Claims Fraud Detection System - Project Report

1. Project Overview

The Claims Fraud Detection System project was developed with the objective of building a predictive model to detect fraudulent insurance claims. Fraudulent activities in insurance lead to significant financial losses, making fraud detection a critical business problem. This project simulates a real-world use case using a synthetic insurance claims dataset.

2. Dataset Description

The dataset used in this project was synthetically generated to represent real-world insurance claim scenarios. It contains 1,000 records with the following features: - claim_id: Unique claim identifier - customer_age: Age of the claimant - policy_type: Type of insurance policy (Auto, Health, Home) - claim_amount: Amount claimed (numeric) - claim_history: Number of previous claims by the customer - incident_type: Type of incident (Theft, Accident, Natural Disaster, Fire) - reported_delay: Days between incident occurrence and reporting - is_fraud: Target variable (1 = Fraudulent, 0 = Legitimate)

3. Responsibilities

The project workflow consisted of the following responsibilities: 1. Data Cleaning: Checked for missing values, removed irrelevant identifiers, and standardized formats. 2. Exploratory Data Analysis (EDA): Conducted visual and statistical analysis to understand feature distributions, identify outliers, and study correlations. 3. Feature Engineering: Encoded categorical variables (policy_type, incident_type) and derived risk-related patterns. 4. Model Development: Applied Logistic Regression and Decision Tree classifiers to predict fraud likelihood. 5. Evaluation: Compared models using accuracy, confusion matrices, and classification reports. 6. Insights & Results: Identified key fraud-indicating features and quantified improvements in fraud detection accuracy.

4. Exploratory Data Analysis

EDA provided several insights into the dataset: - Fraudulent claims were relatively fewer than legitimate claims (class imbalance). - Fraudulent claims often involved higher claim amounts, no prior claim history, and delayed reporting. - Boxplots highlighted outliers in claim_amount, which were often associated with fraud. - Correlation analysis revealed meaningful relationships between claim_amount, claim_history, and fraud.

5. Model Development and Evaluation

Two classification models were applied: - Logistic Regression: Served as a baseline linear model. It achieved reasonable accuracy but struggled with capturing non-linear fraud patterns. - Decision Tree Classifier: Captured non-linear relationships effectively. Provided higher accuracy and interpretability via feature importance scores. Evaluation Metrics: - Accuracy: Both models achieved strong performance, with Decision Trees outperforming Logistic Regression. - Precision & Recall: Decision Trees showed better recall for fraudulent claims, crucial in reducing false negatives (missed fraud cases). - Feature Importance: Claim amount, claim history, and reporting delay were

the strongest predictors of fraud.

6. Results and Insights

Key outcomes of the project: - Improved claims validation accuracy by simulating early detection of fraudulent claims. - Identified fraud indicators: high claim amounts, no prior history, and reporting delays. - Demonstrated the importance of balancing predictive accuracy with business needs (minimizing false negatives). Business Impact: - Early fraud detection can significantly reduce financial losses. - Enhanced efficiency in claims processing by flagging high-risk claims for investigation.

7. Conclusion

The Claims Fraud Detection System successfully demonstrated how data-driven approaches can address fraud detection in insurance. Through data cleaning, EDA, feature engineering, and machine learning models, the project showcased a practical workflow that can be scaled with real-world datasets. Decision Trees proved to be a strong candidate model, offering both accuracy and interpretability. Future Work: - Apply advanced models such as Random Forests, Gradient Boosting, or Neural Networks. - Address class imbalance using oversampling (SMOTE) or anomaly detection methods. - Deploy the model into a real-time claims processing system for proactive fraud prevention.