Data Collection and Preprocessing Phase

Date	06 July 2024
Team ID	SWTID1720082372
Project Title	Early Prediction of Chronic Kidney Disease 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
Chronic Kidney Disease Dataset	Missing values in various columns	High	Implement targeted imputation strategies: mean for numeric columns like 'blood_pressure', median for 'age', mode for categorical columns like 'diabetes'. Use KNN imputer for complex relationships.
Chronic Kidney Disease Dataset	Incorrect data types (e.g., numeric stored as strings)	Moderate	Convert columns to appropriate data types using pd.to_numeric() for numeric columns and

			.astype('category') for categorical columns.
Chronic Kidney Disease Dataset	Presence of outliers in numeric columns	Moderate	Identify outliers using IQR method. Decide whether to cap outliers or remove them based on domain knowledge.
Chronic Kidney Disease Dataset	Potential irrelevant columns (e.g., patient ID)	Low	Drop irrelevant columns that don't contribute to the analysis or prediction task.
Chronic Kidney Disease Dataset	Inconsistent categorical values (e.g., 'yes'/'no' vs 'y'/'n')	Moderate	Standardize categorical values using string replacement. Map all variations to consistent values (e.g., 1 for yes/present, 0 for no/not present).