Data Collection and Preprocessing Phase

Date	11 June 2024
Team ID	SWTID1749620488
Project Title	Chronic Kidney Disease Detection Using Machine Learning
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
Dataset	Missing values in both categorical and numerical features (e.g., age, hemoglobin, pus_cell)	High	Used mean() for numerical columns and mode() for categorical columns to impute missing values.
Dataset	Inconsistent labels in categorical columns (e.g., \tno, \tyes in diabetesmellitus and coronary_artery_disease)	Moderate	Replaced inconsistent tab characters using .replace() to standardize all values.
Dataset	Columns with object data type actually contain numeric values	High	Converted such columns using pd.to_numeric(...,

	(e.g., packed_cell_volume, white_blood_cell_count)		errors='coerce') to enable numerical processing.
Dataset	Ambiguous target label (ckd\t instead of ckd) in the class column	Low	Replaced the value using .replace('ckd\\t', 'ckd') to clean the label.
Dataset	Some features had low variance or high correlation (e.g., albumin, sugar, specific_gravity)	Low	Explored via heatmap and included in correlation analysis to inform model feature selection.