Data Collection and Preprocessing Phase

Date	18 June 2025
Team ID	SWTID1749641473
Project Title	Early Prediction for Chronic Kidney Disease Detection: A Progressive Approach to Health Management
Maximum Marks	2 Marks

Data Quality Report Template

The Data Quality Report Template will summarize data quality issues from the selected source, including severity levels and resolution plans. It will aid in systematically identifying and rectifying data discrepancies.

Data Source	Data Quality Issue	Severity	Resolution Plan
Chronic Kidney Disease Dataset (400 records, 26 attributes)	Missing values in multiple columns (e.g., age, bp, rbc, sod)	High	Use median imputation for numeric fields, mode for categorical fields
	Incorrect data types in columns like pcv, wc, rc	Moderate	Convert using pd.to_numeric(..., errors='coerce')

	Inconsistent categorical values (e.g., " yes" vs "yes" in dm, htn, cad)	Moderate	Apply .str.strip().str.lower() to standardize categories
	Presence of outliers in columns like sc, bgr, bu	Moderate	Detect using IQR method and treat by capping or removing
	Irrelevant or non-informative column id	Low	Drop column using df.drop('id', axis=1)
	Possible class imbalance in target variable classification	Moderate	Use SMOTE for oversampling or set class weights in models
	Categorical variables not encoded for ML	Moderate	Apply Label Encoding or One-Hot Encoding based on model needs