Data Collection and Preprocessing Phase

Date	1 August 2025
Skillwallet ID	SWUID20250194750
Project Title	Anemia Sense: Leveraging Machine Learning For Precise Anemia
Maximum Marks	2 Marks

Data Collection Plan & Raw Data Sources Identification Report:
Elevate your data strategy with a well-structured Data Collection Plan and comprehensive Raw Data Sources report, ensuring meticulous curation and data integrity to support reliable, data-driven anemia diagnosis.

Data Collection Plan:

Section	**Description**
Project Overview	The machine learning project aims to predict the presence of anemia based on patient hematological parameters. Using datasets that include features such as hemoglobin levels, etc. components, the objective is to build a robust model that accurately classifies anemia status—facilitating early detection and better clinical decision-making.
Data Collection Plan	● Search for datasets related to anemia diagnosis, including hematological test results and demographic patient data. ● Prioritize datasets with labeled outcomes (anemia vs. non-anemia) and diverse population samples. ● Ensure inclusion of common clinical features such as Hemoglobin (Hb), MCV, MCH and MCHC.
Raw Data Sources Identified	The raw data sources for this project include publicly available medical datasets from platform such as **Kaggle**. These repositories provide anonymized patient blood test records suitable for machine learning analysis. The datasets typically include clinical variables crucial for anemia classification, such as hemoglobin concentration, MCHC, and demographic features like sex .

Raw Data Sources Report:

Source Name	Description	Location/URL	Format	Size	Access Permissions
Kaggle Dataset	The dataset comprises patient details (gender) and hematological metrics (hemoglobin, MCHC, MCV, MCH), along with anemia diagnosis outcomes. It is used to predict if a patient is likely to suffer from anemia using a binary classification algorithm.	https://www.kaggle.com/datasets/biswaranjanrao/anemia-dataset	CSV	34 kB	Public