Data Quality Report

Date	**03 August, 2025**
Project Title	**Anemia Sense - Machine Learning for Precise Anemia Recognition**
Maximum Marks	**2 Marks**

Data Quality Report:

The Data Quality Report 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: Kaggle Dataset (e.g., anemia.csv)

Data Quality Issues Summary

Data Source	**Data Quality Issue**	**Severity Level**	**Resolution Plan**
Kaggle Dataset	Missing values in key features such as 'Hemoglobin', 'RBC', 'MCV', 'MCH'	Moderate	Impute missing values using mean or median; consult clinical domain knowledge if needed.
Kaggle Dataset	Imbalanced classes in target variable ('Result'): minority anaemic cases	High	Use resampling techniques like SMOTE or under sampling to balance the dataset.
Kaggle Dataset	Outliers present in numerical columns such as abnormal Hemoglobin readings	Moderate	Apply outlier detection and treatment methods; consider domain thresholds.
Kaggle Dataset	Categorical features like 'Gender', 'Stage' requires proper encoding	Moderate	Encode categorical variables with OneHot or Label Encoding prior to model training.
Kaggle Dataset	Inconsistent data types or formatting across columns	Low	Standardize data types and formats before processing.
Kaggle Dataset	Potential multicollinearity among features	Low	Use correlation analysis to identify and remove or combine redundant features.