**Data Collection and Preprocessing Phase**

|  |  |
| --- | --- |
| Date | 15 March 2024 |
| Team ID | SWUID20250184320 |
| Project Title | Online Payment Fraud Detection |
| Maximum Marks | 2 Marks |

**Data Collection Plan & Raw Data Sources Identification**

Elevating the data strategy with the Data Collection plan and the Raw Data Sources report, ensuring meticulous data curation and integrity for informed decision-making in every analysis and decision-making endeavor.

**Data Collection Plan Template**

|  |  |
| --- | --- |
| **Section** | **Description** |
| Project Overview | This machine learning project aims to build an optimized fraud detection system that identifies suspicious transactions with high accuracy and low false positives. Using historical transaction data, the model will learn to classify patterns of legitimate and fraudulent behavior. The objective is to automate fraud detection, enhance financial security, and support real-time decision-making using efficient and scalable ML algorithms like Random Forest and XG Boost. |
| Data Collection Plan | The data for this project is collected from publicly available and reliable sources such as Kaggle. The dataset chosen is specifically curated for fraud detection tasks and contains anonymized transaction records. These sources ensure diverse, labeled, and real-world data suitable for training and evaluating machine learning models for fraud detection. |
| Raw Data Sources Identified | Kaggle: Online Payment Fraud Detection Dataset |

**Raw Data Sources Template**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Source Name** | **Description** | **Location/URL** | **Format** | **Size** | **Access Permissions** |
| Dataset 1 | Description of the data in this source. | https://www.kaggle.com/datasets/rupakroy/online-payments-fraud-detection-dataset | CSV | 493.53 MB | Public |