**Data Collection and Preprocessing Phase**

|  |  |
| --- | --- |
| Date | 17 June 2024 |
| Team ID | SWTID1749712812 |
| Project Title | Unlocking Silent Signals: Decoding Body Language with Mediapipe |
| Maximum Marks | 2 Marks |

**Data Quality Report:**

This Data Quality Report identifies and outlines key issues in the collected MediaPipe landmark data (coords.csv) and presents strategies for improving data integrity. The dataset contains pose, hand, and face landmarks from video frames, labelled by emotion (e.g., "Happy").

**Data Quality Report:**

|  |  |  |  |
| --- | --- | --- | --- |
| **Data Source** | **Data Quality Issue** | **Severity** | **Resolution Plan** |
| Custom dataset generated using MediaPipe and exported as coords.csv | Missing values in several x, y, z, and v columns for landmarks (due to occasional pose detection failures in video frames) | Moderate | |  |  | | --- | --- | | Apply forward fill (ffill) or drop invalid frames. |  | |
| Custom dataset (coords.csv) | Over 2000 columns with dense numerical features may include redundant/unimportant data | Low | Use PCA or feature selection to reduce dimensionality |
| Custom dataset (coords.csv) | High correlation between some landmarks (e.g., left and right arms in neutral poses) | Low | Correlation analysis to identify and optionally drop highly correlated features |
| Custom dataset (coords.csv) | Mixed scales (raw coordinates and visibility values) | Moderate | |  | | --- | |  |  |  | | --- | | Normalize numeric values using StandardScaler or MinMaxScaler | |
| Custom dataset (coords.csv) | Target column (class) is categorical | Moderate | Apply Label Encoding or One-Hot Encoding for modelling |