Data Collection and Preprocessing Phase

Date	14 June 2025
Team ID	SWTID1749713922
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
kidney_disease.csv	Missing values in multiple columns	High	Missing numerical values were addressed using appropriate statistical imputation techniques, while missing categorical values were filled using the most frequent category in each column.
kidney_disease.csv	Unclear or poorly defined feature names	Medium	Renamed columns for clarity and consistency.
kidney_disease.csv	Duplicate rows	Low	Removed using df.drop_duplicates()

kidney_disease.csv	Inconsistent categorical entries like /yes, /no, ckd, ckd\t, notckd	Medium	Cleaned using .str.lower().str.strip() and replaced variations with standard labels
kidney_disease.csv	Numerical columns stored as strings or containing non-numeric characters	Medium	Used pd.to_numeric(errors='coerce') to convert, followed by handling missing values.
kidney_disease.csv	Columns with inconsistent formatting	Low	Applied .str.strip() and .str.lower() across all categorical columns