

Data Collection and Preprocessing Phase

Date	23 June 2025
Name	Aniket Jingonda Patil
Project Title	Health Classification System
Maximum Marks	2 Marks

Data Quality Report Template

Data Source	Data Quality Issue	Severity	Resolution Plan
Dataset (fetal_health.csv)	Class imbalance (uneven distribution of Normal, Suspect, Pathological classes)	High	Apply stratified sampling during train-test split and use class weights in XGBoost model. Consider SMOTE for oversampling minority classes if needed.
	Potential outliers in CTG features (e.g., baseline value, histogram, variance)	Moderate	Identify outliers using IQR method and either cap them or remove extreme values after validation.
	Missing values (if any, though initial check shows none)	Low	If discovered, impute with mean for numeric features or remove rows with critical missing data using pandas dropna().