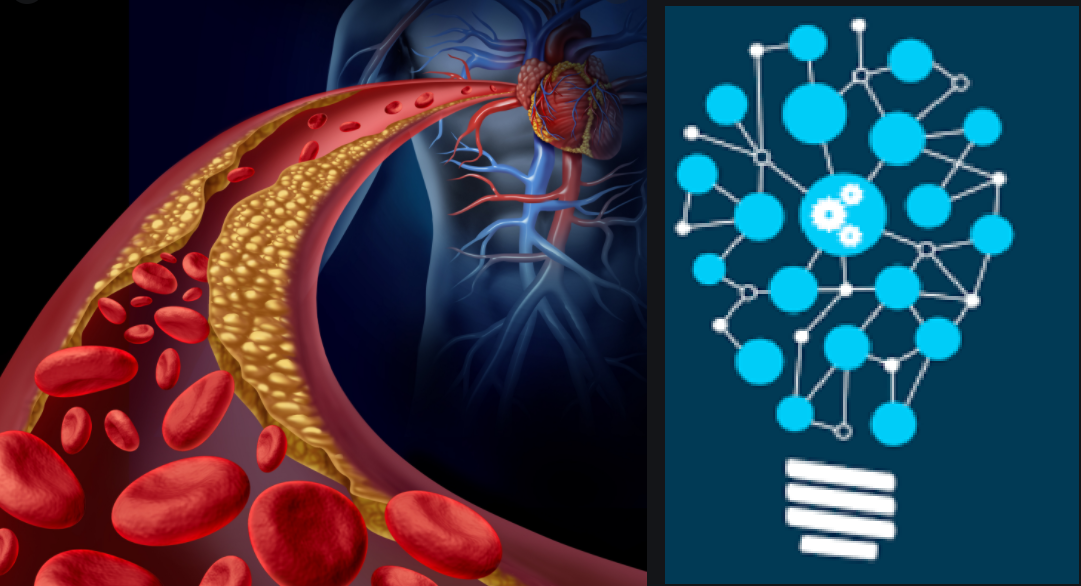


**PYTHON FOR MACHINE LEARNING**

**PROJECT**

**PREDICTION OF CARDIOVASCULAR DISEASE**

**USING MACHINE LEARNING ALGORITHMS**

**Course :19AIE205-Python for Machine Learning**

**Mentor: Dr. Kumaran U**

**(Department of Computer Science and Engineering)**

**BY: TEAM OUTLIERS**

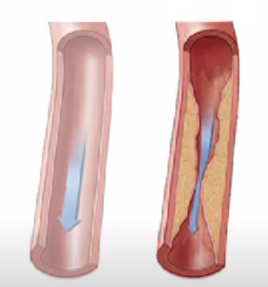
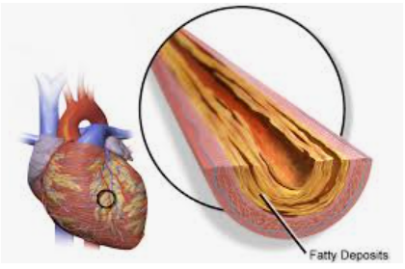
**-APOORVA.M BL.EN.U4AIE19007**

**-M. BHUVANASHREE BL.EN.U4AIE19010**

**-KARNA SAI NIKHILESH BL.EN.U4AIE19027**

**Introduction to Cardiovascular Diseases:**

Cardiovascular disease involves abnormalities of the heart and the blood vessels. The lifetime risk of developing significant cardiovascular disease is greater than the likelihood of developing cancer. Our heart is a muscular pump that requires heart arteries to supply oxygen rich blood to keep it going. Coronary heart disease occurs when these blood vessels become narrowed due to a buildup of plaque. The plaque is made up of cholesterol and other substances. This process is called Atherosclerosis.

Atherosclerosis occurs over a life time, and is influenced by risk factors. Some risk factors can't be changed; however, others can be altered through healthy lifestyle choices and medications if needed.

Symptoms of coronary heart disease occur when heart muscle does not get enough oxygen. Some describe this as chest pressure or chest pain that starts over the left chest and can radiate or travel to the left arm, left jaw, back, or sometimes to the right chest. Some patients do not have any chest pan or pressure symptoms but instead have intense shortness of breath or epigastric discomfort that can feel like a bad heart burn, breathlessness, nausea, vomiting or sweats can also sometimes be associated. These symptoms are called angina. There are many treatments for angina, including medications, angioplasty and stents, and bypass surgery. Aggressive risk factor modification is part of every treatment plan. Coronary heart disease can have serious complications including heart attack and even sudden death. Fortunately, developing significant coronary heart disease is largely preventable. Early prevention is best, but you can still improve how you'll do no matter how old you are by reducing your risk factors!

**Objective of the project:**

What we aim to do is create an efficient classification model that best predicts if a given patient is likely to develop cardio vascular disease or not.

**Overview of our approach towards the problem:**

1. Perform Exploratory Data Analysis to gain insights on the Data
   1. Data Cleaning
   2. Data Summarization: Describe the data and its distributions
   3. Data Visualization: Create graphical summaries of the data
2. Transform the data for training the different classification algorithms
3. Apply the different Machine Learning Algorithms
   1. Train, Cross validate, perform appropriate hyperparameter tuning
   2. Comparative analysis of the accuracy of the models
4. Gain insights from the results obtained

**Dataset Description:**

Source: <https://www.kaggle.com/sulianova/cardiovascular-disease-dataset>

The dataset consists of 70,000 records of patients data, with 11 features and 1 target variable

There are 3 types of input features:

* Objective: factual information
* Examination: results of medical examination
* Subjective: information given by the patient

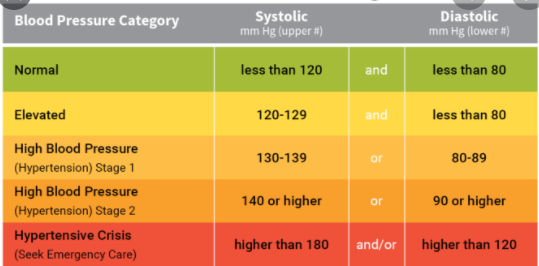
**Features:**

1. Age | Objective Feature | age | int (days)
2. Height | Objective Feature | height | int (cm) |
3. Weight | Objective Feature | weight | float (kg) |
4. Gender | Objective Feature | gender | categorical code |
5. Systolic blood pressure | Examination Feature | ap\_hi | int |

**-**Systolic blood pressure, the top number, measures the force your heart exerts on the walls of your arteries each time it beats

6. Diastolic blood pressure | Examination Feature | ap\_lo | int |

-Diastolic blood pressure, the bottom number, measures the force your heart exerts on the walls of your arteries in between beats.



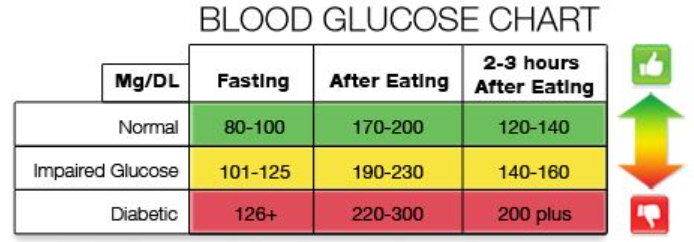
7. Cholesterol | Examination Feature | cholesterol | 1: normal, 2: above normal, 3: well above normal |

Over the years, cholesterol and fat in the blood are deposited in the inner walls of the arteries that supply blood to the heart, called the coronary arteries. These deposits make the arteries narrower, a condition known as atherosclerosis.



8. Glucose | Examination Feature | gluc | 1: normal, 2: above normal, 3: well above normal |

The blood glucose level is the amount of glucose in the blood. Glucose is a sugar that comes from the foods we eat, and it's also formed and stored inside the body. It's the main source of energy for the cells of our body, and it's carried to each cell through the bloodstream.



9. Smoking | Subjective Feature | smoke | binary |

10. Alcohol intake | Subjective Feature | alco | binary |

11. Physical activity | Subjective Feature | active | binary |

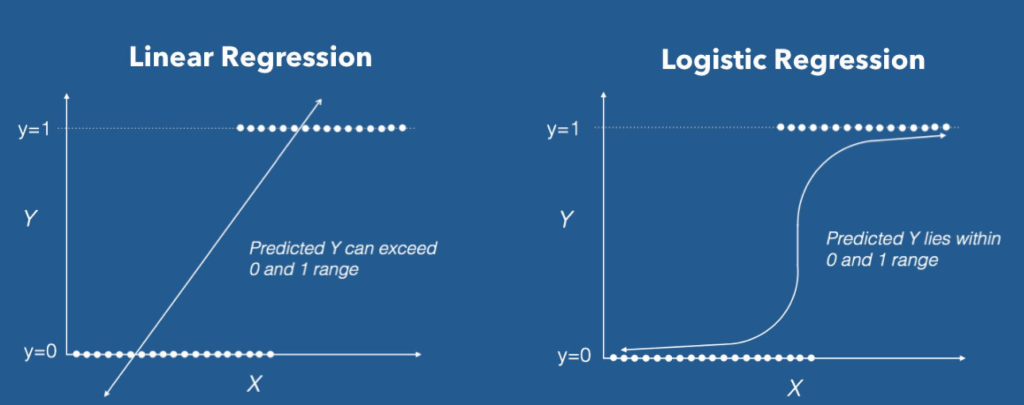
12. Presence or absence of cardiovascular disease | Target Variable | cardio | binary |

All of the dataset values were collected at the moment of medical examination.

Explorato

**Logistic Regression**

Logistic regression (LR) is a statistical method similar to [linear regression](https://www.sciencedirect.com/topics/medicine-and-dentistry/linear-regression-analysis) since LR finds an equation that predicts an outcome for a binary variable, Y, from one or more response variables, X. However, unlike linear regression the response variables can be categorical or continuous, as the model does not strictly require continuous data. To predict group membership, LR uses the log odds ratio rather than probabilities and an iterative maximum likelihood method rather than a [least squares](https://www.sciencedirect.com/topics/medicine-and-dentistry/least-square-analysis) to fit the final model.

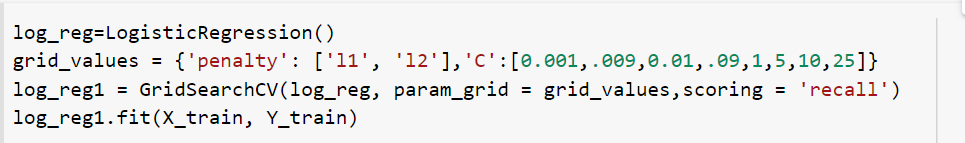


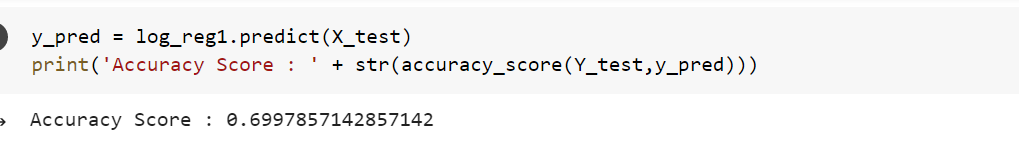
**Implementation :**

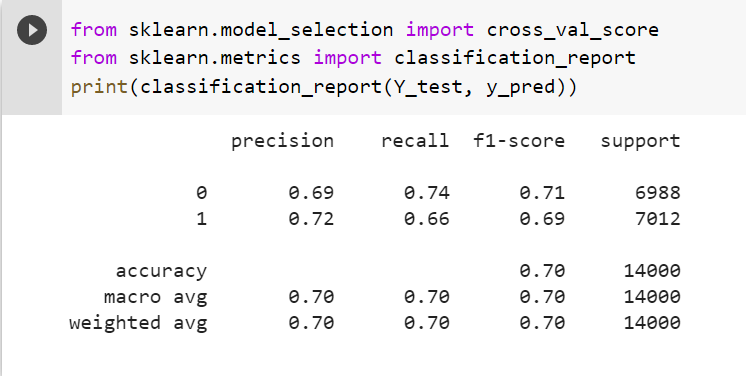
The model was initially created with default hyper parameters, fit to the training data set and later hyperparameter tuning was done using the Grid Search Model and the accuracy was measured.

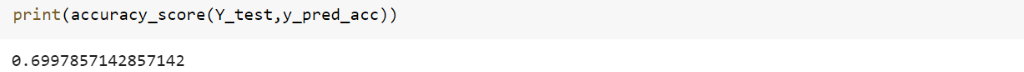
**‘Penalty’**:A regression model that uses L1 regularization technique is called Lasso Regression and model which uses L2 is called Ridge Regression.

**‘C’**  : Inverse of regularization strength; must be a positive float.



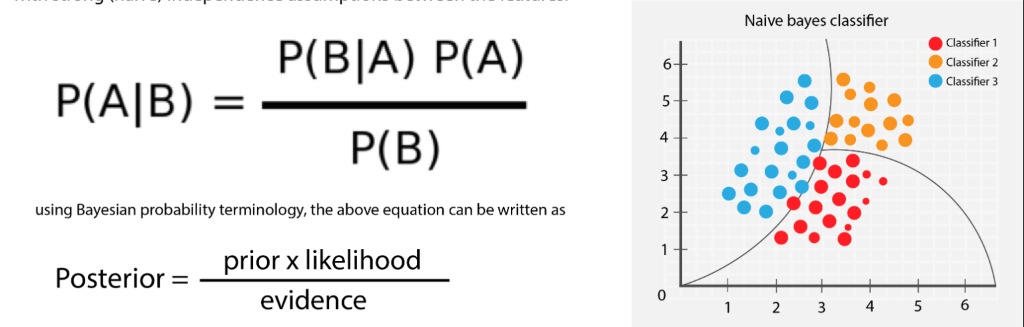






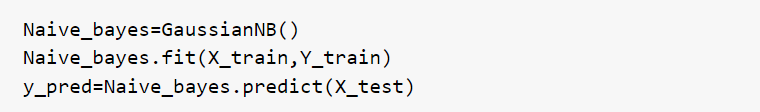
**Naive Bayes Classifier**

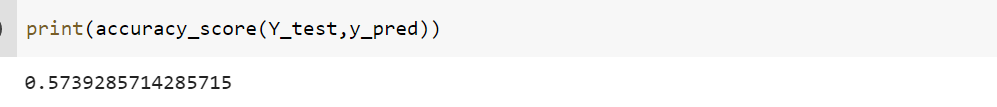
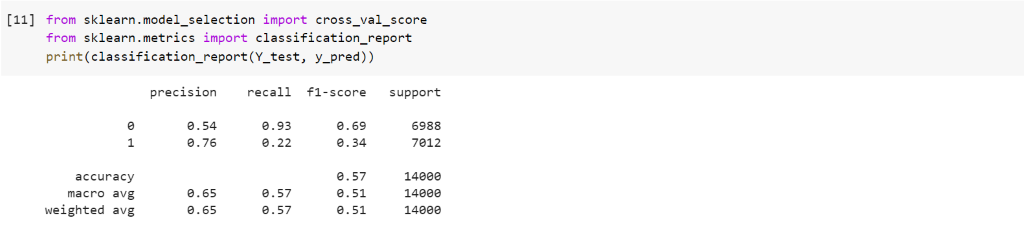
Naive Bayes classifiers calculate the probability of a sample to be of a certain category, based on prior knowledge. They use the Naive Bayes Theorem, that assumes that the effect of a certain feature of a sample is independent of the other features. That means that each character of a sample contributes independently to determine the probability of the classification of that sample, outputting the category of the highest probability of the sample.



**Implementation:**

The model was created fit to the training data set. Grid Search Model was not used for this model since it does not have any hyper parameters.



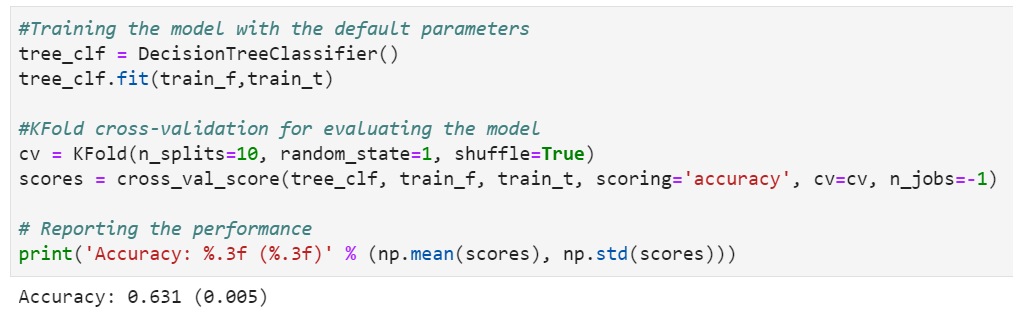


**Decision Tree Classifier:**

Decision Tree algorithm is a supervised learning algorithm that can be used for solving regression and classification problems. The goal of using a Decision Tree is to create a training model that can use to predict the class or value of the target variable by learning simple decision rules inferred from prior data(training data). In Decision Trees, for predicting a class label for a record we start from the root of the tree. We compare the values of the root attribute with the record’s attribute. On the basis of comparison, we follow the branch corresponding to that value and jump to the next node.

**Methodology followed to train the Decision Tree Classifier:**

The model was created using the default hyper parameters, fit to the training dataset and it’s accuracy was measured using the K-Fold Cross Validation technique:



Following the above step, hyper parameter tuning was performed on the Decision Tree Classifier using the Grid Search Method. The hyperparameters in tuned here are:

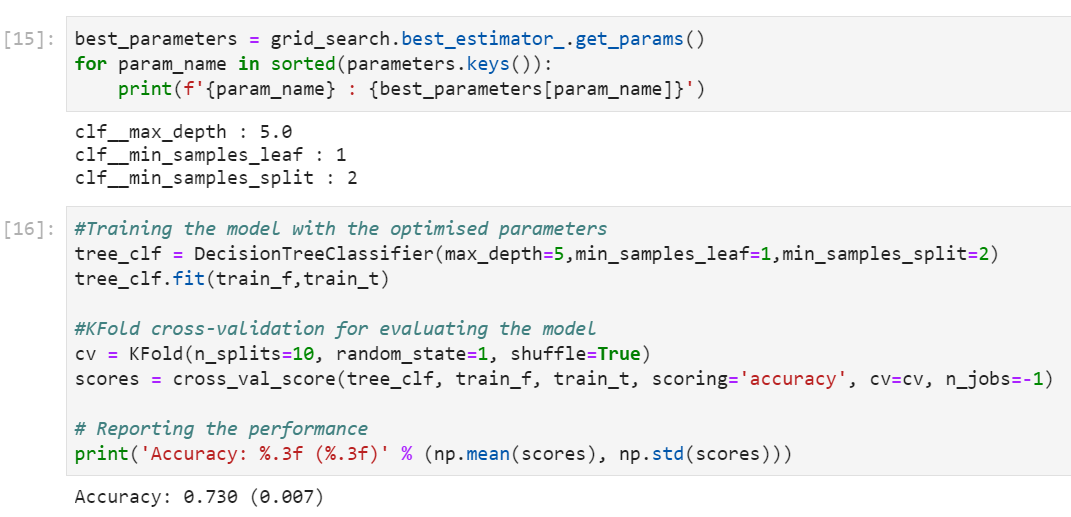
min\_samples\_split : The minimum number of samples required to split an internal node.

min\_samples\_leaf :The minimum number of samples required to be at a leaf node.

max\_depth : The maximum depth of the tree.



The model is then created and trained using the optimized hyperparameters, to result in an improvement of cross validation accuracy score.

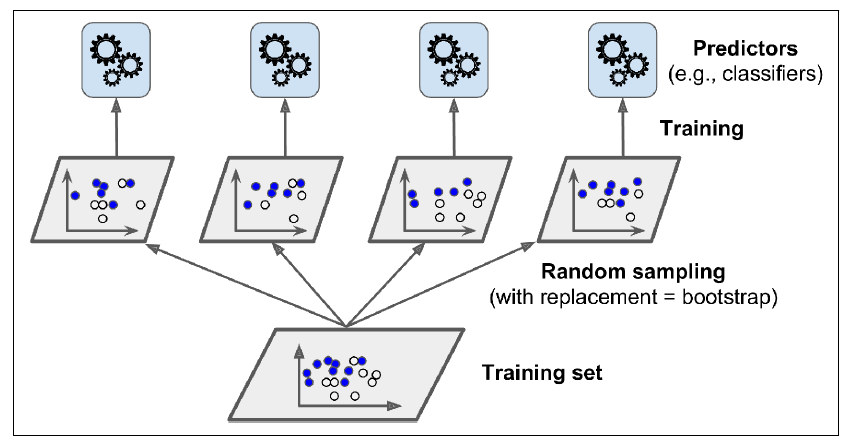


**Ensemble Methods for Classification:**

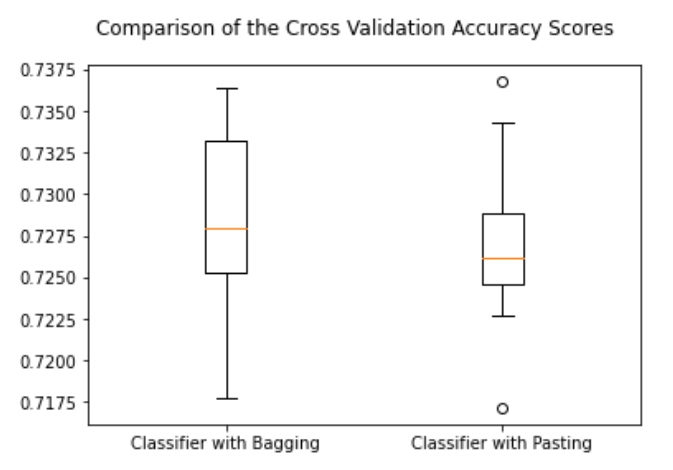
Ensemble learning helps improve machine learning results by combining several models. This approach allows the production of better predictive performance compared to a single model.

**Comparison of Bagging and Pasting:**

Bagging and pasting involves training several predictors on different random samples of the training set. The difference is in bagging selection of training sets is done with replacement whereas in Pasting it is done without replacement.



The methods were compared by training an ensemble of 500 Decision Tree classifiers and having a minimum sample size of 100, and then the models’ cross validation accuracy was compared and plotted as shown below:

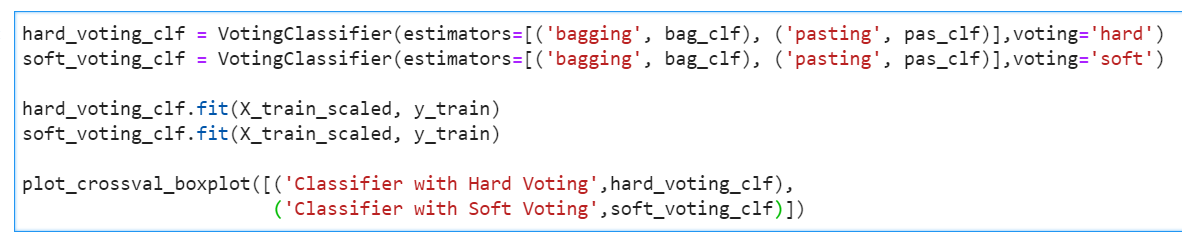


Here, ensemble learning with bagging has a higher mean cross validation accuracy as compared to pasting.

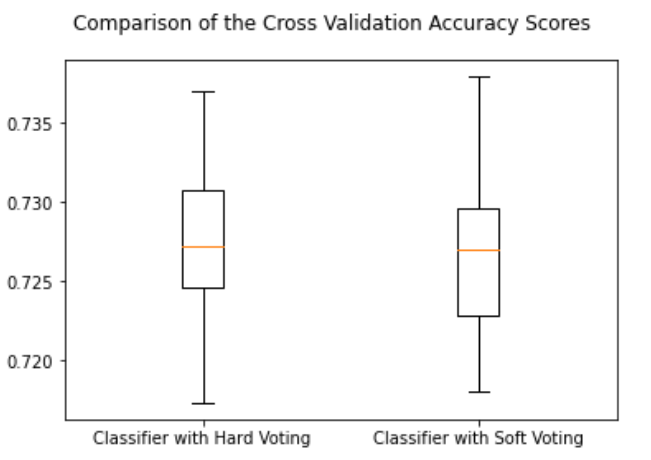
**Comparison of Hard and Soft Voting:**

Predicting the class with the highest class probability, averaged over all the individual classifiers is called soft voting whereas predicting the class probability by taking the statistical mode of the individual classifiers is called hard voting.

The methods were compared by defining two voting classifier taking in an ensemble of the bagging and pasting classifiers from the previous step



Results obtained show that both hard and soft voting almost make the same predictions and hence their accuracy scores are almost same as given in the plot below.

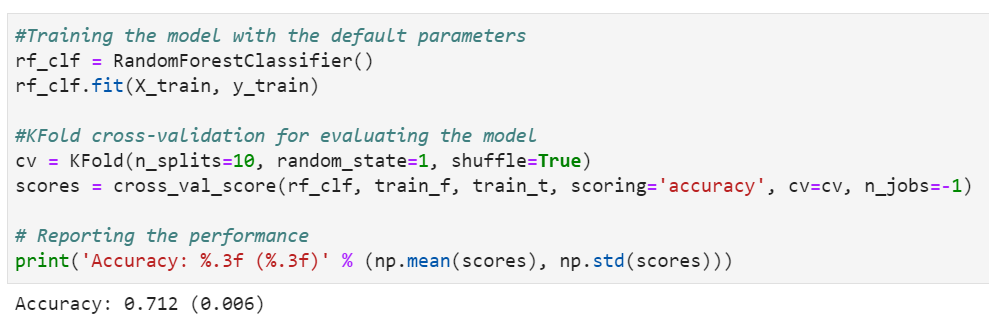


**Random Forest Classifier:**

Ensemble method involving multiple decision tress and using bagging to predict the classes of the various instances.

**Methodology followed to train the Random Forest Classifier:**

The model was created using the default hyper parameters, fit to the training dataset and it’s accuracy was measured using the K-Fold Cross Validation technique:

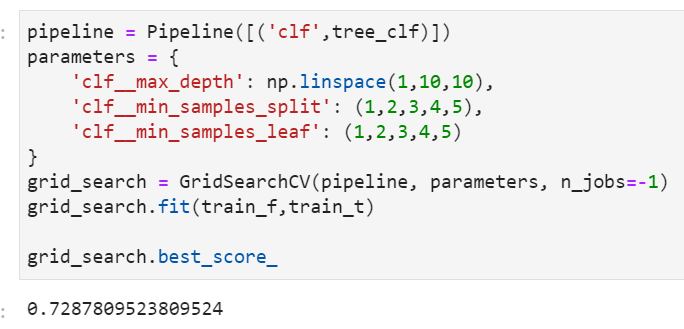


Following the above step, hyper parameter tuning was performed on the Decision Tree Classifier using the Grid Search Method. The hyperparameters in tuned here are:

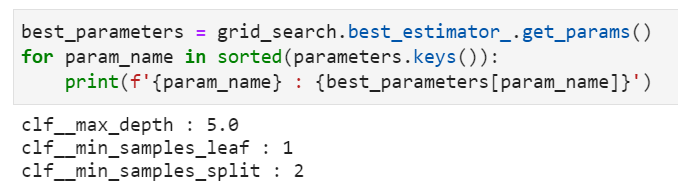
min\_samples\_split : The minimum number of samples required to split an internal node.

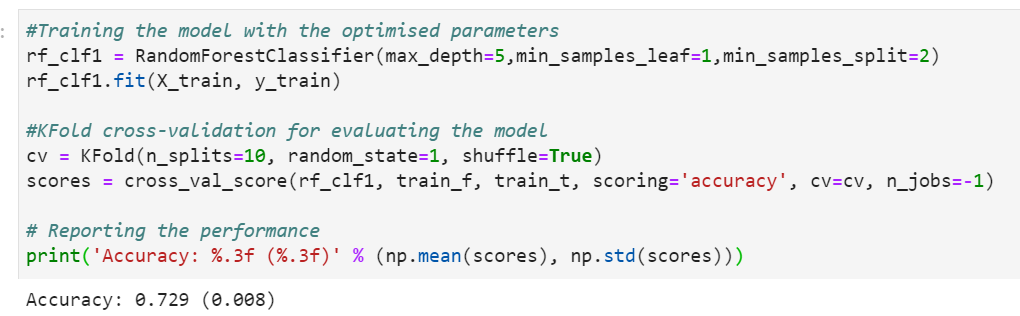
min\_samples\_leaf :The minimum number of samples required to be at a leaf node.

max\_depth : The maximum depth of the tree.



The model is then created and trained using the optimized hyperparameters, to result in an improvement of cross validation accuracy score.





**Final Comparison of the different classification algorithms:**

**Feature Importance in predicting the presence or absence of Cardiovascular Disease:**