Project Description: Fraud Detection System for Sales Data

Objective: The goal of this project is to build a machine learning-based fraud detection system that identifies fraudulent sales transactions in a dataset. The system uses supervised learning techniques to classify transactions as fraudulent (1) or non-fraudulent (0).

Key Features of the System:

Dataset:

- Contains sales-related features such as File_Type, SKU_number, SoldCount, PriceReg, ItemCount, and others.
- The target variable is SoldFlag, indicating whether a transaction is fraudulent.

Data Preprocessing:

- Missing values are handled using fillna(0).
- Categorical features like MarketingType and New_Release_Flag are encoded using LabelEncoder.
- Feature engineering introduces new variables like Price_Variance.

Modeling:

- **Classification models used:**
  - Random ForestLogistic Regression
  - Gradient Boosting
  - SMOTE (Synthetic Minority Oversampling Technique) is applied to handle class imbalance.
  - Models are evaluated using accuracy, precision, recall, F1-score, and cross-validation.

Results:

- All models achieved near-perfect classification metrics (100% accuracy), with no misclassifications.
- Feature importance analysis revealed that File_Type is the most predictive feature.

Insights:

- Dominance of the File_Type feature suggests possible data leakage, which should be addressed for deployment.