CREDIT CARD FRAUD DETECTION

PHASE 2: INNOVATION

EXPLANATION ABOUT TOPIC

* Deep Learning Algorithms: Explore the use of advanced deep learning techniques such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs) for fraud detection. These algorithms can automatically learn complex patterns in transaction data.
* Anomaly Detection: Explore novel anomaly detection techniques, such as isolation forests, one-class SVMs, or auto encoders, to identify unusual patterns in credit card transactions.
* User Behaviour Analysis: Incorporate user behaviour analysis, such as the user’s historical transaction patterns and location-based information, to improve fraud detection accuracy.
* Predictive Analytics: Use predictive analytics to forecast future fraud trends and proactively adapt fraud detection strategies.

DATASET

* Data set source: [www.kaggle.com/data](http://www.kaggle.com/data)
* Diverse Datasets: Kaggle offers a vast and diverse collection of datasets across various domains, including finance, healthcare, natural language processing, computer vision, and more. Users can search for datasets that match their specific interests and project requirements.
* Public and Private Datasets: Kaggle provides both public and private datasets. Public datasets are openly available to the Kaggle community, while private datasets may require certain permissions or participation in competitions to access.
* Collaboration: Kaggle offers collaboration features that allow users to work on projects together. You can join discussions, share notebooks, and collaborate with other data enthusiasts or team members.
* Kernels and Notebooks: Kaggle Kernels, formerly known as Kaggle Notebooks, are a cloud-based environment for coding and analysing data. Users can create, edit, and share Jupyter notebooks directly on Kaggle, which simplifies the process of sharing analysis and code with others.

ABOUT COLUMN

* Time: This column represents the timestamp of each credit card transaction. It usually includes the date and time when the transaction occurred. This information can be valuable for identifying time-based patterns and anomalies in transaction behavior.
* Amount : The “Amount” column contains the monetary value of each transaction. It represents the transaction’s dollar amount, indicating how much money was involved in the transaction. Analysing this column can help identify unusual or high-value transactions that might be indicative of fraud.
* Class : The “Class” column is the target variable in your dataset. It is binary and indicates whether a transaction Is fraudulent or not. Typically, it uses the values 0 for legitimate transactions and 1 for fraudulent transactions. This column is what you aim to predict using machine learning algorithms.

LIBRARIES TO BE USED

* NumPy : NumPy is a fundamental library for scientific computing in Python.
* Pandas : Pandas is a data manipulation library that is often used for data pre-processing.
* Scikit-learn : Scikit-learn is a popular machine learning library in Python. It provides tools for classification, regression, clustering, and more.
* Matplotlib and Seaborn: : These libraries are useful for data visualization.

TRAIN AND TEST

* Data Collection and Pre-processing: Obtain a dataset containing historical credit card transactions. Ensure that sensitive information is anonymized or encrypted to protect privacy.
* Validation and Hyper parameter Tuning: Evaluate the model’s performance on the validation set using metrics like precision, recall, F1-score, and ROC AUC.
* Reporting and Documentation: Document the model’s performance, hyper parameters, and any changes made during the development and deployment process.
* Monitoring and Ongoing Maintenance: Continuously monitor the model’s performance in a real-world setting, and be prepared to adjust the model and retrain it as needed. Stay updated with emerging fraud patterns and adapt the model accordingly.

ABOUT METRICS

* Accuracy: This is the ratio of correctly predicted transactions (both fraudulent and legitimate) to the total number of transactions. While accuracy is important, it can be misleading in imbalanced datasets where the number of legitimate transactions far outweighs fraudulent ones.
* Precision: Precision is the ratio of correctly predicted fraudulent transactions to all predicted fraudulent transactions. It measures the model’s ability to avoid false positives. A high precision indicates that when the model flags a transaction as fraudulent, it’s likely to be correct.
* F1 Score: The F1 score is the harmonic mean of precision and recall. It provides a balance between the two and is particularly useful when dealing with imbalanced datasets.