**HR Analytics Project- Understanding the Attrition in HR**

**1. Problem Definition.**

Problem Statement: 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. But where HR Analytics fit in this? and is it just about improving the performance of employees?

**HR Analytics**

Human resource analytics (HR analytics) is an area in the field of analytics that refers to applying analytic processes to the human resource department of an organization in the hope of improving employee performance and therefore getting a better return on investment. HR analytics does not just deal with gathering data on employee efficiency. Instead, it aims to provide insight into each process by gathering data and then using it to make relevant decisions about how to improve these processes.

**Attrition in HR**

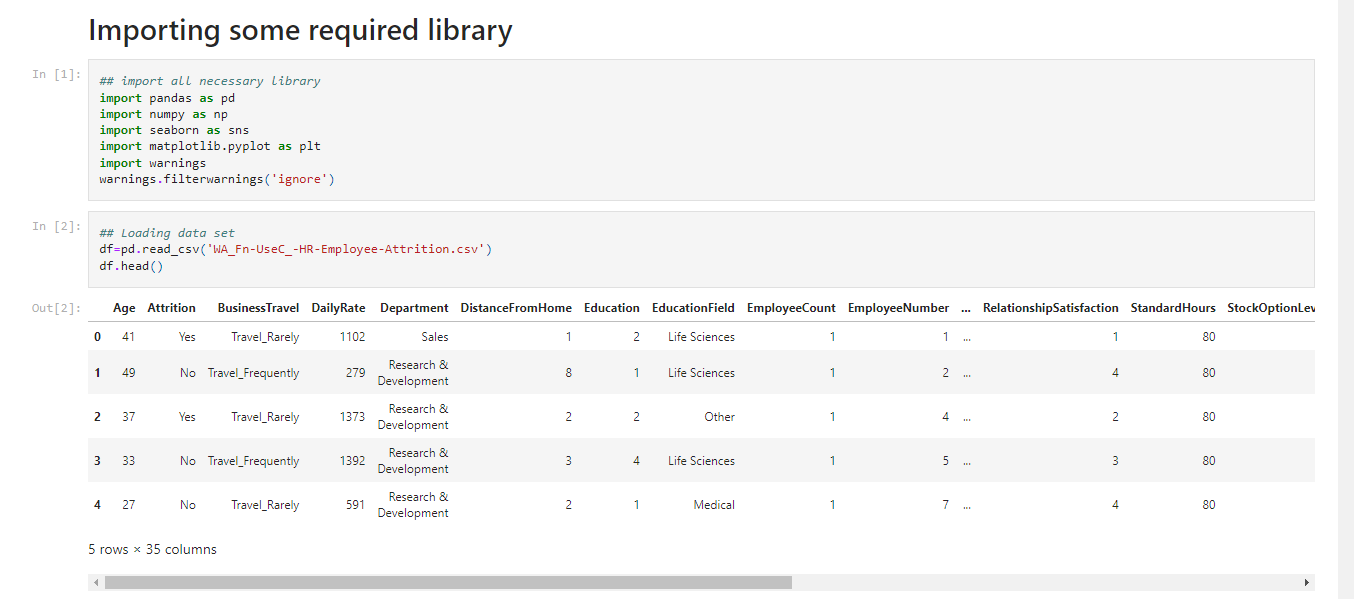
Attrition in human resources refers to the gradual loss of employees overtime. In general, relatively high attrition is problematic for companies. HR professionals often assume a leadership role in designing company compensation programs, work culture, and motivation systems that help the organization retain top employees.

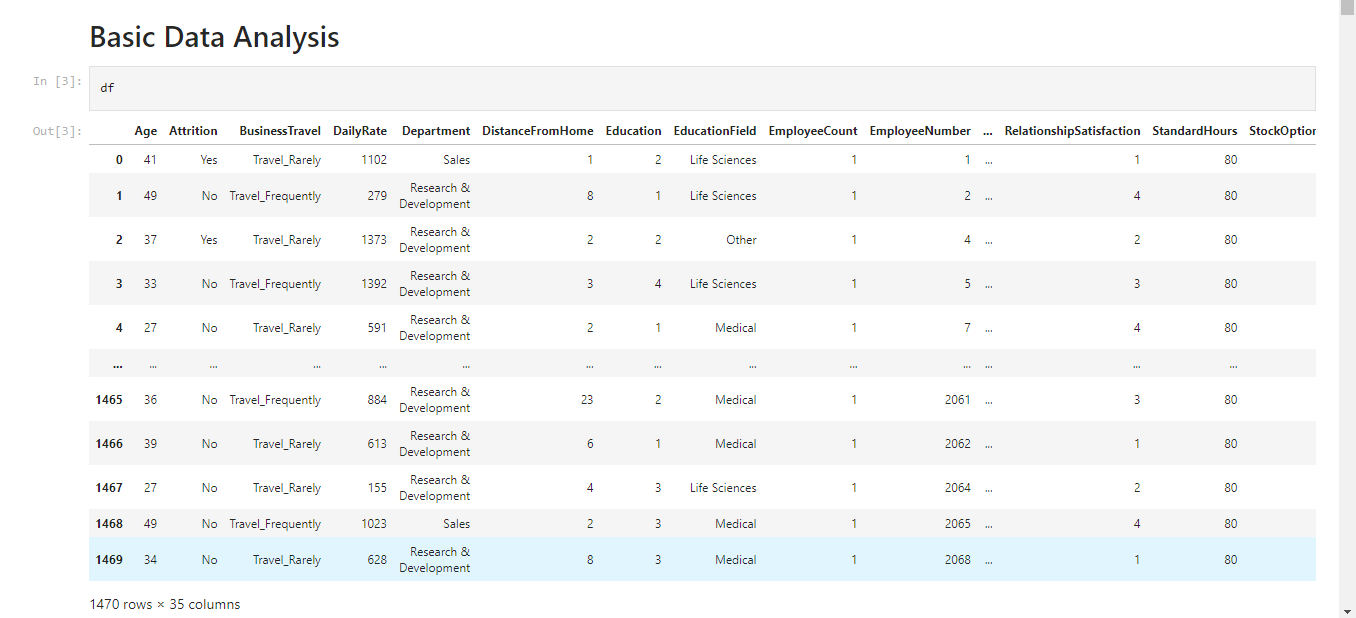
How does Attrition affect companies? and how does HR Analytics help in analyzing attrition? We will discuss the first question here and for the second question, we will write the code and try to understand the process step by step.

**Attrition affecting Companies**

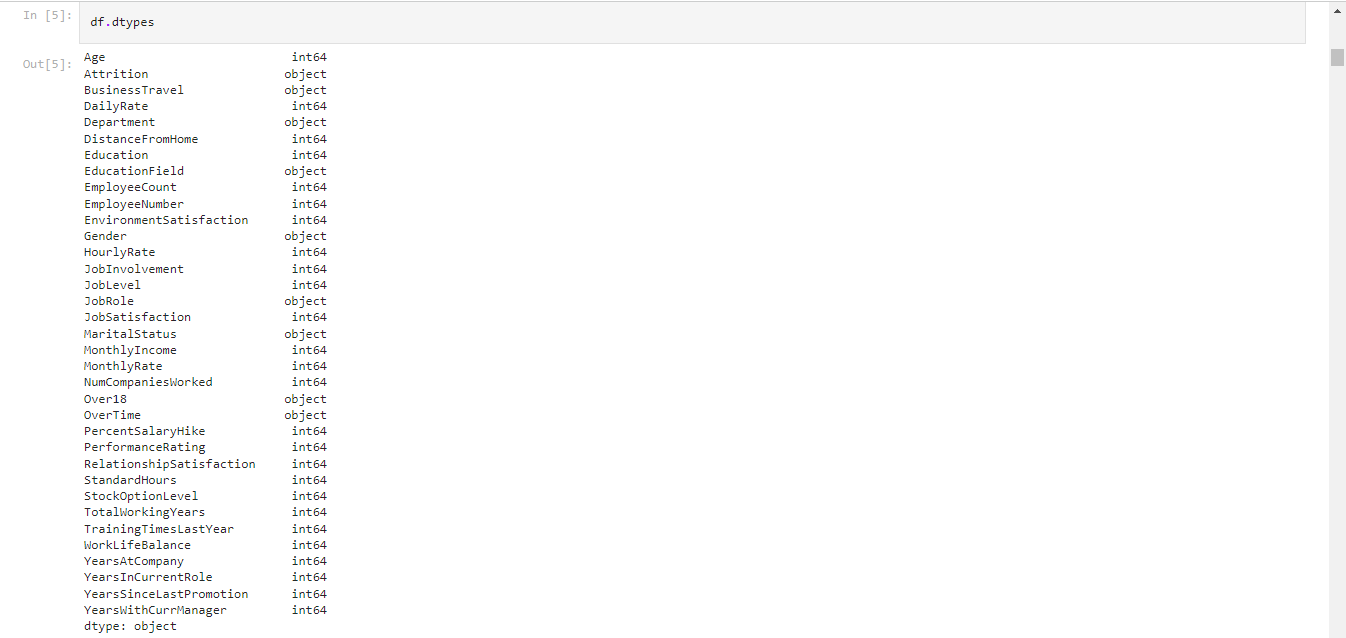
A major problem in high employee attrition is its cost to an organization. Job postings, hiring processes, paperwork, and new hire training are some of the common expenses of losing employees and replacing them. Additionally, regular employee turnover prohibits your organization from increasing its collective 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.

**2. Data Analysis**

****

****

* Here, we can see the data in row and column format.

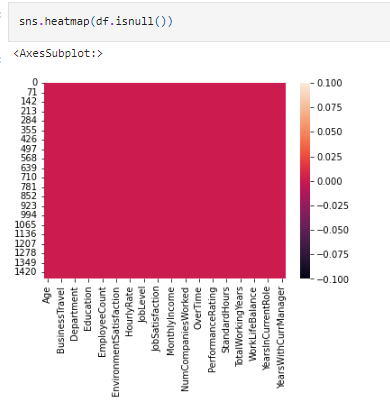


* Categorical and Numerical type of data is available in this data set.

**Check the null values in Dataset**

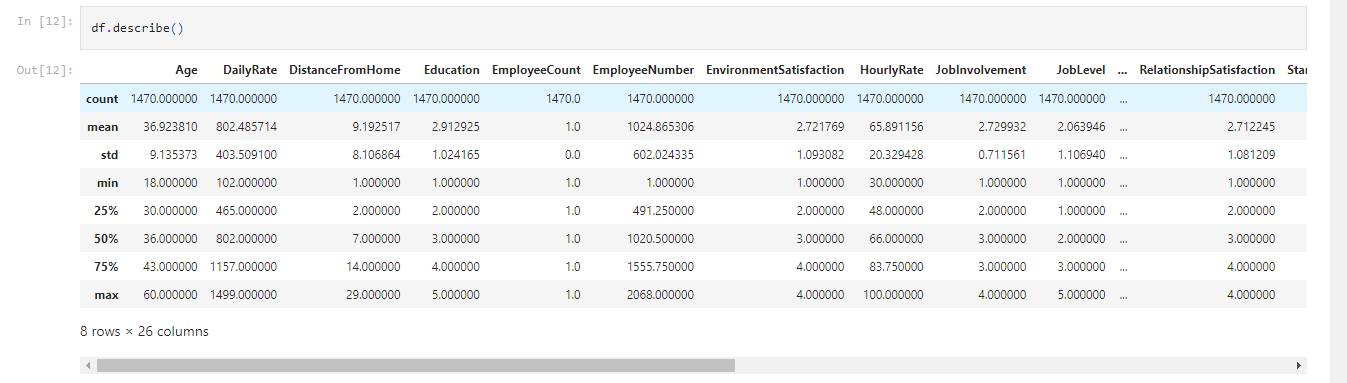
* It can be checked by heatmap.
* It can be checked by isnull() method.
* It can be checked by isnull().sum() also.

**Output:**



**Observation:** It can be seen that there are no any null values in this dataset.

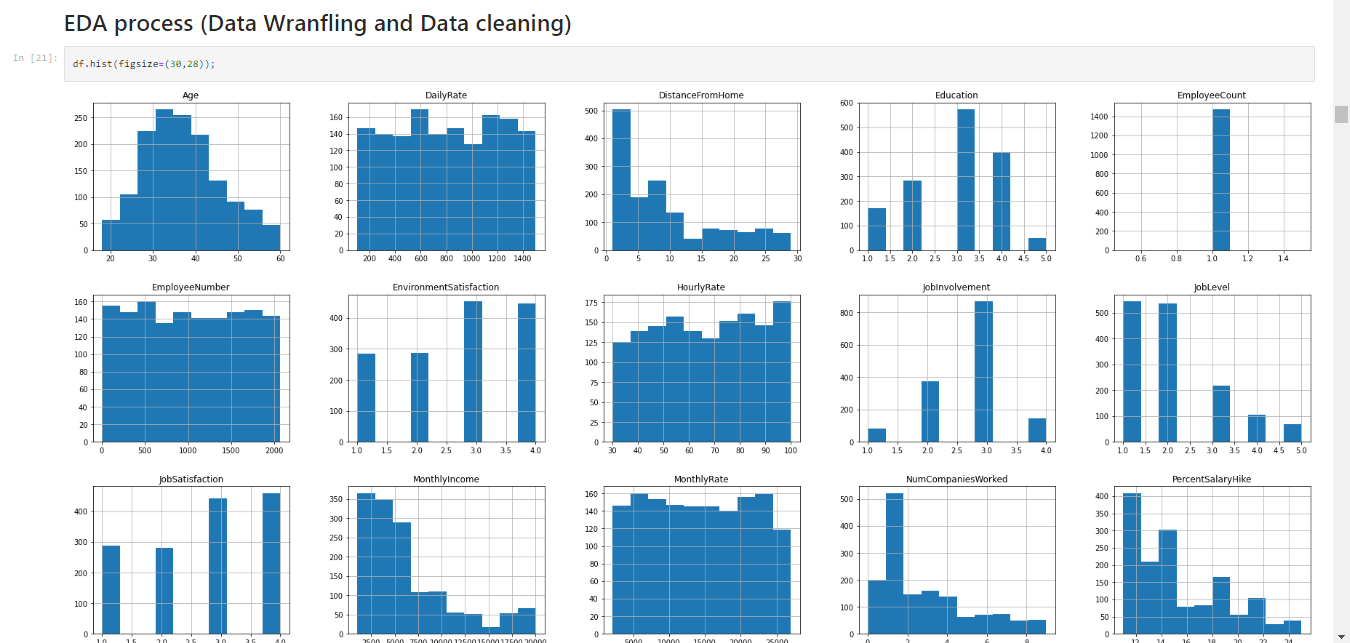
**Summary Statistics:**

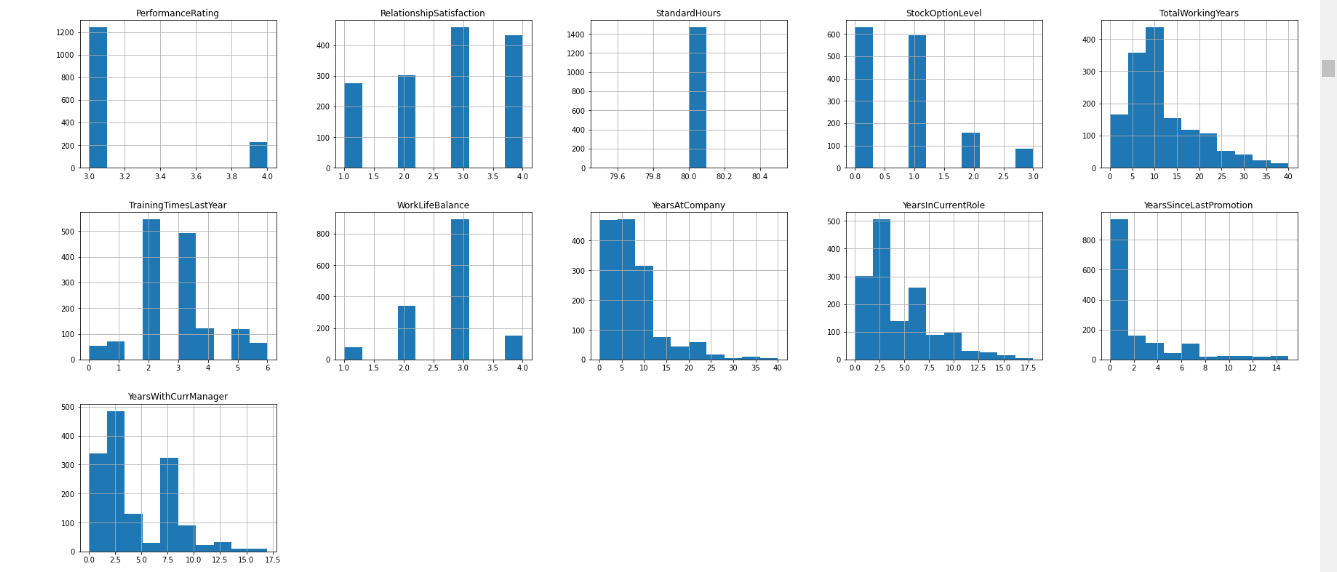


**Observation:**

* In some column mean is more than median so skewness is present and a large gap can be found in some attribute so outliers are there.

**3. EDA Concluding Remark.**

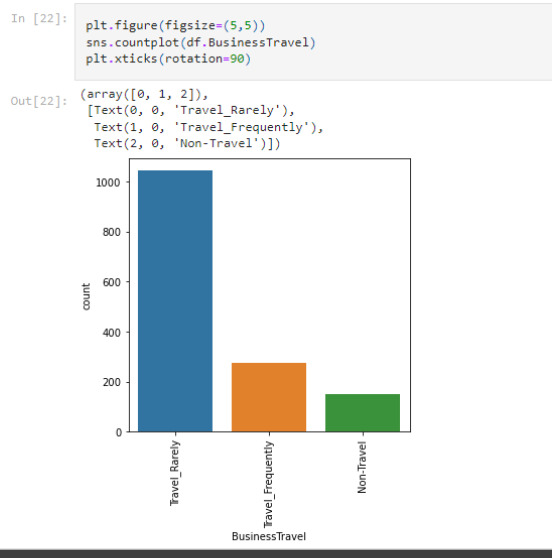




**Observation:**

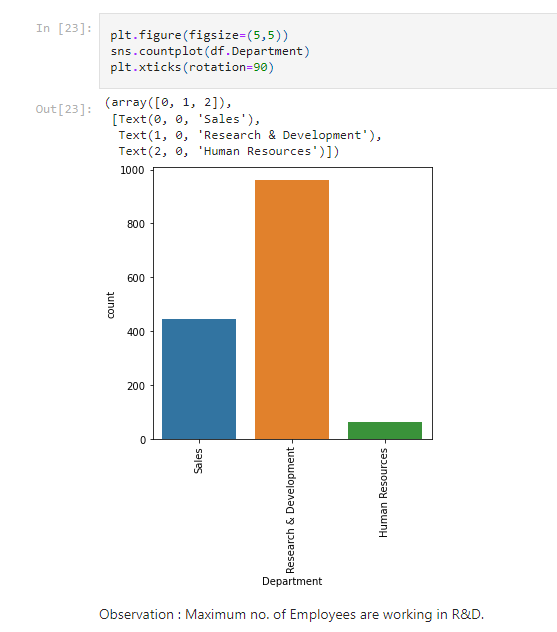
* Above, analysis it can be seen that skewness is there in some attributes. It needs to be remove it to get normally distributed data.

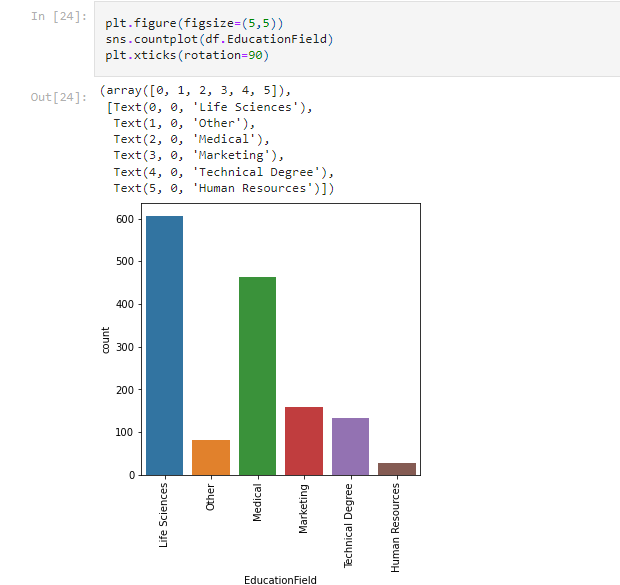
**BAR PLOT**



**Observation:**

* Travel\_Rarely by employees is maximum



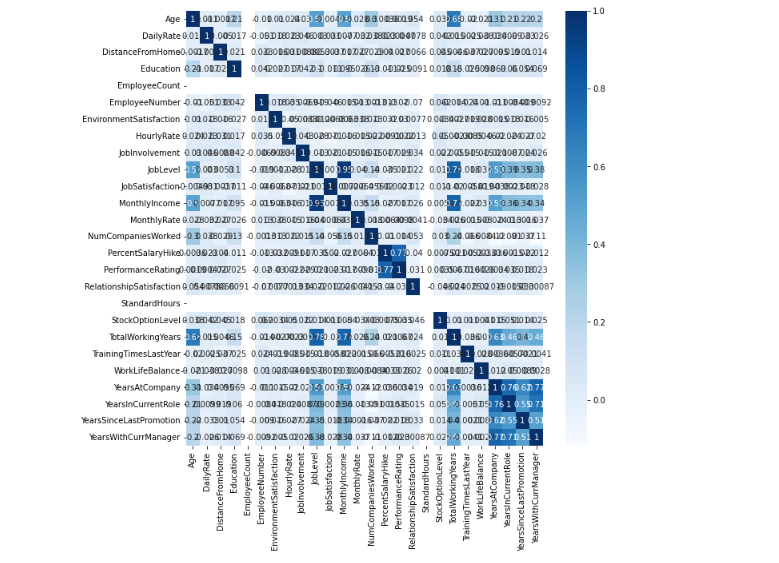


**Observation:**

* In Education Field We can see that employee come from Life Science has maximum count.

**Correlation between attributes**

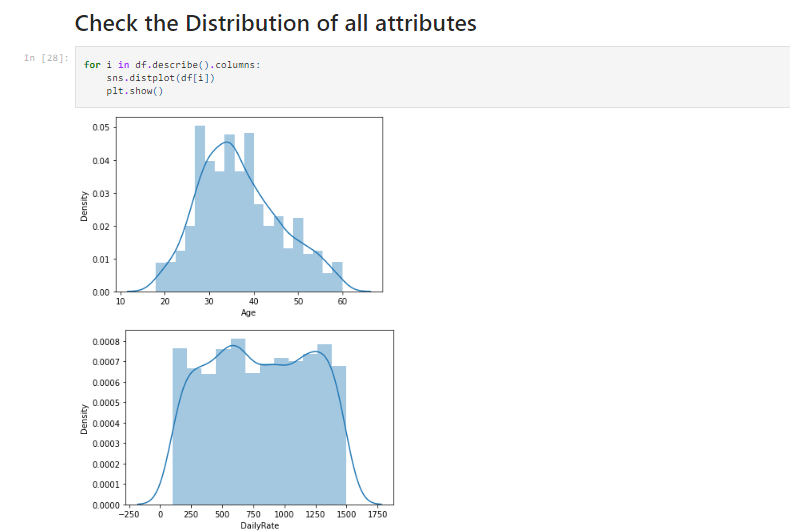


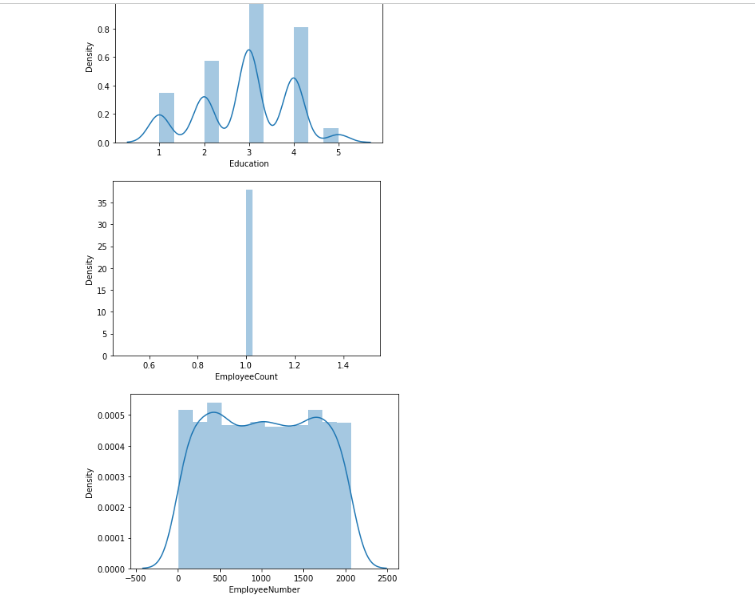


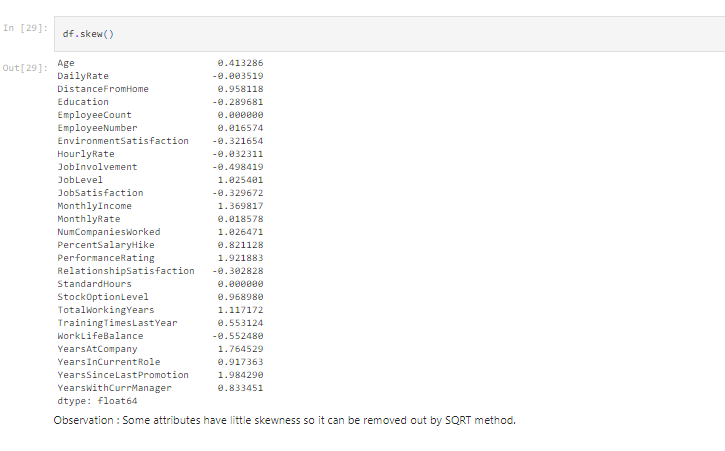
**Observation:**

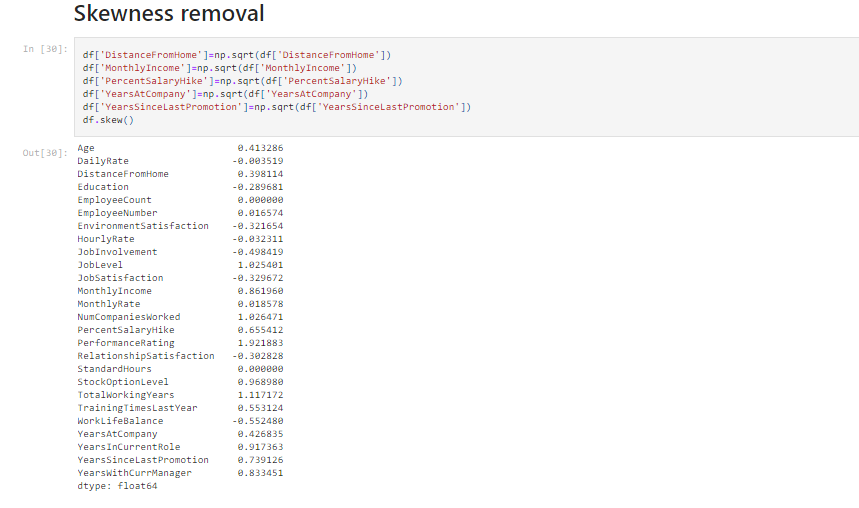
* From the above heat map we can now see which variables are poorly correlated and which ones are strongly correlated.

**Distribution of data**

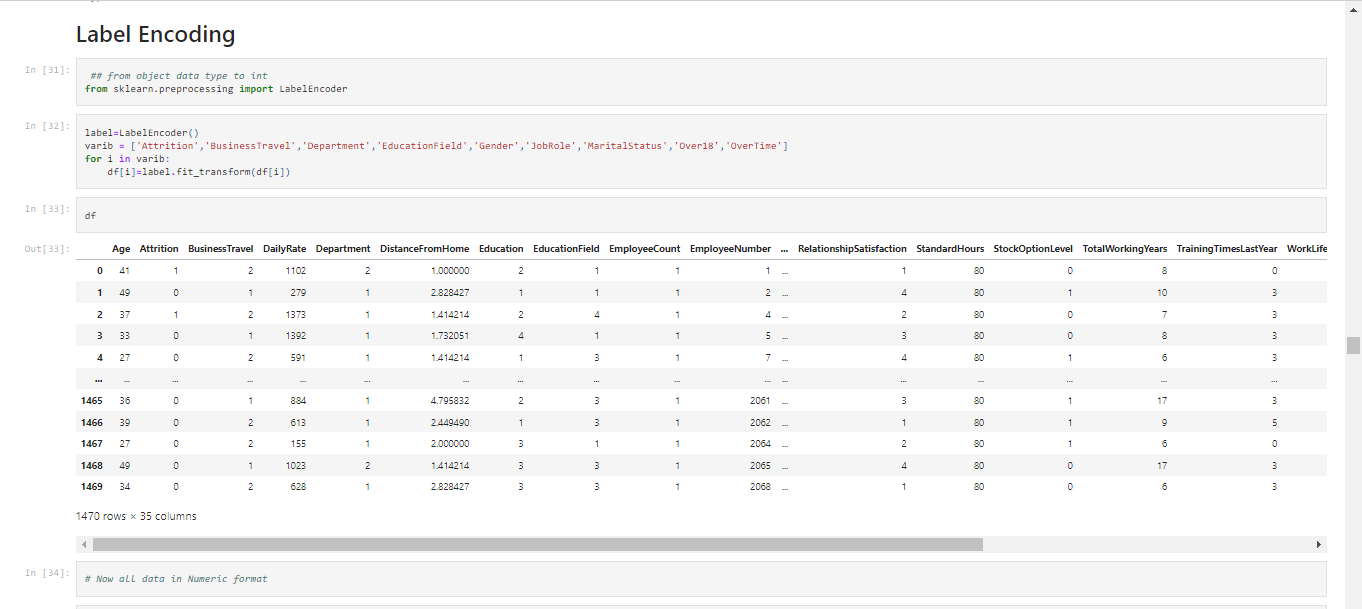


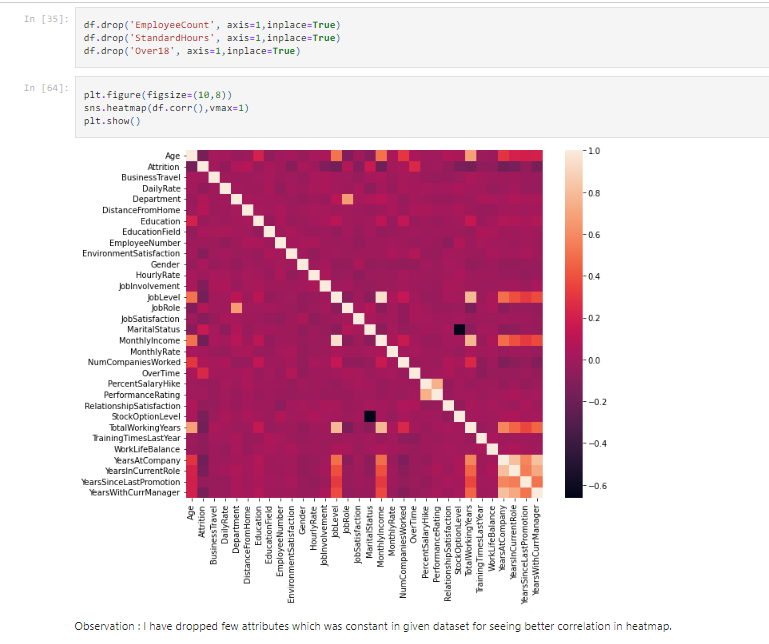




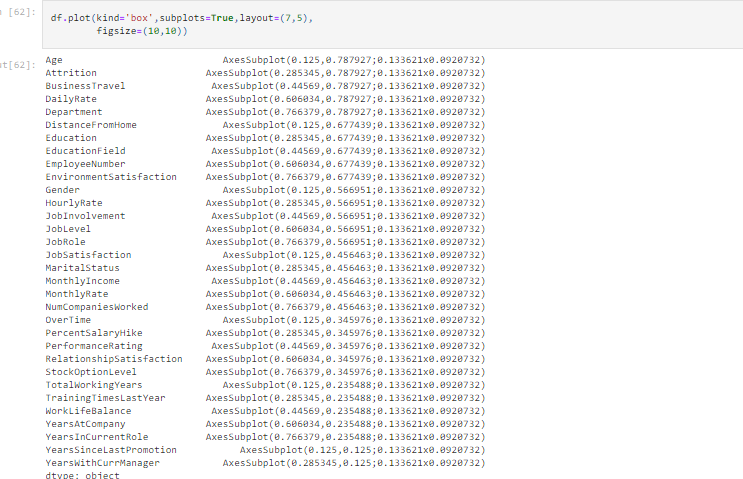


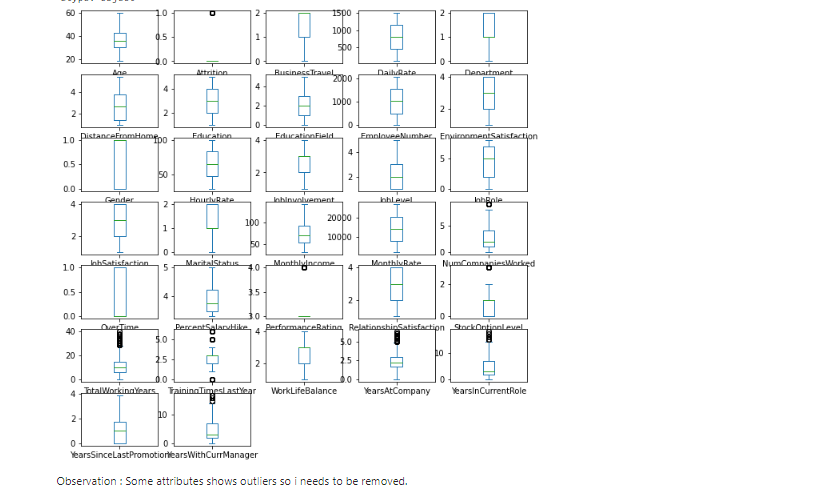
**Label Encoder**





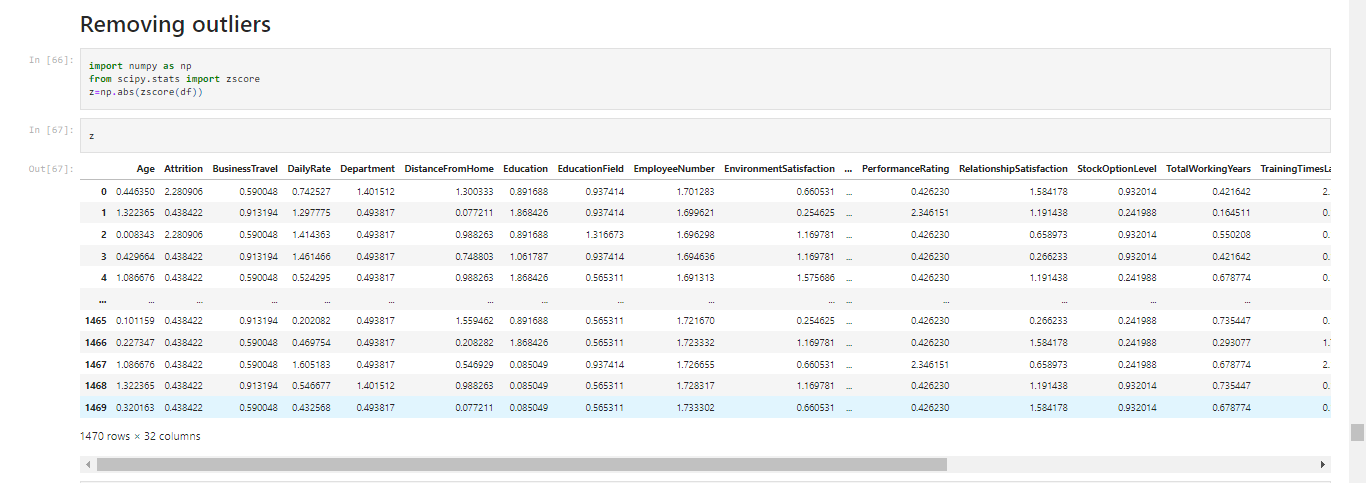
**Checking for outliers**

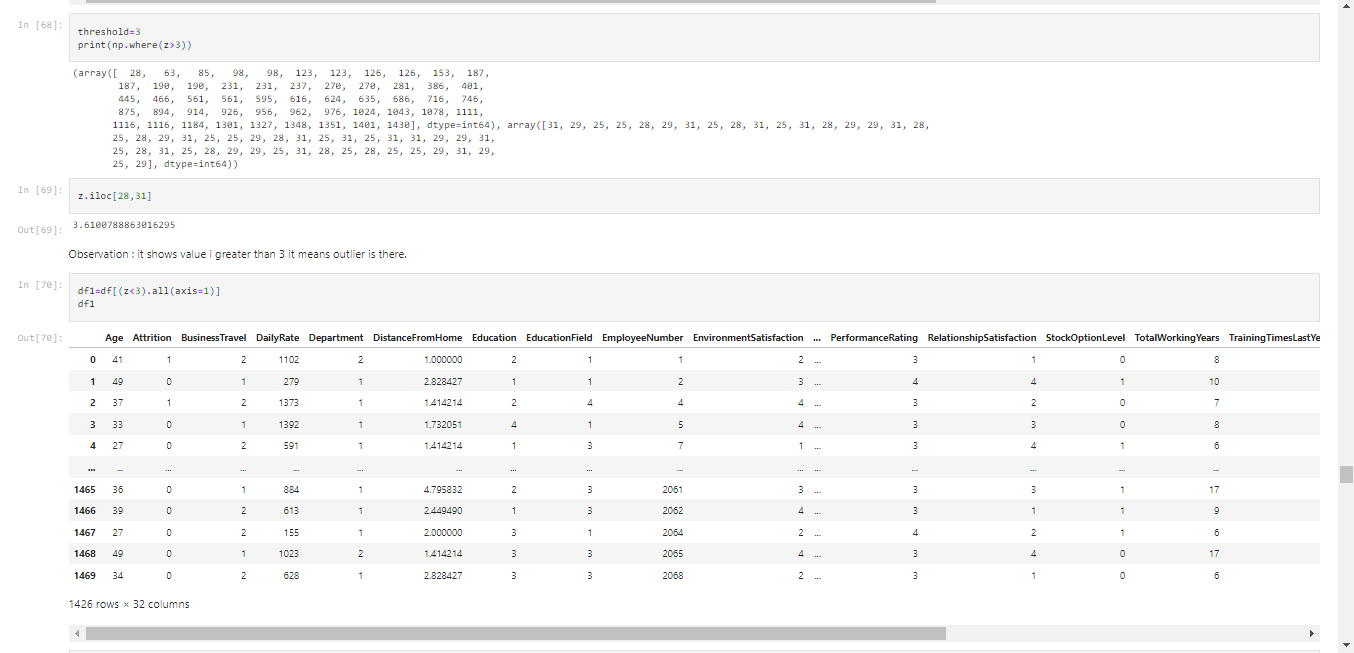




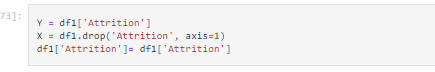
* Outliers can be removed by zscore method.

**Remove outliers**

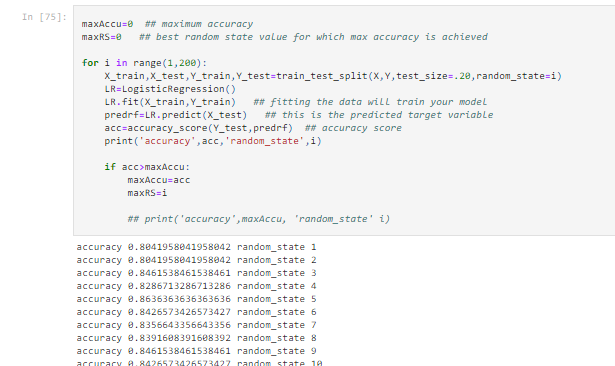


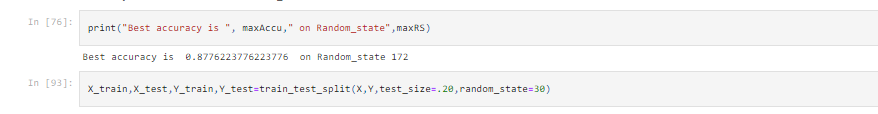


**Train\_Test\_Split Data**



**Code to get best random state**





**ML models**

**About Logistic regression**

This type of statistical model (also known as *logit model*) is often used for classification and predictive analytics. Logistic regression estimates the probability of an event occurring, such as voted or didn’t vote, based on a given dataset of independent variables. Since the outcome is a probability, the dependent variable is bounded between 0 and 1. In logistic regression, a logit transformation is applied on the odds—that is, the probability of success divided by the probability of failure. This is also commonly known as the log odds, or the natural logarithm of odds, and this logistic function is represented by the following formulas:

Logit(pi) = 1/(1+ exp(-pi))

ln(pi/(1-pi)) = Beta\_0 + Beta\_1\*X\_1 + … + B\_k\*K\_k

In this logistic regression equation, logit(pi) is the dependent or response variable and x is the independent variable. The beta parameter, or coefficient, in this model is commonly estimated via maximum likelihood estimation (MLE). This method tests different values of beta through multiple iterations to optimize for the best fit of log odds. All of these iterations produce the log likelihood function, and logistic regression seeks to maximize this function to find the best parameter estimate. Once the optimal coefficient (or coefficients if there is more than one independent variable) is found, the conditional probabilities for each observation can be calculated, logged, and summed together to yield a predicted probability. For binary classification, a probability less than .5 will predict 0 while a probability greater than 0 will predict 1.  After the model has been computed, it’s best practice to evaluate the how well the model predicts the dependent variable, which is called goodness of fit. The Hosmer–Lemeshow test is a popular method to assess model fit.

**Types of logistic regression**

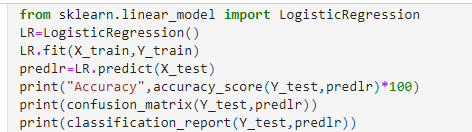
There are three types of logistic regression models, which are defined based on categorical response.

**Binary logistic regression:**In this approach, the response or dependent variable is dichotomous in nature—i.e. it has only two possible outcomes (e.g. 0 or 1). Some popular examples of its use include predicting if an e-mail is spam or not spam or if a tumor is malignant or not malignant. Within logistic regression, this is the most commonly used approach, and more generally, it is one of the most common classifiers for binary classification.

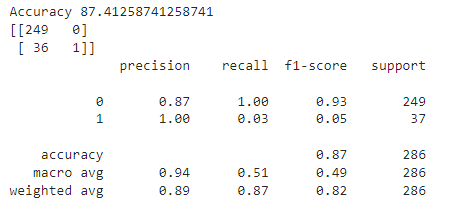
**Multinomial logistic regression:**In this type of logistic regression model, the dependent variable has three or more possible outcomes; however, these values have no specified order.  For example, movie studios want to predict what genre of film a moviegoer is likely to see to market films more effectively. A multinomial logistic regression model can help the studio to determine the strength of influence a person's age, gender, and dating status may have on the type of film that they prefer. The studio can then orient an advertising campaign of a specific movie toward a group of people likely to go see it.

**Ordinal logistic regression:**This type of logistic regression model is leveraged when the response variable has three or more possible outcome, but in this case, these values do have a defined order. Examples of ordinal responses include grading scales from A to F or rating scales from 1 to 5.

**Logistic regression**



**Output:**



**About DecisionTreeClassifier**

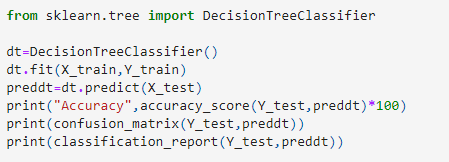
Decision Tree Learning is supervised learning approach used in [statistics](https://en.wikipedia.org/wiki/Statistics), [data mining](https://en.wikipedia.org/wiki/Data_mining) and [machine learning](https://en.wikipedia.org/wiki/Machine_learning). In this formalism, a classification or regression [decision tree](https://en.wikipedia.org/wiki/Decision_tree) is used as a [predictive model](https://en.wikipedia.org/wiki/Predictive_model) to draw conclusions about a set of observations.

Tree models where the target variable can take a discrete set of values are called [classification](https://en.wikipedia.org/wiki/Classification) [trees](https://en.wikipedia.org/wiki/Decision_tree); in these tree structures, [leaves](https://en.wikipedia.org/wiki/Leaf_node) represent class labels and branches represent [conjunctions](https://en.wikipedia.org/wiki/Logical_conjunction) of features that lead to those class labels. Decision trees where the target variable can take continuous values (typically [real numbers](https://en.wikipedia.org/wiki/Real_numbers)) are called [regression](https://en.wikipedia.org/wiki/Regression_analysis) [trees](https://en.wikipedia.org/wiki/Decision_tree).

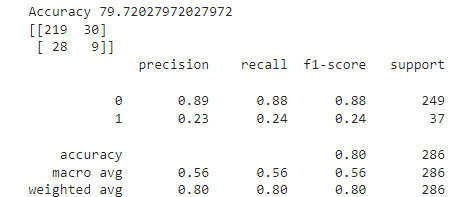
Decision trees are among the most popular machine learning algorithms given their intelligibility and simplicity.

In decision analysis, a decision tree can be used to visually and explicitly represent decisions and [decision making](https://en.wikipedia.org/wiki/Decision_making). In [data mining](https://en.wikipedia.org/wiki/Data_mining), a decision tree describes data (but the resulting classification tree can be an input for [decision making](https://en.wikipedia.org/wiki/Decision_making)).

**DecisionTreeClassifier**



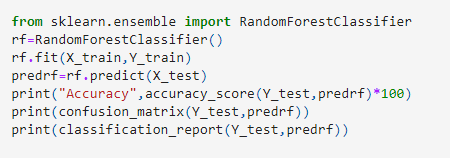
**Output:**



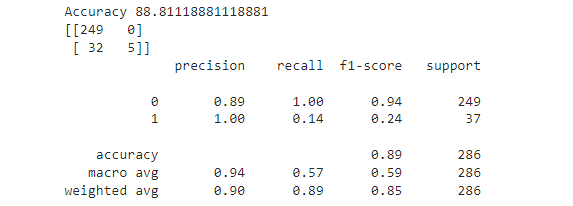
**About RandomForestClassifier**

Random forests or random decision forests is an [ensemble learning](https://en.wikipedia.org/wiki/Ensemble_learning) method for [classification](https://en.wikipedia.org/wiki/Statistical_classification), [regression](https://en.wikipedia.org/wiki/Regression_analysis) and other tasks that operates by constructing a multitude of [decision trees](https://en.wikipedia.org/wiki/Decision_tree_learning) at training time. For classification tasks, the output of the random forest is the class selected by most trees. For regression tasks, the mean or average prediction of the individual trees is returned. Random decision forests correct for decision trees' habit of [overfitting](https://en.wikipedia.org/wiki/Overfitting) to their [training set](https://en.wikipedia.org/wiki/Test_set).   Random forests generally outperform [decision trees](https://en.wikipedia.org/wiki/Decision_tree_learning), but their accuracy is lower than gradient boosted trees [[citation needed](https://en.wikipedia.org/wiki/Wikipedia:Citation_needed)]. However, data characteristics can affect their performance.

**RandomForestClassifier**



**Output :**

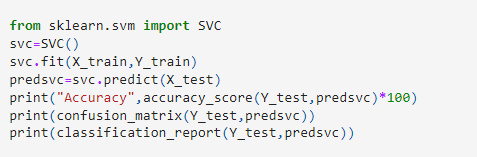


**About Support vector classifier**

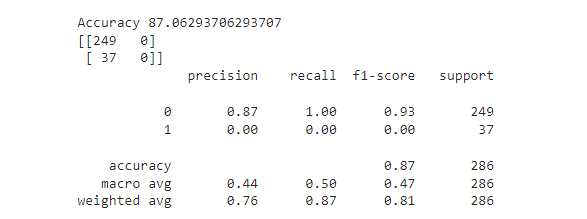
“Support Vector Machine” (SVM) is a supervised [machine learning algorithm](https://courses.analyticsvidhya.com/courses/introduction-to-data-science-2?utm_source=blog&utm_medium=understandingsupportvectormachinearticle) that can be used for both classification or regression challenges. However,  it is mostly used in classification problems. In the SVM algorithm, we plot each data item as a point in n-dimensional space (where n is a number of features you have) with the value of each feature being the value of a particular coordinate. Then, we perform classification by finding the hyper-plane that differentiates the two classes very well (look at the below snapshot).



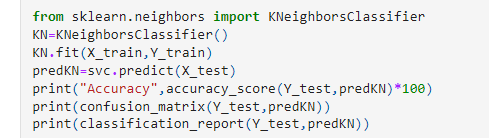
**Support vector classifier**



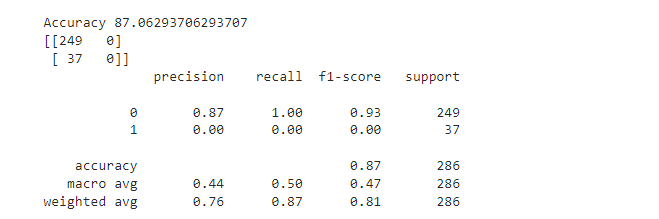
**Output:**



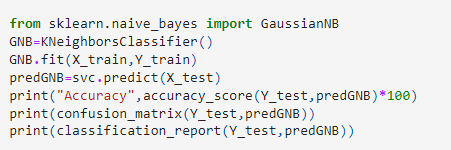
**KNeighborsClassifier**

+

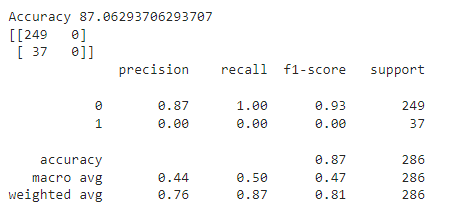
**Output:**



**GaussianNB Classifier**



**Output:**



**About Cross validation**

1 Holdout cross-validation

The hold out technique is an exhaustive cross-validation method. That randomly splits the dataset into train and test data depending on data analysis.

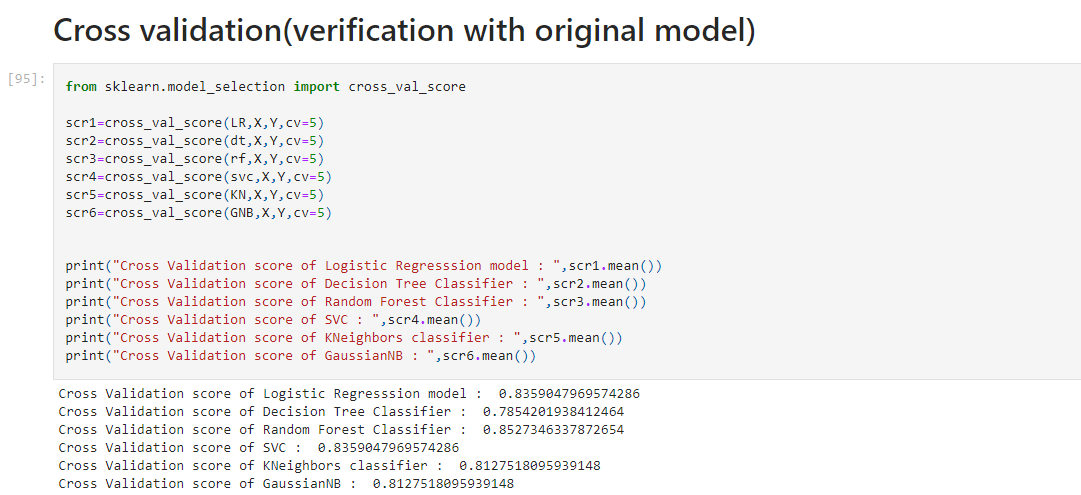
In the case of holdout cross-validation the dataset is randomly split into training and validation data. Generally, the split of training data is more than test data. The training data is used to induce the model and validation data is evaluates the performance of the model.

2 K fold Cross validation

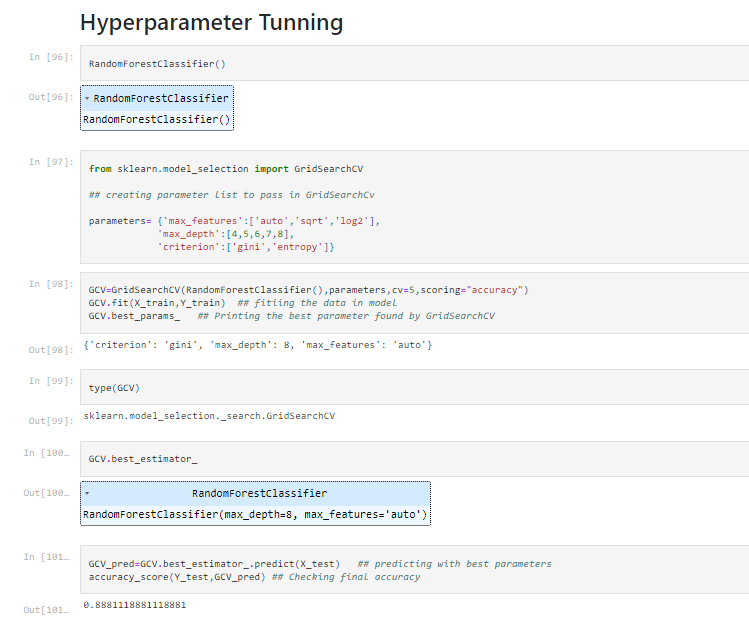
In k-fold cross-validation, the original dataset is equally partitioned into k subparts or folds. Out of the k-folds or groups for each iteration, one group is selected as validation data and the remaining (k-1) groups are selected as training data. The process for k times until each group is treated as validation and remaining as training data.

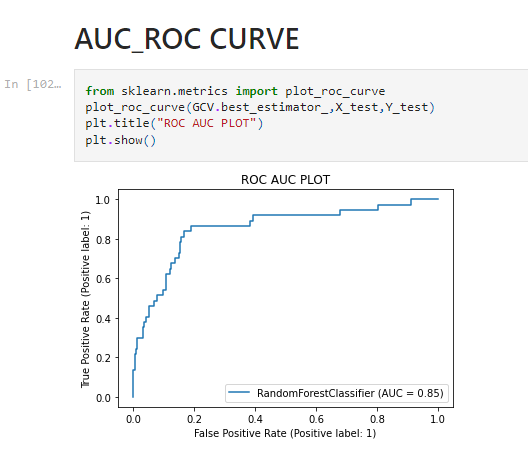
3 Leave one out cross validation

LOOCV is an exhaustive cross validation technique. For a dataset having n rows, 1st row is selected for validation and the rest (n-1) rows are used to train the model. For the next iteration the 2nd row is selected foe validation and rest to train the model. Similarly the process is repeated until n steps or the desired number of operations



**Tunning of the ML model by GridSearchCV**





* The model gave accuracy score of 0.85, not too bad. The random forest works quite well even with the default parameters. That’s one of reason we used RF for this problem. Though this can be improved by tuning hyper parameters of Random Forest classifier. Random forest also doesn’t over fit easily because of its randomness feature.
* One of the best feature Random forest model has- it provides the importance of variables/features in the data/model. For this HR Analytics problem, we are interested in knowing which feature/factor contribute the most in the Attrition and RF’s one function can give us this information. This is just another reason why we have used RF.

**Summary**

* Throughout this post, we saw Data is important in Human Resource department (actually in most of places it is important). We saw how we can avoid using correlated values and why it is important not to use those while modeling. We used Random forest and learned how it can be very advantageous over other available machine learning algorithm. Most of all we found factors which are most important to employees and if are not fulfilled might lead to Attrition.