Data Collection and Preprocessing Phase

Date: 30 July 2025
Project Title: AnemiaSense — Machine Learning Based Anemia Detection
Maximum Marks: 4 Marks

Data Source

It contains 5 attributes: **gender**, **hemoglobin value**, **MCV** (Mean Corpuscular Volume), **MCH** (Mean Corpuscular Hemoglobin), and **MCHC** (Mean Corpuscular Hemoglobin Concentration).
These features are essential for detecting anemia through machine learning models.

Data Quality Report

The dataset for anemia detection was examined for missing values, outliers, and inconsistencies. Data cleaning and preprocessing steps were planned to ensure the accuracy and reliability of the machine learning model.

Data Quality Report Table

Attribute	Issue	Severity	Resolution Plan
gender	Categorical; requires numerical encoding	Medium	Convert gender to numerical values (Male=1, Female=0)
hemoglobin value	Possible missing values or extreme outliers	High	Fill missing values with median; remove unrealistic values
MCV	Possible missing values	Medium	Fill missing values with median value
MCH	Possible missing values	Medium	Fill missing values with median value
MCHC	Possible missing values	Medium	Fill missing values with median value