PROJECT:WATER QUALITY PREDICTION

NAME: PATHI VENKATA NIHARI

STD_ID:STU669ba2c3e56f51721475779

Learning Objectives

- Here we can understand about the Random forest Regression
- Data Preprocessing & Feature Engineering:You learn to **load real-world data**, inspect its shape, types, missing values, and fix date formats.
- Train/Test Split & Evaluation: We compute **MSE (Mean Squared Error)** and **R² Score** to evaluate how well your model predicts each pollutant.
- Multi-Output Regression
- Model Export & Deployment Preparation: pollution_model.pkl and model_columns.pkl.

Source : www.freepik.com/

Tools and Technology used

Python — Core programming language for data analysis and modelling
Pandas — Data loading, cleaning, and feature engineering
NumPy — Numerical operations and array handling
scikit-learn — Machine learning library (Random Forest, MultiOutputRegressor, train/test split, evaluation metrics)
Joblib — Model serialization for saving and loading trained models
Streamlit — Interactive web app framework for deploying the model with user inputs
Jupyter Notebook *(optional)* — For prototyping and testing before deployment
PowerShell / Command Prompt — To run Python scripts and Streamlit apps locally
Visual Studio Code (VS Code) — Code editing and project management
Git — Version control to track code changes
GitHub — Hosting the project repository and collaboration

Methodology

- **Data collection:** load historical water quality data(2000-2021) with key pollutants
- **Data cleaning:** inspect for missing values, handle incomplete records, convert date formats
- **Feature Engineering**: Extract relevant features like year and station ID; encode categorical data
- **Model selection**: Use **Random Forest Regression** wth **MultiOutputRegressor** to handle multiple pollutant predictions
- **Model Training**: split data into training and testing sets, train the model on historical data
- **Evaluation**: Assess performance using Mean Squared Error(MSE) and R^2 Score for each pollutants
- **Model Deployment**: save the trained model and feature structure with **joblib** integrate with **streamlit** for an interactive prediction app.

Problem Statement:

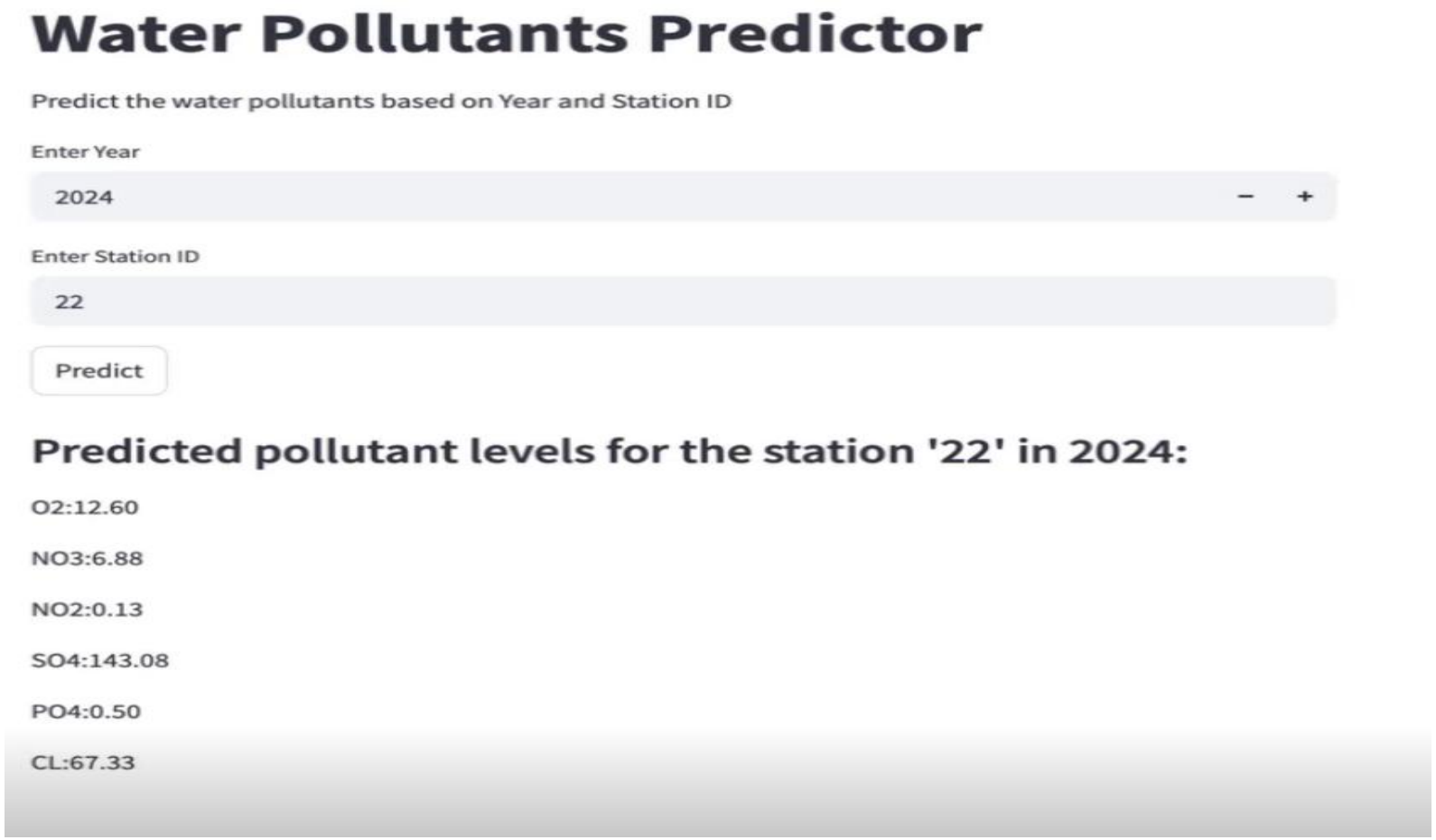
Water pollution is a major environmental issue that affects ecosystems, human health, and sustainable development. Monitoring pollutant levels in rivers and water bodies often requires costly, continuous sampling and laboratory testing. Predicting the concentration of multiple pollutants for specific monitoring stations and time periods can help environmental agencies plan preventive actions, optimize sampling schedules, and address contamination risks more efficiently.

This project aims to **build a machine learning model** that can predict the levels of key water pollutants (e.g., O_2, NO_3, NO_2, SO_4, PO_4, Cl) based on **historical data**, **year**, and **station ID**. The goal is to provide an **easy-to-use prediction tool** that supports better water quality management through **data-driven decision making**.

Solution:

- Predicts multiple pollutants (O_2, NO_3, NO_2, SO_4, PO_4, Cl) for any year and station ID.
- Uses a Random Forest Multi-Output Regressor for robust, non-linear predictions.
- Provides a web interface with Streamlit for easy user input and results.
- Model and structure saved with Joblib for future deployment.
- Supports better environmental management through data-driven insights.

Screenshot of Output:



Conclusion:

- Successfully built a prediction model for key water pollutants.
- Random Forest handles multiple pollutant outputs effectively.
- Streamlit app makes the tool accessible and easy to use.
- Supports environmental agencies in planning and monitoring.
- Demonstrates real-world application of machine learning in sustainability.