**Data Collection and Preprocessing Phase**

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

**Data Quality Report Template**

The Data Quality Report Template will summarize data quality issues from the selected source, including severity levels and resolution plans. It will aid in systematically identifying and rectifying data discrepancies.

|  |  |  |  |
| --- | --- | --- | --- |
| **Data Source** | **Data Quality Issue** | **Severity** | **Resolution Plan** |
| Online Payments Fraud Detection Dataset | Presence of class imbalance (very few fraudulent transactions compared to legitimate ones). | High | Apply resampling techniques like SMOTE (Synthetic Minority Over-sampling Technique) or under sampling to balance the dataset. |

|  |  |  |  |
| --- | --- | --- | --- |
| Online Payments Fraud Detection Dataset | Missing or null values in some columns. | Moderate | Use pandas .isnull() .sum() to identify missing values. Apply imputation methods like mean/median for numerical features or drop rows. |

|  |  |  |  |
| --- | --- | --- | --- |
| Online Payments Fraud Detection Dataset | Unnecessary columns like transaction ID or redundant features that don’t contribute to prediction. | Low | Drop irrelevant columns using df.drop() after manual inspection or feature importance analysis. |
| Online Payments Fraud Detection Dataset | Skewed distributions in numerical features. | Moderate | Apply scaling ortransformationtechniques such as Min-Max Scaling or Log Transformation to normalize distributions. |