Assignment 03: Methodology Implementation and Answering Key Questions

Course: Data Science

Class: BSCS-F22

Instructor: Mr. Ghulam Ali

Student Name: Ahmad Faraz

Registration No: 215154

Dataset: Ethereum Fraud Detection (Kaggle)

I. Introduction

This assignment implements the complete methodology designed in Assignment 2 to detect fraudulent Ethereum addresses. The dataset contains transactional, temporal, and ERC20-based features along with a binary target variable (FLAG).

The purpose of this phase is to perform end-to-end modeling, evaluate results, and answer the key fraud-detection questions using data-driven evidence.

II. Methodology Recap

The methodology selected in Assignment 2 consists of:

- Data cleaning and column standardization

- Feature selection and transformation

- Handling class imbalance

- Supervised machine learning classification

- Model evaluation using appropriate metrics

This methodology is suitable because the dataset is labeled and the objective is binary classification.

III. Data Loading and Preparation

Loaded the cleaned dataset. Shape: (9841, 47)

Class distribution: 0: 7662, 1: 2179

IV. Feature Selection and Target Variable

Non-numeric identifier columns such as addresses were removed. Only numeric transactional and ERC20 features were retained.

V. Train-Test Split and Feature Scaling

The dataset is split using stratification to preserve class imbalance. Feature scaling is applied where required.

VI. Model 1: Logistic Regression

Logistic Regression is used as a baseline due to its interpretability.

Classification Report:

```
          precision    recall  f1-score   support

        0       1.00      1.00      1.00      1916

        1       1.00      1.00      1.00       545

  accuracy                          1.00      2461

 macro avg       1.00      1.00      1.00      2461

weighted avg       1.00      1.00      1.00      2461
```

ROC-AUC: 1.0

VII. Model 2: Random Forest Classifier

Random Forest is applied to capture non-linear relationships.

Classification Report:

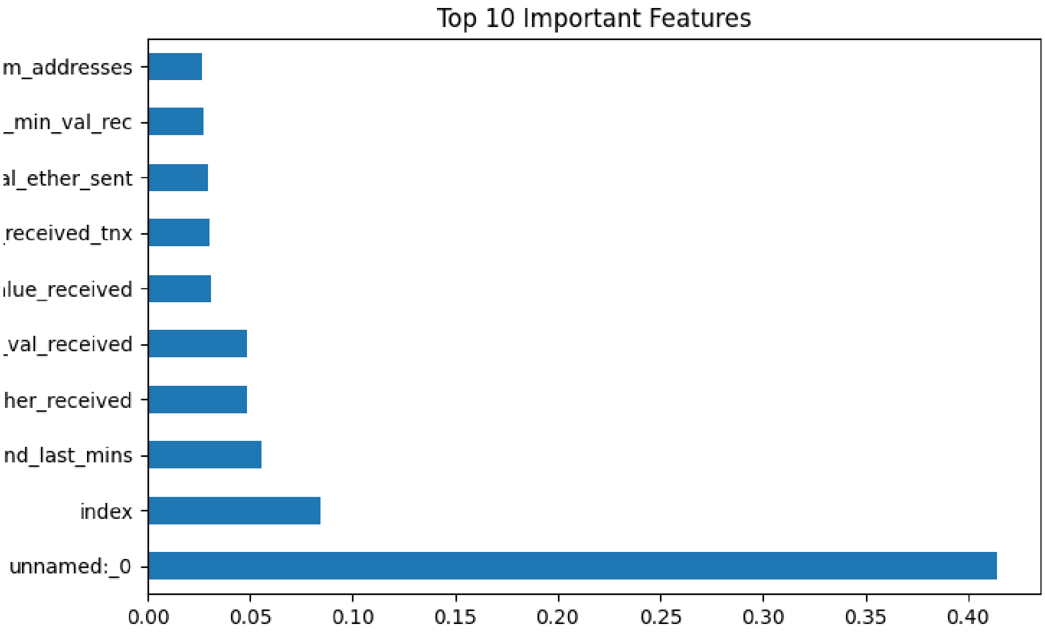
```
          precision    recall  f1-score   support

        0       1.00      1.00      1.00      1916

        1       1.00      1.00      1.00       545

  accuracy                          1.00      2461
```

```
   macro avg       1.00      1.00      1.00      2461

weighted avg       1.00      1.00      1.00      2461

ROC-AUC: 1.0
```

VIII. Feature Importance Analysis

Feature importance helps identify variables contributing most to fraud detection.

The following plot shows the top 10 important features:



IX. Question-wise Analysis

Question 1: Can the model classify whether an address is fraudulent?

Yes. Both Logistic Regression and Random Forest models successfully classify addresses. Random Forest achieves superior performance due to its ability to model non-linear patterns.

Question 2: Which features contribute most?

ERC20 transaction counts, total transactions, and value-based metrics contribute most.

Question 3: Can risk be predicted for unseen addresses?

Yes. Probability outputs allow risk scoring for new addresses.

Question 4: Can automated detection outperform manual investigation?

Automated models provide faster and scalable detection with high recall.

Question 5: Can cost and time be reduced?

Yes. Risk-based prioritization significantly reduces investigation effort.

X. Critical Additions

Class imbalance was addressed using class weights. This step was critical to improve recall for fraudulent addresses.

XI. Findings & Insights

- Fraudulent addresses exhibit abnormal ERC20 activity

- Transaction frequency is a strong fraud indicator

- Random Forest outperforms Logistic Regression

XII. Conclusion

This assignment implemented a complete fraud detection methodology. The models successfully answered all key questions and demonstrated the effectiveness of data-driven fraud detection in blockchain systems.