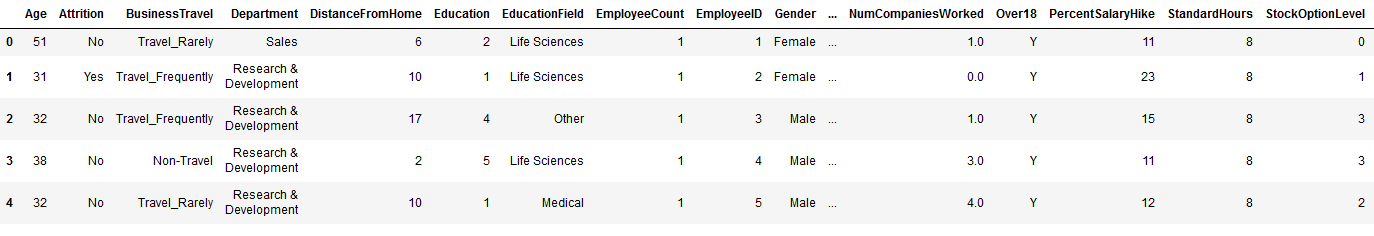
**Employee Attrition Predictor**

**Attrition in a general term means gradual but a significant decrease in the numbers. So when clubbed with employees it basically indicates the number of staff who have left the organization over a particular period of time.**

**Now this attrition is dependent on various parameters within the organization. These parameters can then be analyzed by the HR team to find out what causes this decrease in the work force and what measures could be taken over the course of time to decrease the number of employee attrition.**



**Columns in Dataset:**

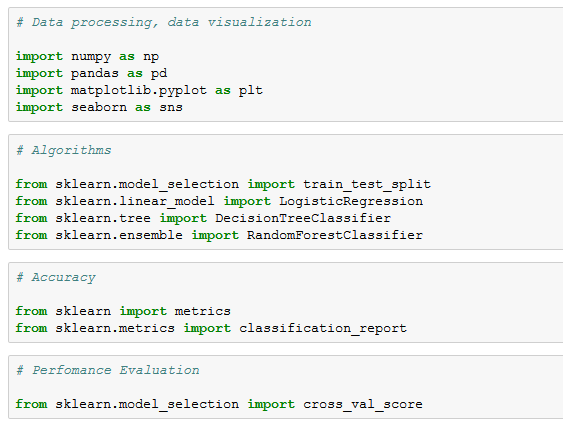
Feature Columns / Input Variables:

|  |  |
| --- | --- |
| Age | Current age of the employee |
| BusinessTravel | Whether or not the employee had to travel to some place for some business related work |
| Department | The division in which the employee is working |
| DistanceFromHome | Distance in kms from his home to the workplace |
| Education | The educational qualification of the employee |
| EducationField | Educational domain of the employee |
| EmployeeCount | The number of employees working in that sector |
| EmployeeID | The ID number of the employee |
| Gender | Gender of the employee |
| NumCompaniesWorked | The total number of companies the employee has worked in |
| Over18 | Whether or not the Employee is over 18 years of age |
| PercentSalaryHike | Salary hike of the employee per year in % value |
| TotalWorkingYears | The overall work experience of the employee |
| TrainingTimesLastYear | The number of times employee had to undergo a training |
| YearsAtCompany | Working years in the current company |
| YearsSinceLastPromotion | How long since the last promotion |
| YearsWithCurrentManager | How long since the employee is working under the current manager |

Target Column / Output Variable:

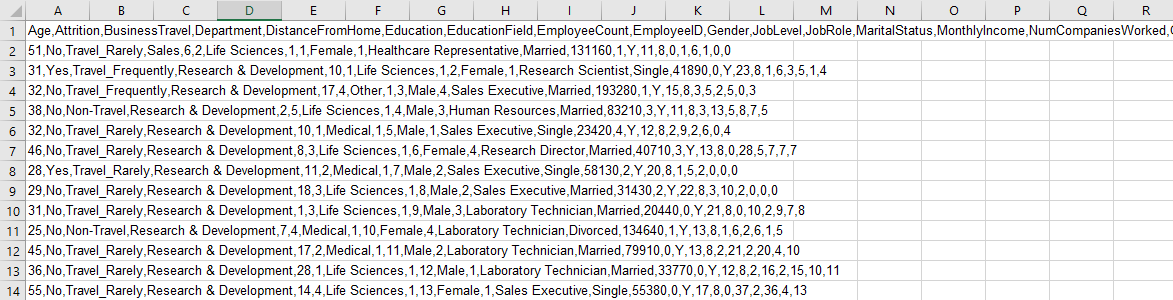
|  |  |
| --- | --- |
| Attrition | Whether or not the employee is going to leave the company |

**Importing Python Libraries:**



**Data Wrangling:**

The Employees Data set was not in the correct tabular format and hence we had to do data wrangling on that dataset



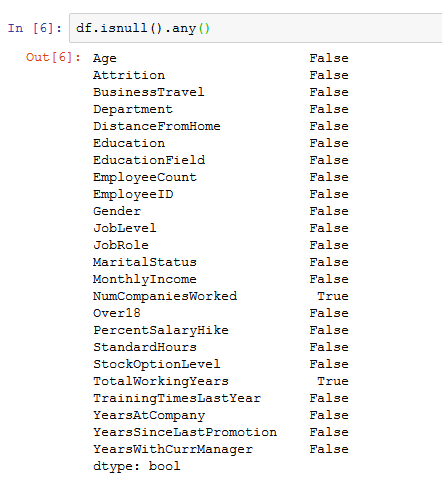
Code Snippet:



**Data Cleaning:**

To check if any null value is present in the Dataset

Code Snippet:



Since there are columns with null values we need to replace the null values.

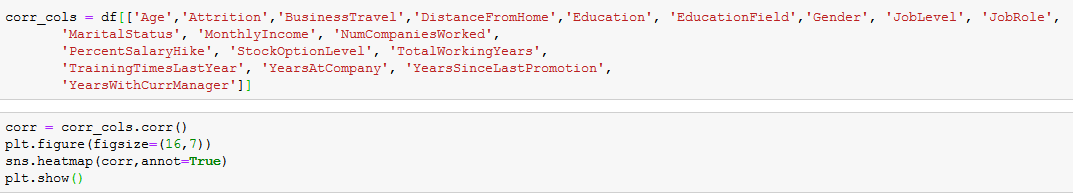
Code Snippet:

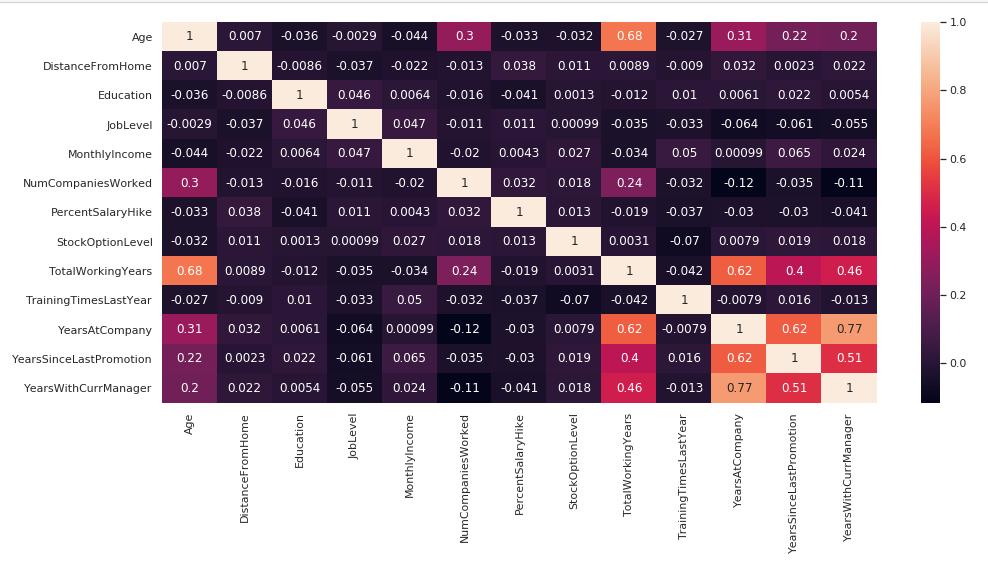


**Data Visualization:**

First we find the co-relation between all the columns

Code Snippet:

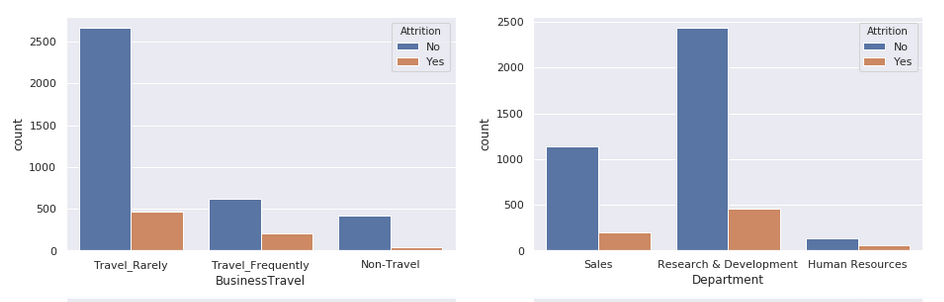


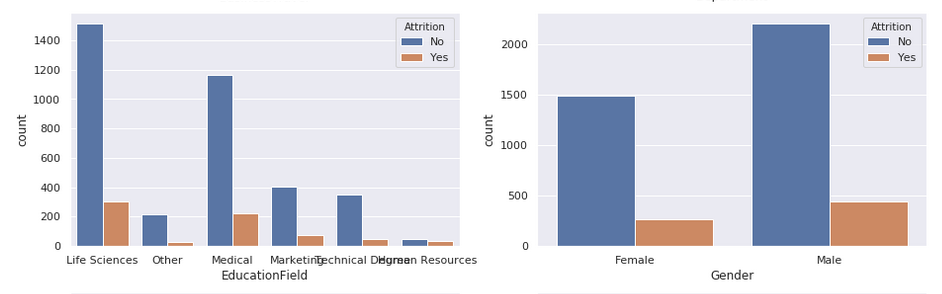


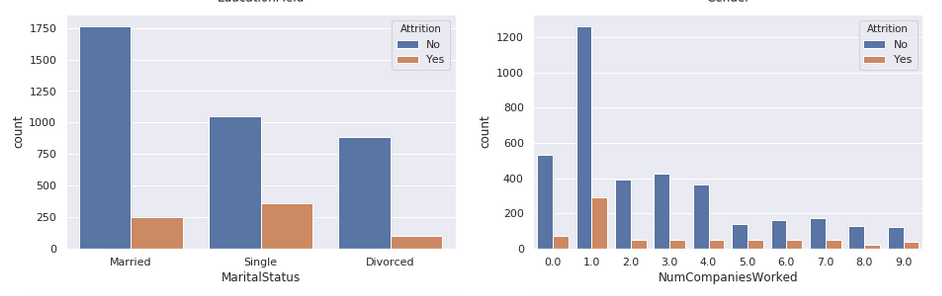
Then we check the relation of the target column (‘Attrition’) with rest of the feature columns:

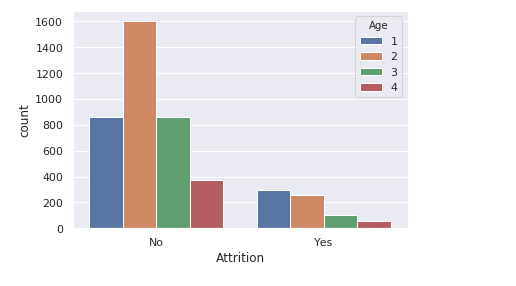
Code Snippet:











## Observation:

#### 1. Employees those travel very frequently are most likely to leave.

#### 2. A lot of people from R&D department has left there can be some issue in this department, which need to be analyzed.

#### 3. People from HR department also leaves the company more frequently.

#### 4. Single Employees has a high tendency to leave.

#### 5. Employees for whom this is second company are most likely to leave.

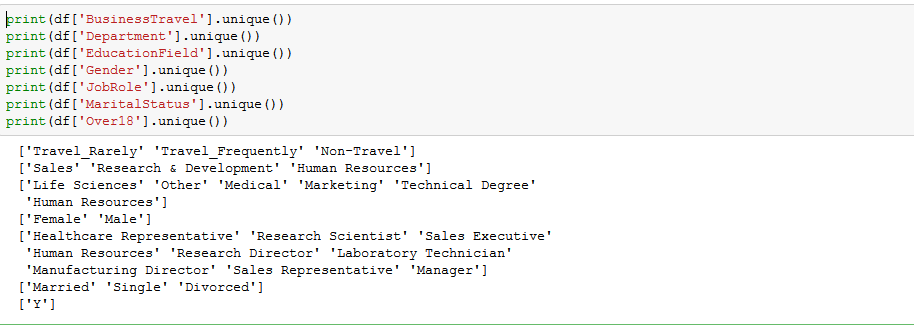
#### 6. If an employee promotion is pending from almost 6-7 years, he is likely to leave.

#### 7. People who are less than 30 years of age have the most tendency to leave the company.

**Converting all Categorical Data into Numeric Data**

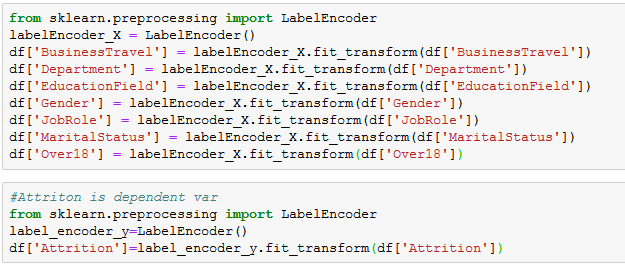
For this we are using Label Encoder where we first find out the unique values in each of the columns

Code Snippet:

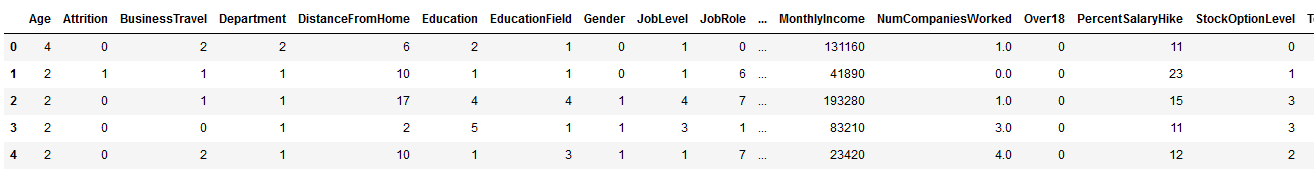


Then we preprocess the data using LabelEncoder where a unique numeric value is assigned to each of those attribute in a column thus converting it to a numeric scheme

Code Snippet:



Now post applying the LabelEncoder we have all the data in numeric format



**Data Modelling:**

Basic steps that needs to be followed while prediction and finding accuracy of a model:

1. Scaling of data 🡪 This includes fetching data from dataset and correction of data in proper readable format.

2. Separate X and Y 🡪 Separate feature and target columns as X and Y.

3. Split train-test 🡪 Splitting of data into train and test data set.

4. Importing libraries 🡪 Import all libraries that are required in algorithm.

5. Modeling data (use algorithm) 🡪 It includes the application of algorithm on the data.

6. Fitting the data 🡪 Use to fit the train data frequently and quickly.

7. Predicting data 🡪 predicting the target data based on the feature test data.

8. Finding Accuracy of Algorithm 🡪 compare the target test data and predicted data and based on that calculating the accuracy.

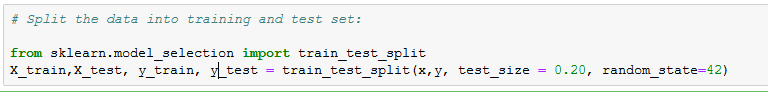
**Building Machine Learning Models using Algorithms:**

## 1. Logistic Regression:

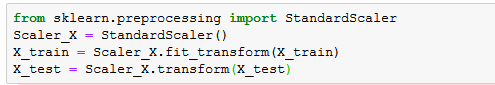
Defining X and Y coordinates where X being the feature columns and Y being the target column



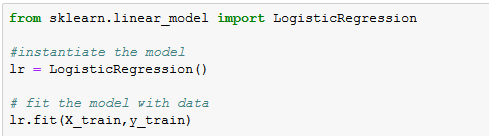
Splitting the data into train and test set



Preprocessing X data using Standard Scalar



Fitting the Data and applying the algorithm

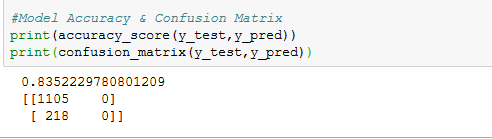


Predicting the response of the data based on X test set



array([0, 0, 0, ..., 0, 0, 0])

Check the model accuracy



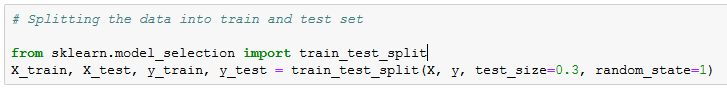
So the above model gives us an accuracy of 0.8352229780801209

## 2. Decision Tree:

Defining X and Y coordinates where X being the feature columns and Y being the target column

## 

Splitting the data into train and test set



## 

Fitting the Data and applying the algorithm

## 

Predicting the response of the data based on X test set

## 

Check the model accuracy

## 

So the above model gives us an accuracy of 0.9750566893424036

## 3. Random Forest:

Defining X and Y coordinates where X being the feature columns and Y being the target column

## 

Splitting the data into train and test set

## 

Fitting the Data and applying the algorithm

## 

Predicting the response of the data based on X test set

## 

Check the model accuracy

## 

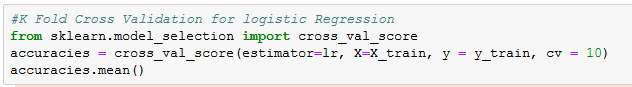
So the above model gives us an accuracy of 0.9841269841269841

# **Evaluating the Model Performance:**

Here we user the K-Fold validation to check the performance of each model

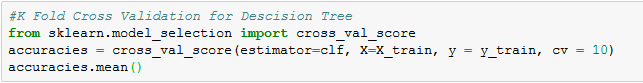
The main reason to use K-Fold Cross Validation is to see how the model will perform to an independent dataset.

1. **K-Fold Cross Validation for Logistic Regression:**



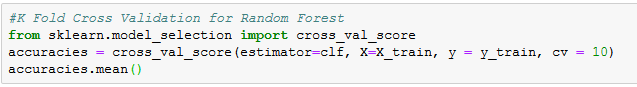
O/P: 0.8412694910267726

1. **K-Fold Cross Validation for Decision Tree:**



O/P: 0.9770026898667675

1. **K-Fold Cross Validation for Random Forest:**



O/P: 0.977649939057706

**Conclusion:**

Below mentioned are the algorithms that we used with their accuracy:

Logistic Regression: 0.8378684807256236

Decision Tree: 0.9773242630385488

Random Forest: 0.9879062736205594

### The performance evaluation for each of these models are:

#### Logistic Regression: 0.8386801588510895

#### Decision Tree: 0.9731160151529525

#### Random Forest: 0.9727923905574833

Here we could see that Random Forest gives the best estimation regarding the "Attribution" factor of an employee and based on the parameters contributing to it the required changes could be brought in to the existing company structure and parameters to reduce the attrition % of the employees working with them.​

Reference:

The data set for analysis has been taken from Kaggle

<https://www.kaggle.com/>

<https://www.kaggle.com/datasets?search=employee+attrition>