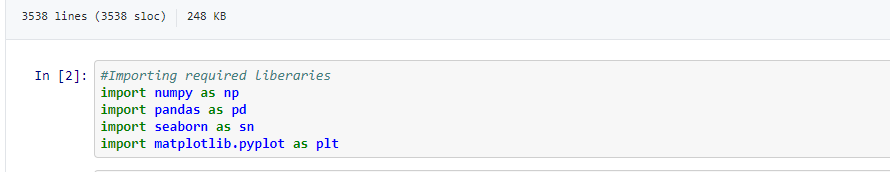
**In this blog-post, I will go through the whole process of creating a machine learning model on the HR Analytic. We will find the attrition for HR. Attrition refers to gradual loss of employee for HR.To build out model we will use many attirubuted such as Department,Age,Distance from home,Education ,education fiesl etc .**

**HR Analytics:**

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.

There are 1470 entries and 35 columns using which we will try to find the attrition for HR.

**Importing the Required Libraries :**



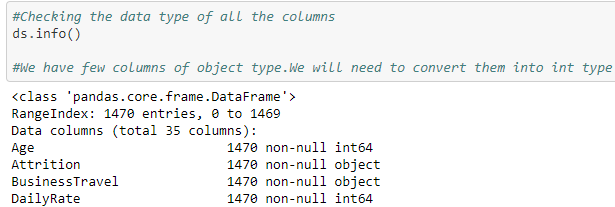
**Getting the Data:**

Loading our dataset from local



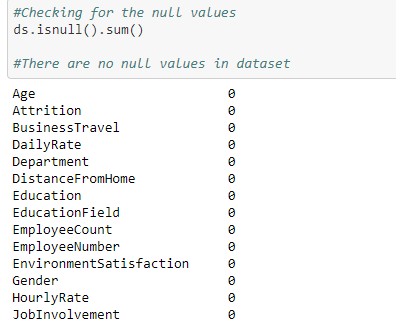
**Data Exploration/Analysis:**

We will observe the data type of each columns and number rows,columns and column names



**Removing the null values :**

We need to check for the null values present in our dataset and If there are any null values we need to process(replace with mean,median or mode) them.



There are no null values in the dataset

**Checking the correlations :**

Using heatmap we can find the correlation between columns.

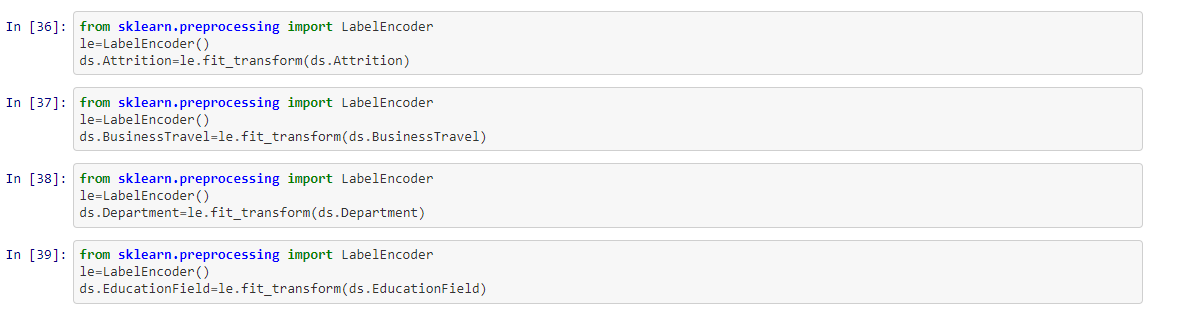
Highest correlated columns are education,department,daily rate,job level and job Role.

Wherease least correlated columns are Employee number and distance from home.



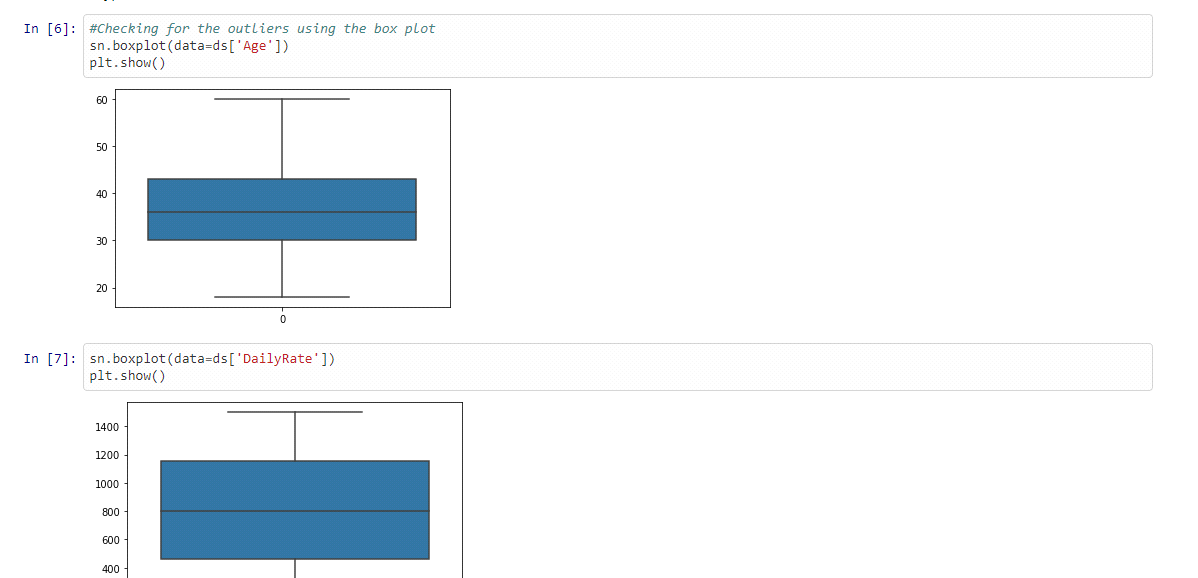
**Converting to integer:**

There are columns like Attrition,BusinessTravel,Department and EducationField which are Char data type.We will convert all char datatype columns to int using the LabelEncoder



**Checking the outlier :**

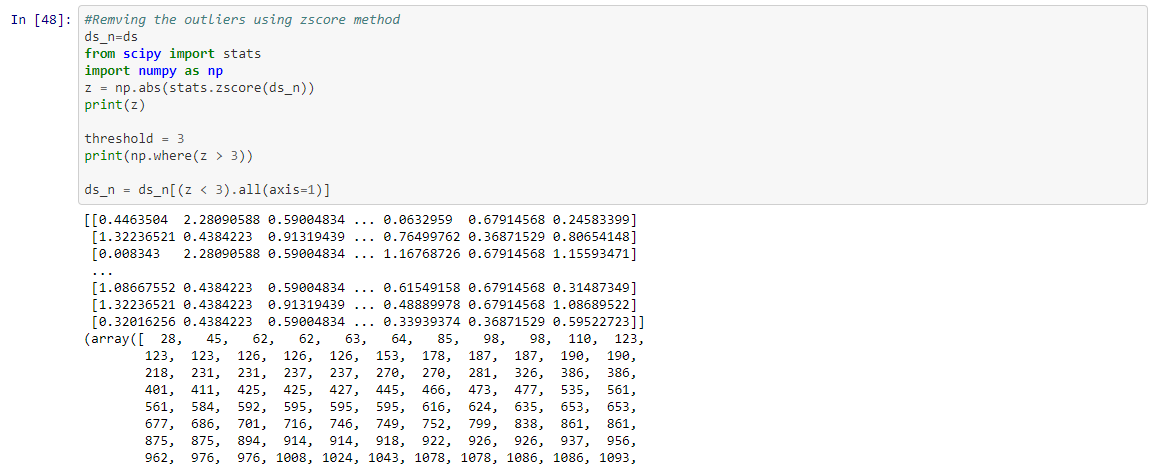
There are outliers in Columns: Monthincome,NumCompaniesWorked,PerformanceRating,StockOptionLevel,TotalWorkingYears,TrainingTimesLastYear,YearsAtCompany and YearsInCurrentRole.





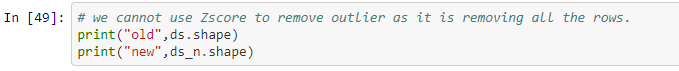
**Removing the Outlier :**

We will remove the outliers using the zscore method.



**Checking the deleted rows:**

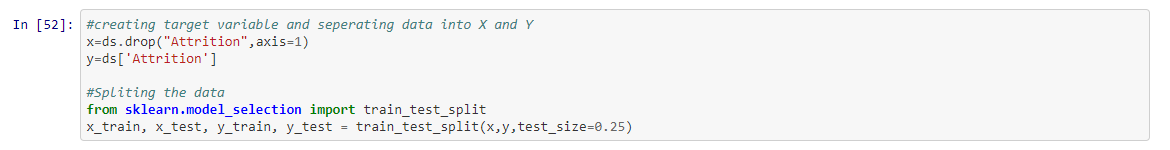
Zscore removes the outliers hence we need to check how many rows are deleted.If deleted rows are too many then we should not remove remove outliers.



Creating X and Y variables :

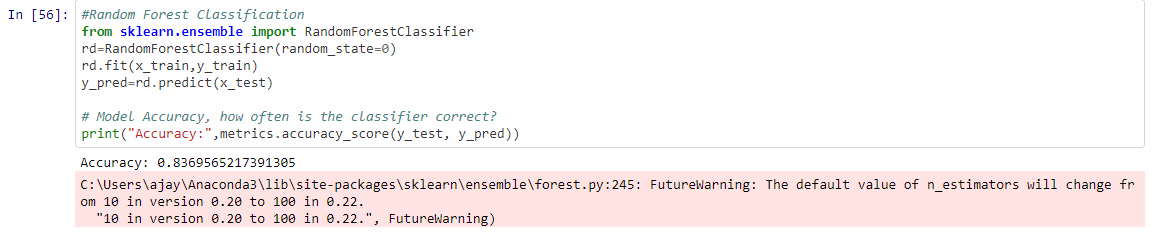
We have preprocessed out dataset.

We are spliting the datatset in x and y axis.



**Training Model:**

Randome Forest



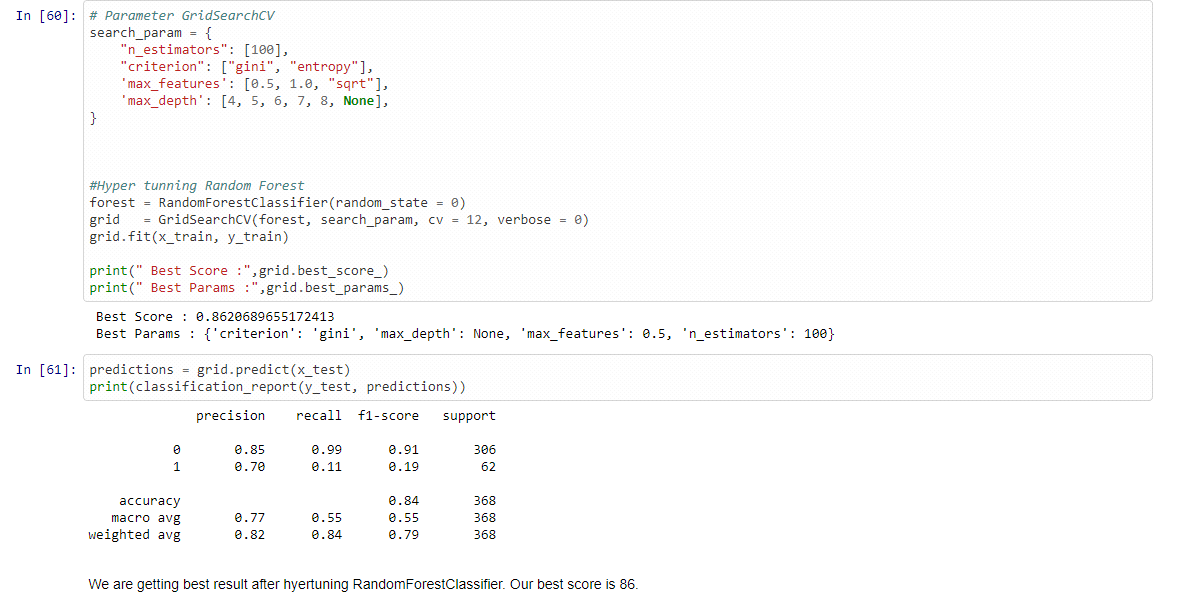
We have used Logistic Regression,SVC and Random Forest to train our data.

The best score 83.69 was given by Randome Forest Regression.

Randome Forest is performing well on our dataset hence now we can start tunig the hyperparmeters for Randome Forest

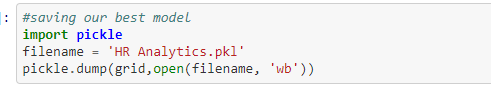
**HyperParameter Tuning :**

Below you can see the code of the hyperparamter tuning for the parameters criterion



Our accuracy has increased to 85% after hypertuning the parameters.

**Saving Mode :**



**Summary :**

We started with loading our dataset and the preprocessing dataset. We checked for null values.There were no null values observed.We found that there are many columns are of Char datatype and we need to convert them to int type before passing the our model. We converted all char type columns to Int type using LabelEncoder.

We also found the correlation using the heatmap .Highest correlated columns were education,department,daily rate,job level and job Role and least correlated columns are Employee number and distance from home.

There were outliers in columns Monthincome,NumCompaniesWorked,PerformanceRating,StockOptionLevel,TotalWorkingYears,TrainingTimesLastYear,YearsAtCompany and YearsInCurrentRole. We removed outliers using the zscore.After completing all the preprocessing we split out data in X and Y axis and passed to our Model.

We used 3 Models Logistic Regression.SVC and Random Forest regression.We found best result with Random forest which was 83%. Randome forest was best suited hence Hypertuned its parameter. After hypertuning the Random Forest accuracy was increased to 83%.