Introduction :- Data is the new oil. More and more data is being captured and stored across industries and this is changing society and, therefore, how businesses work. Traditionally, BI tried to give an answer to the general question: what has happened in my business? Today, companies are involved in a digital transformation that enables the next generation of BI: Advanced Analytics (AA). With the right technologies and a data science team, businesses are trying to give an answer to a new game changer question: what will happen in my business?

We are already listening how AA is helping to increase profits in many companies. However, some businesses are late in the adoption of AA, while others are trying to adopt AA but are just failing for various reasons. an illustrative example, an AA use case involving Machine Learning (ML) techniques to help HR departments to retain talent.

Employee attrition refers to the percentage of workers who leave an organization and are replaced by new employees. A high rate of attrition in an organization leads to increased recruitment, hiring and training costs. Not only it is costly, but qualified and competent replacements are hard to find. In most industries, the top 20% of people produce about 50% of the output.

This use case takes HR data and uses machine learning models to predict what employees will be more likely to leave given some attributes. Such model would help an organization predict employee attrition and define a strategy to reduce such costly problem.

The input dataset is an Excel file with information about 1470 employees. For each employee, in addition to whether the employee left or not (attrition), there are attributes / features such as age, employee role, daily rate, job satisfaction, years at the company, years in current role, etc.

We will cover our analysis on below topics.

1. Data Analysis.

2. EDA Concluding Remark.

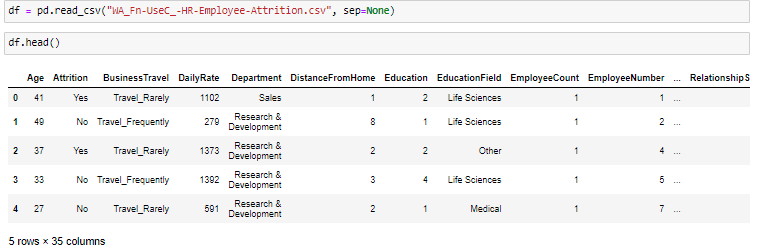
3. Pre-Processing Pipeline.

4. Building Machine Learning Models.

5. Concluding Remarks.

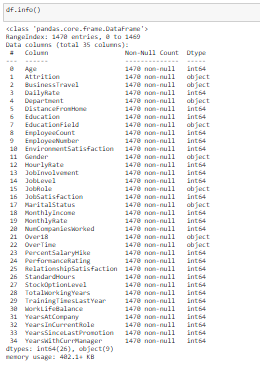
**1. Data analysis: -** At this point we will analyze the relation between independent variables and the target variable, attrition. We have created the following visualizations using Seaborn and Matplotlib library visualization library. Seaboar and malplotlib are visualization library which we can use in Python programing, most of the time we used Count plot(because there are no of variable having binary values), we have also used distribution graph to for Univariate analysis, Bar plot and box plot is also used for multivariate analysis.

* 1. **Create DataFrame:** - At this point it is good to check different variable provided in given dataset, for this we have created as dataframe from csv file.



Data set contain 35 variable, Attrition is our target variable.

* 1. **Check Data Type:** - Lets check what type of data stored in different independent variable, using info we can check datatype and null value count in dataset.



DataSet contain two type of variable, Object type variable store different string value, for example Attrition, BusinessTravel, Gender, MaritalStatus,Over18, OverTime are binary variable, these variable contain all value in ‘Yes’ or ‘No’ or 1 – 0 format, for binary content we will use simple econder to replace string value into numerical 0 or 1.

Other variable like Department, EducationField, JobRole are has multiple values for example JobRole may have different type of role in a department like manager , Clerk, Senior Manager, CEO etc. , we will use onhot encoding to replace these value with 0 and 1 by creating new variables .

Other independent variable contain integer Data type and need not to process further.

* 1. **Shape of Data:** - Dataset contain 1470 records and 34 independent variable/Feature, Target variable is Attrition.

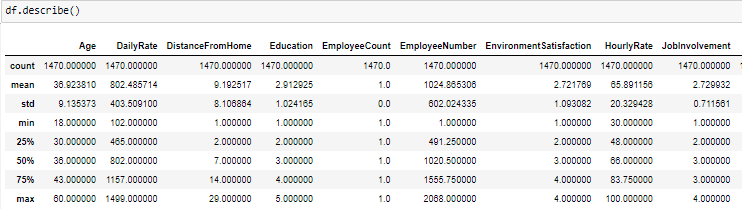


* 1. **Describe numerical value in Dataset:-** multiple columns has numerical value in this dataset, only column which contain regular data need to be analyze.

We can see that dataset do not contain null /missing value.

Some numerical columns like EmployeeNumber are numerical but these are kind of employee Id and has all values unique, hence we can remove such columns.

EmployeeCount and StandardHours has all values 0 hence such column will not be helpful in prediction and we can exclude such variables during preprocessing pipleline.

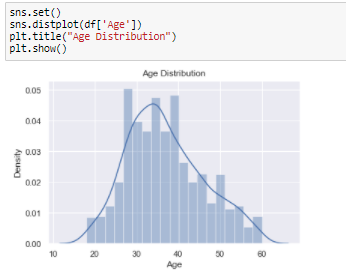


Mean and Max values has no abnormal difference, hence we can say there is no outlier, there could be some value which will be higher, but those could be valid data.

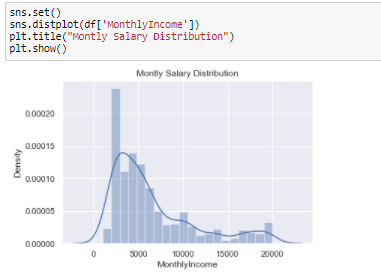
**2. Exploratory Data Analysis**

**2.1 Univariate Analysis: -** it is the simplest form of analyzing data. “Uni” means “one”, so in other words your data has only one variable. It doesn't deal with causes or relationships (unlike regression) and its major purpose is to describe; it takes data, summarizes that data and finds patterns in the data

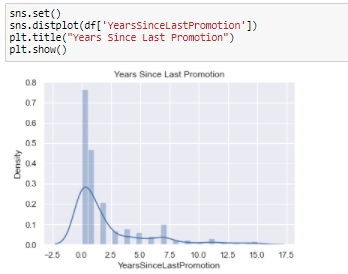
**Age Distribution:-** Below age distribution graph shows that maximum employee working in company belongs to age group 28 to 40 Year.



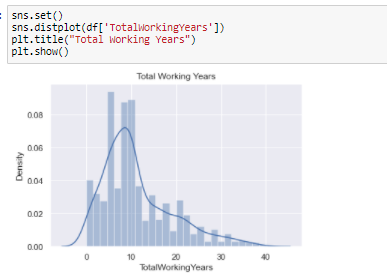
**Monthly Salary Distribution:** - Below Monthly Salary distribution graph shows that maximum employee has salary less than or equal to 5000.



**Year since Last Promotion: -** Below distribution graph shows that maximum employee recently promoted.



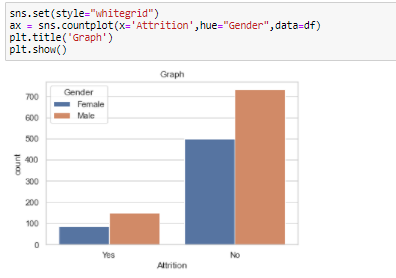
**TotalWorkingYears**: - Below distribution show that maximum employee working in organization are 5 to 12 year old.



* 1. **Bivariate Analysis:** - consists of a group of statistical techniques that examine the relationship between two variables, For example we can check how employee age and attrition are related, In this section we will analyze different independent variable with target variables.

**Gender wise attrition**: - Below graph shows that Male employee attrition rate is comparatively higher then Female Employee,

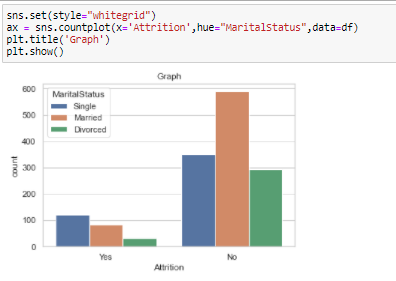
Also we can see data is imbalanced means employee leaving company has very less count compare to employee stay in company.



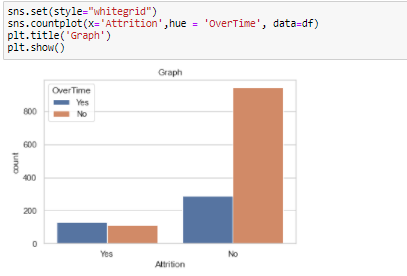
**Age wise Attrition:** - below Graph shows that Male employee has higher Attrition ration compare to Female employee of same age group.



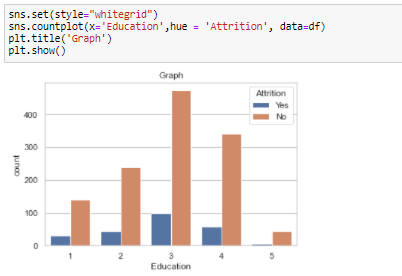
**Marital Status and Attrition:** - We can see that Bachelor employee’s attrition is higher than Married and Divorced employee.



Overtime and Attrition is also one factor which should be analyze here, as per below graph, difference of attrition of employee working in overtime and no overtime is very less but we can see that employee working in overtime has bit here count in attrition.



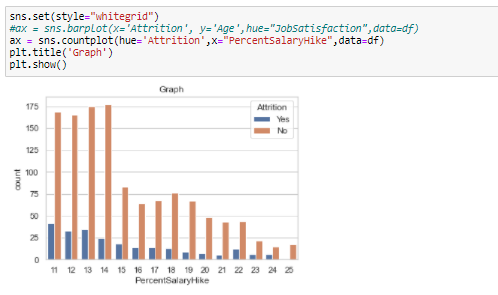
**Eduction Vs Attrition:** - We can see that employee having education level 3 has highest attrition count.



**Attrition Vs Business travel:** - Employee traveling rarely has higher attrition then employee who do not have business travel or travel frequently.



Attrition Vs PercentSalaryHike:- Salary increment play a big role in employee attrition, below graph shows that employee who got less salary hike leave the company. It is observed that most of the company gives good salary hike at the time of joining, but they have their own criteria to increase salary of existing employee, sometimes.



* 1. **Multivariate Analysis**: - Multivariate means involving multiple dependent variables resulting in one outcome. This explains that the majority of the problems in the real world are Multivariate. For example, we cannot predict the weather of any year based on the season. There are multiple factors like pollution, humidity, precipitation, etc. Here we will analyze relation between target variable and 2 independent variable.

**Distance, Gender and Attrition ratio**: - Female employee who travel long distance has higher chance

to leave job.



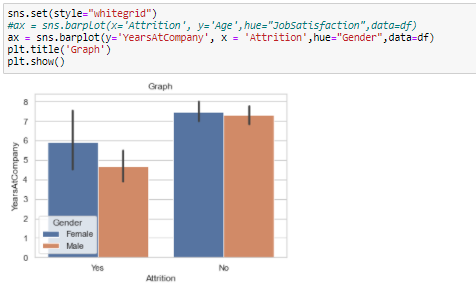
**Education, MonthlyIncome and Attrition: -** If we assume that Education field show 5 for higher education and 1 for lower education level.

1. Higher the education, Attrition will be high.
2. Below graph also show another observation that if an employee having high education and getting less salary then Attrition probability for this employee will be high.
3. most of the time Employee getting lower salary as compare to other employee with same education level have high Attrition rate.

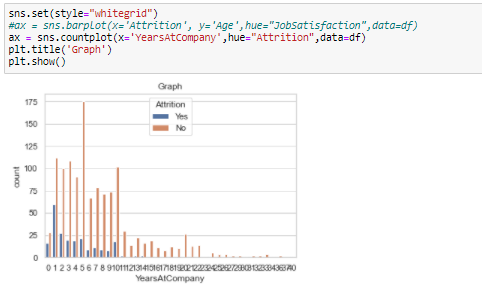


**Gender, YearAtCompany and Attrition:** - In below bar plot visual we have multiple observation.

1. Female Employee stay in company longer then male employee.
2. Most of them Male employee leave company before completion of 5 years company, while most of the Female employee leave company before completion of 6 year. .



We can see Attrition is higher for newly joined employee, means maximum attrition ratio is for the those employee who are 1 or 2 year old in company, this is very huge loss for the company because maximum expanse company do for newly joined employee for their training, hiring process and other HR related task.



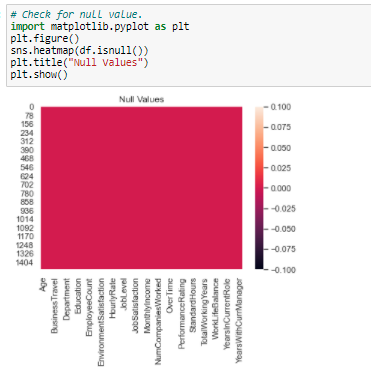
**Pre-Processing Pipeline: -** A real-world data generally contains noises, missing values, and maybe in an unusable format which cannot be directly used for machine learning models. Data preprocessing is required tasks for cleaning the data and making it suitable for a machine learning model which also increases the accuracy and efficiency of a machine learning model.

Pre-Processing required Data Cleanup. We have different steps to clean data. Will check out data using below steps and clen them accordingly.

1. **Null value Check**:- Data set sometime contain missing or null values in it, Null value should be removed if exist in data set before pushing data into Machine learning. There are multiple method to remove Missing value in data, Sometime dataset do not have null balues but it contain some special characters, we should remove those value, normally we replace such characters with NaN value first of all then apply different strategy required to remove Null value.

We can replace null value using Mean Median or Mode method,

Check if Null value exist in Dataset. We can see in below graph that null value not exist in given Data.



1. **Remove unwanted Variable**:- Some variable has all unique value or one value for all records , such variable do not help in prediction because of high Variance or low correlation with Target Variable. We can see below column has same problem and these columns can be removed from data set.

* EmployeeCount
* Over18
* StandardHours
* EmployeeNumber



1. **Data Encoding**: - In Current data set as we already found in our previous section that There are multiple column contain binary string value, which need to be replace with 0 or 1 so that model analyze the data, Machine Learning Model do not understand String or non-numerical value hence this is mandatory to encode data correctly before pushing it into Machine learning model.

There are three types of categorical variables: **binary**, [**nominal**](https://www.scribbr.com/statistics/nominal-data/), and [**ordinal**](https://www.scribbr.com/statistics/ordinal-data/) variables.

|  |  |  |
| --- | --- | --- |
| **Type of variable** | **What does the data represent?** | **Examples** |
| **Binary variables (aka dichotomous variables)** | Yes/no outcomes. | * Heads/tails in a coin flip * Win/lose in a football game |
| **Nominal variables** | Groups with no rank or order between them. | * Species names * Colors * Brands |
| **Ordinal variables** | Groups that are ranked in a specific order. | * Finishing place in a race * Rating scale responses in a survey\* |

Current Dataset contain two type of Categorical variable. **Binary** and **Nominal**.

**Binary Variable** having values in Yes or No. for this type of variable we will simple Encoding technique.

**Nominal Variable** having multiple values with Order, For Nominal Variable we will use One hot encoding which will create a separate independent variable for each type of category in Dataset and mark 1 for those record which has that category yes other wise 0.

**List of Variable are Nominal Variable**.

1. BusinessTravel
2. Department
3. EducationField
4. MaritalStatus

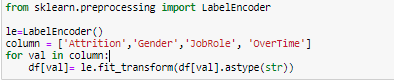
We use Onehot encoding get\_dummies method to encode these variables as shown in below print.



**List of Binary Variable**

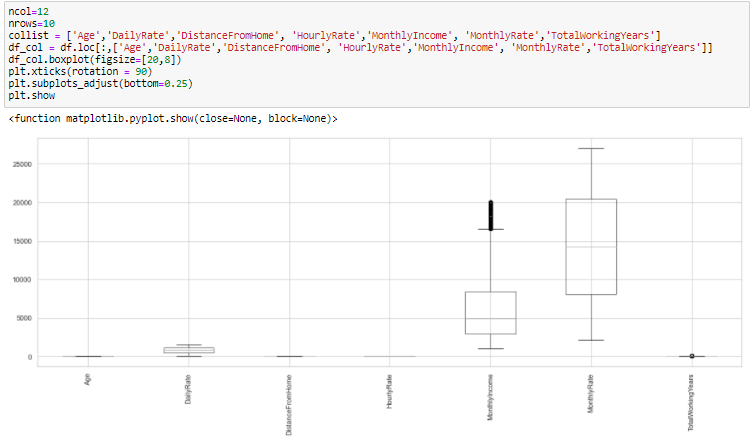
1. Gender
2. OverTime
3. JobRole
4. Attrition

Binary variable Encoding done using simple Encoding method.

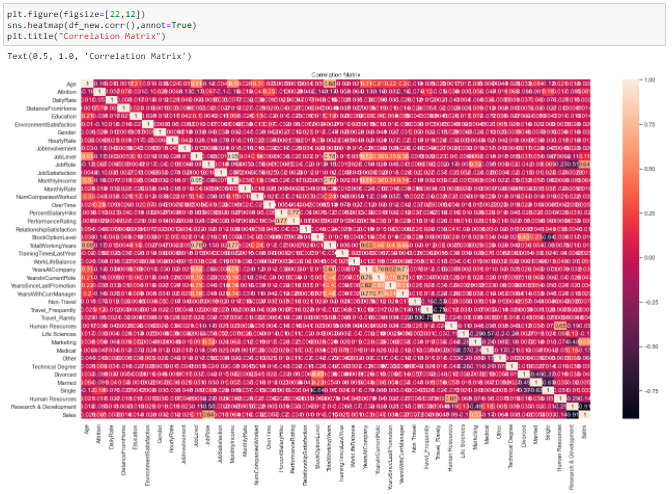


**4. Check for Outliers in Dataset: -** Outlier is an observation that appears far away and diverges from an overall pattern in a sample. Outliers in input data can skew and mislead the training process of machine learning algorithms resulting in longer training times, less accurate models and ultimately poorer results.

Outlier can be checked for variable which contain regular data. As we can see there is no outlier exist in Dataset.



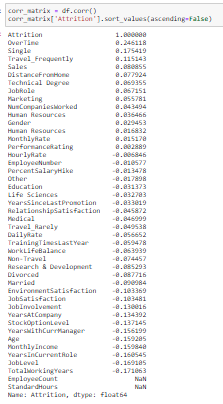
**5. Correlation Matrices**: - Correlation help us to identify relationship of variables, it shows if we have a variable strong positive relation or negative correlation or week correlation. Below Heat map show correlation of different variable, because our dataset contain lots of variable hence identify all correlation is bit cumbersome task.



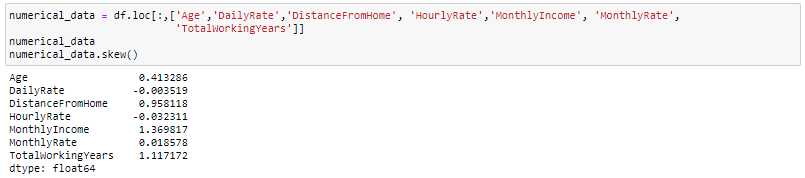
We will simply check correlation of different variable with target variable. We can see that there are few variable having almost 0 correlation with target variable. We can drop these variables and keep only those which is having high Positive or Negative correlation with target because changes in value of such variable will affect the target variable either positively or negatively.

We can delete below Variable from dataset.

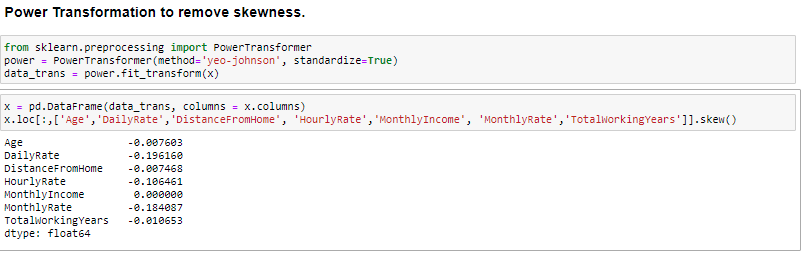
* PerformanceRating
* BusinessTravel
* StandardHours
* EmployeeCount
* EmployeeNumber



**6. Check Skewness:-** We have clean our data, not this is the time to split data in x and y where x contains all independent variable which we can to push into model and y contain the target variable. Before processing data into model we will check skewness of numerical variable. Will use power transformation to remove skewness.

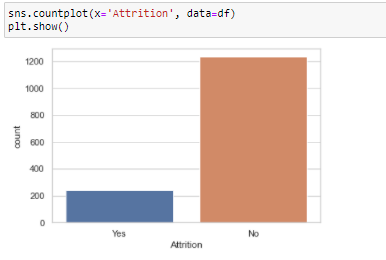


We an see that column DistanceFromHome, MonthlyIncome, TotalworkingYears has skewness. It should be between -0.55 and 0.55.



**Model Building: -** Now we have clean data set split into X and Y, We can process this data into different machine learning algorithm.

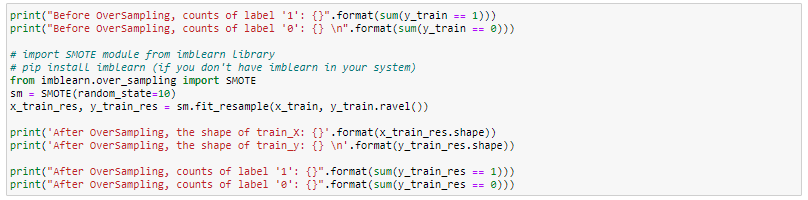
We have observed that data we have received is imbalance , imbalance data is huge difference in categorical output, here we can see Attrition is not very much compare to total no of employees in company.



The Imbalanced classification problem is what we face when there is a severe skew in the class distribution of our training data. Okay, the skew may not be extremely severe (it can vary), but the reason we identify imbalanced classification as a problem is because it can influence the performance on our Machine Learning algorithms,Data Imbalance can be handle using Oversampling or Undersampling .

* **Oversampling**— Duplicating samples from the minority class
* **Undersampling** — Deleting samples from the majority class.

For this we have use imblearn.over\_sampling library and SMOTE function.



Before OverSampling, counts of label '1': 154

Before OverSampling, counts of label '0': 698

After OverSampling, the shape of train\_X: (1396, 40)

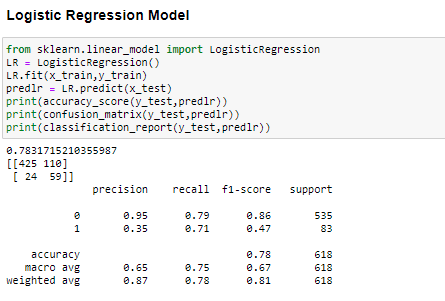
After OverSampling, the shape of train\_y: (1396,)

After OverSampling, counts of label '1': 698

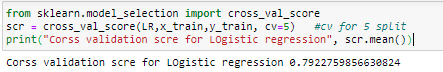
After OverSampling, counts of label '0': 698

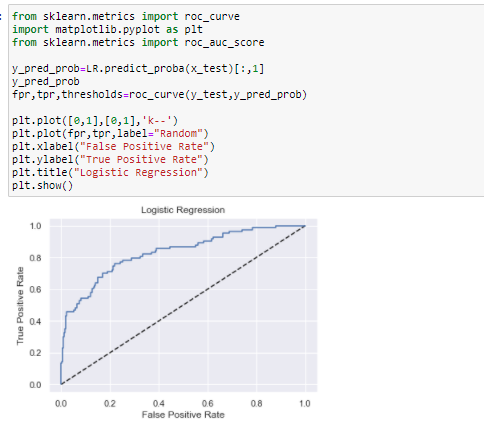
Now we can process our Data in different Machine learning models. We use below Model to learn and predict data.

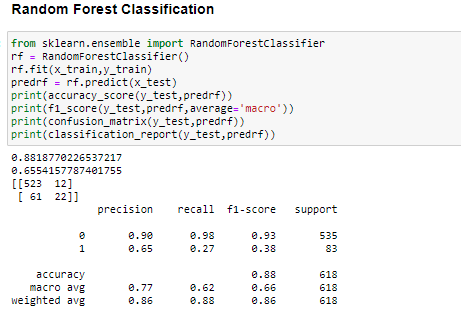
* Logistic Regression
* Random Forest Classifier
* KNeighborsClassifier
* Support Vector Classifier
* AdaBoost Classification

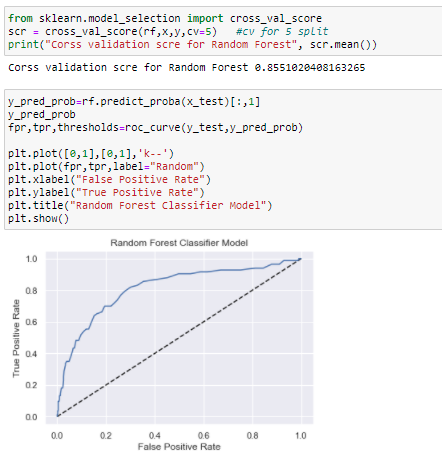


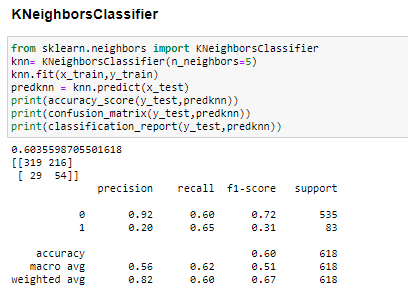
**Cross Validation:-**

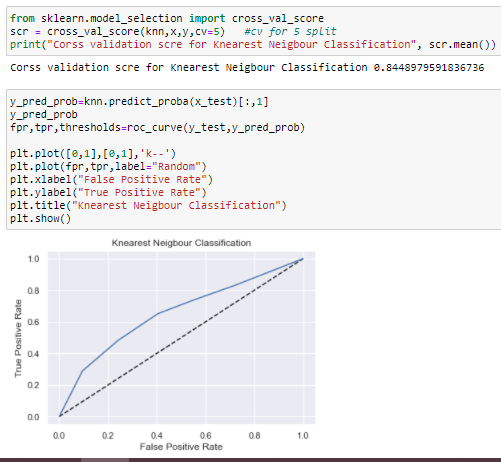


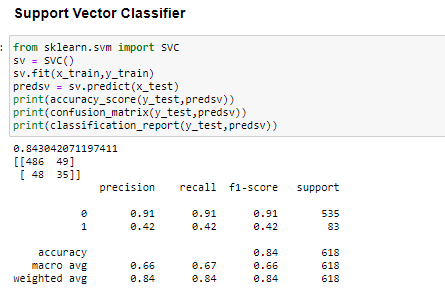


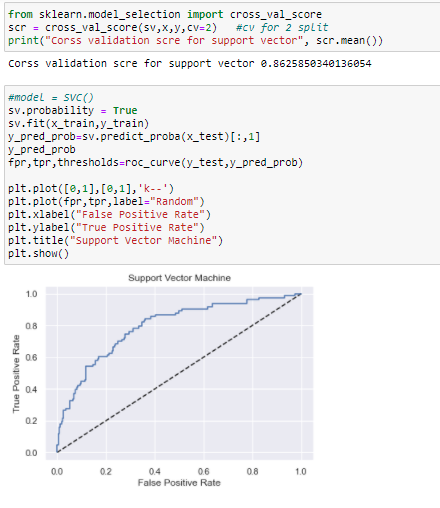


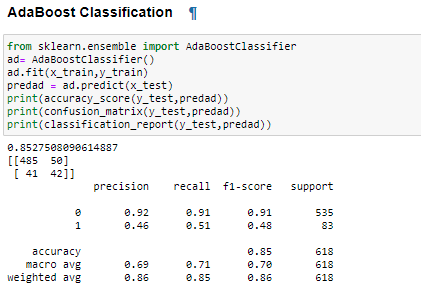


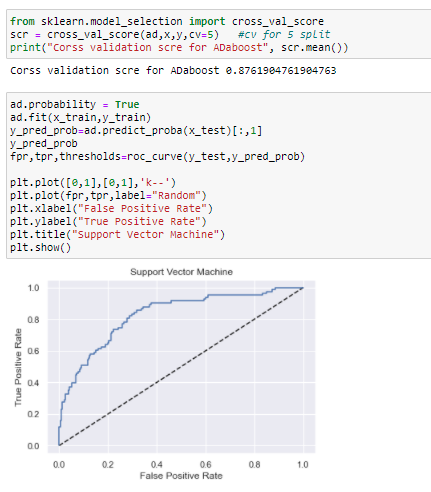




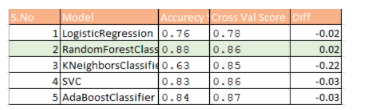






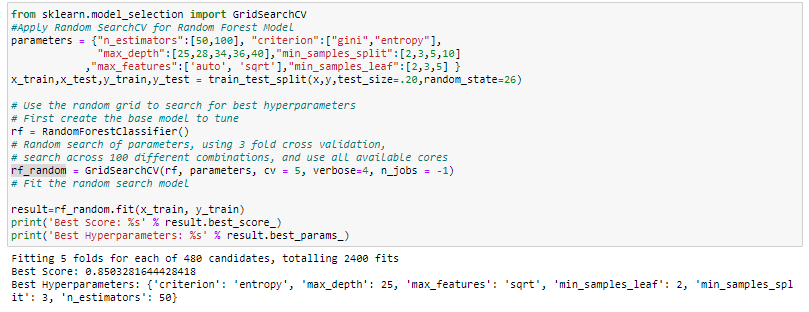


**Compare Model Performance:-**

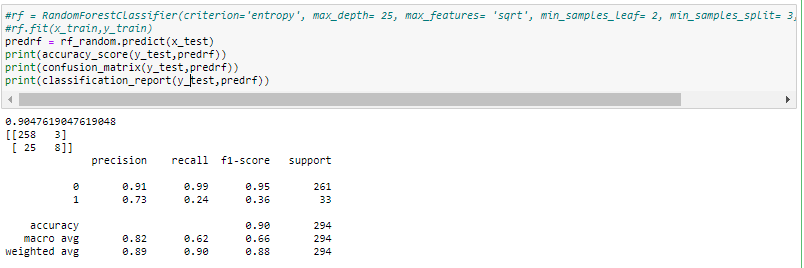


Best Model is Random Search Classification as it is giving highest 88% accuracy and cross validation score is also 86%.

**Hyper parameter Tuning:** - We will further perform Hyper parameter tuning for Random Forest model, for this we will use GridSearchCV method.



So we have received best accuracy score and Parameters, will predict value using these value and check accuracy .



Accuracy score with best parameter coming 90%, which is good score.