Report: Data Preprocessing and Machine Learning Model Evaluation:

I. 1. Introduction

Employee attrition, or employee turnover, is a significant concern for many organizations as it can lead to substantial costs in terms of recruitment, training, and loss of productivity. Understanding and predicting which employees are likely to leave the company can help organizations take proactive steps to improve employee retention, thereby saving costs and maintaining a stable workforce.

This project focuses on predicting employee attrition using a machine learning approach. We aim to build a predictive model that can identify employees at risk of leaving the company based on various features such as demographic information, job role, work environment, and personal factors.

The dataset used for this project is the HR Employee Attrition dataset, which contains information on 2940 employees, including their age, business travel frequency, department, distance from home, education, job role, job satisfaction, and many other factors.

The tasks performed in this project can be categorized into two main parts:

Data Preprocessing and Cleaning:

- Data Exploration: Analyzing the dataset to understand its structure and identify important features.
- Data Cleaning: Handling missing values, removing duplicates, and addressing any inconsistencies in the data.
- Data Encoding: Converting categorical variables into a format suitable for machine learning algorithms.
- Data Standardization: Scaling numerical features to ensure they contribute equally to the model.

Building Machine Learning Models:

- Model Training: Training a logistic regression model to predict employee attrition.
- Model Evaluation: Evaluating the model's performance using metrics such as accuracy, confusion matrix, precision, recall, and F1-score.

By the end of this project, we aim to have a robust predictive model that can help HR departments identify at-risk employees and take necessary actions to improve retention. The insights gained from this model can also guide organizations in making data-driven decisions to enhance employee satisfaction and reduce turnover rates.

I. Data Preprocessing and Cleaning

a. data exploration:

```
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler, LabelEncoder
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import classification_report, confusion_matrix, accuracy_score

# Load the data
df = pd.read_csv('C:/Users/4B/Downloads/WA_Fn-UseC_-HR-Employee-Attrition (1).csv')

# Data exploration
print("First rows of the data:")
print(df.head())
print("\nDescription of the data:")
print(df.describe())
print("\nInformation about the data:")
print(df.info())
```

b. Cleaning data:

```
# Data cleaning: check for missing values
print("\nMissing values per column:")
print(df.isnull().sum())
```

c. data encoding:

```
# Encoding the target variable 'Attrition'
df['Attrition'] = df['Attrition'].map({'Yes': 1, 'No': 0})

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

# Identifying numeric and categorical columns
numeric_features = X.select_dtypes(include=[int, float]).columns
categorical_features = X.select_dtypes(include=[object]).columns

# Label Encoding categorical variables (before One-Hot Encoding for checking)
for col in categorical_features:
    X[col] = LabelEncoder().fit_transform(X[col])

# One-Hot Encoding categorical variables
X = pd.get_dummies(X, columns=categorical_features)
```

d. data Standardization:

```python
# Standardizing numeric features
scaler = StandardScaler()
X[numeric_features] = scaler.fit_transform(X[numeric_features])
```

II. Building Machine Learning Models

```python
# Splitting data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_

# Training a logistic regression model
model = LogisticRegression(max_iter=1000)
model.fit(X_train, y_train)

# Predictions and model evaluation
y_pred = model.predict(X_test)

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

IV. Execution

```
Premières lignes des données :
   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
1                  7                        1                     7
2                  0                        0                     0
3                  7                        3                     0
4                  2                        2                     2
```

```
[5 rows x 35 columns]

Description des données :
              Age    DailyRate  DistanceFromHome    Education  EmployeeCount  \
count  1470.000000  1470.000000       1470.000000  1470.000000         1470.0
mean     36.923810   802.485714          9.192517     2.912925            1.0
std       9.135373   403.509100          8.106864     1.024165            0.0
min      18.000000   102.000000          1.000000     1.000000            1.0
25%      30.000000   465.000000          2.000000     2.000000            1.0
50%      36.000000   802.000000          7.000000     3.000000            1.0
75%      43.000000  1157.000000         14.000000     4.000000            1.0
max      60.000000  1499.000000         29.000000     5.000000            1.0

       EmployeeNumber  EnvironmentSatisfaction   HourlyRate  JobInvolvement  \
count     1470.000000              1470.000000  1470.000000     1470.000000
mean      1024.865306                 2.721769    65.891156        2.729932
std        602.024335                 1.093082    20.329428        0.711561
min          1.000000                 1.000000    30.000000        1.000000
```

```
25%        491.250000                 2.000000    48.000000        2.000000
50%       1020.500000                 3.000000    66.000000        3.000000
75%       1555.750000                 4.000000    83.750000        3.000000
max       2068.000000                 4.000000   100.000000        4.000000

          JobLevel  ...  RelationshipSatisfaction  StandardHours  \
count  1470.000000  ...               1470.000000         1470.0
mean      2.063946  ...                  2.712245           80.0
std       1.106940  ...                  1.081209            0.0
min       1.000000  ...                  1.000000           80.0
25%       1.000000  ...                  2.000000           80.0
50%       2.000000  ...                  3.000000           80.0
75%       3.000000  ...                  4.000000           80.0
max       5.000000  ...                  4.000000           80.0

       StockOptionLevel  TotalWorkingYears  TrainingTimesLastYear  \
count       1470.000000        1470.000000            1470.000000
mean           0.793878          11.279592               2.799320
```

```
std           0.852077          7.780782                1.289271
min           0.000000          0.000000                0.000000
25%           0.000000          6.000000                2.000000
50%           1.000000         10.000000                3.000000
75%           1.000000         15.000000                3.000000
max           3.000000         40.000000                6.000000

       WorkLifeBalance  YearsAtCompany  YearsInCurrentRole  \
count      1470.000000     1470.000000         1470.000000
mean          2.761224        7.008163            4.229252
std           0.706476        6.126525            3.623137
min           1.000000        0.000000            0.000000
25%           2.000000        3.000000            2.000000
50%           3.000000        5.000000            3.000000
75%           3.000000        9.000000            7.000000
max           4.000000       40.000000           18.000000

       YearsSinceLastPromotion  YearsWithCurrManager
```

```
count              1470.000000           1470.000000
mean                  2.187755              4.123129
std                   3.222430              3.568136
min                   0.000000              0.000000
25%                   0.000000              2.000000
50%                   1.000000              3.000000
75%                   3.000000              7.000000
max                  15.000000             17.000000

[8 rows x 26 columns]

Information sur les données :
<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

Valeurs manquantes par colonne :
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
```

```
YearsAtCompany              0
YearsInCurrentRole          0
YearsSinceLastPromotion     0
YearsWithCurrManager        0
dtype: int64
Accuracy: 0.891156462585034
Confusion Matrix:
 [[244  11]
 [ 21  18]]
Classification Report:
               precision    recall  f1-score   support

           0       0.92      0.96      0.94       255
           1       0.62      0.46      0.53        39

    accuracy                           0.89       294
   macro avg       0.77      0.71      0.73       294
weighted avg       0.88      0.89      0.88       294
```

Conclusion:

In this project, we utilized machine learning to predict employee attrition, a crucial issue impacting organizational costs and stability. By preprocessing and cleaning the data, encoding categorical variables, and standardizing numerical features, we prepared the dataset for modeling. A logistic regression model was then trained and evaluated, achieving an accuracy of 89.12%. This predictive model equips HR departments with valuable insights to identify at-risk employees and implement targeted retention strategies. The results underscore the potential of data-driven approaches in enhancing workforce management and decision-making processes within organizations.