1. What is the source of your dataset?

We used open-source healthcare datasets from trusted academic and public repositories:

1) Heart Disease Dataset: https://archive.ics.uci.edu, https://physionet.org, https://www.who.int/data
2) Stroke: https://data.gov.in, https://www.nature.com, https://clinicaltrials.gov
3) Anemia : https://www.nih.gov/health-information, https://data.gov.in, https://physionet.org

2. Why did you choose this dataset for your problem statement?

We chose these datasets because:

- They are directly related to our problem statement: AI-powered early risk prediction for patient health conditions.
- Anemia, Heart Disease, and Stroke are critical and blood-related diseases, making them suitable for a unified prediction framework.
- The datasets contain structured clinical parameters such as hemoglobin levels, blood pressure, cholesterol, glucose, age, and lifestyle indicators.
- These diseases often develop silently, and early prediction can prevent severe complications, disability, or death.
- The data is clean, interpretable, and suitable for machine learning models, especially for explainable AI in healthcare.
- Hence, these datasets clearly support early diagnosis, risk analysis, and preventive healthcare, which aligns perfectly with PS02.

3. How was the data collected?

The data was collected from open-source datasets published by academic institutions and public contributors.

The datasets were made available in CSV format and accessed directly via public repository links.

No web scraping or private APIs were used.

The data represents previously collected and anonymized clinical records, ensuring ethical and safe usage.