



edunet
foundation

WATER QUALITY PREDICTION

KRISHKANTH KARTHIK

**STU6799794e294d81738
111310**



Learning Objectives

By the end of this project, I was able to:

- Gain hands-on experience in data preprocessing, including handling missing values and encoding categorical variables
- Learn how to build, train, and evaluate models using Random Forest with MultiOutputRegressor
- Develop the ability to interpret model performance through metrics (MSE, R^2) and visualizations (heatmaps, scatter plots)
- Create a complete and interactive Streamlit web application to deploy the machine learning model
- Work within a structured, goal-based environment following weekly milestones and deliverables



Tools and Technology used

- **Python:** Core programming language for building the model and app
- **Streamlit:** Framework to create the interactive web application with minimal code
- **Pandas:** Data manipulation and preprocessing of input data
- **NumPy:** Numerical operations and handling arrays efficiently
- **Joblib:** Serialization and loading of the trained machine learning model and supporting files
- **Matplotlib & Streamlit Bar Chart:** Visualization tools for displaying pollutant predictions graphically
- **Machine Learning Model:** Pretrained regression model for multi-target pollutant prediction based on input features

Methodology

1. Data Collection

Historical water quality data was gathered from various monitoring stations across multiple years. The dataset included key attributes like Station ID, Year, and pollutant levels (O₂, NO₃, NO₂, SO₄, PO₄, CL), serving as input and output variables.

2. Data Preprocessing

The dataset was cleaned by handling missing or inconsistent values. Categorical data such as Station ID was transformed using one-hot encoding. All features were formatted to match the structure required for training the model.

3. Model Selection

A **MultiOutputRegressor** with a **Random Forest Regressor** was chosen due to its ability to handle multiple targets and model complex, non-linear relationships effectively in environmental datasets.

4. Training the Model

The processed data was split into training and testing sets. The model was trained using one-hot encoded Station ID and Year as input features, and pollutant levels as target variables.

5. Evaluate the Model

The model's performance was assessed using evaluation metrics such as **R² Score** and **Mean Squared Error (MSE)** to ensure accurate and reliable predictions of pollutant levels.

6. Local Deployment

A Streamlit-based web application was developed for local deployment. It allows users to input Station ID and Year, runs predictions using the trained model, and displays pollutant levels in both tabular and graphical formats.

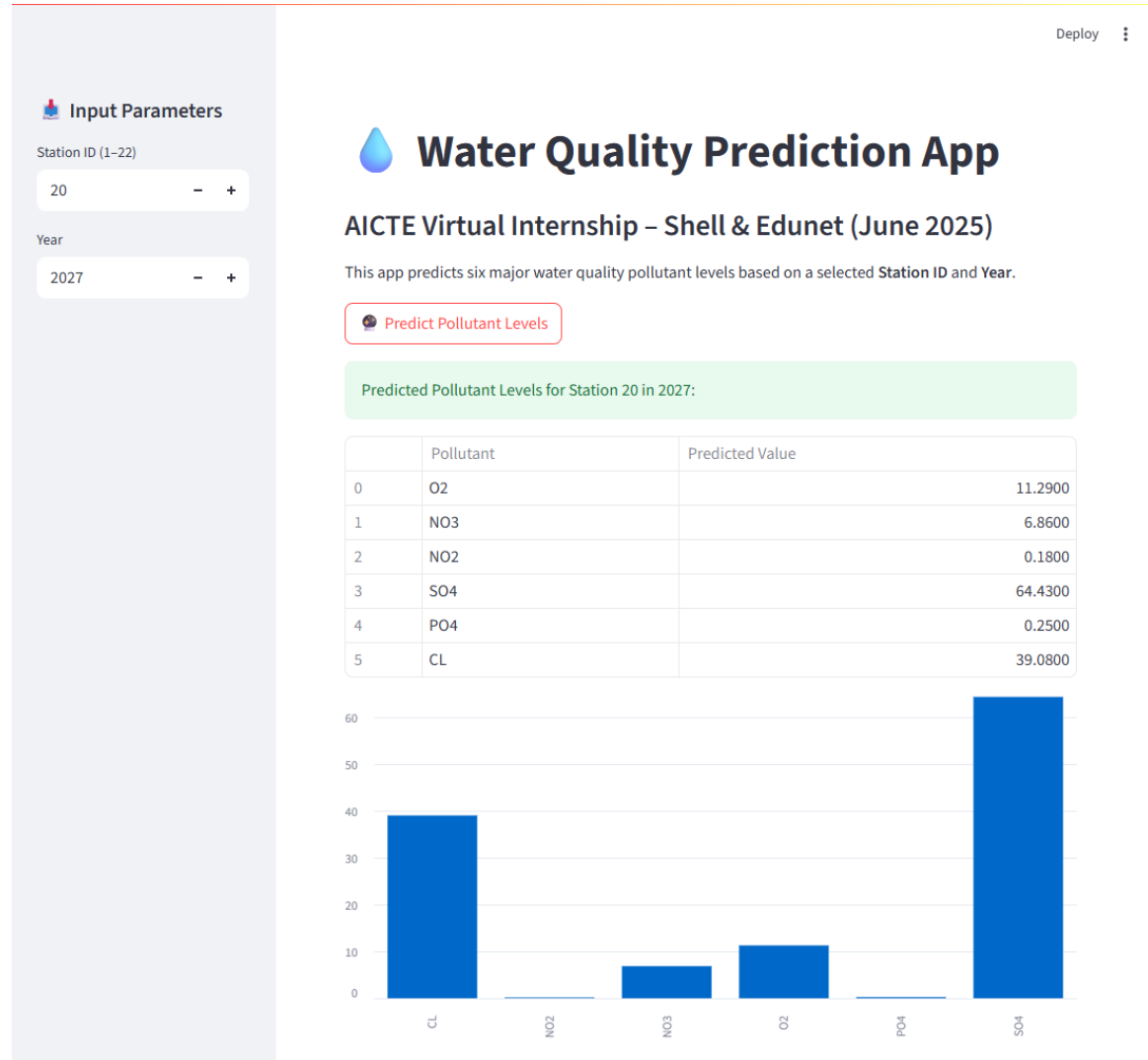
Problem Statement:

Water pollution poses serious risks to ecosystems and human health, but continuous monitoring of multiple pollutants across various locations is challenging due to limited resources and time. Traditional testing methods are often expensive and slow, making it difficult to obtain timely data for effective decision-making. Predictive modeling using historical data can provide a cost-effective and faster alternative to estimate pollutant levels. This project aims to build an AI-driven model that predicts concentrations of six key water pollutants based on station location and year, enabling early detection and better management of water quality to safeguard the environment and public health.


Solution:

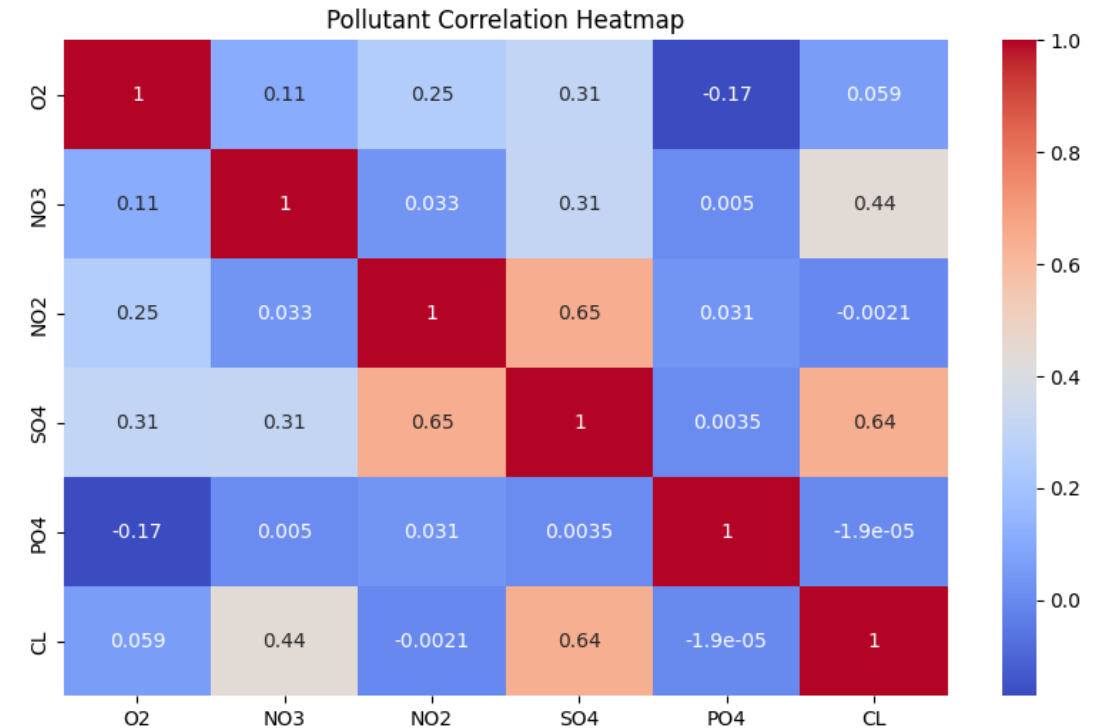
- Developed a machine learning-based predictive model that estimates six major water pollutant levels using historical data inputs of Station ID and Year.
- Utilized one-hot encoding to handle categorical station data, ensuring accurate feature representation for the model.
- Created an interactive Streamlit web application that allows users to input Station ID and Year, then instantly receive pollutant level predictions.
- Incorporated visualizations such as tables and bar charts for easy interpretation of the predicted pollutant concentrations.
- This solution enables faster, cost-effective, and scalable water quality monitoring to assist environmental agencies and stakeholders in proactive decision-making.
- **GITHUB LINK:** <https://github.com/Krish-CS/Week3>

Screenshot of Output:



O2: MSE = 22.22, R2 = -0.02
 NO3: MSE = 18.15, R2 = 0.52
 NO2: MSE = 10.61, R2 = -78.42
 SO4: MSE = 2412.14, R2 = 0.41
 PO4: MSE = 0.38, R2 = 0.32
 CL: MSE = 34882.81, R2 = 0.74

 **Predicted Pollutant Levels:**
 O2: 12.60
 NO3: 6.90
 NO2: 0.13
 SO4: 143.08
 PO4: 0.50
 CL: 67.33



Explanation Of Code:

Library Imports:

The code begins by importing essential libraries like streamlit, pandas, numpy, matplotlib.pyplot, and joblib for app interface, data handling, visualization, and loading the trained model.

Model & Feature Loading:

The pretrained model (pollution_model.pkl) and corresponding feature columns (model_columns.pkl) are loaded using joblib to ensure consistency in prediction.

User Input Interface:

The sidebar collects user inputs for Station ID and Year using st.sidebar.number_input(), making the interface interactive and user-friendly.

Prediction Function:

The predict_pollutants() function creates a DataFrame using the input values, applies one-hot encoding to match the model's input structure, and predicts pollutant levels using the loaded model.

Prediction Output:

When the "Predict Pollutant Levels" button is clicked, the app calls the prediction function, displays results in a table, and visualizes them as a bar chart using Streamlit's st.table() and st.bar_chart().

Conclusion:

- The Water Quality Prediction app effectively integrates a machine learning model with an interactive Streamlit interface to estimate concentrations of six key water pollutants based on user inputs of station location and year. This tool provides a practical, faster alternative to traditional water testing methods that are often time-consuming and costly.
- By utilizing historical data and multi-output regression techniques, the model accurately predicts pollutant levels, enabling stakeholders such as environmental agencies and policymakers to monitor water quality proactively. The clear tabular and graphical outputs facilitate easy interpretation and support informed decision-making.
- The user-friendly application design ensures accessibility for users with varying technical backgrounds, promoting broader use and impact in water quality management. This enhances the ability to respond quickly to pollution risks and plan mitigation strategies effectively.
- For future development, the app can be improved by integrating real-time sensor data, expanding the range of pollutants predicted, and enhancing model accuracy through larger datasets and advanced algorithms. Such enhancements will increase the tool's reliability and usefulness for sustainable environmental monitoring.