Team 25: Anemia Classification Using CBC Data

1 Introduction

The `25.csv` **Complete Blood Count (CBC)** dataset provides critical hematological parameters collected from multiple CBC tests, manually diagnosed and labeled with different types of **Anemia**. This dataset serves as a foundation for developing **machine learning models** to classify anemia types based on blood test results, aiding in early detection and diagnosis. Automated classification can assist healthcare professionals in improving diagnostic accuracy and patient care.

2 Dataset Description

The dataset consists of various CBC parameters that are essential for identifying anemia types. The key features include:

2.1 Hematological Parameters

- **HGB (Hemoglobin)** – Measures the amount of hemoglobin in the blood, crucial for oxygen transport.
- **PlT (Platelets)** – The number of platelets in the blood, involved in blood clotting.
- **WBC (White Blood Cells)** – The count of white blood cells, vital for immune response.
- **RBC (Red Blood Cells)** – The count of red blood cells, responsible for oxygen transport.
- **MCV (Mean Corpuscular Volume)** – Average volume of a single red blood cell.
- **MCH (Mean Corpuscular Hemoglobin)** – Average amount of hemoglobin per red blood cell.
- **MCHC (Mean Corpuscular Hemoglobin Concentration)** – Average concentration of hemoglobin in red blood cells.
- **PDW (Platelet Distribution Width)** – Measurement of variability in platelet size distribution.
- **PCT (Procalcitonin)** – Used to diagnose bacterial infections and sepsis risk.

2.2 Additional Blood Parameters

- **LYMp** – Lymphocyte percentage.
- **NEUTp** – Neutrophil percentage.
- **LYMn** – Lymphocyte count.
- **NEUTn** – Neutrophil count.

2.3 Target Variable (Classification Task)

- **Diagnosis** – Anemia type classification based on CBC parameters.

3 Tasks and Requirements

To develop a machine learning model for anemia classification, the following tasks need to be performed:

3.1 Data Preprocessing and Feature Engineering

- Handle missing values and clean inconsistent records.
- Normalize numerical attributes such as hemoglobin, RBC, and WBC counts.
- Perform feature selection to identify the most relevant CBC parameters for anemia classification.

3.2 Anemia Type Classification (Supervised Learning)

- Train classification models.
- Evaluate model performance using **accuracy, precision, recall, and F1-score**.
- Determine key blood parameters that influence anemia classification.

3.3 Visualization and Reporting

- Generate histograms and scatter plots to visualize CBC parameter distributions.
- Create correlation heatmaps to analyze relationships between blood features and anemia diagnosis.
- Develop confusion matrices to assess classification performance.

4 Submission Requirements

- A well-structured report detailing the methodology, results, and analysis in a given report format.
- Python code is used for implementation.
- A presentation summarizing key findings and recommendations in a given presentation format.