Machine Learning for predicting HR Attrition

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.

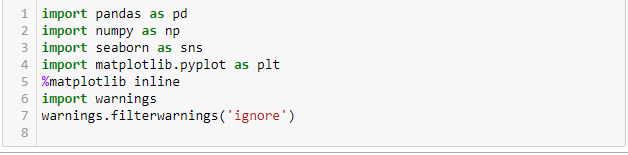
The key for success of any organization are its valuable employees and to retain these employees are the main function of HR of any organization.

In this case study we will build a machine learning model to predict the attrition.

We will proceed with these points in mind

1. Data cleaning, Loading and formatting
2. Checking duplicates in data ( Data integrity check)
3. Exploratory data analysis ( in which we compare data to find out some meaningful insight)
4. Feature Engineering in which we find out the best features which contribute maximum to label and decide outcomes of target variables. We can use PCA
5. Balancing Imbalanced Dataset
6. Machine Learning Model building
7. Dumping or saving model
8. **Data cleaning, Loading and formatting -**

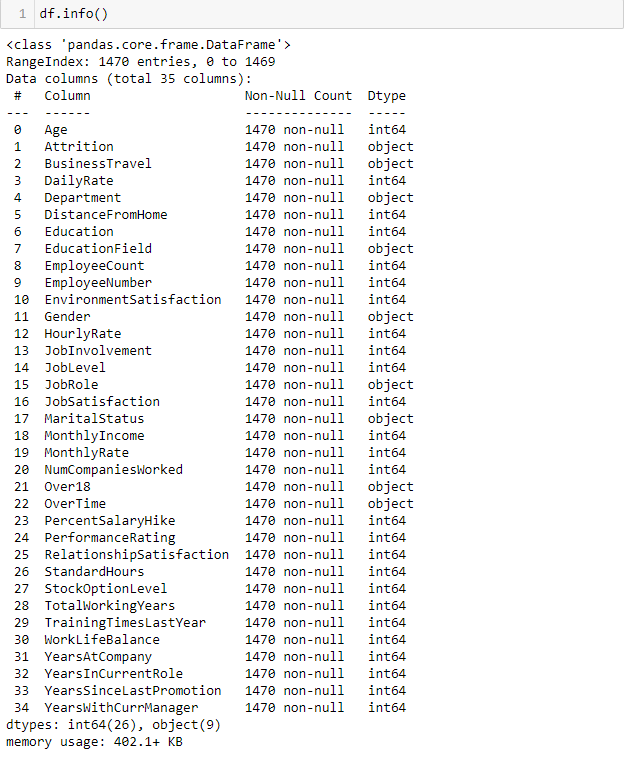
In this we will first import the necessary libraries-



Then we will load the dataset from locally saved file or raw file on Github



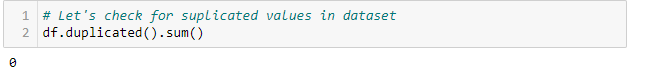
Checking Information regarding dataset columns and datatypes as well as null if present.



Notes -

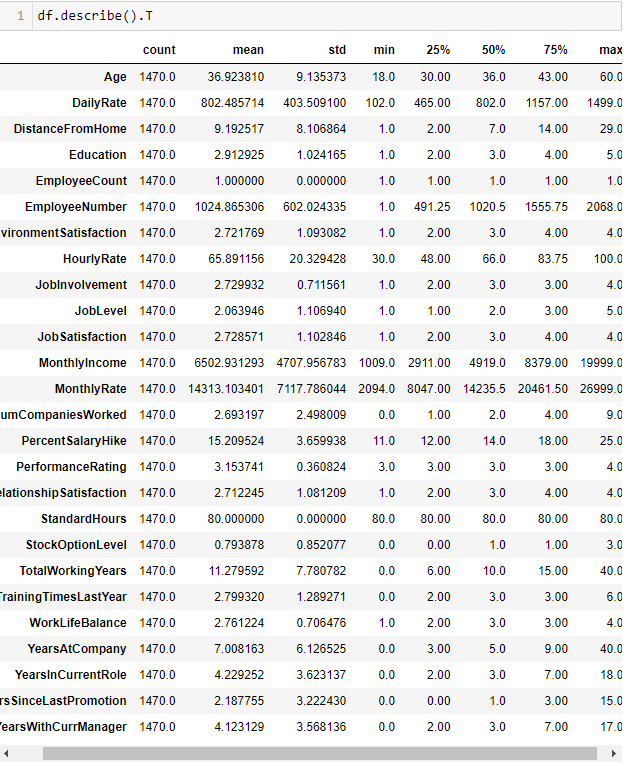
1. There are No Nulls present in dataset
2. There are various object data types present in the dataset.
3. There are total 35 features
4. Our target label is Attrition.

2. **Checking Duplicates in Data -**

****

There are no Duplicates present in data so we are good to go.

**Then we go for checking some statistical analysis by using .Describe() method**

****

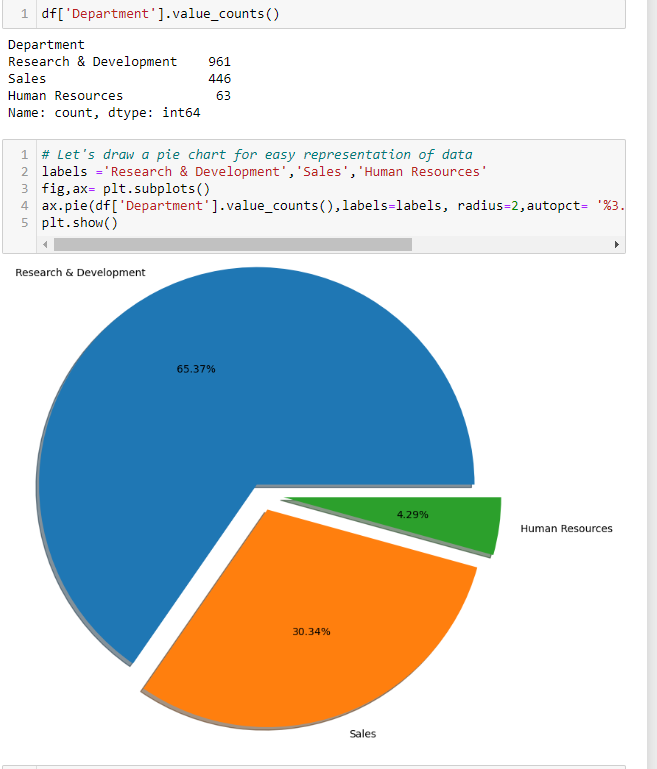
**Notes -**

* **Minimum age of employee is 18 and Maximum age is 60**
* **Average distance from home is 9.19**
* **Employee count and standard hours are having one unique value only**
* **Average hourly rate for employment is 65.89**
* **Average job satisfaction is 2.1 it means there are many employees who are not satisfied with their job**
* **Mean percent salary hike is 15**
* **Average performance rating is 3.0 which means performance of employees are good in organization.**
* **In some features the median is greater than mean which shows skewness is present in the dataset.**

**3. Exploratory Data analysis -**

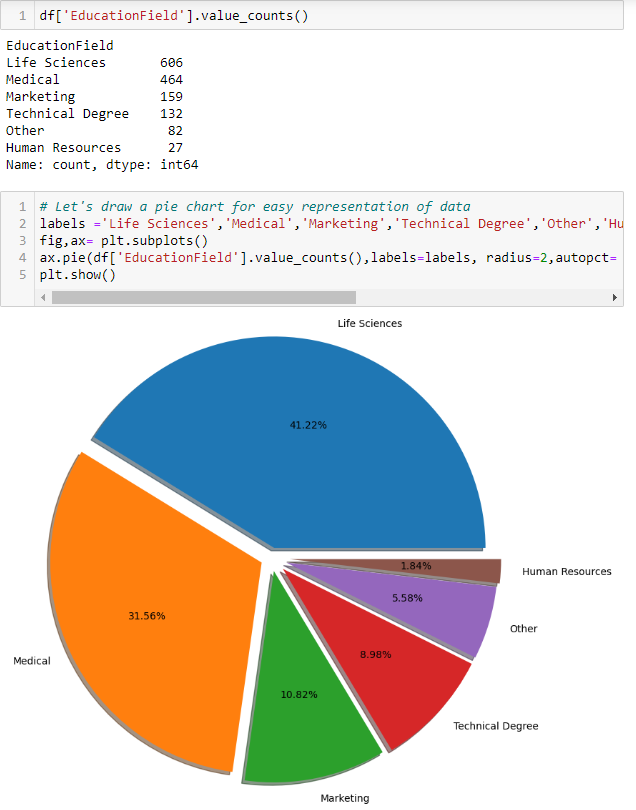
**Exploratory data analysis is done to find out the anomalies in the dataset.**

Let’s take the department column and use value.count()s method

****

We can see there are 3 departments: Research & Development , sales and human resource. In which Highest employees are in Research and Development and some less in Sales and Lastly few in HR.

Using same for Educationfield column:

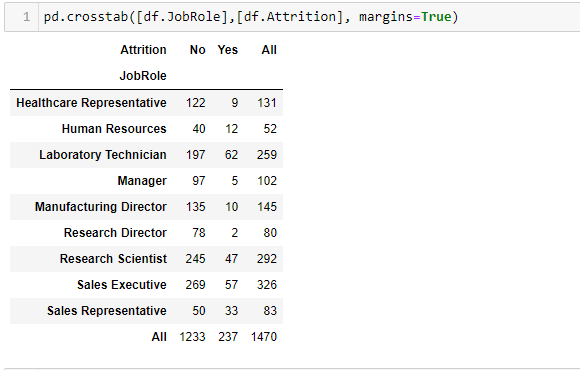


There are total 6 different Educationfield:-

1. Lifesciences
2. Medical
3. Marketing
4. Technical degree
5. Other
6. Human Resources

Most of the employees are from Life Sciences and Medical Background

Using Crosstab method we can see two columns comparison in one



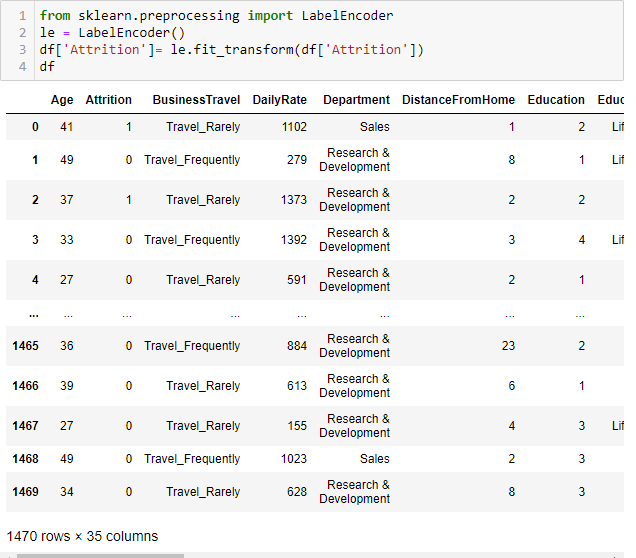
In this we can easily see the Attribution contributors from different Job roles

1. Maximum employees who left are Laboratory Technician and Sales Executive
2. Managers are likely to stick to the company
3. Maximum employees holding Sales Executive position may be that’s why employees are more in this job role as maximum attrition occurs in this.

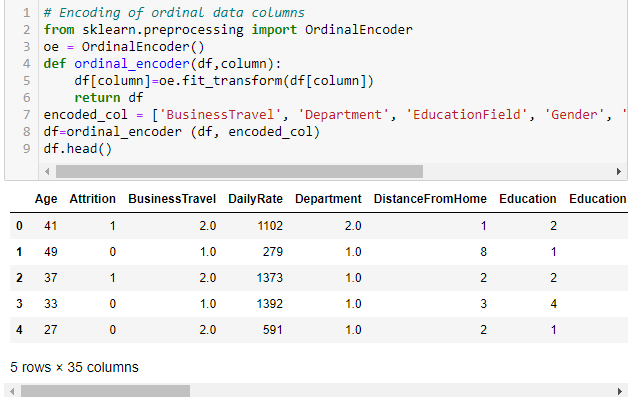
**Encoding Data -**

As most of the features are categorical or having an object datatype we need to convert it into Numeric data points as Machine learning models understand only Numeric data.

We will use Label encoder for our target feature and Ordinal encoder for Ordinal data.



For ordinal data we use Ordinal Encoder



* As our features are totally converted into numeric data points we will now separate Numeric data types and Plot a Boxplot to find out the outliers.

Numeric=['Age', 'DailyRate', 'DistanceFromHome',

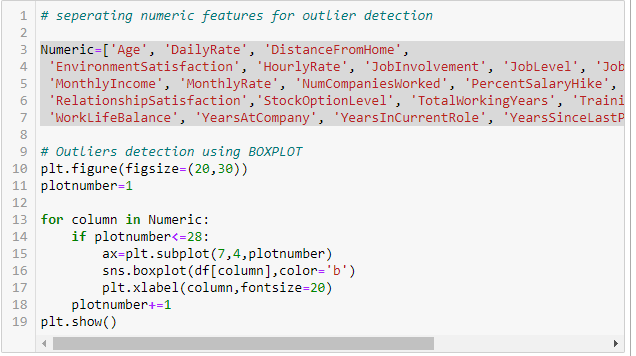
'EnvironmentSatisfaction', 'HourlyRate', 'JobInvolvement', 'JobLevel', 'JobSatisfaction',

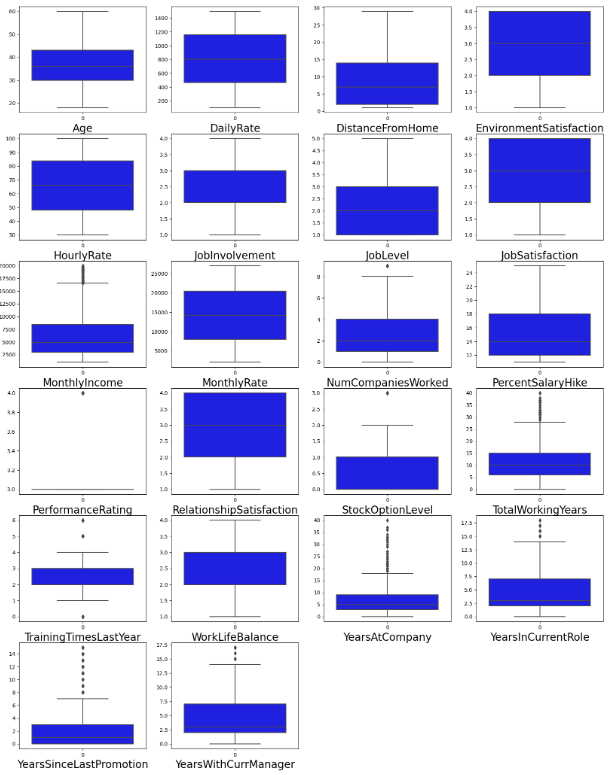
'MonthlyIncome', 'MonthlyRate', 'NumCompaniesWorked', 'PercentSalaryHike', 'PerformanceRating',

'RelationshipSatisfaction','StockOptionLevel', 'TotalWorkingYears', 'TrainingTimesLastYear',

'WorkLifeBalance', 'YearsAtCompany', 'YearsInCurrentRole', 'YearsSinceLastPromotion', 'YearsWithCurrManager']

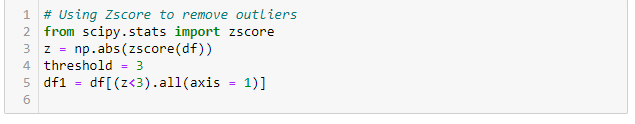
These are Numeric data types



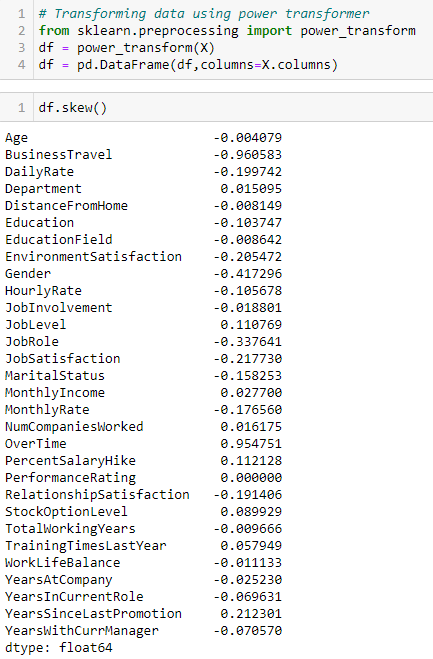


We can clearly see there some columns such as Monthly income, Training Times last year, Years at company Years in current role, Years since last promotion and Years with curr manager has outliers.

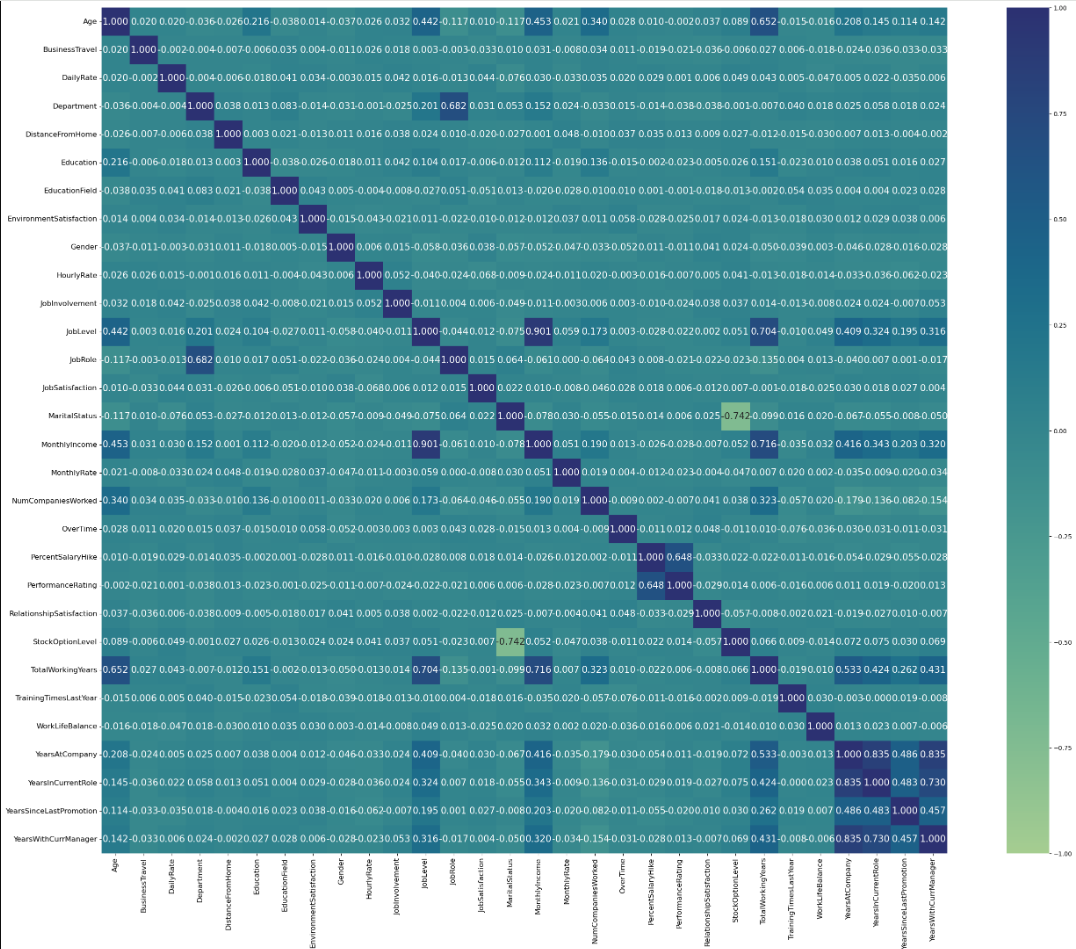
We can remove outliers by using zscore stats



Then we will check for skewness in data by using .skew() method and using power transform we will transform the data to remove the skewness



* Checking multicollinearity and correlation between features and removing irrelevant features by plotting Heat map. Correlation heatmap shows in a glance which variables are correlated, to what degree, in which direction, and alerts us to potential multicollinearity problems.

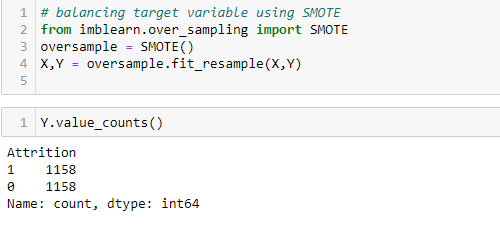


Comments -

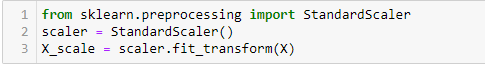
1. Job level is highly correlated with monthly income and there may be multicollinearity
2. Percent salary hike is highly correlated with performance rating.

**5. Handling Imbalanced Dataset -**

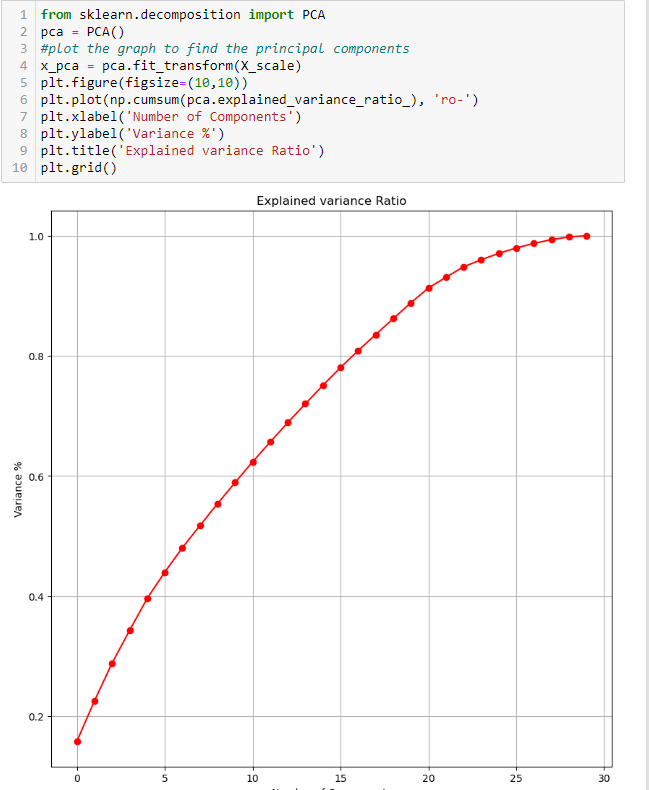
If the dataset is imbalanced for eg one class has 70% data and other has 30% data. Then there might be a possibility that the result may be biased towards the class which has a higher number of data. We can resolve this by Synthetic Minority oversampling Technique (SMOTE) to over sample the minority class and make it to 50% so then both classes are having the same number of Data points.



**Scaling Data using standard scaler -**

****

**Using PCA dimensionality reduction for feature selection -**

****

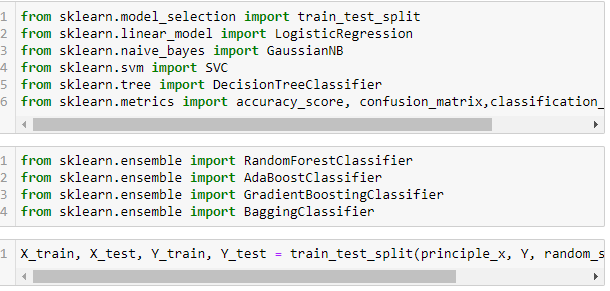
Above plot shows the contribution of features. As we can clearly see 21 components are contributing to 90% of data so we take 21 components and make a dataset.

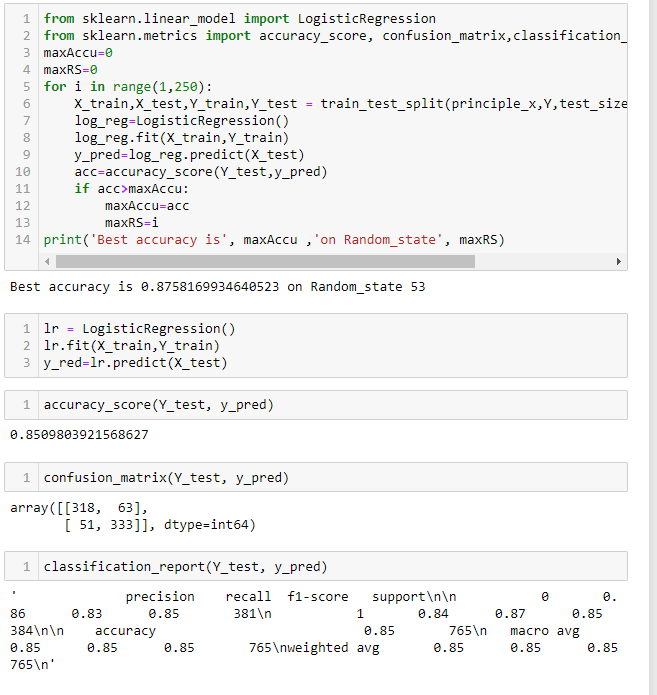
**6. Machine Learning Model Building -**

In this we will build a supervised machine learning model. We first split the data into Train and Test and make the model.

First we will build a Logistic Regression model as our base model.

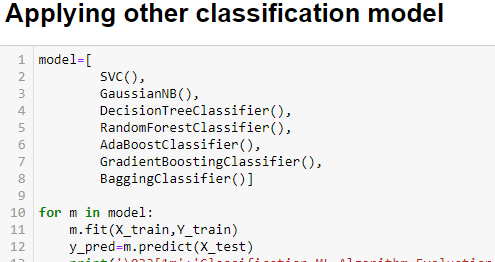
Importing necessary libraries





Logistic regression model is trained and secured an accuracy score of 85%

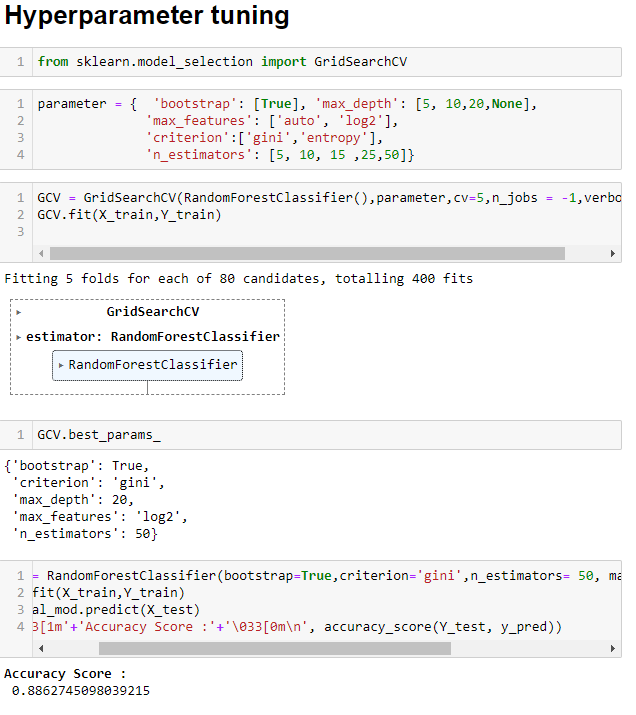
As our base model is ready we will now train other model and select model with best accuracy and validation score



| Algorithm | Accuracy score | CV Mean score | F-1 score | Recall | Precision |
| --- | --- | --- | --- | --- | --- |
| Logistic regression | 0.85 | 0.69 | 83 | 83 | 83 |
| SVC | 0.90 | 0.59 | 90 | 90 | 90 |
| Gaussian NB | 0.83 | 0.74 | 84 | 84 | 84 |
| Decision Tree Classifier | 0.80 | 0.84 | 80 | 80 | 80 |
| Random forest Classifier | 0.88 | 0.91 | 89 | 89 | 89 |
| Ada boost classifier | 0.83 | 0.86 | 83 | 84 | 83 |
| Gradient Boost Classifier | 0.86 | 0.88 | 86 | 86 | 87 |
| Bagging Classifier | 0.83 | 0.89 | 83 | 83 | 83 |

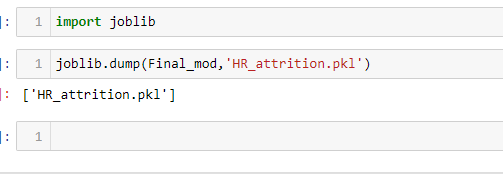
As CV score for Random Forest Classifier is high we choose Random forest classifier and tune hyper parameters to get maximum score.

Hyper parameter directly controls model structure, function and performance. Hyperparameter allow data scientist to tweak the performance of model and retrieve maximum accuracy from the model.



Final model Accuracy score falls below the score we get before so there is a possibility that we should use default parameters to get maximum accuracy from this model.

Then we will finally save this model for further use using joblib



**Conclusion Remarks -**

* **There is high attrition in sales representatives and laboratory technicians.**
* **Random forest classifier model gives maximum accuracy.**
* **Maximum number of employees are from lifesciences**
* **To get the best feature for model feeding PCA is used and 21 components are then used to build the model.**