**HR ANALYTICS USING MACHINE LEARNING**

Machine learning (ML) is nothing but an application of artificial intelligence (AI) that provides systems the ability to automatically learn and grow from experience without being explicitly programmed. **Machine learning focuses on the development of computer programs** that can access data and use it to learn for themselves.

Machine learning algorithms have a wide variety of applications, like Traffic prediction, Image recognition, Products Recommendations etc. One such application of machine learning lies in the ‘HR Analytics Market’, to predict the Attrition in human resources. There are various factors/features which impact the Attrition in human resources— Age, Department, Education etc. These factors help create a pattern to decide the Attrition in human resources, and the machine learning models get trained on this pattern to make the predictions in future, automating the process and making the process quicker.

**PROBLEM STATEMENT**

Attrition in Human Resources are very big problem now a days. It refers to the gradual loss of employees’ overtime. 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.

To solve this problem, we use machine learning models, these models, when provided with the data, can gives us best accuracy and predictions, that will be better for the growth of any organization.

**The Dataset**

**Link for the dataset** — https://github.com/dsrscientist/IBM\_HR\_Attrition\_Rate\_Analytics

**We have 1 dataset here:**

This set contains 1470 records and 34 input features. The output ‘Attrition’ column needs to be predicted in this set. We will use Classification techniques here, since the predicted output will be either ‘Yes’ or ‘No’. Our task is to predict the Attrition of the Employee.

**Following is the description of features available in the dataset**–

1. Age – Age of the Employee
2. Business Travel – Employee Travel
3. Daily Rate – Daily rate of the employee
4. Department – Department of Employee
5. Distance From Home – Office distance of employee from their Home
6. Education – Education of employee
7. Education Field – Field of Education
8. Employee Count – count of the Employee
9. Employee Number – Unique ID number of Employee
10. Environment Satisfaction – rating of Employee satisfaction from 1 to 4
11. **Gender** – Gender of Employee
12. **Hourly Rate** – Hourly rate of the Employee
13. **Job Involvement** – Rating of Job Involvement from 1 to 4
14. **Job level** – Level of job (Fresher to experienced)
15. **Job Role** – Roles in job
16. **Job Satisfaction** – Satisfaction in Job
17. **Marital Status** – Employee marital status
18. **Monthly Income** – Monthly income of Employee
19. **Monthly Rate** – Monthly Rate of Employee
20. **NumCompaniesWorked** – Number of companies worked by Employee
21. **Over18** – Age of Employee over 18 or not
22. **OverTime** – Over Time of Employee
23. **PercentSalaryHike** – Salary Hike of Employee in terms of Percentage
24. **PerformanceRating** – Performance rating of the Employee
25. **Relationship Satisfaction** – Ratings of Relationship Satisfaction of Employee from 1 to 4
26. **Standard Hours**- Standard Working hours of Employee
27. **Stock option level** – Level of Stock options for Employee
28. **Total Working Hours** – Total Working Hours of Employee
29. **Training Times Last Years** – Last Year Training time
30. **Work life Balance**- time and focus between working and family
31. **Years At Company** – Number of Years at present Company
32. **Years In Current Role** – Number of Years at Current role
33. **Years Since Last promotion** – Number of Years since last promotion
34. **Years With CurrManager** – Number of years with Current Manager
35. **Attrition** – Attrition of Employee in terms of ‘Yes’ or ‘No’

# Contents of the article

This article explains the complete process to build a machine learning model. Below mentioned are the various phases that we will go through, throughout the project -

1. Exploratory data analysis and Data Modelling

2. Outlier detection and skewness treatment

3. Encoding the data — Label Encoder

4. Scaling the data — Standard scaler

5. Fitting the machine learning models

6. Cross-validation of the selected model

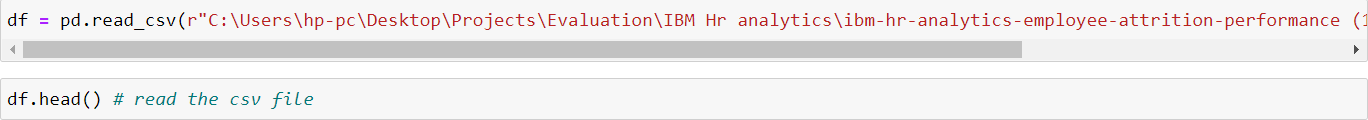
7. Model hyper-tuning

8. Saving the final model and prediction using saved model

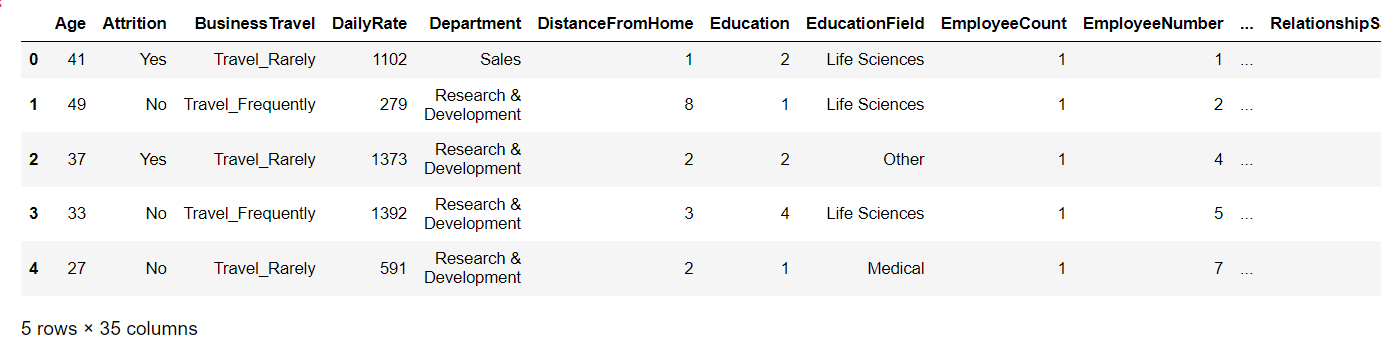
So, let’s begin exploring our data set and start building a prediction model.

# Exploratory Data Analysis and Data Modelling

We load the training dataset using Pandas library –



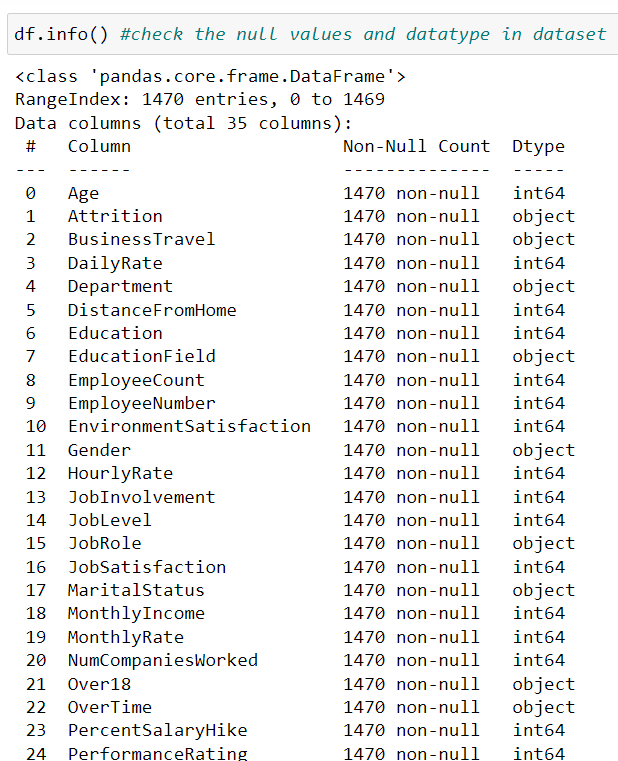
The first step is to have a look at the sample of our data –



We identify the below mentioned points in the first look –

1. Over18 and Employee Count column contains same value.
2. Employee Count, Employee Number are not important for Attrition analysis.
3. The Business Travel, Department, Attrition, Gender, job role, marital status, overtime is in a string format, which we will need to convert to integer type.

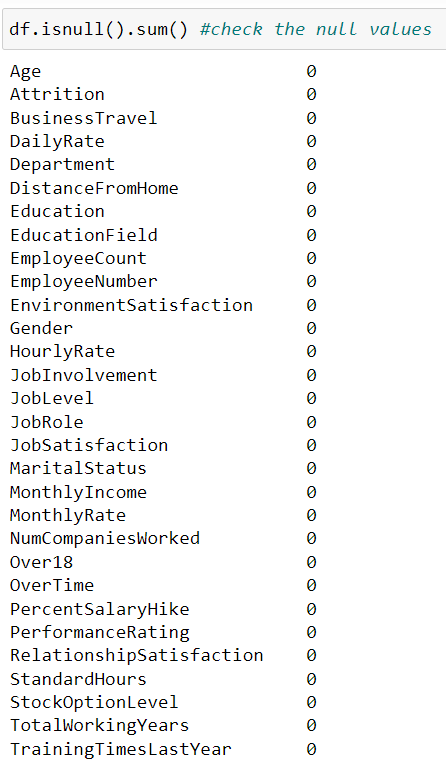
**We further proceed to explore the dataset.**



We run the **df.info()** command, which gives us the information about number of values present in each column, and data types of each column.

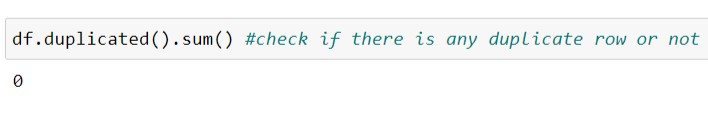
We observe that we have many columns as ‘object’ data types, and ‘Attrition’ column (the output) is also Object type. Since we know what our columns signify, and We can use Label Encoder to treat them.

**We now check the count of NaN (null) values in our dataset, which turns out to give the following result –**



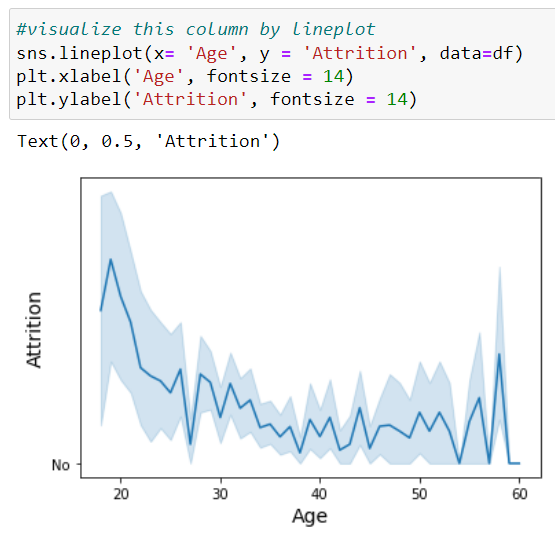
* No null values are found.

**We now check for Duplicate values.**



* No duplicate values are found.

**We now start exploring the important columns available in our dataset one by one.**

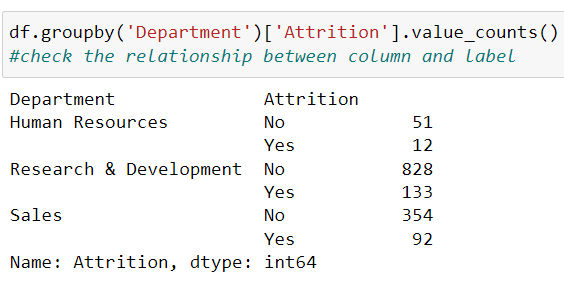


On taking the Line Plot and See the relationship between Age and Attrition, it is noticed that

1. Attrition rate is higher at the age of less than around 25.

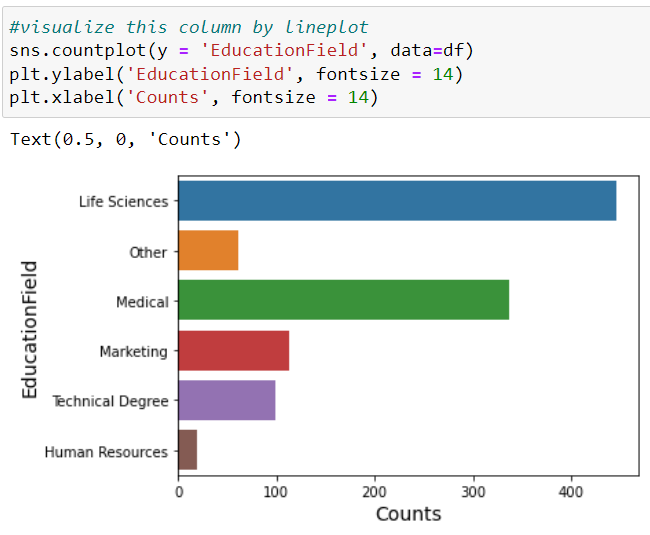
2. As the Age increases, the Attrition is going low.

3. And after the age of 55, Attrition rate is still increasing.

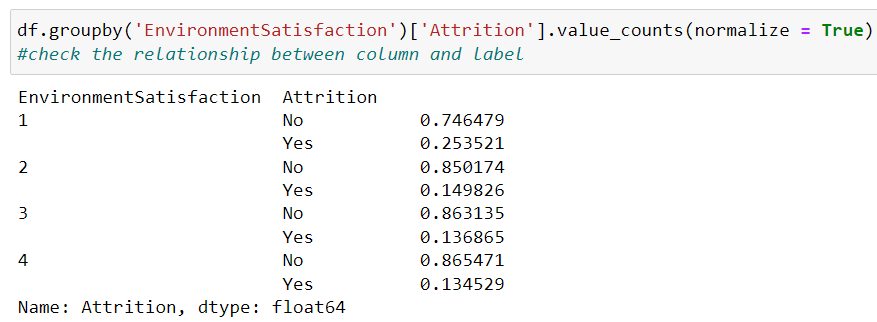
 By using the ‘group by’ method and we can see the relationship between Department and Attrition, and we conclude that

* Attrition is more in sales in terms of percentage

In order to understand the distribution of Education Field, we see this Count Plot, it is noticed that



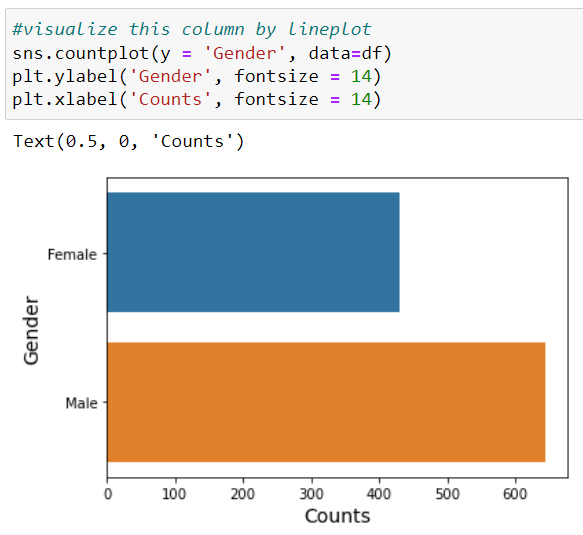
* Life Sciences has a greater number of peoples and Human Resources has least.



By using the ‘group by’ method and we can see the relationship between Environment Satisfaction and Attrition, and we conclude that

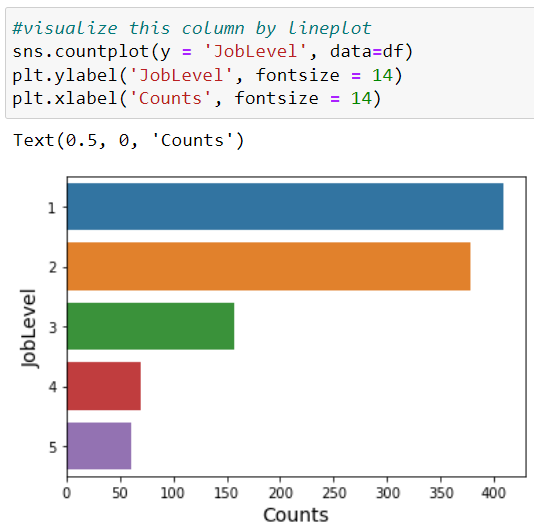
* We can clearly see here; the Attrition is higher where people have less environment satisfaction and it low where people have high environment satisfaction.

Now, we can understand the distribution of Gender by using Count plot.



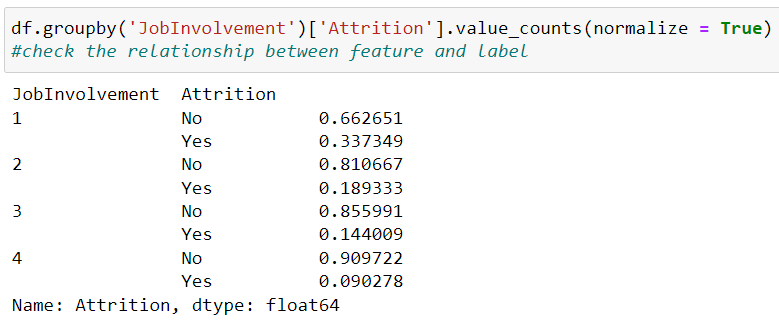
* We can conclude, Males are more in number than Females.

We can understand the distribution of Job involvement by using count plot.

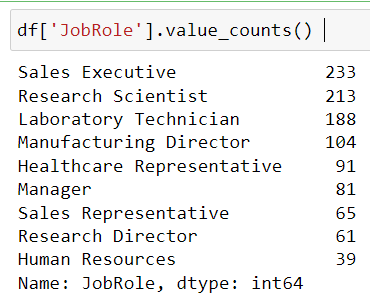


* At job level 1, workforce is higher than others
* And at job level 5, workforce is low

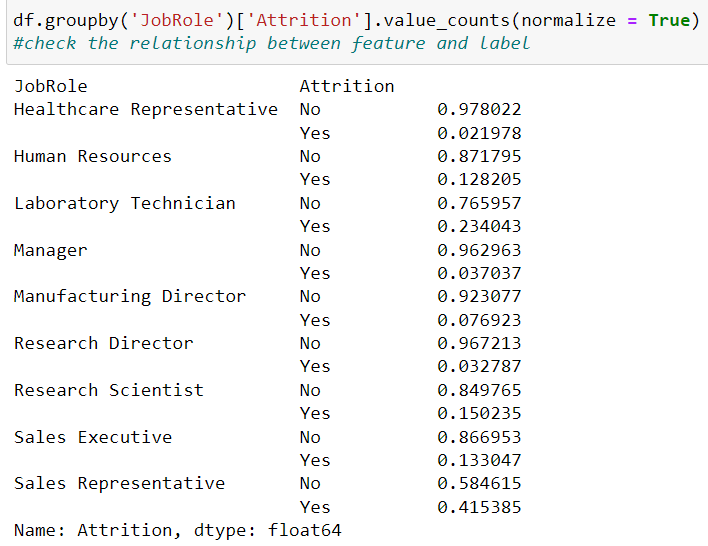
We see the relationship between job involvement and Attrition by using simple group by method



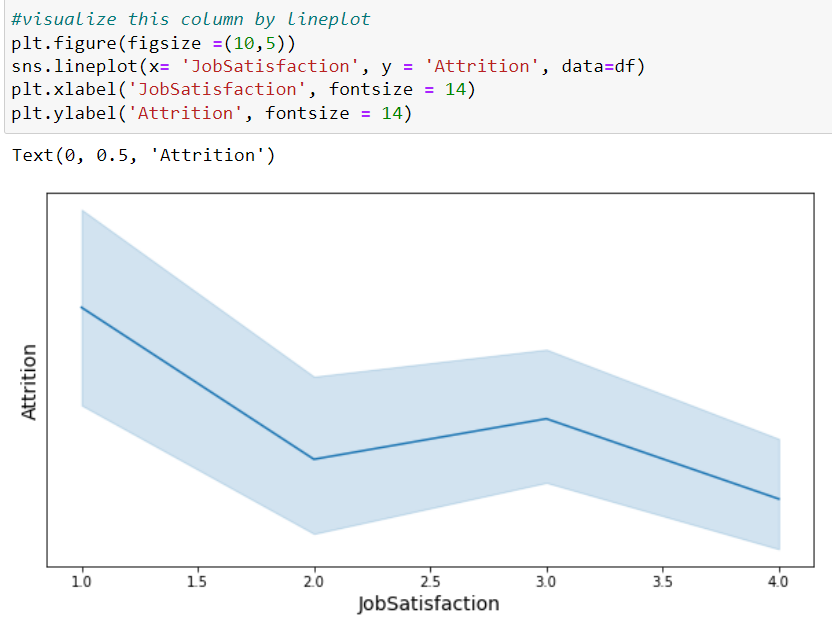
* We can conclude that the attrition of job involvement level 1 is high than others.
* And attrition is low at where job involvement level is high.



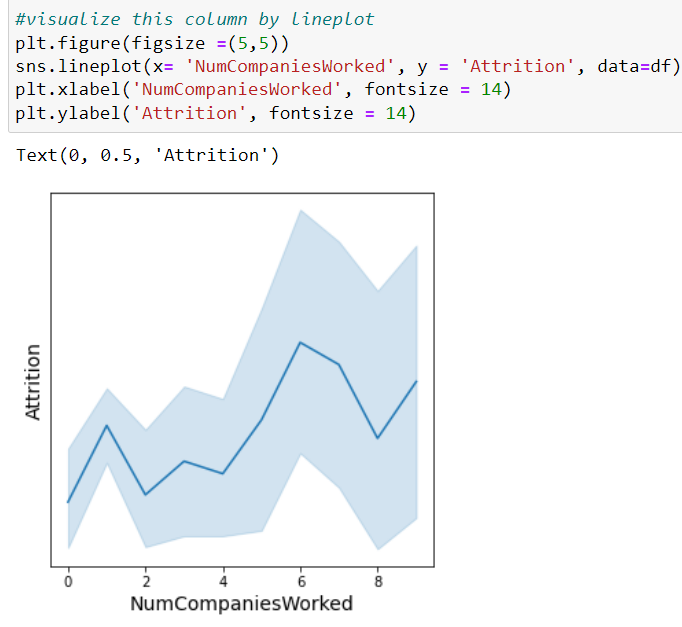
* The workforce is higher for Sales Executive role
* And workforce is less for Human resources



* The attrition is higher in Sales Representative and low in Laboratory Technician.



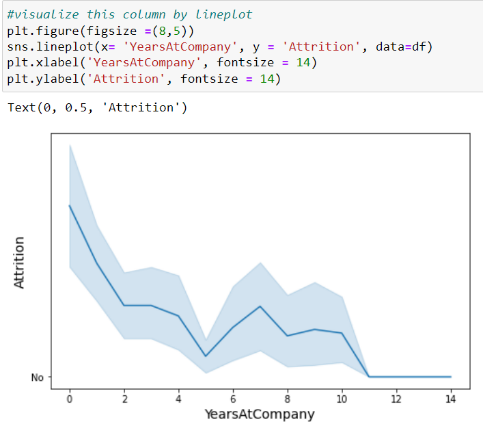
* We can conclude here by visualize line plot, Attrition is higher where job satisfaction is low.



* Attrition rate is higher when people change more companies
* Attrition rate is low when people do not change or change less companies



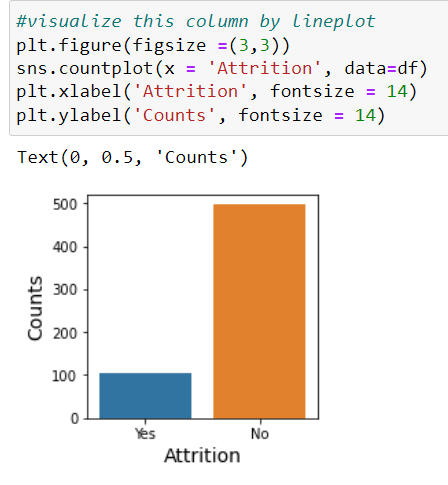
* We can analyse the relationship between Attrition and total working years.
* Attrition is higher where total working years are less than 5 years.
* And Attrition Is low where total working years are higher.



* We can analyse the relationship between Attrition and years at company
* Attrition is higher where people have less years of experience
* Attrition is low where peoples are experienced

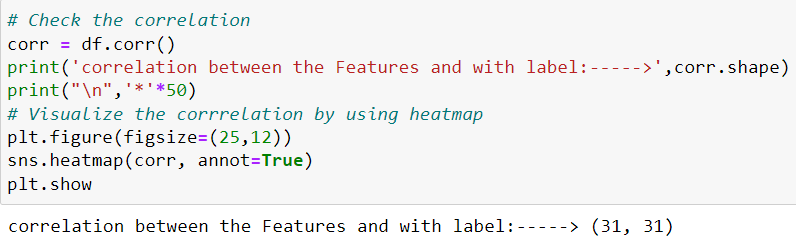
In this way, we can analyse the dataset using various plots, draw meaningful insights from them and provide recommendations based on those insights.

Then, we check the data-type of our target variable: **Attrition**. It turns out that it is a categorical one, consisting of essentially two values (Yes or No), making it a typical **classification problem**.

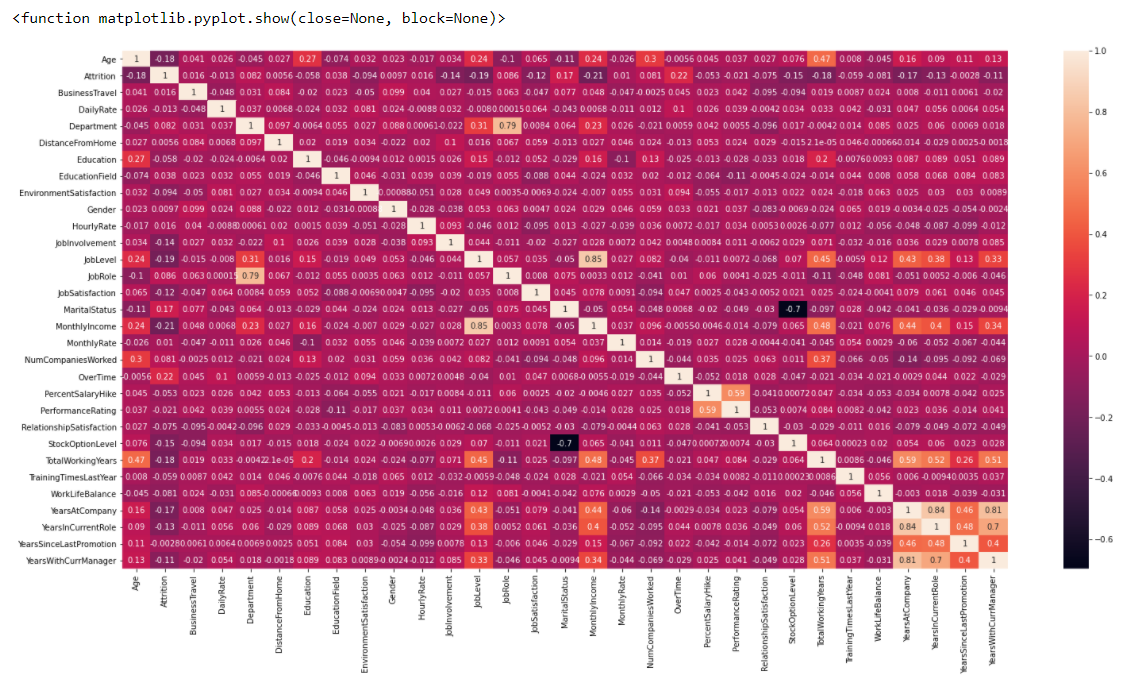


## **Correlation between all Features**

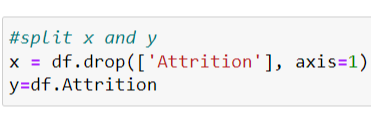
Plotting correlation



**Output:**



* For making our machine learning model and to work on the accuracy, first we need to check the correlation between the features.
* We can check by using heatmap, there is a multicollinearity problem between the features.
* We can also check it by using VIF (Variance Inflation factor) and simply, drop that particular feature and remove multicollinearity problem.

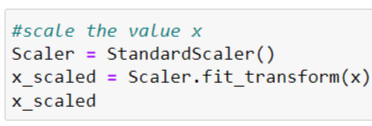
 Now just drop label column from ‘x’ and store it in a variable which is ‘y’

Now, Variable ‘x’ carries all features of our dataset and variable ‘y’ carry only one column that is our target value ‘Attrition’

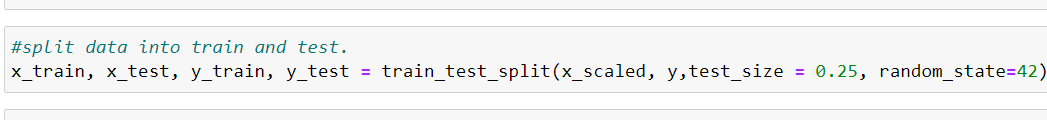
**Data Transformation:**

So, we just need to scale the data which is present inside in variable ‘x’.

We can use other methods also to transform the data like Log transformation etc.

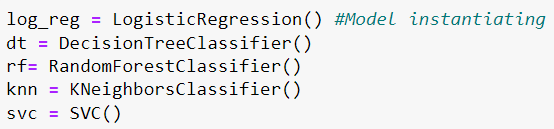


We then split the dataset into two sets: *training set*and*testing set.*As the name implies, the training set will be used when training the model and the testing to test its predictive capability.

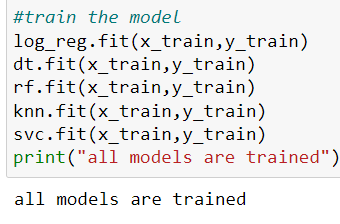


**Build the Model:**

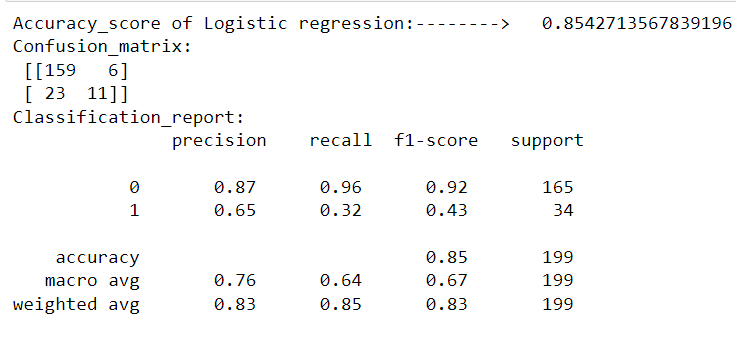
**Here, we check the accuracy by using 5 classification model.**



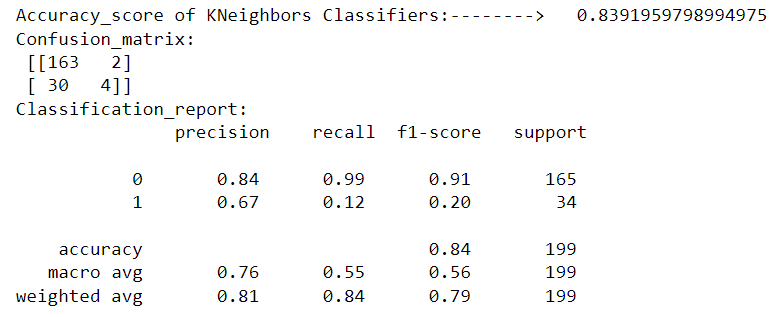
* Check accuracy on 5 models which is logistic regression, Decision tree classifier, Random Forest classifier, kNN and support vector machine



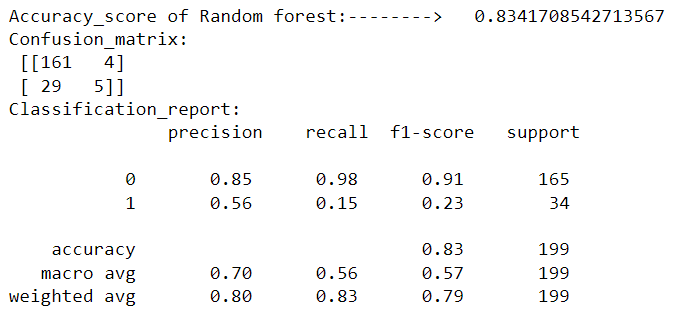
* Train the model, then test the model and check the accuracy score by using these models.



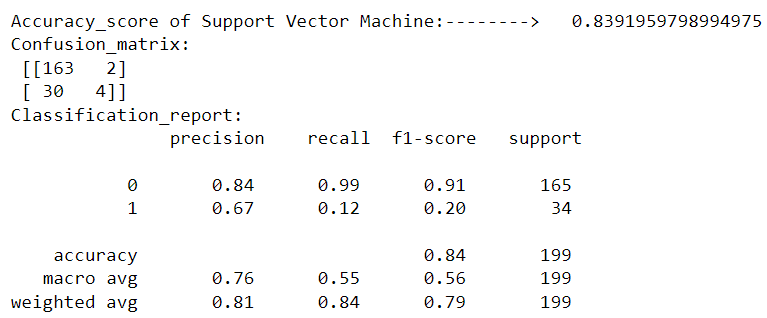
We can see, Logistic Regression is giving us, 85% Accuracy score



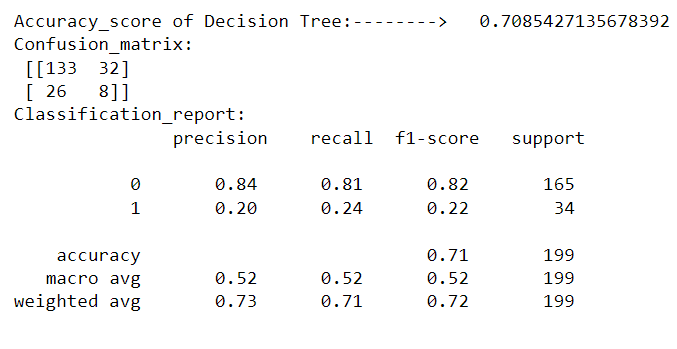
We can see, KNN Classifier is giving us, 84% Accuracy score



We can see, Random Forest is giving us, 83% Accuracy score

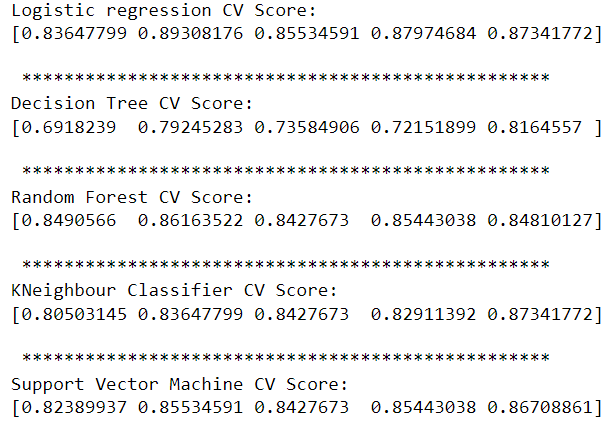


We can see, SVM Classifier is giving us, 84% Accuracy score

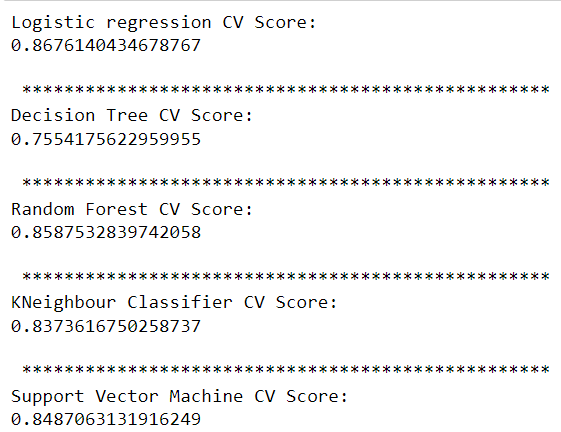


We can see, Decision Tree Classifier is giving us, 71% Accuracy score

Now, we can use Cross Validation Technique to compare and and select an appropriate model for the specific predictive modelling problem. When we select cv=5, then this technique will give us 5 accuracy score.



By calculating the mean value of these cv score, we can select appropriate model for predictive modelling problems



Now, we can easily conclude, Logistic regression is giving us high accuracy as compare to other models but we can improve our accuracy by using Hyperparameter Tuning.

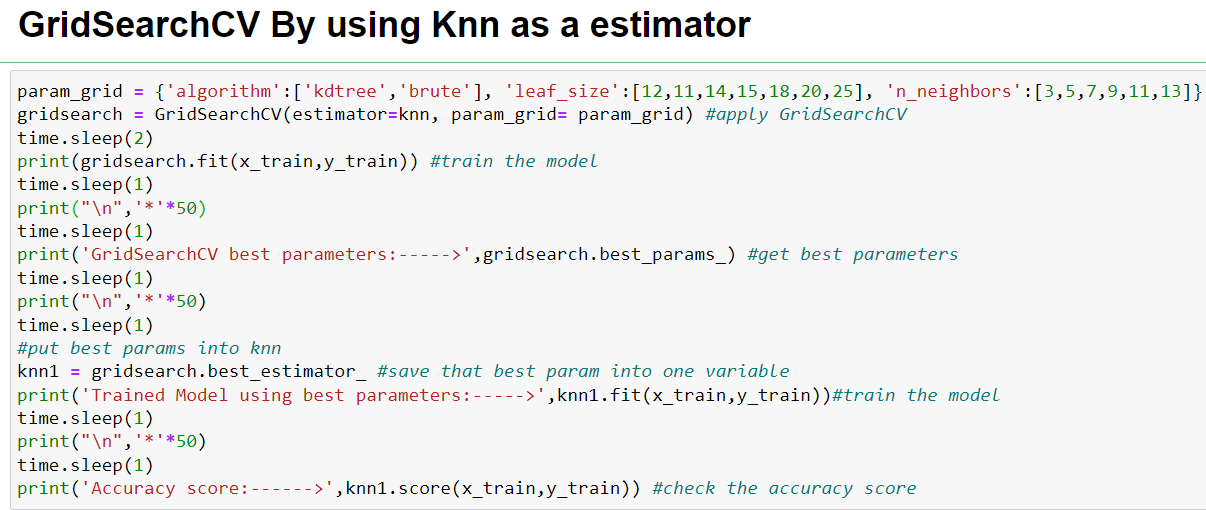
**Hyperparameter Tuning**

* This technique generally used to improve the accuracy score.
* There are of two types:

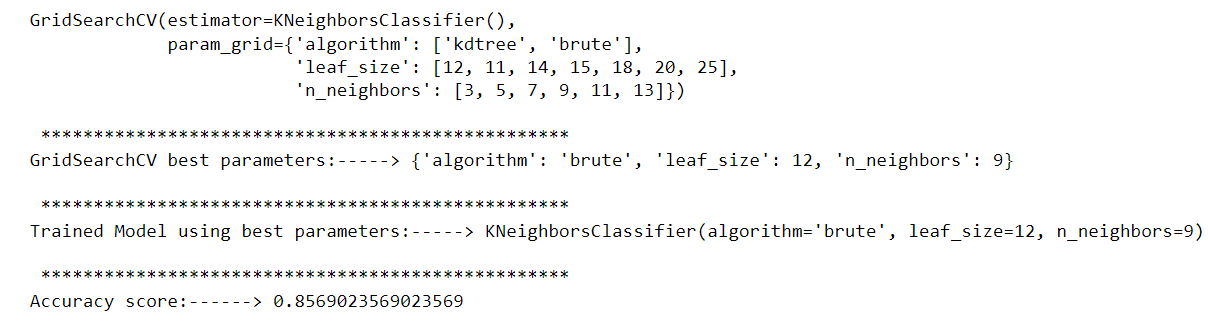
1. Grid search CV
2. Random search CV

* We use Grid search CV on every model one by one, and check the accuracy score is improved or not.

**Input:**

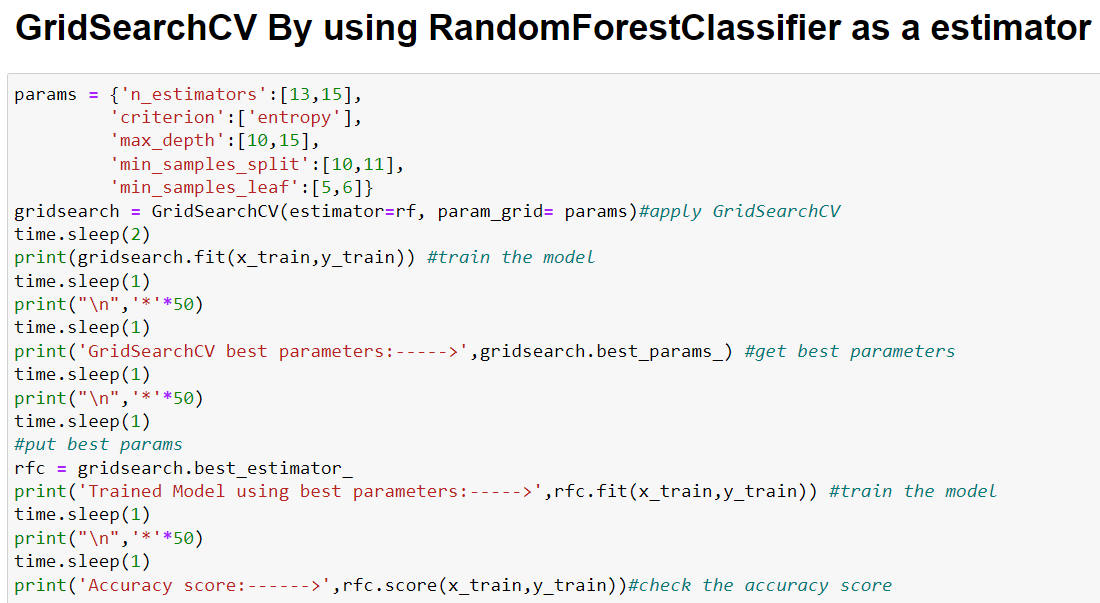


**Output:**

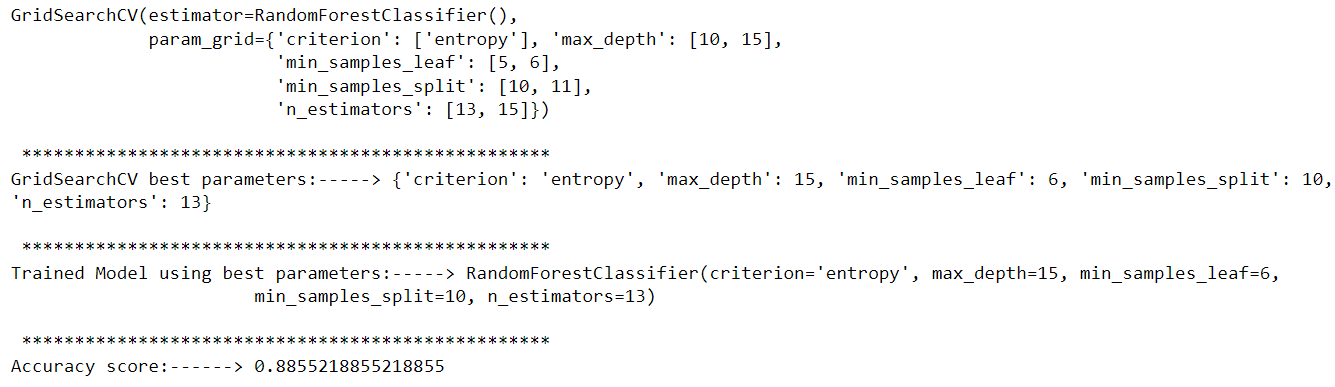


Earlier Accuracy score of KNN Classifier is 83% but After using Grid Search CV technique, it improves to 85.6%

**Input:**

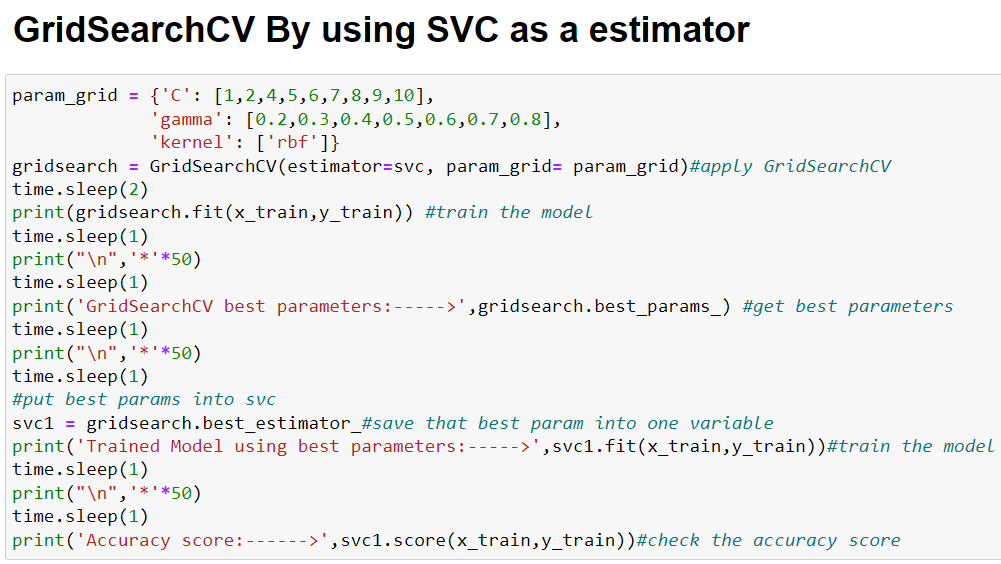


**Output:**

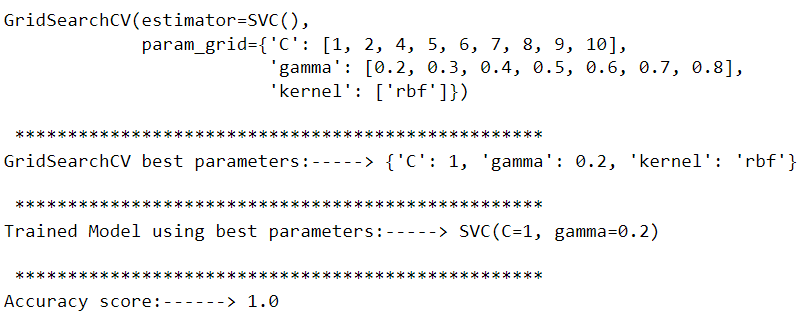


Earlier Accuracy score of Random Forest Classifier is 85% but After using Grid Search CV technique, it improves to 88%

**Input:**

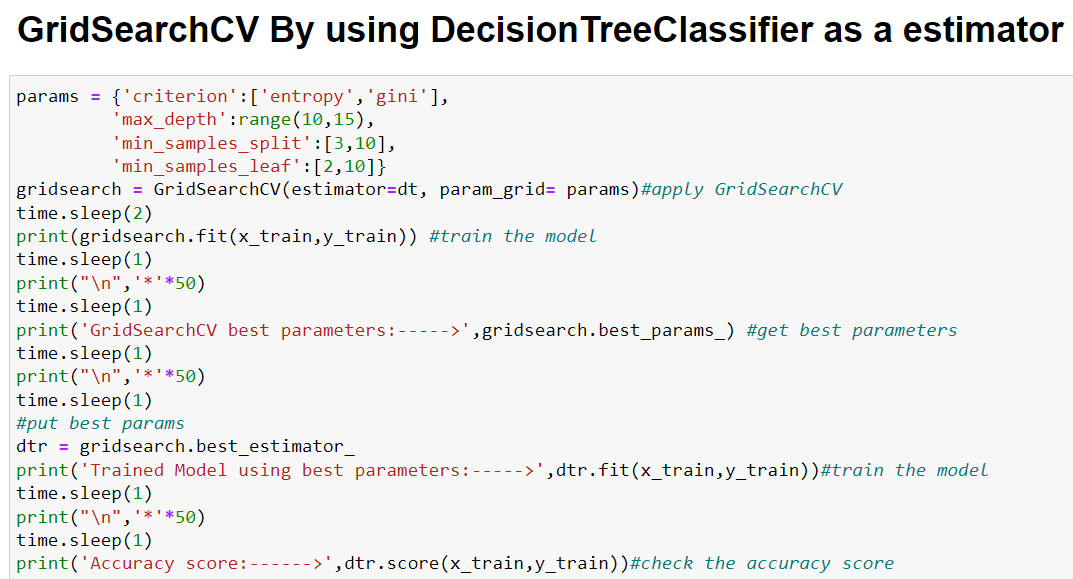


**Output:**

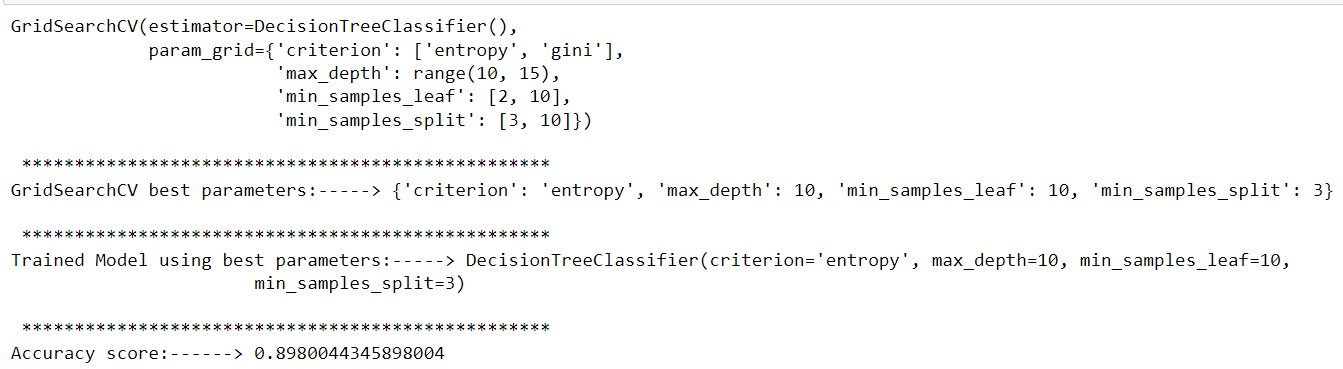


Earlier Accuracy score of SVC Classifier is 84.8% but After using Grid Search CV technique, it improves to 100%

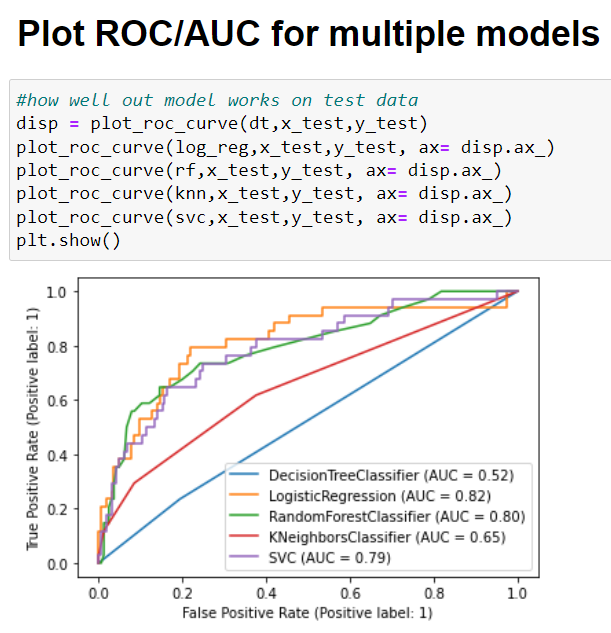
**Input:**



**Output:**



Earlier Accuracy score of Decision Tree Classifier is 75.5% but After using Grid Search CV technique, it improves to 89.8%

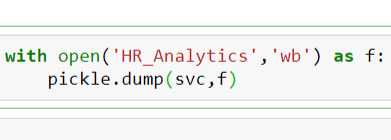


We can also plot AUC – ROC curve for performance measurement for classification problems at various thresholds settings.

* ROC is a probability curve
* AUC represents degree or measure of separability.

**Important points:**

* We can also use bagging Classifier with Bootstrap = True and False to check the accuracy score on these models.
* We can also use PCA Technique to further check the Accuracy Score
* So, Support vector machine when we used it as a estimator it is giving us 100% accuracy, that should be our best model.
* We can save the best model out of it and predict it easily.



**Conclusion:**

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. Attrition in HR is creating problems to the company. This is especially concerning if our business is customer-facing, as customers often prefer to interact with familiar people. Errors and issues are more likely if we constantly have new workers.

**We conclude that by representing some important observations. As we see,**

* As the Age is growing, the Attrition rate is becoming low.
* Attrition is more in sales department.
* Those workforces have less education, their Attrition rate is higher.
* Attrition in Technical degree holder peoples are more than others
* Those peoples have less Environment Satisfaction, their Attrition rates are high
* Peoples who have less job Involvement, their Attrition rates are high
* Peoples who have less job level or experience, their Attrition rates are high
* Attrition in sales representative is high
* Attrition is high who have less job satisfaction
* Attrition is high who have less Monthly income
* Attrition rate is higher when people change more companies
* people who have high salary hike their attrition rate is high
* Attrition rate is high who performance is less
* those who have less relationship satisfaction their attrition rate is high
* those who have less experience, their attrition rate is high
* those who have zero training years their attrition rate is higher
* those who have poor work life balance their attrition rate is high
* those who have less year at company their attrition rate is higher
* those people who have less years at current role, their attrition rate is higher

Now we conclude that, Peoples who are inexperienced or having less years of experience, generally in Sales department, having technical degrees, their Attrition rates are high than others because their Job satisfaction, Environment satisfaction, relationship satisfaction, job level, monthly salary, training years are less than other workforce. These peoples are generally very young in their age, so, we can say that as the maturity grows peoples choose every step carefully and wisely.

Monthly Salary and job satisfaction also plays a major role but we cannot avoid other factors also like environment satisfaction, Trainings etc. Companies or any organizations should provide proper training, salary hikes on time, motivational seminars or webinars, learning sessions, and other facilities also like sports, medical etc. to the employees, so that employees get motivated all the time with their job.