This is the IABAC project given in my datamites institute.

**Requirement**

INX Future Inc, is one of the leading data analytics and automation solutions provider with over 15 years of global business presence. In recent years, the employee performance indexes are not healthy and this has become a growing concern among the top management. The CEO Mr. Brain, decided to initiate a data science project, which analyzes the current employee data and find the core underlying causes of the performance issues. He also expects a clear indicators of non-performing employees, so that any penalization of non-performing employee, if required, may not significantly affect other employee morals.

The following insights are expected from this project:

1. Department wise performances
2. Top 3 Important Factors affecting employee performance
3. A trained model which can predict the employee performance based on factors as inputs. This will be used to hire employees
4. Recommendations to improve the employee performance based on insights from analysis.

# Analysis

Data were analyzed by describing the features present in the data. the features play the bigger part in the analysis. the features tell the relation between the dependent and independent variables. Pandas also help to describe the datasets answering following questions early in our project. The futures present in the data are divided into numerical and categorical data.

### **Categorical Features**

These values classify the samples into sets of similar samples. Within categorical features are the values nominal, ordinal, ratio, or interval based. The categorical features as follows,

* Gender
* EducationBackground
* MaritalStatus
* EmpDepartment
* EmpJobRole
* BusinessTravelFrequency
* EmpEducationLevel
* EmpEnvironmentSatisfaction
* EmpJobInvolvement
* EmpJobLevel
* EmpJobSatisfaction
* OverTime
* EmpRelationshipSatisfaction
* EmpWorkLifeBalance
* Attrition
* PerformanceRating

### **Numerical Features**

These values change from sample to sample. Within numerical features are the values discrete, continuous, or timeseries based. The Numerical Features as follows,

* Age
* DistanceFromHome
* EmpHourlyRate
* NumCompaniesWorked
* EmpLastSalaryHikePercent
* TotalWorkExperienceInYears
* TrainingTimesLastYear
* ExperienceYearsAtThisCompany
* ExperienceYearsInCurrentRole
* YearsSinceLastPromotion
* YearsWithCurrManager

### **Alphanumeric Features**

Numerical, alphanumeric data within same feature. These are candidates for correcting goal. Employee ID number is a mix of numeric and alphanumeric data types.

# Project Summary with Analysis

The given data of employees has the 1200 data to perform a higher level machine learning where it is well structured. The features present in the data are 28 in total. The Shape of the data is 1200x28. The 28 features are classified into quantitative and qualitative where 16 features are qualitative and 11 features are quantitative. The employee ID data is alphanumerical data which doesn't play a role as a relevant feature for performance rating.

The dataset is a complete labelled data and categorical which decides the machine learning algorithm to be used. The important aspects of the data are depending on the correlation of data between features and performance rating. The analysis of the project has gone through the stage of distribution analysis, correlation analysis and analysis by each department to satisfy the project goal.

The main technique used in the preprocessing data using the LabelEncoder method to convert the string-categorical data into numerical data, because, the most of machine learning methods are based on numerical methods where strings are not supportive.

The machine learning model which is used in this project is random forest classifier and Extreme Gradient Boosting classifier(XGBoost) which predicted the nearby higher accuracy of 93%. Since it is categorical labelled data, it has to go through the classifier machine learning techniques which will be suitable for this structured data. The numerical features are the most relevant in the model according to correlation technique.

The overall project was performed and achieved the goals by using the machine learning model and visualization techniques.

# Data Preprocessing

Data Processing is a task of converting data from a given form to a much more usable and desired form i.e. making it more meaningful and informative. Using Machine Learning algorithms, mathematical modelling and statistical knowledge, this entire process can be automated. It has following Steps:-

1. Data Cleaning or Data Wrangling or Data Mugging
2. Label Encoder
3. Feature Selection
4. Feature Scaling

### **1. Data Cleaning or Data Wrangling**

The Data cleaning and wrangling is the part of the Data science project where the workflow the project go through this stage. because the damaged and missing data will lead to the disaster in the accuracy and quality of the model. If the data is already structured and cleaned, there is no need for the data cleaning. In this case, the given data is well structured and cleaned and there are no missing data present in this data.

### **2. Label Encoder**

In machine learning, we usually deal with datasets which contains multiple labels in one or more than one columns. These labels can be in the form of words or numbers. To make the data understandable or in human readable form, the training data is often labeled in words.

* **Label Encoding**:-

It refers to converting the labels into numeric form so as to convert it into the machine-readable form. Machine learning algorithms can then decide in a better way on how those labels must be operated. It is an important pre-processing step for the structured dataset in supervised learning.

* **One Hot Encoding**:-

It refers to splitting the column which contains numerical categorical data to many columns depending on the number of categories present in that column. Each column contains “0” or “1” corresponding to which column it has been placed.

Here I am using Label Encolding to convert Categorical data and alphanumeric data into numerical form.

## **Feature Selection**

In Machine Learning and Statistics feature selection is also known as variable selection, attribute selection and variable subset selection, is a process of selecting best relevant features (variables, predictors) for use in model construction.Feature selection has four different approaches such as:-

* Filter approach
* Wrapper approach
* Embedded approach
* Hybrid approach

Here i am using Wrapper approach method.

This approach has high computational complexity. It uses a learning algorithm to evaluate the accuracy produced by the use of the selected features in classification. Wrapper methods can give high classification accuracy for particular classifiers.

Types of Wrapper Method:-

* Subset Selection (Exhaustive feature Selection)
* Forward Step Selection
* Backward Step Selection (Recursive Feature Elimination)

I am using Feature ranking with Recursive Feature Elimination.

Given an external estimator that assigns weights to features (e.g., the coefficients of a linear model), the goal of recursive feature elimination (RFE) is to select features by recursively considering smaller and smaller sets of features. First, the estimator is trained on the initial set of features and the importance of each feature is obtained either through a coef\_ attribute or through a feature\_importances\_ attribute. Then, the least important features are pruned from current set of features. That procedure is recursively repeated on the pruned set until the desired number of features to select is eventually reached.

And the features that are important for prediction are 'EmpNumber', 'Age', 'EmpDepartment', 'EmpJobRole', 'DistanceFromHome', 'EmpEducationLevel', 'EmpEnvironmentSatisfaction', 'EmpHourlyRate', 'EmpJobSatisfaction', 'NumCompaniesWorked', 'EmpLastSalaryHikePercent', 'EmpRelationshipSatisfaction', 'TotalWorkExperienceInYears', 'TrainingTimesLastYear', 'EmpWorkLifeBalance', 'ExperienceYearsAtThisCompany', 'ExperienceYearsInCurrentRole', 'YearsSinceLastPromotion', 'YearsWithCurrManager'.

### **Feature Scaling**

Feature Scaling is a technique of data preprocessing which is applid to independent variables to normalize the data within particular range

#### **Why and Where to apply feature scaling?**

The dataset contains features that highly vary in magnitudes and units and range. The algorithm which us Euclidean distance are sensitive to magnitudes. Here feature scaling helps to scales all the features equally.Those algorithms are:- 1.KMeans 2.KNearestNeighbors 3.Principle Component Analusis (PCA)

We dont have to use feature scaling for Decision Tree, Random Forest and Xgboost. If we perform feature scaling then also it is not going to make such impact.

Techniques to perform feature Scaling:- (1). Normalization :- It is technique to transfrom the data between 0 and 1. (2).Standardization :- It is very effective technique which rescales a feature value. So that it has distribution with 0 mean and variance equals to 1.

# Visualization

we can able to perform the analysis by the visualisation of the data in two forms here in this project. One is by distributing the data and visualize using the density plotting. The other one is nothing but the correlation method which will visualise the correlation heat map and we can able to achieve the correlation values between the numerical features.

### **Distribution Plot**

In general, one of the first few steps in exploring the data would be to have a rough idea of how the features are distributed with one another. To do so, we shall invoke the familiar kdeplot function from the Seaborn plotting library. The distribution has been done by both numerical and categorical features. it will show the overall idea about the density and majority of data present in a different level.

# Results

### 1. **Department wise performance**

In department wise performance, we have to analyze the data from each department present in the category. The data frame has to be separated or sliced according to department wise. In Employee department feature there are six departments available. The performance analysis by the department as follows,

1. Sales:-

The Performance rating level 3 is more in the sales department. The male performance rating the little bit higher compared to female. The total work experience does not count the performance rating.

1. Human Resources:-

The majority of the employees lying under the level 3 performance. The older people are performing low in this department. The female employees in HR department doing really well in their performance. The total work experience does matter to performance in this department.

1. Development:-

The largest number of employees are level 3 performers. Employees of all age are performing at the level of 3 only. The gender-based performance is nearly same for both.

1. Data Science:-

The highest average of level 3 performance is in data science department. Data science is the only department where less number of level 2 performers. The overall performance is higher compared to all departments. The age does not count as an important factor in their performance. Male employees are doing good in this department. Same like HR, the number of work experience does matter.

1. Research & Development:-

The age factor is not deviating from the level of performance here where different employees with different age are there in every level of performance. The R&D has the good female employees in their performance.

1. Finance:-

The finance department performance is exponentially decreasing when age increases. The male employees are doing good. The experience factor is inversely relating to the performance level.

### **2. The Top 3 important features affecting the employee performance**

The Random Forest classifier and XGBoost classifier in Sklearn also contains a very convenient and most useful attribute feature importances which tells us which features within our dataset has been given most importance through the ML. The top three important features affecting the performance rating are ordered with their importance level as follows,

1. Employment Environment Satisfaction
2. Employee Salary Hike Percentage
3. Years Since the last Promotion

### **3. A Trained model which can predict the employee performance**

The trained model is created using the machine learning algorithm as follows with the accuracy score,

1. Random Forest classifier using Grid Search :- 93.0% accuracy.

Random Forest Classifier accuracy is changing between 91% to 93%. For better accuracy go to xgboost.

1. Extreme Gradient Boosting Claasifier(XGBoost) :- 93.8% accuracy

### **4. Recommendations to improve the employee performance**

1. The overall employee performance can be achieved by employee environment satisfaction. The company needs to focus more on the employee environment.

2. The salary hike will give the boost to the employees to perform well financially and psychologically.

3. The promotion will help the employees to achieve more performance by giving the chance to be more responsible and leadership qualities.

4. The experience years in current role need to be revised while offering the employment to the new employees.

5. Employee's work-life balance affects the performance rating.

6. While recruiting for HR, consider the female candidates where they perform well compared to male.

7. The development and data science department is having an overall higher performance comparing to rest of the departments.