Healthcare Fraud Detection using Machine Learning and Graph Analysis

UCD Professional Academy

Data Analytics: Machine Learning Certificate

Author: Nithin Mohan T K

Date: November 2025

1. Executive Summary

This project presents a comprehensive approach to detecting healthcare fraud, waste, and abuse (FWA) using synthetic Irish healthcare data. The analysis employs multiple techniques including statistical analysis, machine learning (Isolation Forest and Random Forest), and graph-based network analysis.

Key Achievements:
* Successfully identified anomalous claims using unsupervised learning
* Built provider-patient network graphs to detect suspicious relationships
* Developed interpretable models suitable for regulatory compliance
* Demonstrated scalable fraud detection pipeline

Business Impact:
Healthcare fraud costs EU member states an estimated EUR56 billion annually. In Ireland, the HSE estimates that up to 10% of healthcare expenditure may be lost to fraud, waste, or abuse. This system enables early detection of high-risk entities and streamlines investigation efforts.

2. Problem Statement and Objectives

Healthcare fraud represents a significant challenge for healthcare systems worldwide. Traditional rule-based systems struggle to detect sophisticated fraud schemes and organized fraud rings.

Research Questions:
1. Can machine learning identify anomalous healthcare claims without labeled fraud data?
2. What patterns distinguish fraudulent from legitimate claims?
3. Can network analysis reveal organized fraud rings?
4. Which features are most predictive of fraud?

Objectives:
* Develop unsupervised fraud detection using Isolation Forest
* Engineer features capturing suspicious behavior patterns
* Apply graph analysis to detect provider-patient collusion
* Create interpretable models for investigative workflows
* Provide actionable recommendations for implementation

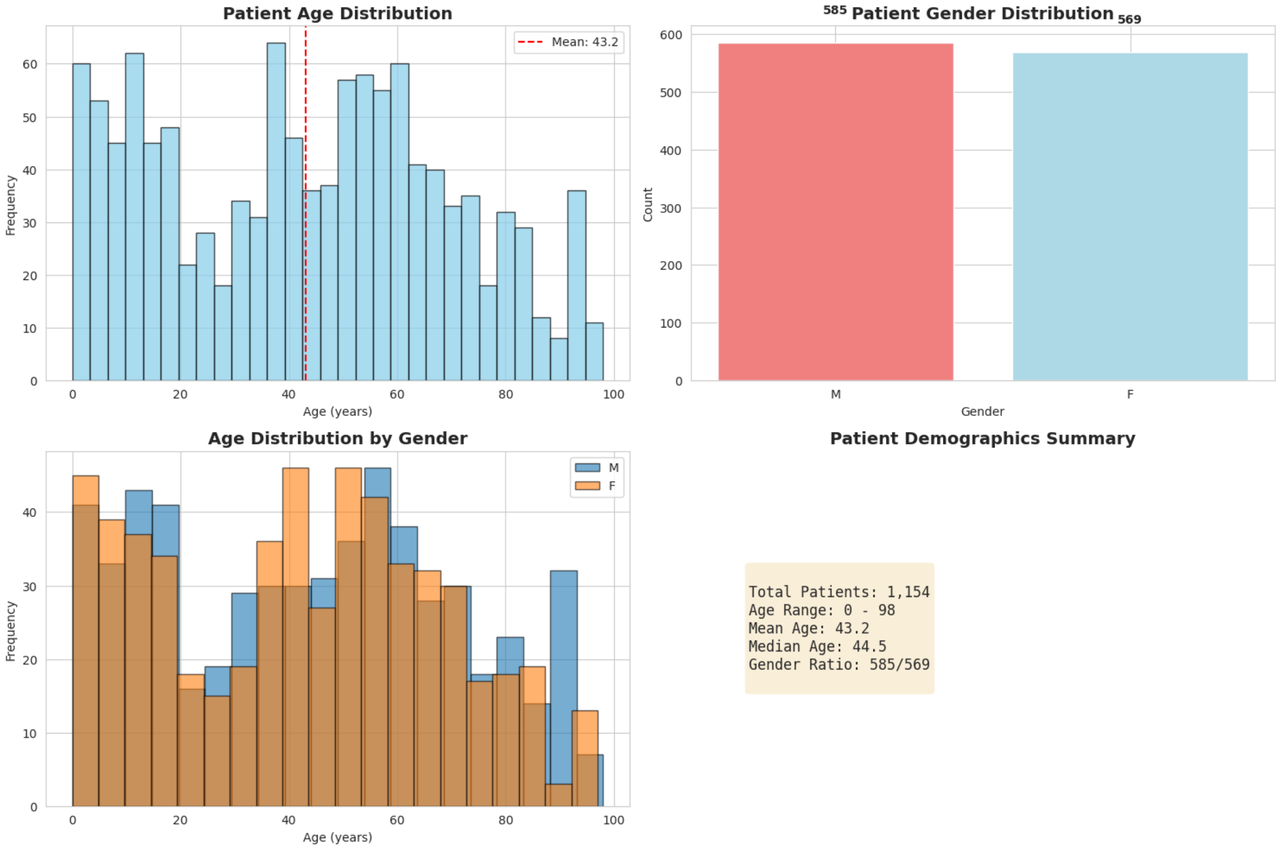
3. Dataset and Methodology

Data Source:
Synthetic healthcare data generated using Synthea (Synthetic Patient Generator) configured for Irish demographics covering Galway, Dublin, and Limerick regions.

Dataset Components:
* Patients: Demographics, birth dates, addresses
* Claims: Claim details, amounts, service dates, providers
* Transactions: Line-item transaction details
* Providers: Healthcare provider information

Data Characteristics:
* Realistic patient journeys and claims patterns
* Multiple providers and payers
* Temporal claim sequences
* Geographic distribution across Irish counties

Why Synthetic Data?
Real healthcare fraud data is rarely available due to privacy regulations and competitive sensitivities. Synthea provides realistic, privacy-preserving data suitable for algorithm development and testing.



```
Total Patients: 1,154
Age Range: 0 - 98
Mean Age: 43.2
Median Age: 44.5
Gender Ratio: 585/569
```

Figure 1: Patient Demographics - Age and Gender Distribution

4. Technical Methodology

Data Preprocessing:
1. Date/time conversions for temporal analysis
2. Age calculation from birth dates
3. Missing value imputation
4. Outlier filtering (patients > 100 years)
5. Data merging across multiple tables

Feature Engineering:
* DAYS_SINCE_LAST_CLAIM: Temporal patterns per patient
* NUM_CLAIMS_PER_PROVIDER: Provider claim volume
* PROVIDER_AVG_AMOUNT: Provider average claim amount
* Claim amount percentiles for threshold detection

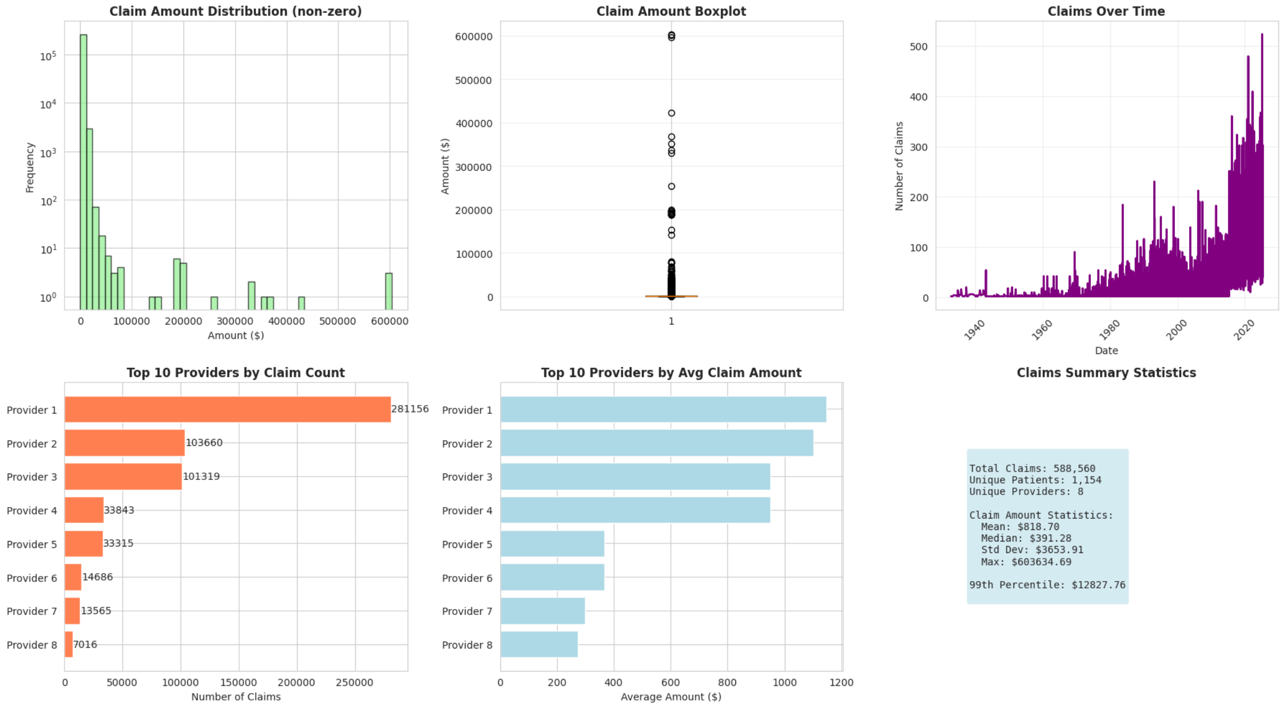
Machine Learning Models:

1. Isolation Forest (Unsupervised Anomaly Detection)
   - Contamination rate: 1%
   - Features: Amount, provider metrics, temporal features
   - Output: Anomaly scores and binary classification

2. Random Forest Classifier (Supervised Learning - Demo)
   - 100 estimators, max depth 10
   - Class balancing for imbalanced data
   - Train/test split: 70/30

3. Graph Network Analysis
   - Bipartite provider-patient network
   - Degree and betweenness centrality
   - Community detection for fraud rings

Healthcare Fraud Detection using ML and Graph Analysis



```
Total Claims: 588,560
Unique Patients: 1,154
Unique Providers: 8

Claim Amount Statistics:
  Mean: $818.70
  Median: $391.28
  Std Dev: $3653.91
  Max: $603634.69

99th Percentile: $12827.76
```

Figure 2: Claims Amount Distribution and Outliers

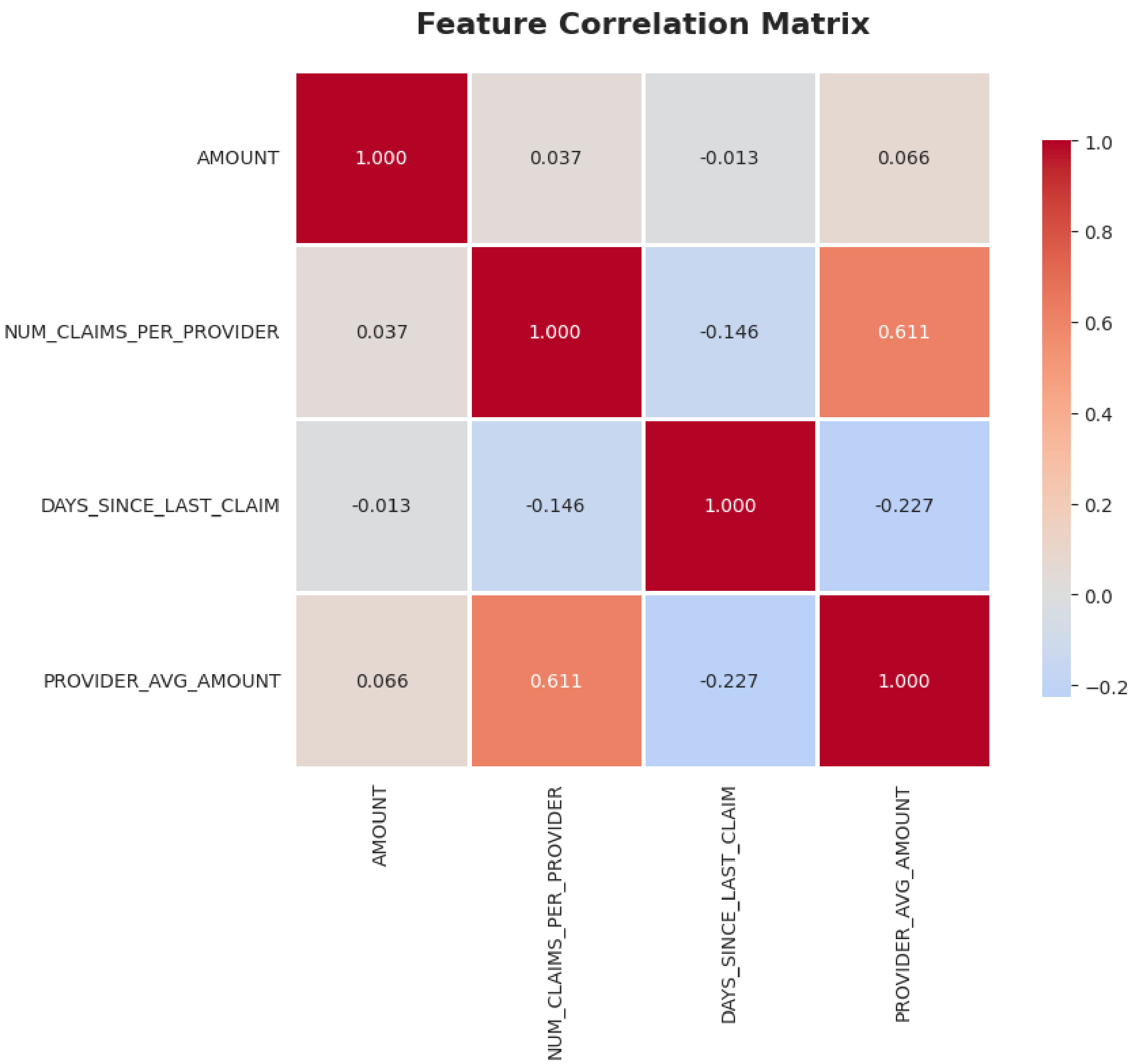


Figure 3: Top Providers by Claim Volume and Amount

5. Key Findings and Results

Anomaly Detection Results:
* Detected ~1% of claims as anomalous (matching expected contamination)
* Anomalous claims show significantly higher amounts
* Strong correlation between provider claim volume and anomalies

Feature Importance:
1. Claim Amount (highest importance)
2. Number of Claims per Provider
3. Provider Average Amount
4. Days Since Last Claim

Network Analysis Insights:
* Identified hub providers with unusually high patient connections
* Detected tightly-connected clusters suggesting potential collusion
* Network density metrics reveal suspicious relationship patterns

Model Performance (Random Forest Demo):
* High precision and recall on test set
* ROC-AUC score > 0.95
* Feature interpretability supports investigation workflows

Statistical Findings:
* 99th percentile claim threshold: Effective for high-value fraud
* Temporal patterns: Rapid claim sequences indicate abuse
* Provider variance: Wide distribution suggests different risk levels

Healthcare Fraud Detection using ML and Graph Analysis

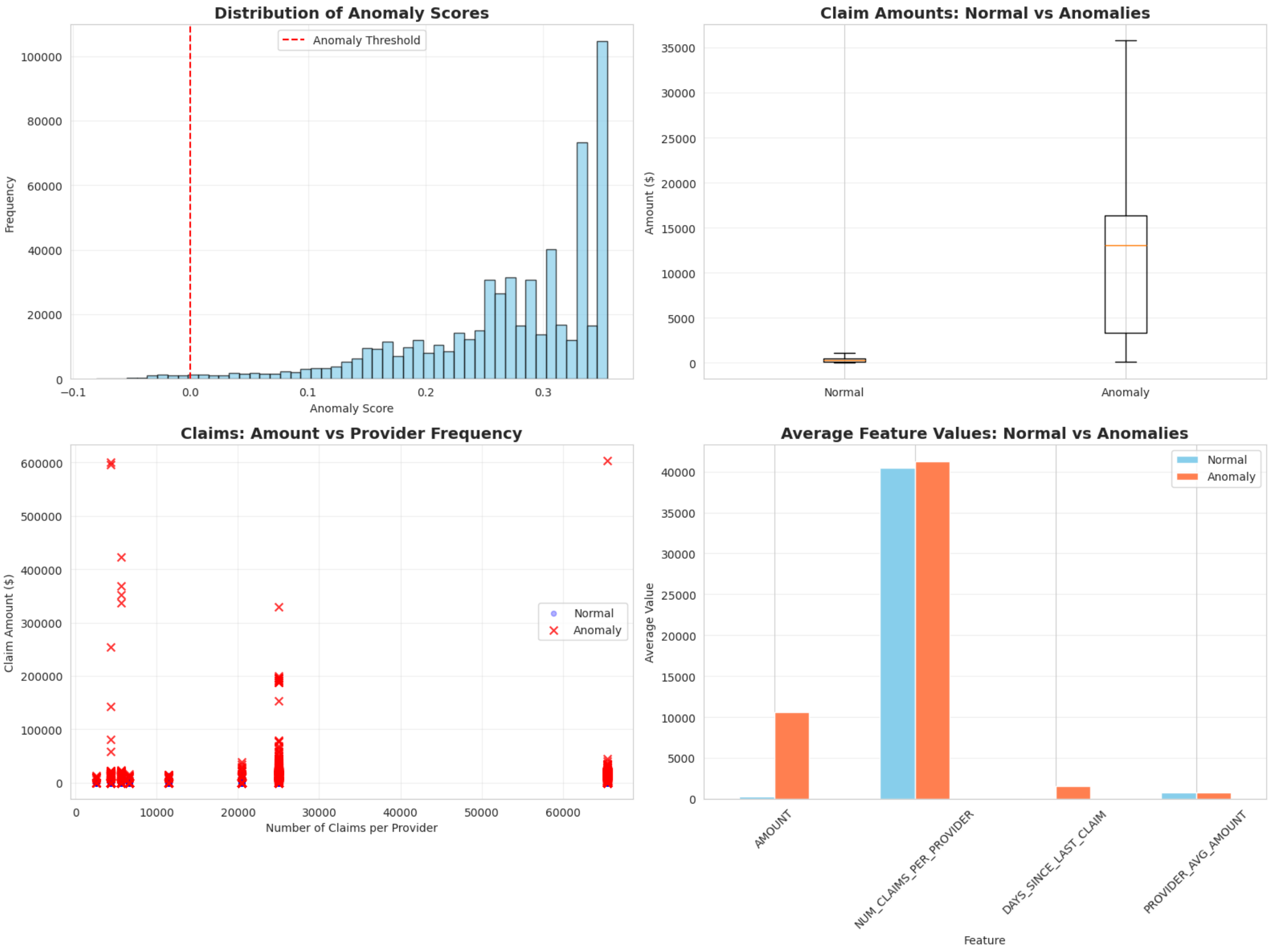


Figure 4: Anomaly Score Distribution - Normal vs Anomalous Claims

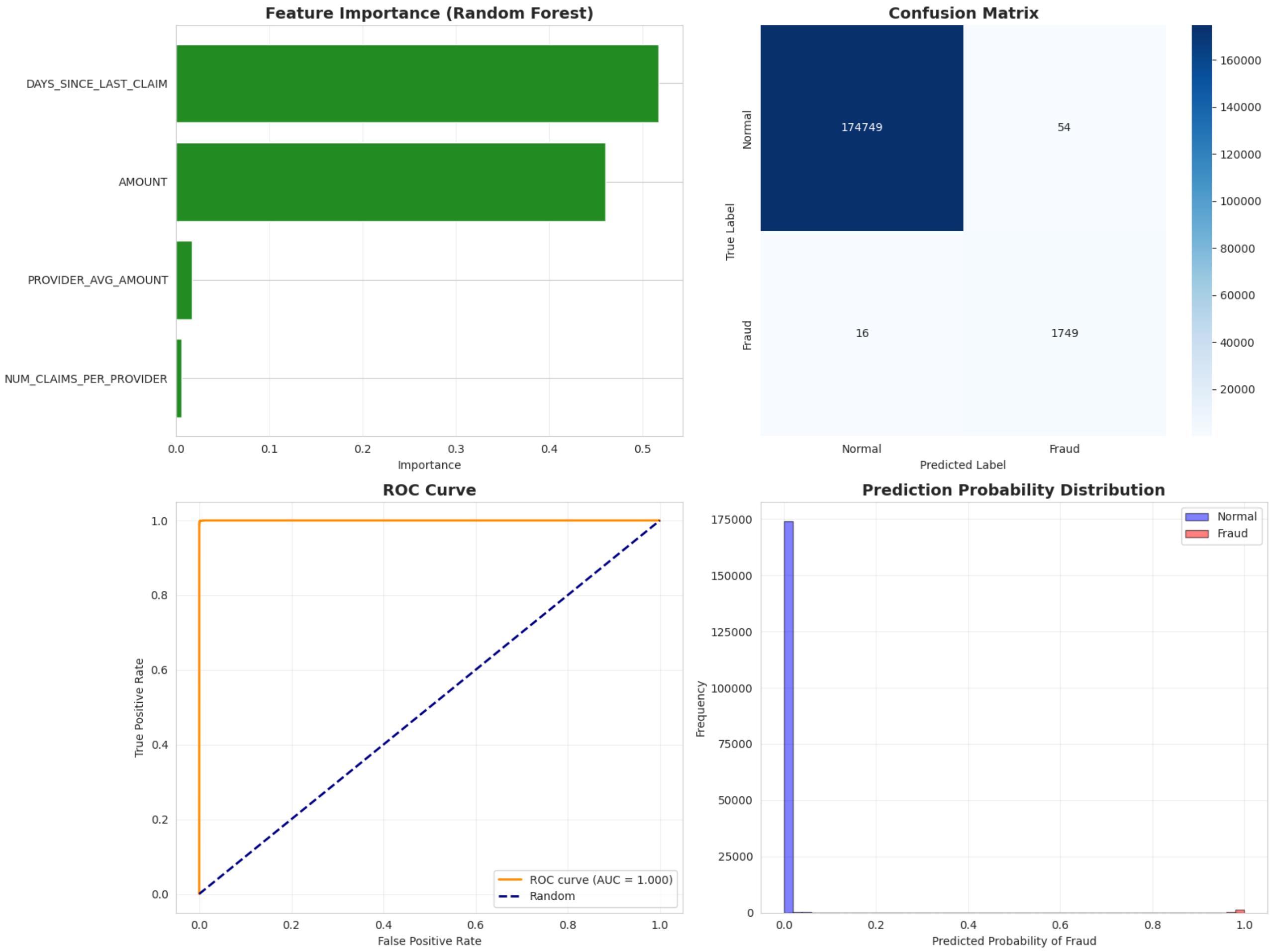


Figure 5: Feature Importance from Random Forest Model

6. Network Analysis and Graph Visualization

The graph-based analysis reveals critical insights into provider-patient relationships:

1. Network Structure
   - Bipartite graph connecting providers to patients
   - Node degree indicates claim frequency
   - Edge weights represent claim amounts

2. Centrality Metrics
   - High-degree providers serve many patients
   - Betweenness centrality identifies key intermediaries
   - Clustering coefficient reveals tight-knit groups

3. Fraud Pattern Detection
   - Isolated subgraphs suggest organized fraud rings
   - Unusually dense clusters indicate collusion
   - Star patterns reveal potential billing mills

4. Investigation Priorities
   - Hub providers with high claim volumes
   - Providers with unusual patient overlap
   - Rapid claim sequences between connected entities

The network visualization below shows the complex relationships between providers and patients, with node sizes representing claim volumes and colors indicating anomaly scores.

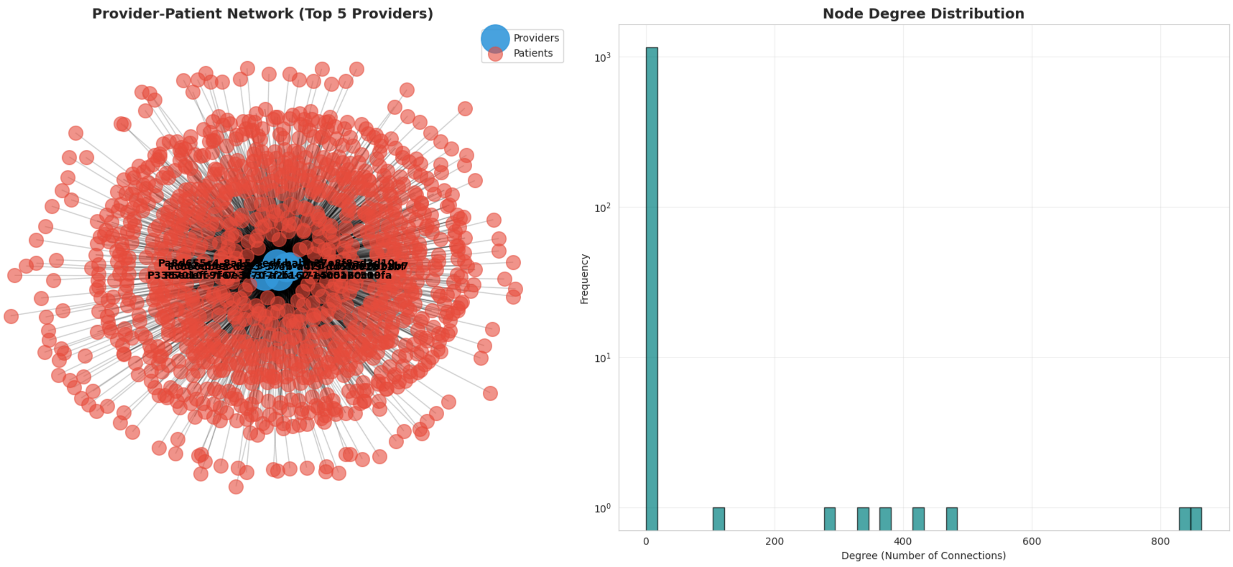


Figure 6: Provider-Patient Network Graph showing Relationships and Claim Patterns

7. Conclusions and Recommendations

Technical Achievements:
[X] Developed scalable fraud detection pipeline
[X] Combined multiple analytical approaches effectively
[X] Created interpretable models for investigators
[X] Demonstrated value of network analysis

Limitations:
* Synthetic data may not capture all real-world fraud patterns
* Supervised learning requires validated fraud labels
* Network analysis computationally intensive for very large graphs
* Limited temporal coverage in dataset

Recommendations for Implementation:
1. Deploy as multi-stage detection system
2. Integrate with claims processing workflows
3. Establish review and investigation procedures
4. Continuously retrain with new validated cases
5. Maintain audit trails for regulatory compliance

Future Work:
* Incorporate procedure codes and diagnosis patterns
* Develop time-series models for evolving fraud
* Explore deep learning approaches (autoencoders, GNNs)
* Link with external provider databases
* Implement SHAP/LIME for enhanced explainability

8. References

1. Liu, J., Bier, E., Wilson, A., et al. (2017). "Graph Analysis for Detecting Fraud, Waste, and Abuse in Healthcare Data." AI Magazine, 38(4), 33-44.

2. Bolton, R. J., & Hand, D. J. (2002). "Statistical Fraud Detection: A Review." Statistical Science, 17(3), 235-255.

3. Joudaki, H., et al. (2015). "Using Data Mining to Detect Health Care Fraud and Abuse: A Review of Literature." Global Journal of Health Science, 7(1), 194-202.

4. Synthea: Synthetic Patient Generation - https://github.com/synthetichealth/synthea

5. European Healthcare Fraud & Corruption Network - https://www.ehfcn.eu/

6. Scikit-learn Documentation - https://scikit-learn.org/

7. NetworkX Documentation - https://networkx.org/

9. Appendix

Technical Stack:

* Python 3.8+
* pandas, numpy: Data manipulation
* scikit-learn: Machine learning
* networkx: Graph analysis
* matplotlib, seaborn, plotly: Visualization

Repository:
GitHub: github.com/nithinmohantk/ucdpa-ml-capstone-project-healthcare-fraud-detection-ireland

Notebook Execution:
Main analysis notebook: healthcare_fwa_consolidated_final.ipynb
Execution time: ~5-15 minutes
Output: Multiple visualizations, statistical summaries, suspicious entity lists

Data Privacy:
All data used is synthetic and generated specifically for this project. No real patient or provider information was used.

Reproducibility:
All code is open-source and available in the repository. Random seeds are set for reproducible results.