**Insurance Claim Fraud Detection: A Machine Learning Approach BY**

***SURYAPRAKASH G***

Insurance claim fraud is a critical issue in the insurance industry, causing substantial financial losses. This project aims to develop a machine learning model that can accurately identify fraudulent insurance claims.

**Problem Statement**

The insurance industry suffers significant financial losses due to fraudulent claims. This study aims to develop a machine learning model to accurately detect fraudulent claims by analyzing various features.

**Data Analysis**

The dataset used contains 39 columns, providing demographic information about the insured individuals, policy details, incident details, and the outcome of the claim (whether it is fraudulent or not). The dataset has 999 entries.

**Exploratory Data Analysis (EDA) Observations**

* Missing Values: There are no missing values in the dataset.
* Data Types: The dataset includes both numerical and categorical variables.
* Variable Distributions: Histograms can be used to analyze the distribution of variables such as `months\_as\_customer`, `age`, and `policy\_annual\_premium`.
* Correlation Analysis: This can help identify relationships between different variables.

**Pre-processing Pipeline**

* + Handling Missing Values: No action required as there are no missing values.
  + Encoding Categorical Variables: Categorical variables such as `policy\_state`, `policy\_csl`, `insured\_sex`, and `insured\_education\_level` need to be encoded using one-hot encoding or label encoding.
  + Scaling Numerical Variables: Numerical variables should be scaled using standardization or normalization techniques.

**Building Machine Learning Models**

* Feature Selection : Use techniques like mutual information or recursive feature elimination to select relevant features.
* Model Selection: Train and evaluate various machine learning models, including logistic regression, decision trees, random forests, and support vector machines.
* Hyperparameter Tuning: Employ grid search or random search to tune model hyperparameters.
* Model Evaluation: Evaluate model performance using metrics such as accuracy, precision, recall, and F1 score.

**Conclusion**

The best performing model will be chosen based on its performance on the test dataset. This model can then be used to predict fraudulent insurance claims, helping insurance companies reduce financial losses by identifying and preventing fraudulent activities.

This study aims to provide a robust approach to fraud detection, contributing to the efficiency and security of the insurance industry.