**Problem Statement:**

Every year a lot of companies hire a number of employees. The companies invest time and money in training those employees, not just this but there are training programs within the companies for their existing employees as well. The aim of these programs is to increase the effectiveness of their employees. But where HR Analytics fit in this? and is it just about improving the performance of employees?

**HR Analytics**

Human resource analytics (HR analytics) is an area in the field of analytics that refers to applying analytic processes to the human resource department of an organization in the hope of improving employee performance and therefore getting a better return on investment. HR analytics does not just deal with gathering data on employee efficiency. Instead, **it aims to provide insight into each process by gathering data and then using it to make relevant decisions about how to improve these processes.**

**Personal Experience:**

I choose HR attrition dataset for writing a blog, the biggest reason behind to select this dataset is that I have the experience of working with such a team(HR team) for 1 year. I personally attached to this dataset I will share my personal experience here also. When an employee leaves the company, what is the reason behind it and what can we improve keeping that reason in mind. This is the biggest challenge in HR.

Whenever an employee leaves the company, one of the HR team member takes a short interview discussion with him about the company, in otherwords he takes a feedback from a employee. That feedback is recorded and based on the empolyees feedbacks management takes  take appropriate measures to prevent it.

A major problem in high employee attrition is its cost to an organization. Job postings, hiring processes, paperwork, and new hire training are some of the common expenses of losing employees and replacing them. The company also provides some training and some additional things their cost is also involve on it. Additionally, regular employee turnover prohibits your organization from increasing its collective knowledge base and experience over time. This is especially concerning if your business is customer-facing, as customers often prefer to interact with familiar people. Errors and issues are more likely if you constantly have new workers.

**About The Dataset:**

In this dataset, we are going to discusss employee attrition prediction i.e. predicting that employee will leave or resign from the current company and for that we will do use some machine learning algorithms.

**Process of HR Analytics:**

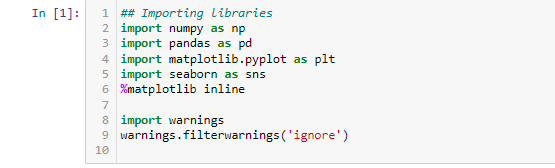
HR Analytics is made up of several components that feed into each other.

* To gain the problem- solving insights that HR Analytics promises, data must be first collected.
* Then the data needs to be monitored and measured against other data, such as historical information, norms and averags.
* This helps the management to identify trends and patterns.
* The last and final step is to apply insight to organizational decisions.

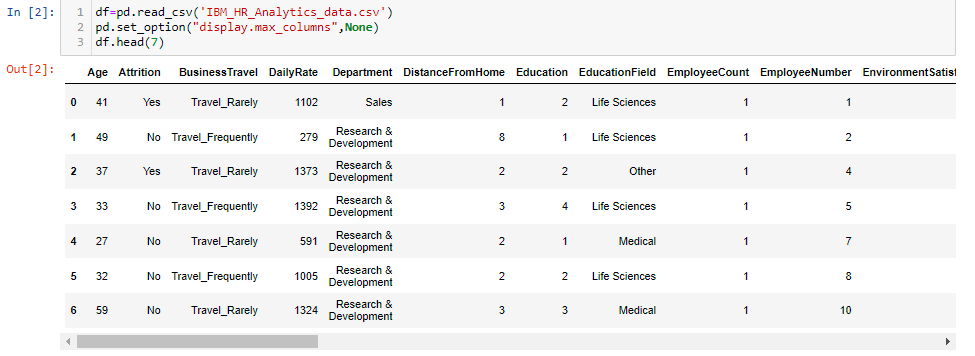
Let’s see that how A Machine Learning will help to predict the attrition of the employee.

Let’s Get Started:::

#### Importing necessary librabries for now:

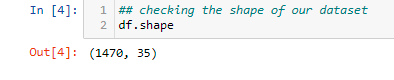


**Importing The Dataset in jupyter notebook:**



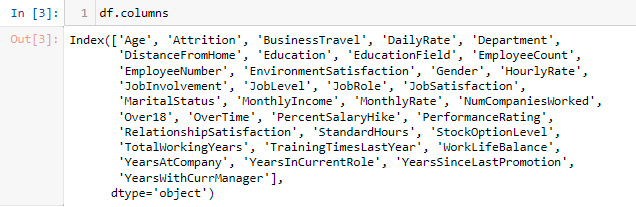
As our dataset is in csv form so we will use pd.read\_csv fuction to call the dataset and give path of the location, where dataset is saved.

Now first of all we will check from below method that how big our dataset is?



We can see that there are 1470 rows and 35 columns present including target. Here we will keep in mind that we have small number of rows or imformation present so we have to think one more time before dropping any data. In real time we have millions of rows present that time we can drop some unnecessary rows , but it is not that type of case.

Now we will which are the attributes/columns present in my dataset:



## About the columns:

**Target:-**

Attrition:- identifying why employees voluntarily leave, what might have prevented them from leaving, and how we can use data to predict attrition risk.

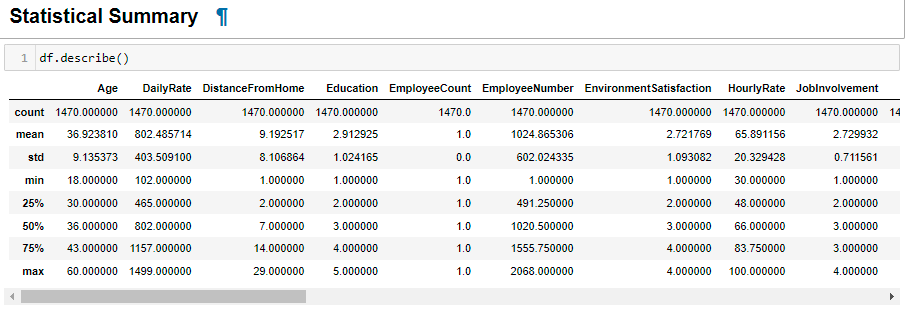
**Features:-**

almost all of the features are self explanatory, here I am explaining some important features which might be new for fresherss.

As

* EnvironmentStatisfaction:- how much satisfied with the environment of the organizations they work for. The bigger the number, the more satisfaction.
* HourlyRate:- it is the amount of money that is charged,paid or earned for every hour worked.
* Job Level:- It is much similar with designation. The higher the number the bigger the post.
* job Satisfaction:- It is much similar as rating.

Now We will find some additional statistics information about the dataset. The rough information we can say!!!

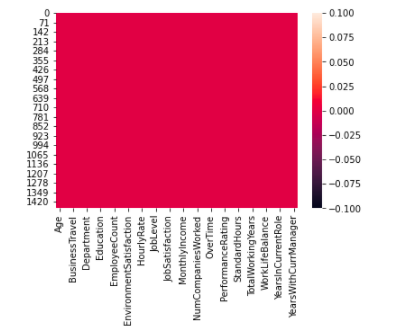


We observe some points in that:

* The maximum age of the employee as working with an organization is 60. Which is obvious that after 60 year of age every one have to get retire from his job.
* The maximum rate of employee per hour is 100 and minimum is 30
* Seems Employee Number is an nomial data. It is only for name sake.
* Seems there is no Nulls present in the dataset.

Now we will Check nulls if present then we have to handle them:





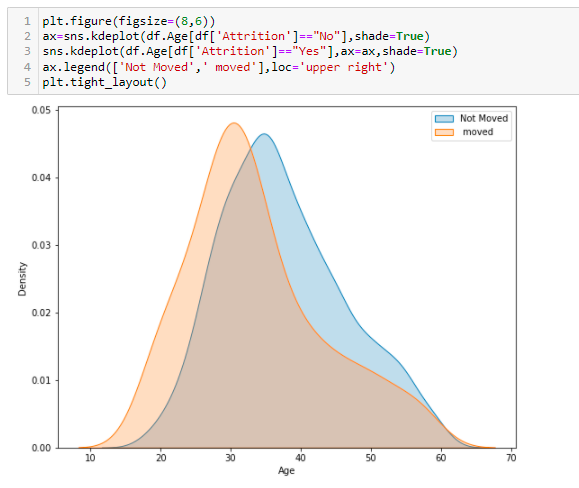
We can see that there is no nulls present in my data set.

**EDA (Exploratory Data Analysis):**

Exploratory Data Analysis is the crucial process of using summary statistics and graphical representations to perform preliminary investigations on data in order to uncover patterns , detect anomalies and verify assumptions.

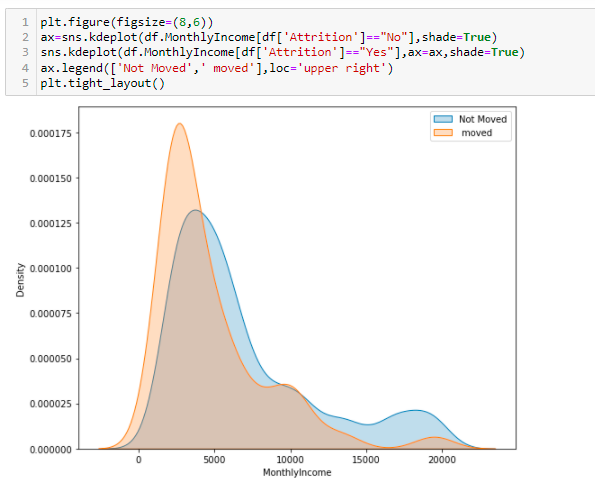
We will find the patterns in data through data visualization.

* **Attrition VS Age:**

****

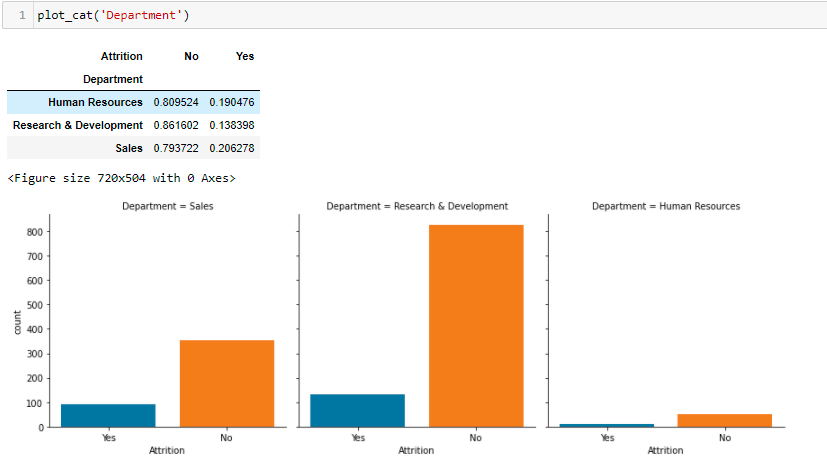
**Observation:-**  As we can see that above chart the attrition is maximum between the age groups 27 to 33. The attrition rates keep falling down while increasing age, as people look stability in their jobs at these point of times. Also we can see the very younger aged employees have far more chances to leaving the organization since they are exploring at that point of time .

**Monthly Income Vs Attrition:** Is income the main factor for employee attrition?

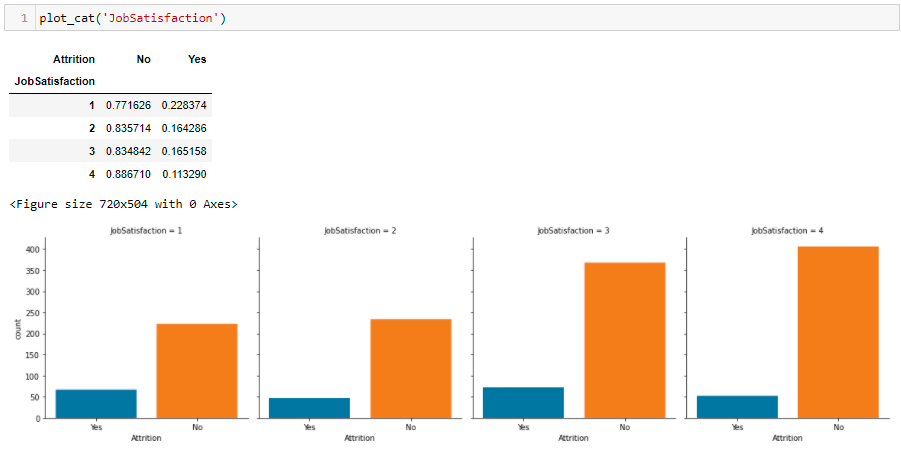


**Observations:-** As we can see in the above graph, the attrition rate is high at very low income as 5k per month. When the monthly income is quite good , the chances of an employee leaving the organization is very low as seen by downward lines. But between 8k to 13k the middle class employee tend to shift towards a better standard of living.

**Department Vs Attrition:-**

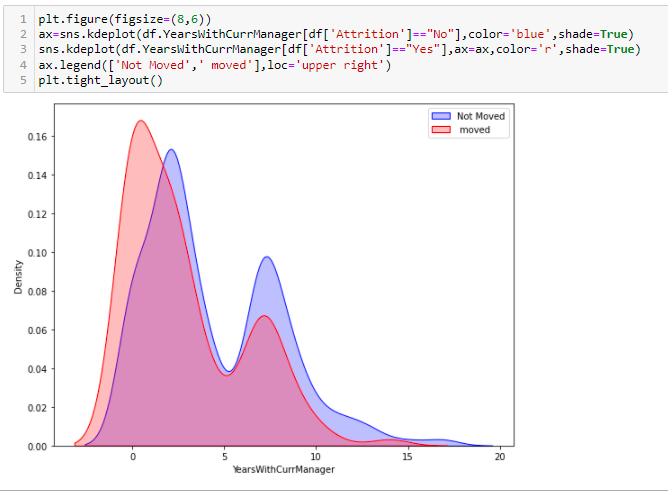
 Our dataset is distributed only 3 main departments. We can see from above table that Sales department has the highest attrition rate(20.6%), Human Resources department has second highest attrition rate(19.04%) and R&D department has least attrition rate(13.8%).

**Job Statisfaction Vs Attrition:-**

****

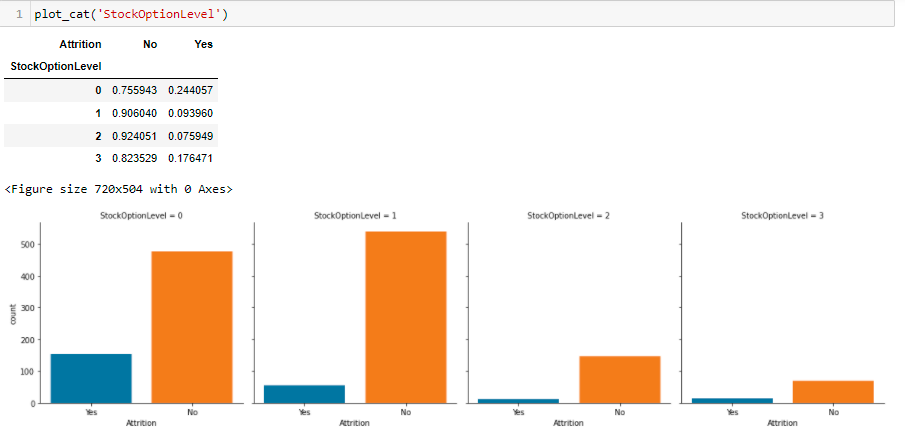
**Observation:-** Here we can see a kind of up-trend. As JobSatisfaction rate is increasing the employee don’t want to leave organizations. The attrition rate is high as (22%) at job satisfaction rate 1.

**YearsWithCurrManager Vs Attrition:**  Here we are analyzing the relationship of an employee with their manager.

****

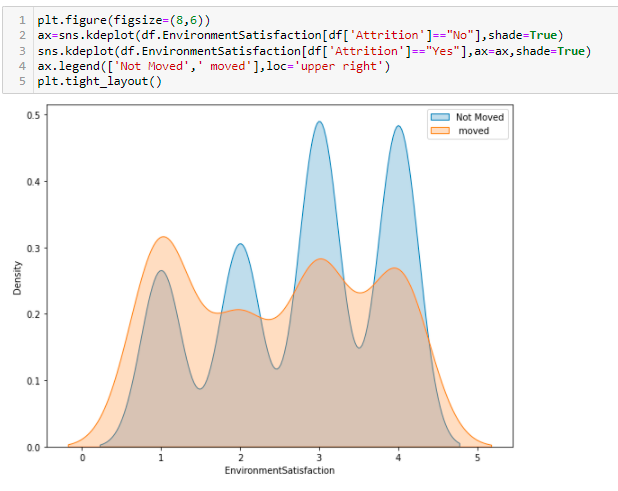
**Observations:-**  We can see there are 3 major spikes in attrition here. At the very start , when the time spent with the manager is very less people want to leave their jobs, may be their understanding does not match. After spending 5 or mare then 5 year with the manager the employee feel they need an improvement, therefore they also want to go for a new job. But when the quite decent time spend with the manager then the prople did not go for a change, may be they are statisfied with their jobs.

**StockOptionLevel Vs Attrition:-** Stock options are a type of alternative compensation that some companies, including many startups, offers as a part of their packages for employee.



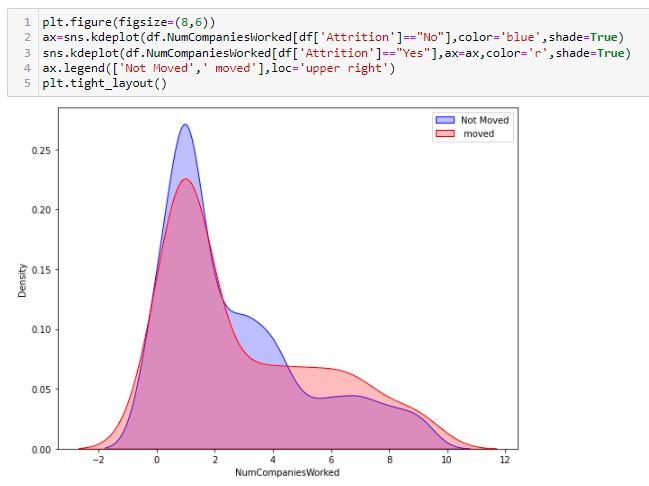
**Outputs of above graph:-** Employee has high chance to leave the organization when stock availing options are limited we can see in above graph and table as well. When StockOption level is zero the attrtion rate is highest and as level is increases the chance of attrition is also decreases.

**EnvironmentSatisfaction Vs Attrition:-**



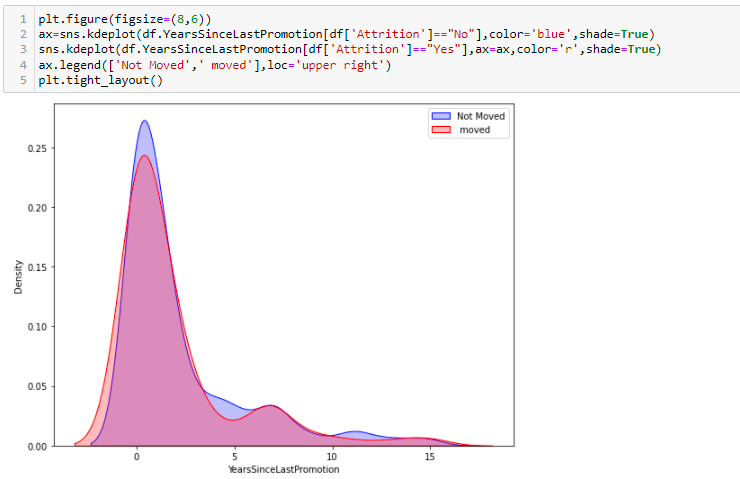
**Observation:**  When Environment Satisfaction rate is 1 the people tend to move on to get better opportunities and experiences. But as the Envionment Satisfaction rate is increases the chance of people leaving the organization slightly decreases.

**Experience VS Attrition:-**

****

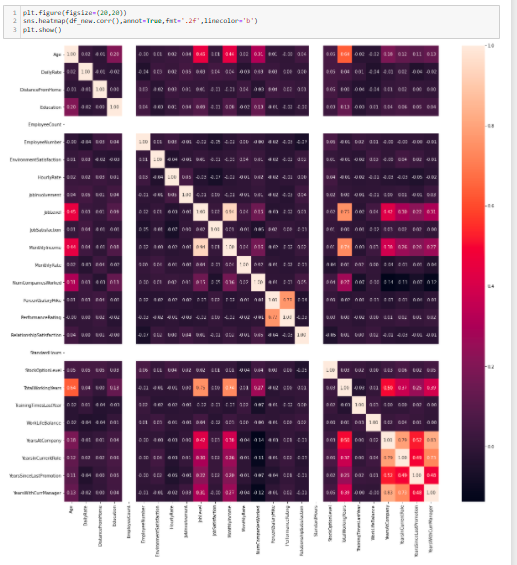
**Observation:-** As we can see from above graph, those employees who have started their career with the company, they leave the company at their initial year of career. Those people have more companies experience their attrition rate is least.

**Years In Last Promoted Vs Attrition:-**

****

**Observations:-** We can clearily see in above graph that there is less chance to leave the company by employee after initial years of promotions. But as the no. of year is increasing the chance of attrition is also increasing and it is obvious that if company didn’t take care them , then they will definitely leave the company. Here one thing to notice that at 7-8 year we can see there a little spike it means that after seven to eight year from he last promoted their chance of leaving is increases.

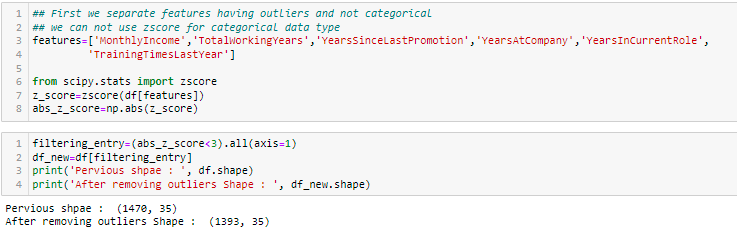
**Plotting Heatmap:** Heatmap helps us to find the relationship between the features with features and features with target:



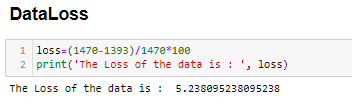
**Observations of Heatmap:-**

* Monthly Income is very high correlated with joblevel. It could be happen as senior employees must have earn more .
* performanceRating is highly correlated with percentsalaryhike. It is because every employee salary hike depends on one of the factor as performance ratings.
* Also total working year is highly correlated with monthly income and job level. which is obvious that senior employees must have pay paid larger of time.
* YearAtCompany is quite enough related with YearswithCurrmanager , which is fine.
* and also yearAtCompany is related to yearinCurrenRole. Which could be obvious.

**Checking Outlier:-** An outlier is an observation that appears to deviate markedly from other observations in the sample. An outlier may indicate bad data so that we have to remove them. For that we simple use zscore method to remove outliers.

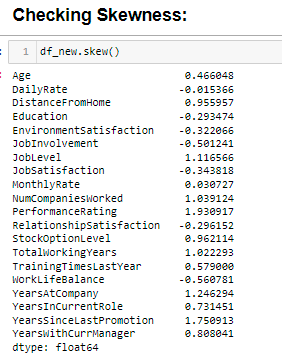


Here we can see that there are 77 outliers present in our dataset and we remove them. But before going ahead we have to check the dataloss, as our dataset is very small so we have to take care of dataloss.

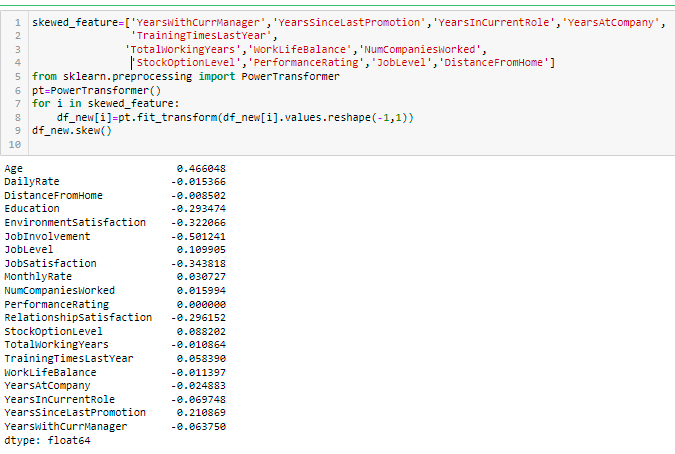


As we see that dataloss is 5.2%which could be acceptable. So we can move ahead.

**Checking Skewness:-** skewness is used along with kurtosis to better judge the likelihood of events falling in the tails of probability distribution.



We can see there are many features have highly skewed data. We will handle their skewness using power transformation.



We can see that we have successfully removed the skewness from the dataset. We have check and handle all the necessary things for our dataset.

# Encoding For Categorical data.

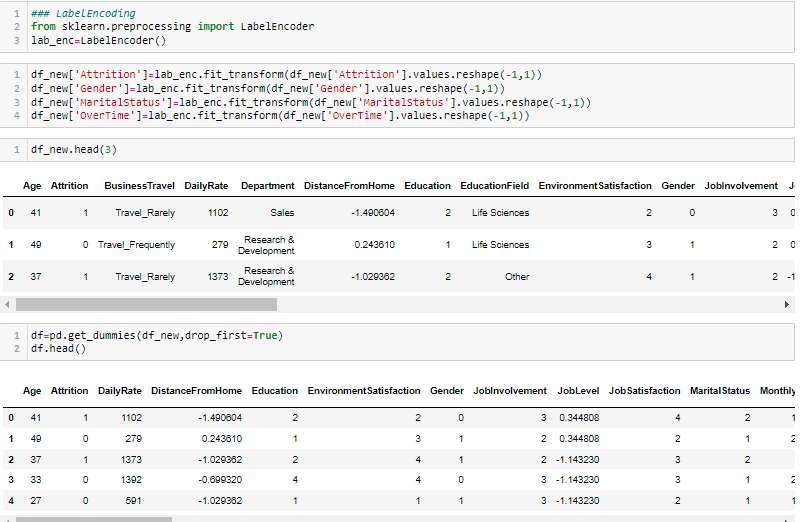
We have some feature having object data type so before building a model we have to encode it so that our model understand these features as well.

#### Label Encoding

* Attrition
* Gender
* MaritalStatus
* OverTime

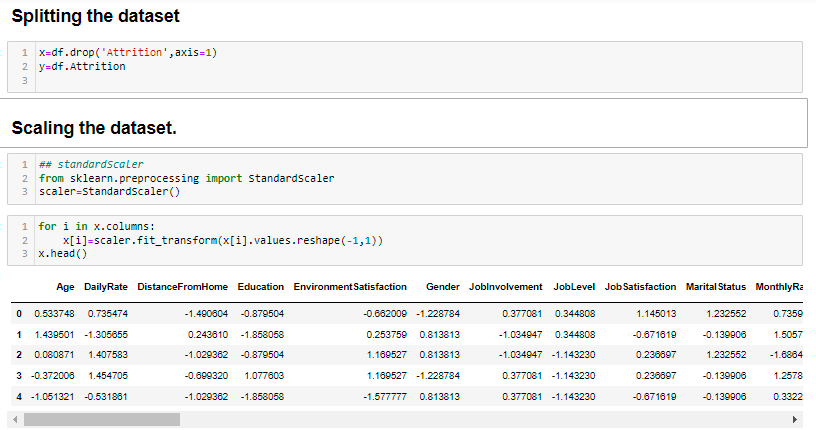
### get\_dummies for

* BusinessTravel
* Dapartment
* EducationField
* JobRole

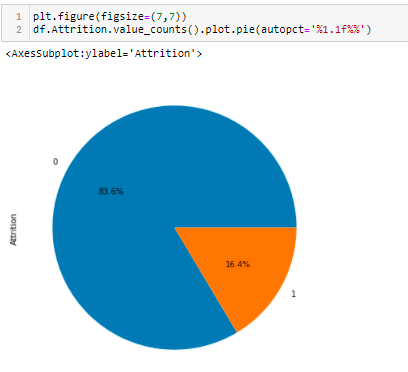


**Splitting and Scaling:-** The machine learning algorithm works on numbers and does not know what that number represents i.e. if the number is higher than others model give more importance it than others and this leads to our model biased So we have to scaling the dataset to similar range pattern for better result.

These are the steps showing below for splitting and scaling the dataset:

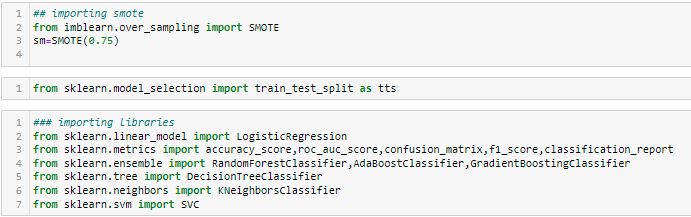


**Checking Imbalance Dataset:** Before building the model we have to check either our data is balanced or not. Imbalanced dataset refer to those type of datasets where the target class has an uneven distribution of observation.

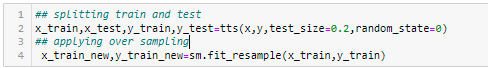


**Observation:-** As we can see from above graphs that 84% Employees not want to leave company and 16.1% employees want to leave the organization. So that we can say that we have biased class, which is the distinctive example of Imbalanced Classification Problem. We will use Over-Sampling or Under-Sampling technique to handle this type of problem.

**Model Building:-** Importing necessary libraries:-

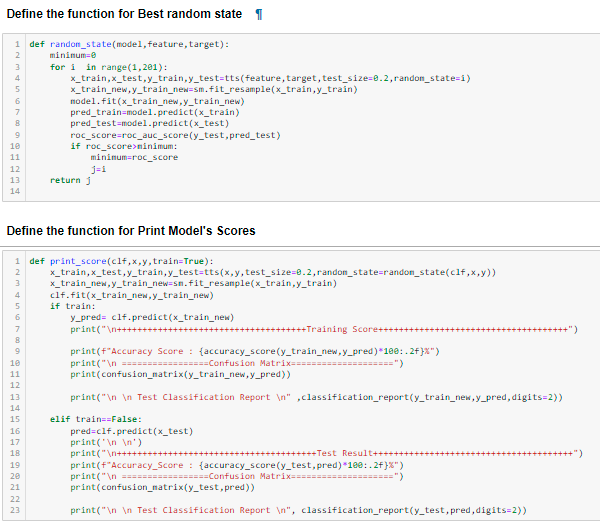


Now we will split our dataset into train and test then apply SMOTE function for handle imbalance data.

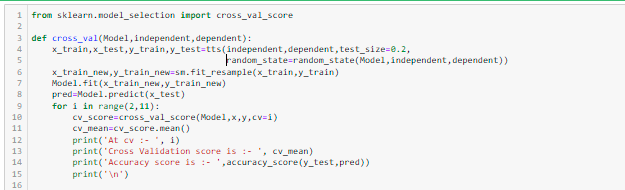


We will give all models one by one and check the accuracy.

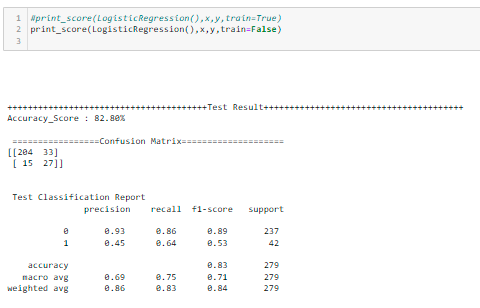
We will define a function for repetitive code:

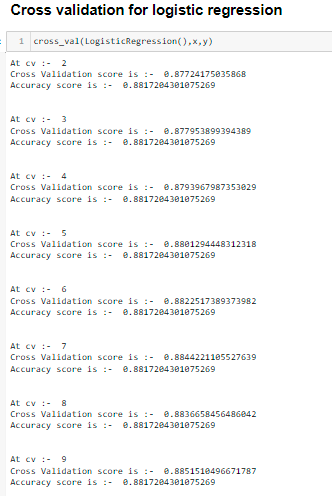


Define a Function for Checking cross validation:

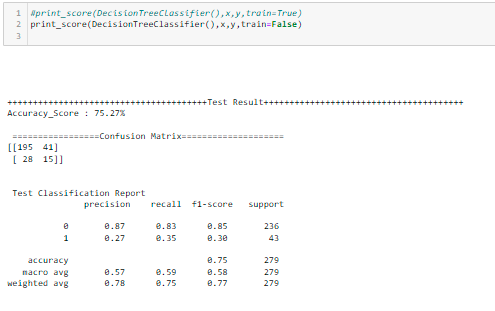


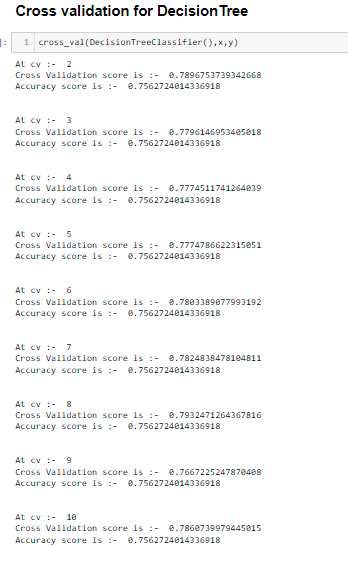
1.Logistic Regression:-



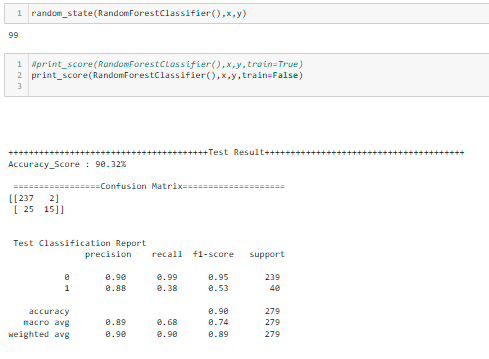


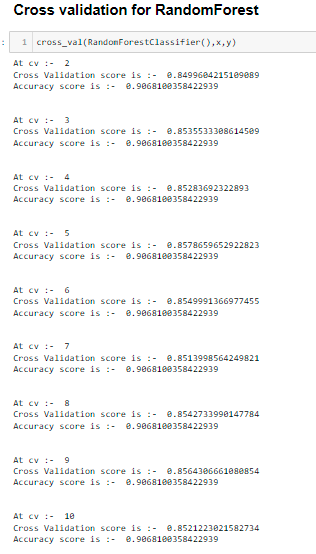
2. DecisionTreeClassifier:-



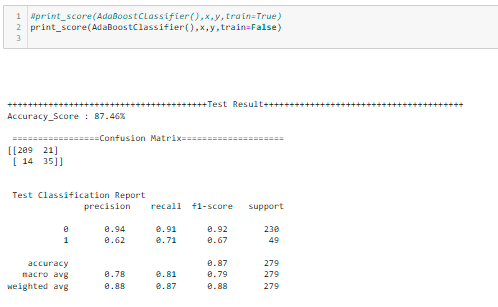


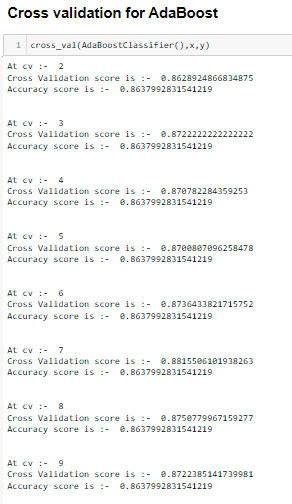
3. RandomForestClassifier:-



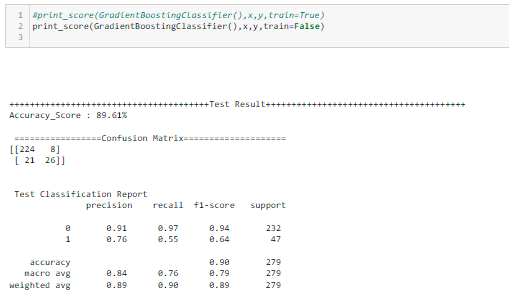


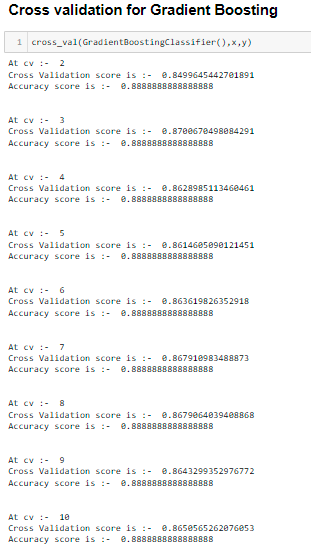
4. AdaBoostClassifier:-



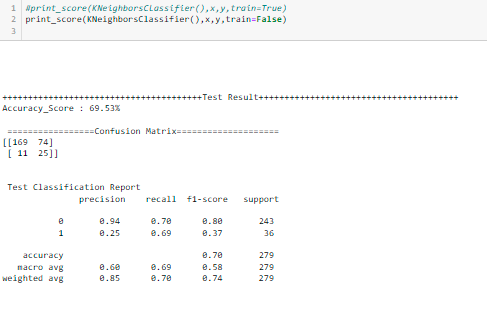


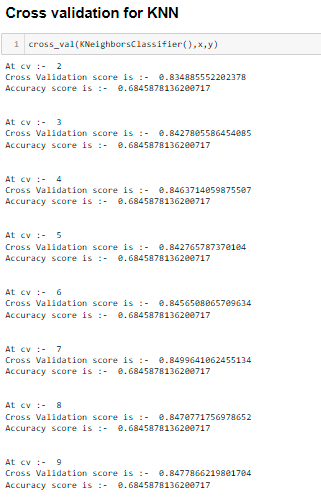
5. GradientBoostingClassifier:-





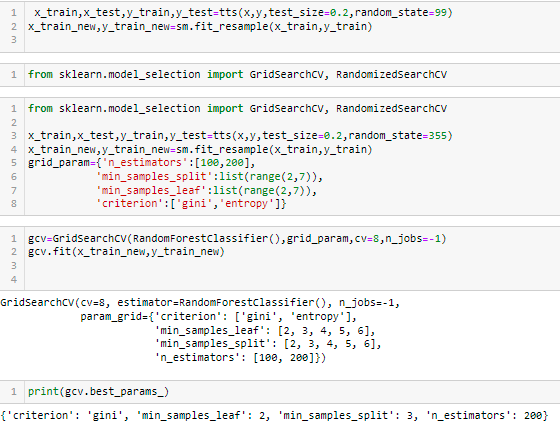
**6. KNN Model:**



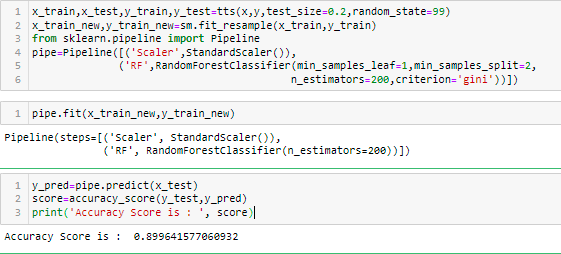


**As we can see that “RandomForest” gives us the Best Accuracy. Now we will tune the parameter for that:-**

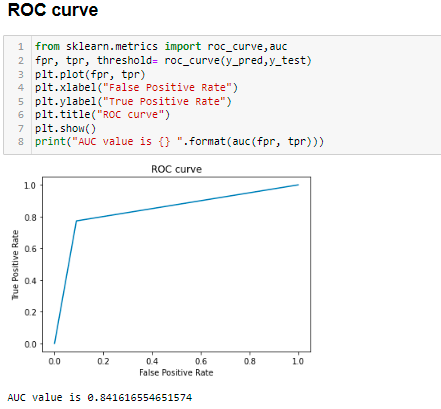
**Hyper parameter Tuning:-**

****

**Creating Pipeline:-**  In real world the final model is built with pipeline. We work on all preprocessing steps , do EDA, make analysis etc. Once we find all the hyper parameter and feature selection techniques etc, we use the main techniques and create pipeline. This will be the clean and better flow of data through series of sequences.



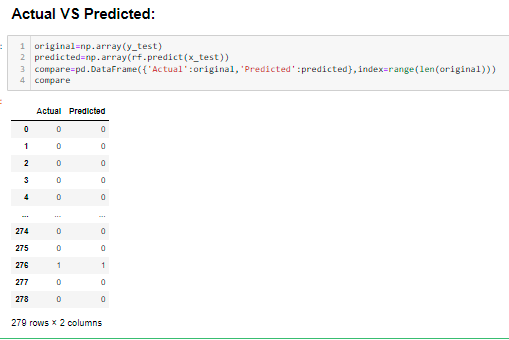
Accuracy score is 90% which is quite good. Let’s Plot ROC &AUC curve.



**Conclusion:-** In this type of problems the main part is data analysis and data cleaning, luckily here we don’t need to clean the dataset, but all the time we may not find clean data that time we have to take care to missing data.

In other hand we must have to manage categorical data and numerical data, as we can see that here we generate many other columns as well, that leads our model to curse of dimensionality.Its resulting exponential increase in computational efforts required for its processing and/or analysis.

The Random Forest shows us the highest 90% accuracy. But we have to take care to False positive and False Negative values. HR analytics, also referred to as people analytics, workforce analytics, or talent analytics, involves gathering together, analyzing, and reporting HR data. It enables your organization to measure the impact of a range of HR metrics on overall business performance and make decisions based on data. Let’s compare the Actual vs Prediction score:



The prediction is quite matching with Actual. We have successfully trained our model to predict Employee data from Sample Data Sets with the goal of constructing and evaluating other Employee Attrition prediction models.