Fraudulent Insurance Claim Detection Report

Problem Statement:

Identify whether an insurance claim is fraudulent using various policy, incident, and customer-related variables.

Approach:

1. Data Cleaning
   - Removed irrelevant and identifier columns
   - Handled missing values in categorical fields

2. EDA
   - Visualized distributions, correlations, and class imbalance
   - Found class imbalance and weak correlations

3. Feature Engineering
   - One-hot encoding of categorical variables
   - Standardized numerical features

4. Model Building
   - Logistic Regression with RFECV for feature selection
   - Random Forest with class_weight='balanced'

5. Evaluation
   - Logistic Regression performed better with 85.3% accuracy

- Recall for fraud detection: 86.5% (vs. 14.9% for Random Forest)

- ROC AUC: 0.84

Recommendations:

- Use logistic regression for real-time fraud detection

- Improve data features and collect more fraud-specific variables

- Automate alerts on high-risk claims