**HR Analytics Project- Understanding the Attrition in HR**

**Introduction:**

Attrition, from an overall perspective, is a progressive decrease or diminishing of a thing. From a business perspective, there are two different ways to characterize Attrition. Attrition for a business can be portrayed as either employee attrition or client attrition – both vital to comprehend as a business owner. Employee attrition is utilized to portray the decrease of employees. Employee attrition can occur for a large number of reasons. The reasons may include employees resigning, securing other position openings, or leaving because of misery.

Nonetheless, it is vital to note that for it to be characterized as employee attrition and not simply a piece of employee turnover, business owners or administrators should choose not to top off the particular position that is presently unfilled. Inside employee attrition, there is either voluntary or involuntary employee attrition. The distinctive factor for the two sorts of Attrition versus employee turnover is that with Attrition, the positions are not immediately topped off or topped off by any means. Voluntary employee attrition models incorporate employees leaving to seek after other open positions or resigning. Then again, involuntary employee attrition incorporates work position disposal because of business cutting back.

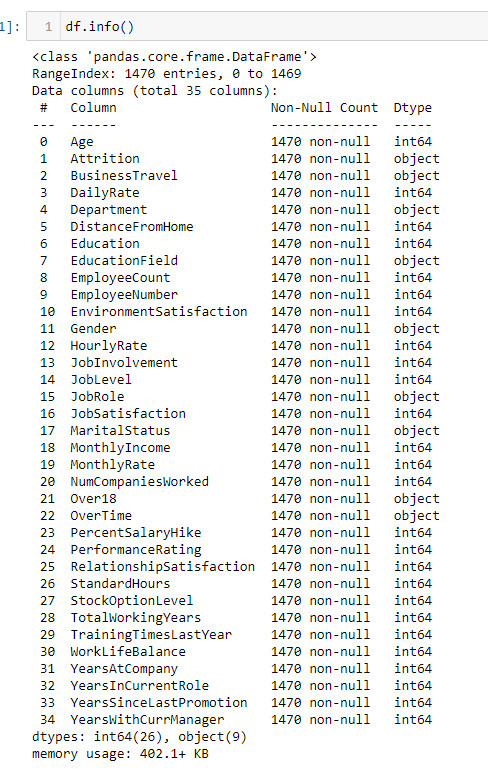
In this paper, we will make a model based on the information provided to predict the Attrition of an employee. Also, we will try to answer How does Attrition affects companies? Moreover, how does HR Analytics help in analysing Attrition? We will discuss the first question, and for the second question, we will write the code and try to understand the process step by step.

**Dataset:**

The dataset used in this project has been provided by Data trained for the assessment process. The dataset contains 1470 rows and 35 columns. In this dataset, our dependent, or we can say target variable is the Attrition column.

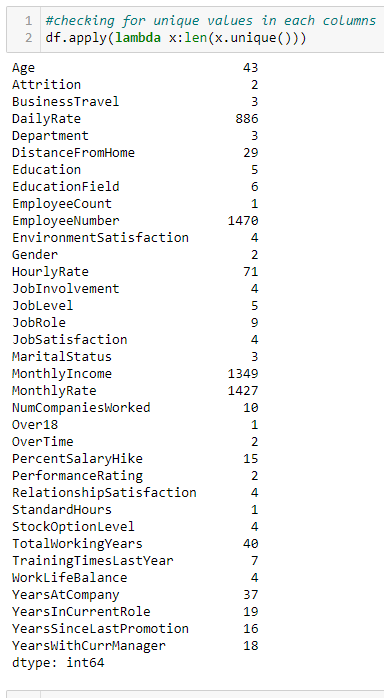
**EDA:**

It is idle to explore a data set with the various exploratory method, especially when they can be done together for comparison. Every data scientist should compile a cookbook of techniques in exploratory data analysis. Once we fully understand our data set, we may need to revisit one or more data munging tasks to refine or transform the data. The purpose of EDA is to obtain confidence in our data to a point where we will be ready to engage a machine learning algorithm. The very first thing to check in the data given is the type of the given variable.



As shown in the image shown above, there are nine object type variables and 26 integer type variables. Looking at the Non-null Count, it is clear there is no null value in the dataset. As the total number of rows, this dataset contains 1470, and every variable has a 1470 non-null count.

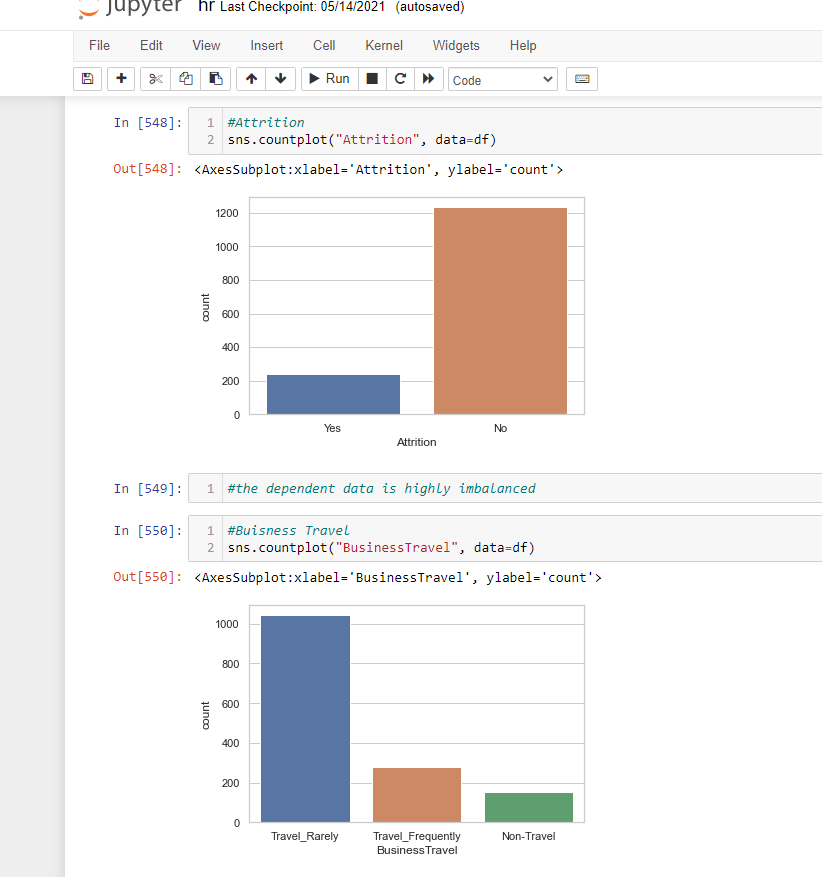
Further, we will try to examine the unique value of each variable.



As the Attrition columns contain only two unique values, we already know it is a binary classification problem. However, some columns only contain one unique value; these columns are Standard Hours, Over18, and employee count. These columns will add no insight; it will also not need the model building; we will drop these values.

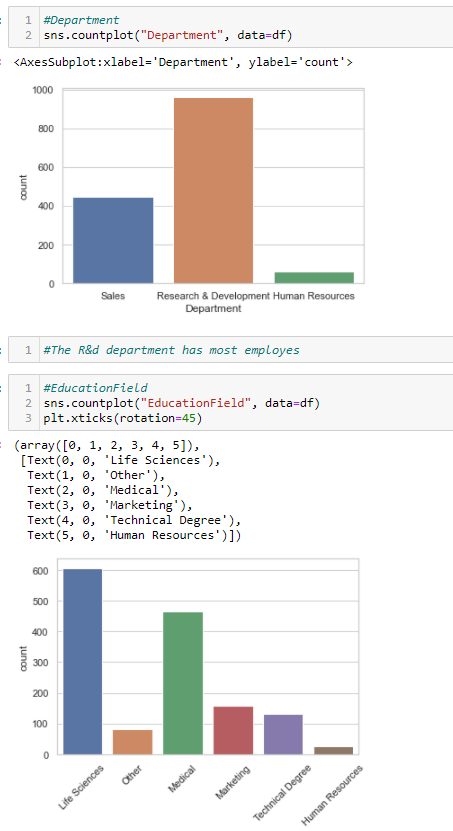
* **Univariate Analysis**

Attrition:



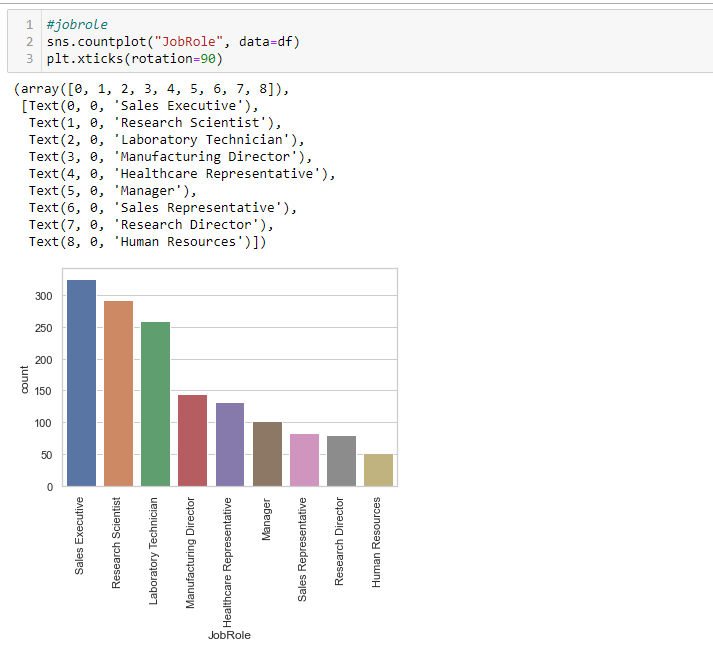
* Attrition is our dependent variable. Also, from the image, we can see that the data is highly imbalanced. We need to do data balancing; otherwise, we will not make an effective model.

Department:



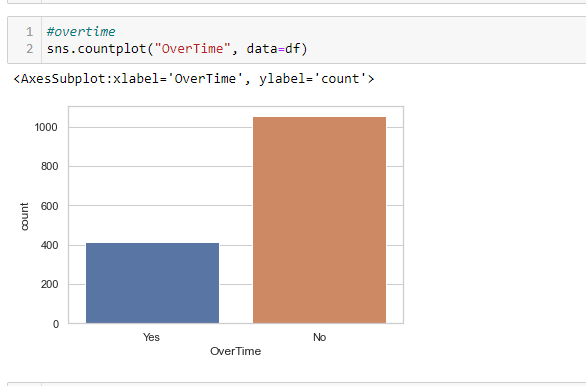
* Majority of the employee works in Research & Development Department.

Job role:

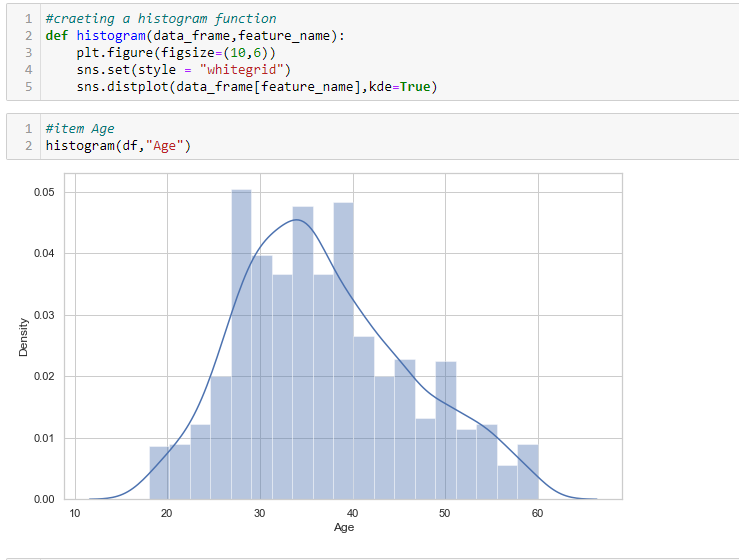


* Though Majority of the employee works in Research & Development department, the Majority of employees have job role of Sales Executive.

Over Time:

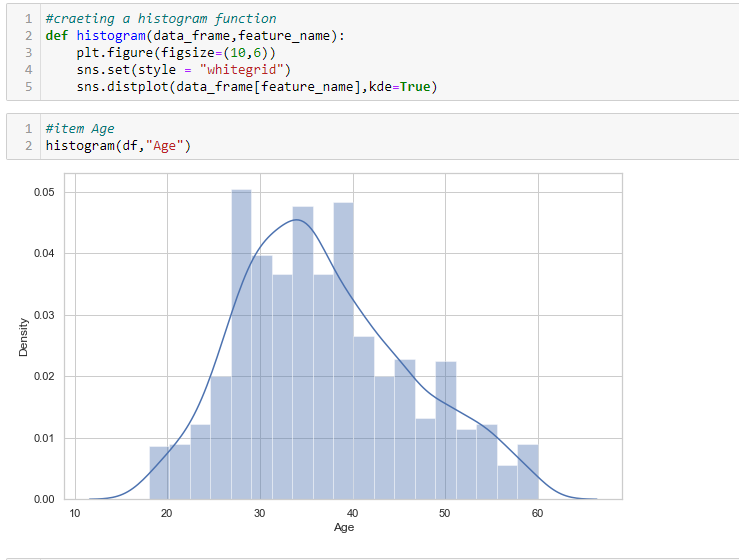


* Though most employees are not doing overtime, however, there is a reasonable number of overtime employees.
* To analyse the integer variable, we are making a histogram function with the help of a distplot.



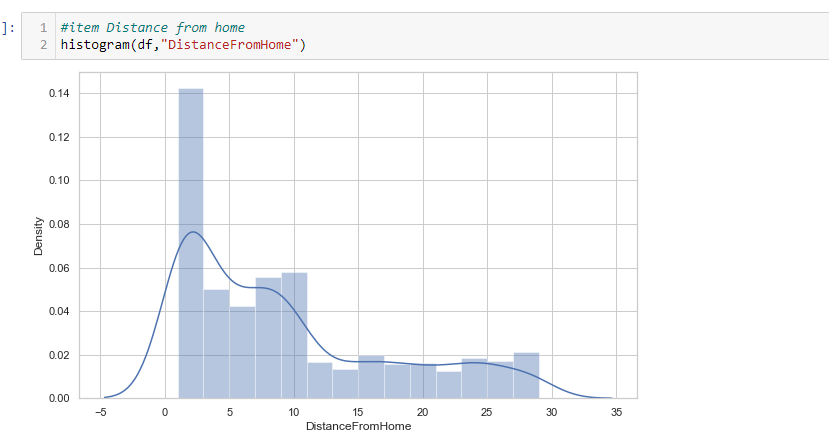
Now with the help of this function, we will analyze the Integer variables.

Age:



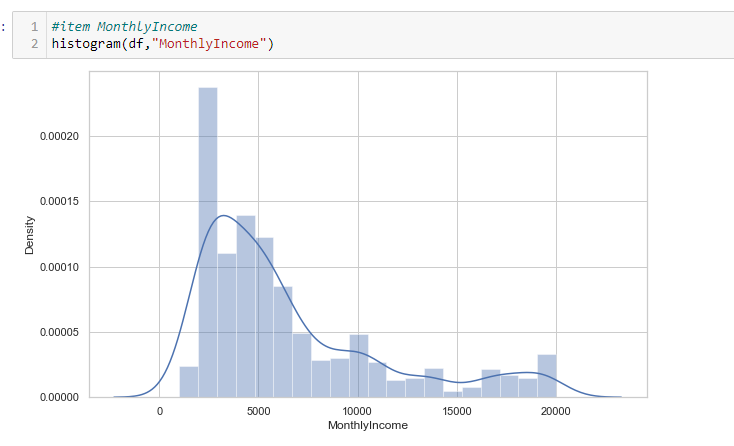
* Majority of the employee is in the age group of 26to 40 years. The youngest employee is above 18, and the oldest employee is 60 years old.

Distance from home:



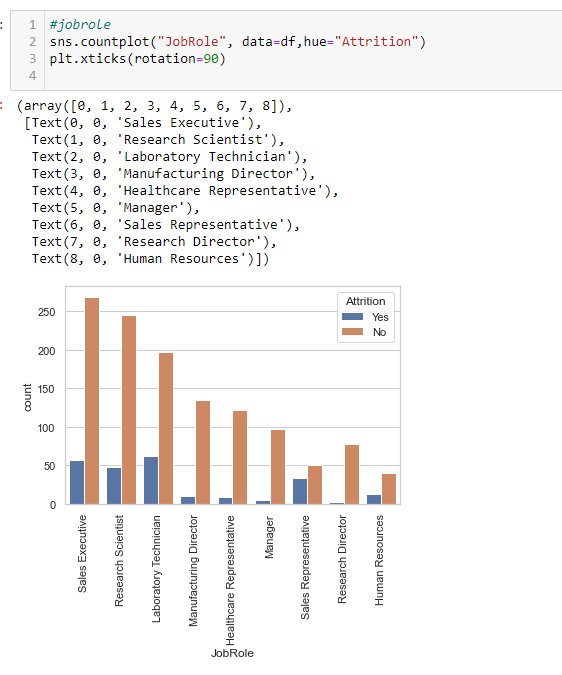
* The Majority of the employees live near the office.

Monthly Income:



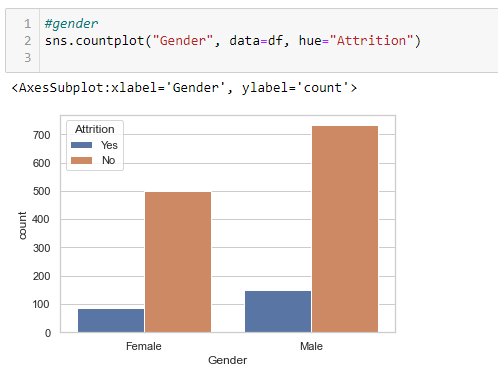
* The Majority of the employees have a salary in the range of 5k per month, while few employees earn nearly 20k per month.
* **Bivariate Analysis**:

Job role/Attrition



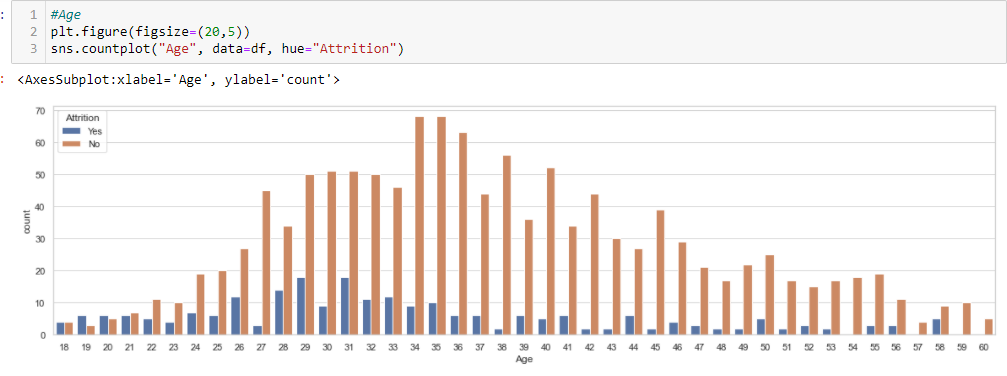
* Employees who are Sales Executive, Research Scientist, and Lab rotary Technician are more prone towards Attrition than other job role employs

Gender/Attrition



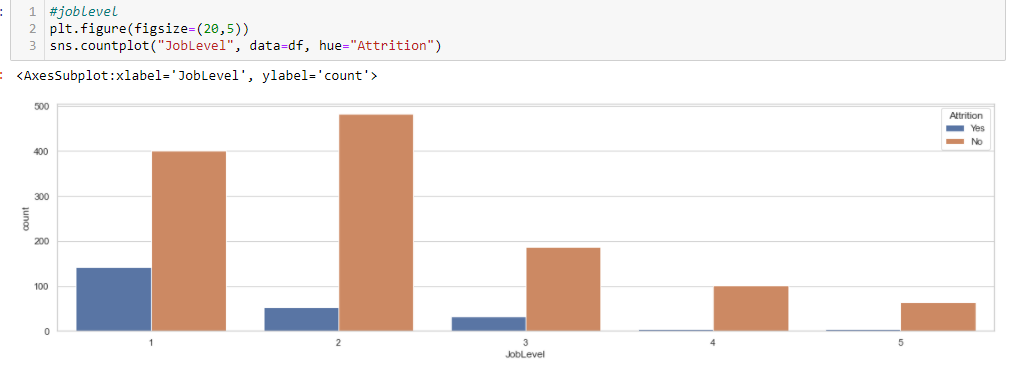
* As indicated in the image, Male employees are more inclined towards Attrition than female.

Age/Attrition



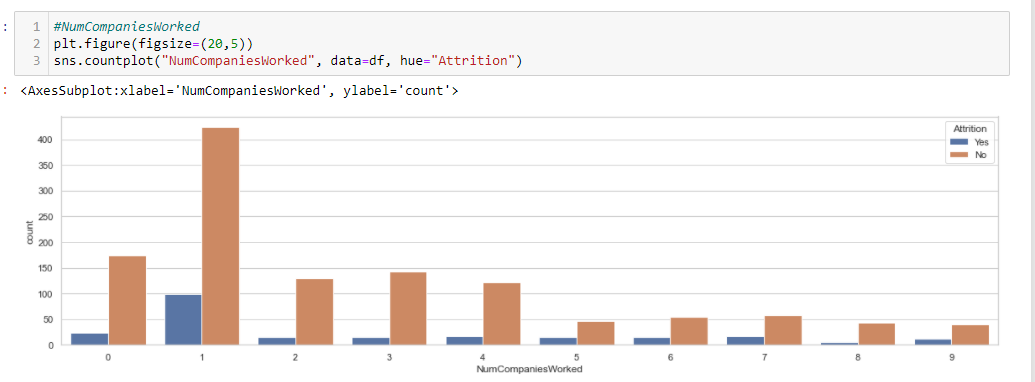
* Employees of age group 25 to 30 are more inclined towards Attrition in comparison to other age groups.

Job level/Attrition



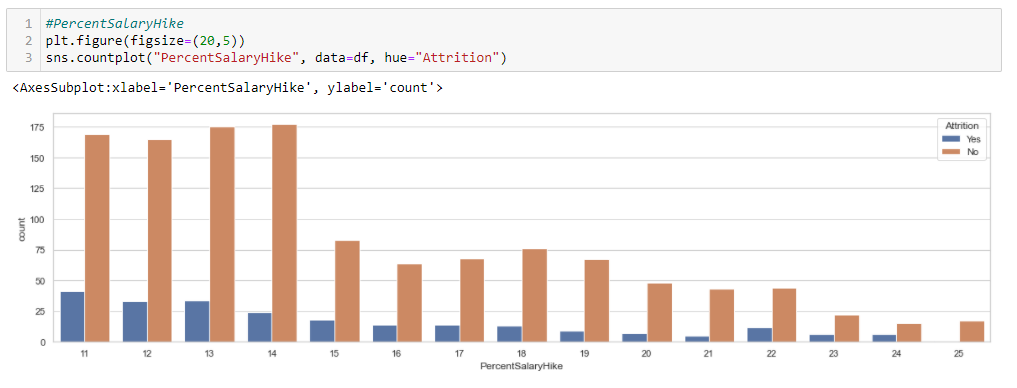
* As the image indicates, employees on a lower job level are more prone to Attrition, while employees on the higher post are significantly less inclined towards Attrition.

Number of companies worked/Attrition



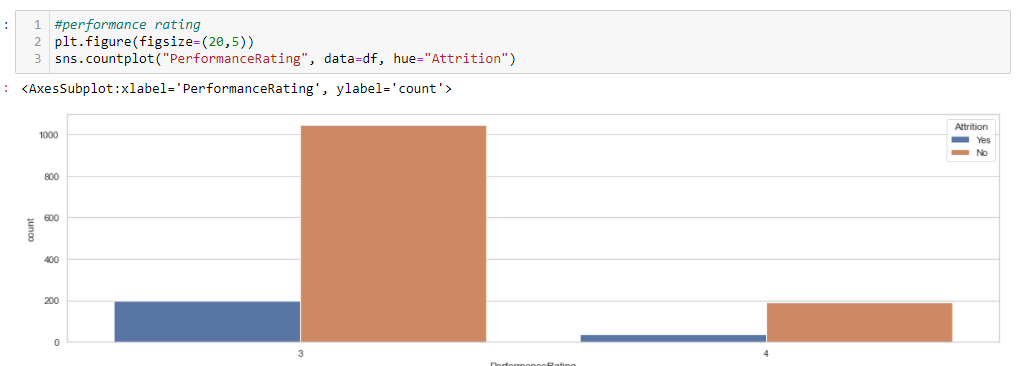
* The employee who has worked with fewer companies is more inclined towards Attrition, while the employee who has worked with many companies is less prone to Attrition.

Salary Hike/Attrition



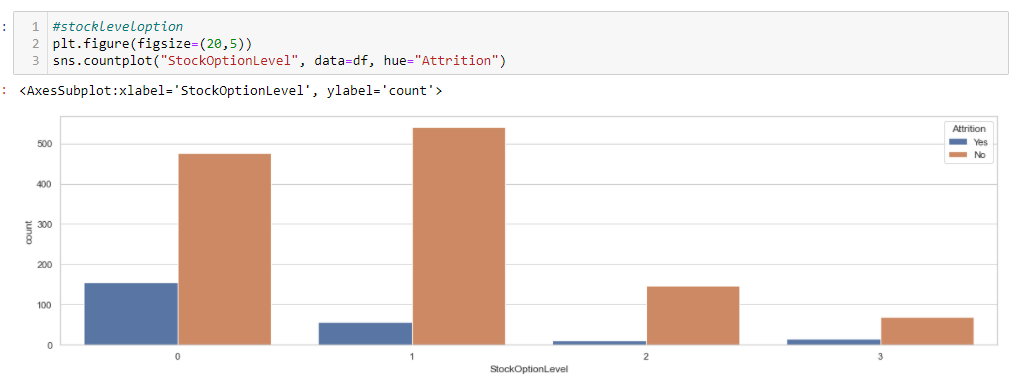
* Salary hike plays a vital role in the attrition process. As shown in the image, with the increase in salary hikes, employees are becoming less prone to Attrition. In contrast, employees with fewer salary hikes are more inclined towards Attrition.

Performance rating/Attrition



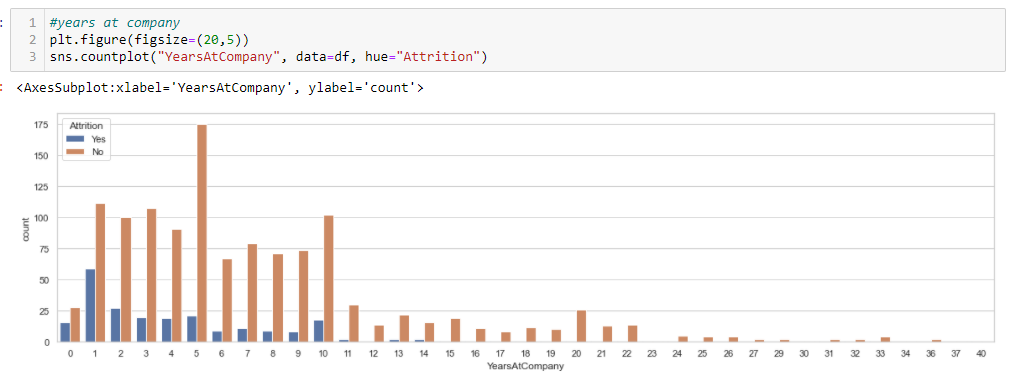
* Employees with less performance rating are more inclined towards Attrition.

Stock Options Level/Attrition



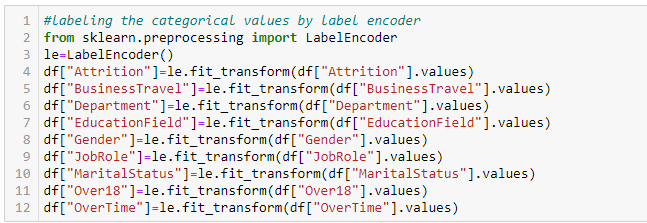
* Employees who have less access or no access to stock options are more prone to Attrition. At the same time, the possibility of Attrition is significantly less with employees who have more access to stock options.

Years at company/Attrition



* Employees working with the company for a very long time are significantly less inclined towards Attrition. In contrast, employees who have been working for the last one or two are more inclined towards Attrition.

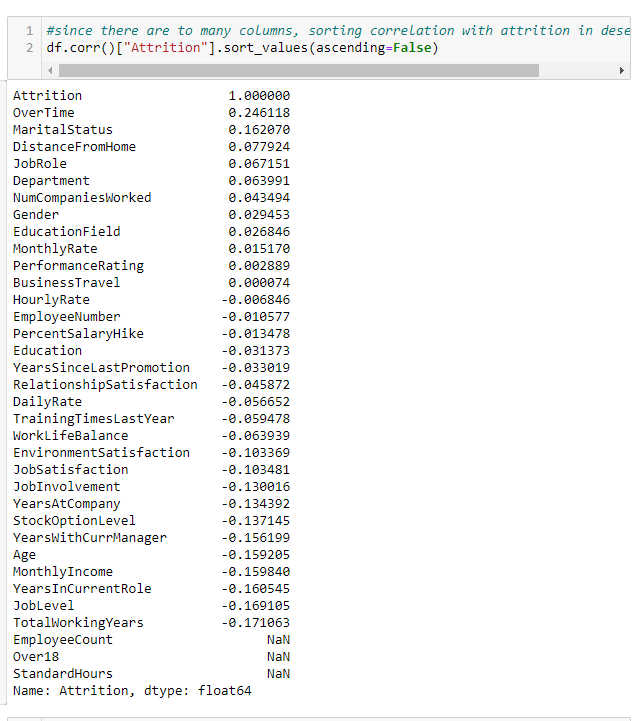
As we completed the analysis part, it is time to go for feature selection. We will use the correlation method to get the importance of the feature concerning the dependent variable. So, before proceeding, we need to encode all of the categorical data.



Here we have used Label Encoder to transform the value of the given variables.

**Correlation/Feature selection:**

As there are many columns so heatmap will not be a good idea to check the correlation. So, we will use a different approach.

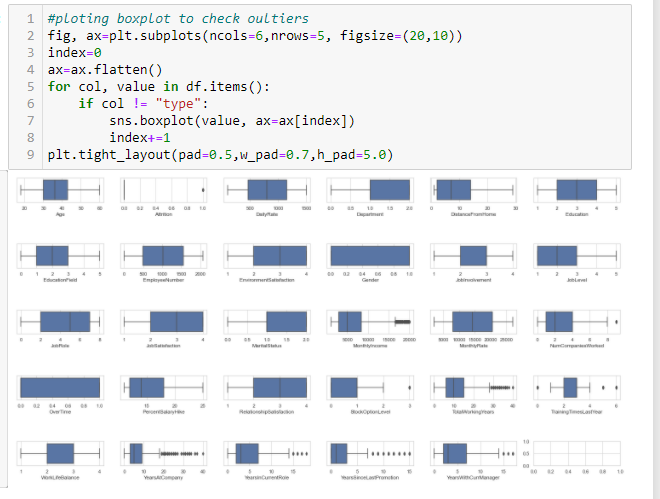


As we can see in the image, over time and marital status are the most positively correlated variable with Attrition, while entire working years and job level are the most negatively correlated variables. Also, we can see in Performance rating, Business travel and Hourly rate are close to zero and significantly less correlated to Attrition we will drop these columns. Further, we can see that Over 18, Employee count, and Standards hours show nan because this variable contained only a single unique value which resulted in no correlation with Attrition, so that we will drop these variables.

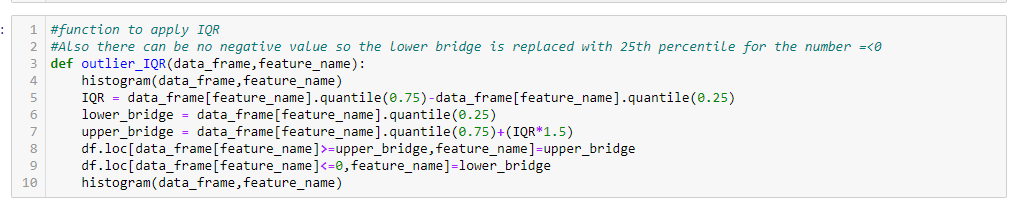
df.drop(["EmployeeCount","Over18","StandardHours","PerformanceRating","BusinessTravel","HourlyRate"],axis=1,inplace=True)

**Checking Outliers**:

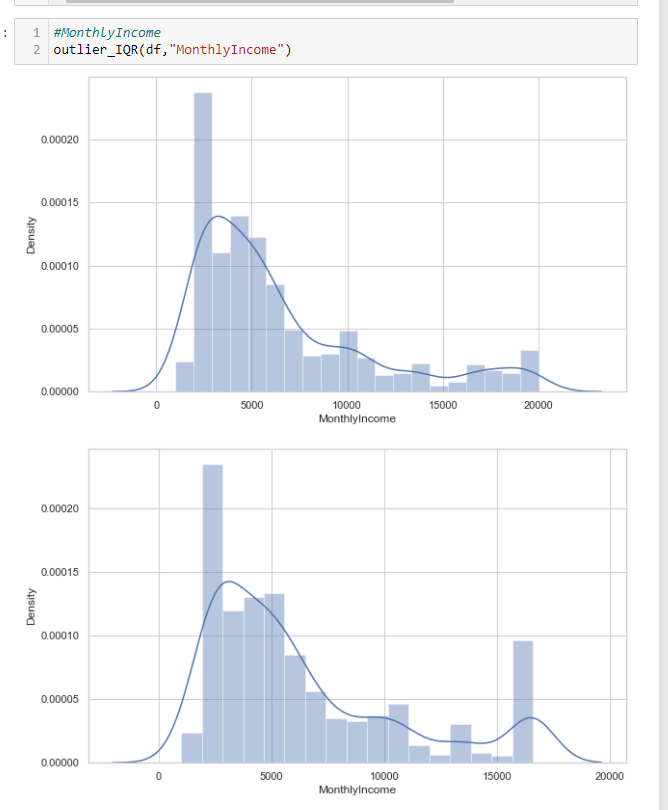
The next step is to see whether our data contains outliers or not.



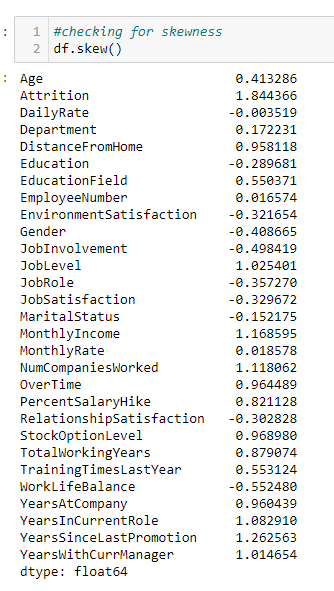
There are outliers in some of the columns, so we will use the IQR method to replace the outliers; we are not using Z-score because we already have very little data, and Z-score will result in data loss. We are creating a function name outlier IQR to fix the outliers.



We can apply this function to all variables individually, which contains outliers in the image shown above.

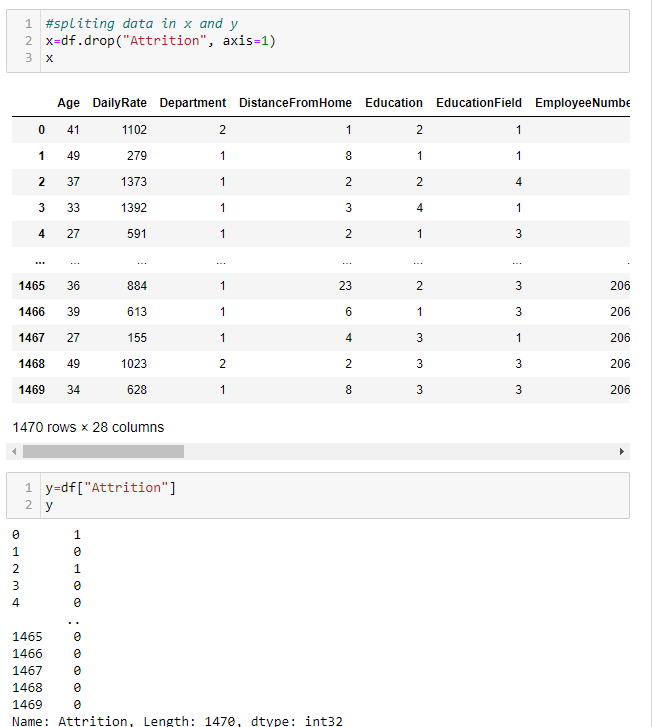


As we can see while applying the outlier IQR function on monthly income, the value has shifted from 20 k to a nearby 15k. We can do the same with all other variables containing outliers. Further, before feeding the data for the prediction, we need to check the skewness of the data and treat it.

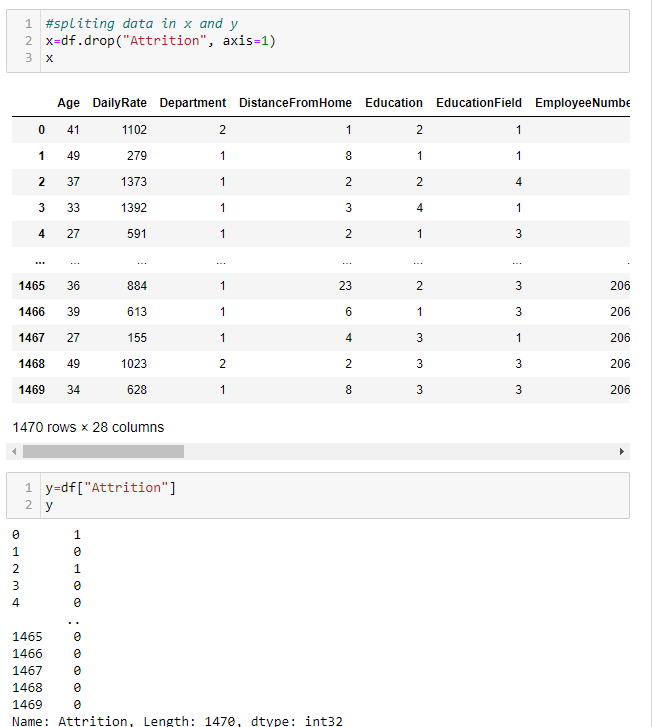


We will consider any value as skewness that is more than 0.5. As we can see, there is skewness in some columns. Price is our dependent variable, so we will not be changing anything in that column. So before proceeding further, we need to data in independent(x) and dependent(y) datasets.

**Splitting the data into input and target variable:**

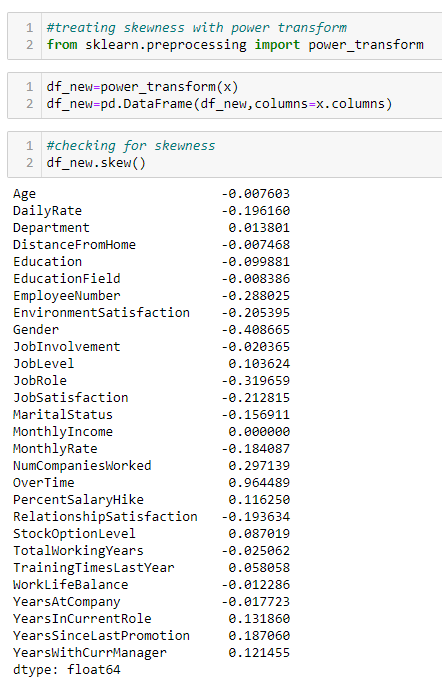


Except for Attrition, we have saved all the data in x.



We have saved Attrition in y

Now we will treat skewness with the help of the Power Transform method.

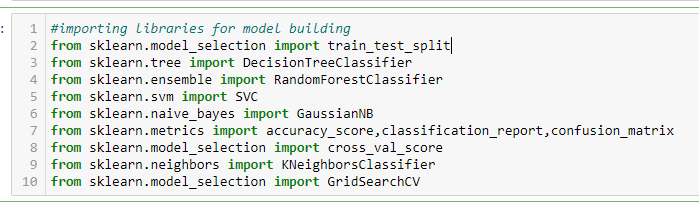


Except for overtime, all the variable's skewness less than 0.5, implying the skewness problem has been treated. Overtime skewness will not be affecting our model as it is a categorical variable. Now before proceeding to the model-building part, we need to balance our dependent variable.

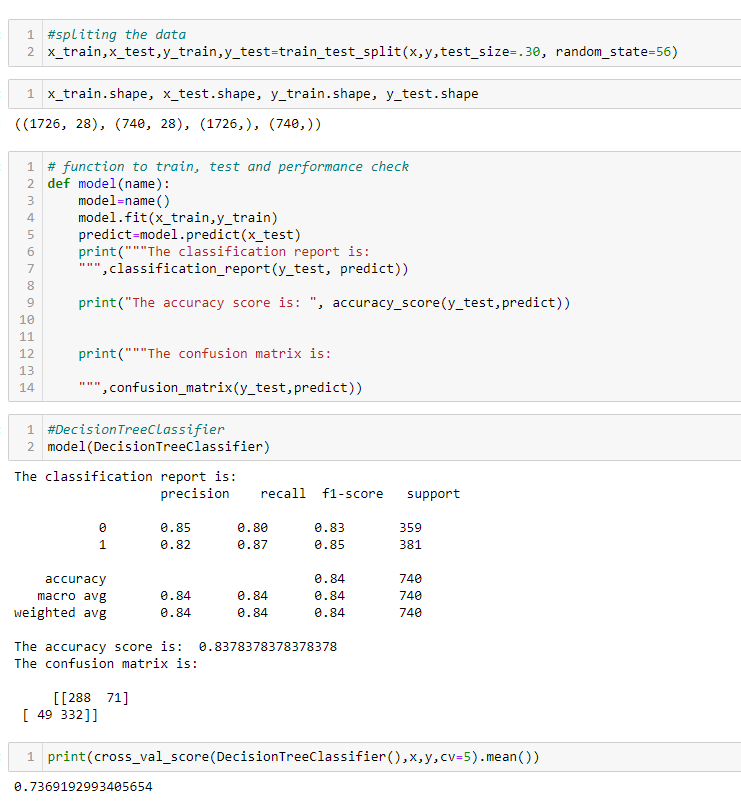


As we can see before and after the graph of the dependent variable, SMOTE function has been able to upscale the data. Further, we will import all the necessary libraries for the model building process.

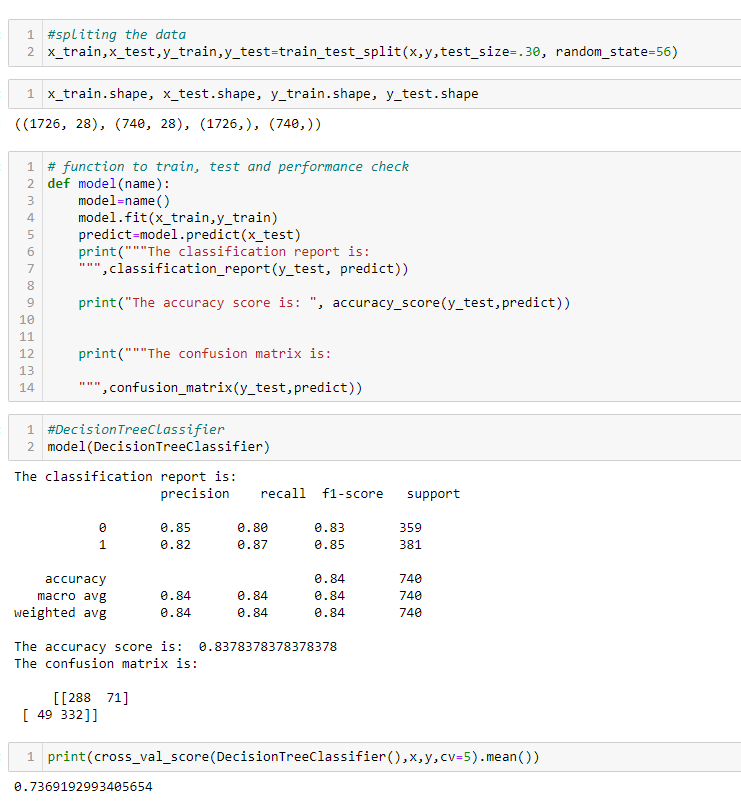
**Importing libraries**:



We have used an accuracy score, classification report, and confusion matrix to evaluate the model. We are also using the cross-validation method to ensure our model is not under or overfitted. Now we will split the data into train and test data to train the model and determine its performance.

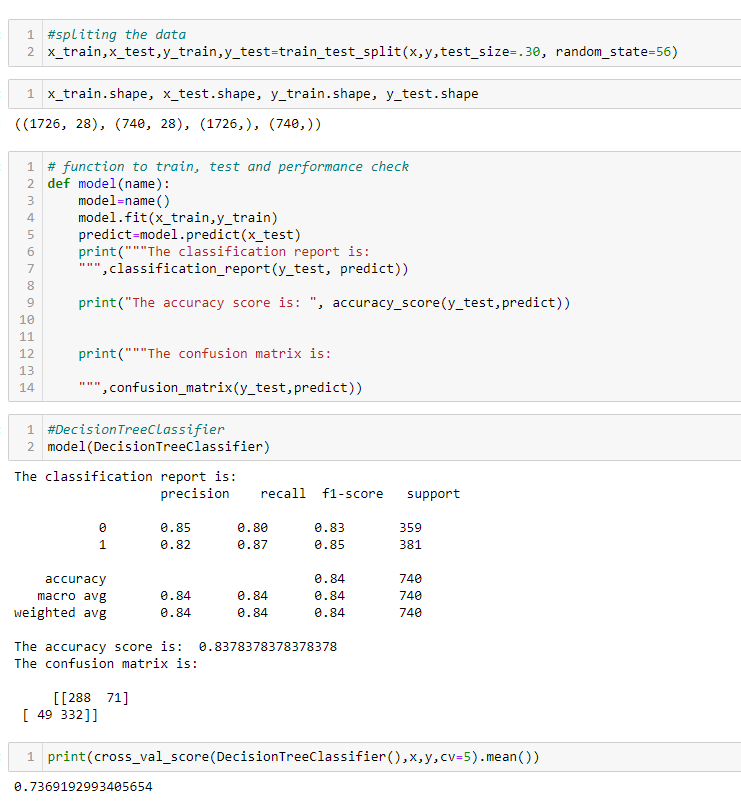


The training data contains 1726 rows, while the testing data contains 740 rows.



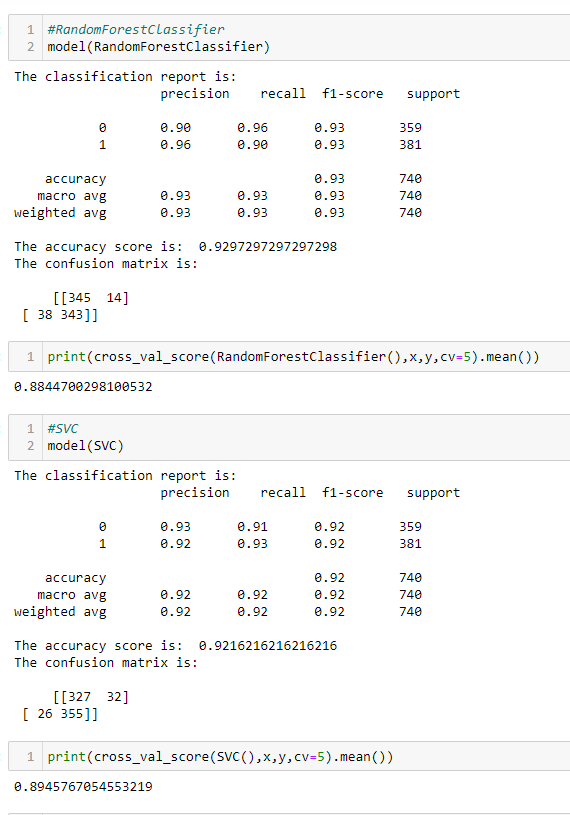
Here we have created a function model that will take the classification model's name, apply all the processes, and give information about the model's efficiency.

**Decision Tree Classifier:**



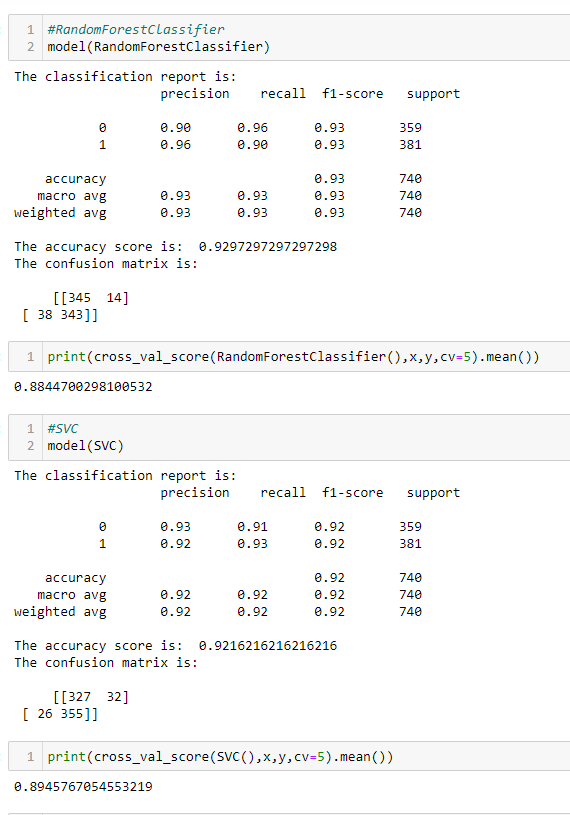
* Decision Tree Classifier accuracy score is 83 percent.

**Random Forest Classifier:**



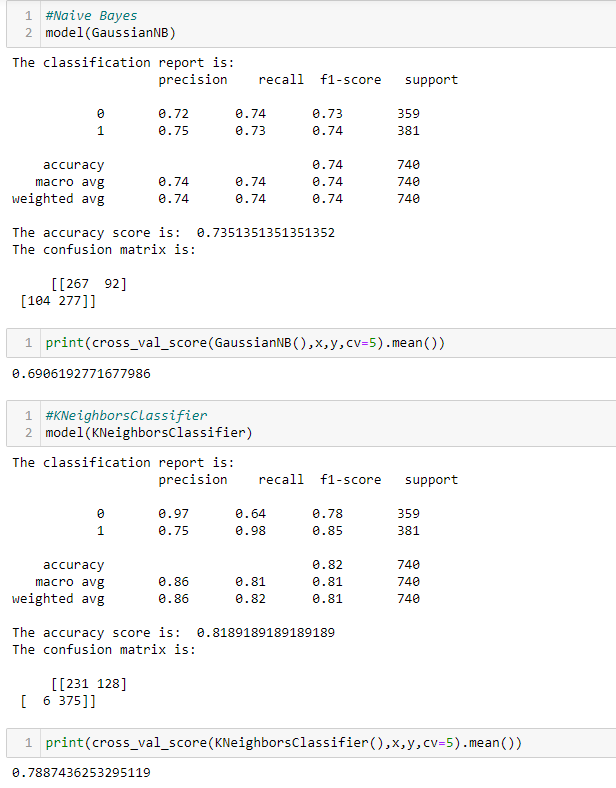
* Random Forest Classifier accuracy score is 92 percent.

**SVC:**



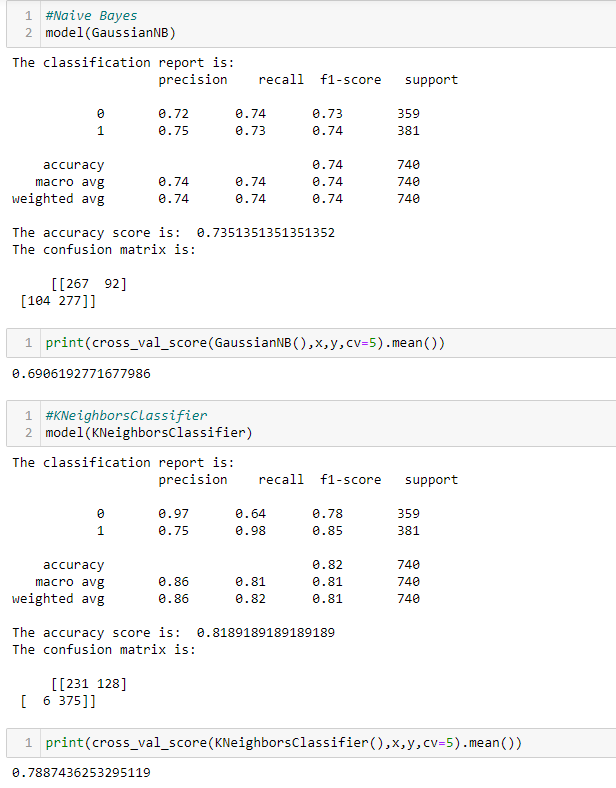
* SVC accuracy score is 92 percent.

**Naive Bayes:**



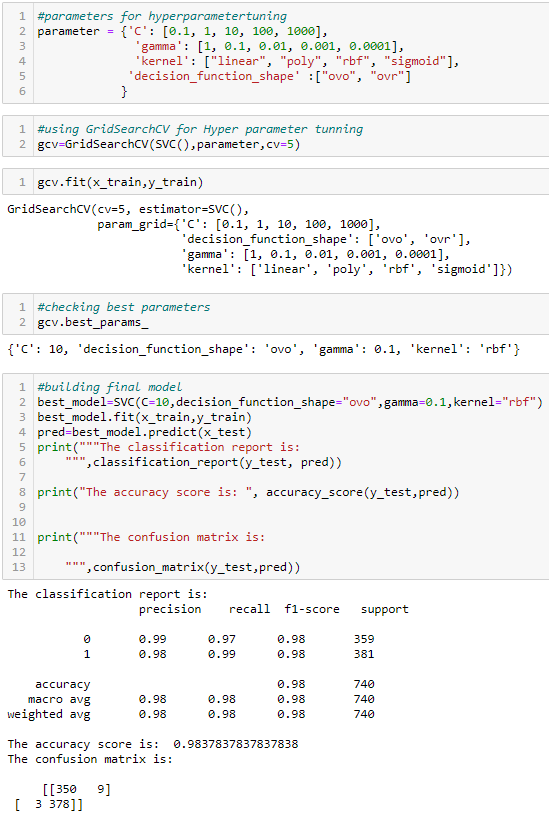
* Gaussian NB accuracy score is 73 percent.

**K Neighbors Classifier:**



* K Neighbours Classifier accuracy score is 81 percent.

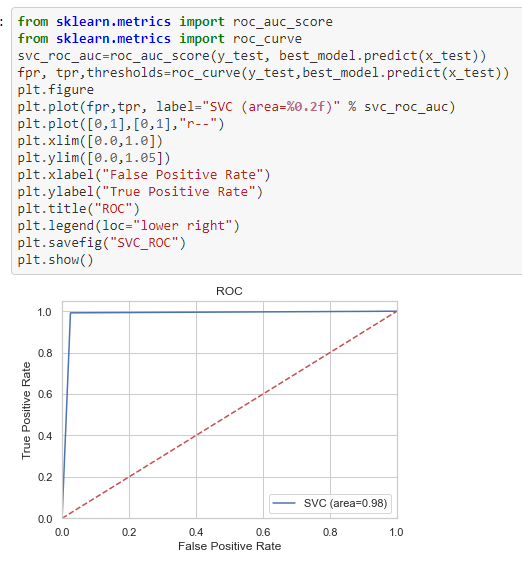
There are two models with an accuracy score of 92 percent; however, SVC is our best model as the difference between cross-val-score and accuracy score is minimum in SVC. Now we will perform hyper parameter tuning to enhance our model.



As we can see, we have set several different parameters to do hyper parameter tuning; with the help of Grid Search CV. We trained our model with each parameter, then with the best parameters, we trained our model, and the accuracy score has resulted in 98 percent. It is a 6 percent increase in the effectiveness of the model. Further, we will use roc\_auc\_score to check the effectiveness of the model.

**AUC - ROC Curve:**

AUC - ROC curve is a presentation estimation for the characterization issues at different threshold settings. ROC is a probability curve, and AUC addresses the degree or proportion of distinctness. It tells how much the model is fit for recognizing classes. Higher the AUC, the better the model is at predicting 0s as 0s and 1s as 1s. By similarity, the Higher the AUC, the better the model recognizes employees is slanted towards Attrition.



As the image shows, the classifier can perfectly distinguish between all the Positive and the Negative class points with the effectiveness of 98 percent.

**Conclusion:**

In this paper, we have gone through the process of prediction model building. The paper showed how to clean data with various techniques and what will happen if we will not clean the data. We have also analysed the data graphically to find out insights from the given data. It was one of the main objectives. Also, we learned how to deal with imbalanced dependent data set as if we will not fix it, our model will not be good and will contain biases. This paper showed how to create a different type of models and how to evaluate them. Also, it showed how to approach with hyperparameter tuning of the selected model. As we can see, due to hyperparameter tuning, our model showed an increase of 6 percent in its effectiveness.