WATER POTABILITY

Kislay

210107045

Date of submission= 25/04/2024

Final Project Submission

Course Name- Application of AI and ML

in Chemical Engineering

Course Code-CL653

Contents

1.Executive Summary

Overview: The Water Potability AI/ML project aims to address the critical issue of assessing the safety and quality of drinking water by leveraging machine learning algorithms. With increasing concerns about water contamination and its adverse effects on public health, accurate prediction of water potability becomes essential for ensuring safe drinking water supplies.

Problem Statement: The project tackles the challenge of reliably determining the potability of water samples based on various physicochemical and chemical properties. Traditional methods for water quality assessment are often time-consuming, costly, and may not provide real-time insights, making them inadequate for addressing the dynamic nature of water contamination.

Proposed Solution: The proposed solution involves developing predictive models using machine learning techniques that can efficiently analyze water quality data and classify water samples as potable or non-potable. By leveraging the power of AI/ML, the project seeks to automate and enhance the process of water potability assessment, enabling timely interventions and decision-making.

Methodologies:

1. Data Collection and Preprocessing: A comprehensive dataset containing information on water quality attributes is collected and pre-processed to handle missing values, outliers, and inconsistencies. Feature engineering techniques may be applied to extract relevant features from the raw data.

2. Model Development: Various machine learning algorithms such as logistic regression, random forest, support vector machines (SVM), and neural networks are explored and trained using the pre-processed data. Hyperparameter tuning and cross-validation techniques are employed to optimize model performance.

3. Evaluation and Validation: The trained models are evaluated using appropriate evaluation metrics such as accuracy, precision, recall, and F1-score. Validation techniques such as holdout validation or cross-validation are utilized to assess the generalizability and robustness of the models.

Expected Outcomes:

1. Accurate Prediction: The developed AI/ML models are expected to accurately predict the potability of water samples based on the provided features, enabling timely identification of unsafe drinking water sources.

2. Efficiency and Automation: Automation of water potability assessment through AI/ML models can significantly reduce the time and resources required for manual inspection and testing.

3. Enhanced Public Health:By ensuring the availability of safe drinking water, the project contributes to safeguarding public health and mitigating the risks associated with waterborne diseases and contamination.

2. Introduction:

Background: In Chemical Engineering, ensuring water safety and quality is critical for various industrial processes and public health. Traditional methods for water quality assessment often lack real-time insights. Thus, there's a need for advanced solutions. Machine learning offers promising avenues for automating water potability assessment, aiding in proactive management of water resources. This aligns with Chemical Engineering's goals of optimizing processes, ensuring compliance, and safeguarding public health and the environment.

Problem Statement:

This project's main goal is to create a machine learning model that can precisely predict a water sample's potability based on its chemical and physical properties, such as:

1. pH value: When assessing the acid-base balance of water, pH is a crucial factor. It also serves as a gauge for the water's acidity or alkalinity. The WHO has advised a maximum pH range of 6.5 to 8.5. The ranges of the current experiment were 6.52–6.83, which correspond to the WHO recommended range.

2. Hardness: The main sources of hardness are the salts of calcium and magnesium. Water passes through geologic strata that contain these salts, which dissolve them. The amount of hardness in raw water is influenced by the amount of time it spends in contact with materials that produce hardness. The ability of water to precipitate soap due to the presence of calcium and magnesium was the original definition of hardness.

3. Solids (Total Dissolved Solids, or TDS): A variety of inorganic and certain organic minerals and salts, including bicarbonates, chlorides, magnesium, sodium, potassium, and sulphates, can be dissolved in water. These minerals gave the water an unpleasant flavour and a diluted hue. This is a crucial factor in how water is used. High TDS values are indicative of highly mineralized water. TDS is recommended to be consumed at a maximum of 1000 mg/l and at a desirable limit of 500 mg/l.

4. Chloramines: The two main disinfectants utilised in public water systems are chlorine and chloramine. Most frequently, ammonia is added to chlorine to treat drinking water, which results in the formation of chloramines. It is deemed safe for drinking water to have up to 4 mg/L, or 4 parts per million (ppm), of chlorine.

5. Sodium: Sulphates are organic compounds that are present in rocks, soil, and minerals. They can be found in food, plants, groundwater, and ambient air. Sulphate is mostly used in the chemical industry for commercial purposes. The amount of sulphur present in saltwater is around 2,700 mg/L. In most freshwater sources, its concentrations vary from 3 to 30 mg/L, yet in certain regions, significantly greater quantities (1000 mg/L) are observed.

6. Electrical Conductivity: Pure water is a good insulator and not a strong conductor of electrical current. Water's electrical conductivity is improved when the concentration of ions rises. Electrical conductivity in water is often determined by the concentration of dissolved

particles in the water. In actuality, electrical conductivity (EC) gauges a solution's ability to transfer electricity through its ionic process.

7.Organic Carbon: Both synthetic and naturally occurring organic matter (NOM) that has decomposed is the source of total organic carbon (TOC) in source waters. The total organic carbon content, or TOC, of pure water is measured. The US EPA states that the TOC in treated or drinking water is less than 2 mg/L, and in source water used for treatment, it is less than 4 mg/Lit.

8. Trihalomethanes (THMs): These substances can be present in water that has undergone chlorine treatment. The amount of organic matter in the water, the temperature of the treated water, and the amount of chlorine needed to treat the water all affect the concentration of THMs in drinking water. THM concentrations in drinking water up to 80 ppm are regarded as safe.

9. Turbidity: The amount of suspended solid matter in water determines how turbid it is. The test is used to determine the quality of waste discharge in relation to colloidal matter. It measures the light-emitting capabilities of water.

Objectives:

a. To comprehend the underlying patterns, distributions, and linkages within the water quality dataset, conduct thorough exploratory data analysis (EDA). Employ correlation analysis, statistical summaries, and visualisation strategies to pinpoint the main variables affecting water potability and provide guidance for the preprocessing and feature selection processes that follow.

b. Create a strong machine learning model that can reliably forecast the potability of water using the features that have been identified and the knowledge gathered from EDA. To attain optimal performance, apply suitable algorithms, like random forests, decision trees, or neural networks, and adjust model designs. Make sure the model can effectively handle the data's complexity and generalises to samples that haven't been seen before.

c. Use strict optimisation and assessment methods to improve the predicted accuracy and dependability of the model. To optimise the model's parameters and evaluate its performance on validation datasets, apply techniques like hyperparameter tuning, cross-validation, and performance metric assessment. To make sure the model works in real-world situations, keep validating it against actual water samples.

3.Methodology:

Description of data source: The primary data source is Kaggle, a widely recognized platform known for hosting datasets. Attributes include chemical and physical parameters, with a target variable indicating potability.

Dataset: https://www.kaggle.com/datasets/uom190346a/water-quality-and-potability

Dataset characteristics

1. ph: pH of 1. water (0 to 14).
2. Hardness: Capacity of water to precipitate soap in mg/L.
3. Solids: Total dissolved solids in ppm.
4. Chloramines: Amount of Chloramines in ppm.
5. Sulfate: Amount of Sulfates dissolved in mg/L.
6. Conductivity: Electrical conductivity of water in µS/cm.
7. Organic_carbon: Amount of organic carbon in ppm.
8. Trihalomethanes: Amount of Trihalomethanes in µg/L.
9. Turbidity: Measure of light emitting property of water in NTU.
10. Potability: Indicates if water is safe for human consumption. Potable - 1 and Not potable – 0

Data Preprocessing:

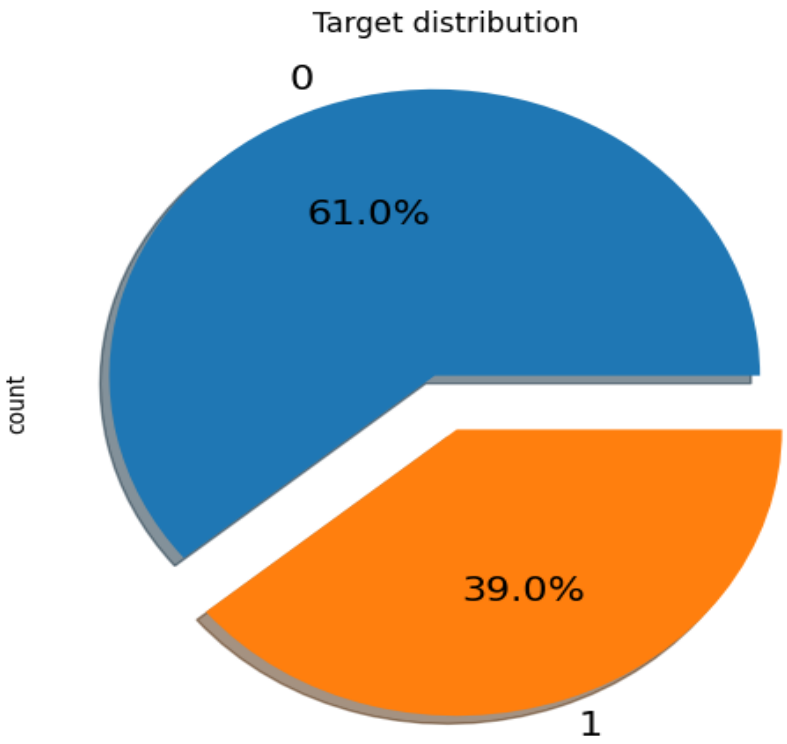
To get the data ready for analysis, a few steps could be needed:

1. Cleaning: Examine the data for errors, outliers, and missing numbers. Imputation procedures (replacement with the mean or median) can be used to handle missing numbers, and trimming or other suitable measures can be taken to deal with outliers.
2. Normalisation: To prevent some variables from dominating the analysis because of their bigger magnitude, normalise the data to ensure that all characteristics have a similar scale. A min-max scaler is employed in this instance.
3. Data Splitting: To accurately assess the model's performance, separate the dataset into test, validation, and training sets. Make that the sample distribution among various sets accurately captures the fundamental properties of the data.

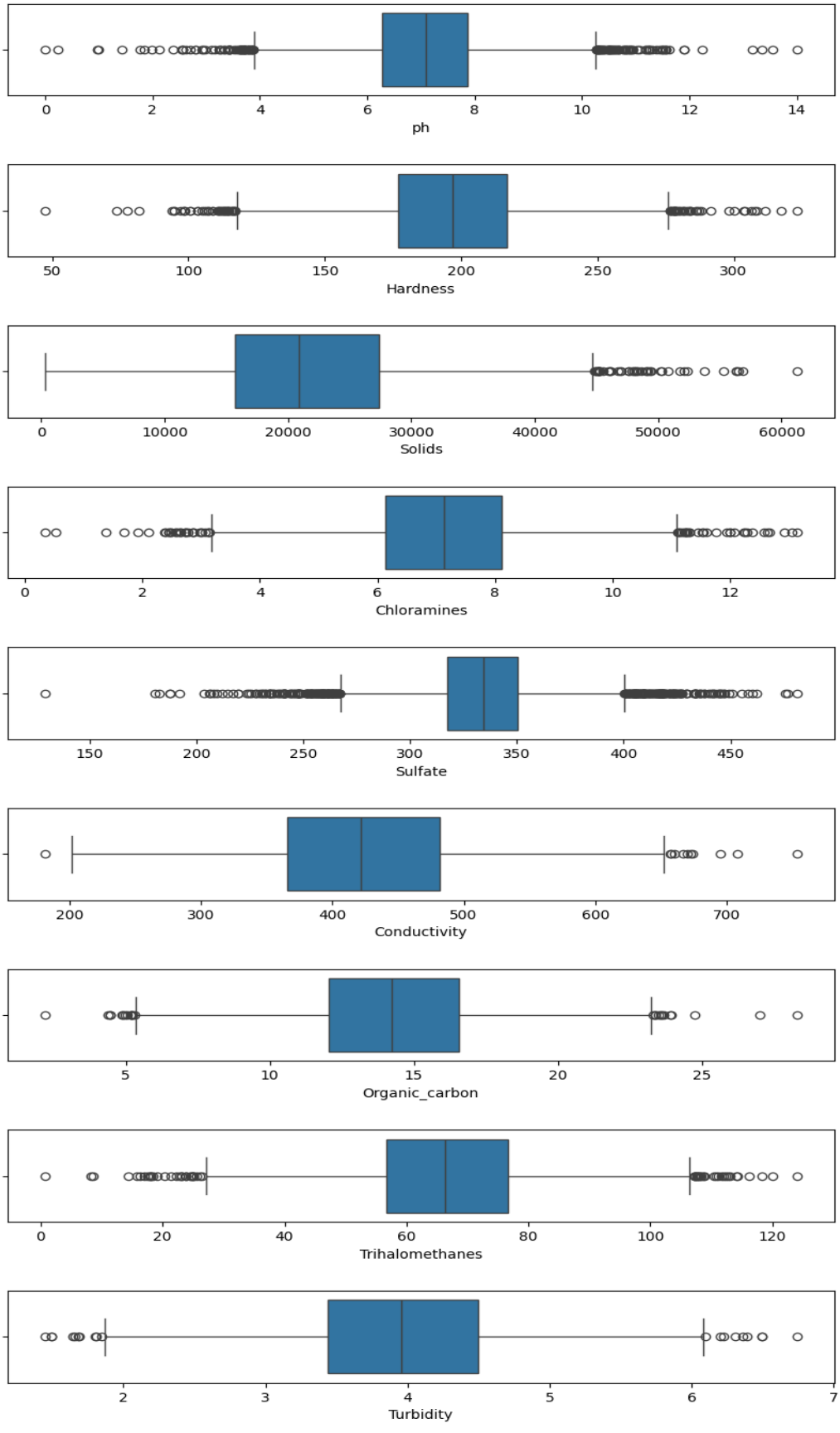
We can guarantee that the data is clean, standardised, and appropriate for analysis and modelling by carrying out these preprocessing procedures, which will ultimately increase the precision and resilience of our machine learning algorithms for predicting water potability.

Results from EDA:

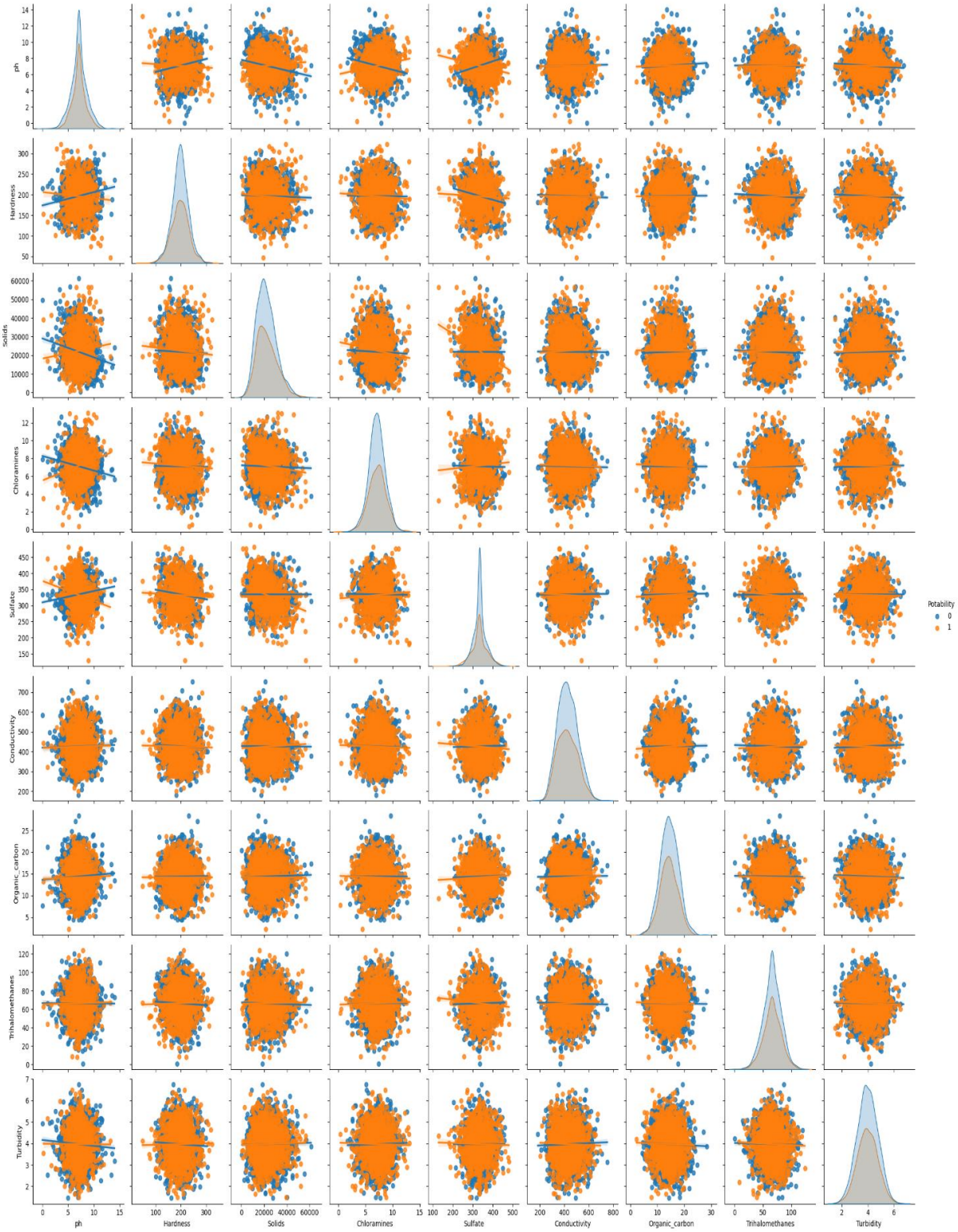
- 61 percent of data corresponds to non-potable and 31 percent corresponds to potable water.



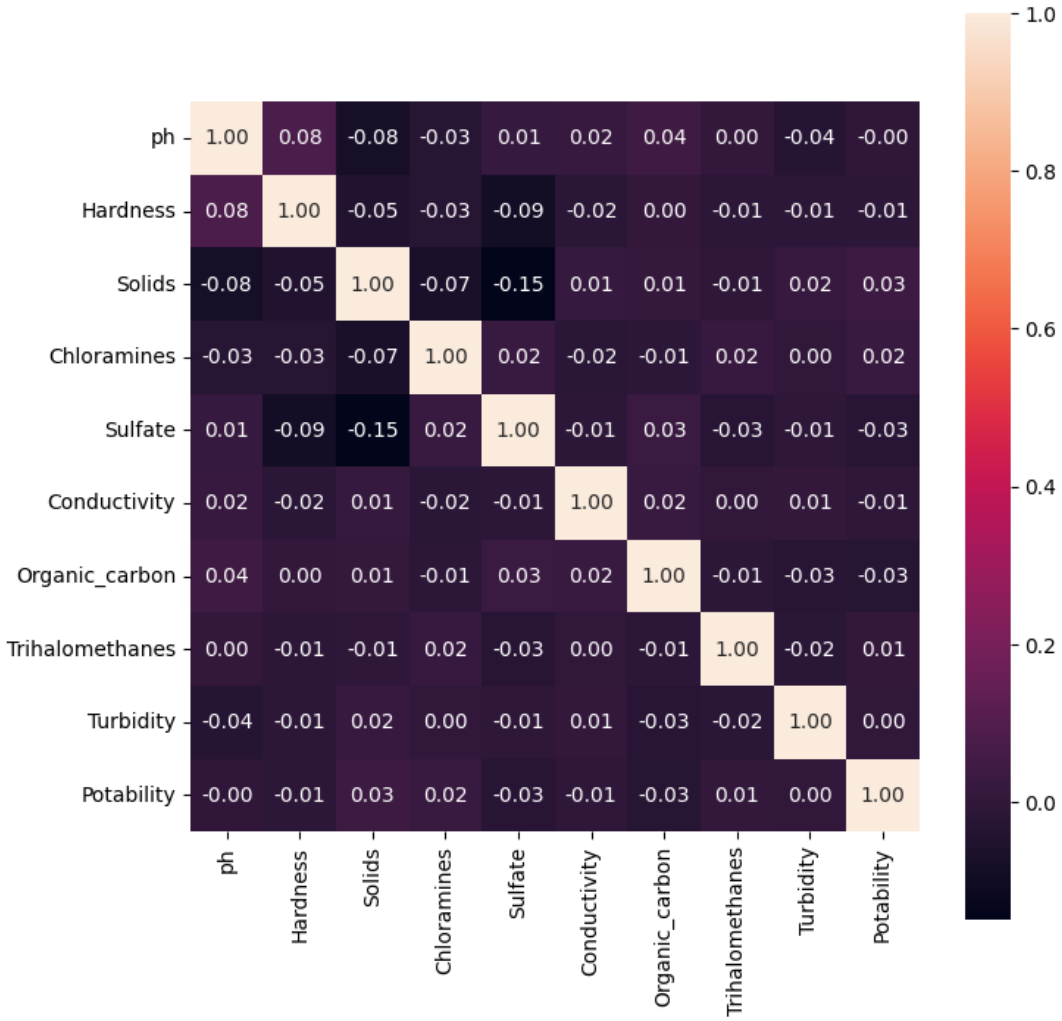
Boxplots:


0 2 4 6 8 10 12 14
ph
50 100 150 200 250 300
Hardness
0 10000 20000 30000 40000 50000 60000
Solids
0 2 4 6 8 10 12
Chloramines
150 200 250 300 350 400 450
Sulfate
200 300 400 500 600 700
Conductivity
5 10 15 20 25
Organic_carbon
0 20 40 60 80 100 120
Trihalomethanes
2 3 4 5 6 7
Turbidity

Pairplot:

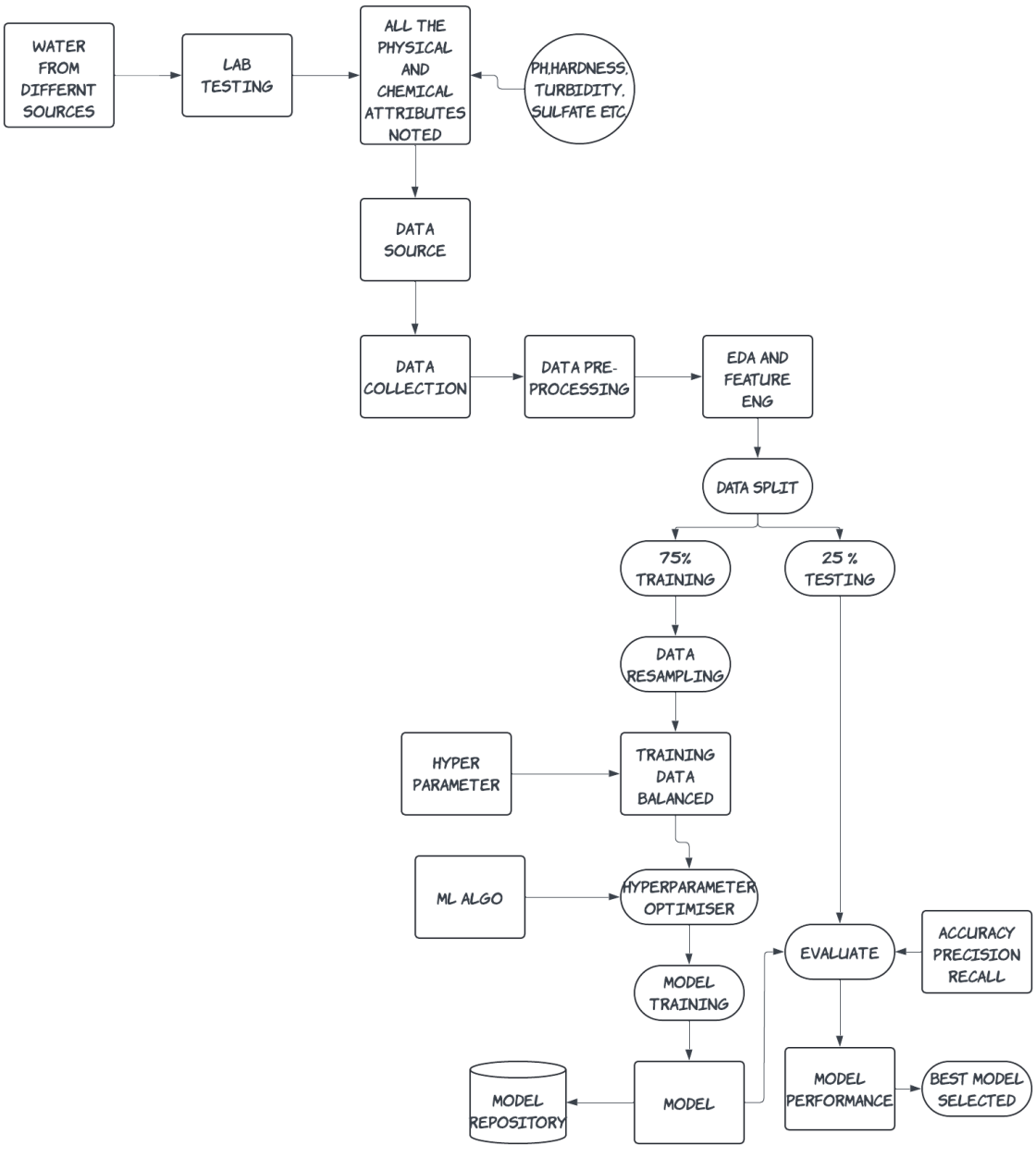

ph
Hardness
Solids
Chloramines
Sulfate
Conductivity
Organic_carbon
Trihalomethanes
Turbidity
Potability
0
1

Heatmap:



Model Architecture:

The proposed AI/ML model architecture combines supervised learning algorithms like logistic regression, random forest, and support vector machines (SVM) to predict water potability. This choice offers a versatile approach capable of handling the complex relationships in water quality data. Ensemble techniques, such as bagging and boosting, further enhance model performance. This architecture strikes a balance between accuracy and interpretability, making it well-suited for real-time applications in Chemical Engineering.



Tools and Technologies: Python with libraries like

•Pandas: Powerful for data manipulation, cleaning, and exploration.

•NumPy: Provides numerical computing foundation for data analysis.

•Matplotlib/Seaborn: Create informative visualizations of your data.

•Scikit-learn: Python library with a wide range of machine learning algorithms for classification, regression, and more.

4. Implementation plan:

Development Phases:

1. Data Collection and Preprocessing : Gather historical data on pipeline characteristics, environmental factors, and maintenance records. Clean the data, handle missing values, and encode categorical variables.
2. Feature Engineering : Extract relevant features from the dataset, such as pipeline age, material type, corrosion levels, and operational parameters. Conduct exploratory data analysis (EDA) to identify significant features.
3. Model Development: Train classification models using machine learning algorithms such as Random Forest, Support Vector Machines (SVM), Decision Tree, Naïve Bayes, . Tune hyperparameters using techniques like grid search and cross-validation.
4. Model Evaluation : Evaluating model performance using metrics such as accuracy, precision, recall, F1-score.
5. Deployment and Testing : Deploy the trained model in a production environment. Test the model's predictions on unseen data and fine-tune if necessary.

Model Training:

a. Algorithms: In the water potability prediction project, the selection of machine learning models depends on the characteristics of the data and the specific problem being addressed. Among the models under consideration are:

Certainly, here are brief descriptions of each of the 16 classification models:

1. Logistic Regression: Linear model estimating probabilities for binary classification tasks.

2. Ridge Classifier: Linear classifier with ridge regression regularization to prevent overfitting.

3. SGD Classifier: Stochastic Gradient Descent classifier suitable for large datasets.

4. Support Vector Classifier (SVC): Constructs hyperplanes to separate classes in high-dimensional spaces.

5. NuSVC: Variant of SVC allowing tuning of the 'nu' hyperparameter for support vectors.

6. Decision Tree Classifier: Non-parametric model using tree-like structures for classification.

7. Gaussian Naive Bayes: Assumes feature independence and Gaussian distribution for classification.

8. Bernoulli Naive Bayes: Assumes binary features and Bernoulli distribution.

9. Perceptron: Single-layer neural network for binary classification.

10. Nearest Centroid: Classifies samples based on nearest centroids of each class.

11. Random Forest Classifier: Ensemble method constructing multiple decision trees.

12. AdaBoost Classifier: Combines multiple weak learners to create a strong classifier.

13. XGB Classifier: Gradient boosting implementation known for speed and performance.

14. Passive Aggressive Classifier: Online learning algorithm for large-scale classification tasks.

15. Bagging Classifier: Ensemble meta-estimator that fits base classifiers on random subsets of data.

16. GradientBoostingClassifier: Boosting algorithm building trees sequentially in a gradient descent manner.

Each of these models offers different strengths and is suitable for various types of datasets and classification problems

b. Parameter tuning- Use of techniques like grid search and randomized search to optimize hyperparameters for GradientBoosting Classifier and AdaBoost Classifier.

GradientBoosting Classifier:

```
Best parameters: {'learning_rate': 0.3, 'max_depth': 7, 'n_estimators': 200}
Best score: 0.798855860263014
Accuracy on test set: 0.8048780487804879
```

AdaBoost Classifier:

```
Best parameters: {'learning_rate': 0.3, 'n_estimators': 150}
Best score: 0.7479727202094903
Accuracy on test set: 0.7560975609756098
```

c. Ensemble methods- Employ ensemble methods like bagging and boosting to improve model performance and reduce overfitting

Model Evaluation: The model's accuracy, precision, recall, F1-score, and area under the ROC curve (AUC-ROC) are measured in order to evaluate its performance. These metrics offer a thorough evaluation of the model's capacity to correctly categorise water samples and differentiate between potable and non-potable samples.

5.Testing and Deployment

1. Testing Strategy: The model will be tested against unseen data using a holdout dataset that was not used during training. This dataset will contain new instances of pipeline characteristics, environmental factors, and maintenance records. Model performance will be evaluated using metrics such as accuracy, precision, recall, and F1-score to ensure its generalization to new data and to detect any overfitting.

2. Deployment Strategy: The model will be deployed as a scalable and efficient web service, allowing real-time predictions on incoming pipeline data. It will be hosted on a cloud platform such as AWS or Google Cloud for scalability and performance. Regular monitoring and maintenance will be conducted to ensure the model's continued effectiveness and reliability in production environments.

3. Ethical Considerations: Ethical considerations in deploying the model include ensuring fairness and transparency in decision-making, avoiding biases in data and predictions, and protecting sensitive information. It's crucial to communicate the limitations and uncertainties of the model to stakeholders and to have mechanisms in place for handling potential errors or biases that may arise. Additionally, data privacy and security measures must be implemented to safeguard confidential information. Regular audits and reviews will be conducted to assess the model's impact and address any ethical concerns.

6.Results and Discussions:

These are the accuracy scores of different models

	Score
Random Forest Classifier	0.804878
Bagging Classifier	0.792683
XGB Classifier	0.780488
GradientBoostingClassifier	0.780488
Decision Tree	0.759146
Ada Boost Classifier	0.740854
Logistic Regression	0.600610
Ridge	0.600610
SGD Classifier	0.600610
Support Vector Classifier	0.600610
Gaussian NB	0.600610
Bernoulli NB	0.600610
Perc	0.600610
Passive Aggressive	0.600610
Nearest Centroid	0.512195
NuSVC	0.500000

Classification report and Confusion matrix:

The vertical axis is true label and horizontal axis is predicted in all confusion matrixes.

a.Gradient Boosting Classifier :

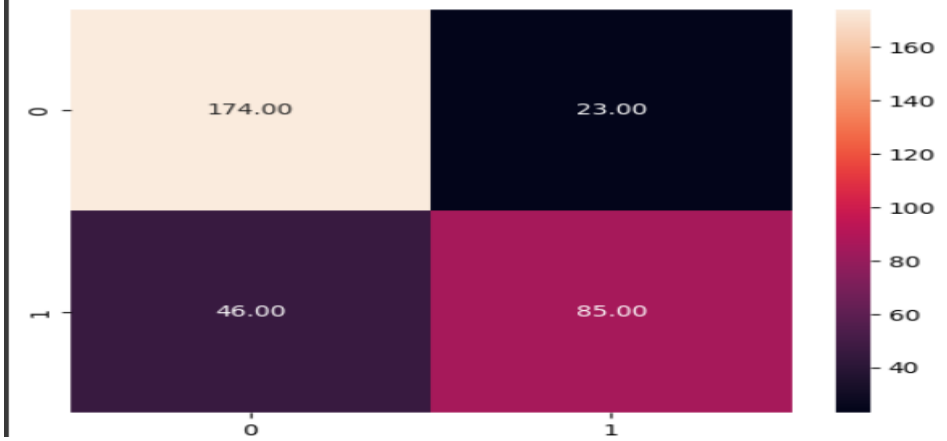
	precision	recall	f1-score	support
0	0.77	0.90	0.83	197
1	0.80	0.60	0.69	131
accuracy			0.78	328
macro avg	0.79	0.75	0.76	328
weighted avg	0.78	0.78	0.77	328

<Axes: >

b.Bagging Classifier:

	precision	recall	f1-score	support
0	0.79	0.88	0.83	197
1	0.79	0.65	0.71	131
accuracy			0.79	328
macro avg	0.79	0.77	0.77	328
weighted avg	0.79	0.79	0.79	328

<Axes: >

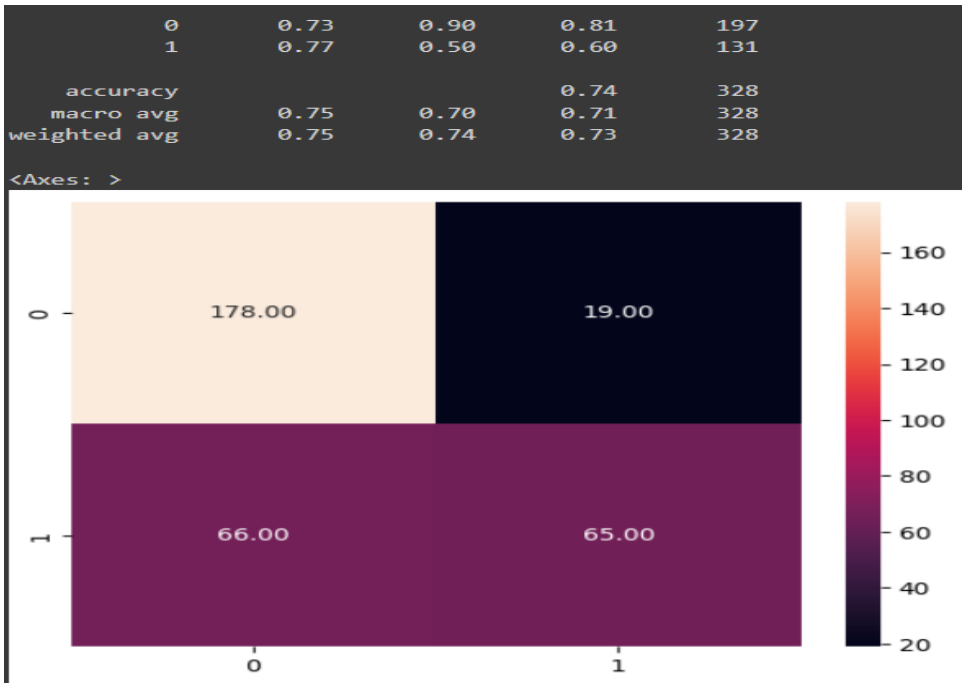


c .Random Forest Classifier:

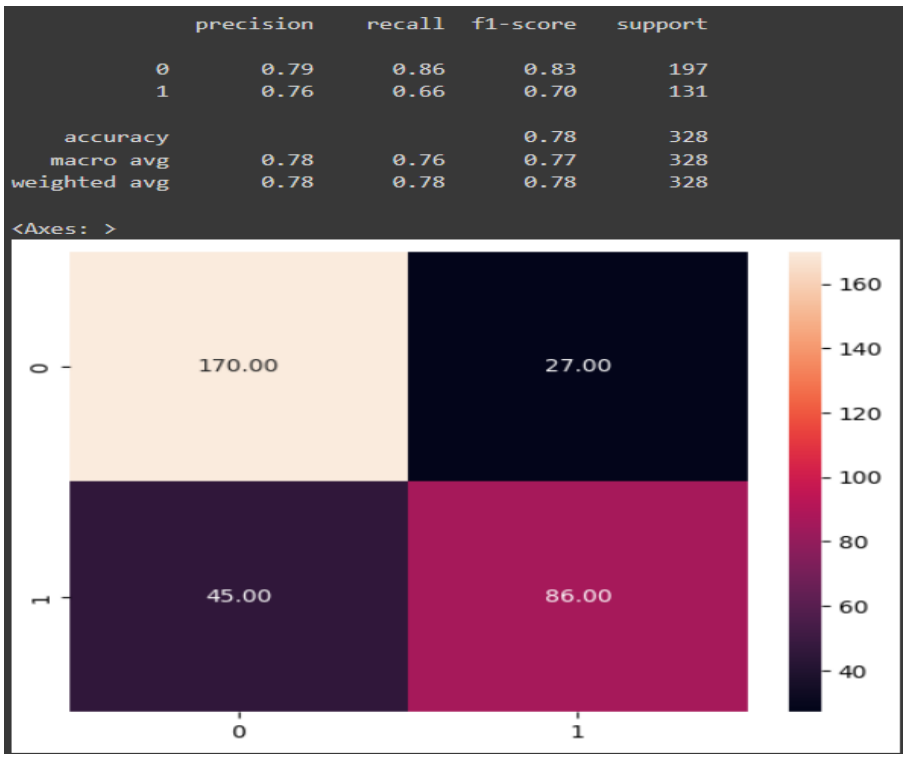
	precision	recall	f1-score	support
0	0.79	0.91	0.85	197
1	0.83	0.64	0.72	131
accuracy			0.80	328
macro avg	0.81	0.78	0.79	328
weighted avg	0.81	0.80	0.80	328

<Axes: >

d. Ada Boost Classifier:



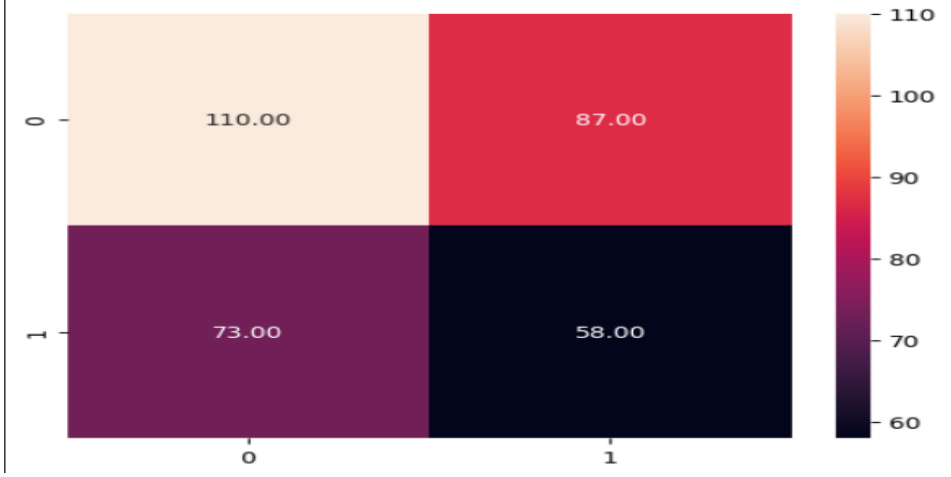
e. XGB Classifier:



f. Nearest Centroid:

	precision	recall	f1-score	support
0	0.60	0.56	0.58	197
1	0.40	0.44	0.42	131
accuracy			0.51	328
macro avg	0.50	0.50	0.50	328
weighted avg	0.52	0.51	0.52	328

<Axes: >



7. Conclusion and Future Work :

The Water Potability project successfully addressed the critical issue of assessing the safety and quality of drinking water through the development of predictive machine learning models. By leveraging various algorithms and techniques, the project achieved accurate predictions of water potability based on physicochemical and chemical properties. These models provide valuable insights into the factors influencing water quality and enable timely interventions to ensure safe drinking water supplies.

Impact:

The project's outcomes have significant implications for various stakeholders, including industries, public health agencies, and environmental organizations. By automating and enhancing water potability assessment, the developed models contribute to:

1. Ensuring Public Health: By identifying and mitigating risks associated with contaminated water sources, the models help safeguard public health and prevent waterborne diseases.

2. Operational Efficiency: Industries, such as pharmaceuticals, food and beverage production, and wastewater treatment plants, benefit from improved operational efficiency and product quality assurance.

3. Environmental Sustainability: By promoting the sustainable management of water resources, the project supports environmental conservation efforts and minimizes pollution.

Future Work: Future research directions for the Water Potability project include:

1. Integration of Additional Data Sources: Incorporating additional data sources, such as meteorological data, geographical information, and historical water quality records, could further enhance model performance and predictive accuracy.

2. Development of Real-time Monitoring Systems: Implementing real-time monitoring systems using sensor networks and IoT devices would enable continuous monitoring of water quality parameters and early detection of contamination incidents.

3. Exploration of Advanced Machine Learning Techniques: Investigating advanced machine learning techniques, such as deep learning and ensemble methods, could lead to more sophisticated models capable of capturing complex relationships in water quality data.

4. Deployment in Resource-limited Settings: Adapting the developed models for deployment in resource-limited settings, such as rural areas or developing countries, would extend their accessibility and impact on global water quality management efforts.

In conclusion, the Water Potability project represents a significant step towards leveraging AI/ML technologies for ensuring safe and sustainable water resources. By addressing the challenges of water quality assessment, the project contributes to the advancement of public health, environmental sustainability, and industrial operations.

8.References:

https://www.researchgate.net/publication/365495813_Water_Potability_Analysis_and_Prediction

[1] Roberto F., et al. Evaluation of a GFP reporter gene construct for environmental arsenic detection.
Talanta, 2002, 58(1): 181-188.
[2] Erdogan O., et al. Critical evaluation of wastewater treatment and disposal strategies for Istanbul
with regards to water quality monitoring study results. ELSEVISE, 2008, 226: 231-248.
[3] Lourenco N.D., et al. UV spectra analysis for water quality monitoring in a fuel park wastewater
treatment plant. Chemosphere, 2006, 65: 786-791.
[4] Kim B. C. Multi-channel continuous water toxicity monitoring system: its evaluation and

application to water dis-charged from a power plant. Environmental Monitoring and Assessment,2005, 109(3): 156-164.
[5] ISO/TC 147 /SC5. ISO 11348-1 2007(E)Water quality – Determination of the inhibitory effect of water samples on the light emission of Vibrio fischeri (Luminescent bacteria test) - Part 1: Method using freshly prepared bacteria (ISO 11348-1). Geneva, Switzerland: ISO, 2007

9.Appendices:

Any supplementary material, including code snippets, detailed data analysis, or additional plots and graphs.

10.Auxiliaries

Web link: (if deployed as live website give website link)

Data Source:
https://drive.google.com/file/d/1zPibZ2au0Vi8N8uodxcfLTOkDZkUZvAa/view?usp=sharing

Python File : https://colab.research.google.com/drive/18HV9KL-pk3P9PpHDmeQq-50KG5JX4tBw?usp=sharing