**Blog**

on

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



# Problem Definition

# Overview: -

Employee attrition refers to the percentage of workers who leave an organization and are replaced by new employees. A high rate of attrition in an organization leads to increased recruitment, hiring and training costs. Not only it is costly, but qualified and competent replacements are hard to find. In most industries, the top 20% of people produce about 50% of the output.

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.

**Attrition Rate**

Attrition Rate is formulated as: -

Attrition Rate (%) = (Number of separations/ Average Number of employees) \* 100

# 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.

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.

# Aim of this Project: -

In this project, we will test out several machine learning models to predict when employees are going to quit by understanding the factors that lead to employee attrition.

This use case takes HR data and uses machine learning models to predict what employees will be more likely to leave given some attributes. Such model would help an organization predict employee attrition and define a strategy to reduce such costly problem.

Our main goal is to predict the probability of an employee leaving the company and our target variable is "Attrition" which is categorical in nature so this is a Binary classification problem and we will use classification algorithms to make our model.

# Steps used in this project: -

1- Define the Problem

2- Data Gathering

3- Data Cleaning

4- Data Exploration and Visualization

5- Train the algorithm

6- Evaluate our model using evaluation metrics & etc.

# Dataset Information: -

Dataset provided in \*.csv format.

HR dataset was sourced from [IBM HR Analytics Employee Attrition & Performance](https://www.ibm.com/communities/analytics/watson-analytics-blog/hr-employee-attrition/) which contains employee data for 1,470 employees with various information about the employees. We will use this dataset to predict when employees are going to quit by understanding the main drivers of employee churn

We have used IBM HR Analytics Employee Attrition & Performance Dataset in this project. This dataset contains employee data for 1,470 employees with various information about employees. The dataset includes multiple input features as Age, Business Travel, Employee Number, Daily Rate, Environment Satisfaction, Job Role, Job Satisfaction, Over Time, Percent Salary Hike, Years at Company, Years in Current Role, Work Life Balance etc.

Output/Target Variable: Attrition.

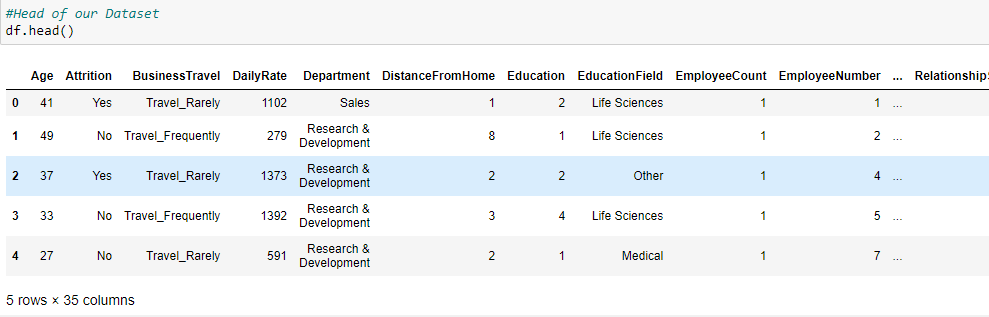
**Data Analysis**

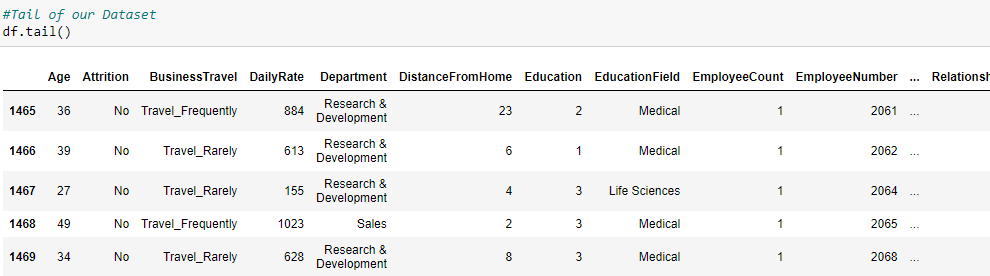
First step in any project is importing basic libraries to do data analysis and import dataset from data source on which we will work, so we are importing it.



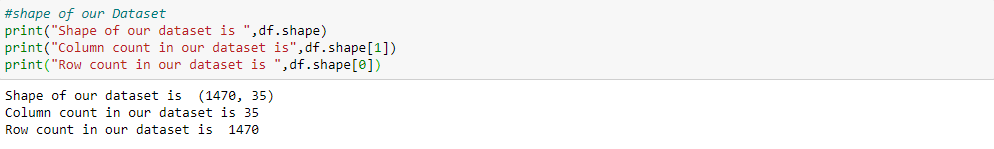


Now we will check Head and Tail of our Dataset: -

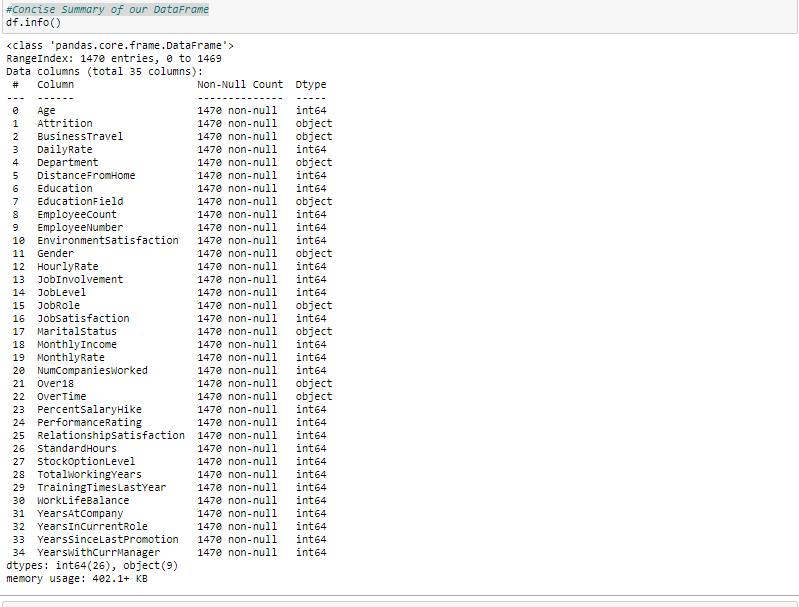




Now we will check shape of our dataset means rows and columns count in our dataset.



Now we will check Concise Summary of our Dataset. It will show us column information, datatype of all features, also length of each column and most important presence of null values.

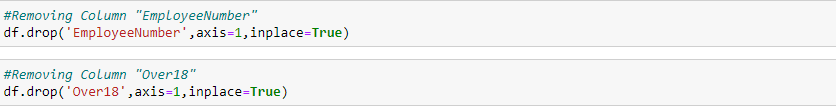


Our Dataset has no missing values and 9 columns have object type data.

**Removing Unnecessary columns**: -

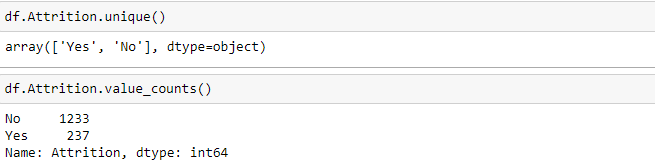
1-Column over18 has only 1 value so it will not help us in understanding the attrition rate so we can drop it.

2- Column Employee Number have id number provided by company to their employees so it will not help us in understanding the attrition rate so we can drop it.

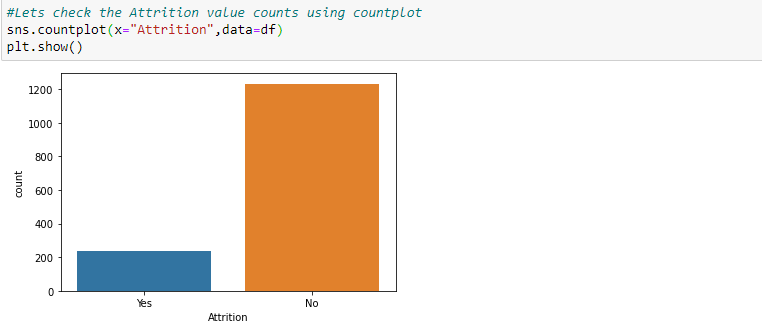


# Exploring Target Variable

Our target variable is "Attrition" which is categorical type data in nature.



Now we will do analysis of our Target variable using Count Plot.



Observations:

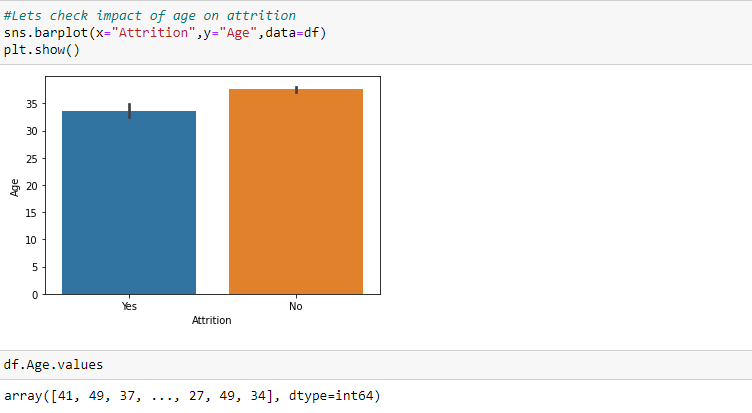
1-As per above calculation, attrition has been done for 237 employees which is near around 16% of whole dataset.

2-As per above calculation, attrition has not been done for 1233 employees which is near around 84% of whole dataset.

3-As per Count plot we can say that out of 7 employees, 1 employee is facing attrition process.

**Impact of input features on Attrition: -**

1.**Age**- We will check role of age in attrition process.



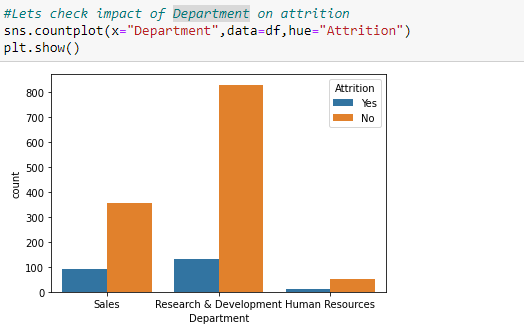
Observations:

1- As per above calculation, attrition has been done for employees having average age between 33 and 34 years.

2- As per above calculation, attrition has not been done for employees having average age between 35 and 38 years.

3- As per these observations, we can say that people who are staying in company have higher average age.

2. **Department:** - We will check role of department in attrition process.



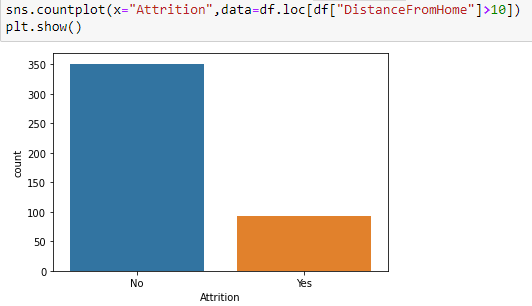
Observations:

1-As per Count plot we can say that out of 4 employees, 1 employee is facing attrition process in Sales Department.

2-As per Count plot we can say that out of 8 employees, 1 employee is facing attrition process in R&D Department.

3-Hr Department have minimal attrition rate.

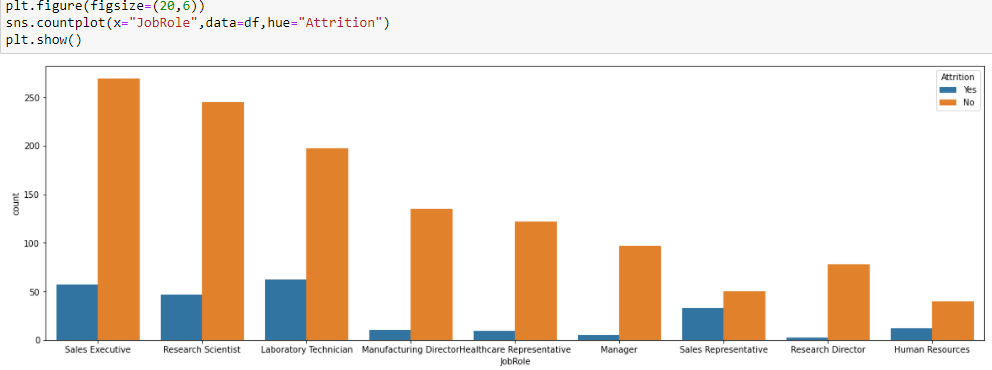
3. **Distance from home**: - If distance from home is above 10 km, then what is value of attrition.



Observation:

Out of 3 to 4 employee, 1 employee is facing attrition process when distance is above 10 km. So, if we count of distance of home for all employees then it will not put high impact.

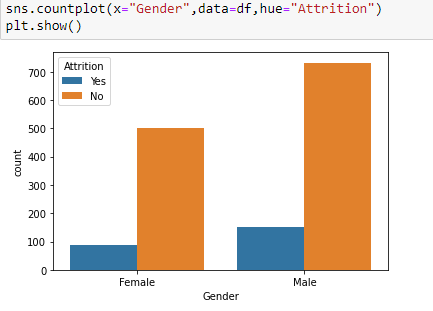
4. **Job Role**: - Let's check impact of job role in attrition process.



Observation:

Several Job Roles are listed in the dataset: Sales Executive, Research Scientist, Laboratory Technician, Manufacturing Director, Healthcare Representative, Manager, Sales Representative, Research Director & Human Resources.

5. **Gender:** - Let's check impact of gender on attrition process.

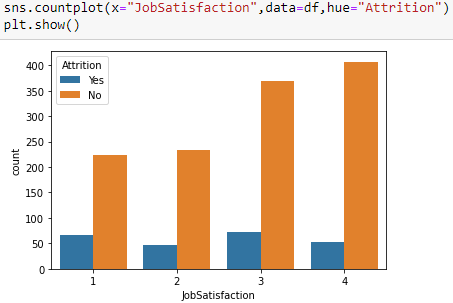


Observation:

1-As per Count plot, relative proportion of male ex-employees is higher than female ex-employees.

2-As per calculation, attrition has been done for 17% male employees and 14.5% for female employees.

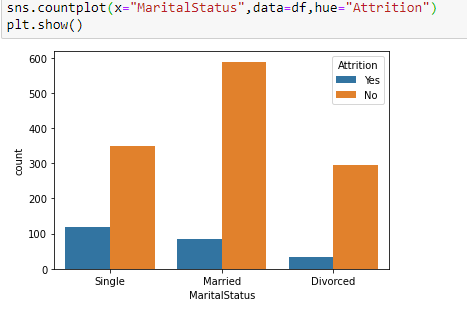
6. **Job Satisfaction**: - Let's check impact of Job Satisfaction on attrition process.



Observation:

As per count plot, if job satisfaction increases then proportionally attrition decreases.

7. **Marital Status**: - Let's check impact of Marital Status on attrition process.

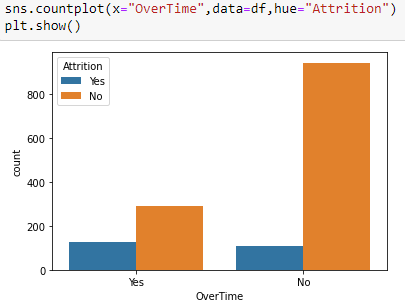


Observations:

1-Emloyees with single marital status have highest proportion of attrition at 25%.

2-Married Employees have higher tenure.

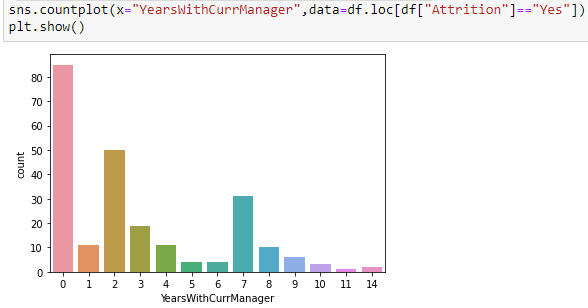
8. **Overtime**: - Let's check impact of Overtime on attrition process.



Observation:

As per above Count plot, if overtime is more then attrition proportion increases.

9. **Years with Current Manager**: - Let's check impact of Years with Current Manager on attrition.



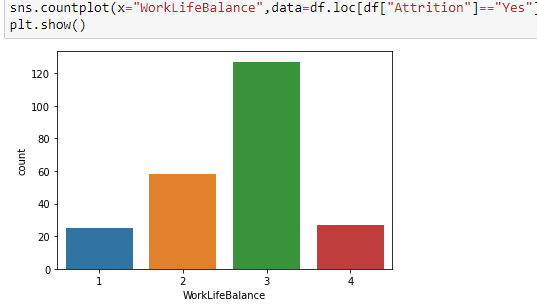
Observations:

1-If years working with current manager is 0 then attrition is highest.

2-If years working with current manager is 2 then attrition decreases.

3-As per calculations, average number of years with current manager is 4.5 for active employees and 2.9 for ex-employees.

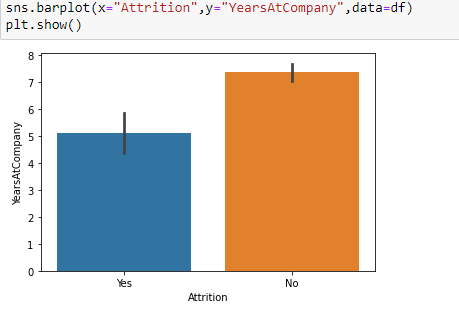
10. **Work life balance**: - Let's check impact of work life balance in attrition process.



Observation:

As per plot & calculations, employees who have "Bad" Work-Life Balance are more likely to face attrition process.

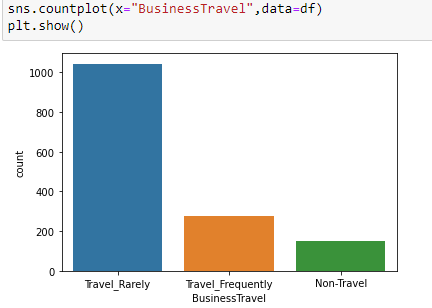
11. **Years at Company**: - Let's check impact of working years on attrition process.



Observation:

The average number of years at the company for currently active employees is 7.5 years and ex-employees is 5 years.

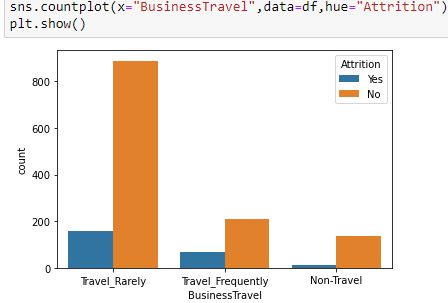
12. **Business Travel: -** First checking how many employees do travel for business.



Observation:

Most of the employees travel very rarely for business.

**Let's check impact of Business Travel on attrition process: -**

****

Observation:

As per plot & calculation, Due to frequent business travel Attrition proportion increases.

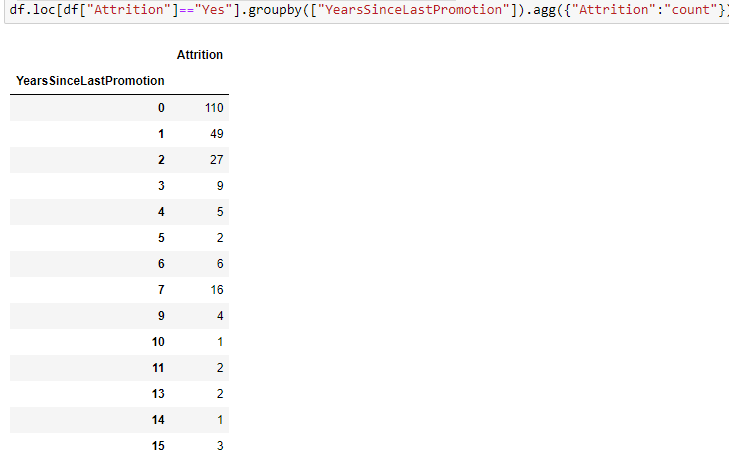
13. **Monthly Income**: - Let's check impact of Monthly Income on attrition process.



Observation:

High Monthly Income decreases the chance of attrition process.

14. **Years Since Last Promotion**: - Let's check impact of Years Since Last Promotion on attrition process.



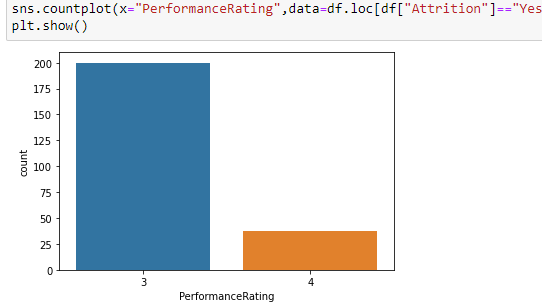
Observations:

1-Majority of employees left company have not completed a year since last promotion.

2-Employees completed more years in company are less like to leave it.

3- As per above calculations, promotion is not prime factor behind attrition.

15. **Performance Rating**: - Let's check impact of performance rating on attrition process.



Observation:

Performance rating plays an important role in attrition process.

**Removing Unnecessary columns**: -

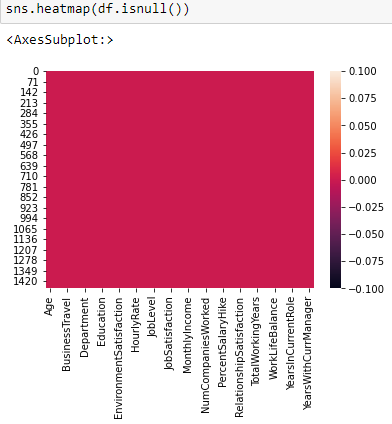
1-Column Employee Count has same value in entire dataset so it will not help us in understanding the attrition rate so we can drop it.

2- Column Standard Hours also have same value in entire dataset so it will not help us in understanding the attrition rate so we can drop it.

After dropping these features, Let’s check updated shape of dataset.



**Let’s check Null values using Heatmap: -**



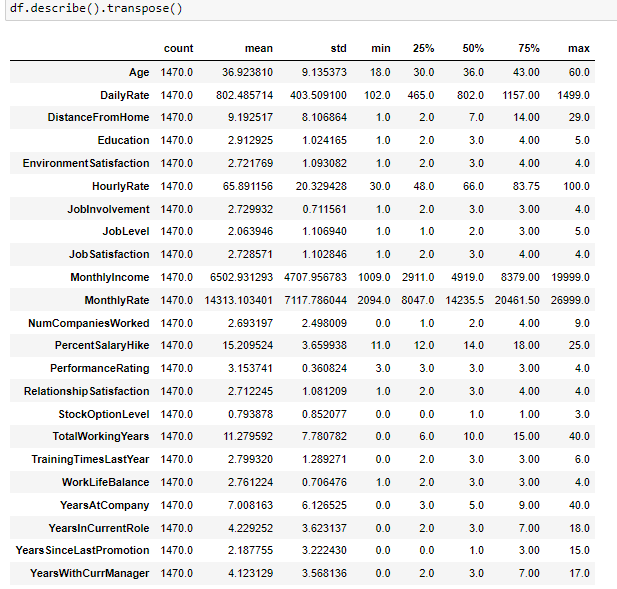
Observation: -

Dataset has no missing values.

**EDA Concluding Remarks**

In statistics, exploratory data analysis is an approach of analysing data sets to summarize their main characteristics, often using statistical graphics and other data visualization methods.

# 1- Summary Statistics: - In descriptive statistics, summary statistics is used to summarize set of observations, in order to communicate the largest amount of information as simply as possible. It includes central Tendency, dispersion, skewness, variance, range, deviation etc.



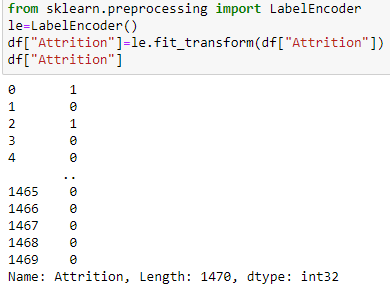
Observations on basis of Summary Statistics: -

1-The Mean is more than median (50th Percentile) in 14 columns and median is more than mean in 9 columns.

2-Count of every column is 1470 and there are no missing values.

2- **Feature Engineering: -** Feature engineering is the process of transforming raw data into features that better represent the underlying problem to the predictive models, resulting in improved model accuracy on unseen data. We are doing it for target variable as it is of object type so changing it in to numeric format using label encoder.

**Label Encoding**: - Label Encoding refers to converting the labels into numeric form so as to convert it into the machine-readable form.



# 3-Correlation Matrix: - A correlation matrix is a table showing correlation coefficients between variables. Each cell in the table shows the correlation between two variables. A correlation matrix is used to summarize data, as an input into a more advanced analysis, and as a diagnostic for advanced analyses. A correlation matrix consists of rows and columns that show the variables.

# 

# Checking correlation using Heatmap with annotations: -

# Heatmap: - It is a data visualization technique that shows magnitude of a phenomenon as colour in two dimensions. The variation in colour may be by hue or intensity, giving obvious visual cues to the reader about how the phenomenon is clustered or varies over space.

# 

Observations on basis of Correlation Matrix & Heatmap: -

1- Dark shades are highly correlated.

2-Columns Distance from Home, Monthly Rate, Number of Companies Worked & Performance rating are positively correlated to attrition.

3-Columns Job Level, Monthly Income, Total Working Years & Years in Current Role are negatively correlated to attrition.

4- **Plotting Outliers: -** An outlier is an observation that lies an abnormal distance from other values in a random sample from a population. Outliers should be investigated carefully. So, we will use graphical techniques Box Plot and Scatter Plot for identifying outliers.

**Box Plot-** The box plot is a useful graphical display for describing the behaviour of the data in the middle as well as at the ends of the distributions. The box plot uses the [median](https://www.itl.nist.gov/div898/handbook/eda/section3/eda351.htm) and the lower and upper quartiles (defined as the 25th and 75th [percentiles](https://www.itl.nist.gov/div898/handbook/prc/section2/prc252.htm)).

A- **Univariate Analysis**: - Univariate involves the analysis of a single variable. First, we are going to do univariate analysis using Box Plot method.



# 

Observations:

1- Some columns have outliers.

2- Some columns have no outliers and some have minimal outliers.

# Now we are creating histogram of each input variable to get a broader idea of the distribution.

# 

Observation:

Presence of unusual values in above histograms & also distribution is not normal in some columns and these things denote the possibility of potential outliers.

# B- Multivariate Analysis: - Multivariate analysis is used to study more complex sets of data. It is a statistical method that measures relationships between two or more response variables.

# Scatter plot matrix: -Scatter plot matrix is a grid (or matrix) of scatter plots used to visualize bivariate relationships between combinations of variables.

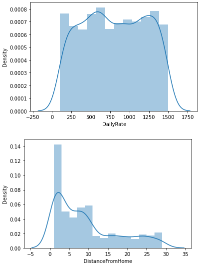
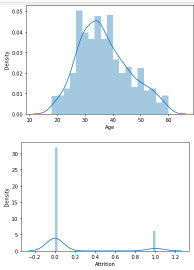
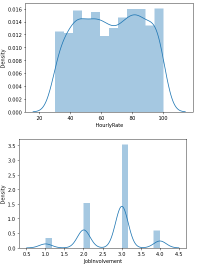
# 

Observation:

Using multivariate analysis, we can look at the interactions between the variables. Scatterplots of all pair of attributes helps us to spot structured relationship between input variables.

4- **Skewness**- Skewness is a measure of asymmetry or distortion of symmetric distribution. It measures the deviation of the given distribution of a random variable from a symmetric distribution, such as normal distribution. A normal distribution is without any skewness, as it is symmetrical on both sides. Hence, a curve is regarded as skewed if it is shifted towards the right or the left. Skewness is of 2 types: - 1- Positive Skewness 2- Negative Skewness.

# Distplot to check Distribution of Skewness: - Distplot plots a univariate distribution of observations. The distplot () function combines the matplotlib hist function with the seaborn kde plot () and rug plot () functions. For individual columns we are using Distplot.

Observations:

1-We have checked skewness in all columns but here showing only few columns. In next step we will check and try to remove skewness in more detail for our input dataset.

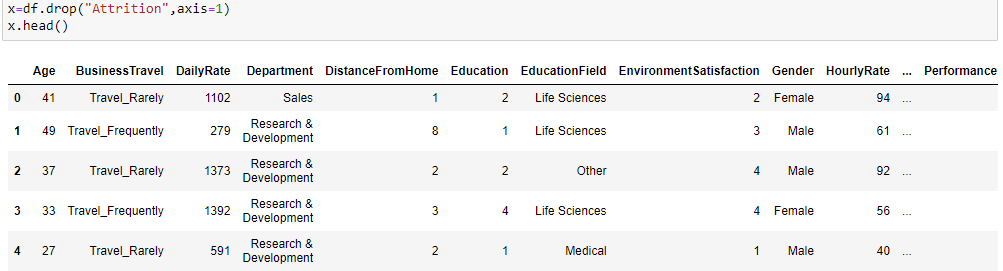
2-Huge Skewness present in some columns.

3- Except these columns, others have minimal deviation.

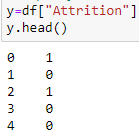
4- To avoid data loss we are not removing outliers while we will remove skewness from dataset.

**Pre-processing Pipeline**

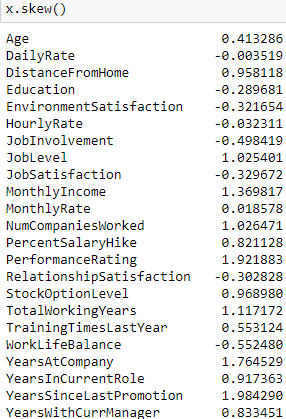
Data pre-processing is a data mining technique which is used to transform the raw data in a useful and efficient format. Pipelines are a simple way to keep our data pre-processing and modelling code organized. Many stages involved in this process are Data Cleaning, Data Integration, Data Transformation & Data Reduction. We will start this process by dividing dataset into input & output.



Output Feature: -



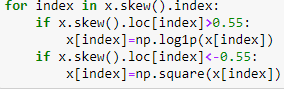
**Let's check skewness in input dataset:-**

****

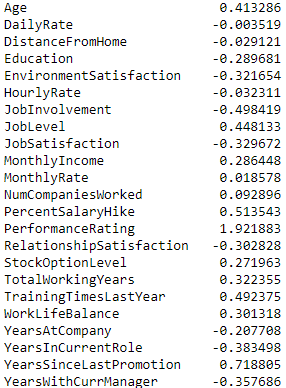
**Data Transformation by Removing skewness from input dataset: -** We will use cube root & log transform method to remove skewness in input dataset.

**Log transformation**: - Log transformation is a data transformation method in which it replaces each variable x with a log(x). It is commonly used for reducing right skewness and is often appropriate for measured variables. We will use NumPy log1p () method for this process.

**Square Root transformation**: - The square root, x to x^ (1/2) = sqrt(x), is a transformation with a  
moderate effect on distribution shape.  It is also used for reducing right skewness, and also  
has the advantage that it can be applied to zero values.



After using transformation methods, we will again check skewness in input dataset.



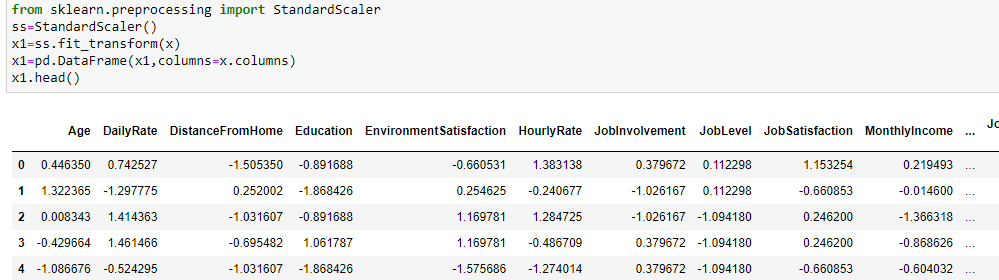
Observation:

We have reduced most of the skewness from input dataset and currently only 2 features have presence of some skewness.

**Pandas. get\_ dummies: -** We have 7 object type features in our input dataset so for better results it's going to be easy to transform the object type features to dummies/indicator variables.

**Standardization: -** Standardization of a dataset is a common requirement for many machine learning estimators: they might behave badly if the individual features do not more or less look like standard normally distributed data (e.g., Gaussian with 0 mean and unit variance).

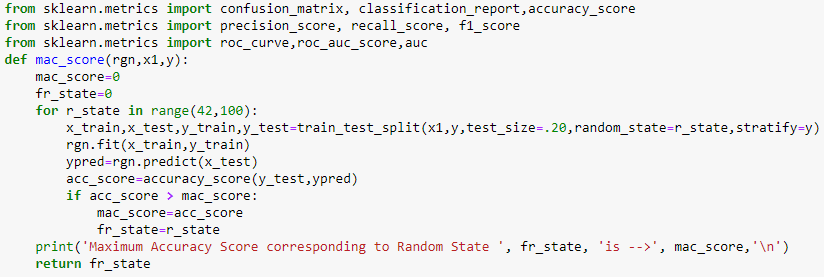
**Standard Scaler**: - Standard Scaler is a class from Sklearn.preprocessing which is used for standardization. It will Standardize features by removing the mean and scaling to unit variance.



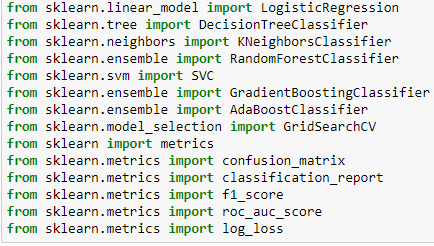
Now our Input and output features are ready to apply machine learning algorithms on them.

**Building Machine Learning Models**

Target variable is categorical in nature so we will use different classification algorithms. First, we are making function to calculate maximum accuracy score at best random state.



**Importing machine learning Libraries for classification.**

****

**Classification Evaluation metrics: -** The key classification metrics used in this project are: -

**Learning Score**: - It is the score calculated for x\_ train & y\_ train after applying classification algorithms. It is used to judge the quality of the fit (or the prediction) on new data.

**Accuracy Score**: - Accuracy score is a metric for evaluating classification models that summarizes the performance of a classification model as the number of correct predictions divided by the total number of predictions.

**Cross validation Score**: - Cross-validation is a resampling procedure used to evaluate machine learning models on a limited data sample. We are using k-Fold Cross-Validation procedure for classification models. The cross\_ val\_ score () function will be used to perform the evaluation, taking the dataset and cross-validation configuration and returning a list of scores calculated for each fold.

**F1 Score**: - The F1 Score is 2\*((precision\*recall)/ (precision +recall)). It is also called the F Score or the F Measure. F1 score conveys the balance between the precision and the recall.

**Roc\_ Auc Score**: - The roc\_ auc\_ score function computes the area under the receiver operating characteristic (ROC) curve, which is also denoted by AUC or AUROC. By computing the area under the roc curve, the curve information is summarized in one number.

**Log Loss**- It is important classification metrics indicative of how close the prediction probability is to the corresponding actual/true value (0 or 1 in case of binary classification). It can be calculated as: -

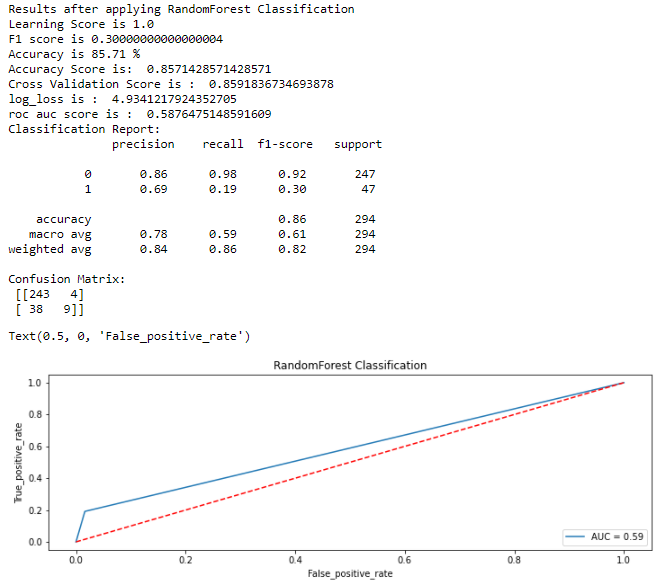


**Classification Report**: - It is used to measure the quality of prediction from classification algorithm. It is a text summary of the precision, recall, F1 score & support for each class.

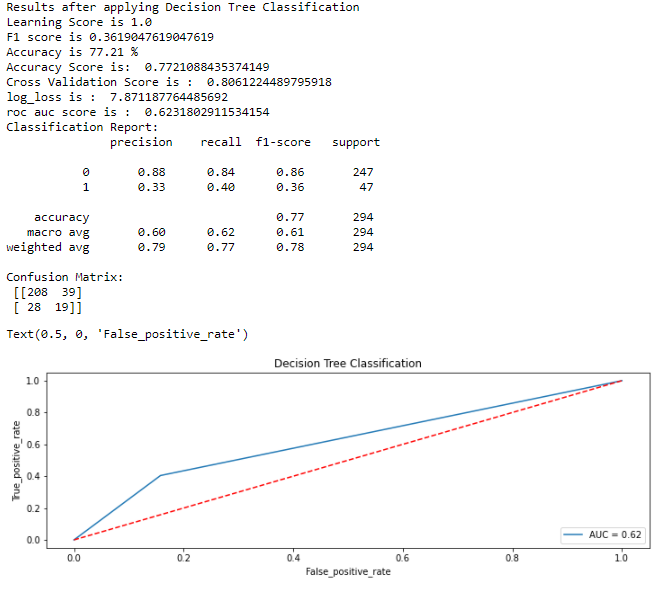
**Confusion Matrix:** - A confusion matrix is a summary of the predictions made by a classification model organized into a table by class. Each row of the table indicates the actual class and each column represents the predicted class. A value in the cell is a count of the number of predictions made for a class that are actually for a given class. The cells on the diagonal represent correct predictions, where a predicted and expected class align.

The different classification algorithms used in this project are-

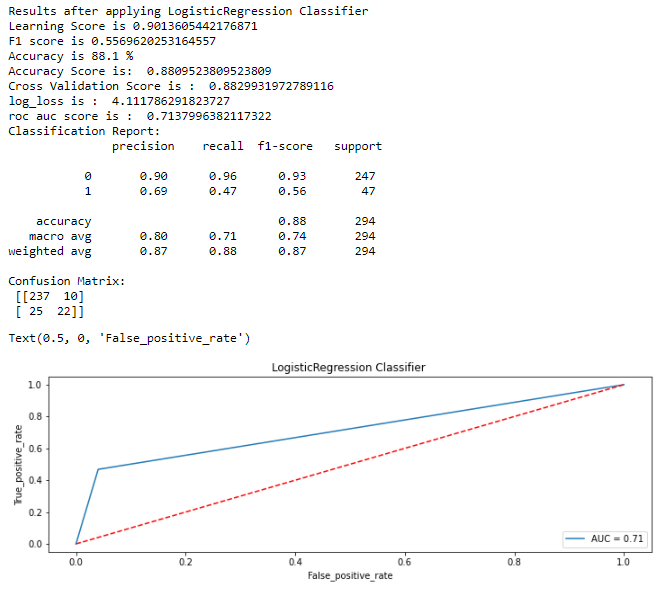
1- **Random Forest Classification:** - A random forest is a meta estimator that fits a number of decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting.



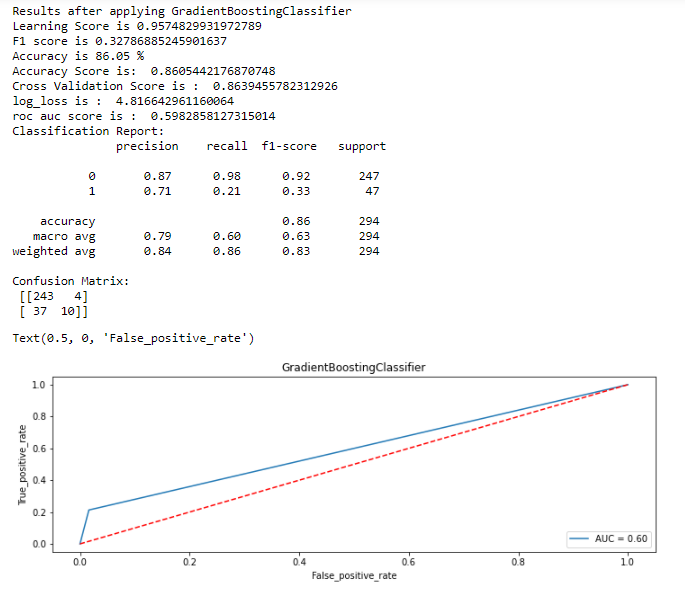
**2- Decision Tree Classification: -** Decision Tree Classifier is a class capable of performing multi-class classification on a dataset.



3- **Logistic Regression**: - Logistic regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable, although many more complex extensions exist. A binary logistic model has a dependent variable with two possible values, such as pass/fail which is represented by an indicator variable, where the two values are labelled "0" and "1".

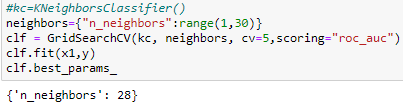


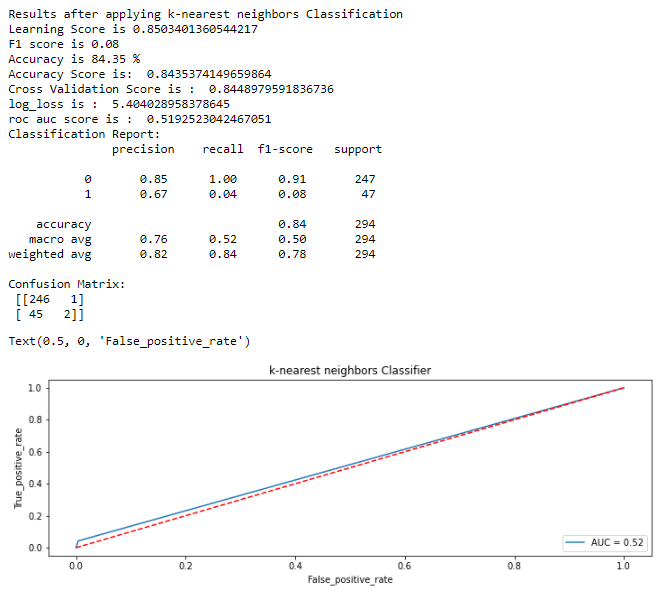
4- **Gradient Boosting Classifier**: - Gradient Boosting Classifier supports both binary and multi-class classification. It builds an additive model in a forward stage-wise fashion.



5- **k-nearest neighbors**: - K nearest neighbors implements learning based on the k nearest neighbors of each query point, where k is an integer value specified by the user.

# Grid Search CV: - It helps to loop through predefined hyperparameters & fit our estimator on our training set so in the end we can select best parameters. We are using it to find n\_neighbors: -

****

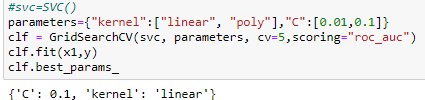
****

6- **Support vector Classification**: - The main goal of this supervised method is to find a function in a multidimensional space that is able to separate training data with known class labels. The classifier separates data points using a hyperplane with the largest amount of margin.

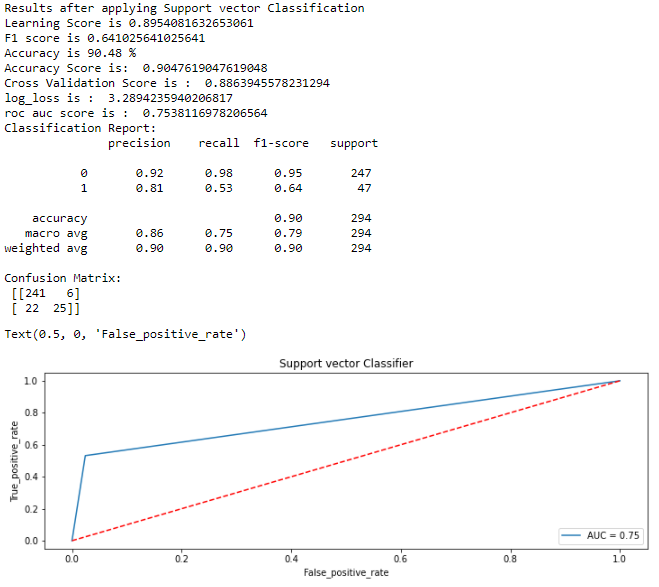
**Hyperparameter Tuning:** - Hyperparameter tuning is the problem of choosing a set of optimal hyperparameters for a learning algorithm. A hyperparameter is a parameter whose value is used to control the learning process. We are using Grid search approach to do this task.

**Grid Search CV:** - It helps to loop through predefined hyperparameters and fit our estimator on our training set. So, in end, we can select the best parameters from the listed hyperparameters.

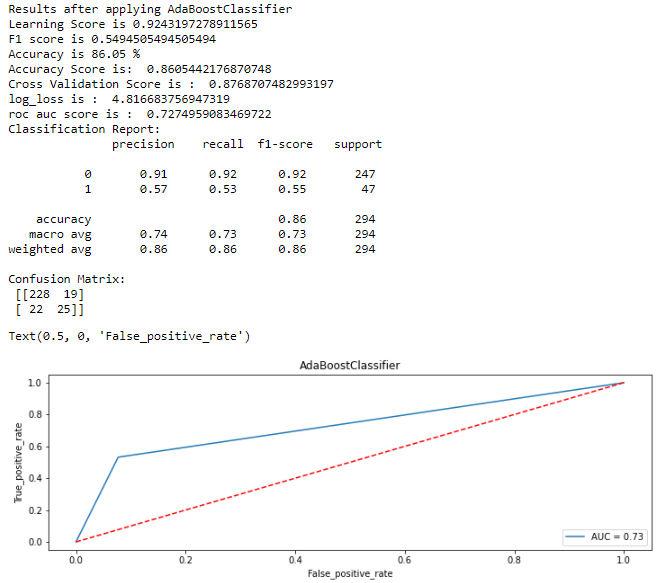
**Using Grid Search CV to find best parameters for Support vector Classification: -**



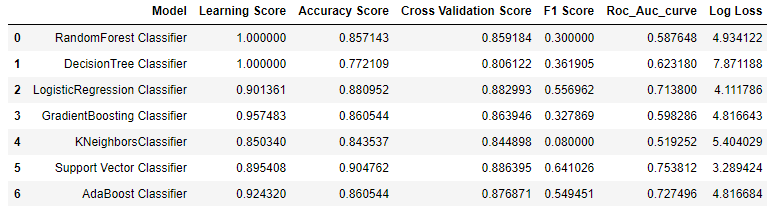
**Applying best parameters on Support Vector Classification: -**



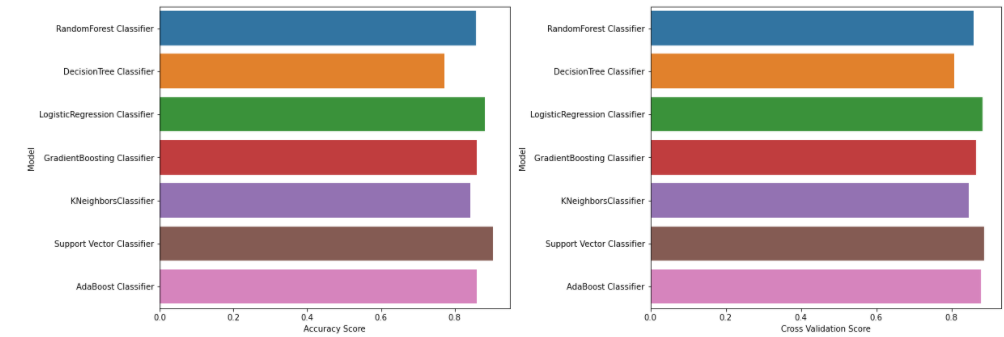
7**- AdaBoost Classification**: - AdaBoost, short for Adaptive Boosting, is a statistical classification meta-algorithm used in conjunction with many other types of learning algorithms to improve performance. The output of the other learning algorithms ('weak learners') is combined into a weighted sum that represents the final output of the boosted classifier.



After applying different classification algorithms on model, we have obtained following scores: -



**Bar plot of Accuracy Score & Cross Validation Score of various models**: -

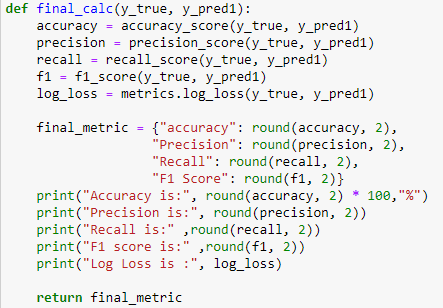


Observations on basis of Scores & Bar Plots: -

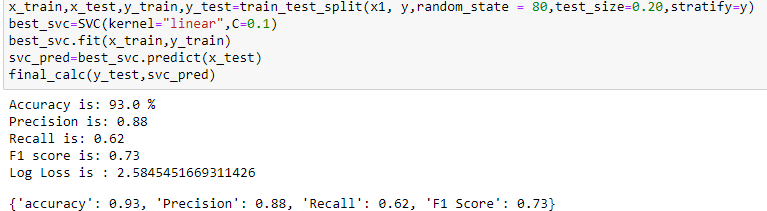
1- After comparing above 7 models on basis of Learning Score, Accuracy Score, Cross Validation Score, F1 Score, Roc\_ Auc\_ curve & Log Loss parameters Support Vector Classifier gives best result.

2- We already did Hyperparameter tuning to find best parameters using Grid Search CV for SVC and we will use them in implementing Support vector classification for final model building. SVC gives best scores when value of **C** parameter is 0.1 & kernel type is linear and random state is 80.

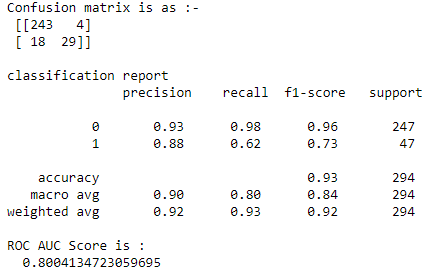
**Function to calculate accuracy, precision, recall and f1 score for our Final Model**: -



**Building Final model with best parameters: -**



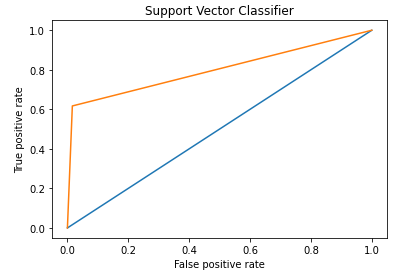
**Confusion Matrix, Classification Report & Roc\_ Auc Score for final Model: -**



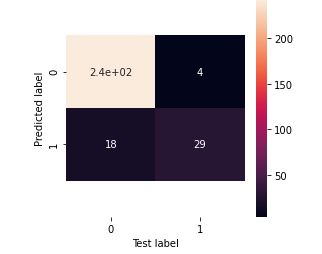
**Evaluating Predictions**: -

**True Positive Rate**: - True positive rate, also referred to sensitivity or recall, is used to measure the percentage of actual positives which are correctly identified.

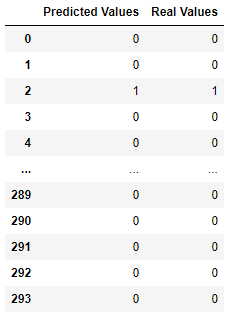
**False Positive Rate**: - The false positive rate is calculated as the ratio between the number of negative events wrongly categorized as positive (false positives) and the total number of actual negative events (regardless of classification).



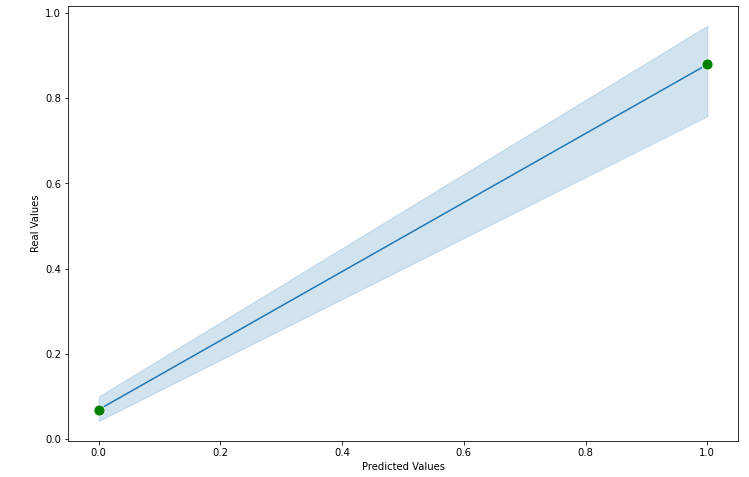
**Confusion matrix using Heatmap for Predicted & Test Labels**: -



**Evaluating Prediction with real values: -**



**Let's plot Graph to check relation between predicted and Real Values using Line Plot: -**

****

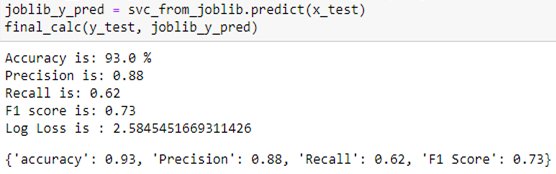
Observation: -

1- Graph is linear and it shows the best relation between predicted and real values.

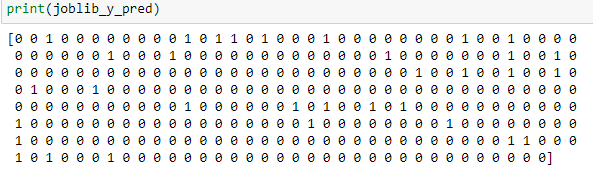
2- Predicted and Real values are very much close to each other.

We have saved our final model using JOBLIB and also save predicted values in CSV file using pandas. Now we are testing the saved model.

1- Loading the saved model and check Classification Evaluation metrics: -



2- Checking Predicted values: -



**Concluding Remarks**

After building, saving, testing, re-loading our final model, Classification Evaluation metrics & on basis of relation between real and predicted values, we have reached on following conclusions: -

1- We spent a great deal of time organizing your data and carefully choose our features.

2- Our Target variable Attrition is categorical in nature so we have used different classification algorithms on dataset and calculated different scores for them.

3- We have used Grid Search CV to find best parameters k-nearest neighbors Classification & Support vector Classification.

4- We have used best parameters obtained by hyperparameter tuning using Grid Search CV method in building of our final model.

5- We have checked Maximum Accuracy Score, best random state, Learning Score, Accuracy Score, Cross Validation Score, F1 Score, Roc\_ Auc\_ score, Log Loss, classification report & confusion matrix for all models.

6- Predicted and Real values are very much close to each other and this shows our model is correct. Also, line plot between them is linear.

7- While using Support vector Classification, learning score, accuracy score, Cross validation score, F1 score &Roc\_ Auc Score are maximum & Log Loss is minimum.

8- To avoid data loss we have not removed outliers but removed skewness very carefully.

9- During pre-processing we have used Standard Scaler to bring features at common scale.

10- We have tested 7 classification algorithms and checked their performance very carefully.

11- Support vector Classifier is best choice for our final model.

12- We have also done data analysis to check impact of input features age, Department, Distance from Home, Job Role, Gender, Job Satisfaction, Marital Status, Overtime, Years with Current Manager, Work Life Balance, Business Travel, Monthly Income, Performance rating and Years Since Last Promotion on target feature Attrition and also calculated points where chances of attrition are high.

13- Strong indicators of **attrition** are: Performance Rating, Monthly Income, job satisfaction, Work Life Balance, Business Travel, Years with Current Manager, Overtime, Age, Distance from Home, Total Working Hours & Years at Company.

14- Meeting of HR team with employees regarding Performance Rating, Job Satisfaction, Work Life balance & Monthly Income will definitely help in reducing the count of attrition.

15- Meeting of Manager with employees regarding Business Travel, Total working Hours and Years with Current Manager topics will definitely reduce the count of job leaving.

16- Overtime is one of the biggest factors behind attrition so HR team & manager should try to discuss the deadline of work & need of overtime with employee. Also, they should provide some extra benefits/perks for overtime so that employee will get positive motivation.

17- We have tried to achieve our goal at best level as result is good as per our model.

**||Thanking You||**