**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 | 6 Marks |

**Data Exploration and Preprocessing Template**

Identifies data sources, assesses quality issues like missing values and duplicates, and implements resolution plans to ensure accurate and reliable analysis.

|  |  |
| --- | --- |
| **Section** | **Description** |
| Data Overview | The dataset has 400 entries and 26 features, with both numerical and categorical data. Some columns like rbc and sod have missing values. The target is classification, used for CKD prediction |
| Univariate Analysis | Exploration of individual variables (mean, median, mode, etc.).  All Numeric columns |
| Bivariate Analysis | Relationships between two variables (correlation, scatter plots).  Age-gender,hemoglobin-place,age-place,location-place,hemoglobin-age,gender-diabetes,hemoglobin-place,hdl-diabetes, type of alcohol consumed\_country liquor-duration of alcohol consumption(years) |
| Multivariate Analysis | Patterns and relationships involving multiple variables.  All numeric variables |
| Outliers and Anomalies | Identification and treatment of outliers. |
| **Data Preprocessing Code Screenshots** | |
| Loading Data | Code to load the dataset into the preferred environment (e.g., Python, R). |
| Handling Missing Data | Code for identifying and handling missing values. |
| Data Transformation | Code for transforming variables (scaling, normalization). |
| Feature Engineering | Code for creating new features or modifying existing ones. |
| Save Processed Data | Code to save the cleaned and processed data for future use. |