**“Python machine learning project - detecting the loan application status will be passed or not”**

**About the data set**

This data set includes details of applicants who have applied for loan. The data set includes details like credit history, loan amount, their income, dependents etc.

**Independent Variables**

Loan\_ID

- Gender

- Married

- Dependents

- Education

- Self\_Employed

- Applicant-income

- CoapplicantIncome

- Loan\_Amount

- Loan\_Amount\_Term

- Credit History

- Property\_Area

**Dependent Variable (Target Variable)**

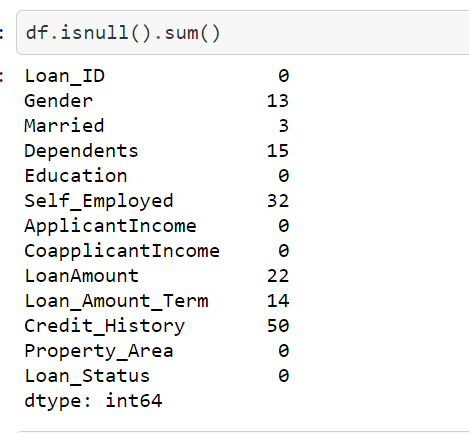
One of the major challenge today our economy , government , banking sector facing is the increase of NPAs (Non Performing Assets )

**What is NPA?**

A nonperforming asset (NPA) refers **to a classification for loans or advances that are in default or in arrears**. A loan is in arrears when principal or interest payments are late or missed. A loan is in default when the lender considers the loan agreement to be broken and the debtor is unable to meet his obligations. Therefore the goal of the project is to predict on various variables weather to approve the loan approval status or not as this will help to detect the defaulters in advance

**Missing values**

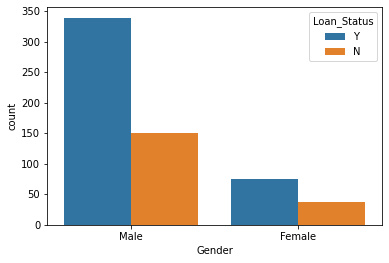
Missing Data can occur when no information is provided for one or more items or for a whole unit. Missing Data is a very big problem in a real-life scenarios. Missing Data can also refer to as NA(Not Available) values in pandas. In DataFrame sometimes many datasets simply arrive with missing data, either because it exists and was not collected or it never existed



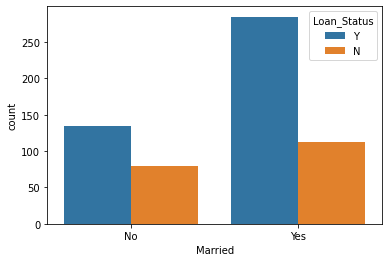
These were the various missing values present in the data set and were filled by various alternatives like mean , median and mode

**Then performing EDA (Exploratory data analysis )**

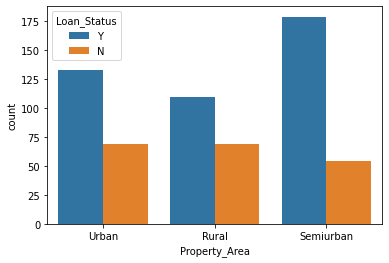
EDA in data science is quite like the service advisor doing a rough inspection of your car, asking a few preliminary questions, setting expectations and then taking the car in for service. It is one of the first things done with the data, so it is a critical phase, as many inferences and consequent actions depend on this exploration

* 

The above graph represents that more male applied for the loan approval as compared to women



Mostly married couples has applied for the loan approval but the approval rate was less as we compared with the non married or single

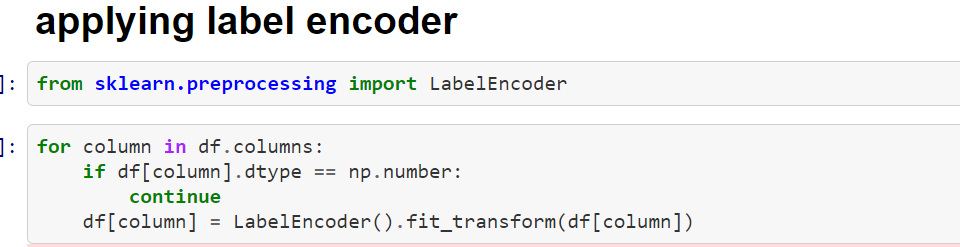


Mostly people applied form semi urban region then followed by urban and rural areas but the loan approval was slightly less in semi urban region as compared to urban and rural areas



This is our target or label column that is loan status here Y represents yes approved and N represents not approved

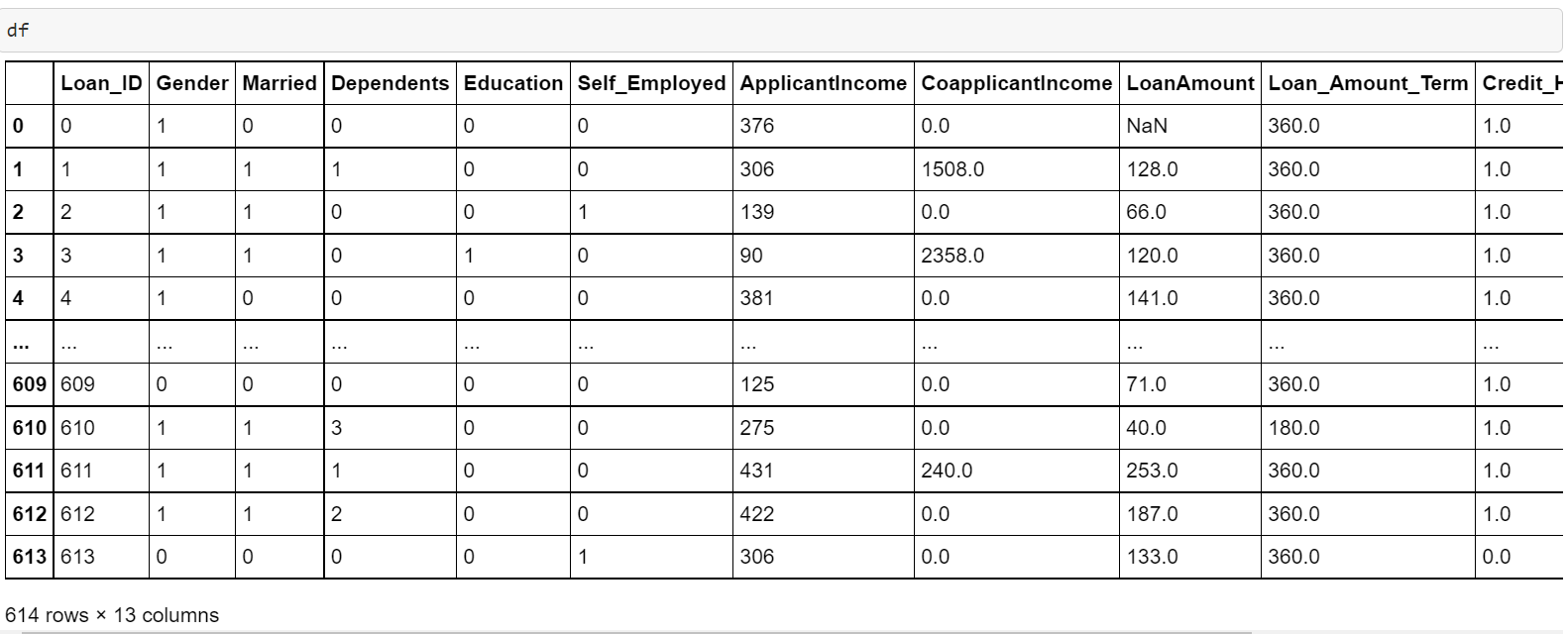
**Label encoder**

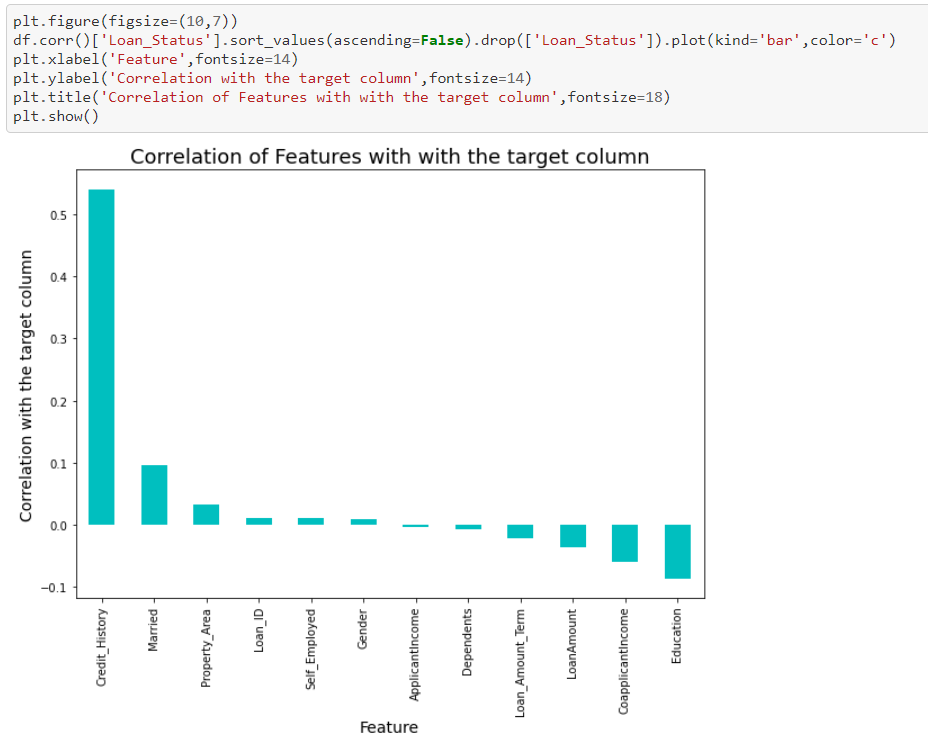


**Label Encoding**

It refers to converting the labels into a numeric form so as to convert them into the machine-readable form. Machine learning algorithms can then decide in a better way how those labels must be operated. It is an important pre-processing step for the structured dataset in supervised learning.

Our data set after label encoding is like this :



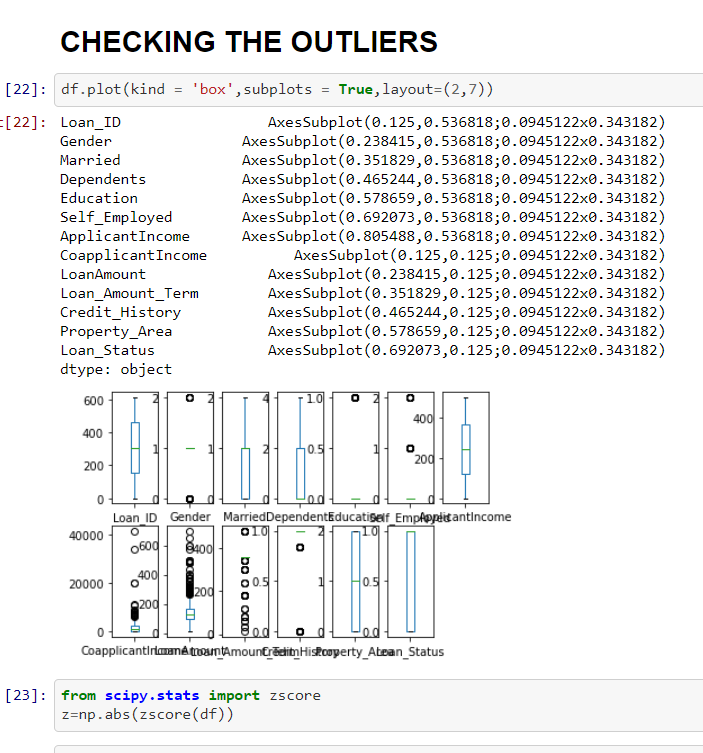


Checking the correlation with target column ( Loan status ) and from here we analyzed that there is less correlation between application income and Loan status there for it can be modified or droped as per the requirements

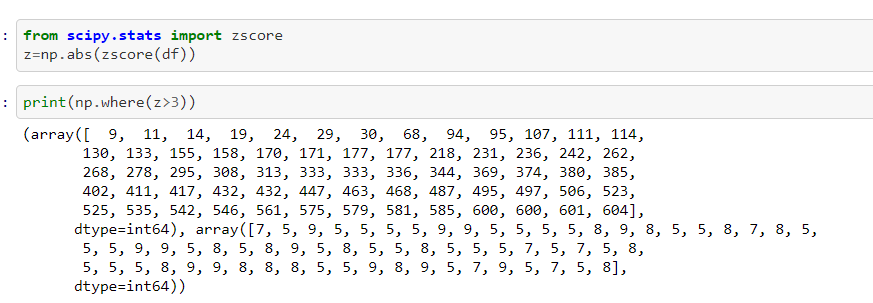
**Outliers**

An **outlier** is an object that deviates significantly from the rest of the objects. They can be caused by measurement or execution error. The analysis of outlier data is referred to as outlier analysis or outlier mining.

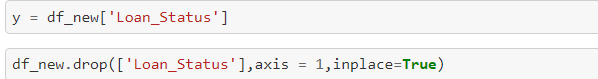
Identified outliers in the dataset



These are the identified outliers and they will be removed using the library called Scipy. Stats using the Z score and we removed the data beyond the standard deviation 3 which mean the outliers beyond 99.7 % is removed



The above are the values beyond standard deviation 3 or above the range of 99.7 % that now will be removed around 10% percent of the data or we can say the outliers were removed from the data set this will help in making the better predictions

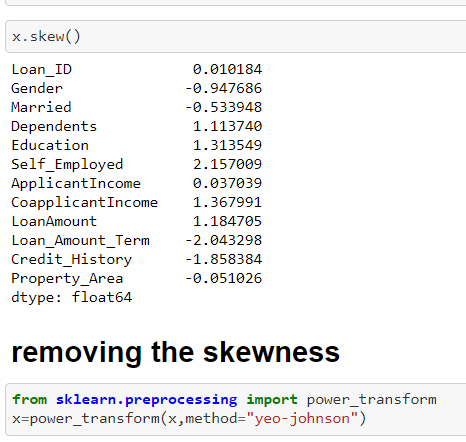


Then we selected our target column and other columns , target column will be represented with y variables

**Skewness**

Skewness is **a measure of the asymmetry of data distribution**. Skewness is an asymmetry in a statistical distribution, in which the curve appears distorted or skewed either to the left or to the right. Skewness can be quantified to define the extent to which a distribution differs from a normal distribution.

Removing the skewness



The skewness is removed from the X named variable not from the Y variable and it was removed using the power transform library which is presented In the sklearn it is important to remove the skewness as it is the part of data cleaning and it also improved the accuracy level of machine learning model

Standard scaler

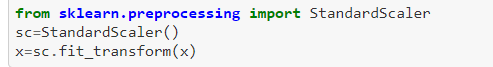
Standard Scaler helps to get standardized distribution, with a zero mean and standard deviation of one (unit variance). It standardizes features by subtracting the mean value from the feature and then dividing the result by feature standard deviation.

The standard scaling is calculated as:

z = (x - u) / s

Where :

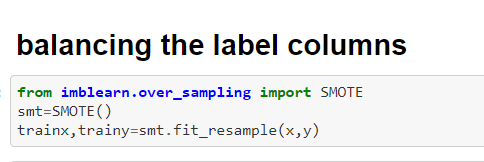
* z is scaled data.
* x is to be scaled data.
* u is the mean of the training samples
* s is the standard deviation of the training samples



There is a library called standard scaler in sklearn by using this we standardized our data which was present in the X variables

**Data balancing**

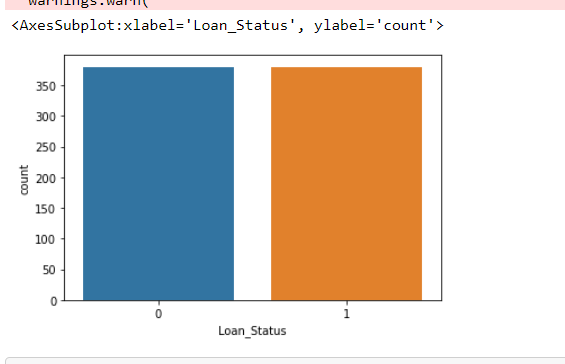
In Machine Learning and Data Science we often come across a term called **Imbalanced Data Distribution**, generally happens when observations in one of the class are much higher or lower than the other classes. As Machine Learning algorithms tend to increase accuracy by reducing the error, they do not consider the class distribution. This problem is prevalent in examples such as **Fraud Detection**, **Anomaly Detection**, **Facial recognition**, etc



## ****SMOTE (Synthetic Minority Oversampling Technique) – Oversampling****

SMOTE (synthetic minority oversampling technique) is one of the most commonly used oversampling methods to solve the imbalance problem.   
It aims to balance class distribution by randomly increasing minority class examples by replicating them.

After smote our data looks like (only to be applied over label or target column )



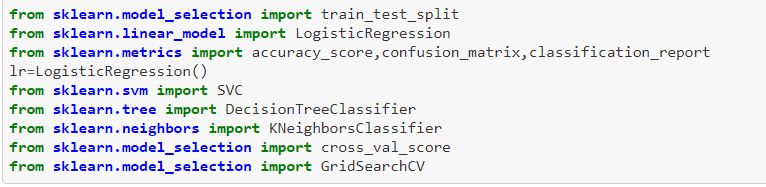
**Data bifurcation**

Here, the data the been divided into training and testing data sets in the ratio 80:20

Best Random state chosen is 997



Here in our data we have to predict weather to approve loan or not means that this data set is a classification based model so we will import or apply libraries as per the need of the data set

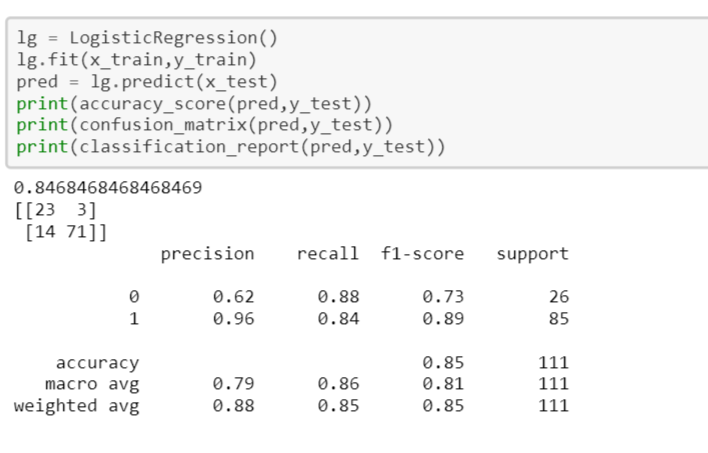


Applying the various machine algorithms to the data-set and predicting weather the loan application will be accepted or not accepted , various algorithms applied to the data set are :

* Logistic regression
* Support vector classifier
* Decision tree
* K nearest neighbors
* Xg boast
* Grid search cv (cross validation )

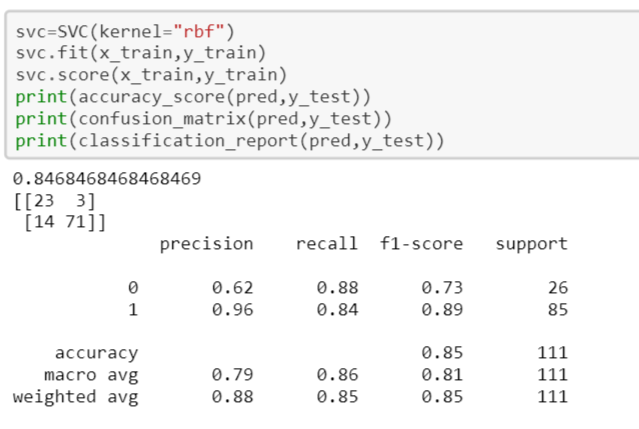
**Explanation of above algorithms**

**Logistic regression**



Logistic Regression was used in the biological sciences in early twentieth century. It was then used in many social science applications. Logistic Regression is used when the dependent variable(target) is categorical.

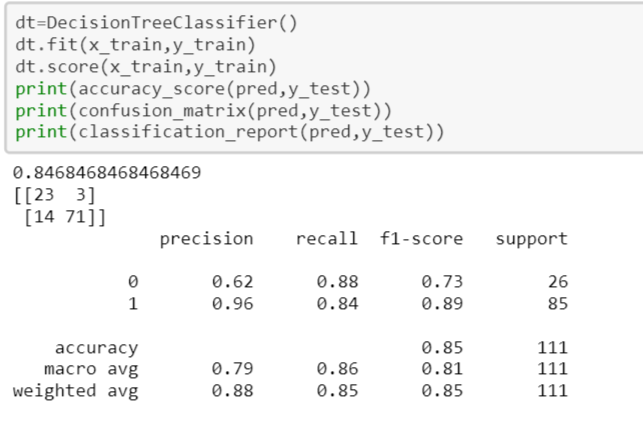
## What is Support Vector Machine?



The objective of the support vector machine algorithm is to find a hyperplane in an N-dimensional space(N — the number of features) that distinctly classifies the data points.

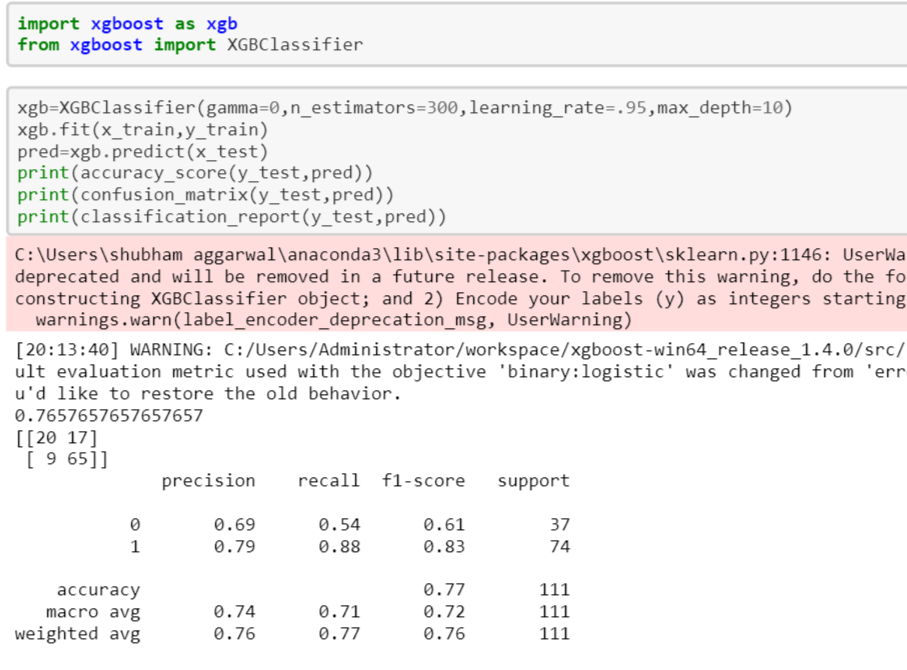
To separate the two classes of data points, there are many possible hyperplanes that could be chosen. Our objective is to find a plane that has the maximum margin, i.e the maximum distance between data points of both classes. Maximizing the margin distance provides some reinforcement so that future data points can be classified with more confidence.

**Decision tree**



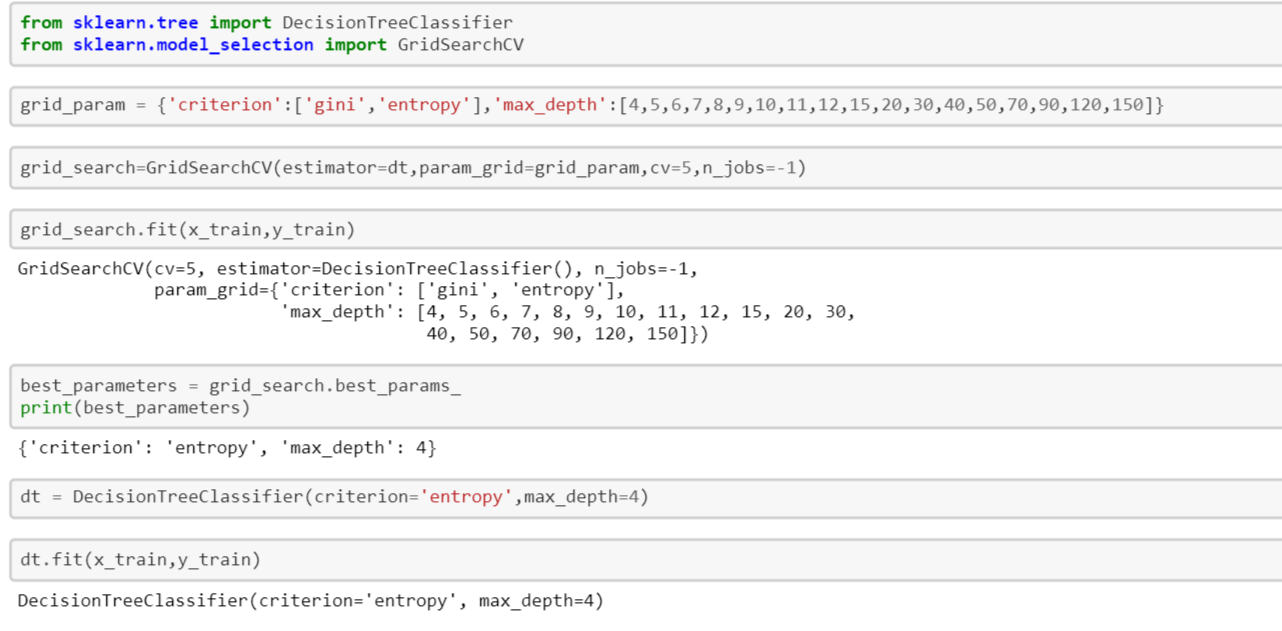
A decision tree is **a flowchart-like structure in which each internal node represents a "test" on an attribute** (e.g. whether a coin flip comes up heads or tails), each branch represents the outcome of the test, and each leaf node represents a class label (decision taken after computing all attributes).

**XG Boast**



XGBoost is a decision-tree-based ensemble Machine Learning algorithm that uses a **gradient boosting framework**. In prediction problems involving unstructured data (images, text, etc.) ... A wide range of applications: Can be used to solve regression, classification, ranking, and user-defined prediction problems.

**Grid search CV**



ridSearchCV is a function that comes in Scikit-learn’s(or SK-learn) model\_selection package.So an important point here to note is that we need to have Scikit-learn library installed on the computer. This function helps to loop through predefined hyperparameters and fit your estimator (model) on your training set. So, in the end, we can select the best parameters from the listed hyperparameters.

1. ****estimator****: Pass the model instance for which you want to check the hyperparameters.
2. ****2.params\_grid****: the dictionary object that holds the hyperparameters you want to try
3. ****3.scoring****: evaluation metric that you want to use, you can simply pass a valid string/ object of evaluation metric
4. ****cv****: number of cross-validation you have to try for each selected set of hyperparameters
5. ****5.verbose****: you can set it to 1 to get the detailed print out while you fit the data to GridSearchCV
6. ****6.n\_jobs****: number of processes you wish to run in parallel for this task if it -1 it will use all available processors

**Conclusion**

**The accuracy score or the best score predicted is 84.68 percent with the SVM , Xg boast classifier , K nearest neighbors**

**Saving the best predicted model**

**Using the pickle library the model is saved and can be used for further use**

