

## Data Collection and Preprocessing Phase

|               |   |
|---------------|---|
| Date          | 9 June 2024   |
| Team ID       | SWTID1720112707   |
| Project Title | Anemia Sense: Leveraging Machine Learning For Precise Anemia Recognitions |
| Maximum Marks | 2 Marks   |

### Data Collection Plan & Raw Data Sources Identification Template

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 Template

| Section                     | Description   |
|-----------------------------|---|
| Project Overview            | Anemia Sense aims to leverage machine learning techniques to develop a precise anemia recognition model. The project's objective is to utilize key hematological parameters such as Age, MCH, MCHC, MCV, Hemoglobin, and Gender to accurately detect anemia in individuals.   |
| Data Collection Plan        | <p>The data will be collected from the following sources:</p> <ol style="list-style-type: none"> <li>1. Public health datasets available online.</li> <li>2. Hospital and laboratory records (with necessary permissions).</li> <li>3. Existing medical research databases and journals.</li> <li>4. Synthetic data generated to supplement training data where real data is insufficient.</li> </ol> |
| Raw Data Sources Identified | Kaggle  |

### Raw Data Sources Template

| Source Name | Description  | Location/URL  | Format | Size  | Access Permissions     |
|-------------|--|---|--------|-------|------------------------|
| Kaggle      | Comprehensive health data including hematological parameters from various hospitals. | <a href="https://www.kaggle.com/datasets/biswaranjanrao/anemia-dataset">https://www.kaggle.com/datasets/biswaranjanrao/anemia-dataset</a>                         | CSV    | 5 kB  | Public                 |
| Kaggle      | Research dataset containing detailed blood test results and diagnoses.               | <a href="https://www.kaggle.com/datasets/ehababoelnaga/anemia-types-classification">https://www.kaggle.com/datasets/ehababoelnaga/anemia-types-classification</a> | Excel  | 22 kB | Private (with access)  |
| Kaggle      | Synthetic data generated to balance class distribution and enhance model training.   | <a href="https://www.kaggle.com/datasets/saieemmohammed/anemia-detection">https://www.kaggle.com/datasets/saieemmohammed/anemia-detection</a>                     | CSV    | 3 kB  | Private (internal use) |
| Kaggle      | Anemia-specific dataset from a medical research journal.                             | <a href="https://www.kaggle.com/datasets/saurabhshahane/anemia-diagnosis-dataset">https://www.kaggle.com/datasets/saurabhshahane/anemia-diagnosis-dataset</a>     | CSV    | 8 kB  | Public                 |