AI-Powered River Pollution Monitoring System

I. Abstract

This project presents an **AI-powered system for monitoring and classifying river pollution** across three major Indian rivers — **Ganga**, **Sangam**, and **Yamuna**. By integrating structured sensor datasets and image-based pollution annotations, the system provides a hybrid approach to environmental monitoring. Through the use of supervised machine learning and visual analysis, it enables accurate assessment of water quality and helps identify pollution trends.

The system is designed to be extendable for real-time IoT integrations, drone imagery, and smart city applications focused on **sustainable river management**.

II. Dataset Description

- **Ganga and Sangam Rivers**: Sensor data in CSV format capturing:
  - Dissolved Oxygen (DO)
  - pH Levels
  - Oxidation Reduction Potential (ORP)
  - Temperature (Temp)
  - Conductivity (Cond)
  - Water Quality Index (WQI)
  - Status (Target Label: e.g., *Very Poor*, *Fair*, *Good*)
- **Yamuna River**: Annotated image regions from JSON files, labeled as:
  - `"polluted": "yes"`
  - `"polluted": "no"`
  - Each image contains multiple labeled regions using polygon annotations.
- Together, these datasets provide both **numerical evidence** and **visual validation** for AI-based pollution analysis.

III. Methodology

1. **Data Cleaning & Preprocessing**
   - Missing value handling
   - Outlier filtering
   - Data normalization

2. **Feature Selection**
    - Input features: `DO`, `pH`, `ORP`, `Cond`, `Temp`, `WQI`
    - Target: `Status` label encoded into numerical classes
3. **Model Training**
    - Algorithm: `RandomForestClassifier`
    - Training/Testing split with evaluation using:
        - Accuracy
        - Classification report
        - Confusion matrix
4. **Model Saving**
    - Trained model exported using `joblib` as `rf_model.pkl`
5. **Prediction on Unseen Data**
    - Results exported to `river_pollution_predictions.csv` for analysis
6. **Visual Annotation Integration**
    - Analyzed Yamuna image region labels to support model results
    - Labeled image areas provide qualitative pollution indicators

IV. Key Visualizations

- **DO vs WQI**
  → Bubble chart with temperature as bubble size
- **pH vs Status**
  → Boxplot comparing pH ranges across pollution categories
- **WQI vs Temperature**
  → Scatter plot highlighting trends between heat and water quality
- **ORP vs Conductivity**
  → Bi-variable scatter plot with `Status` and `River` color/style
- **Average WQI per River**
  → Comparative bar chart to measure overall river health

V. Model Output & Prediction

- Trained ML model is saved as `rf_model.pkl`
- Can predict pollution status from unseen water quality data
- Predicted values validated with real-world images and pollution scores
- Output saved in `river_pollution_predictions.csv` for reporting

VI. Visual Analysis of Yamuna (Image Annotations)

- Each image labeled with one or more regions marked as:
  - `Polluted (yes)`
  - `Not Polluted (no)`
- This annotation data helps:
  - Qualitatively verify the ML predictions
  - Enable future development of **AI-based image classifiers** for river monitoring
  - Combine **sensor + visual data** for hybrid AI models

VII. Tools & Technologies

Category	Tools
Programming	Python 3.x
Data Handling	pandas, numpy
ML & Modeling	scikit-learn, joblib
Visualization	seaborn, matplotlib, plotly
Platform	Jupyter Notebook / Google Colab
Version Control	GitHub
Future Scope	Streamlit, TensorFlow (for image-based models)

Conclusion

The **AI-powered River Pollution Monitoring System** provides an end-to-end pipeline for environmental analysis using both quantitative sensor data and qualitative image evidence. It demonstrates:

- Practical use of machine learning in environmental monitoring
- Visual validation using manually labeled image regions
- Scalable design for smart city and research applications

This project sets the foundation for future integrations with **drone footage**, **satellite imaging**, and **real-time IoT sensors** to create a more intelligent and responsive water quality monitoring system.