**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 Collection Plan & Raw Data Sources Identification Template**

Elevate your data strategy with the Data Collection plan and the Raw Data Sources report, ensuring meticulous data curation and integrity for informed decision-making in every analysis and decision-making endeavor.

**Data Collection Plan Template**

|  |  |
| --- | --- |
| **Section** | **Description** |
| Project Overview | This project aims to build a machine learning model to predict the presence of Chronic Kidney Disease (CKD) based on clinical and laboratory data. The goal is to support early detection and timely intervention, thereby improving patient outcomes. |
| Data Collection Plan | |  | | --- | |  |   The dataset used in this project was provided by project mentors as part of the SmartInternz mini-project framework. It includes anonymized medical records of patients with various clinical attributes relevant to CKD diagnosis. No external data collection was required. |
| Raw Data Sources Identified | |  | | --- | |  |   The primary dataset was shared in .csv format, containing 400 patient records and 26 features. These include numerical values (e.g., hemoglobin, blood urea) and categorical attributes (e.g., hypertension, diabetes). The dataset is clean but contains missing values and mixed data types that required preprocessing. |

**Raw Data Sources Template**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Source Name** | **Description** | **Location/URL** | **Format** | **Size** | **Access Permissions** |
| Dataset 1 | Description of the data in this source. | <https://drive.google.com/file/d/1mPl4yaTKuKZ3017YfYC19Ni7Y964eCNI/view?usp=sharing> | Excel | 48 kB | Public |