PROJECT OVERVIEW

Problem Statement: This project addresses the challenge of detecting fraudulent transactions using machine learning classification techniques. The goal is to build a predictive model that accurately identifies fraudulent transactions based on user behavior and transaction attributes. By learning patterns in historical data, the model aims to flag suspicious activity and reduce financial risk.

Dataset Summary: The dataset contains records of financial transactions with both legitimate and fraudulent instances. Key features include:

- **Transaction Details:** Time, amount, transaction type
- **User Behavior:** Origin and destination, balance before and after transactions
- **Label:** A binary variable (isFraud) indicating whether the transaction was fraudulent

The dataset provides a realistic basis for building a fraud detection model using supervised learning approaches.

Tools Used:

- **Python –** Core programming language for analysis and modeling
- **Jupyter Notebook –** For interactive coding and documentation
- **Pandas & NumPy –** For data loading, cleaning, and transformation
- **Seaborn, Matplotlib, Altair –** For data visualization and EDA
- **Scikit-learn –** For preprocessing, model building, and evaluation
- **Plotly Express –** For interactive visualizations
- **GridSearchCV –** For hyperparameter tuning

Steps Taken So Far:

Data Cleaning

The dataset was loaded and examined for basic quality issues such as missing values and duplicates. Preprocessing steps included label encoding of categorical features and verification of class imbalance (fraud cases are rare).

Feature Engineering

New variables were derived from transaction attributes, including differences in balances and categorization of transaction types. Label encoding was used to convert categorical data into numeric form for modeling.

Exploratory Data Analysis (EDA)

Visual and statistical analyses were conducted to explore the dataset:

- Fraudulent transactions were found to occur mostly in specific transaction types (e.g., "TRANSFER" and "CASH_OUT").
- Imbalanced class distribution was confirmed, requiring attention during model training.
- Distributions of amounts and balance differences were compared between fraudulent and non-fraudulent cases.

Modeling

A RandomForestClassifier was implemented as the primary predictive model. It was trained and evaluated using key performance metrics such as:

- **Accuracy**
- **Precision**
- **Recall**
- **F1 Score**
- **ROC AUC**

Model selection and tuning were supported by GridSearchCV.