**Data Collection and Preprocessing Phase**

|  |  |
| --- | --- |
| Date | 9 JULY 2024 |
| Team ID | 739661 |
| Project Title | Anemiasense: Leveraging Machine Learning For Precise Anemia Recognitions |
| Maximum Marks | 2 Marks |

**Data Quality Report Template**

This report provides an overview of the data quality assessment conducted for the Anemiasense project, which aims to leverage machine learning for precise anemia recognition. The report covers various aspects of data quality to ensure the reliability and suitability of the data for machine learning modeling.

|  |  |  |  |
| --- | --- | --- | --- |
| **Data Source** | **Data Quality Issue** | **Severity** | **Resolution Plan** |
| Dataset | Missing Values | High | Impute missing values using mean or median for numerical features. For categorical features, use mode. |
| Dataset | Outliers | Moderate | Use statistical methods like Z-score or IQR to detect and handle outliers, either by removing or transforming them. |
| Dataset | Duplicate Records | Low | Remove duplicate records from the dataset to ensure each instance is unique |
| Dataset | Inconsistent Format | Moderate | Standardize date formats and numerical representations across columns using data preprocessing techniques. |
| Dataset | Irrelevant Features | Low | Perform feature selection techniques (e.g., SelectKBest) to identify and retain only relevant features for anemia recognition. |
| Dataset | Data Currency | High | Ensure the dataset is up-to-date by verifying the date range and considering additional data collection if necessary. |