Data Collection and Preprocessing Phase

Date	26 July 2025
Team ID/ Skill Wallet ID	**SWUID20250195143**
Project Title	Anemia Sense: Leveraging Machine Learning For Precise Anemia Recognitions
Maximum Marks	2 Marks

Data Collection Plan & Raw Data Sources Identification Report:

This report outlines the data collection plan and identifies raw data sources to ensure high-quality, reliable input for the anemia prediction model. Careful curation of data supports accuracy, consistency, and integrity throughout the analysis and modeling process.

Data Collection Plan:

Section	Description
Project Overview	The machine learning project aims to predict anemia status based on patient health indicators. Using features such as Gender, Hemoglobin, MCH, MCHC, and MCV, the objective is to develop a model that accurately classifies whether a patient is anemic or not, enabling early detection and preventive healthcare measures.
Data Collection Plan	• Search for datasets related to anemia diagnosis and blood test results. • Prioritize datasets with diverse demographic and clinical information to ensure model generalizability.
Raw Data Sources Identified	The raw data for this project was obtained from Kaggle, a widely used platform for data science and machine learning datasets. The collected dataset contains patient health metrics, including Gender, Hemoglobin, MCH, MCHC, and MCV, along with a target label indicating anemia status.

Raw Data Sources Report:

Source Name	Description	Location/URL	Format	Size	Access Permissions
Kaggle Dataset	The dataset contains patient health indicators including Gender, Hemoglobin, MCH, MCHC, and MCV, along with anemia diagnosis labels.	https://www.kaggle.com/code/emreiekyurt/anemia-classification-with-eda-100-acc/input	CSV	34.63 kB	Public