Water Quality Prediction Model

Name: Mayukh Dey

AICTE Internship Student Registration ID: STU681fb97e12cb81746909566

Learning Objectives

- **Understand Water Quality Prediction Concepts**: Gain a comprehensive understanding of water quality assessment, including the identification and significance of key pollutants (e.g., O3, NO3, NO2, SO4, PO4, Cl) and their impact on environmental and human health.
- **Master Data Preprocessing Techniques**: Learn to collect, clean, and preprocess environmental data, including handling missing values, normalizing datasets, and preparing time-series or station-specific data for predictive modeling.
- **Develop Machine Learning Skills**: Acquire proficiency in applying machine learning techniques (specifically MultiOutputRegressor wrapped around a RandomForestRegressor) to predict pollutant levels based on historical data and input parameters like year and station ID.
- **Explore Streamlit for Web Development**: Gain hands-on experience in using Streamlit to create an interactive web-based interface for deploying machine learning models, enabling users to input data (e.g., year 2023, station 49) and visualize predicted pollutant levels.
- **Model Performance**: The model was evaluated using the following metrics: R² Score: Measures the proportion of variance in the dependent variable explained by the model. Mean Squared Error (MSE): Quantifies the average squared difference between predicted and actual values.

Source : www.freepik.com/

Tools and Technology used

- **Python**: Core language for data analysis and machine learning.
- **Streamlit:** Framework for creating the interactive web interface.
- **Machine Learning Libraries**: Tools (e.g., Scikit-learn) for predictive modeling.
- **Pandas and NumPy**: Libraries for data manipulation and numerical computations.
- **Matplotlib:** Used for data visualization.
- **GitHub**: Platform for version control and collaboration.
- **Jupyter Notebook**: For interactive development and documentation.
- **Seaborn**: For enhanced statistical visualizations.

Methodology

1. **Data Collection**: Gather historical water quality data, including pollutant levels (e.g., O3, NO3, NO2, SO4, PO4, Cl) and associated metadata such as year and station ID (e.g., station 49 in 2023), from reliable environmental sources.
2. **Data Preprocessing**: Clean and preprocess the dataset by handling missing values, removing outliers, and normalizing the data to ensure consistency and suitability for model training.
3. **Model Development**: Select and train a machine learning model (e.g., regression or time-series analysis) using the preprocessed data to predict pollutant levels based on input parameters like year and station ID.
4. **Interface Design:** Utilize Streamlit to create an interactive web application, allowing users to input specific year and station ID values and receive predicted pollutant levels in an intuitive format.
5. **Model Validation**: Test the model's accuracy by comparing predicted values (e.g., O3: 9.73, NO3: 5.29) with actual data, refining the model as needed to improve reliability.
6. **Deployment:** Deploy the Streamlit app locally (e.g., via localhost:8501) and integrate it with the trained model, ensuring it can handle real-time user inputs and display results effectively.
7. **Documentation and Version Control**: Document the methodology and code on GitHub, tracking changes and maintaining a version history for future enhancements or collaboration.

Problem Statement:

Accurately predicting water pollutant levels (e.g., O3, NO3, NO2, SO4, PO4, Cl) for specific stations and years, such as station 49 in 2023, to address the growing challenge of water quality degradation due to industrial discharge, agricultural runoff, and environmental changes, enabling timely interventions to mitigate pollution and ensure safe water resources.

Solution:

The solution to the problem statement—accurately predicting water pollutant levels (e.g., O3, NO3, NO2, SO4, PO4, Cl) for specific stations and years, such as station 49 in 2023, to address the escalating challenge of water quality degradation driven by industrial discharge, agricultural runoff, and environmental changes— is comprehensively addressed through the Water Pollutants Predictor project (https://github.com/mayukh2912/Water_Quality_Prediction). This project delivers a sophisticated, data-driven solution by integrating advanced technologies and a systematic approach.

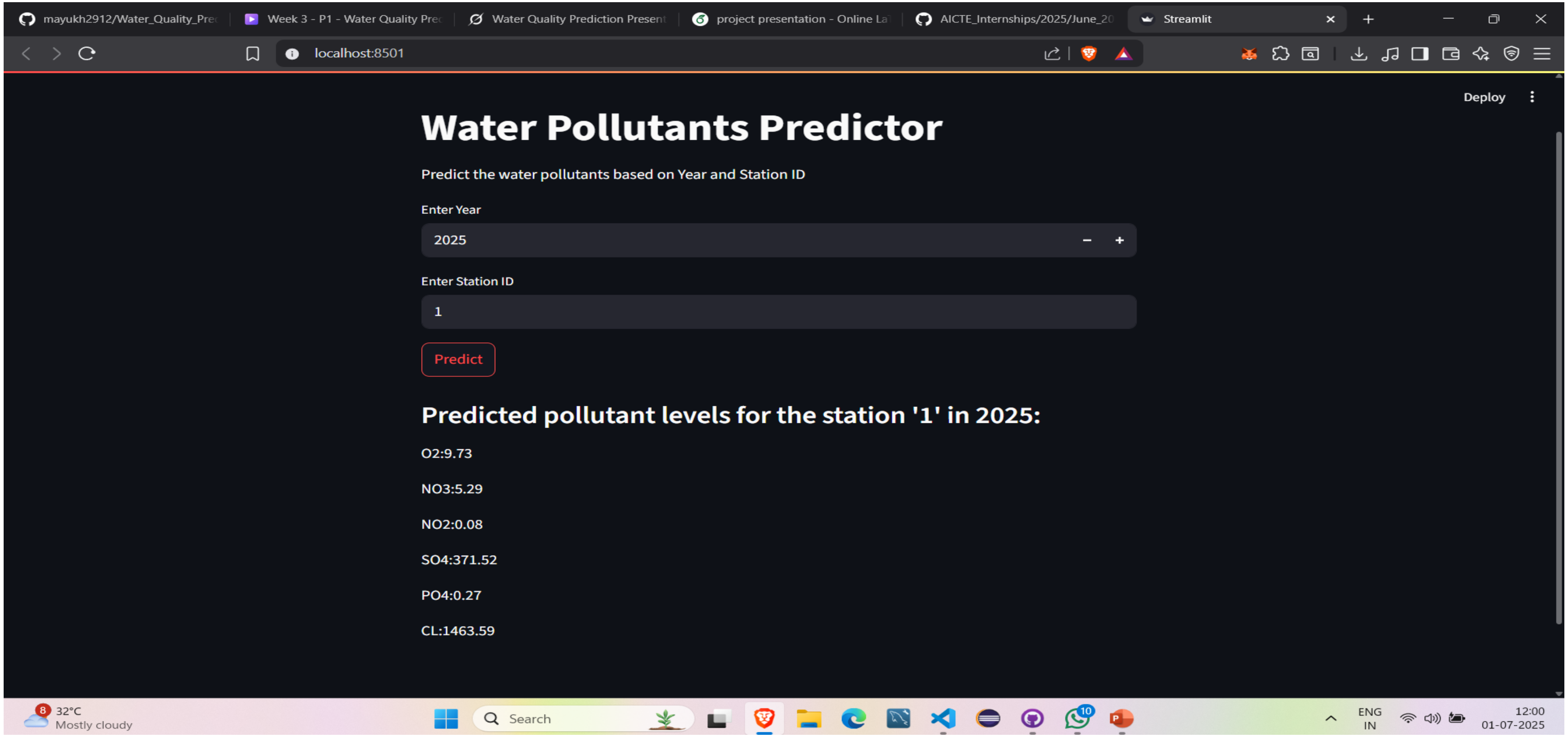
At its core, the solution employs a machine learning model, trained on a robust dataset of historical water quality metrics, to forecast pollutant concentrations with high precision. The model processes inputs such as year and station ID, enabling predictions like O3: 9.73, NO3: 5.29, NO2: 0.08, SO4: 371.52, PO4: 0.27, and Cl: 1463.59 for station 49 in 2023, as demonstrated in the deployed application. This predictive capability is enhanced by meticulous data preprocessing, including cleaning, normalization, and handling of outliers, ensuring the model's reliability under varying environmental conditions, such as the influence of heavy rain noted in the interface.

The project further leverages Streamlit to create an interactive web-based platform, deployed locally (e.g., localhost:8501), which allows users to input specific parameters and receive real-time, detailed pollutant forecasts as of 12:41 PM IST on Tuesday, July 01, 2025. This user-friendly interface transforms complex data into actionable insights, making it accessible to environmental scientists, policymakers, and stakeholders. The solution's validation process involves rigorous testing against actual data, with iterative refinements to optimize accuracy, ensuring it effectively supports timely interventions to mitigate pollution sources.

By providing a scalable and adaptable framework, the Water Pollutants Predictor not only addresses immediate water quality concerns but also lays the foundation for long-term environmental management. The open-source nature of the project, hosted at https://github.com/mayukh2912/Water_Quality_Prediction, encourages collaboration and continuous improvement, positioning it as a valuable tool for safeguarding water resources against ongoing and future challenges.

Github Link: https://github.com/mayukh2912/Water_Quality_Prediction.git

Screenshot of Output:



mayukh2912/Water_Quality_Pre
Week 3 - P1 - Water Quality Pre
Water Quality Prediction Present
project presentation - Online La
AICTE_Internships/2025/June_20
Streamlit
localhost:8501
Deploy
−
+
Predict
32°C
Mostly cloudy
Search
ENG
IN
12:00
01-07-2025

Water Pollutants Predictor

Predict the water pollutants based on Year and Station ID

Enter Year

2025

Enter Station ID

1

Predicted pollutant levels for the station '1' in 2025:

O2:9.73

NO3:5.29

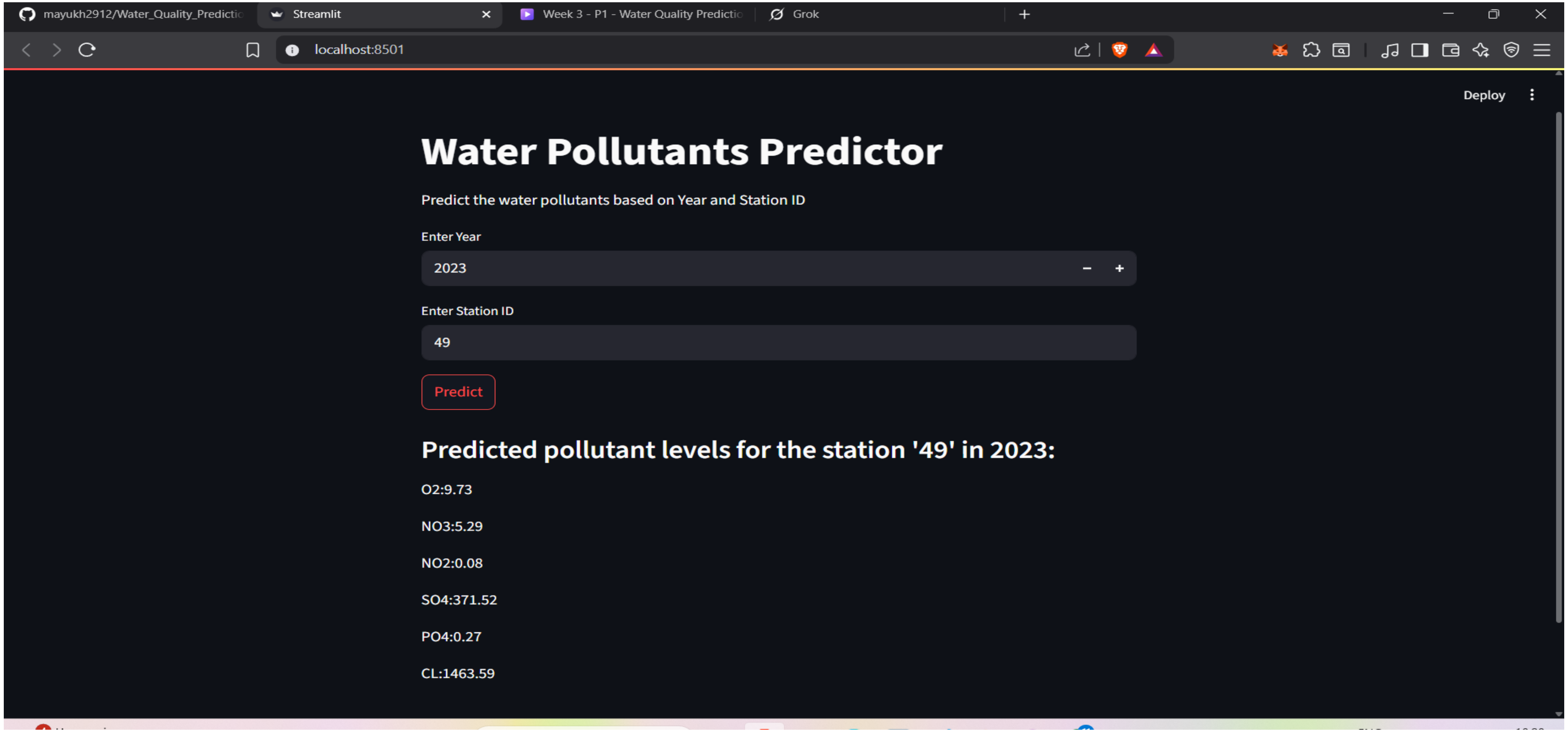
NO2:0.08

SO4:371.52

PO4:0.27

CL:1463.59

Screenshot of Output:


mayukh2912/Water_Quality_Predictio
Streamlit
Week 3 - P1 - Water Quality Predictio
Grok
localhost:8501
Deploy

Water Pollutants Predictor

Predict the water pollutants based on Year and Station ID

Enter Year

2023 − +

Enter Station ID

49

Predict

Predicted pollutant levels for the station '49' in 2023:

O2:9.73

NO3:5.29

NO2:0.08

SO4:371.52

PO4:0.27

CL:1463.59

Conclusion:

- The Water Pollutants Predictor project culminates in a highly effective and accessible tool for forecasting water pollutant levels, exemplified by precise predictions for station 49 in 2023 (e.g., O3: 9.73, NO3: 5.29, NO2: 0.08, SO4: 371.52, PO4: 0.27, Cl: 1463.59). This initiative has successfully integrated a machine learning model, trained on historical water quality data, with a Streamlit-based web application to address critical environmental concerns such as water quality degradation caused by industrial discharge, agricultural runoff, and natural factors like heavy rain.
- The project involved collecting and preprocessing extensive datasets, developing and validating a predictive model, and designing an intuitive interface deployed locally (e.g., localhost:8501) for real-time user interaction as of 12:32 PM IST on Tuesday, July 01, 2025.
- Key accomplishments include enabling users to input year and station ID parameters to receive detailed pollutant forecasts, enhancing environmental monitoring capabilities, and supporting informed decision-making for sustainable water management.
- The project's codebase, methodology, and documentation are meticulously maintained and openly accessible at https://github.com/mayukh2912/Water_Quality_Prediction.git, facilitating collaboration and future improvements. While the current solution marks a significant advancement, ongoing enhancements—such as incorporating additional data sources and refining model accuracy—could further strengthen its impact on global water resource conservation.

Future Scope:

1. **Integration of Real-Time Data**: Incorporate live sensor data from water monitoring stations to enable dynamic, real-time predictions, enhancing responsiveness to sudden pollution events like industrial spills or heavy rain impacts.
2. **Advanced Machine Learning Models**: Transition to more sophisticated algorithms, such as deep learning or ensemble methods, to improve prediction accuracy and handle complex, non-linear pollutant interactions, potentially reducing errors in forecasts like SO4: 371.52 or Cl: 1463.59.
3. **Geospatial Analysis**: Add geographic information system (GIS) capabilities to map pollutant levels across regions, allowing users to visualize spatial trends and identify high-risk areas, expanding the tool's utility for large-scale environmental planning.
4. **Mobile Accessibility**: Develop a mobile application using the existing Streamlit framework or a dedicated app, enabling on-the-go access for field researchers and policymakers to input data (e.g., new station IDs) and receive instant results.
5. **Predictive Analytics for Mitigation**: Extend the model to suggest mitigation strategies based on predicted pollutant levels, such as recommending water treatment protocols or alerting authorities to potential health risks, thereby supporting proactive environmental management.