Assessment Report

on

"Employee Attrition Prediction"

submitted as partial fulfillment for the award of

BACHELOR OF TECHNOLOGY DEGREE

SESSION 2024-25

in

CSE(AIML)

By

Group 1

1. Prakhar Rai
2. Prakhar Tiwari
3. Prajesh Singh Meena
4. Krrish Kumar
5. Hemang Singh
6. Divyansh Yadav

Under the supervision of

"Mr. Abhishek Shukla"

KIET Group of Institutions, Ghaziabad

May, 2025

1. **Introduction**

   Employee attrition is a critical challenge for organizations, affecting productivity and operational stability. This project focuses on building a machine learning model using the IBM HR Analytics dataset to predict whether an employee is likely to leave the company. By applying a Random Forest classifier, the goal is to classify employees into "stay" or "leave" categories based on various features like job role, overtime, income, and satisfaction. The project emphasizes classification accuracy and feature importance to support data-driven HR decisions and improve employee retention.

2. Problem Statement

Develop a machine learning model to predict if an employee is likely to leave the company using IBM HR Analytics data. Focus on classification techniques and visualize feature importance.

3. Objectives

- Use the **IBM HR Analytics dataset** that contains various employee information.
- Build a **machine learning model** that predicts whether an employee will **stay or leave**.
- This is a **binary classification problem**:

Class 0 = Employee stays

Class 1 = Employee leaves (attrition)

4. Methodology

- **Data Collection:**
  The user uploads the IBM HR Analytics dataset in CSV format.
- **Data Preprocessing:**

    - Missing numerical values are handled using mean imputation.
    - Categorical variables are converted using label encoding.
    - Features are scaled (optional for tree-based models like Random Forest, though not strictly required).
- **Model Building:**
    - The dataset is split into training (80%) and testing (20%) sets.
    - A **Random Forest Classifier** is trained on the processed training data.
- **Model Evaluation:**
    - Evaluate the model using accuracy, precision, recall, and F1-score metrics.
    - Generate a confusion matrix and visualize it using a heatmap to analyse prediction performance.
    - Visualize feature importance to understand which employee features affect attrition the most.

5. Data Preprocessing

- Numerical missing values in features are filled with the mean of respective columns.

- Categorical features are encoded into numeric format using **Label Encoding**.

- Feature scaling is generally not necessary for Random Forest, but can be applied for consistency or future comparison with other models.

- The dataset is then split into 80% training and 20% testing sets for model validation.

6. Model Implementation

- A Random Forest Classifier is used due to its robustness, ability to handle mixed data types, and built-in feature importance mechanism.

- The model is trained on the training data.

- Predictions are made on the test set to classify whether employees are likely to leave (attrition) or stay.

7. Evaluation Metrics

- **Accuracy:** Measures the overall correctness of the attrition predictions.
- **Precision:** Indicates the proportion of employees predicted to leave who actually left.
- **Recall:** Shows the proportion of actual employees who left that were correctly predicted.
- **F1 Score:** The harmonic mean of precision and recall, balancing both metrics.
- **Confusion Matrix:** Visualized with a heatmap using Seaborn to interpret true positives, true negatives, false positives, and false negatives clearly.

8. Results and Analysis

- The Random Forest model demonstrated **good classification performance** on the test data, effectively identifying employees who are likely to leave.
- The **confusion matrix heatmap** gave clear visual insight into correct and incorrect predictions.
- Feature importance analysis highlighted key drivers of attrition (e.g. **Over Time** , **Job Satisfaction**, **Monthly Income**, etc.), which are useful for HR interventions.
- Precision and recall metrics showed the model's strength in minimizing false predictions and correctly identifying potential attrition cases.

9. 9. Conclusion

The **Random Forest model** successfully predicted employee attrition with strong performance metrics. This project highlights the capability of machine learning to assist HR departments in identifying employees at risk of leaving. By analysing critical features like overtime, satisfaction level, and income, the model supports better workforce planning and retention strategies. Future improvements may include trying other

advanced models (e.g., Gradient Boosting), fine-tuning Random Forest hyperparameters, and handling class imbalance for even better results.

10. References

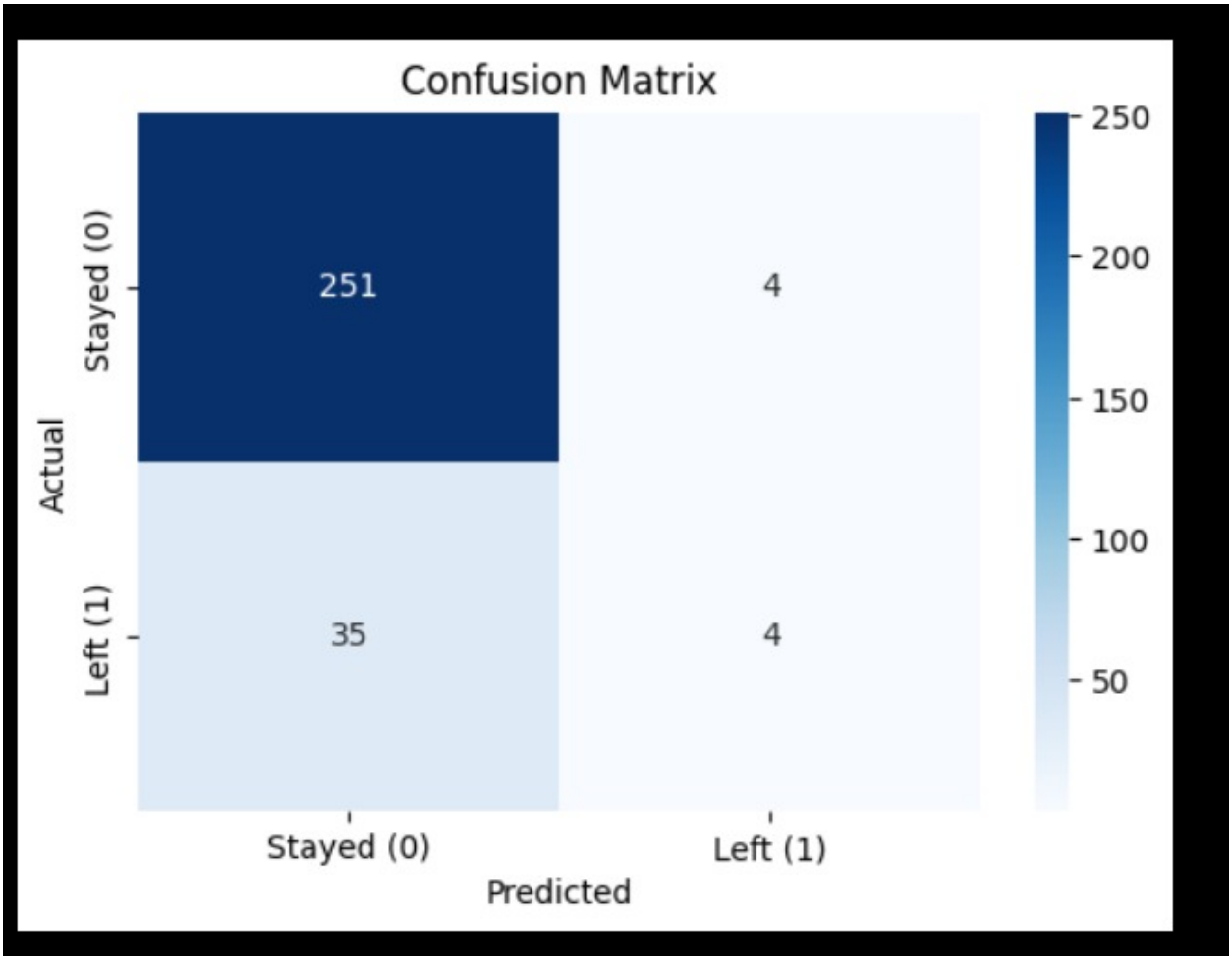
- scikit-learn documentation
- pandas documentation
- Seaborn visualization library
- Matplotlib

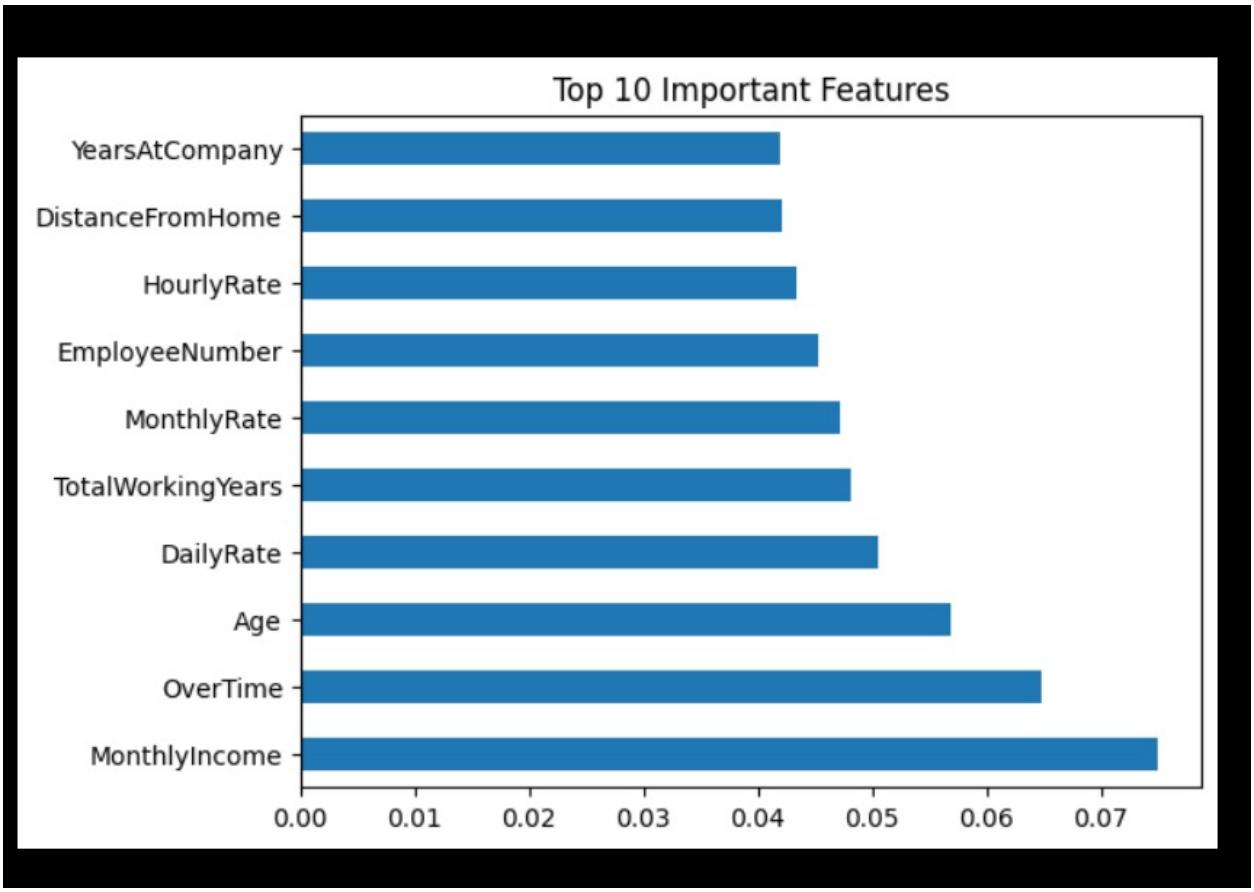
```
Saving archive (3).zip to archive (3) (4).zip
Classification Report:
              precision    recall  f1-score   support

           0       0.88      0.98      0.93       255
           1       0.50      0.10      0.17        39

    accuracy                           0.87       294
   macro avg       0.69      0.54      0.55       294
weighted avg       0.83      0.87      0.83       294

Confusion Matrix:
[[251   4]
 [ 35   4]]
```


Confusion Matrix
Actual
Stayed (0)
Left (1)
251
4
35
4
Stayed (0)
Left (1)
Predicted
250
200
150
100
50


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