Fraudulent Claim Detection

Problem Statement

Insurance fraud is a serious issue in the financial and insurance sectors, leading to significant financial losses annually. The primary objective of this project is to build a machine learning model that can detect potentially fraudulent insurance claims, enabling early intervention and reducing financial losses.

Methodology

The project follows a structured data science pipeline:

1. Data Collection: Dataset contains personal, financial, and claim-specific information.
2. Data Preprocessing: Null checks, label encoding, feature scaling, class balance checks.
3. Exploratory Data Analysis (EDA): Visual relationships between features and target.
4. Model Building: Trained Logistic Regression, Random Forest, Gradient Boosting, and Decision Tree.
5. Model Selection & Evaluation: Used accuracy, precision, recall, F1-score, ROC-AUC.

Techniques Used

- Data Handling: pandas, numpy
- Data Visualization: matplotlib, seaborn
- Modeling: scikit-learn (LogisticRegression, RandomForest, GradientBoosting, DecisionTree)
- Evaluation: Confusion Matrix, Classification Report, ROC Curve, Cross-validation

Visualizations

- Class Distribution Bar Plot
- Feature Correlation Heatmap
- Boxplots & Histograms

- ROC Curves for Model Comparison

Insights

- Class imbalance with fewer fraudulent claims
- Key features like claim amount, age, deductible linked to fraud
- Ensemble models performed better than logistic regression
- Gradient Boosting had the highest AUC

Actionable Outcomes

- Deploy Gradient Boosting for real-time fraud detection
- Create alert systems for suspicious claims
- Collect more fraud samples or use SMOTE
- Enhance features using domain knowledge