# REAL-TIME BUSINESS MONITORING SYSTEM FOR CORPORATES TO IDENTIFY KEY PERFORMANCE INDICATORS (KPI'S) OF THE EMPLOYEES TO PREDICT FUTURE EMPLOYEE'S PROMOTIONS.

DEPARTMENT OF BUSINESS INTELLIGENCE SYSTEMS INFRASTRUCTURE
DATA VISUALIZATION TOPICS

TEAM MEMBERS:

DHRUV PARMAR

A PROJECT UNDER THE GUIDANCE OF
PROFESSOR MR. SIJU PHILIPH

# Introduction & Objectives:

In a service-based company, the HR department aims to enhance its promotion strategy by using Previous Promotion Cycle's Data to identify the most suitable candidates for promotion for the upcoming Promotion Cycle.

The company aims to optimize its promotion strategy, Ensuring that deserving candidates are identified promptly and provided with growth opportunities that are most likely to succeed in higher roles. All the while maximizing overall Organizational Performance and Employee satisfaction.

Enhance transparency in the promotion process, fostering a fair and merit-based environment within the organization. In the process, Gain several Insights about various Factors affecting the Organizational Performance in different Divisions.


## Abstract:

This report delves into a comprehensive analysis of diverse Key Performance Indicators (KPIs) utilizing Exploratory Data Analysis (EDA) techniques within Microsoft Excel and Data Visualization tools via Power BI. Prediction Modelling and Correlation Analysis were done by using Python Programs with the help of Jupyter Notebook.

Our study proposes a strategic solution aimed at streamlining the workflow of the HR Team through the development of a predictive model. This model facilitates the identification of the most suitable candidates for promotion, There by optimizing the promotion process.

Furthermore, our approach encompasses the integration of various parameters linked to Organizational Performance, providing the HR Team with insightful dashboards to effectively monitor and manage these metrics. The combined use of predictive analytics, data visualization, and dashboard tracking presents a holistic framework for enhancing HR operations while empowering informed decision-making processes within the organization.


## Data Collection:

Effective data collection and preparation is a crucial foundation for promotional decisions, development of the employee and overall success management processes within the organization.
During the data collection phase, we collected information from a single source called as "Kaggle" which had two datasets namely:

- Train Dataset
- Test Dataset.

The research methodology followed industry standards and we found the perfect dataset for the project. Although we came across few challenges like null values in the dataset, different formats etc.
To alleviate these challenges, we came up with the strategy to clean the data using Tableau Prep Builder.

Link: https://www.kaggle.com/datasets/bhrt97/hr-analytics-classification

## Data Cleaning:

Data cleaning is a careful process or you can say meticulous step which is aimed at preparing for "Data Analysis". The initial review identified issues like null values, outliers and inconsistencies.
To remove null values we came with the solution to use Tableau Prep Builder. We fed the .csv file to Tableau Prep to clean the data. Once feeding of the data was done, we cleaned the data easily.
Noteworthy challenges we faced during cleaning was to overcome the inconsistency by getting the solution. So one of the inconsistency was re-naming each attribute in a certain way which can be understood, which

was time-consuming. These hurdles were overcome through by using Tableau Prep by using clean step and renaming the attributes according to our will so that it is formatted properly.

**Attributes and their significance:**

| Attributes | Description | Data-types |
|---|---|---|
| Employee ID | Unique identifier to recognize the employee | Numeric |
| Department | A category to which the employee belongs to | String |
| Region | A place to which an employee belongs to | Alphanumeric |
| Education | Employee's educational background | String |
| Gender | Gender of the employee | String |
| Recruitment Channel | Source from where the employee was hired | String |
| No Of Trainings | No of trainings that employee went through for the process | Numeric |
| Age | Age of the employee | Floating Numeric |
| Previous Year Rating | Employee's previous year rating | Numeric |
| Length Of Service | Number of year employees with the company | Floating Numeric |
| KPIs_met >80% | Role based KPI, A target/goal to be achieved/ or achieved by the employee | Boolean |
| Awards won | Awards handed out to employee by the company | Numeric |
| Avg Training Score | Mean Score of the employee | Numeric |
| Is-Promoted | The outcome whether the employee is promoted or not | Boolean |

# Approach:

# Data Analysis:

The Initial Analysis was done using Microsoft Excel. A total of 54808 records were present in the Dataset. Out of them 4668 (8.517%) Employees were Promoted. Each Attribute was segregated to gain a Comprehensive Understanding.

**In the attribute "Recruitment Channel" the No. Of Employees promoted are:**

- Under Referred, 92 employees are promoted i.e., 1.97% of employees that are promoted.
- Under Others, 2651 employees are promoted i.e., 56.79% of employees that are promoted.
- Under Sourced, 1925 employees are promoted i.e, 41.238% of employees that are promoted.

**In the attribute " KPI's Met >80%" the No. Of Employees promoted are:**

- Under Yes , 3262 employees are promoted i.e., 69.88% of employees that are promoted.
- Under No, 1406 employees are promoted i.e., 30.12% of employees that are promoted.

**In the attribute " Number Of Trainings" the No. Of Employees promoted are:**

- For 1 , 3910 employees are promoted i.e., 83.76% of employees that are promoted.
- For 2, 605 employees are promoted i.e., 12.96% of employees that are promoted.
- For 3, 122 employees are promoted i.e., 2.6% of employees that are promoted.
- For 4 or more, 31 employees are promoted i.e., 0.68% of employees that are promoted.

**In the attribute " Age" the No. Of Employees promoted are:**

- From 20 to 30 Years , 1487 employees are promoted i.e., 31.855% of employees that are promoted.
- From 30 to 40 Years , 2436 employees are promoted i.e., 52.185% of employees that are promoted.
- From 40 to 50 Years, 575 employees are promoted i.e., 12.317% of employees that are promoted.
- For More than 50 Years, 170 employees are promoted i.e., 3.641% of employees that are promoted.

**In the attribute " Length Of Service" the No. Of Employees promoted are:**

- For Less than 5 Years, 2654 employees are promoted i.e., 56.85 % of employees that are promoted.
- From 5 to 10 Years, 1567 employees are promoted i.e., 33.569% of employees that are promoted.
- From 10 to 15 Years, 287 employees are promoted i.e., 6.14% of employees that are promoted.
- For More than 15 Years, 162 employees are promoted i.e., 3.47% of employees that are promoted.

**In the attribute " Average Training Score" the No. Of Employees promoted are:**

- For Less than 60, 1980 employees are promoted i.e., 42.416 % of employees that are promoted.
- From 60 to 70, 861 employees are promoted i.e., 18.445% of employees that are promoted.
- From 70 to 80, 693 employees are promoted i.e., 14.845% of employees that are promoted.
- From 80 to 90, 1055 employees are promoted i.e., 22.6% of employees that are promoted.
- For More than 90, 79 employees are promoted i.e., 1.692% of employees that are promoted.

**Correlations:**
From the Dataset we were able to infer the following correlations

- Age is sort of INVERSELY proportional to Promotion as most of the young employees get promoted in the dataset.
- No of Trainings is INVERSELY proportional to Promotion i.e, less no. of trainings the higher chance of an Employee getting promoted.
- Length of service does play an important role. As per the Data to getting promoted as one thinks but you can say it is almost INVERSELY proportional to Promotion.
- KPI's met>= 80% is Directly proportional to Promotion as the employee has to reach the criterion of >= 80% to get a promotion.
- Awards won is INVERSELY proportional to promotion.

**Key Performance Indicators (KPIs):**
Below are the KPI's that play a major role in Promotions of Employees.

- Number Of Trainings
- KPI's Met >80%
- Length Of Service
- Age

## Visualization:

With the help of Power BI, extracting meaningful insights and patterns was speed and accurate. Variety of charts were used in the visuals / dashboard like Histogram / Clustered Column Chart, Pie Chart, Line Graph, Tree Map, Stacked Bar Chart, Table Chart and Cards. Significance of all the charts are listed below as to Why selected these specific charts?

1) **Histogram / Clustered Column Chart:** Histograms are chosen for data visualization for several reasons
   a) **Distribution of Data:** Histograms are particularly useful for visualizing the distribution of a dataset. They provide a way to see the underlying shape of the data, whether it's symmetric, skewed, unimodal, bimodal, or exhibits other patterns.
   b) **Frequency Counts:** Histograms display the frequency or count of data points within predefined intervals or bins. This helps in understanding the concentration of values within specific ranges and identifying peaks or gaps in the data.
   c) **Identifying Central Tendency:** Central tendency measures like the mean, median, and mode can be identified by observing the central region of a histogram. This is especially valuable for understanding where the bulk of the data lies.
   d) **Decision-Making Support:** In situations where decision-making is based on the distribution of a variable, histograms provide a clear visual summary that aids in making informed decisions

   **Insights Achieved:**

   - This chart is applied in **'Dashboard 1'** for variables **'Promotion'** & **'No Of Trainings'** which shows Promotion of employee's based on different no of trainings taken. Here the key insight we notice is people with trainings more than 6 have **NOT** been promoted which proves employees with least number of trainings have been preferred to be promoted.
   - It is also used in **'Dashboard 2'** for variables **'Department'** & **'Promotion'** which shows total employees promoted across different departments with count, where we see highest number of employees promoted are in **'Sales & Marketing'** and lowest in **'Legal'**
   - At last it has also been used in **'Dashboard 3'** for variables **'Employee ID'** & **'Length Of Service'** Where we notice a pattern where majority of people provide service to the company for 3 years which decreases down the line.

2) **Pie Chart:** Pie charts are chosen for data visualization in specific situations due to their unique characteristics and advantages.
   a) **Part-to-Whole Relationship:** Pie charts are ideal for representing the part-to-whole relationship, where each slice of the pie represents a proportion of the whole. This is useful when you want to illustrate how individual components contribute to the total.
   b) **Simple and Intuitive:** Pie charts are simple and intuitive, making them easy to understand for a wide range of audiences, including those with limited experience in data analysis. The circular shape and division into slices provide a clear visual representation.
   c) **Percentage Representation:** Each segment in a pie chart represents a percentage of the whole, making it easy to see the relative size of each category or component at a glance. This can be useful for conveying the distribution of a categorical variable.
   d) **Aesthetically Pleasing:** Pie charts are often considered aesthetically pleasing and can be visually engaging. This makes them suitable for presentations and reports where a visually appealing representation is desired.

   **Insights Achieved:**
   - This chart is applied in '**Dashboard 1'** for variables **'Promotions'** & **'Gender'** which shows promotions of total employees divided in M:F ratio, which is responsive to all the visuals in the dashboard like Employee's Promoted By Region.
   - Pie chart has also been used in **'Dashboard 3'** with variables **'Length Of Service'** & **'Promotion'.** Where we notice a pattern where promotions increase as length of service increases and maximum people have been promoted with length of service 3 after which the promotions decrease, which proves promotions stop after specific years of experience as it increases.

3) **Line Graph:** Line graphs are a popular choice for data visualization in specific scenarios due to their ability to show trends, patterns, and relationships over a continuous or sequential range.
   a) **Trend Analysis:** Line graphs are particularly effective for illustrating trends in data over time. By connecting data points with lines, they provide a visual representation of the overall direction of the data, whether it's increasing, decreasing, or remaining relatively constant.
   b) **Time Series Data:** Line graphs are well-suited for visualizing time series data, where the x-axis represents a continuous time interval. This makes them ideal for showing how a variable changes over days, months, years, etc.
   c) **Forecasting:** Line graphs can be used to visually project trends into the future, aiding in forecasting and making predictions based on historical data.
   d) **Easy Interpretation:** Line graphs are generally easy to interpret, making them accessible to a wide audience. This is especially important when presenting data to individuals who may not have a strong background in statistics.

   **Insights Achieved:**
   - This chart has been used in **'Dashboard 2'** with variables **'Age'** and **'Promotion'** where we clearly see that promotions increase from the age of 20 and maximum promotions happen at the age of 30 which is the peak of ones career generally after which the promotions decrease, which means promotions nearly stop as the experience and age increases.

4) **Tree Map:** Tree maps are a type of data visualization that represents hierarchical data structures through nested rectangles.
   a) **Hierarchical Data Representation:** Tree maps are designed to visualize hierarchical structures where each level of the hierarchy is represented by nested rectangles. This makes them suitable for displaying categorized data in a structured and organized manner.
   b) **Proportional Representation:** The size of each rectangle in a tree map represents a quantitative value, allowing for a proportional representation of data. This helps users quickly grasp the relative contribution of each category or sub-category.

c) **Efficient Use of Space:** Tree maps efficiently use space to display a large amount of information in a compact format. This is especially advantageous when dealing with datasets that have many categories or sub-categories.
d) **Colour Encoding:** Colours can be used to encode additional information in tree maps, such as highlighting specific categories, indicating values, or conveying qualitative information. This enhances the interpretability of the visualization.

**Insights Achieved:**

- Tree Map is used in **'Dashboard 2'** with variables **'Department'** & **'Role Based KPI'** which depicts KPI'S achieved by department. The visual shows number of employees achieved KPI'S in Sales & Marketing are the highest with R&D as the lowest.

5) **Stacked Bar Chart:** Stacked bar charts are chosen for data visualization in certain scenarios due to their ability to represent the composition of a whole across different categories and subcategories.
a) **Comparison Across Categories:** Stacked bar charts make it easy to compare the total sizes of different categories as well as the relative proportions of subcategories within each category. This visual comparison aids in understanding the distribution of values.
b) **Cumulative Representation:** The cumulative nature of stacked bars helps in understanding the cumulative impact of each category. Viewers can see how the total changes as they move along the horizontal axis.
c) **Trend Analysis:** Stacked bar charts are useful for trend analysis, especially when you want to observe changes in the composition of categories over time or across different conditions.

**Insights Achieved:**

- This chart is used in **'Dashboard 1 & 2'** with variables **'Gender'**, '**Promotion'** and **'Region'**, which depicts promotion of male and female employees across different regions where we see Region 2 has the highest number of promotions and with lowest is Region 18, which is integrated in Dashboard 1 and 2 with various parameters acting responsive.

6) **Table Chart:** A table chart is a visual representation of data in tabular form, where information is organized into rows and columns. Unlike many other types of charts that use graphical elements to represent data, a table chart presents the data in a structured, text-based format.
a) **Structure:** A table chart consists of rows and columns, with each row representing a record or observation, and each column representing a variable or attribute.
b) **Data Presentation:** Data in a table chart is presented in a clear and organized manner. Each cell within the table contains a data value, and the arrangement of rows and columns helps users easily locate and compare information.
c) **Numerical and Textual Data:** Table charts can accommodate both numerical and textual data. Numeric values, text, or a combination of both can be included in the cells of the table.
d) **Sorting and Filtering:** Table charts often come with features for sorting and filtering data. This allows users to organize the data based on specific criteria, making it easier to find relevant information.

**Insights:**

- This chart is used in **'Dashboard 1 & 2'** with variables **'Gender'**, **'Employee ID'** and **'Department'** in 1st and **'Gender'** and **'Employee ID'** in 2nd, which shows Employee ID across different departments and their gender to identify individual employees statistics.

7) **Cards:** In Power BI, cards are a type of visualization that displays a single, key metric or value prominently. Cards are useful for emphasizing and summarizing important information, allowing users to focus on specific key performance indicators (KPIs) or metrics.
a) **Single Metric Focus:** Cards are designed to showcase a single metric or value. This helps in highlighting key numbers, such as total sales, revenue, or any other important performance indicator.
b) **Text and Numeric Representation:** A card can display both textual and numeric representations of the metric. You can customize the formatting to make the presentation clear and meaningful, including options for currency symbols, percentage formats, and more.

c) **Interactivity:** Cards can be interactive, allowing users to click on them to drill down into more detailed information. This interactivity can be set up based on your specific data model and requirements.

**Insights:**

- This chart is used in **'Dashboard 3'** with variables **'Employee ID'** & **'Promotion',** which shows total number of employees in the organizations and promoted employees out of them.

**KPI'S Identified (Variables):**

- No Of Trainings
- Age
- KPI'S met >80%
- Length Of Service

**Note: All the above parameters / KPI'S and insights / patterns identified majorly help to answer question or predict. Who should we Promote?** OR **Who can be Promoted?**

# Predictive Modelling: Here We answer Who will be Promoted in Future?

**Preparation for modelling:**

### 1. Dropping Unwanted Features

| Employee id |
|---|
| recruitment_channel |
| region |

### 2. Handling Null Values

```
In [7]: df.isnull().sum()

Out[7]: department                0
        education              2409
        gender                    0
        no_of_trainings           0
        age                       0
        previous_year_rating   4124
        length_of_service         0
        KPIs_met >80%             0
        awards_won?               0
        avg_training_score        0
        is_promoted               0
        dtype: int64
```

All these respective null values have been dropped as per the analysis.

### 3. Transforming Features

```
#Early Career = 1
#Mid-Career = 2
#Established Professionals = 3
#Experienced Leaders = 4

age_catagory = []
for row in df['age']:
```

```
        if row < 30:   age_catagory.append('1')
        elif row < 40:  age_catagory.append('2')
        elif row <50 : age_catagory.append('3')

        else:          age_catagory.append('4')
    df['age_catagory'] = age_catagory
    df.tail()
```

```
#poor = 1
#Below Average = 2
#Average = 3
#Above Average = 4
#Excellent = 5


avg_training_score_catagory = []
for row in df['avg_training_score']:
    if row < 60:   avg_training_score_catagory.append('1')
    elif row < 70:  avg_training_score_catagory.append('2')
    elif row < 80:  avg_training_score_catagory.append('3')
    elif row < 90:  avg_training_score_catagory.append('4')

    else:          avg_training_score_catagory.append('5')
df['avg_training_score_catagory'] = avg_training_score_catagory
df.head()
```

```
#Novice = 1
#Intermediate = 2
#Advanced = 3
#Expert = 4


length_of_service_catagory = []
for row in df['length_of_service']:
    if row <= 5:   length_of_service_catagory.append('1')
    elif row <= 10:  length_of_service_catagory.append('2')
    elif row <= 15:  length_of_service_catagory.append('3')

    else:          length_of_service_catagory.append('4')
df['length_of_service_catagory'] = length_of_service_catagory
df.head()
```

Here, three new feature categories(age_catagory, avg_training_score_catagory, length_of_service_catagory) were created based on current features(age, avg_training_score, length_of_service) and that features were dropped.

## 4. Splitting Data

**Spliting the dataset into train and test set (MAIN DATASET)**

```
In [24]: import matplotlib.pyplot as plt
         from sklearn.model_selection import train_test_split

In [25]: X = df[[ 'department_encoded', 'education_encoded', 'gender_encoded', 'no_of_trainings', 'age_catagory', 'previous_year_rating',
                'length_of_service_catagory', 'KPIs_met >80%', 'awards_won?',
                'avg_training_score_catagory' ]]

In [26]: y = df[['is_promoted']]

In [27]: X = pd.DataFrame(X)

In [28]: y = pd.DataFrame(y)

In [29]: X_train, X_test, y_train, y_test = train_test_split(X, y, random_state=0, train_size = 0.75)
```

Data frame was splitted into X(Features) and y(Predicted value). Train and test splitted into 75:25 ratio.

## 5. Standardize Data

**standerdize the data**

```
In [32]: #Now standerdize the value tp make them between o to 1
         #using standard scaler or minmax scaler
         #from sklearn.preprocessing import MinMaxScaler
         #however i am gonna use standard scaler
         from sklearn.preprocessing import StandardScaler

In [33]: #first fit then transform train data
         scaler=StandardScaler()
         scaler.fit(X_train)

Out[33]: StandardScaler()

In [34]: scaled_train_data=scaler.transform(X_train)

In [35]: scaled_train_data
```

## 6. PCA

Principal Component Analysis (PCA) is a dimensionality reduction technique used in machine learning and statistics. Its main objective is to transform a high-dimensional dataset into a lower-dimensional space while retaining as much of the original variability or information as possible.

In [39]:
```
# import PCA from sklearn.decomposiotion
from sklearn.decomposition import PCA
```

In [40]:
```
#define how many of dimention you want
pca=PCA(n_components=5)
```

In [41]:
```
#again fit and transform the scaled data
pca.fit(scaled_train_data)
```

Out[41]:
```
PCA(n_components=5)
```

In [42]:
```
x_train_pca=pca.transform(scaled_train_data)
```

In [43]:
```
# do the same for test data
#fit and transform test data
pca.fit(scaled_test_data)
x_test_pca=pca.transform(scaled_test_data)
```

In [44]:
```
#check the dimensions
scaled_train_data.shape
```

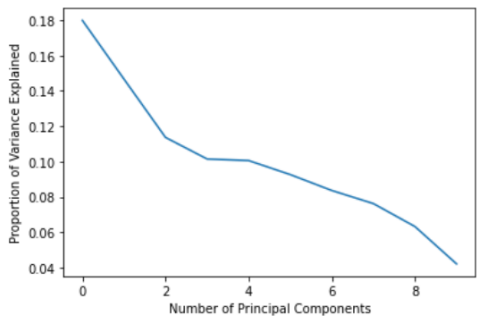
Out[44]:
```
(36495, 10)
```

In [45]:
```
x_train_pca.shape
```

Out[45]:
```
(36495, 5)
```

**Generate Scree plot to find the best number of principal components (denote it as d).**

In [46]:
```
pca = PCA()
pca.fit(scaled_train_data)

# Generate the scree plot
plt.plot(pca.explained_variance_ratio_)
plt.xlabel('Number of Principal Components')
plt.ylabel('Proportion of Variance Explained')
plt.show()
```



## 7. LDA

**Let's start for LDA**

In [51]:
```
from sklearn.discriminant_analysis import LinearDiscriminantAnalysis as LDA

# Apply FDA to reduce dimensionality
lda = LDA(n_components= 1)
x_train_lda = lda.fit_transform(scaled_train_data , y_train)
x_test_lda = lda.fit_transform(scaled_test_data , y_test)
```

```
c:\users\katha\appdata\local\programs\python\python37\lib\site-packages\sklearn\utils\validation.py:993: DataConversionWarning:
A column-vector y was passed when a 1d array was expected. Please change the shape of y to (n_samples, ), for example using rav
el().
  y = column_or_1d(y, warn=True)
c:\users\katha\appdata\local\programs\python\python37\lib\site-packages\sklearn\utils\validation.py:993: DataConversionWarning:
A column-vector y was passed when a 1d array was expected. Please change the shape of y to (n_samples, ), for example using rav
el().
  y = column_or_1d(y, warn=True)
```

In [52]:
```
#check the dimensions
scaled_train_data.shape
```

Out[52]: (36495, 10)

In [53]:
```
x_train_lda.shape
```

Out[53]: (36495, 1)

# Model Training:

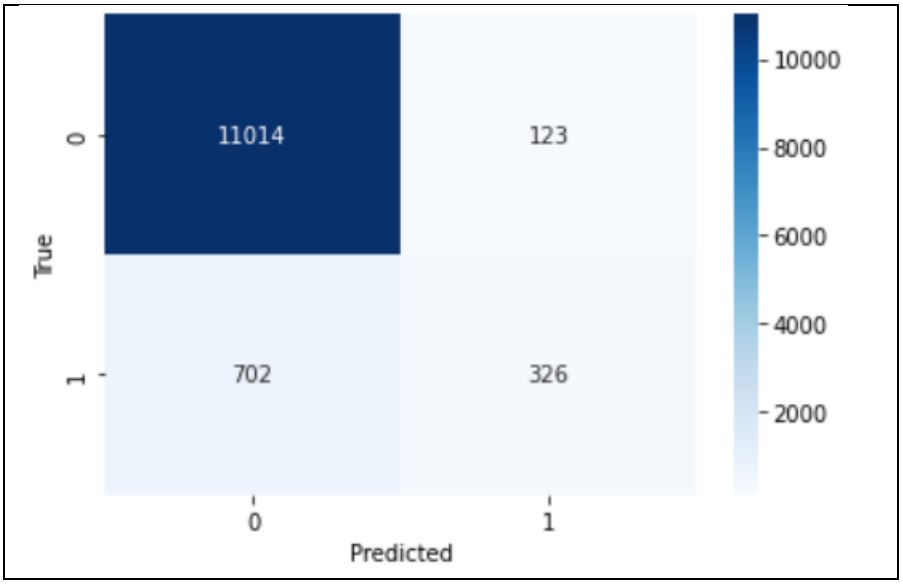
### 1)Random Forest:

Random Forest is a versatile and powerful machine learning algorithm that belongs to the ensemble learning family. It's widely employed for both classification and regression tasks, renowned for its robustness and ability to handle complex datasets.

```
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import confusion_matrix, accuracy_score

# Create an instance of the RandomForestClassifier
clf = RandomForestClassifier()

# Fit the classifier to the train data
clf.fit(scaled_train_data, y_train)

# Make predictions on the test data
y_pred = clf.predict(scaled_test_data)
```

This model has been used for all 3 prepared data.

**Confusion Matrix:**

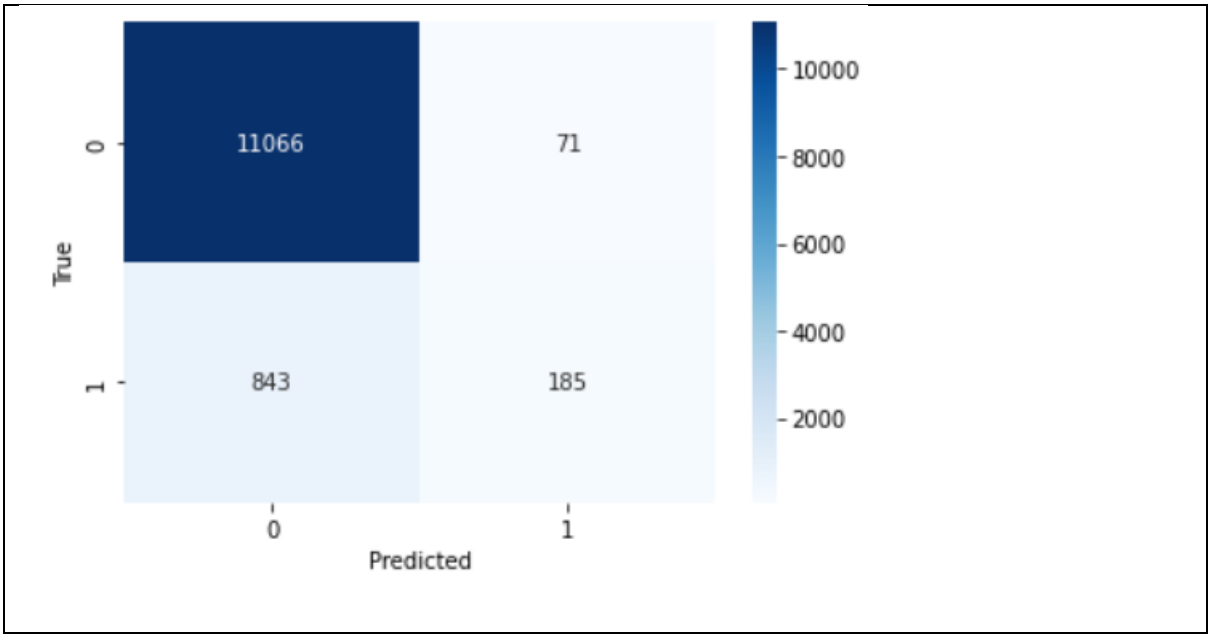



**2) Multilayer Perceptron:**

Multilayer Perceptron (MLP) is a fundamental and widely used type of artificial neural network, playing a pivotal role in various machine learning applications.

```
from sklearn.neural_network import MLPClassifier

# Create an instance of MLPClassifier
mlp = MLPClassifier(hidden_layer_sizes=(50,), max_iter=10, alpha=1e-4,
                    solver='sgd', verbose=10, tol=1e-4, random_state=1,
                    learning_rate_init=.1)

# Fit the model on the training data
mlp.fit(scaled_train_data, y_train)

# Use the model to predict on the test data
y_predN1 = mlp.predict(scaled_test_data)

# Evaluate the model's performance
print("Accuracy:", accuracy_score(y_test, y_predN1))
```

**Confusion Matrix:**




**Results:**

| Model / Data | Standardize Data | PCA | LDA |
|---|---|---|---|
| Random Forest | 93.31% | 87.92% | 85.60% |
| MLP | 92.49% | 89.66% | 91.52% |

## Conclusions:

- **We can say that No. Of Trainings and Role Based KPI's met are the most important factors in determining the chances of Promotion.**
- **The least number of trainings you take, the higher chances of Promotion.**
- **Age also plays an Important role, As per the Analysis an Employee should be Upskilled upto a certain extent to Improve their chances of promotion. The most chances for a Promotion is for Employees between 25 to 35 Years of Age.**
- **The Best Predictive Model upon analysis is found to be Random Forest Model with 93.31%.**
- **Awards Won by Employees didn't account to much of Promotions. We could almost say that they are Consolation Prizes.**