Data Collection and Preprocessing Phase

Date	18 June 2025
Team ID	xxxxxx
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 machine learning project aims to assess the probability of Chronic Kidney Disease (CKD) onset in a patient based on clinical and laboratory data. Chronic Kidney Disease (CKD) is a potentially lethal health issue that necessitates early diagnosis to prevent severe complications. The technology assists healthcare practitioners in identifying at-risk populations by training the classification of patient health data. **Objectives:**

	● Develop a robust classification model (Random Forest, XGBoost, etc.) to predict CKD presence. ● Handle missing data and class imbalance effectively. ● Identify the most important clinical features influencing CKD. ● Provide explainable outputs to aid clinical decisions.
Data Collection Plan	UCI CKD Dataset, Kaggle – CKD Prediction
Raw Data Sources Identified	Raw Data Sources 1. UCI Chronic Kidney Disease Dataset ○ Link: UCI CKD Dataset ○ Description: Contains 400 patient records with 24 clinical features and a binary label (CKD or not). Includes values like blood pressure, albumin, hemoglobin, etc. 2. Kaggle Notebook Reference ○ Link: Kaggle – CKD Prediction ○ Description: A practical implementation of the UCI dataset with preprocessing, model training, and performance evaluation.

Raw Data Sources Template

Source Name	Description	Location/URL	Format	Size	Access Permissions
UCI CKD Dataset	400 patient records with 24 clinical features and a binary CKD label.	https://archive.ics.uci.edu/ml/datasets/chronic_kidney_disease	CSV	~15 GB	Public, free to use
Kaggle – CKD Prediction Dataset	Description of the same UCI dataset rehosted on Kaggle for easy access and collaboration.	https://www.kaggle.com/datasets/mansoordaku/ckdisease	Excel	~15 GB	Public (login required)
SmartWallet guided projects - CKD Prediction Dataset	Description of the same UCI dataset rehosted on Kaggle for	https://drive.google.com/file/d/1mPl4yaTKuKZ3017YfYC19Ni7	Excel	~15 GB	Private (with access)

	easy access and collaboration.	Y964eCNI/view			