**Data Collection and Preprocessing Phase**

|  |  |
| --- | --- |
| Date | 25 June 2025 |
| Team ID | SWTID1749753590 |
| Project Title | Early Prediction for Chronic Kidney Disease Detection: A Progressive Approach to Health Management |
| 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** |
| CKD Dataset (SmartInternz-provided CSV) | The dataset contains missing values in key clinical features such as rbc, pcv, wc, and sod, which are crucial for disease prediction. Some columns like pcv and wc are stored as object type instead of numeric. Additionally, inconsistent categorical labels (e.g., yes, Yes) were found | High – Missing or misformatted data can mislead the model and reduce prediction accuracy. | The CKD dataset has missing values in several columns like rbc, pcv, and sod (high severity), which were handled using median/mode imputation. Some features had incorrect data types (moderate severity), resolved by converting objects to numerics. Minor inconsistencies in categorical values were fixed by standardizing text. The id column was dropped as it holds no predictive value |