Report on Data Preprocessing and Model Performance

- **Data Exploration:**

The dataset contains several features related to employee characteristics and their job environment. The target variable is Attrition(Yes or No), indicating whether an employee has left the company.

- **Columns in the Dataset:**

Age, Attrition, BusinessTravel, DailyRate, Department, DistanceFromHome, Education, EducationField, EmployeeCount, EmployeeNumber, EnvironmentSatisfaction, Gender, HourlyRate, JobInvolvement, JobLevel, JobRole, JobSatisfaction, MaritalStatus, MonthlyIncome, MonthlyRate, NumCompaniesWorked, Over18, OverTime, PercentSalaryHike, PerformanceRating, RelationshipSatisfaction, StandardHours, StockOptionLevel, TotalWorkingYears, TrainingTimesLastYear, WorkLifeBalance, YearsAtCompany, YearsInCurrentRole, YearsSinceLastPromotion, YearsWithCurrManager.

- **Data Cleaning**

Missing values were handled by dropping any rows with missing data.

The target variable Attrition was converted to a binary format where 'Yes' was mapped to 1 and 'No' to 0.

- **Data Encoding**

Ordinal Columns: Encoded using LabelEncoder.

I. Education
II. EnvironmentSatisfaction
III. JobSatisfaction
IV. RelationshipSatisfaction
V. WorkLifeBalance

Nominal Columns: Encoded using One-Hot Encoding.

I. BusinessTravel
II. Department
III. EducationField
IV. Gender
V. JobRole
VI. MaritalStatus
VII. Over18

VIII. OverTime

- **Data Labelling**

Features (X) and target (y) were separated.

Ensured all categorical columns were encoded properly.

- **Train-Test Split**

Data was split into training and testing sets with an 80-20 split.

- **Model Building**

A Random Forest Classifier was used with 100 estimators and a random state of 42 for reproducibility.

Model Performance

Confusion Matrix

[[253 2]

[35 4]]

Accuracy

0.8741 (87.41%)

Classification Report

	precision	recall	f1-score	support
0	0.88	0.99	0.93	255
1	0.67	0.10	0.18	39
accuracy			0.87	294
macro avg	0.77	0.55	0.55	294
weighted avg	0.85	0.87	0.83	294

- **Interpretation**
  - Confusion Matrix:

True Negatives (TN): 253 employees did not leave the company (correctly predicted).

False Positives (FP): 2 employees were predicted to leave but did not.

False Negatives (FN): 35 employees were predicted to stay but actually left.

True Positives (TP): 4 employees left the company (correctly predicted).

- Accuracy:

The overall accuracy of the model is 87.41%. This means that out of all the predictions made by the model, 87.41% were correct.

Precision, Recall, and F1-Score:

For class 0 (employees who did not leave):

Precision: 0.88 (88%)

Recall: 0.99 (99%)

F1-Score: 0.93 (93%)

For class 1 (employees who left):

Precision: 0.67 (67%)

Recall: 0.10 (10%)

F1-Score: 0.18 (18%)

The macro average shows a significant imbalance in recall between the classes, which is reflected in the lower scores for class 1 (attrition).

- **Conclusion**

The Random Forest Classifier performed well in predicting the majority class (employees who did not leave). However, it struggled with predicting the minority class (employees who left), as indicated by the low recall and F1-score for class 1. This could be due to class imbalance in the dataset.