**Project Name: Health Insurance Claim Fraud Detection**

**Problem Statement:**

**Fraudulent claims in health insurance lead to significant financial losses for insurance companies. The goal of this project is to build a machine learning model that can effectively identify fraudulent claims and reduce the risk of financial fraud.**

**Dataset Description**

The dataset used for this project is derived from the **PaySim1 dataset**, which simulates financial transactions to detect fraudulent activities. It contains various features representing transaction details.

**Key Features in the Dataset:**

* step: Time step at which the transaction was recorded.
* type: Type of transaction (e.g., CASH-IN, CASH-OUT, TRANSFER, etc.).
* amount: The amount of money involved in the transaction.
* nameOrig: Identifier for the origin account.
* oldbalanceOrg: Initial balance of the origin account before the transaction.
* newbalanceOrig: Balance of the origin account after the transaction.
* nameDest: Identifier for the destination account.
* oldbalanceDest: Initial balance of the destination account before the transaction.
* newbalanceDest: Balance of the destination account after the transaction.
* isFraud: Target variable indicating if the transaction was fraudulent (1) or not (0).
* isFlaggedFraud: Flag indicating if a transaction was marked as potentially fraudulent.

**Dataset Preprocessing Steps:**

* Checked for missing values and handled inconsistencies.
* Selected relevant features for fraud detection.
* Split the dataset into **training and testing sets** to evaluate model performance.

**Implementation and Model Comparison**

**1st Model: Random Forest Classifier (Project 1)**

* Used a **Random Forest Classifier** to detect fraudulent health insurance claims.
* Processed **100,000** sampled records from the dataset.
* Dataset preprocessing steps:
  + Checked for missing values and data inconsistencies.
  + Selected relevant features: step, amount, oldbalanceOrg, newbalanceOrig, oldbalanceDest, newbalanceDest.
  + Target variable: isFraud.
* Split the dataset into training and testing sets.
* Trained the **Random Forest Classifier**.
* Evaluated the model using:
  + **Accuracy Score**: (to be determined by execution)
  + **Classification Report**: Precision, Recall, F1-score analysis.

**2nd Model: Multiple Algorithm Comparison (Multiple Algorithm Project)**

* Implemented three different models to compare performance:
  1. **Random Forest Classifier** → Accuracy: (to be determined)
  2. **Logistic Regression** → Accuracy: (to be determined)
  3. **XGBoost Classifier** → Accuracy: (to be determined)
* Kept the same dataset preprocessing and feature selection steps.
* Split the dataset into training and testing sets.
* Trained all three models and compared their performances using:
  1. **Accuracy Score** for each model.
  2. **Classification Report** to analyze precision, recall, and F1-score.
  3. **Visualization of accuracy scores** using bar charts.

**Comparison and Conclusion**

* The first model used only one algorithm (**Random Forest**) and achieved an initial accuracy.
* The second model tested multiple algorithms and identified the best-performing one.
* Based on accuracy results, we determine which model is most effective for fraud detection in health insurance claims.
* **Final Recommendation:** The model with the highest accuracy and best classification metrics should be used for deployment in fraud detection systems.