Healthcare Provider Fraud Detection — Final Report Ahmed Hisham 13007014

1. Introduction

Healthcare fraud causes billions in losses every year.
This project analyzes provider-level claim patterns and uses ML to identify providers likely to commit fraud.

Dataset: CMS Healthcare Provider Fraud Detection
Model target: **PotentialFraud (Yes/No)**

The final workflow includes:

- Data exploration
- Feature engineering
- Modeling
- Evaluation
- Interpretation

---

2. Dataset Overview

We used 4 datasets:

1. Beneficiary details
2. Inpatient claims
3. Outpatient claims
4. Fraud labels

After cleaning and aggregation:

- Total providers: **5,410**
- Fraudulent providers: **506**
- Non-fraudulent providers: **4,904**

The dataset is **imbalanced**, affecting model design.

---

3. Feature Engineering

We aggregated inpatient & outpatient claims *per provider*.

Inpatient features:

- LOS (mean, max, min)
- Reimbursement mean & sum
- Claim count

Outpatient features:

- Reimbursement mean & sum
- Claim count

Final dataset:
9 numerical features + PotentialFraud label

Saved as:
`provider_features_final.csv`

---

4. Modeling

Two models were trained:

Logistic Regression

- Serves as baseline

- Works well with scaled data
  Results:
  Accuracy: **94.5%**
  F1 Score: **0.66**

Random Forest Classifier

- Best performance

- Handles nonlinear fraud patterns

- Robust to imbalanced data
  Results:
  Accuracy: **94%**
  Recall (fraud): **0.677**
  Precision (fraud): **0.688**
  F1 Score: **0.682**

---

5. Evaluation

5.1 Confusion Matrix

Model detects most fraudulent providers
with minimal false positives.

5.2 ROC Curve

High AUC value shows excellent class separation.

5.3 Feature Importance

Top indicators:

1. Maximum LOS

2. Total inpatient reimbursement
3. Outpatient claim count
4. Total outpatient reimbursement

These match real-world fraud patterns.

---

6. Business Value

Deploying this model:

- Flags suspicious providers for audit
- Reduces fraudulent payouts
- Supports insurance & government healthcare oversight
- Optimizes investigative resources

---

7. Conclusion

This project demonstrates a full ML pipeline for fraud detection:

- Data prep
- Feature engineering
- Model training
- Evaluation
- Interpretation

The Random Forest model performs strongly and can support real-world fraud surveillance.

Future improvements:

- XGBoost / LightGBM
- SMOTE for balancing
- Diagnosis code embeddings
- Deployment as API for real-time scoring