Blog on Heart Diseases

## 1. Introduction

The most crucial work in the healthcare industry is disease diagnosis. Heart disease is the major cause of medical problem and mortality globally: it accounts for more deaths annually than any other disease. Heart disease covers a range of different conditions that could affect your heart. It is one of the most complex diseases to predict given the number of potential factors in your body that can lead to it. According to the [WHO](https://www.who.int/health-topics/cardiovascular-diseases/#tab=tab_1), an estimated 17.9 million people died from heart disease in 2016, representing 31% of all global deaths. Over three quarters of these deaths took place in low- and middle-income countries. The number one killer in the world today is heart disease, it becomes one of the most difficult diseases to diagnose.  Machine learning classification techniques can significantly give benefit to the medical field by providing an accurate and quick diagnosis of heart diseases.

## 2. Data Description

There are 200 rows (records) in the dataset, and it contains 14 continuous attributes. The goal is to predict the presence of heart disease in the patient. The “goal” field refers to the presence of heart disease in the patient. It is integer valued from 0 (no presence) to 4.

Dataset contains following features:

**age:** The person’s age in years  
**sex:** The person’s sex (1 = male, 0 = female)  
**cp:** The chest pain experienced (value 1: typical angina, value 2: atypical angina, value 3: non-anginal pain, value 4: asymptomatic)  
**trestbps:** The person’s resting blood pressure  
**chol:** The person’s cholesterol measurement in mg/dl  
**fbs:** The person’s fasting blood sugar (> 120 mg/dl, 1 = true; 0 = false)  
**restecg:** Resting electrocardiographic measurement (0 = normal, 1 = having ST-T wave abnormality, 2 = showing probable or definite left ventricular hypertrophy by Estes’ criteria)  
**thalach:** The person’s maximum heart rate achieved  
**exang:** Exercise induced angina (1 = yes; 0 = no)  
**oldpeak:** ST depression induced by exercise relative to rest (‘ST’ relates to positions on the ECG plot)  
**slope:** The slope of the peak exercise ST segment (value 1: upsloping, value 2: flat, value 3: downsloping)  
**ca:** The number of major vessels (0 – 3)  
**thal:** A blood disorder called thalassemia (3 = normal; 6 = fixed defect; 7 = reversable defect)  
**target:** Heart disease (0 = no, 1 = yes)

**Binary**

* sex (0 = female; 1 = male)
* fbs: Fasting blood sugar > 120 mg/dl
* exang: Exercise induced angina (0 = no; 1 = yes)

**Categorical**

* cp: Chest pain type (0 = Asymptomatic angina; 1 = Atypical angina; 2 = Non-angina; 3 = Typical angina)
* restecg: Resting ECG (0 = Left ventricular hypertrophy; 1 = Normal; 2 = ST-T wave abnormality)
* slope: Slope of the peak exercise ST segment (0 = downsloping; 1 = upsloping; 2 = flat)
* thal: Thalium stress test result (0 = NA; 1 = Fixed defect; 2 = Normal; 3 = Reversible defect)

**Ordinal**

ca: number of major vessels (0-3) colored by flourosopy

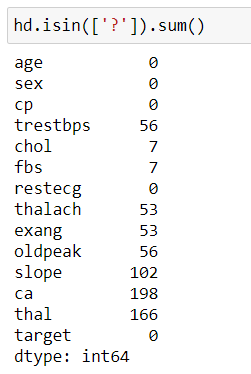
**Numeric**

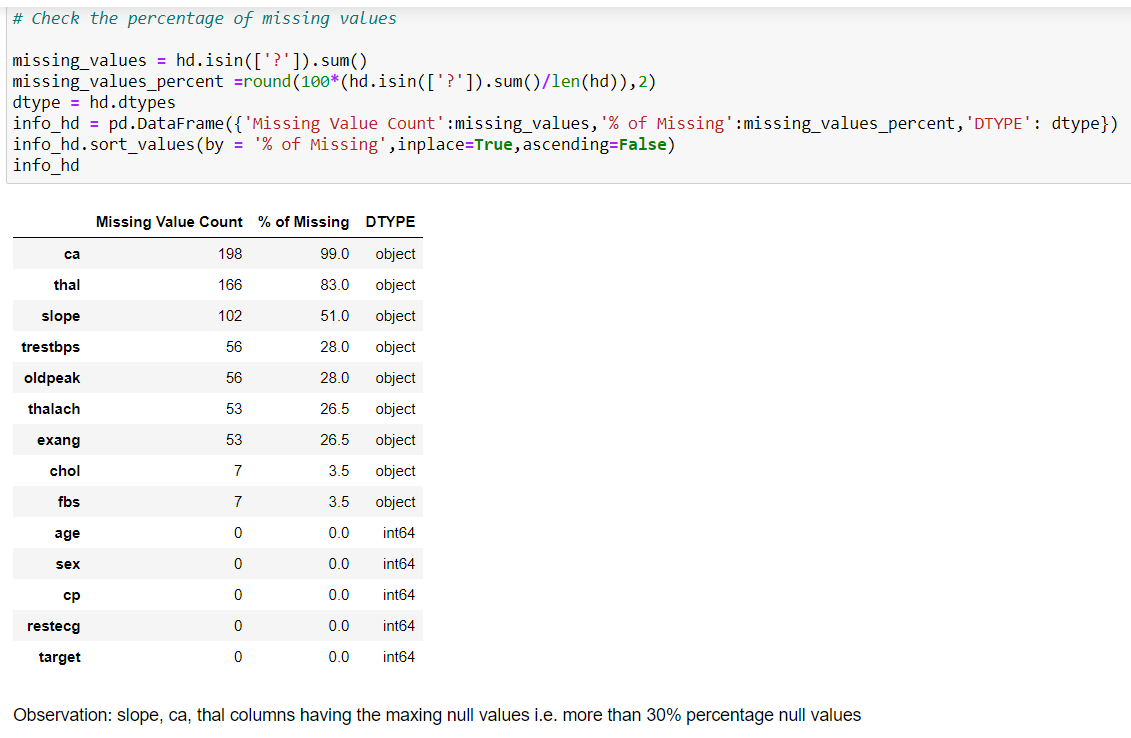
* age
* oldpeak: ST depression induced by exercise relative to rest
* trestbps: Resting blood pressure
* chol: Serum cholestoral in mg/dl
* thalach: Maximum heart rate achieved during thalium stress test

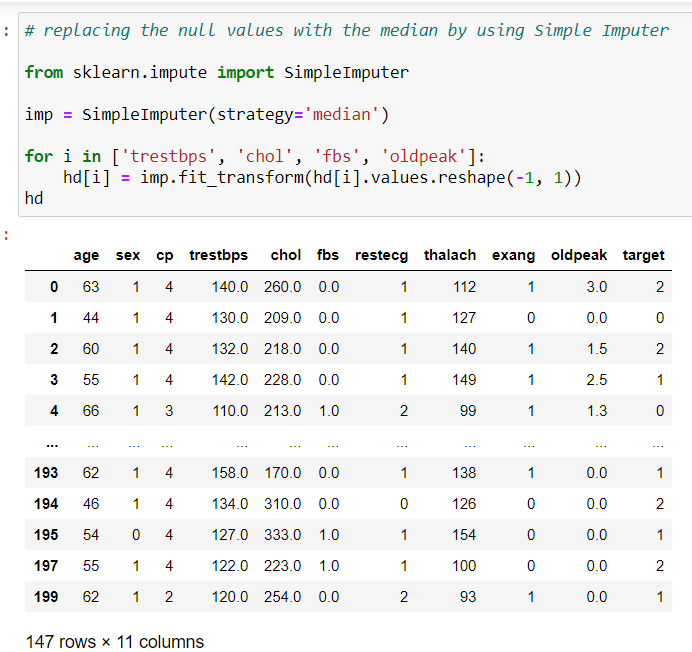
**Target**

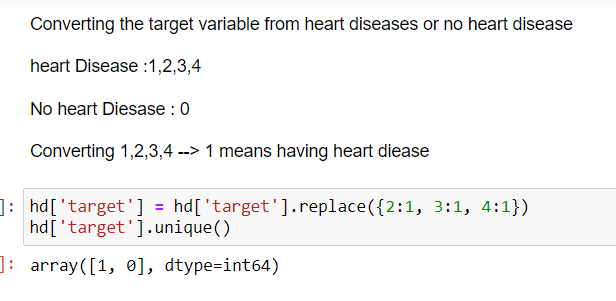
target: 1 = heart disease; 0 = no heart disease

## 3. Data Analysis:

1. Checking the missing values in the dataset and with the percentage of missing values  
   



1. Replacing the null with the SimpleImputer technique   
   
2. Converting the target variable heart disease and no hear disease



## 4. EDA Concluding Remark

1. Filled all the null values with the numeric values by using Simple Imputer
2. Convert the heart disease from 0 to 4 🡪 1 (heart disease) or 0 (no heart disease)
3. Removed the unwanted columns or features from the dataset
4. Removed the outliers whose having more skewness
5. Checked the correlations with the dataset features

## 5. Model Building and Validation

Model selection is the process of combining data and prior information to select among a group of statistical models.

So, we need to focus on four parameters:

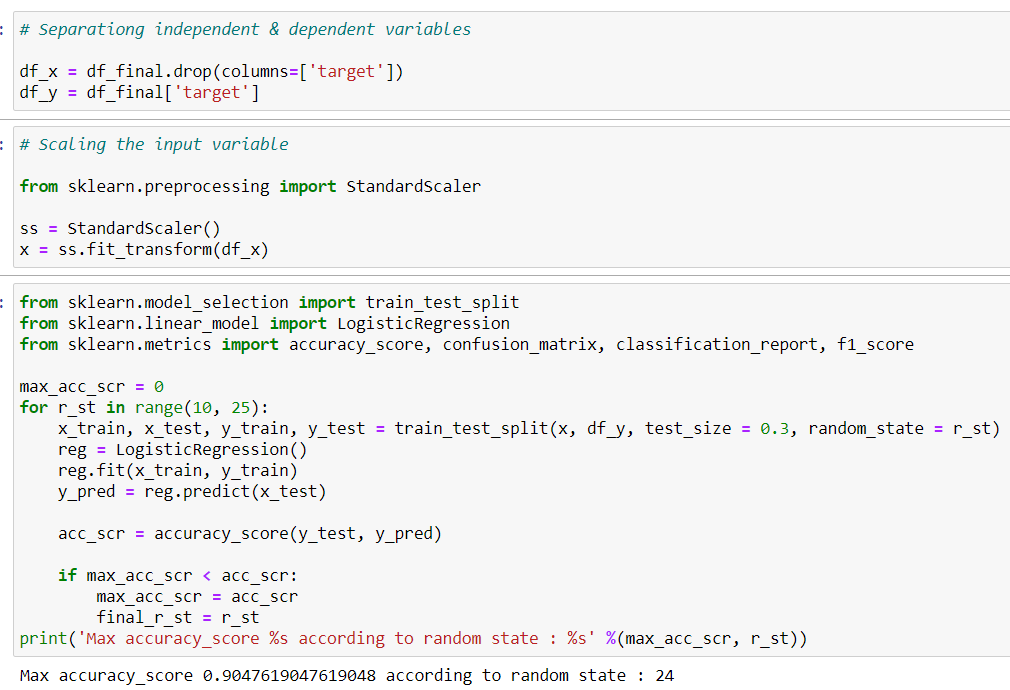
* **Accuracy:** Overall, how often is the classifier correct? i.e {(TP+TN)/Total}
* **True Positive Rate:** When it's yes, how often does it predict yes? default\_ind = 1, {TP/Actual YES}, this is also known as "Sensitivity" or "Recall"
* **Precision:** When it predicts yes, how often is it correct? i.e. {TP/(TP+FP)}
* **Specificity:** When it's actually no, how often does it predict no? default\_ind = 0, {TN/actual NO}
* **Cross Validation Score:** Cross Validation is a technique which involves reserving a particular sample of a dataset on which you do not train the model. Later, you test your model on this sample before finalizing it. Do this for k folds and take mean of accuracy scores of the k fold models.
* **F1 Score:** This is a weighted average of the true positive rate (recall) and precision.
* **ROC Curve:** This is a commonly used graph that summarizes the performance of a classifier over all possible thresholds. It is generated by plotting the True Positive Rate (y-axis) against the False Positive Rate (x-axis) as you vary the threshold for assigning observations to a given class.

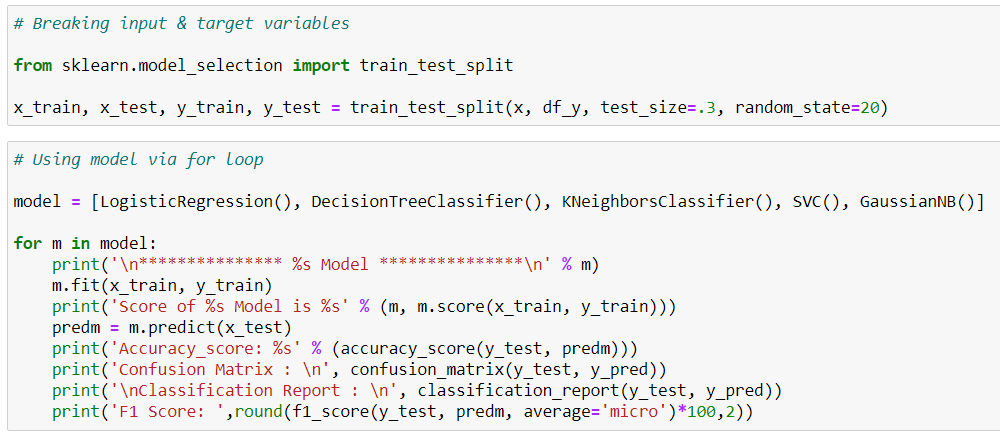
Now our data are ready!

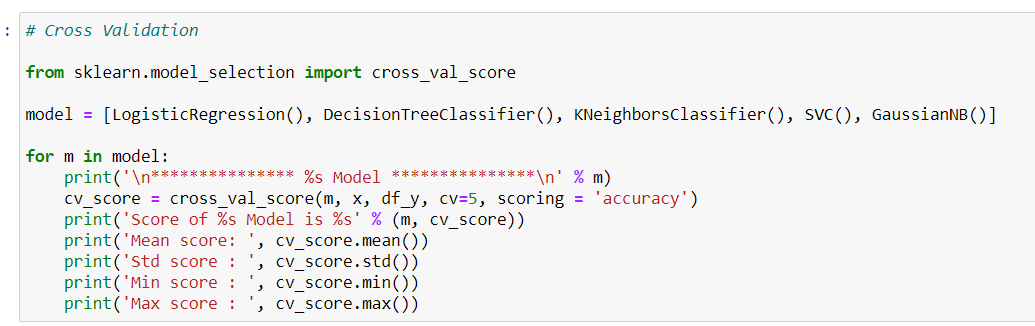
According to the data, our Target variable 'target' is discrete, so our model is Classification.

Let's now begin to train out Classification model!

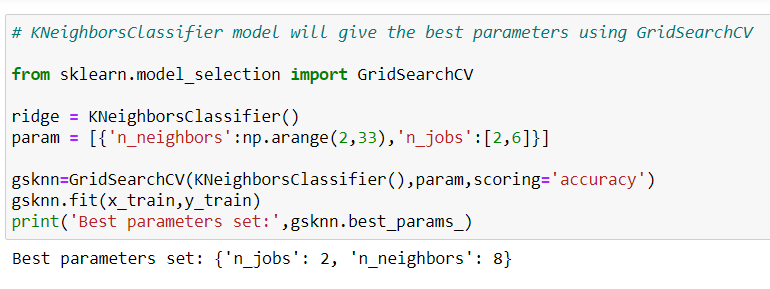
We will need to first split up our data into an X array that contains the features to train on, and y array with the target variable

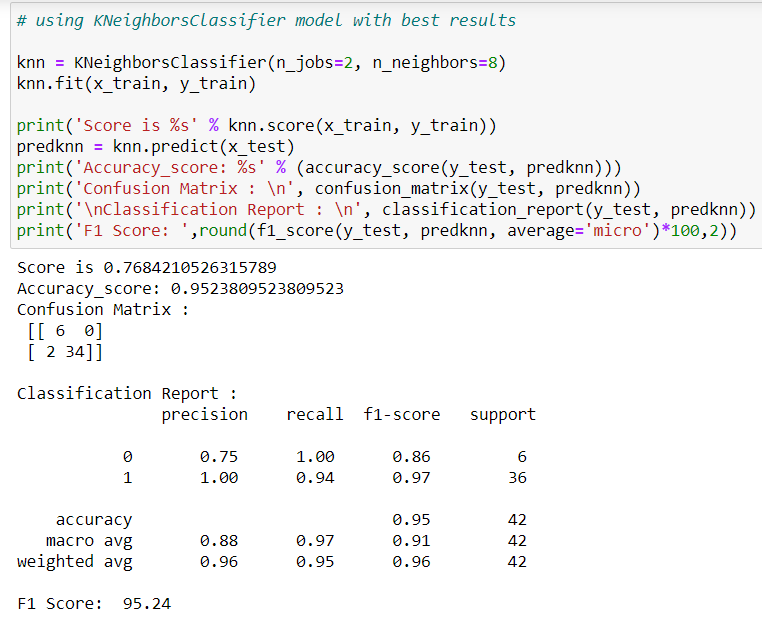


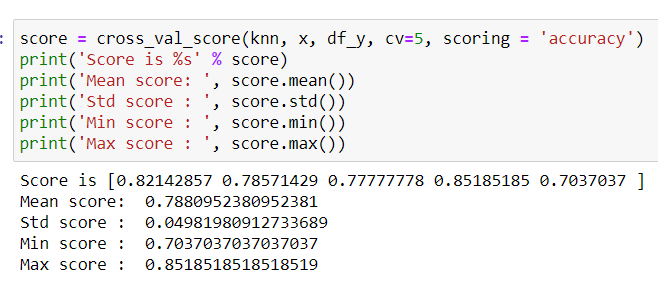




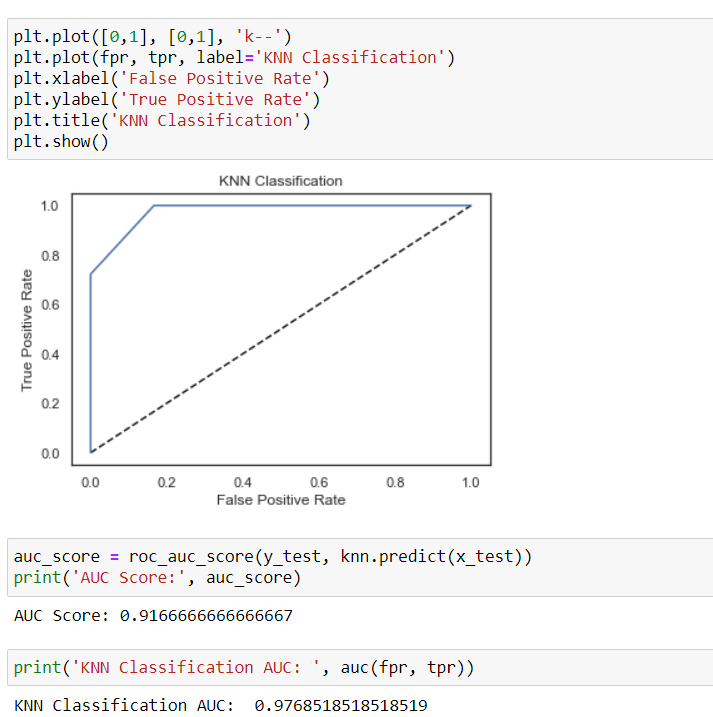
Pipeline and GridSearchCV preprocessing tools will help to streamline the process of model training and hyperparameter optimization.







AUC ROC Curve



## 6. Conclusion

The Area under the ROC curve is 97.7% with KNN Classifier model which is working well beacuse it gives the best score

The model predicted with 95.2% accuracy. The model is more specific than sensitive.