**Fraud Detection Assignment – VISA Assignment(Puranjay Kwatra)**

**Objective**

The goal is to identify fraudulent users based on their activity records, where each user has multiple transactions (activity records) described by 166 features. A user is labeled fraudulent if **any** of their activity records is fraudulent.  
**Bonus**: Predict fraud at the **activity level** as well.

**Data Understanding & Preparation**

* Loaded activity-level data with **476 records** and **92 unique users**.
* Each record includes features from f1 to f166 and a user-id.
* Merged with externally provided **user-level fraud labels** to build the classification target.

**Feature Aggregation**

* Since each user has multiple activities, features were aggregated per user using:
  + **Mean**, **Maximum**, and **Standard Deviation**
* This resulted in a single feature vector per user, producing **498 aggregated features** (166 × 3).

**Model Selection**

* Trained and evaluated the following models:
  + **Random Forest** – Works well with complex feature spaces; used as a baseline.
  + **Logistic Regression** – Lightweight, interpretable model.
  + **XGBoost** – High-performance model, especially effective with imbalanced data.
* Imbalance handling techniques used:
  + class\_weight='balanced' for Random Forest and Logistic Regression

**Hyperparameter Tuning**

* Applied **GridSearchCV** on XGBoost to tune key parameters such as:
  + max\_depth, learning\_rate, n\_estimators, and subsample

**Evaluation Metrics**

Given the class imbalance, the following evaluation metrics were prioritized:

* **Accuracy**
* **ROC AUC Score**

Visualizations generated:

* **Confusion Matrix** – Shows actual vs predicted labels
* **ROC Curve** – Evaluates model performance across different thresholds

**Bonus Task – Activity-Level Predictions**

* Built a separate **XGBoost classifier** to classify each **individual activity record**.
* Used the original 166 features (f1–f166) and activity-level fraud labels.
* Evaluated using:
  + **F1 Score**
  + **Confusion Matrix**
* Created a helper function to extract and display predictions for a specific user, enabling micro-level analysis and validation.