Project: Task 9 - Credit Card Fraud Detection Model: Random Forest vs Logistic Regression (Baseline)

1. Data Analysis

Dataset: Imbalanced credit card transaction data.

Imbalance: The dataset is highly skewed, with only ~1% of transactions representing fraud. This mimics real-world sc

Splitting: Stratified sampling was used to ensure the Test set also had exactly 1% fraud cases, preventing testing bias

2. Model Comparison

Baseline (Logistic Regression): Produced decent accuracy but struggled with Recall (identifying actual fraud cases).

Random Forest: Significantly outperformed the baseline.

Precision: High (Few false alarms).

Recall: High (Caught most fraud cases).

F1-Score: The Random Forest achieved a higher F1-score, making it the superior choice for this task.

3. Feature Importance

The Random Forest identified specific "V" features (e.g., V14, V10, V4) as the most critical indicators of fraud. This hel
terns to trigger alerts.

4. Conclusion Random Forest is highly effective for fraud detection because its ensemble of trees handles the compl
linear models.