# Data Collection and Preprocessing Phase

## Data Collection Plan and Raw Data Source Identification Report:

Elevate your data strategy with the Data Collection plan and the Raw Data Sources report, ensuring meticulous data curation and integrity for informed decision-making in every analysis and decision-making endeavor.

## Data Collection Plan:

| Section | Description |
|---|---|
| Project Overview | The **Anemia Predictor** project aims to develop a machine learning-based system to detect the presence of anemia in individuals using basic hematological features. Anemia is a common health condition, especially in developing countries, and early detection is crucial for timely treatment. |
| Data collection Plan | Data collection is fundamental to machine learning, providing the raw material for training algorithms and making predictions. This process involves gathering relevant information from various sources such as databases, surveys, sensors, and web scraping. The quality, quantity, and diversity of collected data significantly impact the performance and accuracy of ML models. |
| Raw Data Sources Identified | There are many popular open sources for collecting the data. Eg: kaggle.com, UCI repository, etc. In this project, we have used .csv data. This data is downloaded from kaggle.com. |

## Raw Data Source Report – Anemia Predictor:

| Attribute | Details |
|---|---|
| **Data Source** | Public healthcare dataset (synthetic or real), collected from clinical labs or repositories focused on anemia screening. |

| Attribute | Details |
|---|---|
| **Data Type** | Structured, tabular CSV format |
| **No. of Records** | 1,421 samples |
| **Features Used** | Gender, Hemoglobin, MCH, MCHC, MCV, Result |
| **Feature Types** | - Gender: Categorical (0: Female, 1: Male) - Hemoglobin, MCH, MCHC, MCV: Numerical - Result: Binary target (0: Not Anemic, 1: Anemic) |
| **Collection Method** | Collected via blood tests using standard lab equipment |
| **Missing Values** | None |
| **Outliers** | Slight variance in extreme ranges (e.g., MCHC min: 27.8, max: 32.5) but within physiological norms |
| **Data Cleanliness** | Data is clean, preprocessed, and normalized as needed for model input |
| **Format Received** | CSV file loaded into Pandas DataFrame during preprocessing |