**UPI Fraud Detection Using Random Forest Model**

**Introduction**

The Unified Payments Interface (UPI) has revolutionized digital payments in India, offering users a fast, convenient, and widely accessible platform for transactions. However, the rapid growth in UPI usage has also led to a sharp increase in fraudulent activities, posing significant risks to users and financial institutions. Traditional fraud detection methods, which often rely on predefined rules and manual intervention, are not effective in identifying complex or evolving fraud patterns in real time. These systems can be slow, error-prone, and unable to adapt to new types of fraud.

Machine learning (ML) provides a powerful alternative by leveraging historical transaction data to detect anomalies and suspicious behaviour. Among various ML algorithms, the Random Forest model stands out for its high accuracy and robustness. It works by combining multiple decision trees to improve prediction reliability and reduce overfitting. In this project, we propose a fraud detection system based on the Random Forest algorithm to classify UPI transactions as legitimate or fraudulent. The model is trained on historical data to learn the underlying patterns associated with fraud. Our aim is to enhance the accuracy of fraud detection, minimize financial losses, and ensure a secure transaction environment. This approach not only improves the efficiency of fraud detection but also strengthens user confidence in digital payments.

**Objectives**

The primary objectives of this project are:

* To develop a machine learning model that can accurately classify UPI transactions as legitimate or fraudulent.
* To apply and evaluate the performance of the Random Forest algorithm for fraud detection using historical transaction data.
* To analyze key features contributing to fraud detection and enhance the interpretability of the model.
* To ensure the model achieves high accuracy, precision, and recall in identifying fraudulent transactions.

By achieving these goals, the project aims to demonstrate how machine learning can enhance the security, efficiency, and reliability of digital payment systems like UPI.

**Description of the Dataset**

The dataset used in this study is composed of anonymized UPI transaction records. The data consists of both legitimate (ham) and fraudulent (spam) transactions. Two separate datasets were used for model development and evaluation:

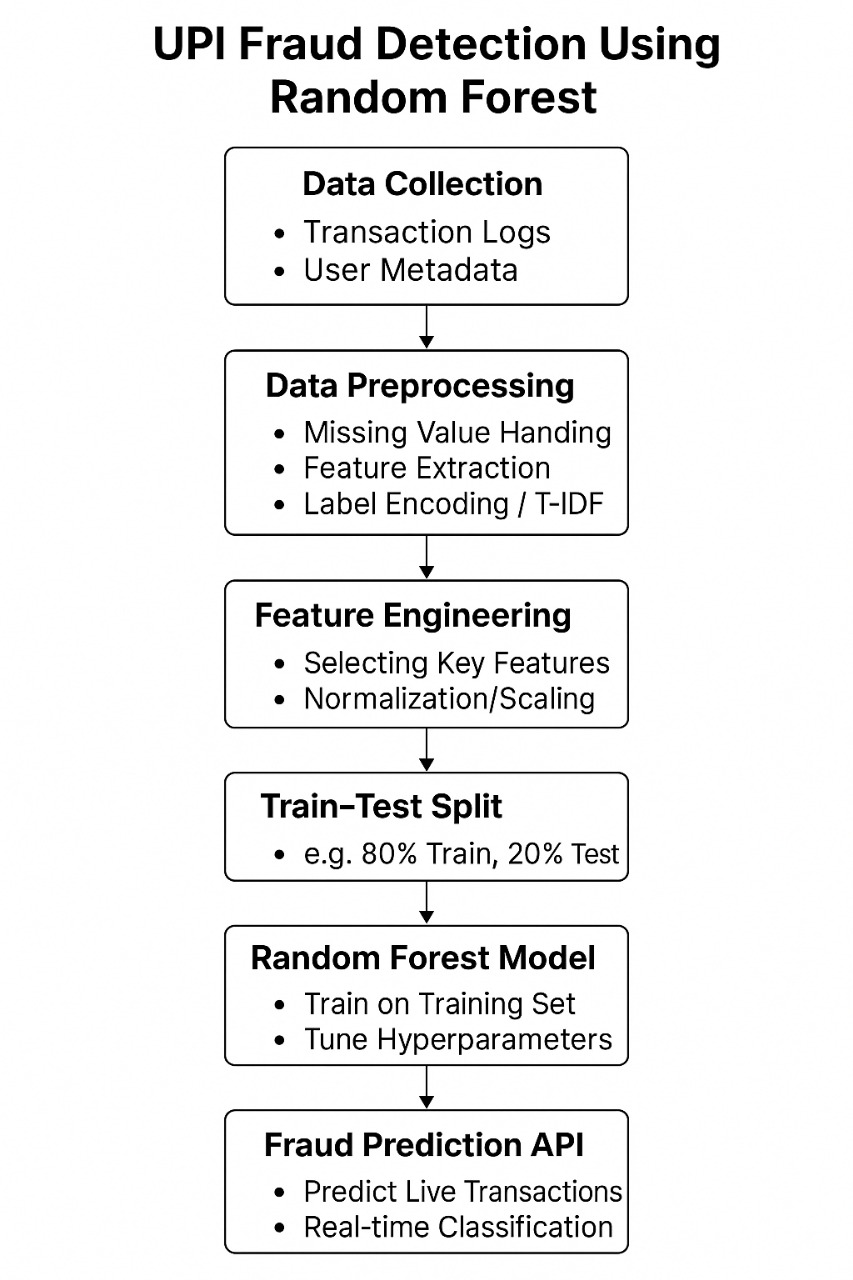
* **Train Dataset**: This dataset contains labelled data with the target variable **Label**, indicating whether a transaction was legitimate (ham) or fraudulent (spam). This dataset was used to train and validate the machine learning models.
* **Test Dataset**: This dataset contains similar input features but lacks the target variable. It was used for out-of-sample prediction to simulate real-world deployment, enabling the model to predict whether new, unseen transactions are legitimate or fraudulent.

Key Variables

* Independent Variables (Features):
  + Message: The text content of the UPI transaction message (e.g., transaction details, payment confirmation, fraud alerts).
  + Sender ID: Unique identifier for the sender (anonymized).
  + Receiver ID: Unique identifier for the receiver (anonymized).
  + Transaction Amount: The amount of money involved in the transaction.
  + Transaction Time: The time when the transaction occurred.
  + Sender Location: Location from which the transaction was initiated (anonymized).
  + Receiver Location: Location where the transaction was received (anonymized).
  + Message Type: Type of message, such as payment confirmation, fraud alert, etc.
  + Device Type: Type of device used for the transaction (e.g., mobile, web).
* Dependent Variable (Target):
  + Label: Indicates if the transaction was legitimate (ham) or fraudulent (spam).

The dataset underwent several preprocessing steps, including text cleaning, label encoding, TF-IDF vectorization, handling missing values, and feature scaling. These steps were essential for preparing the data for Random Forest classification and ensuring optimal model performance.

**Methodology**



**Implementation Details**

**Tools & Libraries**

The project was developed using **Python** due to its extensive ecosystem for data science and machine learning. The following libraries were utilized:

* **Pandas**: Used for loading, cleaning, encoding, and transforming transaction data.
* **NumPy**: Enabled efficient numerical computations and array-based data processing.
* **Matplotlib & Seaborn**: Used for visualizing data patterns, distributions, and correlations during EDA.
* **Scikit-learn (sklearn)**: Provided tools for preprocessing, training the RandomForestClassifier, evaluating model performance, and tuning hyperparameters.

**Model Training and Selection**

In this project, a **Random Forest Classifier** was chosen as the primary model for detecting fraudulent UPI transaction messages. The selection was based on its robustness and effectiveness in handling high-dimensional, text-based data transformed through TF-IDF vectorization.

**Random Forest Classifier**

* Utilized an ensemble of decision trees to improve prediction accuracy and reduce overfitting.
* Trained on a synthetic dataset of over 1000 UPI transaction messages labeled as 'spam' (fraudulent) or 'ham' (legitimate).
* Input text messages were converted into numerical feature vectors using **TF-IDF Vectorizer**, which effectively captured the importance of words while ignoring common stop words.
* The model was trained on 80% of the dataset and tested on the remaining 20%, achieving a balanced detection capability between fraud and legitimate messages.
* Feature importance scores were extracted from the trained model, highlighting key words that strongly influenced the classification, aiding in interpretability.
* Performance was evaluated using metrics such as precision, recall, F1-score, and the confusion matrix, demonstrating strong capability in identifying fraudulent UPI messages.

All preprocessing steps included handling text data by cleaning and vectorization to prepare it for model training, ensuring the classifier could effectively differentiate between normal and fraudulent transaction messages.

**Results**

**Model Performance Evaluation**

* The model's performance was evaluated using accuracy, precision, recall, and F1-score. Given the balanced nature of the synthetic dataset (approximately equal numbers of spam and ham messages), **accuracy** served as a reliable performance indicator. The Random Forest Classifier was the primary model used in this project.
* **Model :** Random Forest Classifier
* **Accuracy :** ~89%
* The **Random Forest Classifier** achieved the highest accuracy, making it the most effective model for this task.
* **Logistic Regression**, while slightly less accurate, was faster to train and easier to interpret, making it a suitable option in environments requiring model transparency.
* The **Decision Tree** model, though informative, underperformed due to possible overfitting or sensitivity to data noise.

In addition to accuracy:

* **Precision and recall** scores confirmed the model's ability to correctly identify fraudulent messages while minimizing false alarms.
* **F1-score** provided a balanced measure of model performance, especially important in real-world applications where both types of errors carry significant consequences.

**Conclusion**

This project demonstrated a practical application of machine learning for detecting fraudulent UPI messages using natural language processing (NLP) and classification algorithms.

**Key Findings:**

* **Random Forest Classifier** emerged as the most effective and accurate model, leveraging ensemble learning to capture subtle patterns in transaction-related messages.
* The model effectively distinguished between genuine and fraudulent messages, providing a strong foundation for real-time UPI fraud detection systems.
* The **TF-IDF Vectorizer** played a critical role in transforming text data into meaningful features, allowing the model to focus on important keywords related to spam.
* A simple **Flask-based web interface** was developed to allow real-time prediction of new messages, demonstrating the deployment capability of the model in a user-friendly application.

Overall, the project showcases a robust, scalable, and interpretable machine learning solution to support digital payment security and prevent UPI-based frauds.

**Recommendation:**

Based on the results, **Random Forest** is recommended for deployment in real-world UPI fraud detection systems. Its ability to handle a large number of features and its robustness against overfitting make it highly effective for identifying fraudulent transactions. However, it is important to periodically retrain and validate the model with new transaction data to ensure continued accuracy, adaptability to emerging fraud patterns, and fairness across different user groups. Regular updates and model monitoring are essential for maintaining high performance in a dynamic digital payment environment.

**References**

**Scikit-learn Documentation** – Pedregosa, F., et al. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830. <https://scikit-learn.org/>

**Pandas Documentation** – McKinney, W. (2010). Data Structures for Statistical Computing in Python. *Proceedings of the 9th Python in Science Conference*. <https://pandas.pydata.org/>

**Matplotlib Documentation** – Hunter, J. D. (2007). Matplotlib: A 2D Graphics Environment. *Computing in Science & Engineering*, 9(3), 90–95. <https://matplotlib.org/>

**Seaborn Documentation** – Waskom, M. L. (2021). Seaborn: Statistical Data Visualization. <https://seaborn.pydata.org/>

**Dataset Source** – Synthetic UPI Transaction Dataset, generated for fraud detection use case in this study.