Data Collection and Preprocessing Phase

Date	08 July 2024
Team ID	SWTID1720190579
Project Title	Early Prediction of Chronic Kidney Disease Using Machine Learning
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	Chronic Kidney Disease (CKD) is one of the growing health threats globally because diagnosis is almost always made late and the disease progresses slowly, resulting in increased health complications and treatment costs. The early prediction of CKD is the focus of this project, and the project proposal entails the use of machine learning in accomplishing this task. Specifically, through the extraction of the patient information such as demographics, medical history and laboratory data, the model aims to find patterns or characteristics as regards to early stage of CKD. Early prediction is actually beneficial in allowing doctors to come up with early

	medical interventions, enhance the overall quality of health of patients, and even reduce the overall cost of health delivery.
Data Collection Plan	Mention from which sources the data are going to be collected.
Raw Data Sources Identified	List the raw data sources with relevant details (as a short description).

Raw Data Sources Template

Source Name	Description	Location/URL	Format	Size	Access Permissions
Kaggle Dataset	This dataset comprises electronic medical records of **491 patients (241 women, 250 men**) collected at Tawam Hospital, Al-Ain, UAE, in 2008. Patients, averaging **53.2 years,** have charts of 22 clinical variables detailing lab tests, exams, and	https://www.kaggle.com/datasets/davidechicco/chronic-kidney-disease-ehrs-abu-dhabi	CSV	9 KB	Public

	medical history. All patients either had or were at risk of **cardiovascular disease,** per hospital standards.				
Google Drive	The **Chronic Kidney Disease (CKD) dataset** contains 29 columns capturing various medical attributes for clinical analysis. These attributes have data like age and bp (blood pressure) etc….. The goal is to analyze and predict CKD based on these medical features.	https://drive.google.com/file/d/1mPl4yaTKuKZ3017YfYC19Ni7Y964eCNI/view	Excel (XLS worksheet)	48 KB	Public (with in SmartInternz group)