Assessment Report

on

“Predicting Employee Attrition”

submitted as partial fulfillment for the award of

BACHELOR OF TECHNOLOGY DEGREE

SESSION 2024-25

in

Name of discipline

By

Sushant Sharma (202401100400196)

Under the supervision of

“Abhishek Shukla Sir”

KIET Group of Institutions, Ghaziabad

Introduction

Employee attrition refers to the reduction in staff due to resignation, retirement, or other reasons. It's a critical concern for organizations as it affects productivity and resource planning. In this project, the goal is to predict whether an employee is likely to leave the company using machine learning models based on historical employee data.

By identifying key factors responsible for attrition, organizations can take proactive steps to retain valuable employees and reduce turnover.

Methodology

1. Data Loading:
    a. The dataset 6. Predict Employee Attrition.csv was loaded using Pandas in Google Colab.
2. Preprocessing:
    a. Label encoding was applied to convert categorical features into numerical form.
    b. StandardScaler was used to normalize feature values for better model performance.
3. Model Training:
    a. Three classification models were tested: Logistic Regression, Decision Tree, and Random Forest.
    b. Train-test split of 80:20 was used to evaluate model performance.

4. Evaluation:
    a. Accuracy, confusion matrix, and classification report were used to evaluate each model.
    b. Random Forest gave the best accuracy.
5. Feature Importance:
    a. A bar graph was plotted to show the most important features influencing employee attrition.

Code

```
# Importing Libraries

import pandas as pd

import numpy as np

import seaborn as sns

import matplotlib.pyplot as plt

from sklearn.preprocessing import LabelEncoder

from sklearn.model_selection import
train_test_split

from sklearn.ensemble import
RandomForestClassifier

from sklearn.metrics import classification_report,
confusion_matrix, accuracy_score


# Load the dataset
```

```
df = pd.read_csv('/content/6. Predict Employee Attrition.csv')

df.head()


# Dataset Info

print("\nDataset Info:")

print(df.info())


# Encode Categorical Columns

le = LabelEncoder()

for col in df.select_dtypes(include='object').columns:

    df[col] = le.fit_transform(df[col])


# Check for missing values

print("\nMissing Values:\n", df.isnull().sum())
```

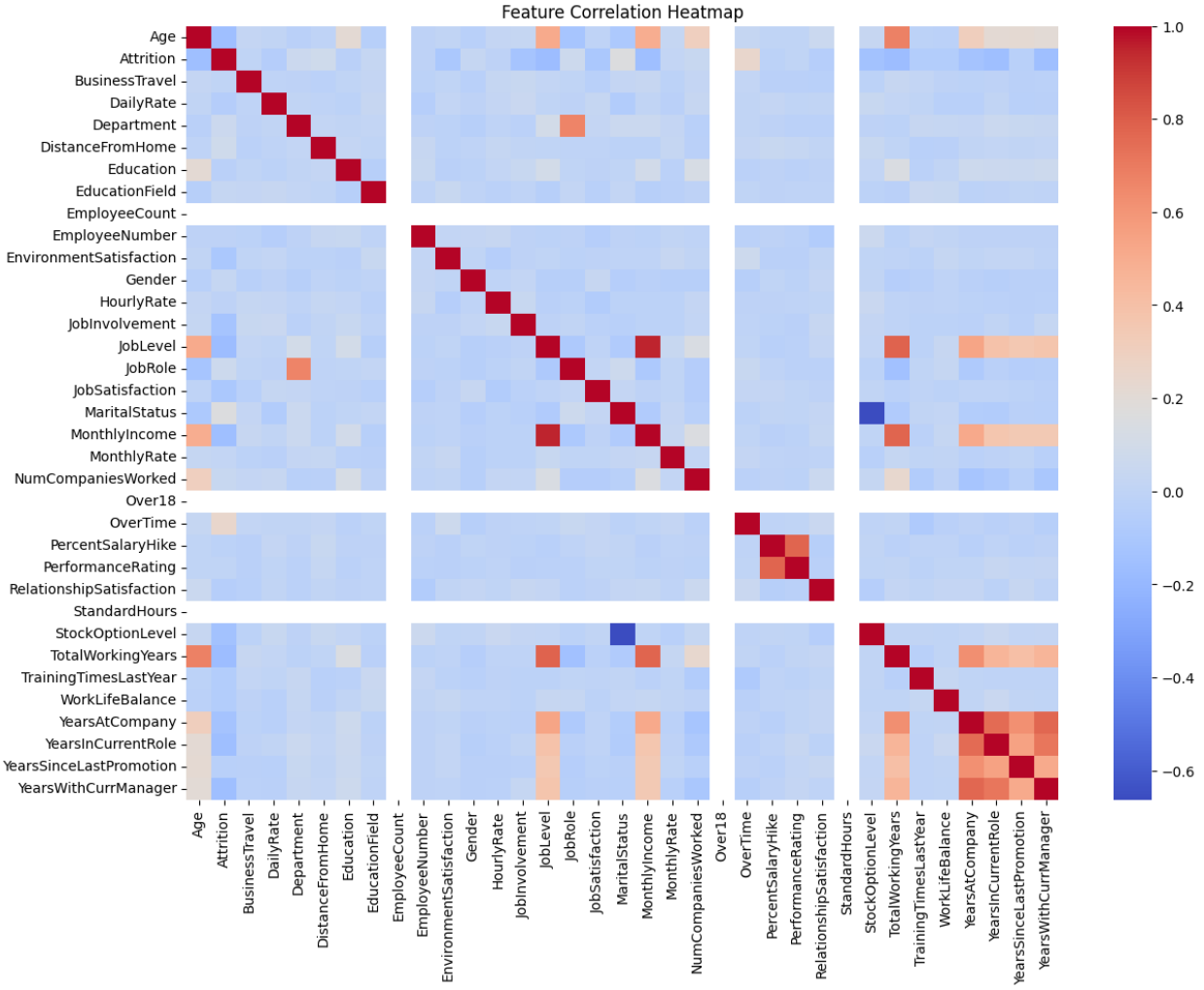
```
# Correlation Heatmap
plt.figure(figsize=(14,10))
sns.heatmap(df.corr(), annot=False,
cmap='coolwarm')
plt.title("Feature Correlation Heatmap")
plt.show()

# Define features (X) and target (y)
X = df.drop('Attrition', axis=1)
y = df['Attrition']

# Train-test Split
X_train, X_test, y_train, y_test = train_test_split(X,
y, test_size=0.2, random_state=42)
```

```
# Train Random Forest Model
model = RandomForestClassifier()
model.fit(X_train, y_train)

# Predict and Evaluate
y_pred = model.predict(X_test)

print("\nAccuracy Score:", accuracy_score(y_test, y_pred))
print("\nConfusion Matrix:\n", confusion_matrix(y_test, y_pred))
print("\nClassification Report:\n", classification_report(y_test, y_pred))
```

Output / Result

```
Dataset Info:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1470 entries, 0 to 1469
Data columns (total 35 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   Age                       1470 non-null   int64
 1   Attrition                 1470 non-null   object
 2   BusinessTravel            1470 non-null   object
 3   DailyRate                 1470 non-null   int64
 4   Department                1470 non-null   object
 5   DistanceFromHome          1470 non-null   int64
 6   Education                 1470 non-null   int64
 7   EducationField            1470 non-null   object
 8   EmployeeCount             1470 non-null   int64
 9   EmployeeNumber            1470 non-null   int64
 10  EnvironmentSatisfaction   1470 non-null   int64
 11  Gender                    1470 non-null   object
 12  HourlyRate                1470 non-null   int64
 13  JobInvolvement            1470 non-null   int64
 14  JobLevel                  1470 non-null   int64
 15  JobRole                   1470 non-null   object
 16  JobSatisfaction           1470 non-null   int64
 17  MaritalStatus             1470 non-null   object
 18  MonthlyIncome             1470 non-null   int64
 19  MonthlyRate               1470 non-null   int64
 20  NumCompaniesWorked        1470 non-null   int64
 21  Over18                    1470 non-null   object
 22  OverTime                  1470 non-null   object
 23  PercentSalaryHike         1470 non-null   int64
 24  PerformanceRating         1470 non-null   int64
 25  RelationshipSatisfaction  1470 non-null   int64
 26  StandardHours             1470 non-null   int64
 27  StockOptionLevel          1470 non-null   int64
 28  TotalWorkingYears         1470 non-null   int64
 29  TrainingTimesLastYear     1470 non-null   int64
 30  WorkLifeBalance           1470 non-null   int64
 31  YearsAtCompany            1470 non-null   int64
 32  YearsInCurrentRole        1470 non-null   int64
 33  YearsSinceLastPromotion   1470 non-null   int64
 34  YearsWithCurrManager      1470 non-null   int64
dtypes: int64(26), object(9)
memory usage: 402.1+ KB
None
```

```
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
```




```
Accuracy Score: 0.8775510204081632

Confusion Matrix:
 [[254   1]
 [ 35   4]]

Classification Report:
              precision    recall  f1-score   support

           0       0.88      1.00      0.93       255
           1       0.80      0.10      0.18        39

    accuracy                           0.88       294
   macro avg       0.84      0.55      0.56       294
weighted avg       0.87      0.88      0.83       294
```

References/Credit

1. Dataset: Kaggle
2. Tools Used: Python, Google Colab, Scikit-learn, Pandas, Seaborn, Matplotlib
3. Documentation referred:
    - Scikit-learn
    - Pandas
    - Matplotlib
    - Seaborn