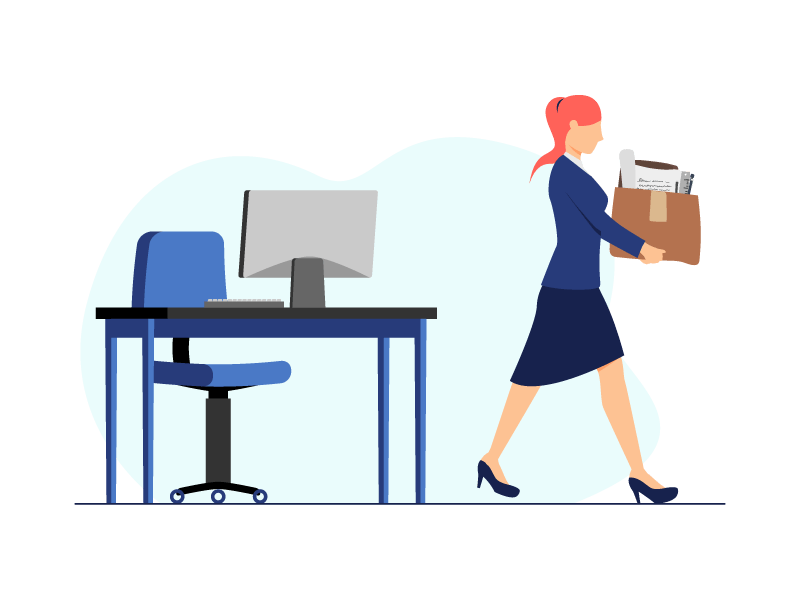
HR Analytics Project- Understanding the Attrition in HR



Blog Submitted To :- Submitted By :-

Data Trained Arti Sharma

Batch - 1836

**Table Content**

* Problem Definition
* Data Analysis
* EDA Concluding Remarks
* Pre-processing Pipeline
* Building Machine Learning Models
* Concluding Remarks

**Problem Definition**

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.**

**Attrition in HR**

Attrition in human resources refers to the gradual loss of employees overtime. In general, relatively high attrition is problematic for companies. HR professionals often assume a leadership role in designing company compensation programs, work culture, and motivation systems that help the organization retain top employees.

How does Attrition affect companies? and how does HR Analytics help in analyzing attrition? We will discuss the first question here and for the second question, we will write the code and try to understand the process step by step.

**Attrition affecting Companies**

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. 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.

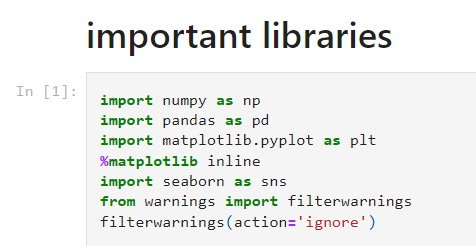
**Overview**

In this article , we are going to discuss employee attrition in the organization that is by predicting that if the employee will leave the organization or not by using some machine learning models.

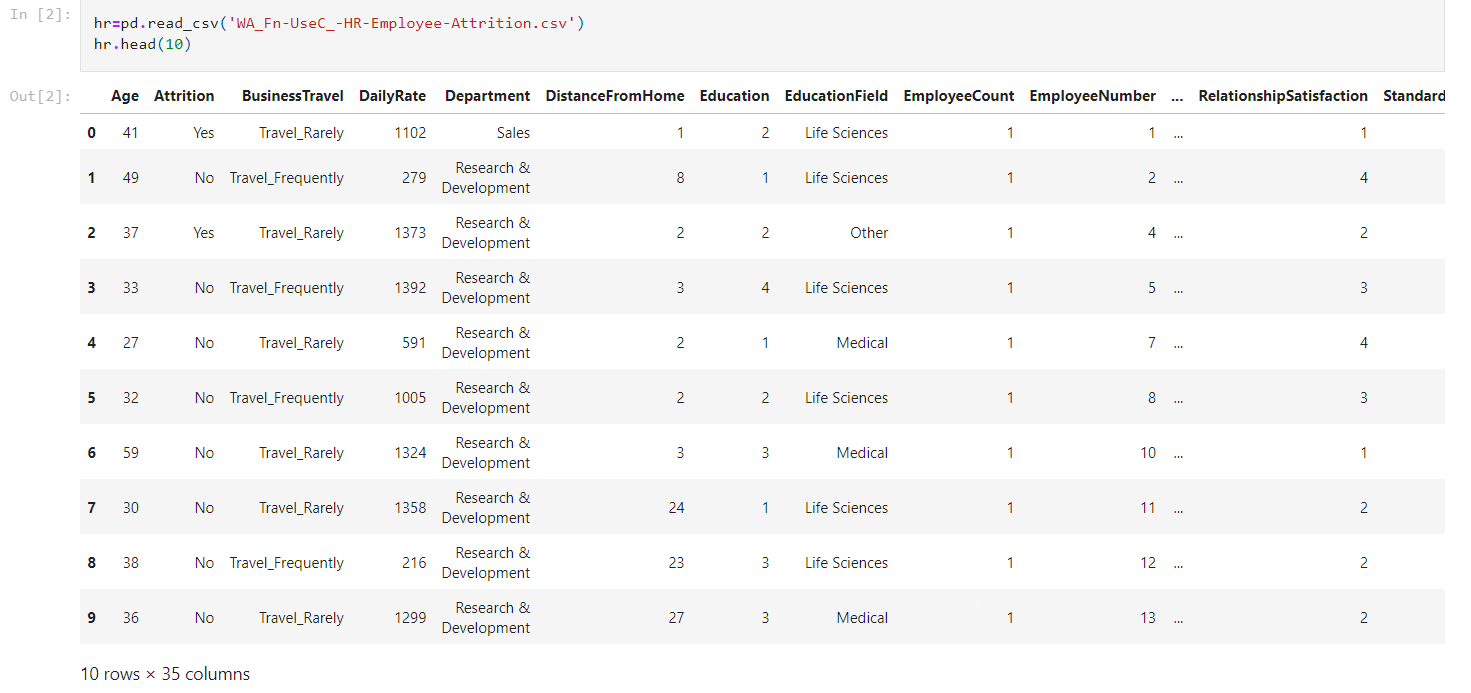
Lets get started….

**Data Analysis**

The very first is step to import the important libraries for data manipulation and wrangling.



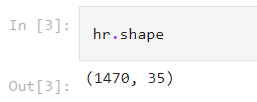
Then , After importing the important libraries we need to load the particular dataset by using respective extensions .for example- dataset can be in csv format, excel format so will import the dataset accordingly.



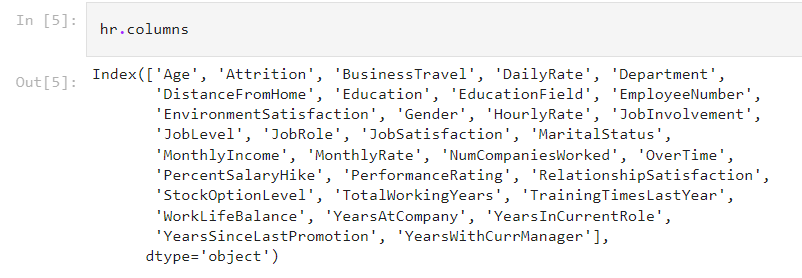
Note:- hr is the variable name.

Our next step to understand the data that means what’s the dataset consists of.. let’s understand that,

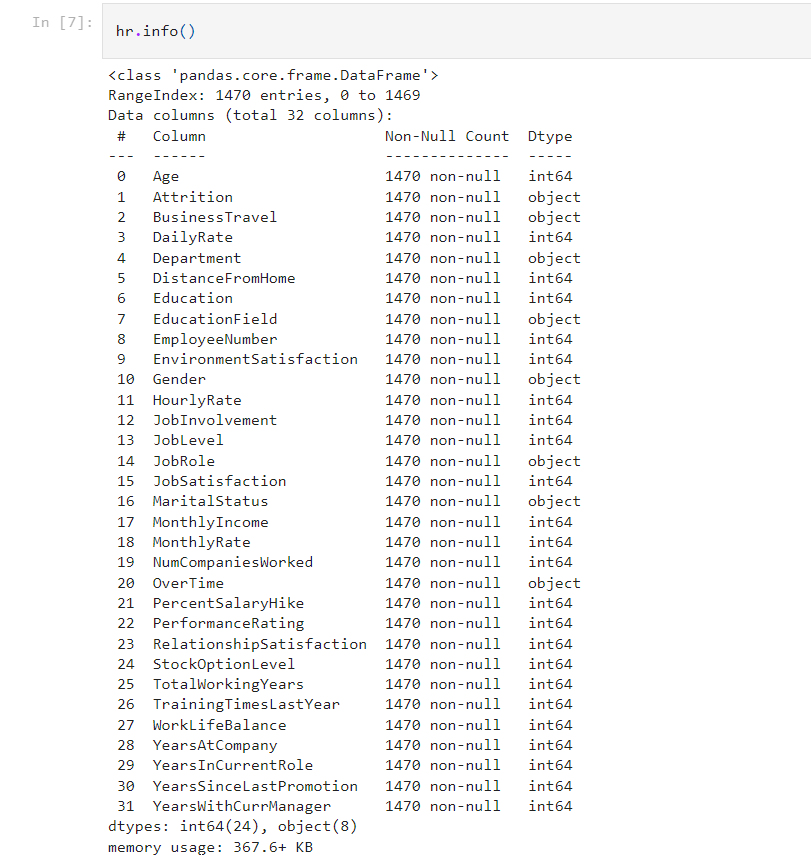
The shape of the dataset:



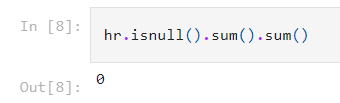
So firstly we have checked that how many rows and columns our dataset have. We found that we have 1470 rows and 35 columns by using hr.shape method.



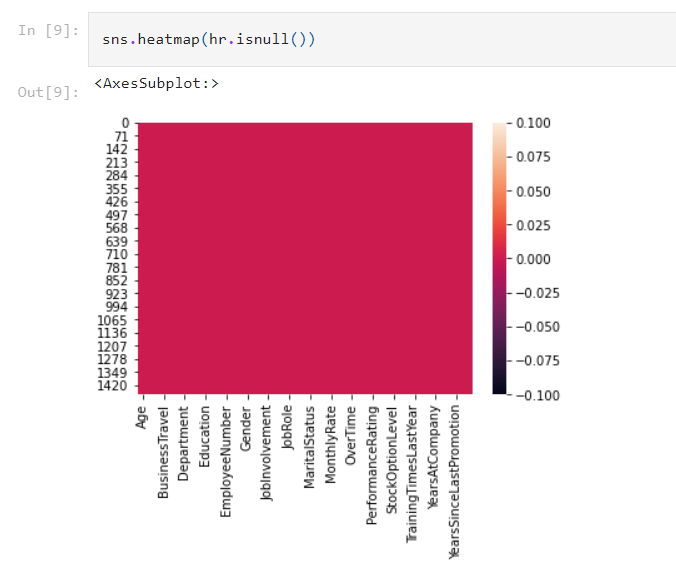
By using hr.columns we get to know about our columns name.



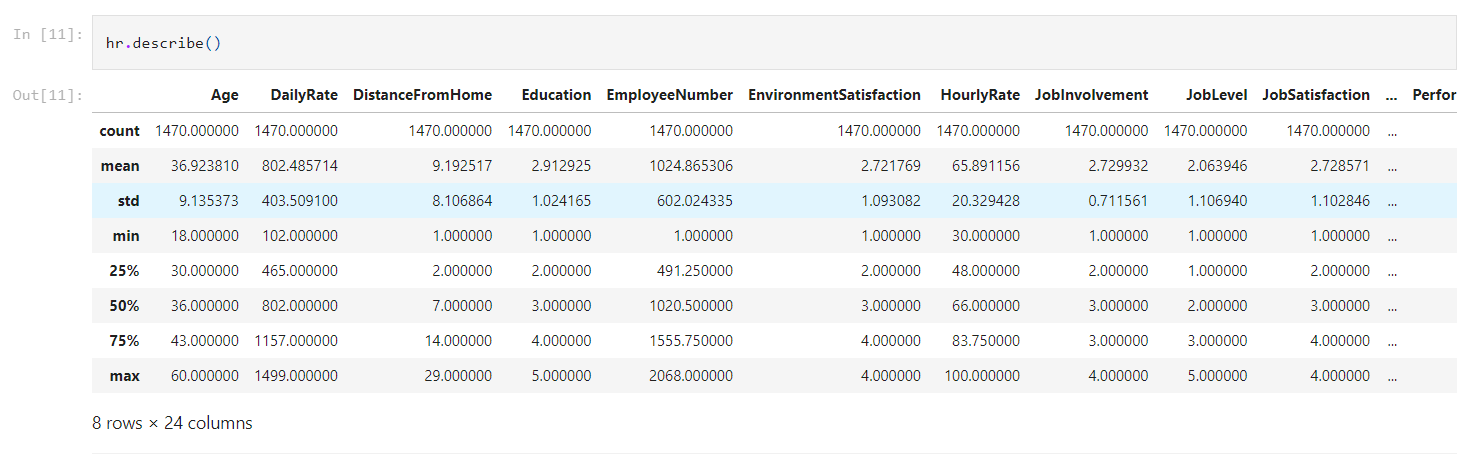
By applying hr.info method we will be knowing about detailed information of our dataset.



By using .isnull or .isna method we got to know that there is no-null values present in our dataset.



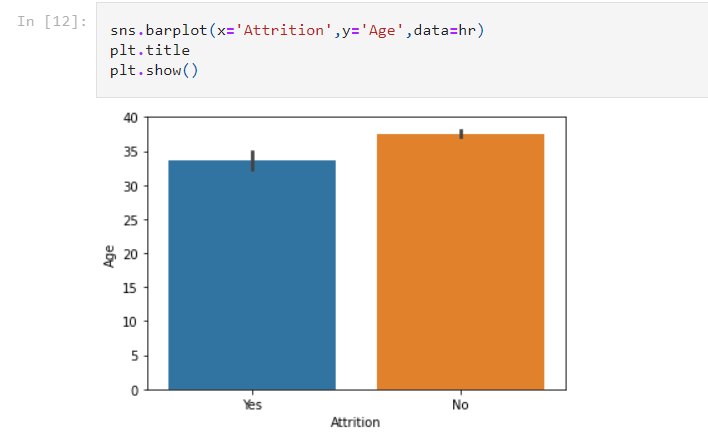
So here plotted the heatmap for null value and it is of same colour which represents no null values.



Now the ( .describe) method which is going to tell all statistical summary of the dataset.

**Exploratory Data Analysis:-**

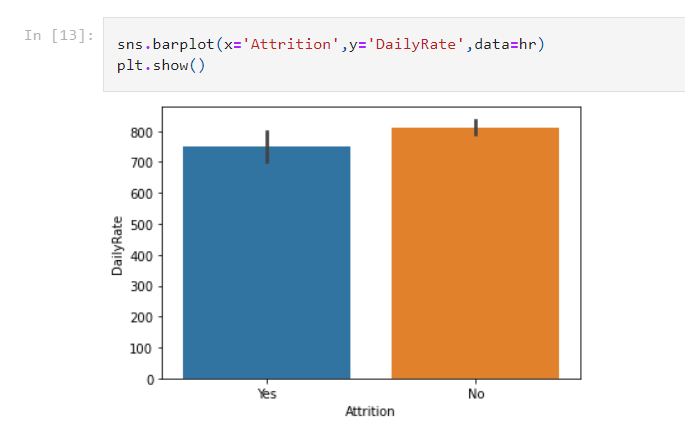
**Age Factor**:

****

if we see the attrition rate with respect to age so we can find that employee aged between 30-35 leave the job so often whereas employee aged after 38 stays.

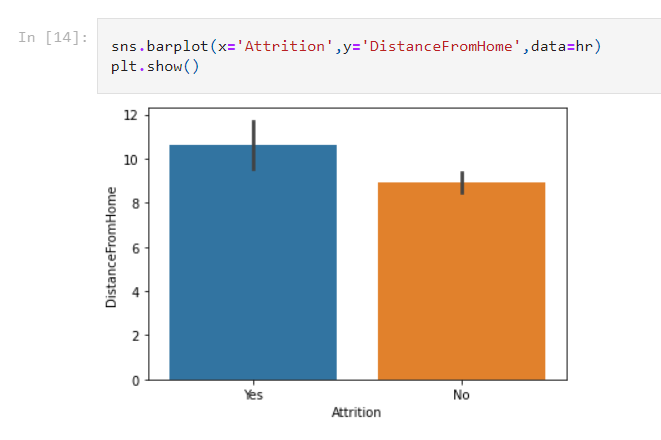
It may be result of keenness of employee of getting promoted, good pay ,job satisfaction and there may numbers of reason to change the jobs.

**Daily Rate:**



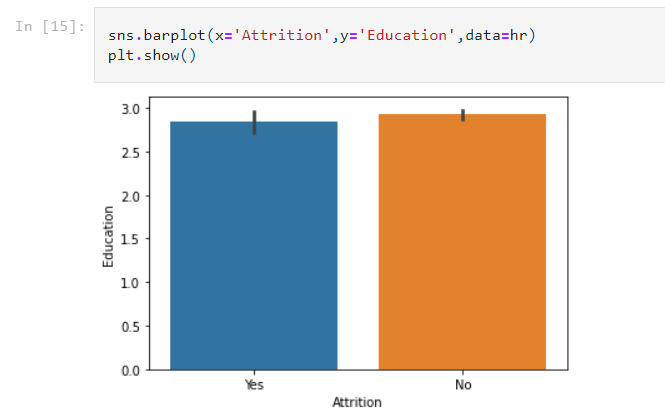
* Daily rate refers the employee’s salary on daily basis.
* Higher the rate less they leave.

**Distance from home:**



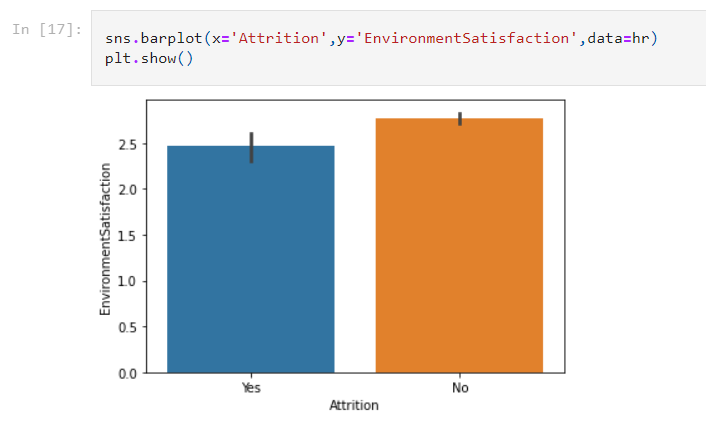
* Distance from home definitely bothers the employee.
* Distance lead to increase in daily expenses.
* 10 out of 12 employee leaving the job just because they travel far from their home.
* Distance cause a lot of problem including their travel expense, consumes time, affects the work life balance.

**Education:**

****

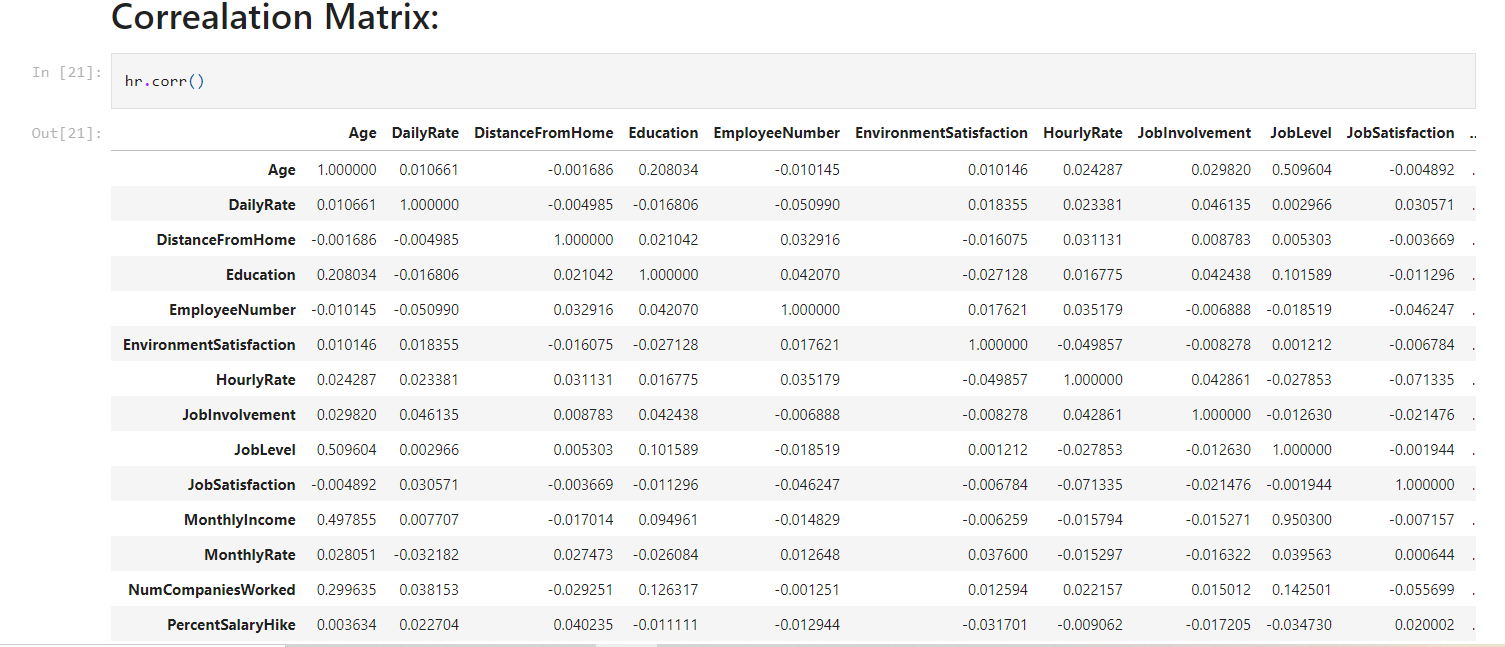
* Attrition rate with respect to education rarely matters.
* Education doesn’t affect the attrition rate.

Environment Satisfaction:



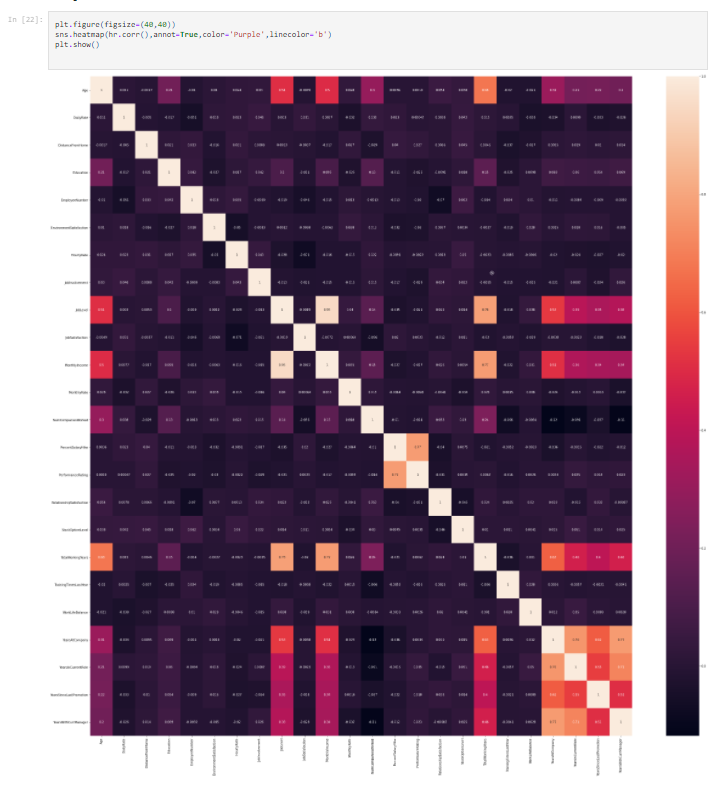
* The working environment is one of the crucial factors which influence the level of satisfaction as well as the motivation of its employees.
* As the bar suggests attrition of satisfied employees is higher that unsatisfied.

**Pre-Processing pipeline:**

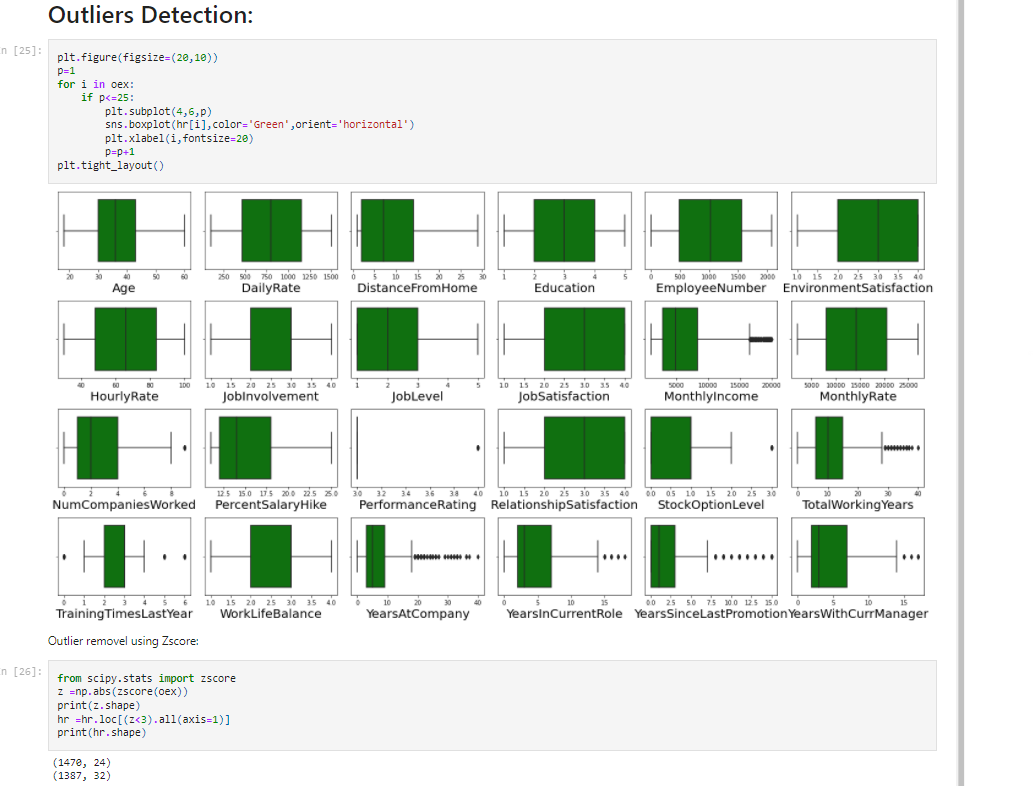
****

* Here we are trying to get the correlation between variables.

**Let’s see correlation between the variables by plotting the heat map.**

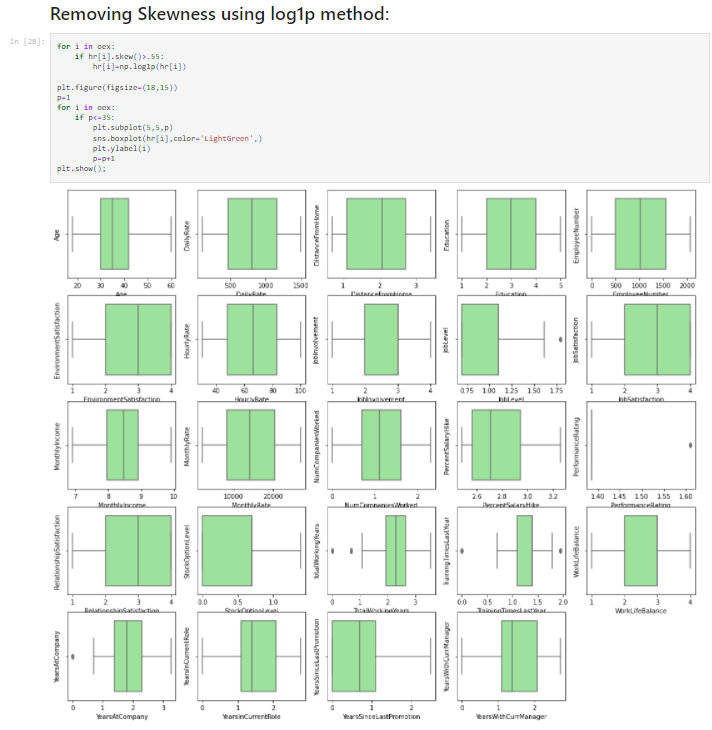
****

**Outlier Detection using boxplot:**

****

* Wedetectoutliers to get to know about extra data that does not fit in the rest of the data.
* As we can see MonthlyRate, NumComapniesWorked, TotalWorkingYears, TrainingTimesLastYear, YearsAtCompany, YearsinCurrentRole, YearsSinceLAstPRomotion,YearsWithCurrentManager has most outliers.
* We have treated the outliers by using ZScore method.

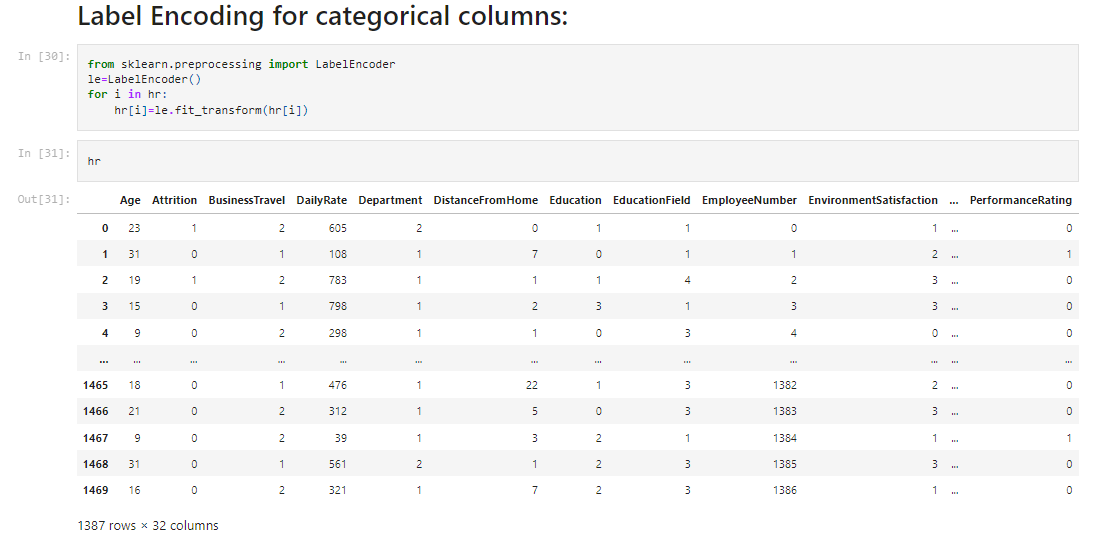
**Then will check if the data is skewed .**

****

* Skewness in data can affect the result of the model performance.
* Skewness should lies between -0.5 to +0.5.
* Now, Removed the skewness by using log transformation.
* Now the outliers and skewness of the data has been treated.

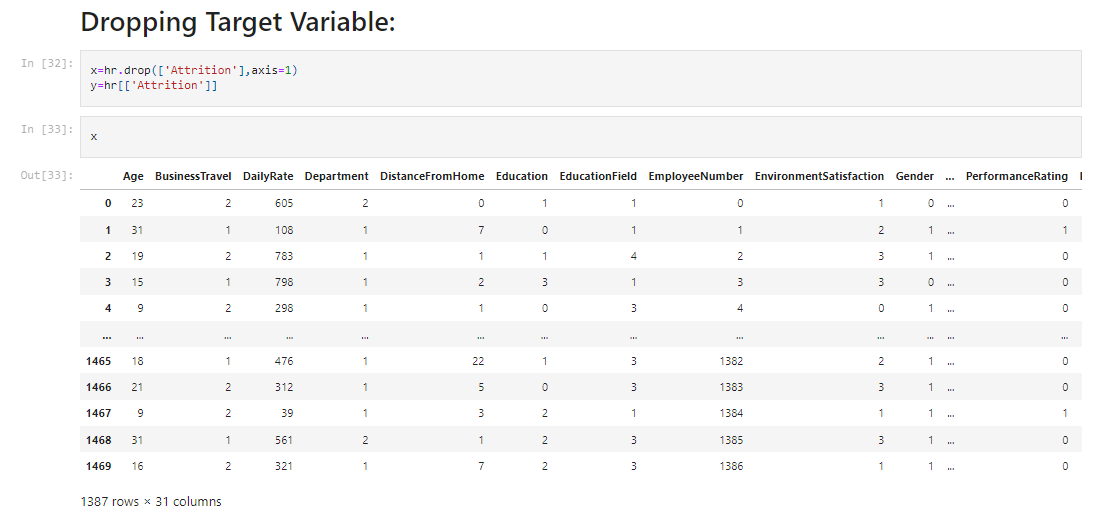
**Encoding:**

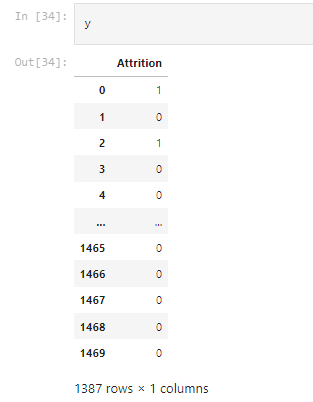
As we know programming language understands only numbers. So now we are going to encode the categorical variables to numerical by using LabelEncoder.



* Now the data is ready for further validation.
* Each columns has been converted into numbers now we can build the model.

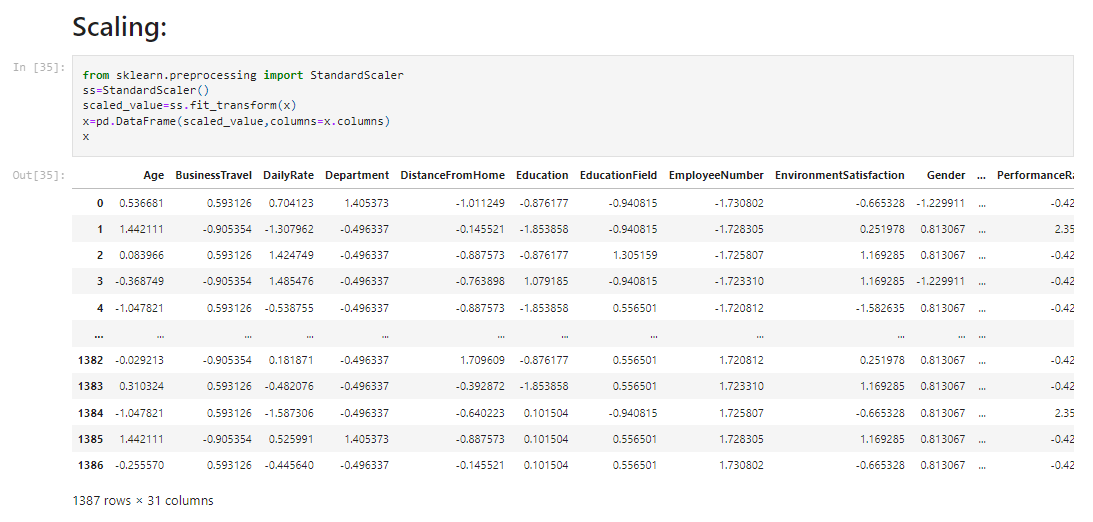
**Now it time to build the model, So now I am going to drop the Target Variable.**

****

****

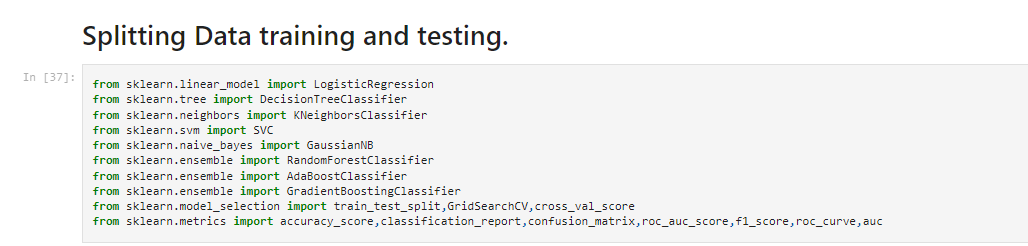
**Feature Scaling:**

* Feature scaling is important for machine learning algorithms that calculate the distances between data.
* Range of all the features should be normalized so that each feature contributes approximately proportionately to the final distance.

****

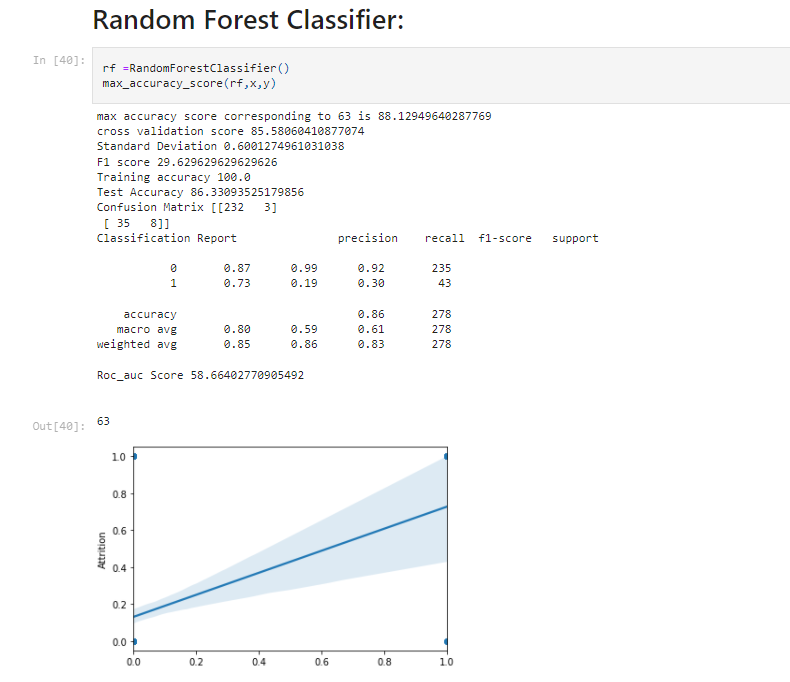
**Building Machine Learning Models:**

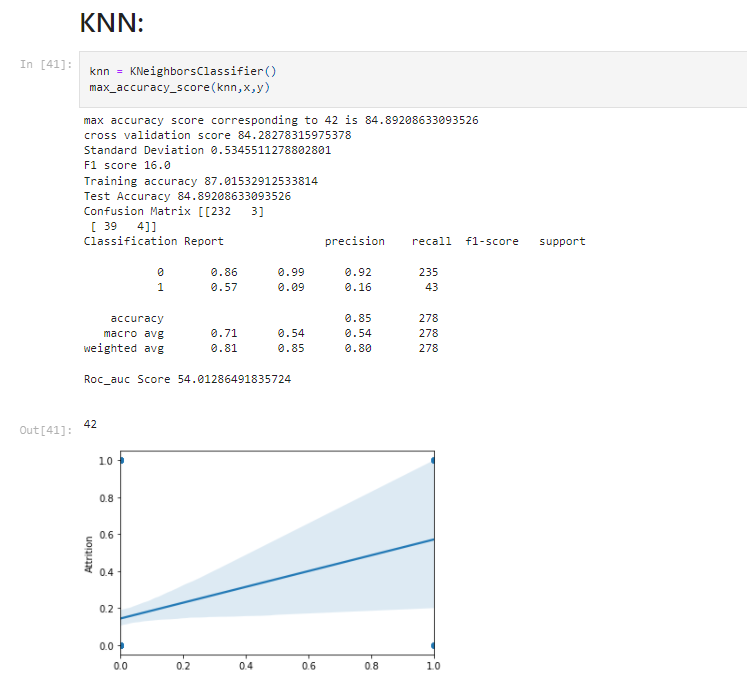
* To build the model we need to import the important libraries.

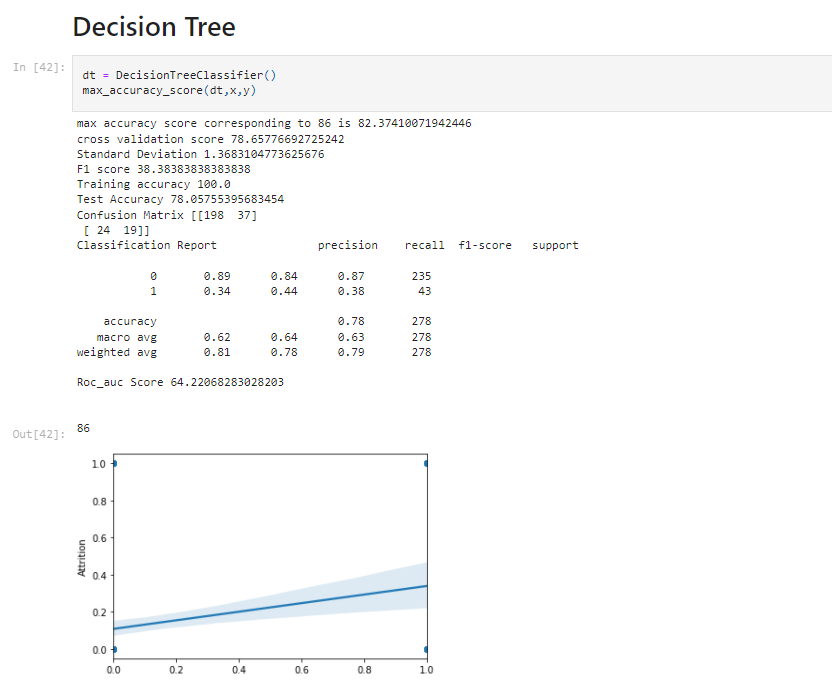


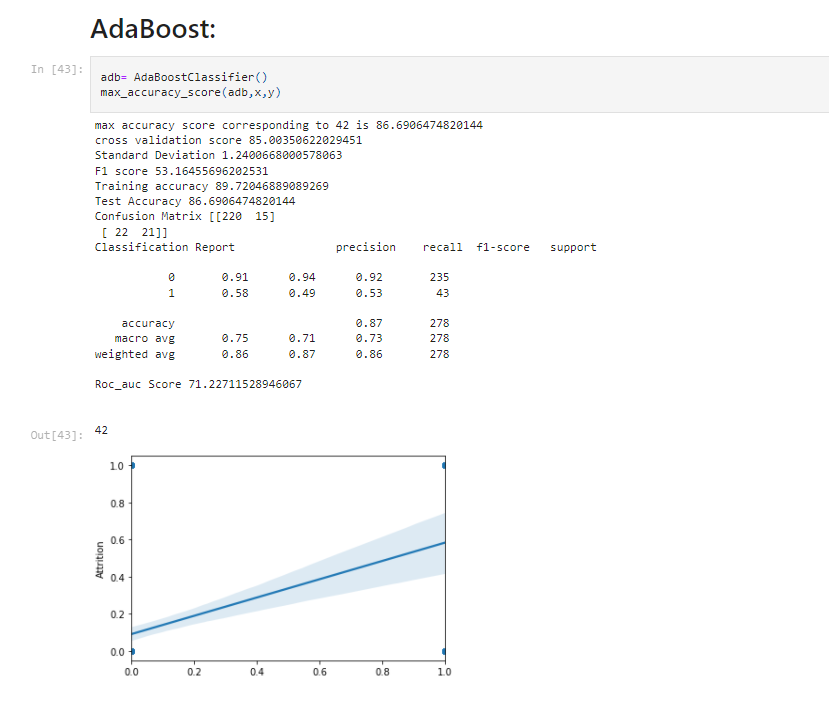
****

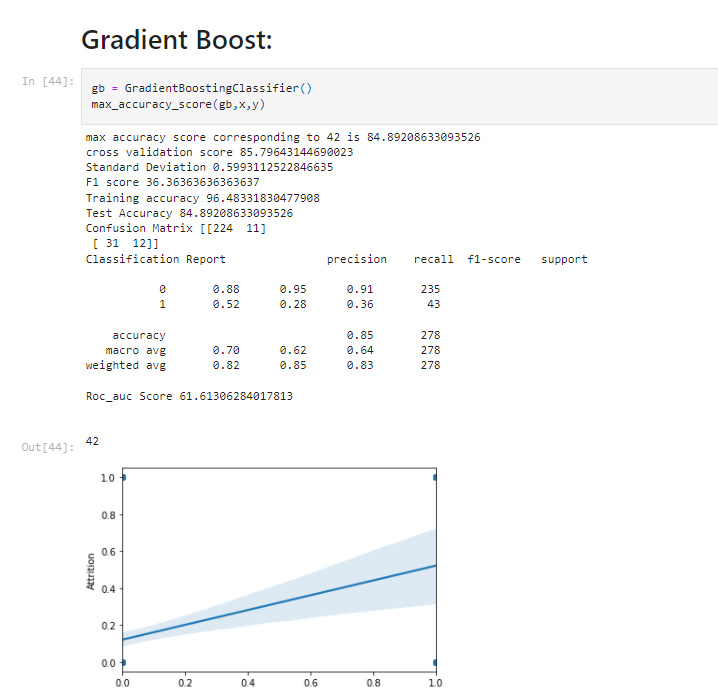
****



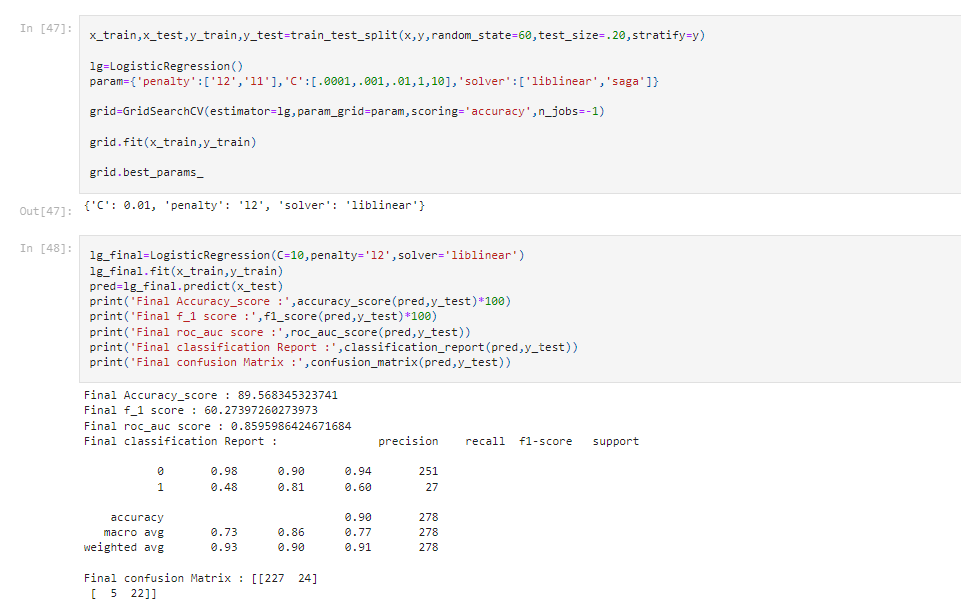
****

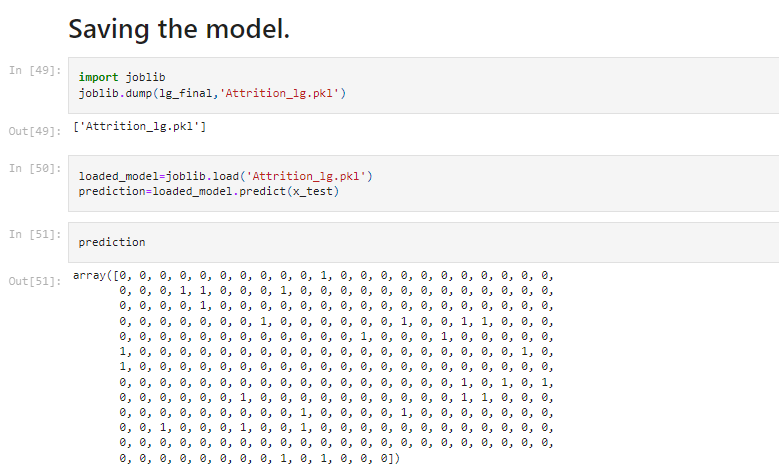
****

****



**Hyperparameter tuning:**

****

**Save the model.**

**Remarks:**

* 0 marks as stays.
* 1 marks as left.
* Data is biased towards non-attrition
* Number of employee that’s stays is more than those left.
* Build the model with best possible approach.
* Performed hyper parameter tuning on logistic regression.
* Printed the Roc-Auc curve to get good visual of the model.

****