# Blog on Implementing Machine Learning On HR Analytics Dataset.

Contents:

1. Problem Definition
2. Data Analysis
3. EDA
4. Pre-Processing Pipeline
5. Building Machine Learning Models
6. Conclusion

## Problem Definition:

In this Machine Learning predictive case study, a HR dataset was taken from the GitHub site from dsrscientist repository and we have attached the link of the dataset from where we got the dataset: - “https://github.com/dsrscientist/IBM\_HR\_Attrition\_Rate\_Analytics” which contains data of the 1,470 employees with 34 different features/attributes about the employees of the company. The dataset contains different features like their Age, Education, Department, business Travel, distance from home, Total working Hours etc. which we will be using to see whether these attributes contribute for predicting that the employees are going to quit the company or not. Here we have chosen the Target variable as “Attrition” column of the dataset which contains two values Yes or No. so, this problem is **classification** type. We will use different classifiers to predict the attrition of the company.

## Data Analysis:

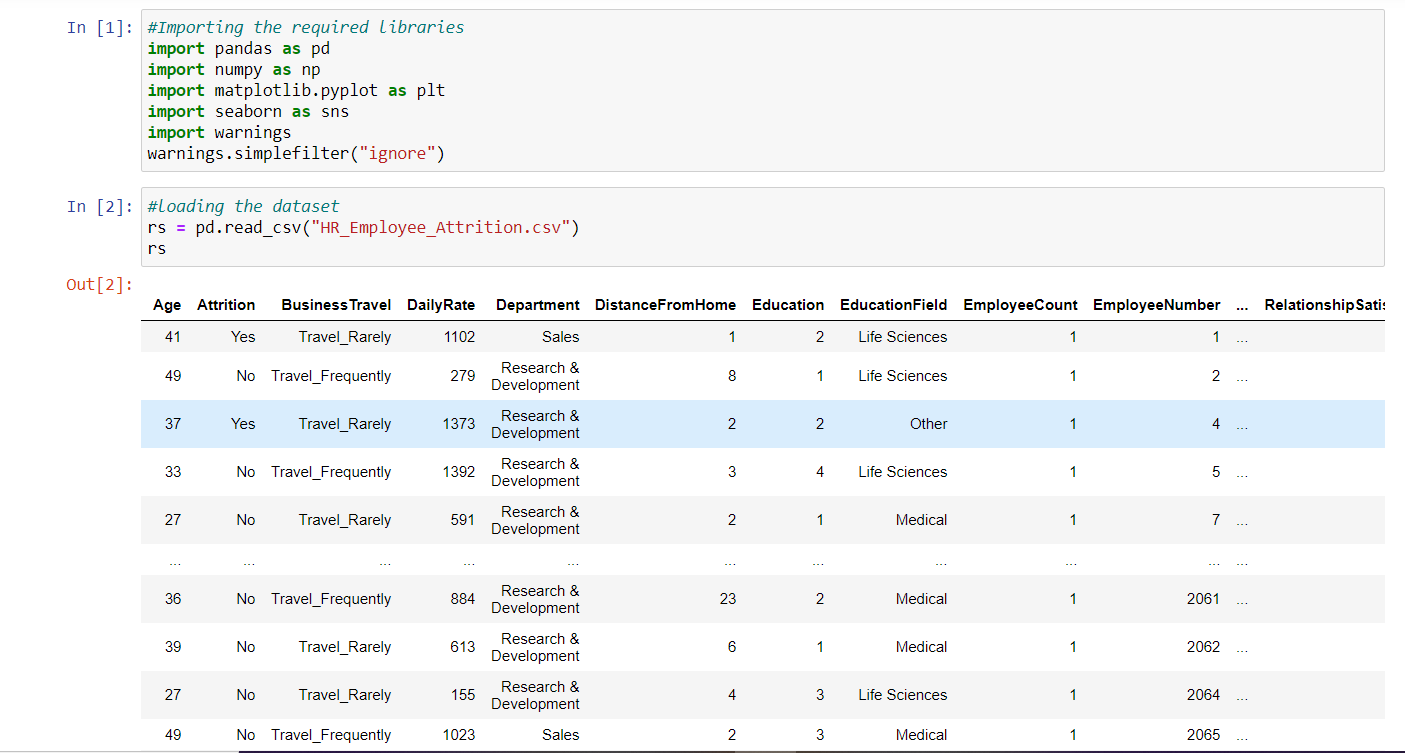
Here in this section, we will analyse the dataset by loading it from jupyter notebook, checking for data types of attributes of the dataset, checking for information of the dataset to see how many numerical and categorical columns are present in the dataset, checking for the null (NaN) values in the dataset to see whether data cleaning is required or not.

First of all, we will import the required libraries which we will be used for coding the dataset. We will import pandas library to import the dataset as csv (comma separated value) file from the Jupyter Notebook.

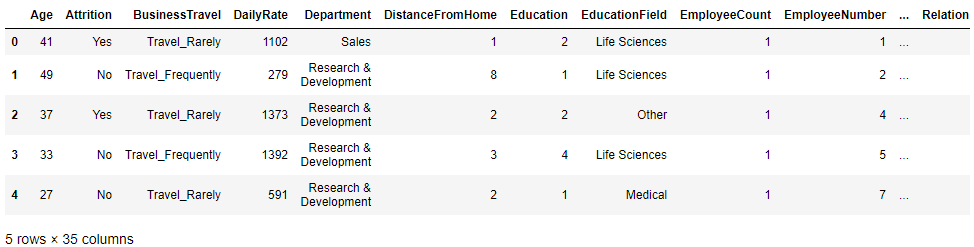
We have loaded the dataset using pandas library and saved the dataset in the “rs” variable. We have printed the dataset by calling instance of the dataset “rs”.

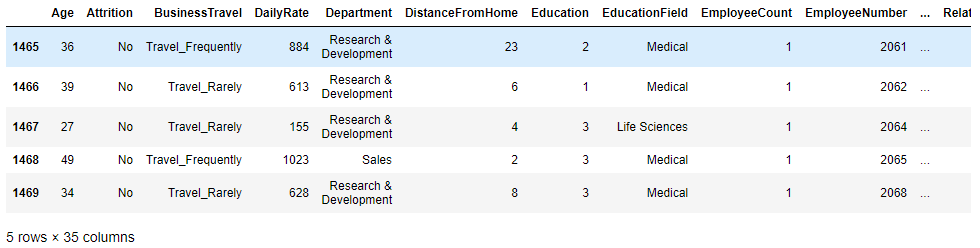
We have used the command rs.head() to display the first five rows of the dataset, rs.tail() to display the last five rows of the dataset, rs.sample() to display a random sample of the dataset

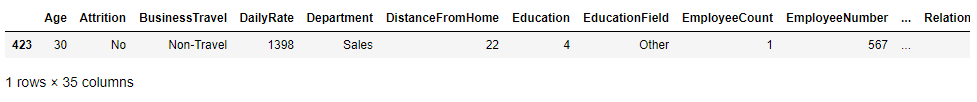
Which is displayed in the figures below.

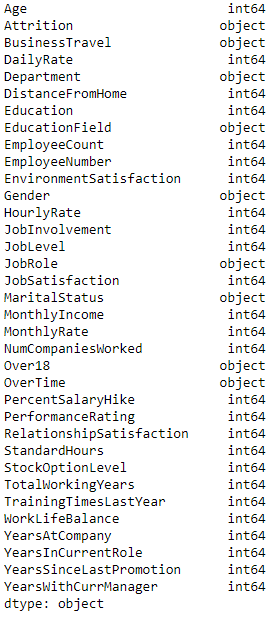


After loading the dataset, we check the information of the dataset by using rs.info () command, then we check for the dimension of the dataset by calling the rs.shape, which tells us that 1470 rows and 35 columns. Then we check for the data types of the dataset using the rs. dtypes which tell us that the dataset contains 26 columns of int(numerical) data type and 9 columns of object(categorical) datatype.

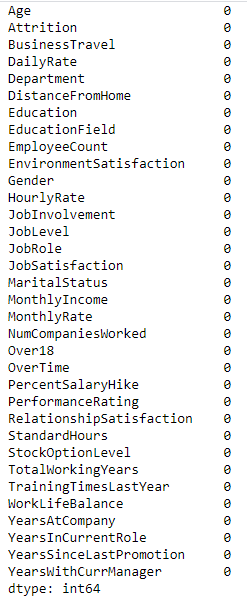




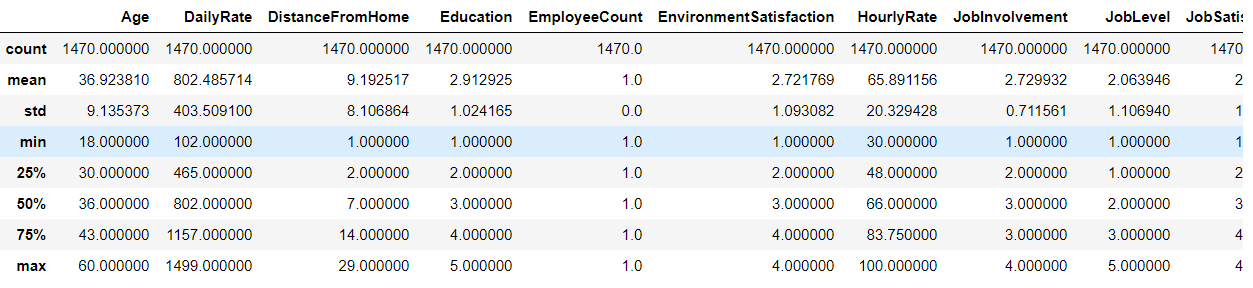




Then we proceed to check for the cleaning of the data if required i. e, by checking if the dataset contains the null values by using rs.isnull().sum() command and using graph with the help of heatmap also we check for the null values, which tells us that there are no null values present in the dataset. So, data cleaning is not required.



Then we check the statistical summary using the describe method i.e., rs.describe(). The figure is attached below.



The observations of the statistical summary are as follows:

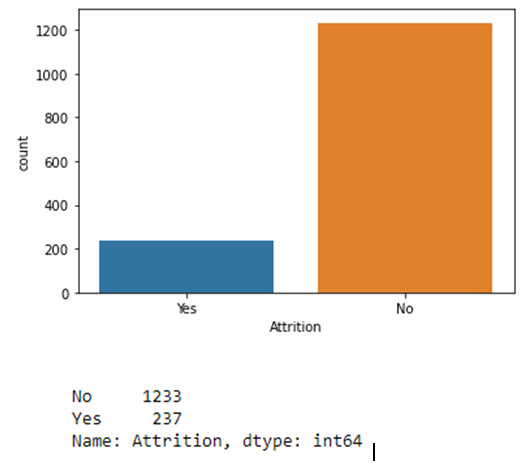
1. first of all, the count of all the columns is 1470. so, there are no missing values in the data.
2. In the columns, Age, Daily Rate, Distance from Home, Education, Environment Satisfaction, Hourly Rate, Job Involvement, Job satisfaction, Relationship Satisfaction, Work life Balance have mean value greater than the median values. So, we can say that the data in these columns are skewed data.
3. In the columns Standard Hours, Employee Count columns all the values are equal to 80 and 1.0. so, it is not providing any info. So, we can drop these columns.
4. In this columns Job Level, total Working Years, Years at company, Years in current role, Year since last promotion, Years with the current manager shows that it has median values greater than the mean values.

From the conclusion of the statistical summary, we can drop the following columns: - Standard Hours, Employee count, Employee number using drop command(rs.drop()) of pandas library as they are not contributing in the analysis of the dataset.

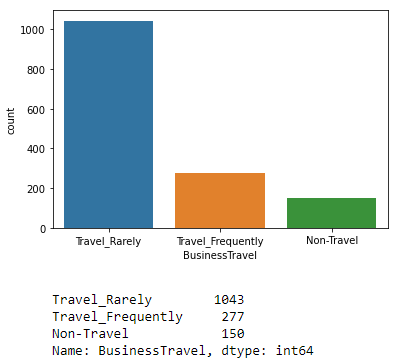
## Exploratory Data Analysis:

Here in this section, we will visualize the dataset by using some libraries like Matplotlib which is imported as instance plt and Seaborn which is imported as instance sns. Both these libraries help us to plot the different attributes of the dataset by using bar plot, histogram, count plot, scatterplot, swarm plot, etc.

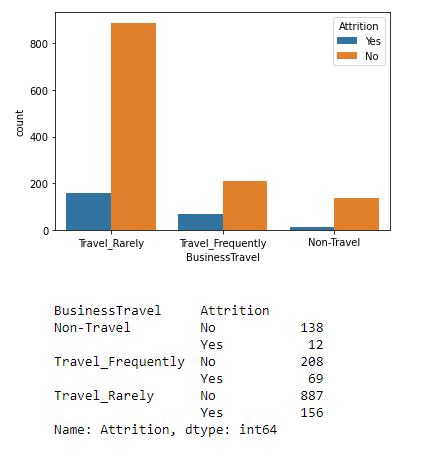
First, we have seen the “attrition” columns wit help of count plot which is shown below in the figure and it tells us that attrition rate is very low. And it also shows that attrition column is imbalanced.



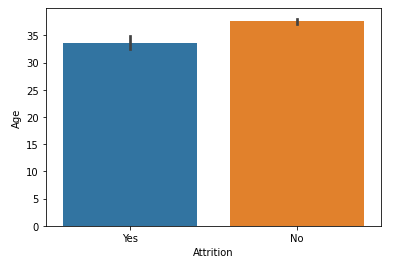
Then we have plotted the “Business Travel” column using the count plot method which show that’s the how does the employee travel for the purpose of business. The employees who travel rarely has the highest count of 1043 followed by who travel frequently with the count of 277 and thos who not travel with count of 150.



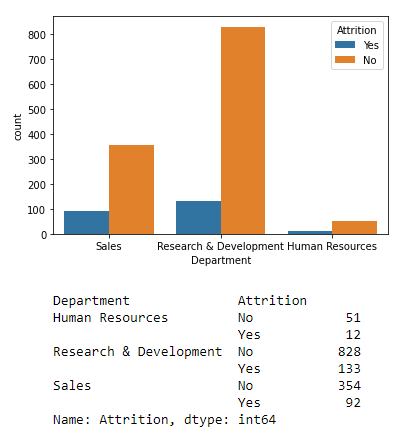
Here we have checked for how does the Business Travel affect the Attrition of the company. The count plot tells us that attrition rate is higher where the employees travel frequently followed by those who travel frequently.



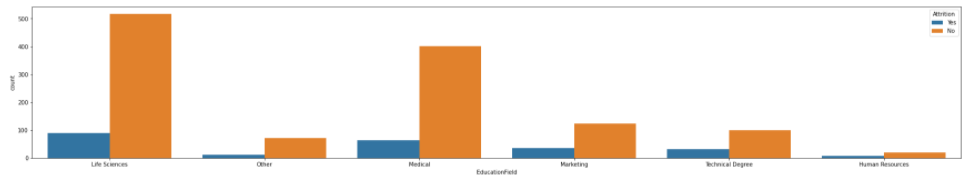
Here we have plotted the bar plot to see how the age affect the attrition. It shows the employees below age of 35 has left he company and age above 40 has not left the company.

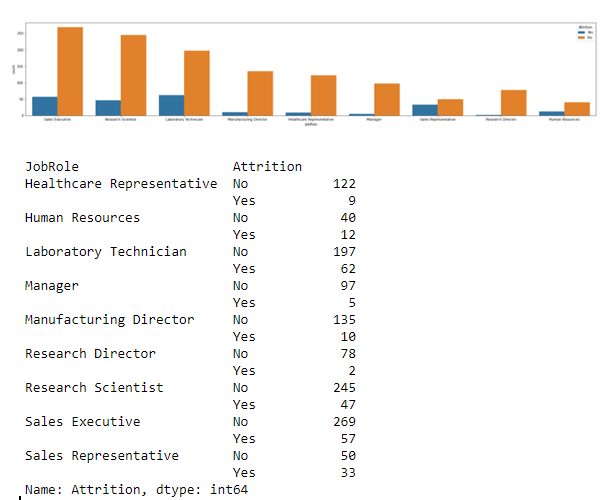


Next we have checked for the “Department” column how it affect the attrition of the company. We have plotted the count plot of attrition with respect to attrition which tell us that attrition rate is 1:4 in Human Resources,1:8 in Research & Development,1:4 in sales.

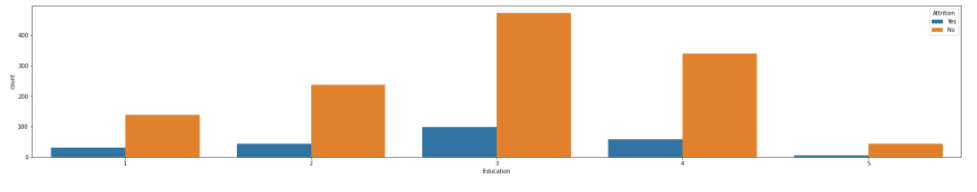


Next, we have plotted the count plot for the “Attrition” with respect to “Education field”. The plot tells us that there are six types of education fields in the dataset: “life Science”, “Other”, “Medical”, “Marketing”, “Technical Degree”, “Human Resource”. The plot tells us that attrition rate is higher in Medical and Life science columns compared to others.

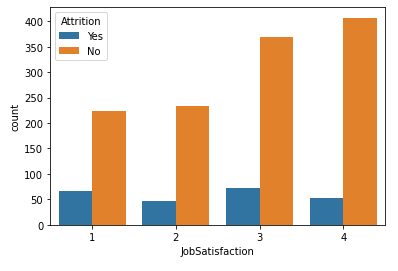




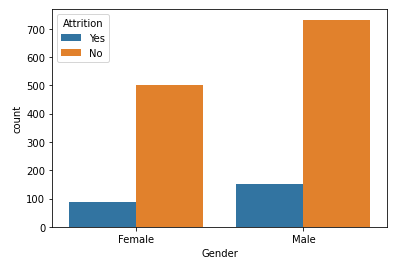
We have plotted the “Education” column how it affects the attrition of company. It shows it has five different levels of education: - 1- Below college, 2- college, 3- Bachelor, 4- Master and 5- doctor and the plot shows that the attrition rate if higher in the employees with “Bachelor” or 3 education level.



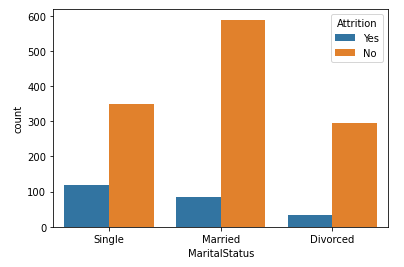
Plotting the “job Satisfaction” for seeing the attrition rate of company we have used the count plot. The plot shows that job satisfaction has 4 different levels as 1-Low, 2- Medium, 3 – High, 4- Very high. The plots shows that the attrition rate is higher in the where the job satisfaction is very low as compared to other columns.



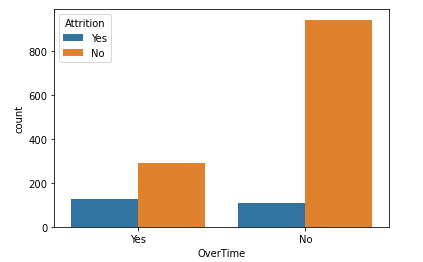
The count plot of the “Gender” column shows that the male’s employees of the company higher attrition rate as compared to the female employees.



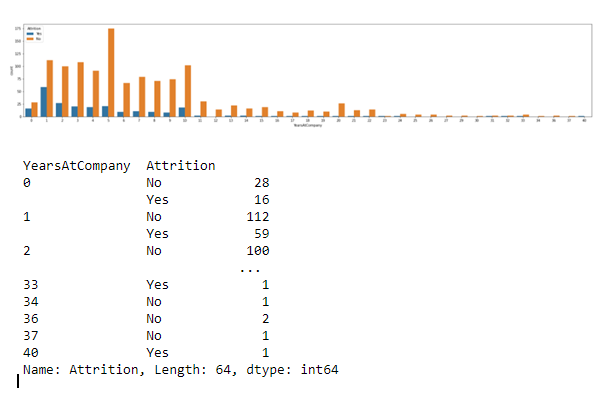
Then we have plotted the attrition rate with respect to the martial status of employees using the count plot. The plot shows that the it has three different type: single, Married, Divorced. The plots tell us that the attrition rate is higher in the single type of employees as compared to the Married and divorced type of employee.



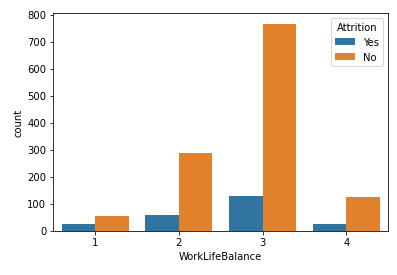
Then we have plotted “overtime” column with respect to “attrition” using the count plot and the plot shows that attrition rate is higher when the employees are doing overtime.



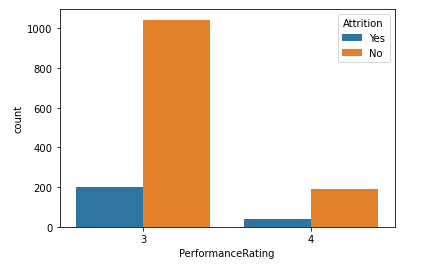
Then we have plotted attrition with respect to “years at company” using the count plot and the plots shows that the employees have left the company in the first five years of joining the company.



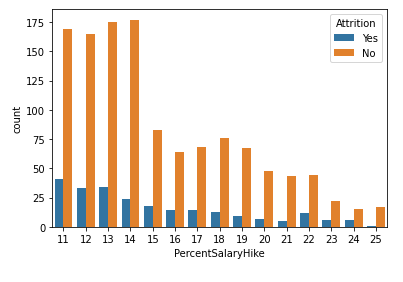
In the columns “year with current Manager” the values count shows that attrition rate is higher in the starting i.e. 0 years. Then we have count plot for the “work life balance” and attrition. it shows that’s the attrition rate is higher where the work life balance is 1.



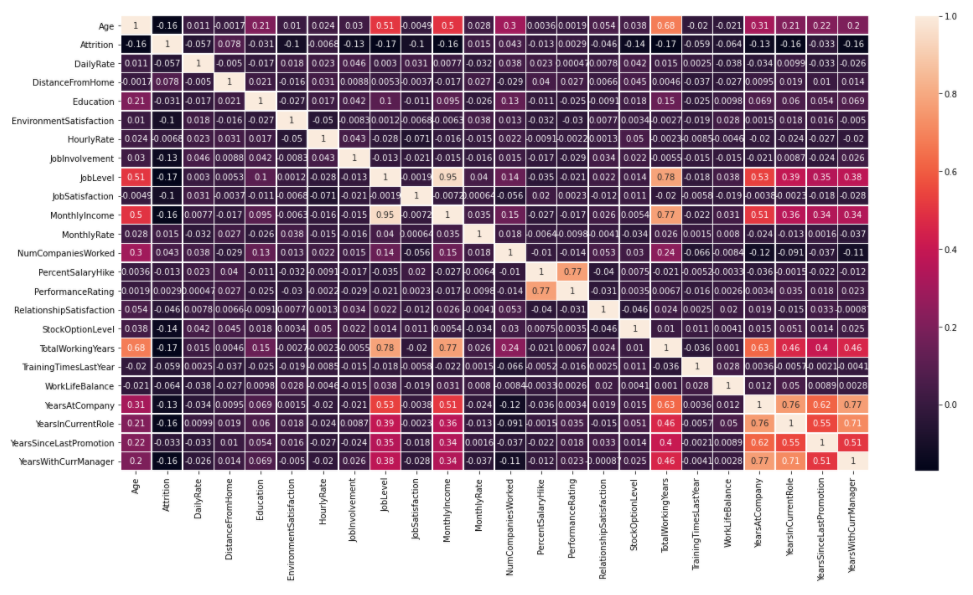
Then we have count plot for the “performance rating”. it shows that employees with performance rating 3 have higher attrition rate.



Then we have a plot of percent salary hike and attrition and it shows that attrition rate is higher where percent salary is low.



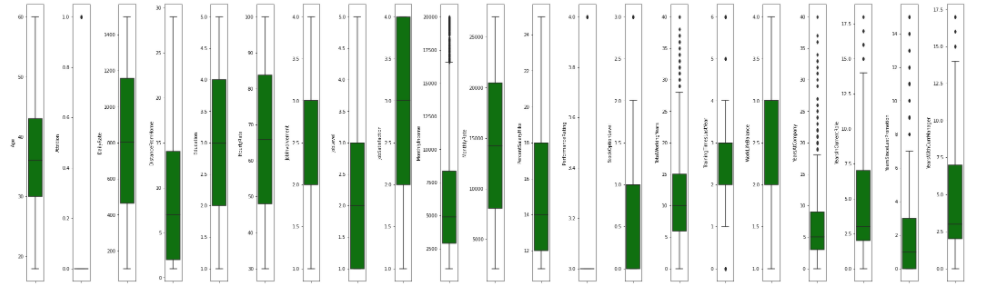
Then we have multi variate analysis by using the correlation function. We have used rs.corr () command to find the correlation of the different attributes with respect to the target variable “attrition”. Then using a heatmap we have plotted it.



The observations are as follows: -

1. We can see that most of the columns are weakly correlated to each other.
2. There is positive high correlation between percent salary hike and performance ratings.
3. There is positive high correlation between Monthly income and total working years.
4. There is positive high correlation between job level and total working years.

Then we have plotted the boxplots and used z-score to find the outliers and remove them from our dataset to improve the model.



## **Pre - Processing Pipeline:**

In this section, we undertake data pre-processing steps to prepare the datasets for Machine Learning algorithm implementation. We can do encoding, feature scaling, split the train test data, handle the skewness and then go for the machine learning algorithms.

1. Encoding

Machine Learning algorithms can handle only the numerical values as their predictor variables. Hence **Label Encoding** becomes necessary as they encode categorical labels with numerical values. To avoid introducing feature importance for categorical features with large numbers of unique values, we can use both **Label Encoding** if they have more than 2 types of variables and **One-Hot Encoding** for less than or equal to 2 types as shown below.

Here we have converted the following columns into numerical values: -

Education 1 'Below College' 2 'College' 3 'Bachelor' 4 'Master' 5 'Doctor'

Environment Satisfaction 1 'Low' 2 'Medium' 3 'High' 4 'Very High'

Job Involvement 1 'Low' 2 'Medium' 3 'High' 4 'Very High'

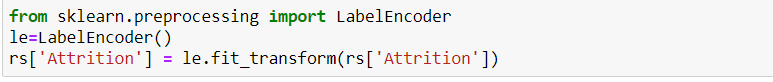
Job Satisfaction 1 'Low' 2 'Medium' 3 'High' 4 'Very High'

Performance Rating 1 'Low' 2 'Good' 3 'Excellent' 4 'Outstanding'

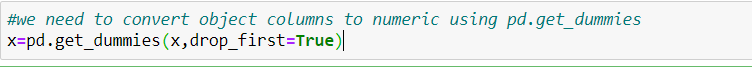
Relationship Satisfaction 1 'Low' 2 'Medium' 3 'High' 4 'Very High'

Work Life Balance 1 'Bad' 2 'Good' 3 'Better' 4 'Best'

Attrition 1-yes, 2-no



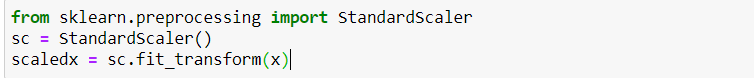
We can use get. Dummies() to convert the categorical data into numerical type.



1. Feature Scaling

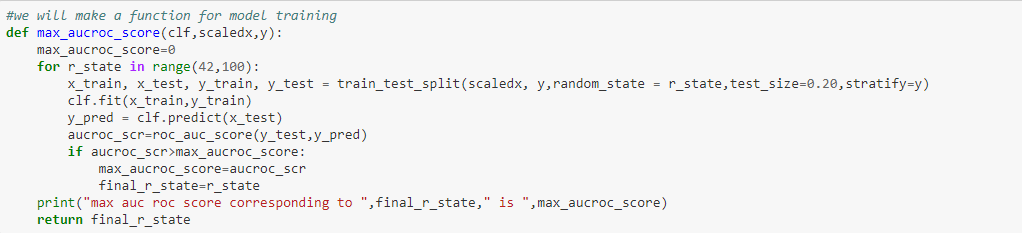
We can do feature scaling with standard Scaler or min max scaler which will standardise the input features. With help of Standrad scaler we will convert the attributes from the range 0 to 1 and in min max scaler we can convert the attributes from the range 0 to n.

Here we have used the standard Scaler which will convert the data from 0 to 1.



1. Train Test Data split

Prior to implementation or applying any Machine Learning algorithms, we must break the data into training and testing data frame from our original dataset.we have split the data into 80:20 ratio.



## ***Machine learning Algorithms:***

There are many machine learning algorithms available in the scikit leran library. We have used four types of algorithms i.e Logistic Regression, Decision tree Algorithm, KNN, SVC. As our target column is imbalanced so we have used roc auc score as our metrics.

We have also used metrics like accuracy score, roc auc score, confusion matrix, classification report from the sklearn.metrics .

First, we have a defined a function so it is simplified for calling the different algorithms by just calling the name of the function and passing the x and y in it.

1. Logistic Regression:



We have got max roc auc score for logistic regression: 78.37

Mean cross val score for logistic classifier: 0.8425184815957902

standard deviation in cross val score for logistic classifier: 0.018609160244716556

1. KNN



We have got max roc auc score for KNN: 55.31

Mean cross val score for KNN classifier: 75.26

Standard deviation in cross val score for KNN: 0.02204465330537351

1. DTC:



We have got max roc auc score for DTC is: 69.35

Mean cross val score of DTC is: 63.03

Standard deviation in cross val score for DTC is: 0.015555

1. SVC



We have got max roc auc score for SVC is : 80.05

Mean cross val score of SVC is: 76.03

Standard deviation in cross val score for SVC is: 0.07615555

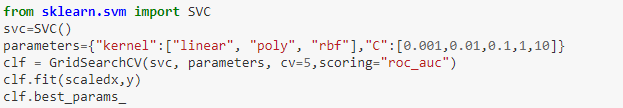
## ***Conclusion:***

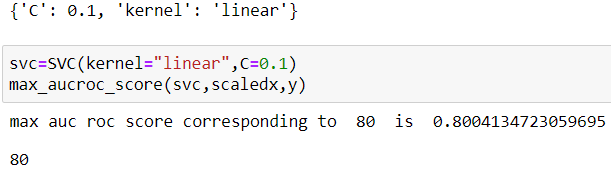
From the above algorithms we have chosen the SVC classifier as our best model. Before

saving it we have done the hyper parameter tuning of SVC to check the overfitting and

underfitting of the data. Then we got these as the classifier best parameter {'C': 0.1,

'kernel': 'linear'} which has helped in improving the accuracy score of the model.



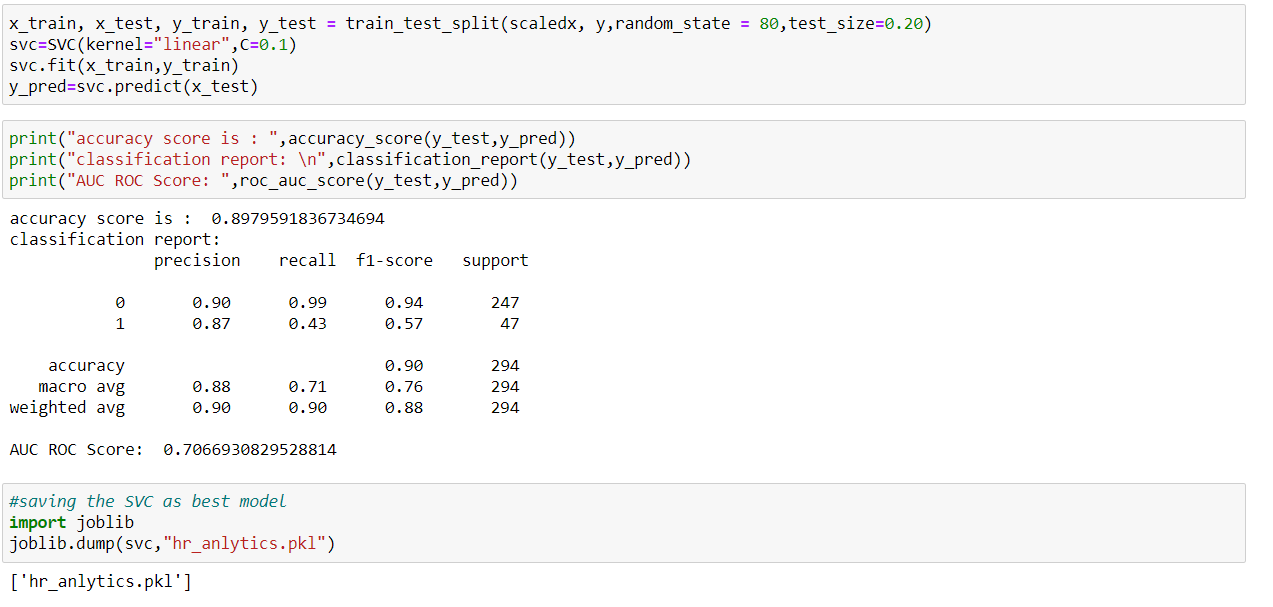


So we

So, we have chosen best random state as 80 and test size as 0.2, kernel = “linear”, c =0.1 and

passed in SVC classifier and found out the best accuracy score, classification report and auc

roc score and saved the model.



The max roc auc score for SVC we got is 70.66 and **Accuracy score** of the SVC model is **89.79%**. then we have saved the model using joblib as “hr\_analytics.pkl” as pickle file. We can

load the pickle file and predict the future data with the help of SVC algorithm whenever we

require it in the future.