Fraud Detection using Support Vector Machines (SVM)

# Your Details:

Name: P.Bhargav

JNTU No: 21341A4538

# Introduction:

Fraud detection is a critical area in the field of financial security, especially in mobile money transactions. With the increasing use of digital financial services, the need for robust fraud detection systems has become more pressing. Support Vector Machines (SVM) are a powerful tool in machine learning, particularly for classification tasks such as fraud detection. In this project, we explore the use of SVM for detecting fraudulent transactions in a synthetic mobile money dataset.

# Tasks Involved:

The following tasks were involved in the project:  
1. Data Collection: Obtaining the dataset and understanding its structure.  
2. Data Preprocessing: Cleaning and preparing the data for model training.  
3. Model Training: Training an SVM model using the prepared dataset.  
4. Model Evaluation: Evaluating the performance of the trained model using various metrics.  
5. Result Analysis: Analyzing the results and identifying areas for improvement.

# Implementation:

The implementation of this project involves the following steps:  
1. Loading the dataset and performing exploratory data analysis (EDA) to understand the structure and identify relevant features.  
2. Preprocessing the data by handling missing values, encoding categorical variables, and normalizing numerical features.  
3. Splitting the data into training and testing sets.  
4. Training an SVM model with the training data.  
5. Evaluating the model using accuracy, precision, recall, F1-score, and other metrics.  
6. Visualizing the results using confusion matrices, ROC curves, and other relevant charts.

# Dataset Overview:

Introduction:  
This dataset presents a synthetic representation of mobile money transactions, created to mirror real-world financial activities while integrating fraudulent behaviors for research purposes. It is derived from a simulator named PaySim, which uses aggregated data from actual financial logs of a mobile money service in an African country. This dataset is specifically designed for fraud detection studies.

Dataset Details:  
The dataset is synthesized using financial logs from a mobile money service operating in an African country. It includes transaction types such as CASH-IN, CASH-OUT, DEBIT, PAYMENT, and TRANSFER over a simulated 30-day period. Fraudulent transactions have been nullified to ensure that the focus is on analyzing legitimate transactions for fraud detection.

Dataset Structure:  
1. step: Represents a unit of time, with each step equating to 1 hour.  
2. type: Transaction types including CASH-IN, CASH-OUT, DEBIT, PAYMENT, and TRANSFER.  
3. amount: The transaction amount.  
4. nameOrig: The customer initiating the transaction.  
5. oldbalanceOrg: The balance before the transaction.  
6. newbalanceOrig: The balance after the transaction.  
7. nameDest: The recipient of the transaction.  
8. oldbalanceDest: The recipient's balance before the transaction.  
9. newbalanceDest: The recipient's balance after the transaction.  
10. isFraud: Indicates if the transaction is fraudulent.  
11. isFlaggedFraud: Flags large, unauthorized transfers.

# Support Vector Machine (SVM):

Introduction to SVM:  
Support Vector Machines (SVM) are supervised learning models used for classification tasks. They work by finding the optimal hyperplane that separates different classes in the feature space. The goal of SVM is to maximize the margin between the classes, making the model more robust to new data.

Mathematics of SVM:  
SVM operates by constructing a hyperplane in a high-dimensional space that separates data points of different classes. The decision boundary is defined by support vectors, which are the data points closest to the hyperplane. The optimization problem aims to maximize the margin between these support vectors and the hyperplane.

SVM for Fraud Detection:  
In the context of fraud detection, SVM can be used to classify transactions as either fraudulent or non-fraudulent based on features like transaction type, amount, and customer balance. The SVM model is trained on historical transaction data and is evaluated on its ability to correctly identify fraudulent transactions.

# Machine Learning Overview:

Introduction to Machine Learning:  
Machine learning is a field of artificial intelligence that involves training models to make predictions or decisions based on data. It can be divided into supervised learning, unsupervised learning, and reinforcement learning. In supervised learning, the model is trained on labeled data to predict an output.

Model Selection:  
For this project, we chose SVM due to its effectiveness in binary classification tasks, especially in cases where the data is not linearly separable. SVM has been proven to work well in fraud detection scenarios, where fraudulent and non-fraudulent transactions need to be distinguished.

# Evaluation Metrics:

Precision, Recall, F1-Score:  
These metrics are crucial in evaluating fraud detection models. Precision measures the accuracy of positive predictions, recall measures the ability to identify all fraudulent transactions, and F1-score is the harmonic mean of precision and recall, providing a balance between the two.

Confusion Matrix:  
A confusion matrix is a table used to evaluate the performance of classification models. It shows the true positives, false positives, true negatives, and false negatives.

# Results:

Model Performance:  
The model was evaluated using accuracy, precision, recall, and F1-score. The results showed that the SVM model performed well in detecting fraudulent transactions, with a high precision and recall rate. The confusion matrix and ROC curve further confirmed the model's effectiveness.

Visualization:  
The confusion matrix and ROC curve are included to visually assess the model's performance. These plots provide insights into how well the model is distinguishing fraudulent transactions from legitimate ones.

# Conclusion:

In conclusion, the SVM model proved to be an effective tool for detecting fraudulent transactions in the mobile money dataset. By utilizing the features available in the dataset, the model was able to correctly classify transactions as fraudulent or non-fraudulent. Future improvements could involve tuning the model's hyperparameters or using more advanced techniques like ensemble methods to further improve performance.