Data Collection and Preprocessing Phase

Date	5 July 2024
Team ID	SWTID1720082525
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	A health insurance company's data is underutilized for chronic kidney disease (CKD) detection. Existing models miss early signs, hindering preventive care and driving up costs. This project aims to develop a new approach using machine learning. By analyzing vast amounts of policyholder data, the project will identify subtle patterns that predict early-stage CKD risk. This will lead to better detection, allowing for earlier intervention and improved health outcomes for policyholders, along with reduced healthcare costs for the company.
Data Collection Plan	• We'll gather detailed information on existing diagnoses (diabetes, hypertension etc.), medications prescribed, and hospitalization history to identify potential risk factors. • Blood tests (creatinine levels) and urine tests (protein-to-creatinine ratio) will provide crucial biomarkers for early detection of CKD.

Raw Data Sources Identified	Data sources include Skillwallet's Chronic Kidney Disease Database. The sample data contains demographics, medical history, lab results, and lifestyle factors for machine learning analysis in predicting chronic kidney disease.

Raw Data Sources Template

Source Name	Description	Location/URL	Format	Size	Access Permissions
Skill Wallet	The dataset comprises patient blood related details (blood glucose, albumin, wbc, haemoglobin), urine details (urea, albumin), and age.	https://drive.google.com/file/d/1mPl4yaTKuKZ3017YfYC19Ni7Y964eCNI/view?usp=sharing	CSV	48 KB	Public