**Credit Card Fraud Detection with Machine Learning**

This project tackles the critical issue of identifying fraudulent credit card transactions using machine learning. It addresses the challenges of imbalanced data and high dimensionality to build effective fraud detection models.

**1. Data Acquisition and Exploration:**

* The project starts with acquiring a historical credit card transaction dataset containing information like amount, location, time, and potential cardholder details.
* Exploratory Data Analysis (EDA) helps understand the data. This involves:
  + Identifying data types, checking for missing values, and analysing their distribution.
  + Examining the distribution of transaction amounts and timestamps for both fraudulent and legitimate transactions (histograms).
  + Visualizing potential relationships between features using correlation heatmaps.

**2. Data Preprocessing:**

* Data cleaning ensures model effectiveness. Removing duplicates and handling missing values.
* Feature scaling is necessary (e.g., scaling transaction amount) to ensure features contribute equally during model training.

**3. Dimensionality Reduction with PCA:**

* The project utilizes Principal Component Analysis (PCA) to address potentially high dimensionality in the data. PCA identifies a smaller set of features (principal components) that capture most of the information from the original features. This can improve model performance and training efficiency by reducing computational complexity.

**4. Imbalance Handling with Oversampling:**

* Credit card fraud data is typically imbalanced, with far more legitimate transactions than fraudulent ones. The project addresses this by employing an oversampling technique to balance the data for better model training. Common oversampling techniques include replicating existing minority class (fraudulent transactions) samples.

**5. Model Building and Evaluation:**

* The project focuses on building and evaluating two machine-learning models:
  + **Logistic Regression:** This model establishes a statistical relationship between transaction features and the likelihood of fraud. It calculates a probability score for each transaction, aiding in identifying suspicious activity.
  + **Decision Tree:** This model creates a tree-like structure where each branch represents a decision based on a specific feature. Transactions are classified by traversing the tree based on their features, ultimately reaching a leaf node that denotes fraudulent or legitimate.
* The project implements functions to train and evaluate these models. Training involves fitting the model to the pre-processed data (including PCA-transformed features and oversampled data), allowing it to learn the patterns that differentiate fraudulent and legitimate transactions.
* Evaluation assesses the model's performance on unseen data. Metrics like precision, recall, F1-score, and ROC-AUC curves are used to analyse how well the model identifies fraudulent transactions while minimizing false positives. Confusion matrices further detail the model's classification accuracy across different categories.

**6. Conclusion and Future Enhancements:**

* By comparing the performance of Logistic Regression and Decision Tree models, the project aims to identify the one that most effectively detects fraudulent credit card transactions. This chosen model can then be integrated into real-world systems to monitor transactions and flag suspicious activity for further investigation.
* Future enhancements could involve exploring additional machine learning models (e.g., Random Forests, deep learning) and incorporating real-time transaction data streams for continuous fraud detection.