Datasets and Usage Plan

1. Air Quality and Health Impact Dataset – Kaggle

🔗 https://www.kaggle.com/datasets/rabieelkharoua/air-quality-and-health-impact-dataset
Description: Contains ~5,800 records linking pollutant levels ($PM_{2.5}$, NO_2, etc.) with health outcomes like respiratory illnesses.
Plan to Use:

- Use as a **baseline dataset** to test pollution–health correlation models.
- Apply data-cleaning and feature-engineering to design the structure for our own real-time data pipeline.
- Validate model accuracy before integrating IoT and clinic data.

2. Air Pollution and Emergency Room Visits Study – PMC

🔗 https://pmc.ncbi.nlm.nih.gov/articles/PMC10519391/
Description: Peer-reviewed study showing how short-term pollutant exposure increases emergency-room visits for respiratory problems.
Plan to Use:

- Extract exposure–response relationships and time-lag patterns.
- Use these as **reference parameters** for training our central AI model.
- Support our system's methodology with scientific validation.

3. Health Data Research UK Gateway

🔗 https://healthdatagateway.org/en
Description: A national platform offering access to anonymized health and clinical datasets for research.
Plan to Use:

- Reference its **data-governance and privacy frameworks** to strengthen our Federated Learning design.
- Identify compatible **aggregated health datasets** for model benchmarking.
- Adopt its metadata standards for structuring clinic data in Project Swasthya.

Overall Usage Strategy

1. Prototype correlations using the Kaggle dataset.
2. Refine model parameters using the PMC study insights.
3. Integrate real-world or simulated clinic data following Gateway standards.

4. Deploy the refined model on our AWS-based AI pipeline for real-time predictions.