A

Assessment Report

on

“Problem Statement”

submitted as partial fulfillment for the award of

BACHELOR OF TECHNOLOGY DEGREE

SESSION 2024-25

in

CSE(Artificial Intelligence & Machine Learning)

By

Vaishnavi Sahu (202401100400206)

Under the supervision of

“Abhishek Shukla”

KIET Group of Institutions, Ghaziabad

Affiliated to

Dr. A.P.J. Abdul Kalam Technical University, Lucknow

(Formerly UPTU)

May, 2025

Introduction

Employee attrition, also known as employee turnover, is a major concern for organizations across the globe. It refers to the gradual reduction in a company's workforce when employees leave and are not immediately replaced. High attrition rates can negatively impact a company's performance, increase hiring and training costs, and lower overall employee morale. Therefore, being able to predict attrition in advance allows companies to take proactive measures to retain valuable employees, improve workplace satisfaction, and optimize HR strategies.

In recent years, with the rise of data-driven decision-making, machine learning techniques have proven to be extremely useful in solving real-world business problems, including attrition prediction. By analyzing historical employee data, organizations can uncover patterns and factors that contribute to employee exits. This includes key attributes such as job satisfaction, monthly income, years of experience, department, work-life balance, and more.

In this project, we aim to build a classification model using Logistic Regression, a widely used algorithm for binary classification tasks. The model will learn from past employee data and predict whether a current employee is likely to leave the organization. The dataset used for this task contains various features that influence attrition, such as work environment conditions, salary levels, distance from home, and performance ratings.

The overall goal is to assist human resource departments by providing a reliable predictive tool that can help identify at-risk employees early. With this information, companies can implement timely interventions such as employee engagement programs, role reshuffling, or compensation adjustments to reduce the likelihood of attrition.

This report outlines the end-to-end process followed to develop the model — starting from data preprocessing and exploration, to model training and evaluation, and finally visualizing the results. The outcomes of this project can play a vital role in strategic HR planning and workforce optimization.

Methodology

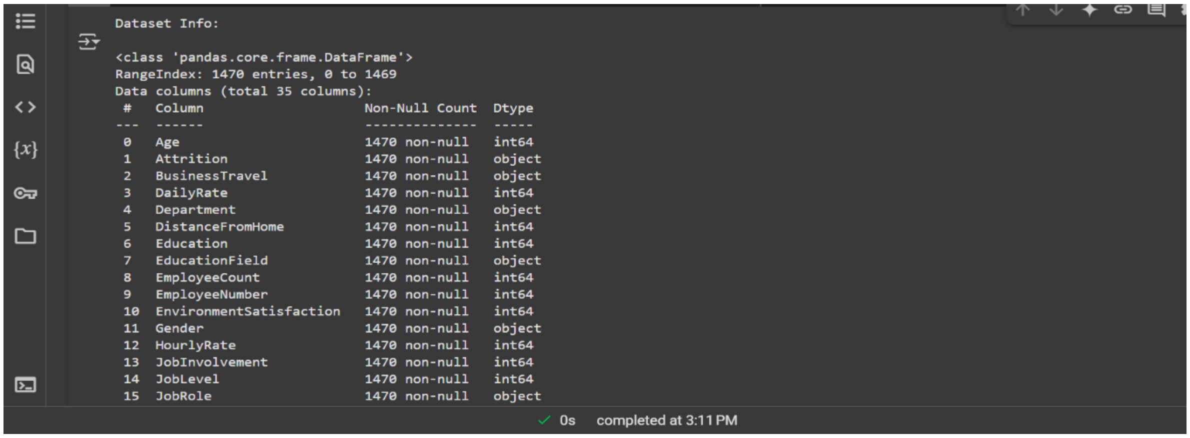
1. **Data Loading**: The dataset was loaded using the pandas library.
2. **Initial Exploration**: We reviewed dataset structure and values using .info() and .head() to understand the data.
3. **Missing Value Handling**: Numerical missing values were replaced with their median to maintain data integrity.
4. **Categorical Encoding**: All categorical variables were encoded using LabelEncoder to convert them into numerical form.
5. **Splitting the Dataset**: The data was split into features (X) and target (y). The target column was Attrition, which was also encoded.
6. **Train-Test Split**: The dataset was split into 80% training and 20% testing sets using train_test_split.
7. **Model Training**: A Logistic Regression model was trained on the training dataset.
8. **Prediction**: The trained model was used to predict attrition on the test dataset.
9. **Model Evaluation**: Accuracy score, confusion matrix, and classification report were generated to evaluate the model's performance.
10. **Visualization**: A heatmap of the confusion matrix was plotted for better visual understanding.

Code

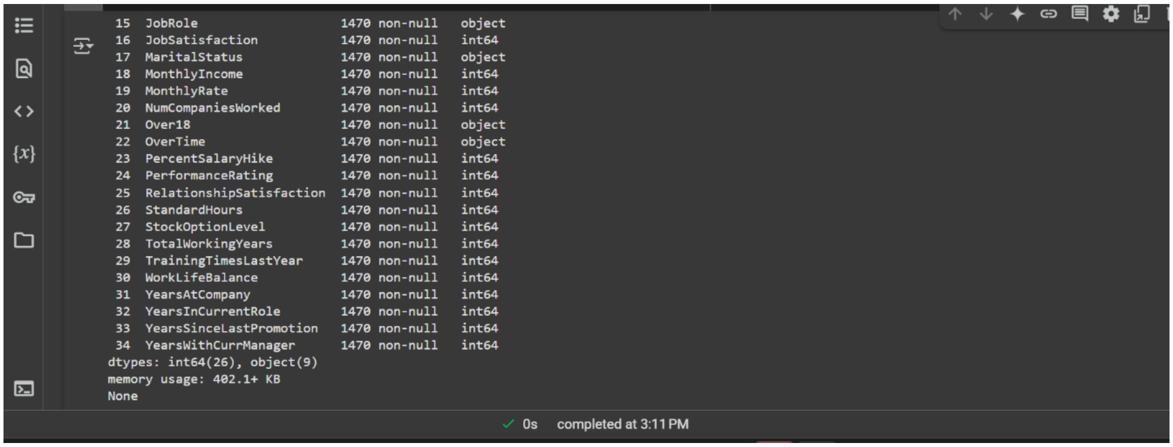
```
# Import necessary libraries
import pandas as pd
import numpy as np
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import LabelEncoder
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import confusion_matrix, classification_report, accuracy_score
import seaborn as sns
import matplotlib.pyplot as plt

# STEP 1: Load the dataset
df = pd.read_csv('/content/6. Predict Employee Attrition.csv')  # Update path if needed

# STEP 2: Basic data check
print("Dataset Info:\n")
print(df.info())
print("\nDataset Head:\n")
print(df.head())

# STEP 3: Handle missing values (if any)
df.fillna(df.median(numeric_only=True), inplace=True)

# STEP 4: Encode categorical columns
categorical_columns = df.select_dtypes(include=['object']).columns  # Automatically find categorical columns
label_encoders = {}

for col in categorical_columns:
    le = LabelEncoder()
    df[col] = le.fit_transform(df[col])
    label_encoders[col] = le

# STEP 5: Split data into features (X) and target (y)
X = df.drop('Attrition', axis=1)  # Use the correct column name for target variable
y = df['Attrition']

# Encode target variable (if it's categorical)
y = LabelEncoder().fit_transform(y)

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

# Proceed with model training and evaluation

# STEP 6: Train logistic regression model
model = LogisticRegression()
model.fit(X_train, y_train)

# STEP 7: Make predictions
y_pred = model.predict(X_test)

# STEP 8: Evaluate model performance
print("\nAccuracy:", 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))

# STEP 9: Heatmap of confusion matrix
conf_matrix = confusion_matrix(y_test, y_pred)
```

```
sns.heatmap(conf_matrix, annot=True, fmt='d', cmap='Blues', xticklabels=['No', 'Yes'], yticklabels=['No', 'Yes'])

plt.xlabel('Predicted')

plt.ylabel('Actual')

plt.title('Confusion Matrix Heatmap')

plt.show()
```

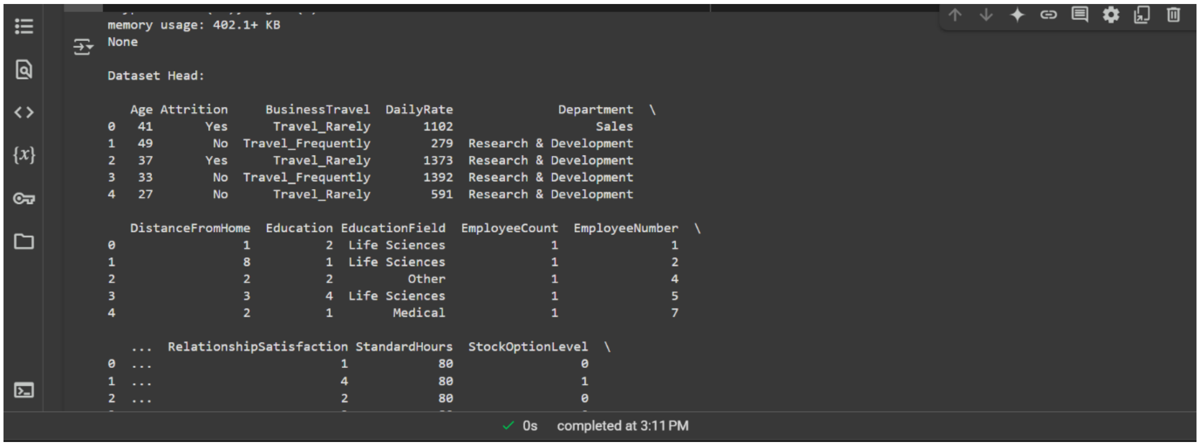
Output



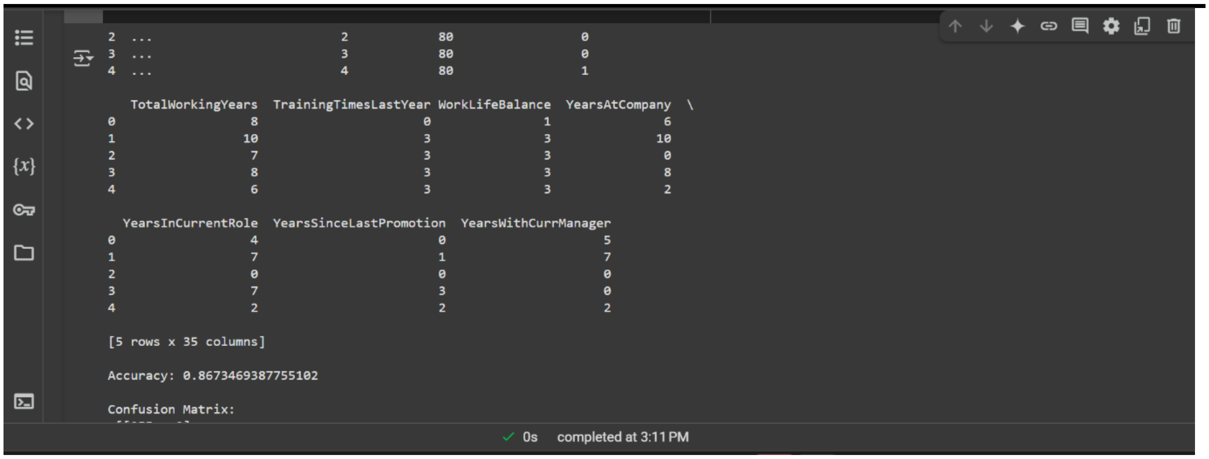
```
Dataset Info:

<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
```



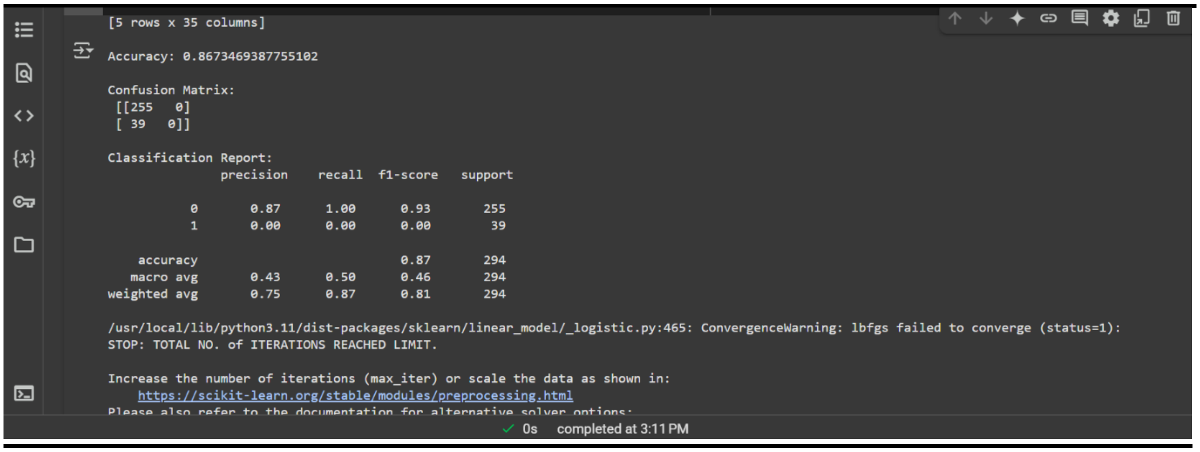
```
 15  JobRole                   1470 non-null   object
 16  JobSatisfaction           1470 non-null   int64
 17  MaritalStatus             1470 non-null   object
 18  MonthlyIncome             1470 non-null   int64
 19  MonthlyRate               1470 non-null   int64
 20  NumCompaniesWorked        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
```



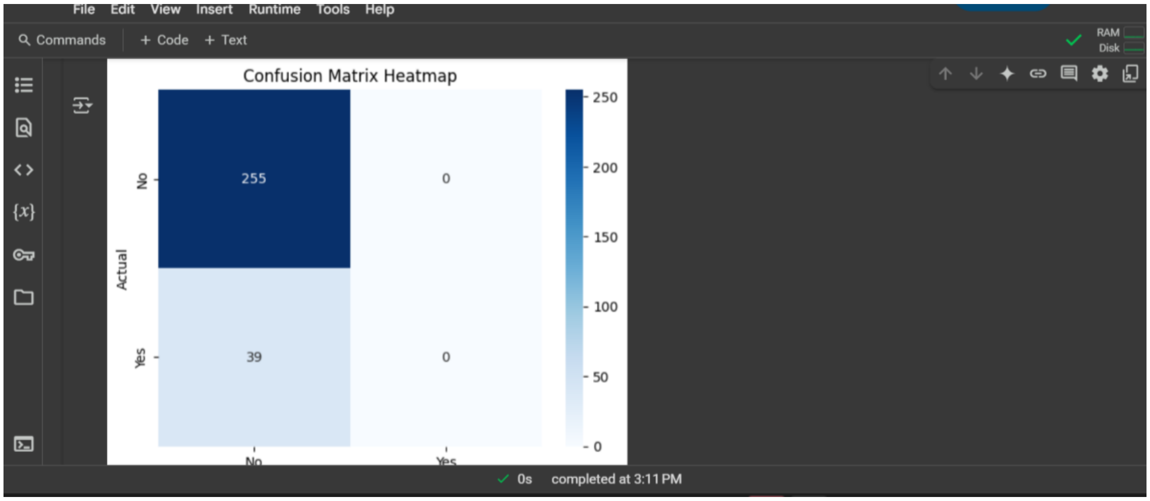
```
memory usage: 402.1+ KB
None

Dataset Head:

   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
```



```
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
1                  7                        1                     7
2                  0                        0                     0
3                  7                        3                     0
4                  2                        2                     2

[5 rows x 35 columns]

Accuracy: 0.8673469387755102

Confusion Matrix:
```



```
[5 rows x 35 columns]

Accuracy: 0.8673469387755102

Confusion Matrix:
 [[255   0]
 [ 39   0]]

Classification Report:
               precision    recall  f1-score   support

           0       0.87      1.00      0.93       255
           1       0.00      0.00      0.00        39

    accuracy                           0.87       294
   macro avg       0.43      0.50      0.46       294
weighted avg       0.75      0.87      0.81       294

/usr/local/lib/python3.11/dist-packages/sklearn/linear_model/_logistic.py:465: ConvergenceWarning: lbfgs failed to converge (status=1):
STOP: TOTAL NO. of ITERATIONS REACHED LIMIT.

Increase the number of iterations (max_iter) or scale the data as shown in:
    https://scikit-learn.org/stable/modules/preprocessing.html
Please also refer to the documentation for alternative solver options:
```



✓ 0s completed at 3:11 PM

References

The following sources, tools, and resources have been used throughout the development of the project:

- **Dataset**:
  The dataset used for this project titled “ Predict Employee Attrition.csv” was provided as part of the coursework. It contains various attributes related to employees such as age, job satisfaction, income, experience, distance from home, and more, which helped in building a robust classification model.

- **Programming Language & Environment**:
  - The entire code was written in Python, a powerful language widely used in AI and data science fields.
  - The implementation was done using Google Colab, which provides an interactive cloud-based Jupyter notebook environment ideal for machine learning development and visualization.

- **Libraries and Tools Used**:
  - *pandas* – for data manipulation and handling datasets.
  - *numpy* – for numerical operations.
  - *matplotlib & seaborn* – for data visualization and plotting.
  - *scikit-learn* – for machine learning algorithms, data preprocessing, model building, and evaluation.
  - *LabelEncoder* – for encoding categorical data into numerical format.
  - *LogisticRegression* – for building the classification model to predict employee attrition.

- **Academic References & Learning Sources**:
  - Online documentation and user guides from:
    - scikit-learn official documentation
    - pandas documentation
    - seaborn documentation