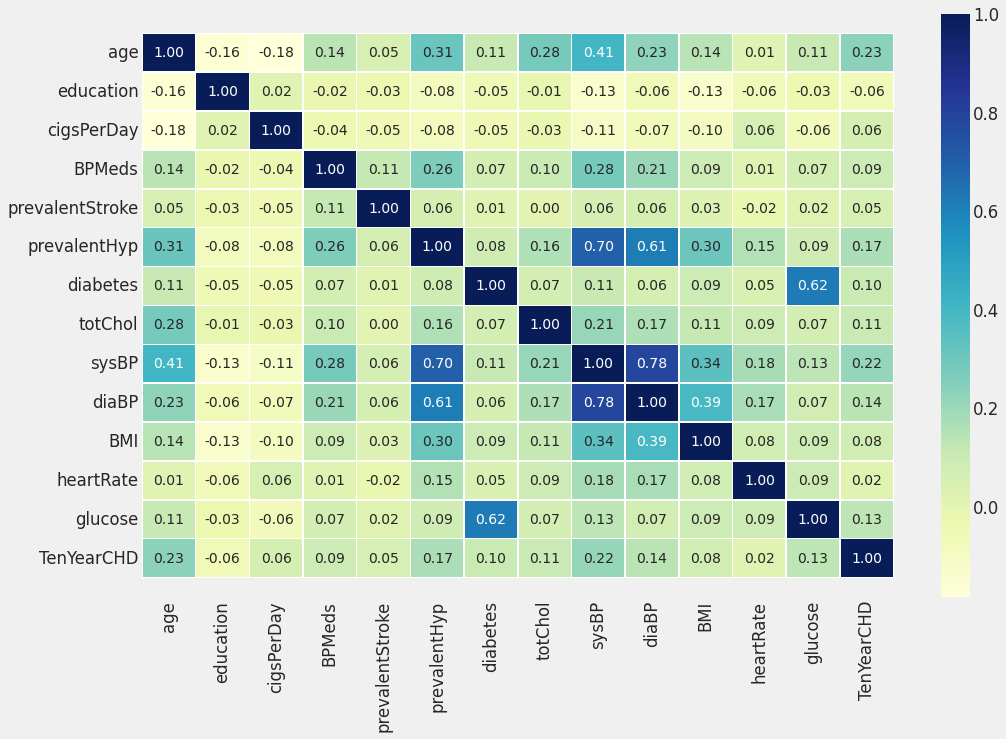
**OUTLIERS DETECTION**

An outlier is a data point that is noticeably different from the rest. They represent errors in measurement, bad data collection, or simply show variables not considered when collecting the data.



**CORRELATION MATRIX (HEATMAP)**

To refine the data further, a correlation matrix was created amongst all the feature variables to analyze interaction effects.



HeartRate and prevalentStroke are the lowest correlated with the target variable.

Also, some of the features have a negative correlation with the target value and some have positive.

**DATA MODELING**

After the data preparation is completed it is ready for the purpose of analysis. Only numerical valued features are taken into consideration. The data were combined and labeled as X and y as independent and dependent variables respectively.

**SPLITTING THE DATASET**

The train\_test\_split was imported from the sklearn.model\_selection. The data is now divided into 70% and 30% as train and test splits respectively. 70% of the data is taken for training the model and 30% is for a test and the random state was taken as 0.

**SCALING THE DATA**

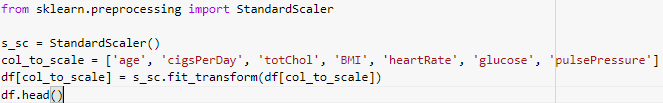
We have used the **Standard Scaler method** to scale the dataset.

Standardize features by removing the mean and scaling to unit variance.

The standard score of a sample x is calculated as:

z = (x - u) / s

where u is the mean of the training samples or zero if with\_mean=False, and s is the standard deviation of the training samples or one if with\_std=False.



**METRICS USED**

* **Classification Report**:A classification report is a performance evaluation metric in machine learning. It is used to show the precision, recall, F1 Score, and support of your trained classification [model](https://thecleverprogrammer.com/2020/11/27/machine-learning-algorithms-with-python/).

1) **Accuracy**: the proportion of total dataset instances that were correctly predicted out of the total instances

accuracy=(true positives+true negatives)/total

2) **Recall (sensitivity)**: the proportion of the predicted positive dataset instances out of the actual positive instances

sensitivity=true positives/(true positives+false negatives)

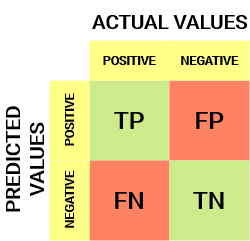
3) **F1 score**: a composite harmonic mean (average of reciprocals) that combines both precision and recall. For this, we first measure the precision, the ability of the model to identify only the relevant dataset instances

precision=true positives/(true positives+false positives)

The F1 score is estimated as

F1=2×(precision×recall)/(precision+recall)

* **Confusion Matrix:**A Confusion matrix is an N x N matrix used for evaluating the performance of a classification model, where N is the number of target classes. The matrix compares the actual target values with those predicted by the machine learning model. This gives us a holistic view of how well our classification model is performing and what kinds of errors it is making.



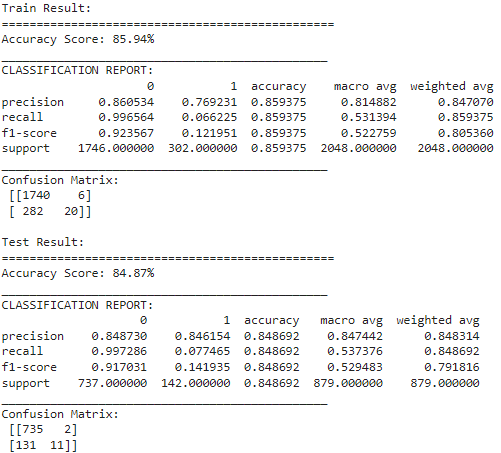
* True Positive (TP) -The predicted value matches the actual value.The actual value was positive and the model predicted a positive value
* True Negative (TN) -The predicted value matches the actual value.

The actual value was negative and the model predicted a negative value

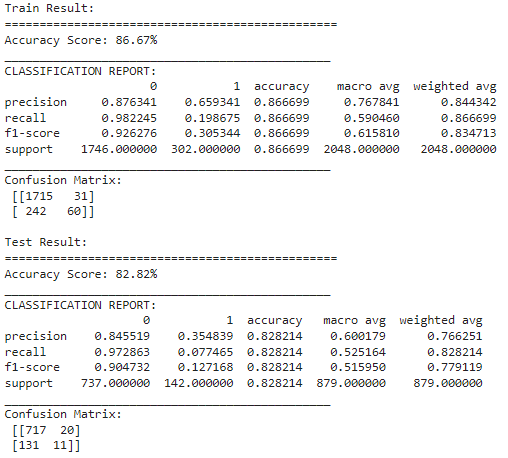
* False Positive (FP) – Type 1 error :-The predicted value was falsely predicted.The actual value was negative but the model predicted a positive value
* False Negative (FN) – Type 2 error:-The predicted value was falsely predicted.The actual value was positive but the model predicted a negative value

**MODEL IMPLEMENTATION**

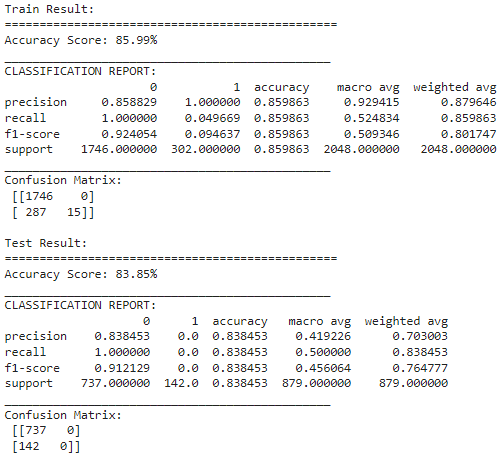
* **Logistic Regression -**A logistic regression is a type of statistical procedure. It is used to refer specifically to the problem in which the dependent variable is binary, that is the number of available categories is two, while the problem with more than two categories is referred to as multi logistic regression.



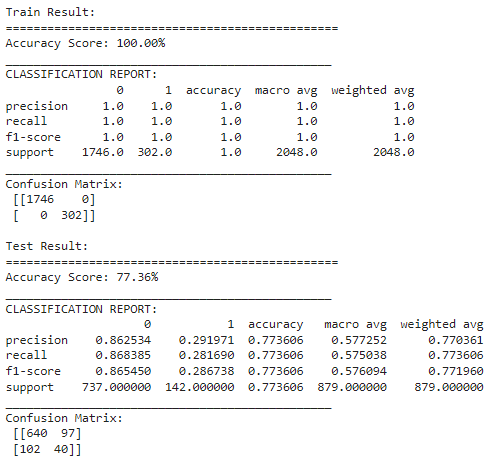
* **K-nearest Neighbors -** The K Nearest Neighbor algorithm falls under the Supervised Learning category and is used for classification (most commonly) and regression. It is a versatile algorithm also used for imputing missing values and resampling datasets.



* **Support Vector Machine -**Support Vector Machine (SVM) is a classification technique used for the classification of linear as well as non-linear data. SVM is the margin based classifier. It selects the maximum margin. This model is further used to perform classification of testing data.

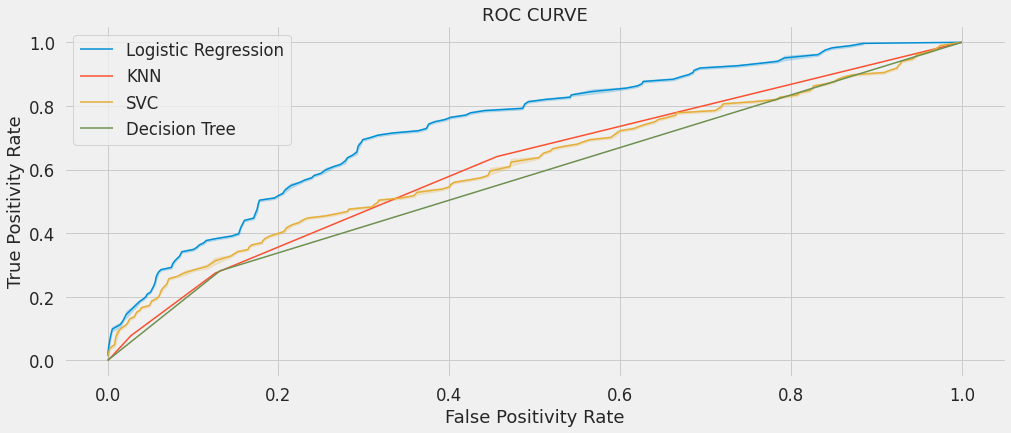


* **Decision Tree Classifier -**Decision tree build classification or regression models in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes.

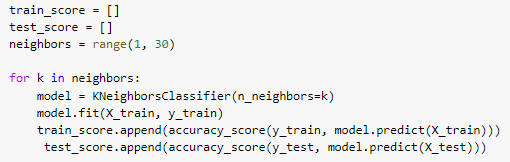


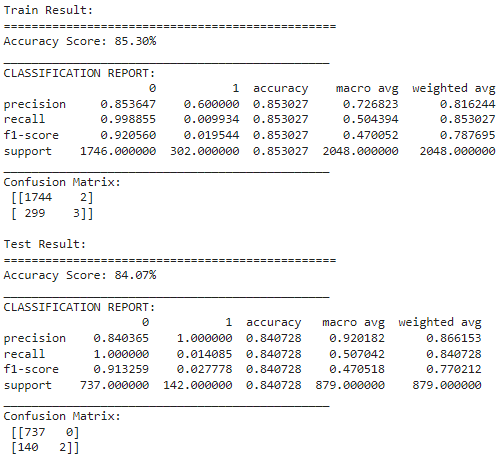
* **ROC Curve:** An ROC curve (receiver operating characteristic curve) is a graph showing the performance of a classification model at all classification thresholds. This curve plots two parameters:

1. True Positive Rate-True Positive Rate (TPR) is a synonym for recall and is therefore defined as follows:  **TPR=TP / (TP+FN)**
2. False Positive Rate-False Positive Rate (FPR) is defined as follows: **FPR =FP / (FP+TN)**



* **Hyperparameter Tuning on K-nearest Neighbors algorithm-**

****

****

**CONCLUSION**

* Patients of age group 32 to 70 years are present among which 38 to 46 years age group have high smoking habits.
* Comparative to male patients’ female patients is more.
* Across the dataset we have 1307 male patients, of which 809 male patients smoke cigarettes.
* We have 1620 female patients out of which 638 smoke cigarettes.
* patients who previously had a stroke and Number of patients with medical history like blood pressure medication, Diabetes is very low.
* Logistic Regression-Nearest Neighbors, Support Vector Machine & Decision Tree Classifier models were implemented.
* From above these models, we found that KNN is the best fitted model compared to other models
* In Hyperparameter tuning, we observed that K-Nearest Neighbors accuracy has improved which shows that KNN (with Hyperparameter Tuning) is the best fitted model for coronary heart disease dataset.

Train Accuracy = **85.30** & Test Accuracy = **84.07**

For future improvement in the model fitting for coronary heart disease, we can perform the Random Forest Classifier, XGBoost models also. Consulting medical people we can analyze the feature in proper and required manner to approach the disease cause and effects.