RAJALAKSHMI ENGINEERING COLLEGE
An AUTONOMOUS Institution
Affiliated to ANNA UNIVERSITY, Chennai

AI19P71
Data Visualization Using Python

PREDICTIVE ANALYTICS FOR EMPLOYEE ATTRITION AND PERFORMANCE ENHANCEMENT

Ranjith Kumar S (211501077)
Sanjay S (211501088)
Santhosh H (211501090)

Problem Statement

- Organizations today face the dual challenge of mitigating employee attrition and enhancing workforce performance in an increasingly competitive and dynamic business environment.

- High attrition rates not only lead to substantial recruitment and training costs but also disrupt workplace morale and productivity.

- Identifying and nurturing high-performing employees is critical for driving organizational success, yet many companies struggle to leverage their HR data effectively.

Dataset Source and Structure

Dataset Source	
No of Features	35
No of Records	1470

Dataset Feature Description

1. **Attrition:** Indicates whether the employee has left the organization (Yes/No), serving as the primary target variable for predicting employee turnover.
2. **JobSatisfaction:** Measures the employee's satisfaction with their role and responsibilities, directly linked to attrition trends.
3. **DistanceFromHome:** Represents the distance between the employee's home and workplace, which could influence work-life balance and attrition rates.

Dataset Feature Description

4. **Education:** Captures the highest education level attained by the employee, reflecting potential career aspirations and professional growth patterns.
5. **MonthlyIncome:** Represents the employee’s monthly income, a key financial factor potentially impacting job satisfaction and loyalty.
6. **JobRole:** Specifies the employee's role within the organization, influencing both job responsibilities and attrition likelihood.
7. **MonthlyIncome:** Represents the employee’s monthly income, a key financial factor potentially impacting job satisfaction and loyalty.

Data Acquisition and Cleaning

CODE

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import hvplot.pandas

%matplotlib inline
sns.set_style("whitegrid")
plt.style.use("fivethirtyeight")

pd.set_option("display.float_format", "{:.2f}".format)
pd.set_option("display.max_columns", 80)
pd.set_option("display.max_rows", 80)
```

```
df = pd.read_csv("https://raw.githubusercontent.com/Santhosh-H/Predictive-Analytics
df.head()
# Display dataset information
print("Dataset Shape:", df.shape)
print("Dataset Columns:\n", df.columns)
print("\nInitial Data Preview:")
print(df.head())
```

OUTPUT

```
Dataset Shape: (1470, 35)
Dataset Columns:
 Index(['Age', 'Attrition', 'BusinessTravel', 'DailyRate', 'Department',
       'DistanceFromHome', 'Education', 'EducationField', 'EmployeeCount',
       'EmployeeNumber', 'EnvironmentSatisfaction', 'Gender', 'HourlyRate',
       'JobInvolvement', 'JobLevel', 'JobRole', 'JobSatisfaction',
       'MaritalStatus', 'MonthlyIncome', 'MonthlyRate', 'NumCompaniesWorked',
       'Over18', 'OverTime', 'PercentSalaryHike', 'PerformanceRating',
       'RelationshipSatisfaction', 'StandardHours', 'StockOptionLevel',
       'TotalWorkingYears', 'TrainingTimesLastYear', 'WorkLifeBalance',
       'YearsAtCompany', 'YearsInCurrentRole', 'YearsSinceLastPromotion',
       'YearsWithCurrManager'],
      dtype='object')

Initial Data Preview:
   Age Attrition     BusinessTravel  DailyRate              Department  \
0   41       Yes      Travel_Rarely       1102                   Sales
1   49        No  Travel_Frequently        279  Research & Development
2   37       Yes      Travel_Rarely       1373  Research & Development
3   33        No  Travel_Frequently       1392  Research & Development
4   27        No      Travel_Rarely        591  Research & Development

   DistanceFromHome  Education EducationField  EmployeeCount  EmployeeNumber  \
0                 1          2  Life Sciences              1               1
1                 8          1  Life Sciences              1               2
2                 2          2          Other              1               4
3                 3          4  Life Sciences              1               5
4                 2          1        Medical              1               7
```

Data Acquisition and Cleaning

CODE

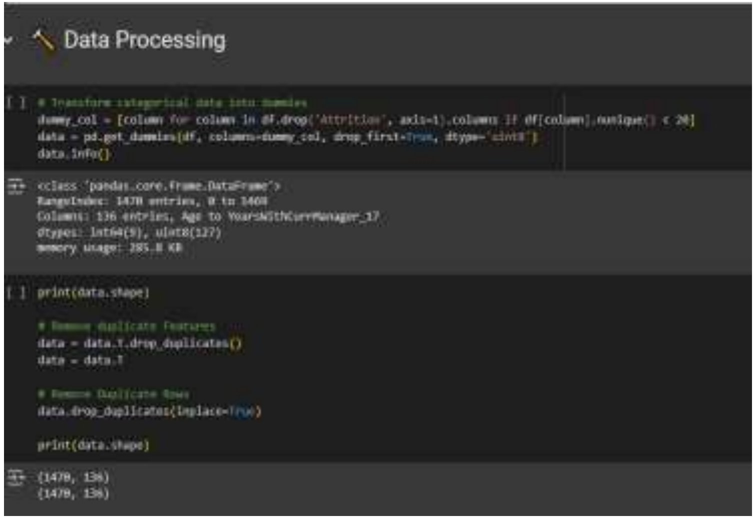
```python
print("\nChecking for Missing Values:")
print(df.isnull().sum())

# Filling missing values
df.fillna(method='ffill', inplace=True)
df.fillna(method='bfill', inplace=True)

print("\nMissing Values After Cleaning:")
print(df.isnull().sum())

# Encoding Categorical Variables
print("\nIdentifying Categorical Variables:")
categorical_columns = df.select_dtypes(include=['object']).columns
print(categorical_columns)

df = pd.get_dummies(df, columns=categorical_columns, drop_first=True)

# Outlier Detection and Handling
print("\nDetecting Outliers:")
numerical_columns = df.select_dtypes(include=['float64', 'int64']).columns
for col in numerical_columns:
    Q1 = df[col].quantile(0.25)
    Q3 = df[col].quantile(0.75)
    IQR = Q3 - Q1
    lower_bound = Q1 - 1.5 * IQR
    upper_bound = Q3 + 1.5 * IQR
    df = df[(df[col] >= lower_bound) & (df[col] <= upper_bound)]

print("\nDataset Shape After Outlier Removal:", df.shape)

# Checking for duplicates
print("\nChecking for Duplicate Rows:")
duplicates = df.duplicated().sum()
print("Number of duplicate rows:", duplicates)

df = df.drop_duplicates()

print("\nCleaned Dataset Preview:")
print(df.head())
```

OUTPUT

```
Checking for Missing Values:
Age                         0
Attrition                   0
BusinessTravel              0
DailyRate                   0
Department                  0
DistanceFromHome            0
Education                   0
EducationField              0
EmployeeCount               0
EmployeeNumber              0
EnvironmentSatisfaction     0
Gender                      0
HourlyRate                  0
JobInvolvement              0
JobLevel                    0
JobRole                     0
JobSatisfaction             0
MaritalStatus               0
MonthlyIncome               0
MonthlyRate                 0
NumCompaniesWorked          0
Over18                      0
OverTime                    0
PercentSalaryHike           0
PerformanceRating           0
RelationshipSatisfaction    0
StandardHours               0
StockOptionLevel            0
TotalWorkingYears           0
TrainingTimesLastYear       0
WorkLifeBalance             0
YearsAtCompany              0
YearsInCurrentRole          0
YearsSinceLastPromotion     0
YearsWithCurrManager        0
dtype: int64
```

Data Preprocessing

CODE AND OUTPUT



Data Processing

```
# Transform categorical data into dummies
dummy_col = [column for column in df.drop('Attrition', axis=1).columns if df[column].nunique() < 20]
data = pd.get_dummies(df, columns=dummy_col, drop_first=True, dtype='uint8')
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1470 entries, 0 to 1469
Columns: 136 entries, Age to YearsWithCurrManager_17
dtypes: int64(9), uint8(127)
memory usage: 285.8 KB
```

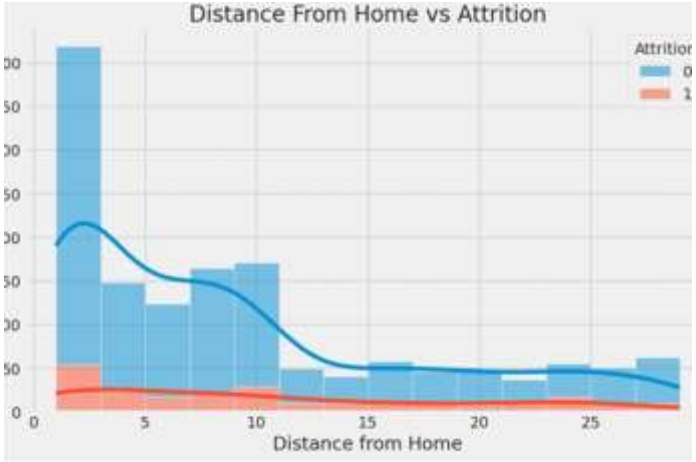
```
print(data.shape)

# Remove duplicate Features
data = data.T.drop_duplicates()
data = data.T

# Remove Duplicate Rows
data.drop_duplicates(inplace=True)

print(data.shape)
```

```
(1470, 136)
(1470, 136)
```

EXPLANATION

- Transforming Categorical Data: Categorical columns with fewer than 20 unique values (except the target, Attrition) are converted into dummy variables using one-hot encoding, which creates binary columns for each category.
- Removing Duplicate Features: Transposing the data allows detection of duplicate columns, which are then removed to avoid redundant information.
- Removing Duplicate Rows: Duplicate records in the dataset are identified and eliminated to ensure each row represents unique data, keeping the dataset's shape unchanged.

Histogram for DistanceFromHome (Attrition Analysis)

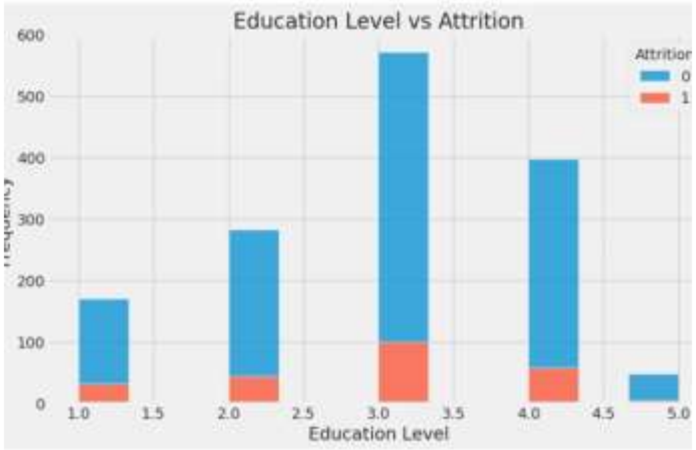


Inference: The histogram shows how the distance from home affects attrition rates, comparing employees who stayed vs. those who left. A greater distance may contribute to higher turnover.

Observation: Employees living farther from work may exhibit higher attrition rates, indicating that long commutes could impact job satisfaction and retention.

Recommendations: Introduce remote work options or transportation support to improve retention among employees with long travel distances.

Education and Relationship Satisfaction

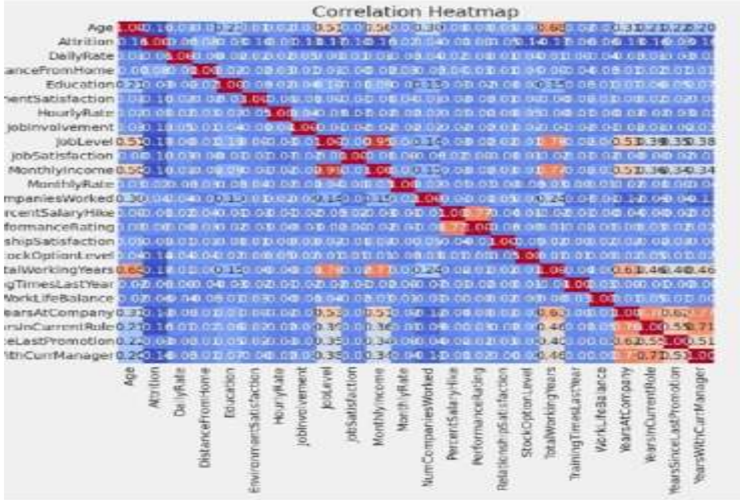


Inference: Employees with lower education levels or low relationship satisfaction tend to have higher attrition rates.

Observation: The data shows that employees with lower education or poor relationship satisfaction are more likely to leave the organization.

Recommendations: Create mentorship programs, improve work-life balance, and foster a culture of collaboration to retain employees.

Correlation Heatmap

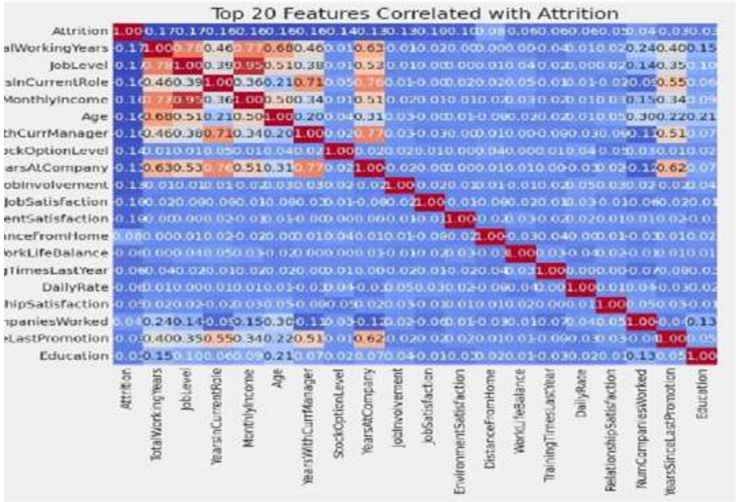


Inference: The correlation heatmap shows the relationships between various numerical features, identifying strong correlations like between JobSatisfaction and WorkLifeBalance.

Observation: Features such as JobSatisfaction and WorkLifeBalance have a high positive correlation, indicating they may jointly influence employee retention.

Recommendations: Focus on holistic employee well-being programs that improve both satisfaction and work-life balance.

Education and Relationship Satisfaction

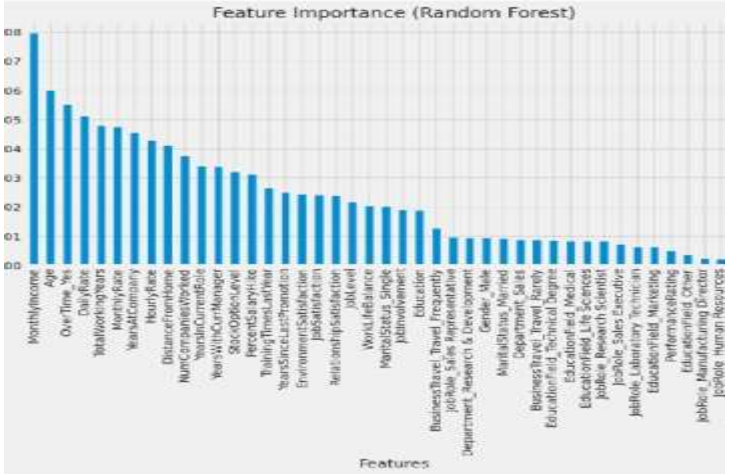


Inference: This heatmap focuses on the top 20 features most correlated with attrition, offering a more targeted view of employee turnover predictors.

Observation: Features like OverTime, MonthlyIncome, and JobRole show strong correlations with attrition, suggesting their critical role in predicting turnover.

Recommendations: Offercompetitive compensation packages, address job role dissatisfaction, and explore flexible working options for employees working overtime.

Feature Importance(Random Forest)

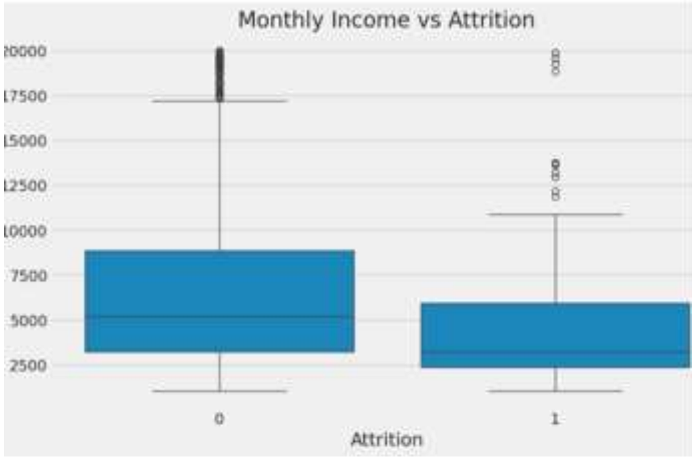


Inference: Random Forest model identifies the most important features influencing attrition, with OverTime and MonthlyIncome emerging as the top predictors.

Observation: Job role-related factors and compensation are critical drivers of employee turnover, according to the feature importance scores.

Recommendations: Enhance employee compensation, recognition programs, and consider job role alignment to retain valuable talent.

Box Plot (MonthlyIncome vs Attrition)

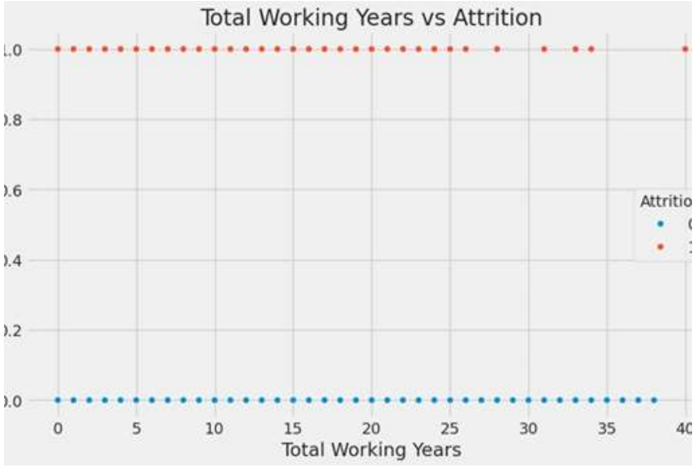


Inference: Employees with higher income tend to have lower attrition rates, as shown by the box plot comparison between employees who stayed and those who left.

Observation: There is a significant difference in income levels between employees who stay and those who leave, suggesting income as a factor in retention.

Recommendations: Offer competitive salaries and benefits to retain top-performing employees.

Scatter Plot

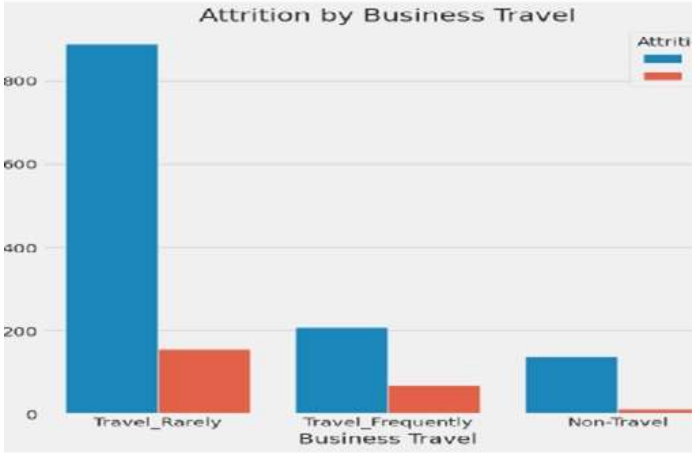


Inference: The scatter plot shows that employees with fewer years in the organization tend to leave more frequently.

Observation: Employees with less experience in the company exhibit higher turnover rates, possibly due to career progression or dissatisfaction.

Recommendations: Create onboarding programs, mentorship opportunities, and career development paths for newer employees.

Pie Chart (Attrition by BusinessTravel)

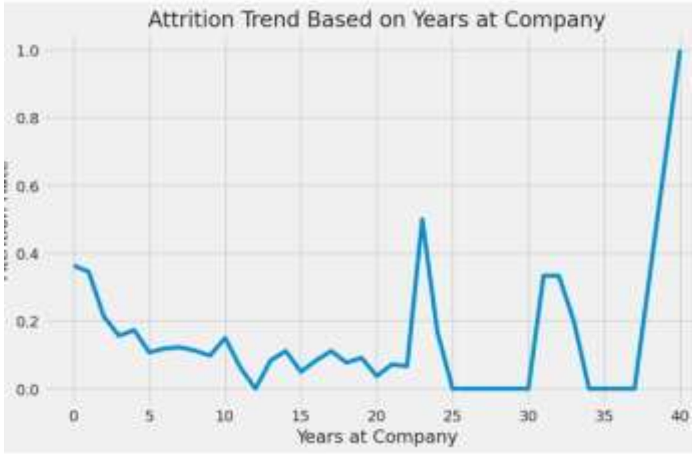


Inference: Employees who travel more frequently for business are less likely to leave compared to those with no travel or occasional travel.

Observation: Frequent business travel may indicate a higher level of engagement or job satisfaction, which correlates with lower attrition.

Recommendations: Consider expanding business travel opportunities or offering alternative engagement programs to employees who do not travel often.

Time-Series Plot(Attrition Trend Over Yrs)

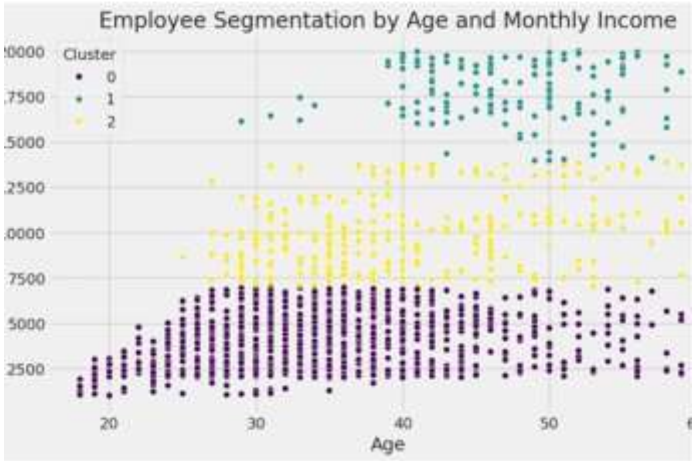


Inference: Attrition trends over time can highlight patterns and identify if turnover has increased due to external factors like economic changes or internal factors like management shifts.

Observation: If the attrition rate rises over certain years, it could suggest a systemic issue within the organization that needs addressing.

Recommendations: Use this information to introduce targeted retention programs at times when turnover rates peak.

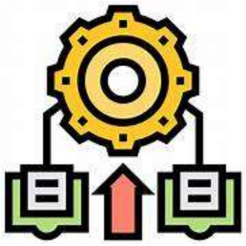
Time-Series Plot (Age & Monthly Income)



Inference: Clustering helps segment employees into groups with different attrition risks, allowing targeted interventions for high-risk clusters.

Observation: Employees in certain clusters with lower income or higher over-time hours tend to have a higher risk of attrition.

Recommendations: Focus retention efforts on high-risk clusters identified through clustering, such as employees with higher overtime and lower salaries.



Machine Learning Model

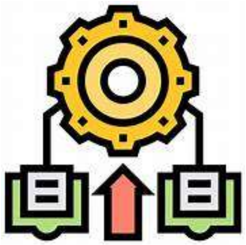
Logistic Regression:

- Logistic Regression was selected for its simplicity and ability to provide interpretable insights into factors influencing employee attrition.

- This model assumes a linear relationship between the independent variables and the log-odds of the target, making it a strong baseline for binary classification tasks like predicting attrition.

- It is particularly effective in highlighting the contribution of features such as OverTime, JobRole, and MonthlyIncome to attrition predictions.

```
TRAINIG RESULTS:
===============================
CONFUSION MATRIX:
[[849  14]
 [ 59 107]]
ACCURACY SCORE:
0.9291
CLASSIFICATION REPORT:
               0       1  accuracy  macro avg  weighted avg
precision   0.94    0.88      0.93       0.91          0.93
recall      0.98    0.64      0.93       0.81          0.93
f1-score    0.96    0.75      0.93       0.85          0.92
support   863.00  166.00      0.93    1029.00       1029.00
TESTING RESULTS:
===============================
CONFUSION MATRIX:
[[348  22]
 [ 43  28]]
ACCURACY SCORE:
0.8526
CLASSIFICATION REPORT:
               0      1  accuracy  macro avg  weighted avg
precision   0.89   0.56      0.85       0.73          0.84
recall      0.94   0.39      0.85       0.67          0.85
f1-score    0.91   0.46      0.85       0.69          0.84
support   370.00  71.00      0.85     441.00        441.00
```



Machine Learning Model

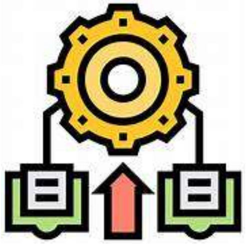
Support Vector Machines (SVM):

- SVM with a linear kernel was used to efficiently classify employee attrition by finding the optimal hyperplane that separates the data into distinct categories.

- This model is well-suited for high-dimensional datasets, ensuring robust performance even with complex feature interactions.

- It helps in identifying critical decision boundaries influenced by factors like OverTime, Age, and JobLevel, which are pivotal in attrition predictions.

```
TRAINIG RESULTS:
===============================
CONFUSION MATRIX:
[[855   8]
 [ 47 119]]
ACCURACY SCORE:
0.9466
CLASSIFICATION REPORT:
                0       1  accuracy  macro avg  weighted avg
precision    0.95    0.94      0.95       0.94          0.95
recall       0.99    0.72      0.95       0.85          0.95
f1-score     0.97    0.81      0.95       0.89          0.94
support    863.00  166.00      0.95    1029.00       1029.00
TESTING RESULTS:
===============================
CONFUSION MATRIX:
[[345  25]
 [ 44  27]]
ACCURACY SCORE:
0.8435
CLASSIFICATION REPORT:
                0      1  accuracy  macro avg  weighted avg
precision    0.89   0.52      0.84       0.70          0.83
recall       0.93   0.38      0.84       0.66          0.84
f1-score     0.91   0.44      0.84       0.67          0.83
support    370.00  71.00      0.84     441.00        441.00
```

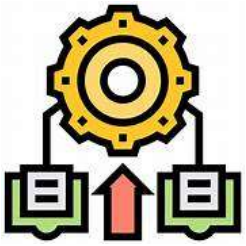


Model Evaluation

The models are evaluated using comprehensive performance metrics to ensure reliability in predicting employee attrition, especially in an imbalanced dataset. Key metrics include **accuracy**, **precision**, **recall**, **F1-score**, and **ROC-AUC**, with **confusion matrices** visualizing prediction outcomes.

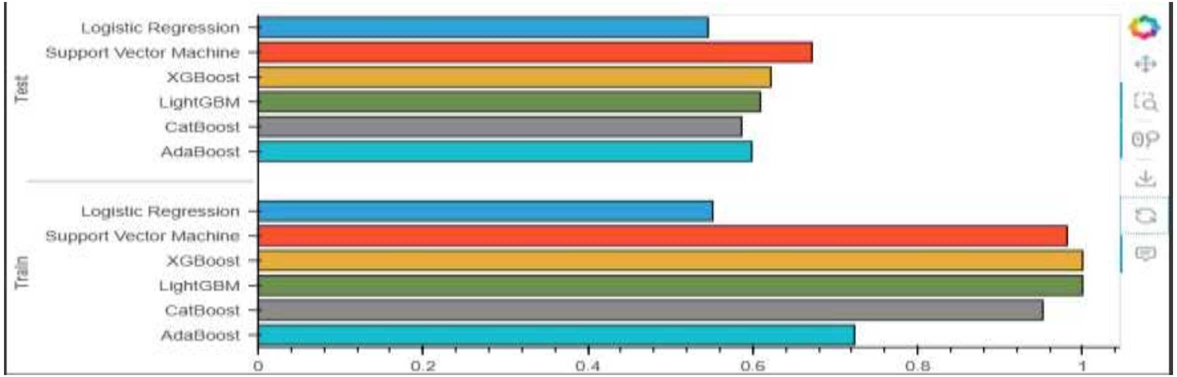
- **PR Curves** analyze the precision-recall trade-off, while **ROC curves** and **AUC scores** measure the models' discriminative power.
- Ensemble methods like **Random Forest** and **XGBoost** outperform simpler models (e.g., Logistic Regression) due to their ability to capture complex, non-linear relationships.

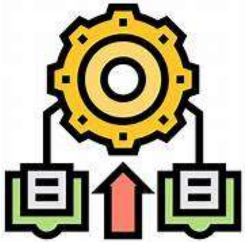
These evaluations provide critical insights into each model's ability to identify employees likely to leave while minimizing false predictions.



Summary of the Findings

The analysis reveals critical insights into employee turnover, derived from machine learning models such as Logistic Regression, SVM, XGBoost, and LightGBM. Among these, **Logistic Regression** and **XGBoost** emerged as the most effective, achieving the highest predictive accuracy and superior **ROC-AUC scores.**





Summary of the Findings

Key Findings:

- The workers with low JobLevel, MonthlyIncome, YearAtCompany, and TotalWorkingYears are more likely to quit there jobs.
- **BusinessTravel** : The workers who travel alot are more likely to quit then other employees.
- **Department** : The worker in Research & Development are more likely to stay then the workers on other department.
- **EducationField** : The workers with Human Resources and Technical Degree are more likely to quit then employees from other fields of educations.
- **Gender** : The Male are more likely to quit.
- **JobRole** : The workers in Laboratory Technician, Sales Representative, and Human Resources are more likely to quit the workers in other positions.
- **OverTime** : The workers who work more hours are likely to quit then others.

Thanks!

Any *questions* ?