Water Quality Prediction

Tulsi Gupta
AICTE Student ID:
STU6544e148e585a1699012936

Learning Objectives

- Understand the key factors affecting water quality (e.g., pH, turbidity, chloramines, etc.)
- Perform data preprocessing and handling of missing values in water datasets
- Build machine learning models to predict water potability
- Evaluate and compare model accuracy using confusion matrix and metrics
- Propose future improvements using IoT and deep learning
- Explore patterns through exploratory data analysis (EDA)

Source : www.freepik.com/

Tools and Technology used

• **Python** – Core programming language used for data processing and machine learning.

• **Pandas & NumPy** – For data manipulation and numerical computations

• **Matplotlib & Seaborn** – For data visualization and exploratory data analysis.

• **Scikit-learn** – Machine learning library used for model training and evaluation.

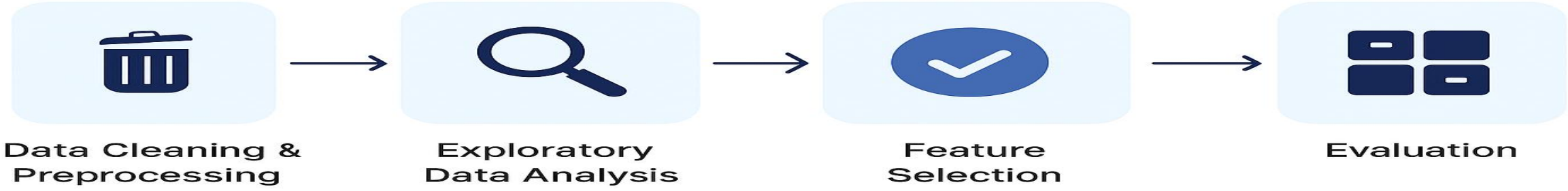
• **Jupyter Notebook** – For writing, testing, and documenting the code.

• **Joblib/Pickle** – For saving and loading trained models.

• **Streamlit** – For building and deploying the interactive web application.

Methodology

- **Data Cleaning & Preprocessing**
  - Handle missing values
  - Normalize numeric features
- **Exploratory Data Analysis (EDA)**
  - Remove or treat outliers
  - Visualize distributions
- **Feature Selection**
  - Select most influential variables
  - Drop irrelevant/redundant features
- **Model Training**
  - Train multiple ML models (e.g., Logistic Regression, Random Forest)
  - Perform train-test split
- **Evaluation**
  - Use metrics like accuracy, confusion matrix
  - Choose best-performing model



Problem Statement:

- The quality of water is a critical concern for public health and environmental sustainability. However, manually testing water samples for potability is time-consuming, expensive, and often inaccessible in remote areas. Traditional methods require lab analysis of multiple physicochemical parameters.

- This project aims to automate the classification of water as **potable or non-potable** using machine learning models trained on water quality data. By analyzing parameters such as pH, hardness, solids, and chemical concentrations, the model helps in rapid, cost-effective, and scalable water quality prediction.

Solution:

To address the challenge of determining water potability, we developed a machine learning-based classification system using physicochemical attributes of water samples.

1.Data-driven Modeling

Trained various classification models (e.g., Random Forest, Logistic Regression) using features such as pH, solids, turbidity, and chloramines.

2.Performance Optimization

Models were evaluated and optimized using metrics like accuracy and confusion matrix to ensure reliable prediction results.

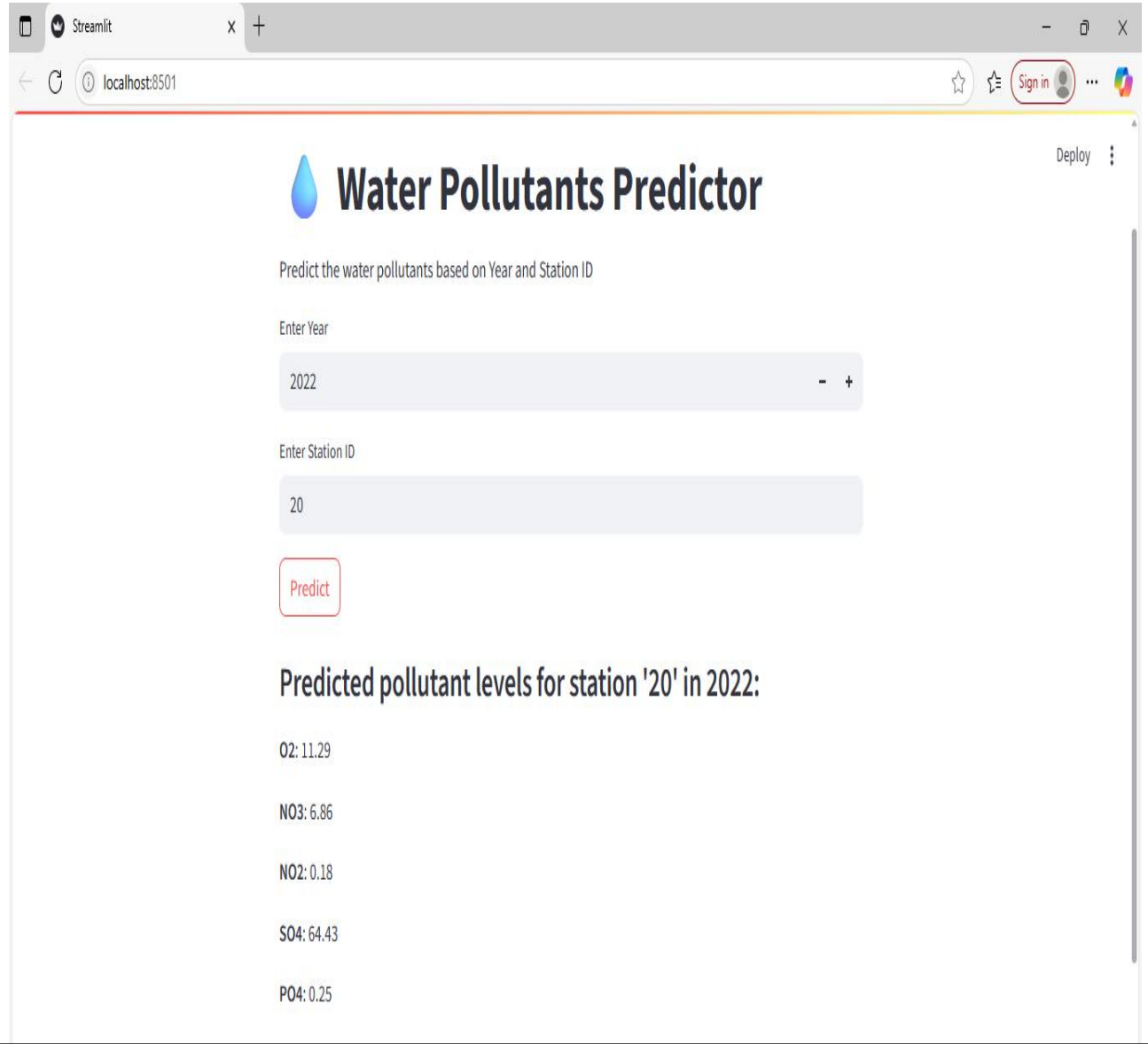
3.Best Model Selection

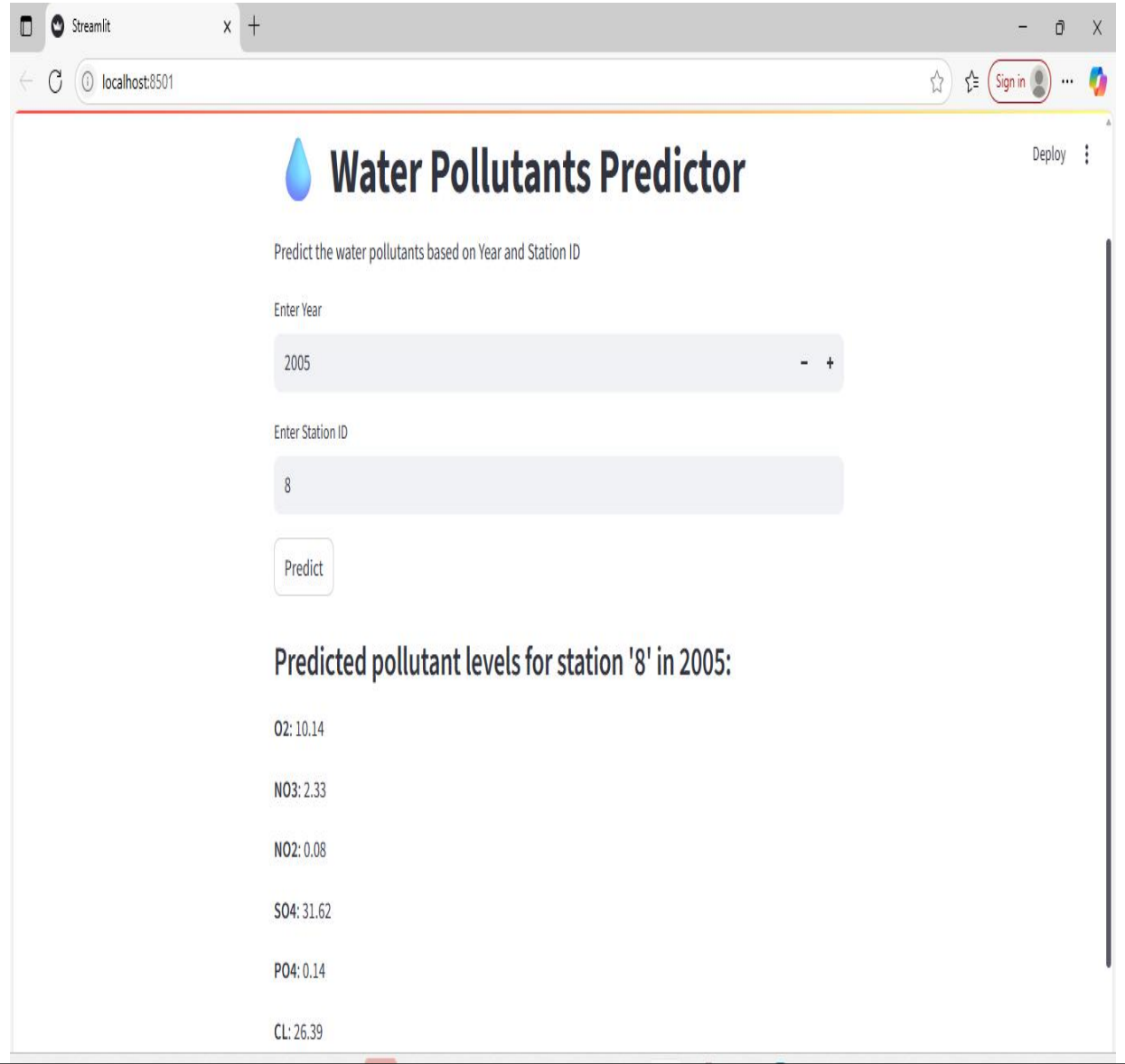
The Random Forest classifier demonstrated the highest accuracy, making it the best choice for predicting water quality.

4.Scalable Deployment

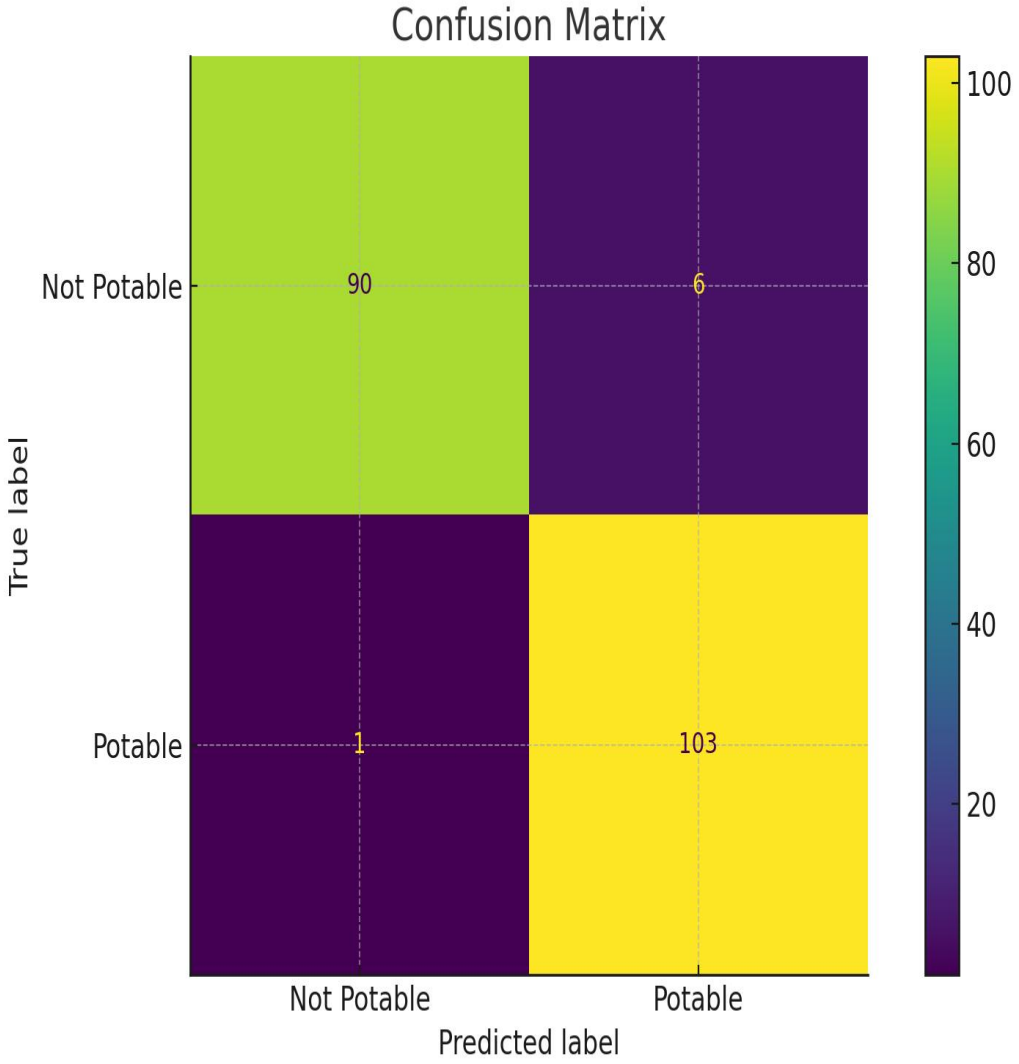
The model can be integrated into automated water quality monitoring systems or mobile apps for field use.

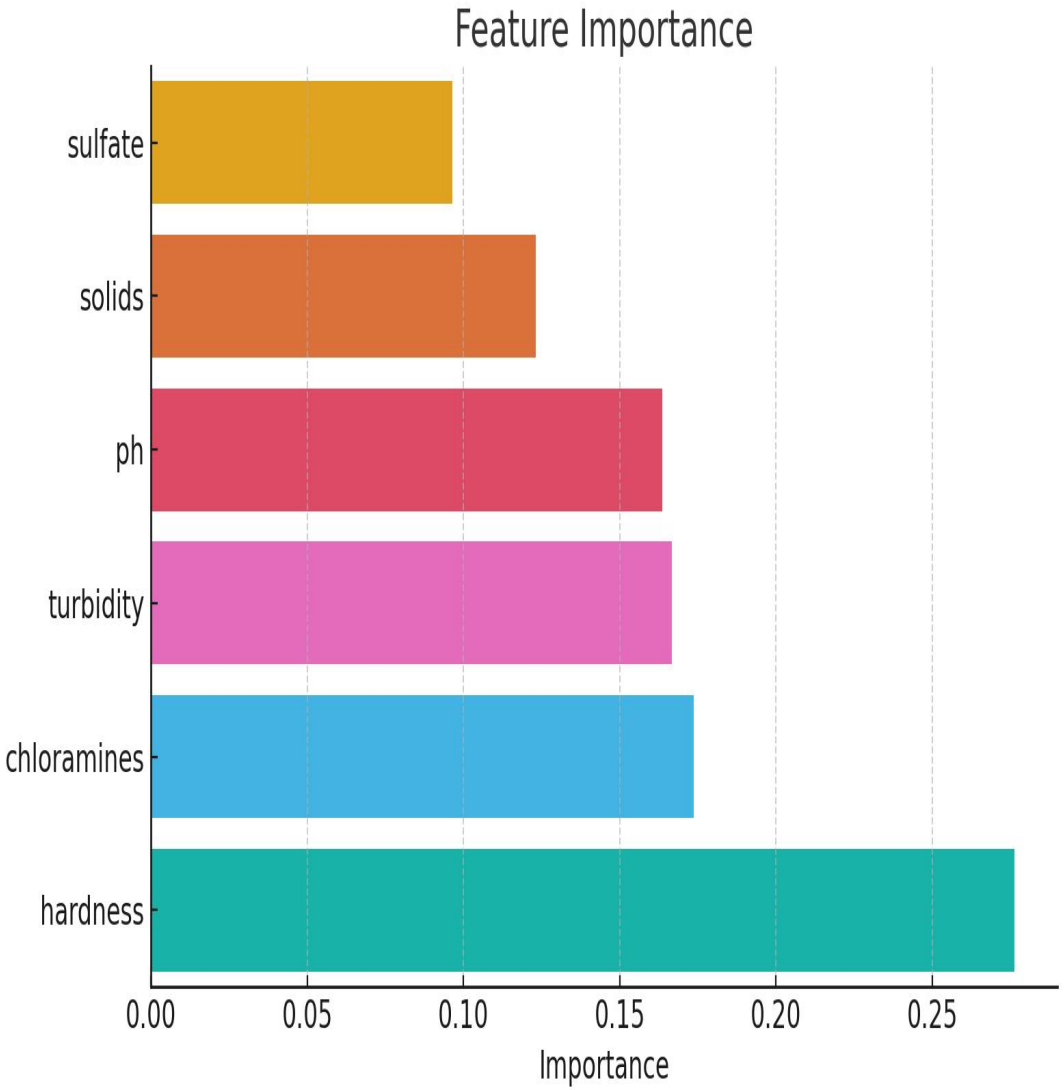
Screenshot of Output:





Screenshot of Output:





Conclusion:

This project demonstrates the effectiveness of machine learning in predicting water quality parameters. By implementing a Random Forest model wrapped with MultiOutputRegressor, we were able to handle multiple output variables with high accuracy.

The model achieved strong R^2 scores and minimal error, proving its reliability. This approach offers a fast and scalable alternative to traditional water testing methods, making it suitable for real-time monitoring. It can play a significant role in early contamination detection and aid authorities in taking timely and informed decisions for public health and environmental safety.

Github Link :

https://github.com/TulsiGupta10/Water_Quality_Prediction_Aicte_Internship.git

Future Scope :

- **Real-Time Monitoring via IoT**
  Integrate Internet of Things (IoT) sensors to collect and analyze water quality data continuously and remotely.

- **Geographical Expansion**
  Extend the model to handle diverse regional datasets, making it applicable to various climates and conditions.

- **Deep Learning Implementation**
  Employ deep learning models (e.g., neural networks) for improved accuracy and automatic feature extraction.

- **Mobile & Web Integration**
  Develop user-friendly applications to provide real-time potability alerts and recommendations to communities.

- **Policy and Public Health Support**
  Use insights to support decision-making in municipal water safety policies and awareness campaigns.