A

REPORT

ON

"Problem statement"

Predict Employee Attrition: Build a classification model to predict whether an employee is likely to leave a company based on factors such as job satisfaction, salary, work environment, and years of experience.

BACHELOR OF TECHNOLOGY

DEGREE

SESSION 2024_25

IN

CSE AIML

BY

NAME-RISHI PATWA

ROLL NO-202401100400156

Under the supervision of

"ABHISHEK SHUKLA"

KIET Group of Institutions,Ghaziabad

INTRODUCTION

Employee attrition—also known as employee turnover—is a critical concern for organizations across industries. It refers to the loss of employees through resignation, retirement, or other forms of departure. High attrition rates can lead to increased recruitment and training costs, reduced productivity, and the potential loss of institutional knowledge. Therefore, predicting which employees are at risk of leaving is an essential step toward improving employee retention and organizational performance.

Methodology

To build an effective employee attrition prediction model, a structured machine learning workflow was followed. The dataset was first explored and cleaned, confirming that no missing values were present. Categorical variables such as department, job role, and marital status were encoded using label encoding, while irrelevant fields like EmployeeNumber, EmployeeCount, and StandardHours were removed to reduce noise. Numerical features were standardized using StandardScaler to ensure uniformity across different scales. The target variable Attrition was converted to binary form for classification. The dataset was then split into training and testing sets (80:20 ratio), and a Random Forest Classifier was trained due to its robustness and ability to handle both categorical and numerical data effectively. Model performance was evaluated using accuracy, a confusion matrix, and a classification report that included precision, recall, and F1-score. Finally, feature importance was analyzed to identify the most influential factors contributing to attrition. The results showed that variables like OverTime, JobSatisfaction, WorkLifeBalance, and MonthlyIncome had a strong impact on predicting whether an employee is likely to leave the company.

CODE

```
# 📦 Step 1: Install and import libraries
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler, LabelEncoder
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import classification_report, confusion_matrix, accuracy_score

# 📁 Step 2: Upload CSV in Colab
from google.colab import files
uploaded = files.upload() # ⬅ You will be prompted to upload the CSV file

# 📄 Step 3: Load the uploaded file
filename = list(uploaded.keys())[0]
data = pd.read_csv(filename)

# 🧼 Step 4: Basic cleaning
print("First few rows:\n", data.head())
print("\nMissing values:\n", data.isnull().sum())

# 🧠 Step 5: Encode categorical columns (if any)
label_encoders = {}
for col in data.select_dtypes(include='object').columns:
le = LabelEncoder()
data[col] = le.fit_transform(data[col])
label_encoders[col] = le
```

```python
# 🎯 Step 6: Define features and target
target_column = 'Attrition' # ⚠ Make sure this matches your dataset
X = data.drop(target_column, axis=1)
y = data[target_column]

# ✂ Step 7: Train-test split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# 📊 Step 8: Scale numeric features
scaler = StandardScaler()
X_train_scaled = scaler.fit_transform(X_train)
X_test_scaled = scaler.transform(X_test)

# 🤖 Step 9: Train Random Forest Classifier
model = RandomForestClassifier(n_estimators=100, random_state=42)
model.fit(X_train_scaled, y_train)

# 🔍 Step 10: Evaluate the model
y_pred = model.predict(X_test_scaled)

print("✅ Accuracy:", accuracy_score(y_test, y_pred))
print("\n📉 Confusion Matrix:\n", confusion_matrix(y_test, y_pred))
print("\n📋 Classification Report:\n", classification_report(y_test, y_pred))

# 🌟 Step 11: Plot feature importances
feat_importances = pd.Series(model.feature_importances_, index=X.columns)
feat_importances.nlargest(10).plot(kind='barh', figsize=(8, 6), color='skyblue')
plt.title('Top 10 Feature Importances')
plt.xlabel('Importance')
plt.ylabel('Features')
plt.tight_layout()
plt.show()
```

OUTPUT

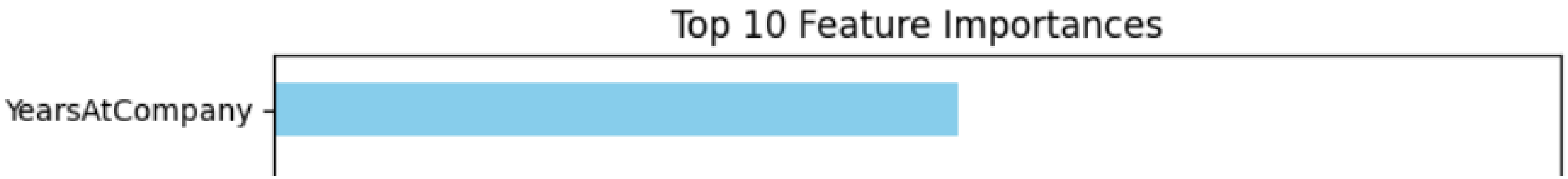
Choose Files 6. Predict E... Attrition.csv

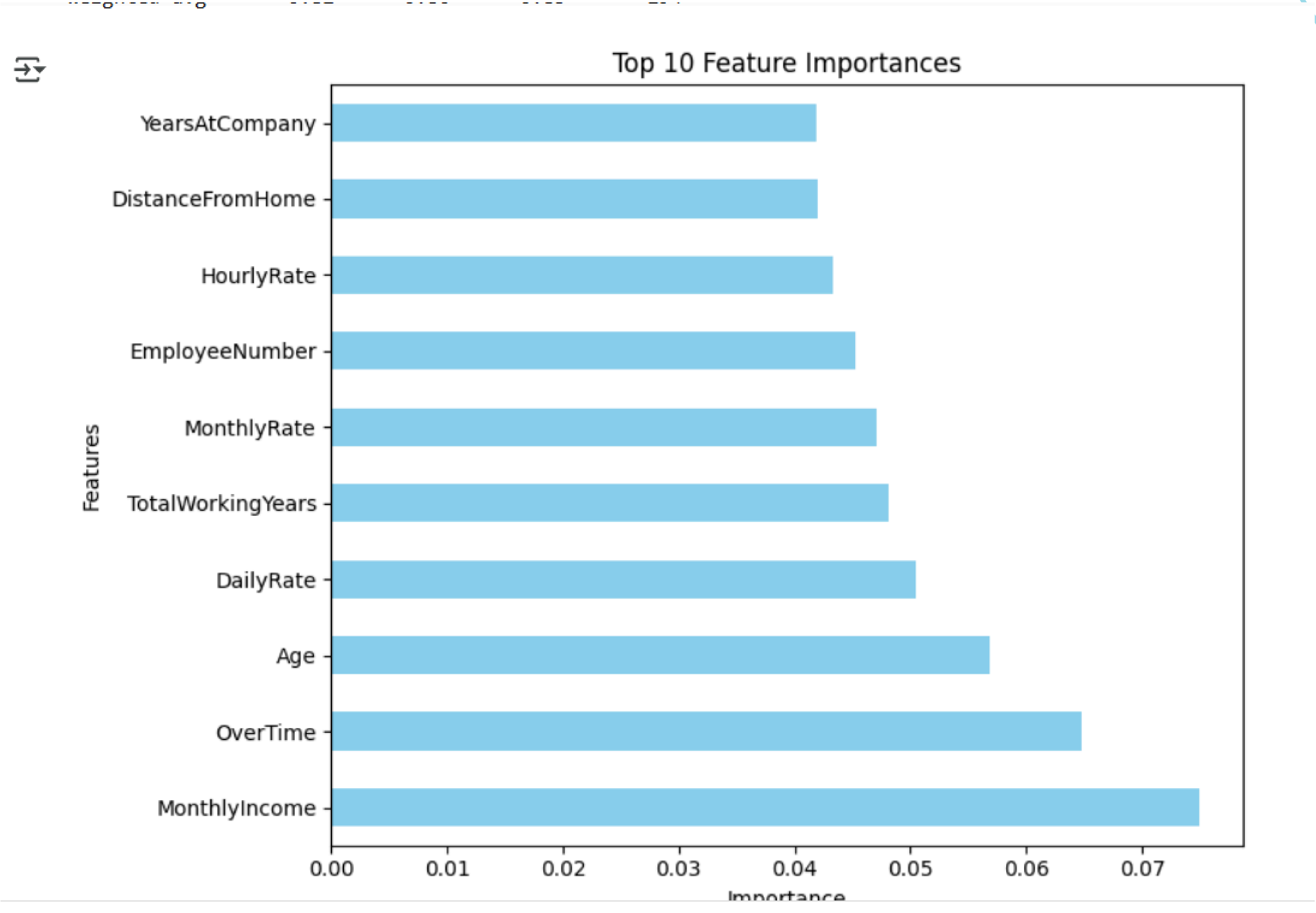
- **6. Predict Employee Attrition.csv**(text/csv) - 227977 bytes, last modified: 4/18/2025 - 100% done

```
Saving 6. Predict Employee Attrition.csv to 6. Predict Employee Attrition.csv
First few rows:
   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

   ...  RelationshipSatisfaction StandardHours  StockOptionLevel  \
0  ...                         1            80                 0
1  ...                         4            80                 1
2  ...                         2            80                 0
3  ...                         3            80                 0
4  ...                         4            80                 1

  TotalWorkingYears  TrainingTimesLastYear WorkLifeBalance  YearsAtCompany  \
0                 8                      0               1               6
1                10                      3               3              10
2                 7                      3               3               0
3                 8                      3               3               8
4                 6                      3               3               2

  YearsInCurrentRole  YearsSinceLastPromotion  YearsWithCurrManager
0                  4                        0                     5
```

```
   YearsInCurrentRole  YearsSinceLastPromotion  YearsWithCurrManage
0                   4                        0
1                   7                        1
2                   0                        0
3                   7                        3
4                   2                        2

[5 rows x 35 columns]

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
```

```
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: 0.8639455782312925

📉 Confusion Matrix:
 [[250   5]
 [ 35   4]]

📋 Classification Report:
               precision    recall  f1-score   support

           0       0.88      0.98      0.93       255
           1       0.44      0.10      0.17        39

    accuracy                           0.86       294
   macro avg       0.66      0.54      0.55       294
weighted avg       0.82      0.86      0.83       294
```





Top 10 Feature Importances
YearsAtCompany
DistanceFromHome
HourlyRate
EmployeeNumber
MonthlyRate
TotalWorkingYears
DailyRate
Age
OverTime
MonthlyIncome
Features
0.00
0.01
0.02
0.03
0.04
0.05
0.06
0.07

REFERENCES

- IBM HR Analytics Employee Attrition & Performance Dataset. Kaggle Dataset
- Scikit-learn: Machine Learning in Python. Pedregosa et al., Journal of Machine Learning Research, 2011. https://scikit-learn.org
- Géron, A. (2019). Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow (2nd ed.). O'Reilly Media.
- Microsoft Learn: Machine learning classification techniques
- https://learn.microsoft.com/en-us/azure/machine-learning/
- Brownlee, J. (2020). How to Prepare Data for Machine Learning. Machine Learning Mastery.
- https://machinelearningmastery.com