# Data Collection and Preprocessing Phase

| | |
|---|---|
| Date | 9 June 2024 |
| Team ID | SWTID1720112707 |
| Project Title | Anemia Sense: Leveraging Machine Learning For Precise Anemia Recognitions |
| 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 |
|---|---|---|---|
| Dataset | Missing values in numerical columns (e.g., Hemoglobin) | Moderate | Fill missing numerical values with the median of the respective columns. |
| Dataset | Missing values in categorical columns (e.g., Gender) | Low | Fill missing categorical values with the mode of the respective columns. |
| Dataset | Imbalanced class distribution | High | Apply undersampling to balance the classes, particularly for the 'Result' column |

| | | | |
|---|---|---|---|
| Dataset | Presence of outliers in numerical columns | Moderate | Use statistical methods to identify and treat outliers or apply robust scaling. |
| Dataset | Mixed data types in some columns | Low | Convert all columns to appropriate data types and handle errors accordingly. |
| Dataset | Redundant or highly correlated features | Low | Perform correlation analysis and drop or combine redundant features. |
| Dataset | Infinite values in some columns | Moderate | Replace infinite values with NaN and then handle them with imputation or removal |
| Dataset | Variability in units of measurement | Moderate | Standardize units of measurement across the dataset using scaling techniques. |