Description:

- Developed a Random Forest-based fraud detection model on a simulated transaction dataset of 414,737 records, achieving >99% precision and recall for fraud detection.
- Implemented a robust feature engineering pipeline, including pre-transaction derived features (rel_amount) while removing leakage-prone features (newbalanceOrig, newbalanceDest) to ensure realistic performance.
- Utilized scikit-learn pipelines to automate preprocessing, encoding, and model training, improving reproducibility and scalability.
- Evaluated model with confusion matrix, precision, recall, and F1-score, demonstrating high accuracy on highly imbalanced data (fraud ratio ~0.1%).
- Highlighted feature importance analysis to identify key predictors such as transaction type, amount, and relative transaction size.

model 1:

With some additional features 'delta_orig','delta_dest','isFlaggedFraud'.It is showing some data Leakage problem as we are providing data after it is diagnosed fraud.

Class	**Precision**	**Recall**	**F1-Score**	**Support**
0.0 (Non-fraud)	1.00	1.00	1.00	414,322
1.0 (Fraud)	1.00	1.00	1.00	415

Metric (Overall)	Precision	Recall	F1-Score	Support
Accuracy	–	–	1.00	414,737
Macro Avg	1.00	1.00	1.00	414,737
Weighted Avg	1.00	1.00	1.00	414,737

model 2:

To solve this problem those features are removed and model is trained which is giving good results without data leakage.

Class	Precision	Recall	F1-Score	Support
0.0 (Non-fraud)	1.00	1.00	1.00	414,322
1.0 (Fraud)	0.99	1.00	0.99	415
Metric (Overall)	**Precision**	**Recall**	**F1-Score**	**Support**
Accuracy	–	–	1.00	414,737
Macro Avg	1.00	1.00	1.00	414,737
Weighted Avg	1.00	1.00	1.00	414,737

Features used and Importance: