Focus: SDG 6 – Clean Water and Sanitation

Title: *Clustering Regions by Water Pollution Levels Using Unsupervised Machine Learning*

1. SDG Problem Addressed

Access to clean and safe water remains a major challenge in many regions, especially where monitoring infrastructure is limited. Polluted water sources cause millions of deaths annually due to diseases such as cholera and typhoid.
To support SDG 6 (Clean Water and Sanitation), this project leverages AI and Machine Learning to analyze water quality data and cluster regions by pollution levels.
This helps policymakers prioritize areas needing urgent intervention and better allocate sanitation resources.

2. Machine Learning Approach

Approach Used: 🧠 Unsupervised Learning
Algorithm: K-Means Clustering

Reason for Choice:
Unsupervised learning is ideal when labeled data (e.g., “clean” or “polluted”) is unavailable. The model automatically groups regions based on similarities in water quality indicators such as pH, turbidity, hardness, and dissolved oxygen. This allows the discovery of hidden patterns in pollution data without manual classification.

3. Dataset and Tools

Datasets:

- *Kaggle Water Quality Dataset*
- *World Bank Open Data (Water and Sanitation Indicators)*
- *UN SDG Global Database (Water Pollution Metrics)*

Tools Used:

Tool	Purpose
Python	Data analysis and modeling
Jupyter Notebook	Interactive coding and visualization
Scikit-learn	ML algorithms (K-Means clustering)
Pandas & NumPy	Data manipulation and preprocessing
Matplotlib / Seaborn	Visualization of pollution clusters

4. Model Building Workflow

1. Data Preprocessing
    - Remove missing values and normalize features (e.g., Min-Max Scaling).
    - Select key variables: pH, Turbidity, Dissolved Oxygen, Conductivity, Nitrate Levels.
2. Model Training
    - Use K-Means algorithm to cluster regions into 3 groups:
        - Cluster 1: *Clean water sources*
        - Cluster 2: *Moderately polluted areas*
        - Cluster 3: *Highly polluted areas*
3. Evaluation & Visualization
    - Apply Elbow Method to determine the optimal number of clusters.
    - Plot scatter diagrams showing clustered regions.
    - Validate clustering using Silhouette Score.

Sample Result:
The model successfully identified clusters where high turbidity and nitrate concentration were strongly correlated with unsafe water levels, guiding where sanitation improvements are most needed.

5. Ethical Reflection

- Bias in Data:
  If water quality data is collected unevenly (e.g., only from urban areas), rural contamination might be underrepresented, leading to biased conclusions.
  To mitigate this, diverse and balanced datasets from multiple sources should be used.
- Fairness and Sustainability:
  This model promotes fairness by using open data and transparent algorithms. It supports sustainable resource management, enabling governments to focus water treatment efforts where communities are most vulnerable.

6. Deliverables

- Code Notebook: Python (Jupyter Notebook) with commented workflow, data preprocessing, and model training steps.
- Report (This Summary): 1-page overview of SDG, ML approach, results, and ethics.
- Presentation: 5-minute demo showing how AI identifies polluted regions and helps prioritize interventions.

7. Expected Impact

This AI-driven clustering system helps governments and NGOs:

- Identify pollution hotspots efficiently.
- Allocate clean water resources more equitably.
- Support SDG 6: Clean Water and Sanitation and SDG 3: Good Health and Well-being by ensuring access to safe drinking water.