Sub Task 4: Real-World AI Application – Fraud Detection System

Dataset Used

A sampled version of the Credit Card Fraud Detection dataset from Kaggle was used. The dataset contains 10,000 transactions, with about 5% labeled as fraudulent. It includes anonymized features V1 to V28, along with Time, Amount, and Class (target).

Preprocessing

The dataset was loaded and preprocessed by scaling the 'Time' and 'Amount' features using StandardScaler. The dataset was split into training and testing sets using stratified sampling to preserve class imbalance.

- The dataset had 4,508 normal transactions and only 492 fraud cases, i.e., about 4.9% frauds, highlighting a significant class imbalance — a common challenge in fraud detection tasks.

Anomaly Detection: Isolation Forest

An Isolation Forest model was trained on the test set to detect anomalies. Predictions were converted such that anomalies (-1) were labeled as 1 (fraud) and normal points as 0. The model was evaluated using classification metrics.

- The Isolation Forest model detected anomalies using an unsupervised approach.
- Precision (1): 0.80 shows that 80% of the detected frauds were actual frauds.
- However, Recall (1): 0.41 indicates that it only caught 41% of total frauds.
- This is typical for unsupervised anomaly detection — high precision but lower recall.
- It achieved an overall accuracy of 93%, but missed many actual frauds.

Binary Classification: Random Forest

A Random Forest Classifier was trained on the training set and evaluated on the test set. This model helped distinguish fraudulent and non-fraudulent transactions based on learned patterns.

- The Random Forest model performed exceptionally well.
- Precision (1): 1.00 means it perfectly avoided false positives for fraud detection.

- Recall (1): 0.89 shows it identified 89% of all actual frauds — a major improvement over Isolation Forest.
- It achieved an overall accuracy of 99%, with strong F1-scores for both classes.

Sample Prediction Test Case

A test sample was selected and passed to the Random Forest model. Output:
True Label: 0, Predicted: 0

Insights & Results

- The Isolation Forest worked reasonably well for detecting outliers.
- The Random Forest classifier provided robust performance with good accuracy on the test set.
- Anomaly detection is useful when labeled data is scarce, but supervised models outperform when labels are available.
- Autoencoder was skipped as it was marked optional.