**Data Collection and Preprocessing Phase**

|  |  |
| --- | --- |
| Date | 16 June 2025 |
| Team ID | SWTID1749709635 |
| Project Title | Mental Health Prediction |
| 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** |
| **Kaggle dataset** | Missing values in 'state', 'self\_employed', 'work\_interfere' columns | Moderate | Fill missing values with default categories ('Unknown', 'No', 'Never') as done in preprocessing. |
| **Kaggle dataset** | Outlier/invalid ages (e.g., Age < 0 or Age > 100) | Moderate | Filter out rows with invalid ages. |
| **Kaggle dataset** | Inconsistent/categorical values in 'Gender' column (e.g., typos, various forms, non-binary, etc.) | Moderate | Standardize using mapping dictionary; remove ambiguous/other entries. |
| **Kaggle dataset** | Categorical variables in many columns (e.g., 'Country', 'self\_employed', etc.) | Moderate | Apply label encoding to all categorical variables. |
| **Kaggle dataset** | Irrelevant columns ('Timestamp', 'comments') | Low | Drop these columns during preprocessing. |
| **Kaggle dataset** | Target variable ('treatment') is categorical | Low | Encode target variable using LabelEncoder. |
| **Kaggle dataset** | Numerical feature ('Age') not standardized | Low | Standardize 'Age' using StandardScaler. |

(Kaggle dataset – survey.csv)