TECHNICAL REPORT

Healthcare Provider Fraud Detection

Author: Omar Ossama – ID 7001032

1. Introduction

This project focuses on detecting fraudulent Medicare healthcare providers using machine learning.

The goal is to analyze multi-source healthcare data, engineer provider-level features, handle class imbalance,

train predictive models, and evaluate their performance.

2. Data Understanding

The dataset includes four files:

- Train_Beneficiarydata.csv

- Train_Inpatientdata.csv

- Train_Outpatientdata.csv

- Train_Labels.csv

BeneID links beneficiaries to claims, and Provider links claims to fraud labels.

3. Data Preparation & Feature Engineering

Data from inpatient, outpatient, and beneficiary files were merged into unified claim-level data.

Then, data was aggregated to provider-level using:

- Counts (claims, beneficiaries)

- Financial statistics (sum, mean, max claim amounts)

- Ratios (inpatient/outpatient share)

- Durations (claim duration, treatment duration)

4. Class Imbalance Strategy

Fraud accounts for around 10% of providers. To address this:

- SMOTE oversampling was used.

- Class weights were applied.

Metrics prioritized: Recall, F1-score, PR-AUC.

5. Modeling

Three models were trained:

- Logistic Regression
- Random Forest
- XGBoost

Hyperparameter tuning was performed using RandomizedSearchCV. Models were compared on precision, recall, F1, ROC-AUC, and PR-AUC.

6. Evaluation

Evaluation included:

- Confusion matrices
- ROC curves
- Precision-Recall curves
- Feature importance

XGBoost showed the best balance between recall and PR-AUC.

7. Error Analysis

False positives and false negatives were analyzed to understand model errors and real-world impact.

High-risk provider patterns were identified through feature importance insights.

8. Conclusion

The final recommended model is XGBoost for its strong recall, interpretability via feature importance,

and robustness with imbalanced data.