Data Collection and Preprocessing Phase

Date	15 June 2025
Team ID	SWTID1749621188
Project Title	Anemia sense: leveraging machine learning for precise anemia
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	Gender column is numerical (0/1), not labeled	Low	Map values: 0 → Female, 1 → Male for better interpretability.
Dataset	No missing values but numerical-only data can limit explainability	Moderate	Consider feature engineering or combining with patient metadata (e.g., age, region).
Dataset	Possible outliers in Hemoglobin and MCV values	Moderate	Use IQR method or Z-score to detect and possibly remove outliers.
Dataset	No clear feature naming (e.g., MCH, MCHC, MCV not explained)	Low	Add feature descriptions in documentation or a separate metadata file.
Dataset	Class imbalance may exist in Result column (needs confirmation)	Moderate	Use value_counts to check. If imbalanced, apply SMOTE or class weighting during training.