**Data Collection and Preprocessing Phase**

| Date | 25 February 2025 |
| --- | --- |
| Team ID | NIL |
| Project Title | Online Payments Fraud Detection Using ML |
| Maximum Marks | 2 Marks |

.

**Data Collection Plan Template**

| **Section** | **Description** |
| --- | --- |
| Project Overview | Develop a machine learning model to detect fraudulent online payment transactions with >99% accuracy. The project will analyze transaction patterns (amount, balances, type) to flag high-risk activities. |
| Data Collection Plan | Data will be collected from **publicly available synthetic financial datasets** that mimic real-world fraud patterns. Primary focus on labeled datasets (fraud/non-fraud) with transaction metadata. |
| Raw Data Sources Identified | <ul><li>**Kaggle**: Synthetic dataset from ["Online Payments Fraud Detection"](https://www.kaggle.com/datasets/volodymyrgavrysh/fraud-detection-bank-dataset-20k-records-binary) (6.3M records, CSV format).</li><li>**IEEE DataPort**: Real-world fraud patterns in ["Financial Fraud Transactions"](https://ieee-dataport.org/)(restricted access).</li><li>**Simulated Data**: Generate additional test cases using scikit-learn’s make\_classification for edge scenarios.</li></ul> |

**Raw Data Sources Template**

| **Source Name** | **Description** | **Location/URL** | **Format** | **Size** | **Access Permissions** |
| --- | --- | --- | --- | --- | --- |
| Kaggle Fraud Dataset | Synthetic financial transactions with labeled fraud cases (isFraud column). Includes transaction types, amounts, and balance changes. | [Kaggle Dataset Link](https://www.kaggle.com/datasets/volodymyrgavrysh/fraud-detection-bank-dataset-20k-records-binary) | CSV | 650 MB | Public |
| Simulated Test Data | Artificially generated fraud scenarios using skewed amount distributions and new transaction patterns. | Generated locally using Python's sklearn.datasets.make\_classification | CSV | 100 MB | Private (Team only) |