HR Analytics

Background and Problem statement

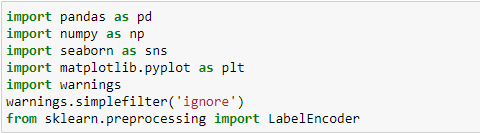
In the initial times, Human Resource department was sidelined and not given enough weightage. In due time, every company has realized its importance and had incorporated in the core group.

Human Resource Department is the most crucial department as it deals with employees and employees are the one who contribute to the health of the organization.

In this project, we have tried to evaluate which factors could trigger attrition and as attrition is directly linked to the profit and accuracy of the organization.

We have taken each column and tried to analyze the correlation to our target factor – Attrition.

Data Analysis



We have imported various libraries to analyze the data set.

Libraries imported are NumPy, Pandas for data analysis. Seaborn and Matplotlib for visualization and warnings for the machine to ignore the version change.

Label Encoder to convert string data to numeric data.

Columns taken into account are: -

Age 1470 non-null int64

1 Attrition 1470 non-null object

2 BusinessTravel 1470 non-null object

3 DailyRate 1470 non-null int64

4 Department 1470 non-null object

5 DistanceFromHome 1470 non-null int64

6 Education 1470 non-null int64

7 EducationField 1470 non-null object

8 EmployeeCount 1470 non-null int64

9 EmployeeNumber 1470 non-null int64

10 EnvironmentSatisfaction 1470 non-null int64

11 Gender 1470 non-null object

12 HourlyRate 1470 non-null int64

13 JobInvolvement 1470 non-null int64

14 JobLevel 1470 non-null int64

15 JobRole 1470 non-null object

16 JobSatisfaction 1470 non-null int64

17 MaritalStatus 1470 non-null object

18 MonthlyIncome 1470 non-null int64

19 MonthlyRate 1470 non-null int64

20 NumCompaniesWorked 1470 non-null int64

21 Over18 1470 non-null object

22 OverTime 1470 non-null object

23 PercentSalaryHike 1470 non-null int64

24 PerformanceRating 1470 non-null int64

25 RelationshipSatisfaction 1470 non-null int64

26 StandardHours 1470 non-null int64

27 StockOptionLevel 1470 non-null int64

28 TotalWorkingYears 1470 non-null int64

29 TrainingTimesLastYear 1470 non-null int64

30 WorkLifeBalance 1470 non-null int64

31 YearsAtCompany 1470 non-null int64

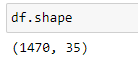
32 YearsInCurrentRole 1470 non-null int64

33 YearsSinceLastPromotion 1470 non-null int64

34 YearsWithCurrManager 1470 non-null int64

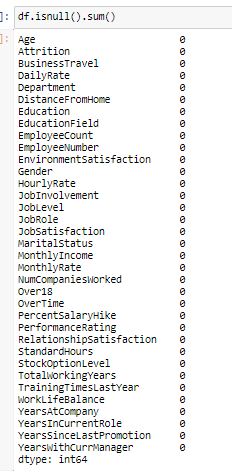
Attrition,BusinessTravel,Department, EducationField , Gender,JobRole,MaritalStatus,Over18,OverTime are object type of data, rest are in integer.

Checking the shape of the dataset:

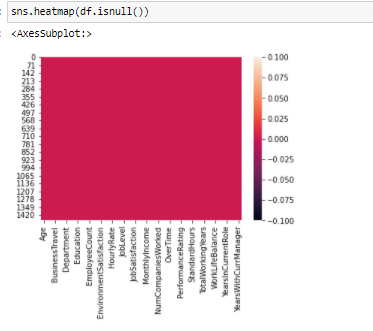


There are 1470 rows and 35 columns

Now checking if there are any null values or not and found the dataset to have no null values.

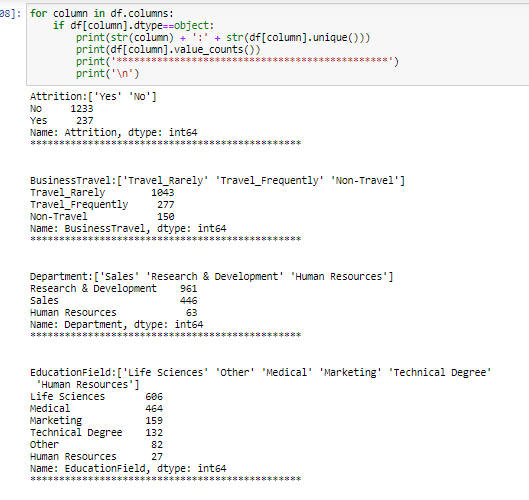


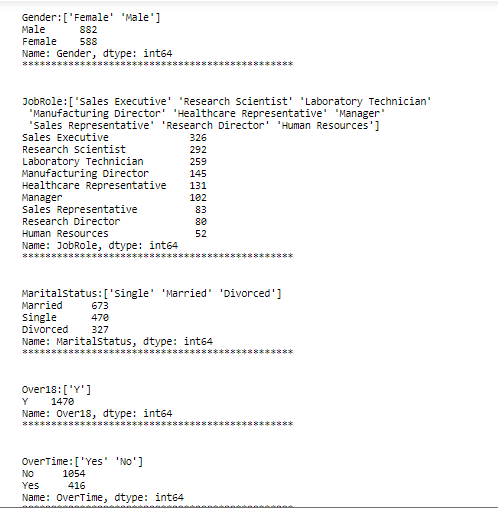
Represented by heatmap: -



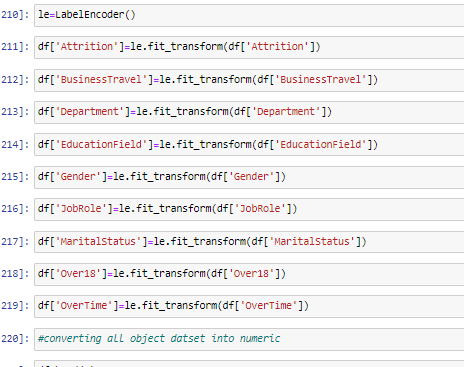
The dataset has not no null values, so uniform in colour.

Checking the attributes of object type of data along with its count: -

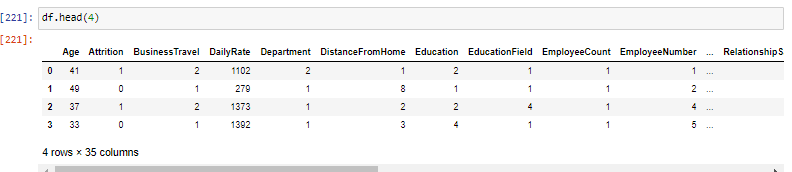




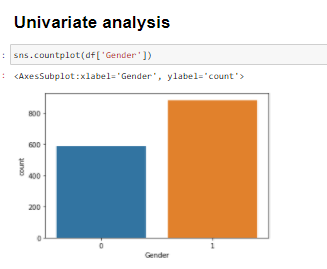
Now, converting all object type of data into numeric for better analysis through LabelEncoder.



Now checking the numeric dataset:-

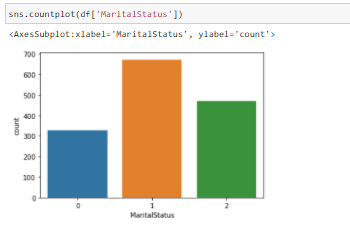


**Now implementing visualization for better clarity:**



There are 882 males and 588 females in the dataset.

Now checking the marital Status:



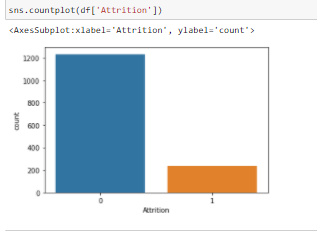
#MaritalStatus:['Single' 'Married' 'Divorced']

#Married 673 (1)

#Single 470 (2)

#Divorced 327 (0)

The Attrition rate: -

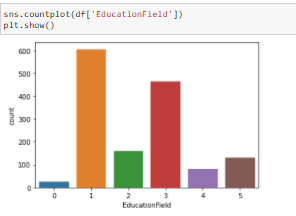


#Attrition:['Yes' 'No']

#No 1233 (0)

#Yes 237 (1)

The Educational Field:



#EducationField:['Life Sciences' 'Other' 'Medical' 'Marketing' 'Technical Degree'

# 'Human Resources']

#Life Sciences 606 (1)

#Medical 464 (3)

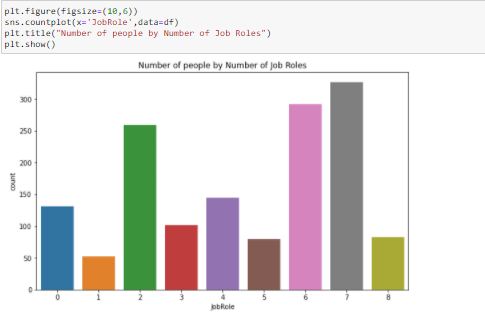
#Marketing 159 (2)

#Technical Degree 132 (5)

#Other 82 (4)

#Human Resources 27 (0)

Next exploring the job role:



#JobRole:['Sales Executive' 'Research Scientist' 'Laboratory Technician'

# 'Manufacturing Director' 'Healthcare Representative' 'Manager'

# 'Sales Representative' 'Research Director' 'Human Resources']

#Sales Executive 326 (7)

#Research Scientist 292 (6)

#Laboratory Technician 259 (2)

#Manufacturing Director 145 (4)

#Healthcare Representative 131 (0)

#Manager 102 (3)

#Sales Representative 83 (8)

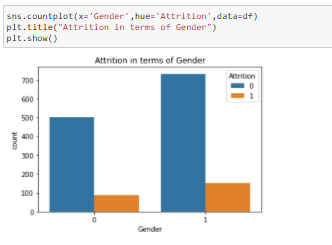
#Research Director 80 (5)

#Human Resources 52 (1)

**Bivariate Analysis**

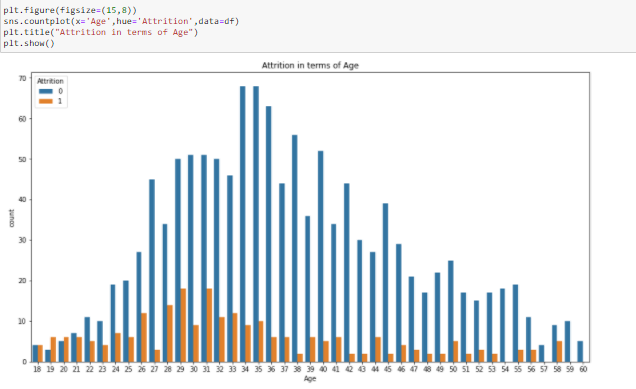
Now finding out the relation between to two columns for better understanding of the dataset

First, checking the gender to attrition ratio:



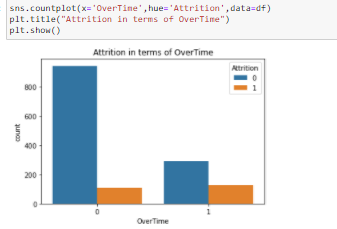
Male has less attrition rate compared to female

Now checking the Attrition rate in terms of age:



#Maximum attrition is early in the ages and with increase in age attrition rate reduces

Now checking the attrition rate over time.

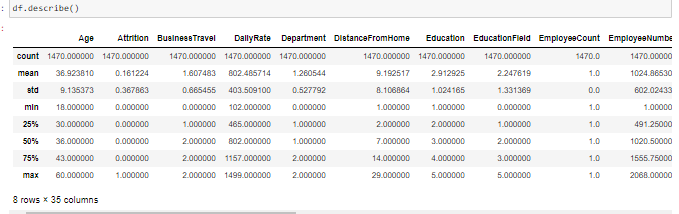


#OverTime:['Yes' 'No']

#No 1054 (0) - if no overtime, attrition rate(leaving=1) is less

#Yes 416 (1) - with overtime, attrition rate is more

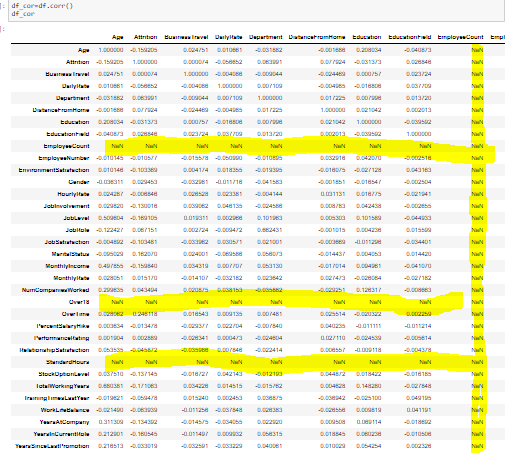
Now checking the variance of the dataset for better clarity through describe.



#There is not much difference between mean and mode

#However, there is difference in range values

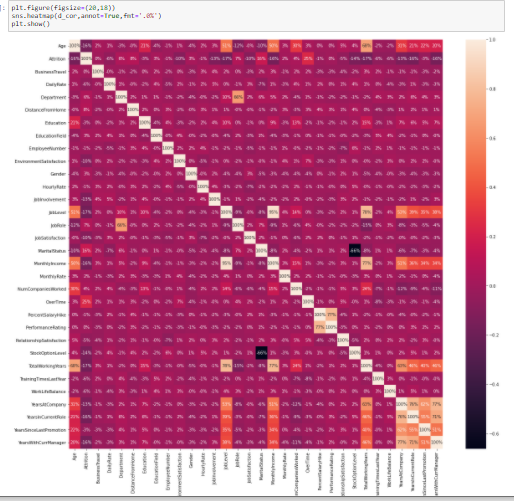
Next, we are checking the correlation between each columns: -



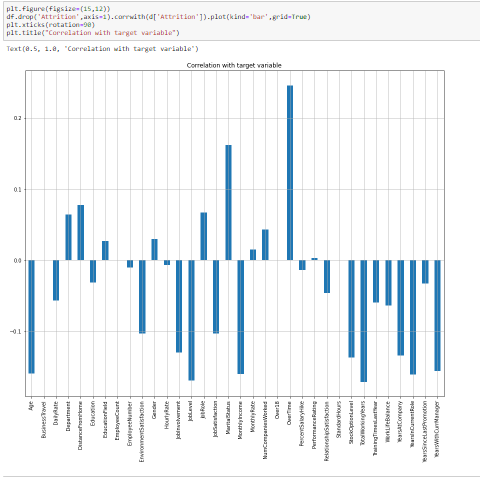
#EmployeeCount,Over18,StandardHours has NAN values as the values are same, does not vary so there is no correlation between them and dropping the columns:



Checking the correlation through heatmap:



For better understanding with the target column, we are visually representing through bar plot.



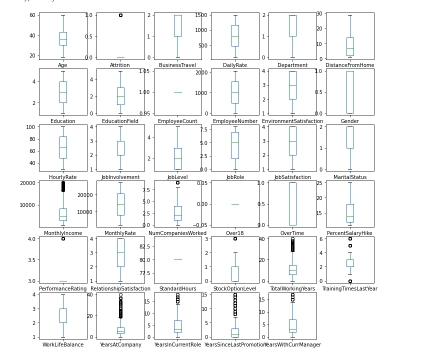
# Attrition is directly linked to the following columns, if these attributes increases, attrition rate increases

#Department,DistanceFromHome,EducationField,Gender,Job Role,Marital Status,Montly rate,NumCompaniesWorked,Overtime,Performance Rating

#Among all - Overtime is the highest

Now, checking the outliers of each column:-

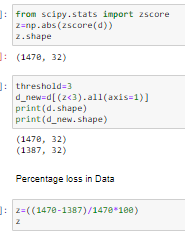




#Attrition, Monthly Income,NumCompaniesWorked,PerformanceRating,StockOptionLevel,TotalWorkingYears,TrainingTimesLastYear,YearsAtCompany

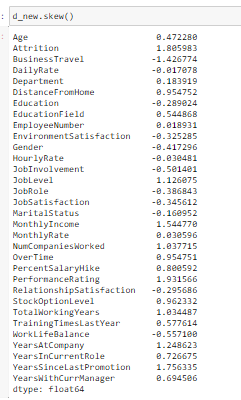
#YearsinCurrentRole,YearsSinceLastPromotion & YearsWithCurrManager has got outliers.

Now removing the outliers and checking the percentage of loss:-



**The percentage loss is** 5.64, so we can go ahead with the new improved data.

Now, Checking the skewness of the dataset:-

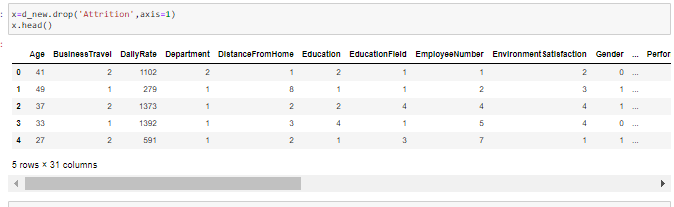


#0.5 being the standard skew value

# the columns-Attriton, Job level,MonthlyIncome,NumCompaniesWorked,PerformanceRating ,TotalWorkingYears,YearsAtCompany YearsSinceLastPromotion have skewed dataset

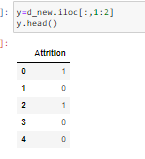
#splitting the dataset for training and testing.

Now, splitting the dataset into x and y and y being the target column(attrition) to build the machine learning model:



Splitted x using drop method.

Dropped the attrition column and in y we have taken attrition as target.



Now checking the shape of x and y.



There are 1387 rows and 31 columns for x and for y 1387 rows and 1 column which is our attrition column.



Now removing the skewness of the featured column as we have seen above the data is skewed.

Then scaling the dataset: -



Now importing various classification algorithms and logistic regression to check the accuracy:

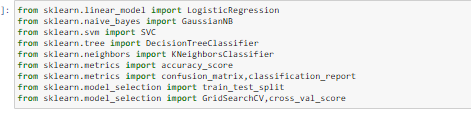
Algorithms used:-

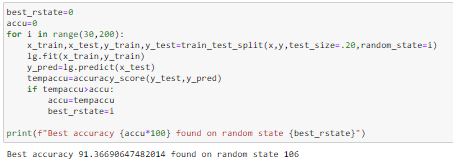
LogisticRegression, GaussianNB, SVC, DecisionTreeClassifier, KNeighborsClassifier.

For scores we have taken:-

accuracy\_score, confusion\_matrix,classification\_report, cross\_val\_score.

We have imported from sklearn.model\_selection to split and train, test the data set and Grid SearchCV to hypertune the model.



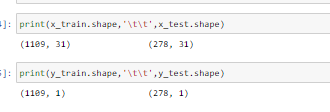


Checking the best random date through Logistic regression and found the state to be 106 with score as 91 percent.

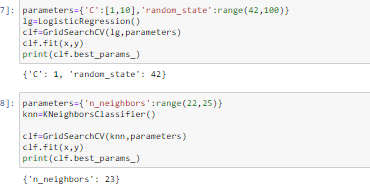
With the best random state, feeding the dataset:-



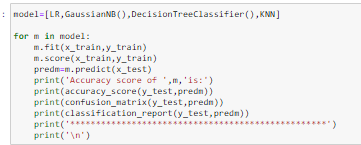
Now, checking the shape of x\_train, x\_test,y\_train and y\_test.



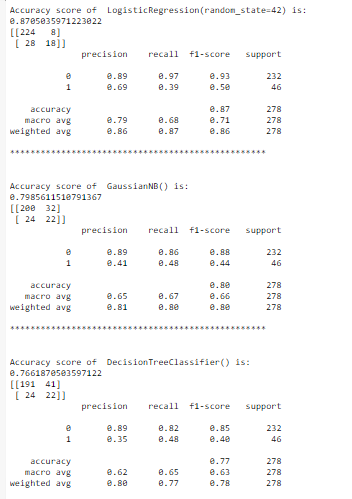
Now finding the best parameter through Grid Search CV

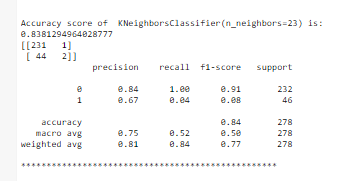


Now, implementing the parameters for different model training to check the score and which model is giving the best score: -



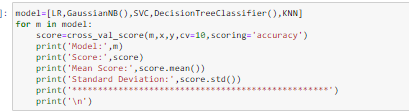
We have taken Logistic Regression,GaussianNB, DecisionTreeClassifier and KNN.

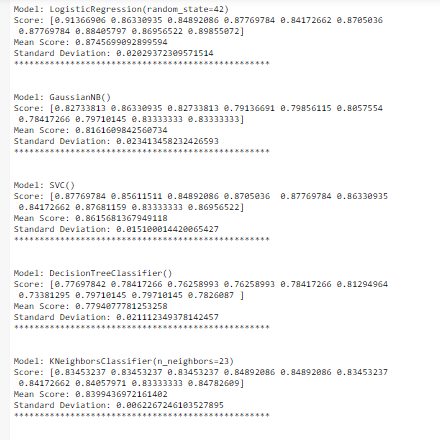




LogisticRegression has the best accuracy score of 87 percent.

Now cross validating the score to check if the model is over-fitted or under fitted.





#Above all LogisticRegression is giving the best score.

# Now using Using Ensemble Technique to boostup the score

# We are importing various libraries to check the score;

# RandomForestClassifier

# AdaBoostClassifier

# GradientBoostingClassifier

# BaggingClassifier

# ExtraTreesClassifier

# 

# 

# 

# RandomForestClassifier() gives the best value of 84 percent.

# Now checking the auc-roc score to to get the relation between True Positive Rate and False Positive Rate

# 

# The score is 61 percent.

# Ultimately saving the model through pickle.

# 