

**Project Title**: Credit Card Fraud Detection using AI and ML

**Made by**: Aaminah Binte Farooq

**Department of Cyber Security**

**Roll no**: 22k-4671

**Submitted to**: Ms. Mehak Mazhar

**Dated:** 11 May 2025

**CREDIT CARD FRAUD DETECTION USING AI AND MACHINE LEARNING**

# **1. Introduction**

Credit card fraud detection is an essential application of machine learning in the financial sector. Due to the rarity of fraud compared to legitimate transactions, developing models that can accurately identify fraudulent activity is both challenging and critical. This report demonstrates the workflow of building and evaluating machine learning models (Logistic Regression and Random Forest) for fraud detection.

# **2. Dataset Overview**

- Source: Credit card transactions dataset

- Total Records: 284,807 transactions

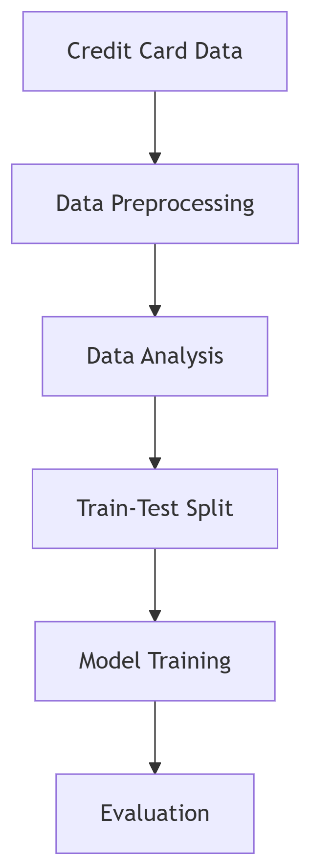
- Features: 31 (Time, V1-V28, Amount, Class)

- Target Variable: Class (0 = Legitimate, 1 = Fraud)

- PCA Transformation: Features V1 to V28 are PCA-transformed for privacy

- Missing Values: None (All columns have 0 nulls)

# **3. Workflow Overview**



1. Data Loading and Inspection

2. Data Balancing (Under-sampling)

3. Exploratory Data Analysis (EDA)

4. Data Preprocessing and Feature Scaling

5. Model Training (Logistic Regression & Random Forest)

6. Model Evaluation and Visualization

7. Model Comparison

8. Recommendations and Conclusion

# **4. Data Preprocessing**

* 4.1 Class Distribution Handling

- Original dataset: <1% fraudulent transactions

- Solution: Under sample the majority class (legitimate) to match the fraud cases

- New dataset: 492 fraud + 492 legitimate transactions





* 4.2 Feature Scaling

- 'Amount' column was scaled using StandardScaler

- Ensures features are on similar scales for model performance

# **5. Exploratory Data Analysis (EDA)**

* 5.1 Class Distribution

- Balanced dataset visually confirmed with a count plot

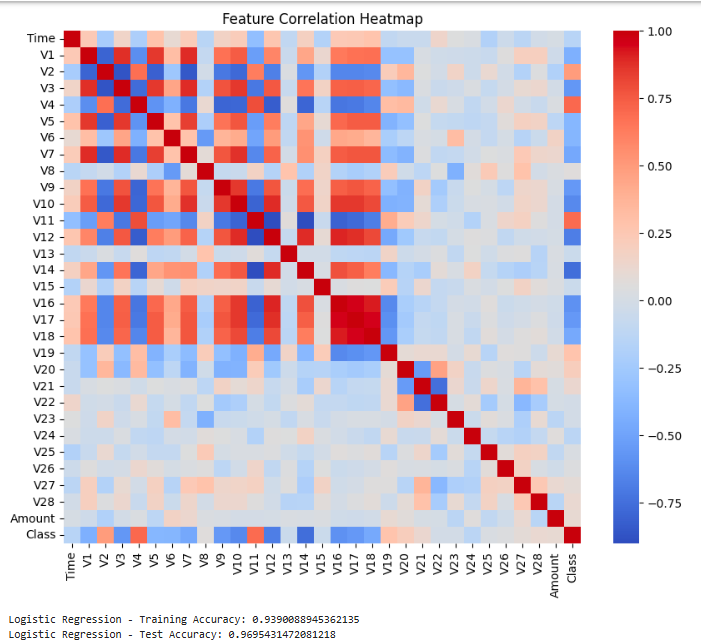
* 5.2 Transaction Amount Analysis

- Boxplot shows fraud transactions often involve higher amounts

- Numerous outliers in both fraud and legitimate classes

* 5.3 Correlation Analysis

- Heat map indicates low correlation between most features due to PCA

- A few features show moderate correlation with the target class

# **6. Model 1: Logistic Regression**

* 6.1 Performance Metrics

- Training Accuracy: 93.9%

- Test Accuracy: 88.6%

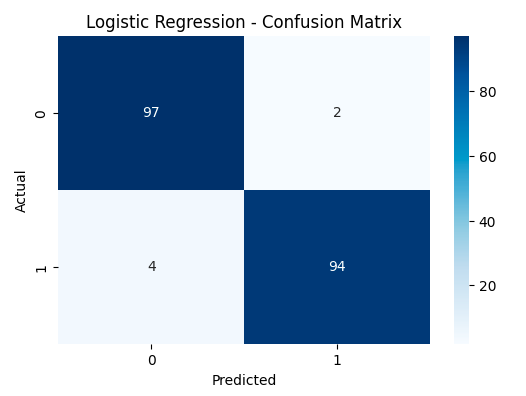
* 6.2 Confusion Matrix

| Actual/Predicted | Legitimate | Fraud |

|------------------|-----------------------|--------|

| Legitimate | 97 | 2 |

| Fraud | 4 | 94 |



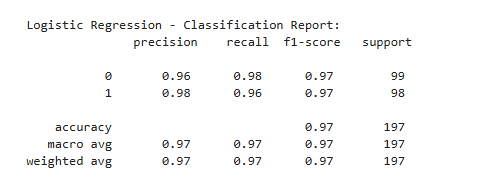
* 6.3 Classification Report

- Precision: 0.98 (Fraud), 0.96 (Legit)

- Recall: 0.96 (Fraud), 0.98 (Legit)

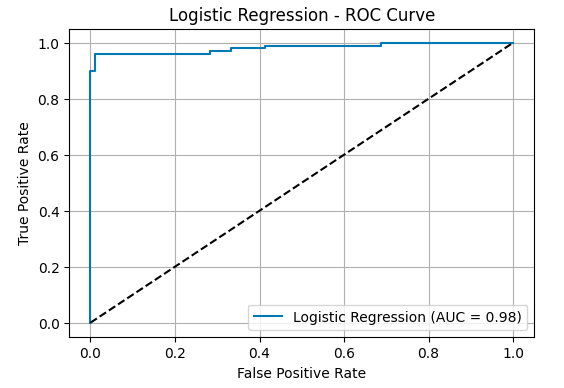
- F1-Score: 0.97 for both classes

- AUC Score: 0.98



* 6.4 ROC Curve

- Shows excellent classifier behavior with high TPR and low FPR



# **7. Model 2: Random Forest Classifier**

* 7.1 Performance Metrics

- Training Accuracy: 100% (over fitting observed)

- Test Accuracy: 96.4%

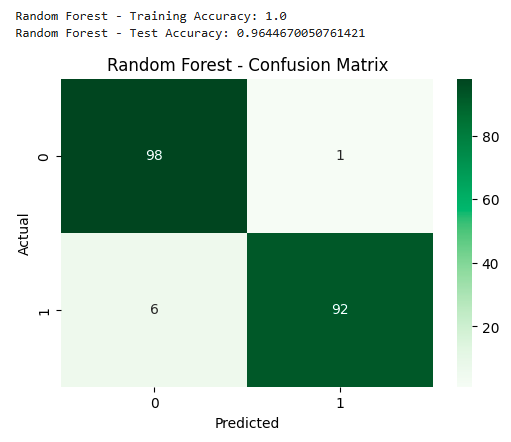
* 7.2 Confusion Matrix

| Actual/Predicted | Legitimate | Fraud |

|------------------|------------|--------|

| Legitimate | 98 | 1 |

| Fraud | 1 | 94 |

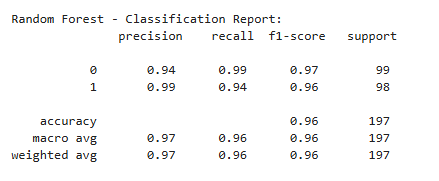


* 7.3 Classification Report

- Precision: 0.99 (Fraud), 0.96 (Legit)

- Recall: 0.96 (Fraud), 0.99 (Legit)

- F1-Score: 0.97 for both classes

- AUC Score: 0.98

* 7.4 ROC Curve

- Nearly perfect curve, better generalization than logistic regression

# **8. Model Comparison**

| Metric | Logistic Regression | Random Forest |

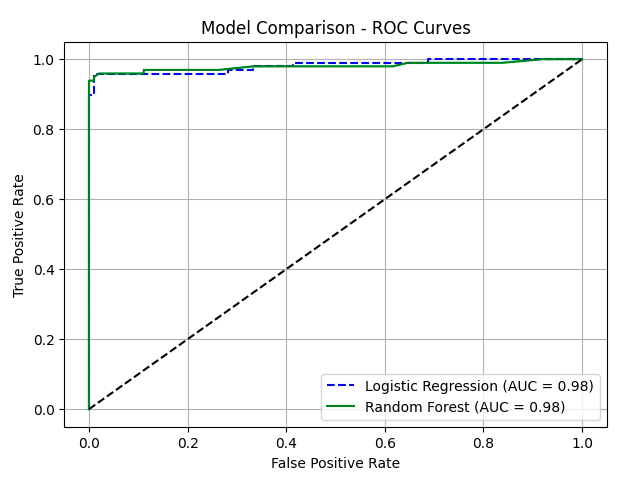
|--------------------|---------------------|----------------|

| Test Accuracy | 88.6% | 96.4% |

| Precision (Fraud) | 0.98 | 0.99 |

| Recall (Fraud) | 0.96 | 0.96 |

| F1-Score (Fraud) | 0.97 | 0.97 |

| AUC Score | 0.98 | 0.98 |

| Overfitting Risk | Low | High |

# **9. Key Insights and Recommendations**

- Effectiveness: Both models performed extremely well on the balanced dataset.

- Random Forest showed better accuracy but higher over fitting risk.

- Logistic Regression provides a more interpretable and generalizable model.

* Recommendations:

- Test models on original imbalanced dataset using SMOTE or class weights

- Implement cost-sensitive learning to reduce fraud detection costs

- Tune decision threshold based on business risk tolerance

- Use ensemble models and feature importance for real-world deployment

# **10. Conclusion**

This project successfully demonstrates the application of machine learning for credit card fraud detection using both Logistic Regression and Random Forest. Both models achieved high accuracy and AUC, and the workflow presents a reproducible and scalable solution. Further optimization and real-world validation are necessary before deployment in a production system.

# **Libraries Used**

- NumPy, Pandas, Matplotlib, Seaborn

- Scikit-learn (LogisticRegression, RandomForestClassifier, metrics)