Data Collection and Pre-processing Phase

Date	6 August 2025
Team ID	SWUID20250185217
Project Name	Anemia Sense: Leveraging Machine Learning For Precise Anemia Recognitions
Maximum Marks	2 Marks

Data Collection Plan & Raw Data Sources Identification Report:

Elevate your data strategy by implementing a comprehensive Data Collection Plan coupled with a detailed Raw Data Sources Report. This approach ensures that every dataset undergoes meticulous curation, validation, and quality checks, safeguarding its integrity from the moment of acquisition. By maintaining accurate, consistent, and well-documented data sources, organizations can lay a robust foundation for insightful analysis, reliable predictions, and data-driven decision-making across all stages of the project lifecycle.

Data Collection Plan:

Section	**Description**
Project Overview	The machine learning project aims to predict anemia risk based on patient blood parameters. Using a dataset with features such as gender, hemoglobin, MCH, MCHC, and MCV, the objective is to build a model that accurately classifies anemia status (present or absent), enabling early detection and proactive health management.
Data Collection Plan	▪ Search for datasets related to anemia diagnosis and hematological health parameters. ▪ Prioritize datasets with diverse demographic and medical information.
Raw Data Sources Identified	The raw data sources for this project include dataset obtained from smartinternz. The provided dataset contains key variables such as gender, hemoglobin level, MCH, MCHC, and MCV values for machine learning analysis.

Raw Data Sources Report:

Source Name	Description	Location/URL	Format	Size	Access Permissions
smartinternz	The dataset comprises patient gender, hemoglobin levels, MCH, MCHC, MCV, and anemia diagnosis results.	https://drive.google.com/file/d/1KMJFNFGwoaQoAouIPabMEHcT1bvqEXau/view	CSV	34 kB	Public