Water Quality Index Prediction using Machine-Learning

A Main Project submitted
in partial fulfilment of the requirements
for the award of the degree of

BACHELOR OF TECHNOLOGY
In
COMPUTER SCIENCE AND ENGINEERING

Submitted by

Under the esteemed guidance
of CH. Lakshmi
Veenadhari Assistant
Professor

DEPARTMENT OF COMPUTER SCIENCE AND
ENGINEERING VISHNU INSTITUTE OF TECHNOLOGY
(Autonomous)
(Approved by AICTE, Accredited by NBA & NAAC and permanently affiliated to JNTU Kakinada)
BHIMAVARAM – 534 202
2023 – 2024

VISHNU INSTITUTE OF TECHNOLOGY

(Autonomous)

(Approved by AICTE, accredited by NBA & NAAC, and permanently affiliated to JNTU Kakinada)

BHIMAVARAM-534202

2023-2024

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

CERTIFICATE

This is to certify that the project entitled "EVENT MANAGEMENT SYSTEM", is being submitted by **M.SATYA KALYANI,Y.MARY SUJATHA,V.SWATHI AND K.KIRAN KUMAR,** bearing the **REGD.NOS: 15PA1A0533, 16PA5A0506, 15PA1A0558 and 15PA1A0532** submitted in fulfilment for the award of the degree of **"BACHELOR OF TECHNOLOGY"** in **"COMPUTER SCIENCE AND ENGINEERING"** is a record of work carried out by them under my guidance and supervision during the academic year 2023-2024 and it has been found worthy of acceptance according to the requirements of university.

Internal Guide

CH. Lakshmi Veenadhari

Head of the Department

Dr. Sumit Gupta

External Examiner

ACKNOWLEDGEMENT

It is natural and inevitable that the thoughts and ideas of other people tend to drift into the subconscious due to various human parameters, where one feels the need to acknowledge the help and guidance derived from others. We express our gratitude to those who have contributed to the fulfillment of this project.

We take the opportunity to extend our sincere thanks to **Dr. D. Suryanaryana**, the director of VIT, Bhimavaram, whose guidance from time to time helped us complete this project successfully.

We also express our sincere gratitude to **Dr. M. Venu**, the principal of VIT, Bhimavaram, for his continuous support in helping us complete the project on time.

We are thankful to **Dr. Sumit Gupta**, Head of the Department of Computer Science and Engineering, for his continuous and unwavering support and guidance. We acknowledge our gratitude for his valuable guidance and support extended to us from the conception of the idea to the completion of this project.

We are very thankful to **CH. Lakshmi Veenadhari**, Assistant Professor, our internal guide, whose guidance from time to time helped us complete this project successfully.

Project Associates

Abstract

Water quality is very dominant for humans, animals, plants, industries, and the environment. In the last decades, the quality of water has been impacted by contamination and pollution. In this paper, the challenge is to anticipate Water Quality Index (WQI) and Water Quality Classification (WQC), such that WQI is a vital indicator for water validity. In this study, parameters optimization and tuning are utilized to improve the accuracy of several machine learning models, where the machine learning techniques are utilized for the process of predicting WQI and WQC. Grid search is a vital method used for optimizing and tuning the parameters for four classification models and also, for optimizing and tuning the parameters for four regression models.

Random forest (RF) model, Extreme Gradient Boosting (Xgboost) model, and Adaptive Boosting (AdaBoost) model are used as classification models for predicting WQC. K-nearest neighbor (KNN) regressor model, decision tree (DT) regressor model,, and Logistic regression model are used as regression models for predicting WQI.

In addition, preprocessing steps including, data imputation (mean imputation) and data normalization were performed to fit the data and make it convenient for any further processing. The dataset used in this study includes 10 features and 3277 instances. To examine the efficacy of the classification approaches, five assessment metrics were computed: accuracy, recall, precision, support, and F1 score. In terms of classification, the testing findings showed that the XGBoost model produced the best results, with an accuracy of 88% when predicting WQC values.

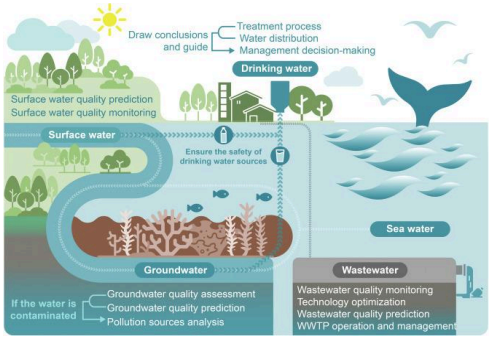
Keywords: **Python, Flask, Machine Learning , XGBoost, HTML5, CSS3**

Contents

Chapter 1 Introduction

Water is among the most precious resources on which all existence is dependent. Water contamination degrades water quality, impacting the health of sea creatures and, by extension, humans that use them. This makes it critical to observe water quality and ensure the survival of nautical life [1]. Comprehension of water quality concerns and issues is also crucial for water pollution mitigation and control. To grasp the condition of the nautical ecosystem, several governments throughout the world have begun to build ecological water management programs. Roughly one billion individuals do not have access to clean water for drinking, and two million individuals perish every year as a consequence of polluted water and poor sanitation and cleanliness. As a result, preserving the freshwater quality is critical [2]. Water quality is critical to the long-term viability of a diversion plan. The water of poor quality may also be costly since resources must be shifted to repair water delivery infrastructure whenever an issue emerges. The demand for enhanced water management and water quality control has been rising for these objectives to assure safe drinking water at reasonable costs. To address these issues, systematic assessments of freshwater, disposal systems, and organizational monitoring issues are necessary [3]. Forecasting water quality entails anticipating fluctuation characteristics in a water system's health at a specific moment. Assessment of water quality is critical for water quality planning and regulation. Water pollution avoidance and regulation methods may be improved by forecasting future updates in water cleanliness at varying degrees of pollution and designing reasonable water pollution prevention and control techniques. The overall consistency of water should be assessed in water diversion plans. To handle everyday drinking difficulties, a considerable quantity of water is carried. Thus, in today's civilization, solutions for anticipating water quality should be researched [4]. The use of artificial intelligence (AI) and machine learning (ML) technologies is currently

critical to security threats [5] and focus on mapping the connection between system inputs and outcomes rather than complex operations strategies [6].



Water quality forecasting is an essential method for water planning, regulation, and monitoring; it is a necessary component of water contamination research to investigate water ecological protection. As a consequence, it is crucial to enhance a realistic and practical strategy for predicting water quality. Simultaneously, forecasting future water quality is necessary for preventing sudden updates in water quality and offering solutions. As a result, precise forecasts of water quality updates may not only assure the health of an individual's potable water but can also help guide fishing productivity and safeguard biodiversity [7]. Furthermore, the typical water quality forecast technique cannot account for the effects of biology, physics, hydraulics, alchemy, and meteorology. At the moment, researchers are primarily concerned with enhancing the practicability and trustworthiness of groundwater forecasting techniques and have presented a range of new techniques, such as artificial neural networks (ANN), stochastic mathematics, fuzzy mathematics, 3S technology, and others, for enhancing water quality forecasting techniques and expand the range of applications [8].

The Water Quality Index (WQI) is a well recognised indicator that gives a thorough assessment of water quality based on various parameters. It gives a quantitative metric that reduces the complicated nature of water quality into a single number, allowing for easy interpretation and comparison across multiple sites and time periods. WQI considers a variety of physical, chemical, and biological characteristics such as pH, dissolved oxygen, turbidity, nutrient levels, hardness, solids, conductivity and the presence of pollutants. WQI gives a thorough evaluation of water quality by aggregating these factors, which supports decision-making processes linked to water resource management. Water quality grading (WQC) is an additional feature that categorizes water samples into specified quality classes based on predefined thresholds. This categorization gives a realistic framework for determining the amount of pollution in water, allowing for targeted actions and regulatory measures. Stakeholders can identify locations or causes of concern, prioritize remediation activities, and adopt necessary actions to safeguard water resources by grading water quality. The study was motivated by the urgent need to address water quality degradation and its effects. Water pollution and contamination pose serious dangers to ecosystems, public health, and long-term development. Water quality monitoring and assessment are essential steps in recognising possible concerns, adopting effective management plans, and maintaining the supply of clean and safe water for

diverse sectors. Traditional techniques of water quality evaluation, which include laboratory analysis and WQI computation utilizing measurable parameters, can be time consuming, costly, and restricted in their capacity to offer real-time information. Predictive modeling provides an alternate method by estimating WQI and WQC based on existing data using machine learning techniques. Water quality may be assessed in a timely way by constructing accurate and effective prediction models, even when direct measurement of all parameters is not possible or practicable.

Machine learning algorithms are used to predict water quality index (WQI). Grid search is a vital method used for optimizing and tuning the parameters for four classification models, namely the random forest (RF) model, extreme gradient boosting (XGBoost) model, gradient boosting (GB) model, and adaptive boosting (AdaBoost) for predicting WQC, and four regression models, namely K-nearest neighbor (KNN) regressor model, decision tree (DT) regressor model, support vector machine (SVM) model, Logistic regression model for predicting WQI. This project's contributions are as follows:

- Data preprocessing is applied, including data imputation (mean imputation), and data cleaning was performed to fit the data and make it convenient for any further processing.
- grid search is used for optimizing and tuning the parameters for four regression models to predict WQI.
- To assess the performance of the classification techniques, **Support, accuracy, recall, precision**, and **F1 score** were computed.

Chapter 2 System Analysis

2.1 Hardware and Software requirements

Hardware Configuration:

- ➔ **Processor**: Intel Core i3
- ➔ **Hard Disk**: 160GB
- ➔ **RAM**: 8GB

Software Configuration:

- ➔ **Operating System**: Windows 7, Windows 8, or Windows 10
- ➔ **Integrated Development Environment (IDE)**: Jupyter Notebook
- ➔ **Libraries Used**: Numpy, Pandas
- ➔ **Technology**: Python 3.6 and above

2.2 Existing system and its disadvantages

Artificial Neural Networks (ANN), Support Vector Regressions (SVR), Grey Systems (GS), Regression Analyses (RA), and other approaches are commonly used to estimate water quality. Liu et al. predicted the Yangtze River Basin's drinking water quality utilizing a long short-term memory (LSTM) network. Dissolved oxygen (DO), pH, chemical oxygen demand (COD), and NH3-N were used to construct the LSTM algorithm. The LSTM technique has proved potential for surveillance water quality. **Disadvantage**: ANN models can suffer from overfitting, especially with complex datasets, and may require significant computational resources for training and inference.

Sakshi Khullar and Nanhey Singh presented a Bi-LSTM model based on deep learning (DLBCL-WQA) to anticipate the water quality variables of the Yamuna River in India. A comparison showed that the suggested approach surpassed all other approaches in terms of error rates and prediction accuracy. Sani Abba et al. examined four machine learning techniques Neuro-Fuzzy Inference (ANFIS), Backpropagation (BPNN), Multilayer Perceptron (MLP), and Support Vector Regressor (SVR) for anticipating the water quality index (WQI). The acquired findings demonstrated the viability of the built smart techniques for forecasting the WQI at the three stations using the neural network ensemble's better modeling outcomes (NNE). The predictive comparison indicated that NNE was successful and hence may be used as a

trustworthy prediction strategy. **Disadvantage:** SVR models may struggle with datasets containing high noise or outliers, requiring careful preprocessing and parameter tuning.

Yafra Khan and Chai Soo See , in their paper, have used Artificial Neural Network and time series analysis to design a water quality prediction model. Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Regression Analysis have been used as a part of evaluating the model performance. Dao Nguyen Khoi et al. , in their paper, have used 12 machine learning models to estimate the quality of water. Model evaluation was done by using 2 statistics, R2 and RMSE. Umair Ahmed et al. have used supervised machine learning algorithms to estimate the Water Quality Index (WQI). Saber Kouadri et al. used 8 artificial intelligence algorithms to generate Water quality Index prediction. Evaluation of models was done using several statistical metrics, which includes correlation coefficient (R), mean absolute error (MAE), root mean square error (RMSE), relative absolute error (RAE), and root relative square error (RRSE). Jitha Nair and Vijaya M S used various prediction models developed using machine learning and big data techniques using sensor networks. **Disadvantage:** Relying solely on statistical metrics like MSE, RMSE, R2, etc., for model evaluation may overlook important factors such as model interpretability and generalization ability, potentially limiting its applicability in real-world scenarios.

In Yang Y & Xiong Q work, a performance assessment of application of three different machine learning techniques, including deep neural networks (DNN), gradient boosting machines (GBM), and extreme gradient boosting (XGBoost), was carried out in order to evaluate the groundwater indices. In the predictions of EWQI and WQI utilizing these three models, it was found that the parameter with the greatest significance was electrical conductivity (EC), while the parameter with the least significance was pH. The model that is presented in reference number takes into account the following nine characteristics about the quality of the water: temperature, pH value, electrical conductivity, oxygen saturation, biological oxygen demand, suspended particles, nitrogen oxides, orthophosphates, and ammonium. Using data spanning the years 2013–2019 from five different places in the Vojvodina Province of Serbia, it is constructed in the Netica programme and then tested and confirmed using this data. In light of this, we can confidently propose it as a reliable instrument throughout the transition from analogue to digital water management. **Disadvantage:** Reliance on historical data from 2013 to 2019 may limit the model's ability to capture recent changes or emerging trends in water quality, potentially reducing its accuracy and adaptability over time.

2.3 Proposed system

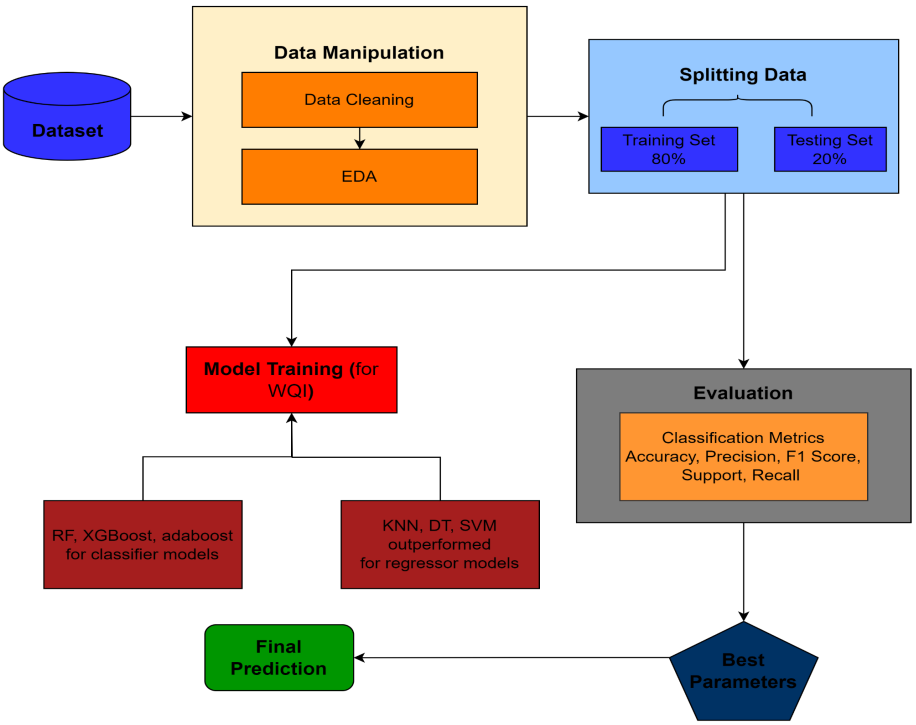
Our proposed methodology incorporates techniques for enhancing model accuracy. In our machine learning workflow, we not only train classifiers such as Random Forest, Decision Tree, XGBoost, and AdaBoost using the Kolleru water attributes dataset but also meticulously optimize their performance through hyperparameter tuning. Leveraging both Grid Search and Random Search methodologies, we meticulously explore the hyperparameter space to identify the most effective combinations. By systematically evaluating various hyperparameter configurations, we aim to maximize the accuracy of our models. This strategic approach elevates the reliability and effectiveness of our water potability classification system, ensuring robust predictions for end-users. Additionally, we seamlessly integrate this optimized machine learning pipeline with a user-friendly interface using Flask, facilitating intuitive interaction and decision-making regarding water safety.

Advantages :

Enhanced Accuracy: Hyperparameter tuning and advanced classifiers improve model accuracy, ensuring reliable predictions.

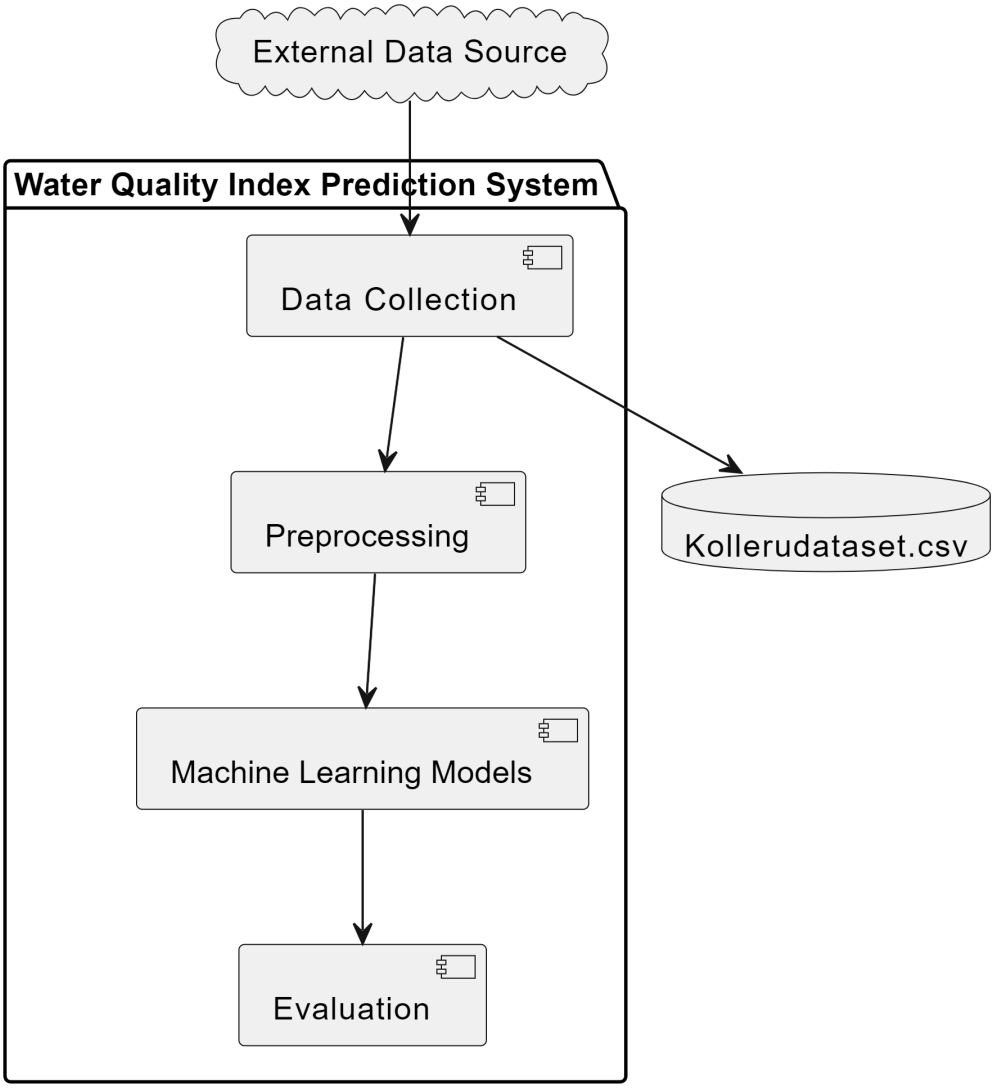
Flexibility: Versatile classifiers and optimization techniques make the method adaptable to various datasets and scenarios.

User-Friendly Interface: Integration with a simple interface enhances accessibility and usability, empowering users to make informed decisions.



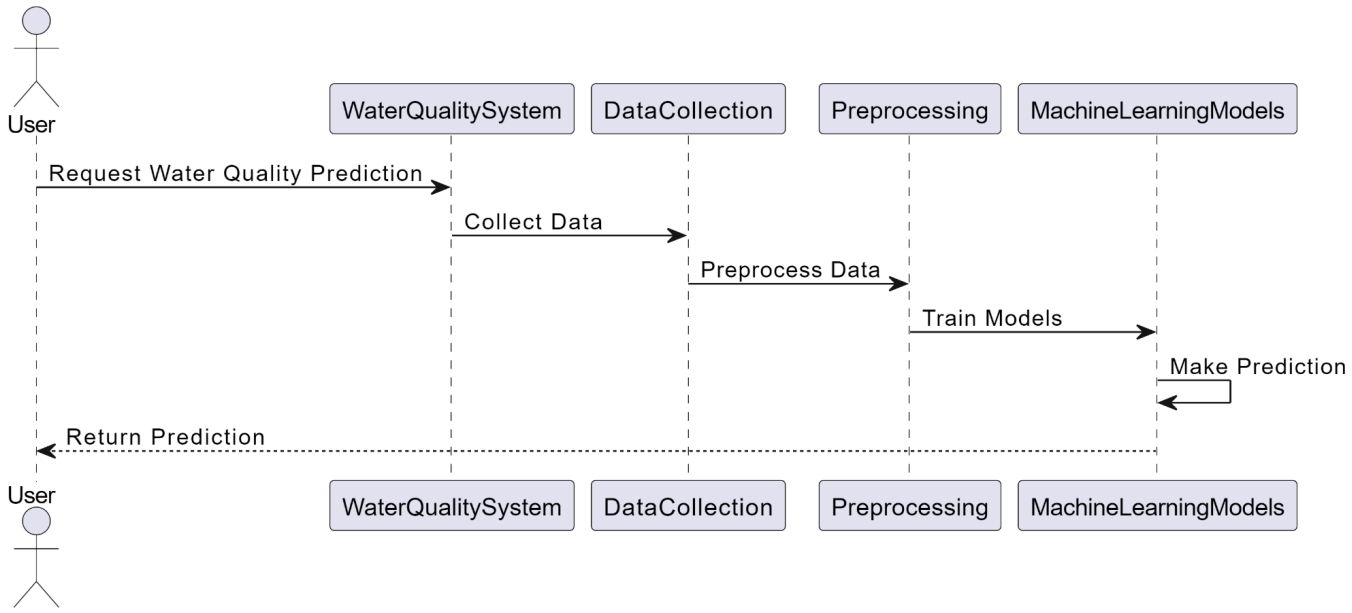
Chapter 3 System Design

3.1 Architecture & Process Flow

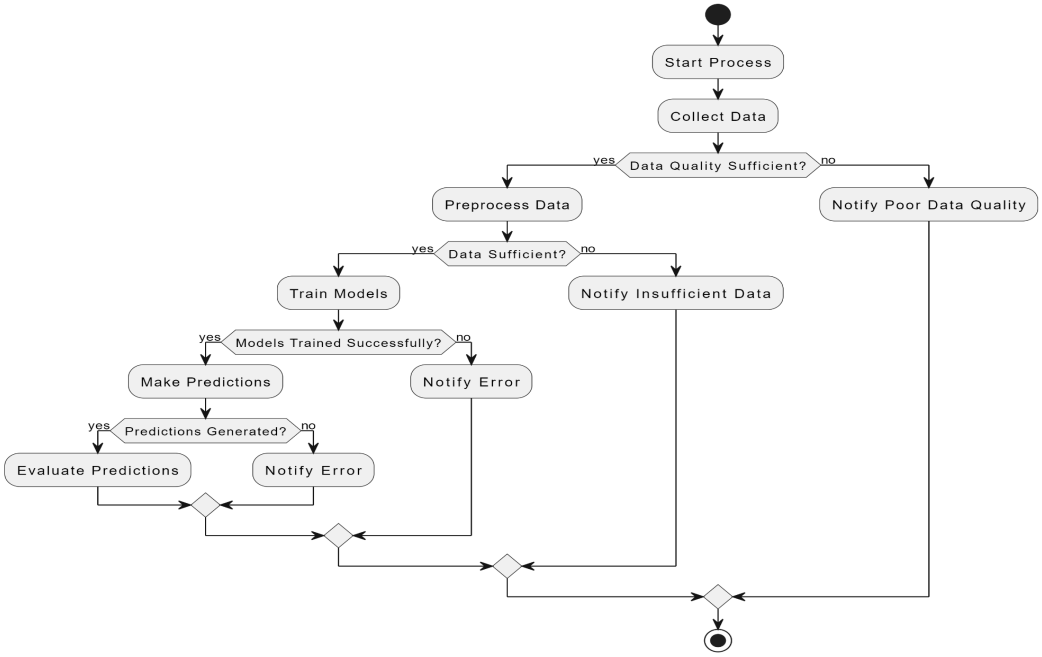


UML Diagrams

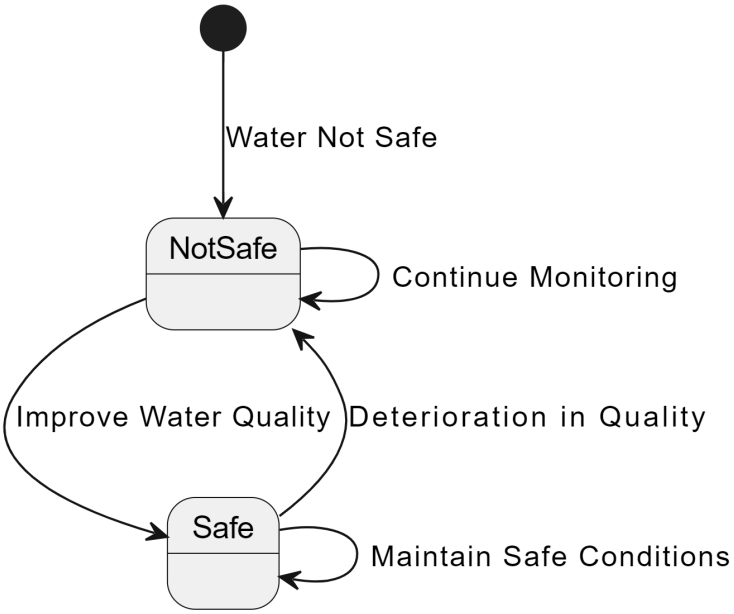
3.2 Sequence Diagram



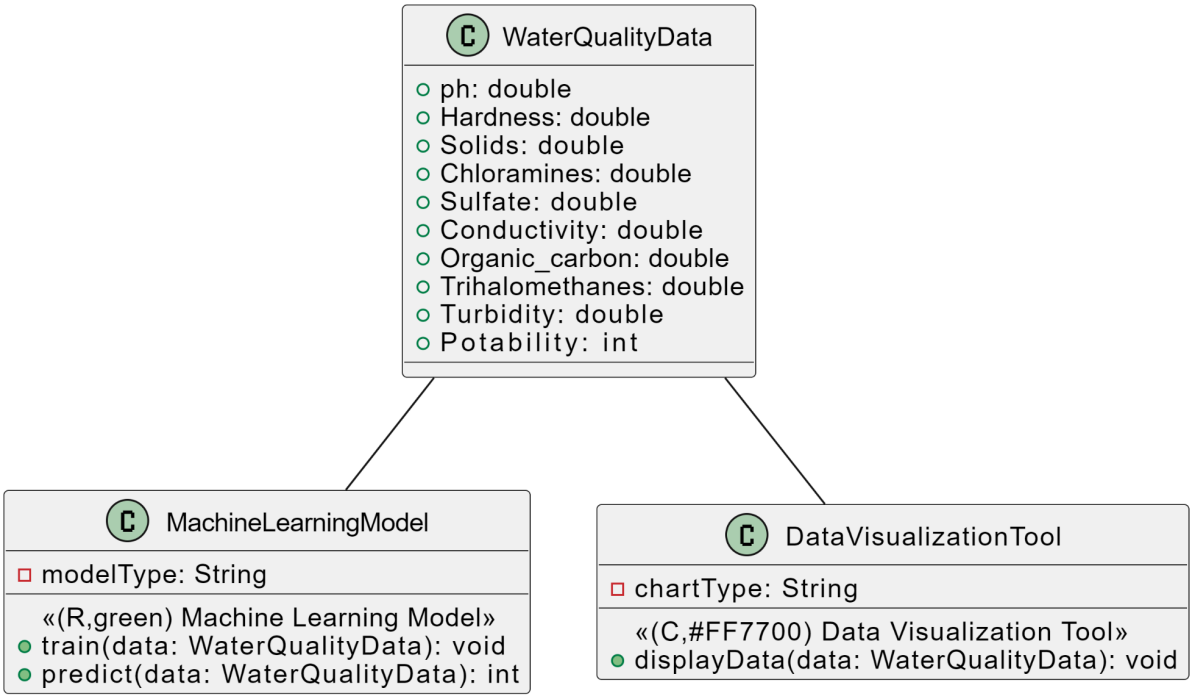
3.3 Activity Diagram



3.4 State Diagram (for potability)



3.5 Class Diagram

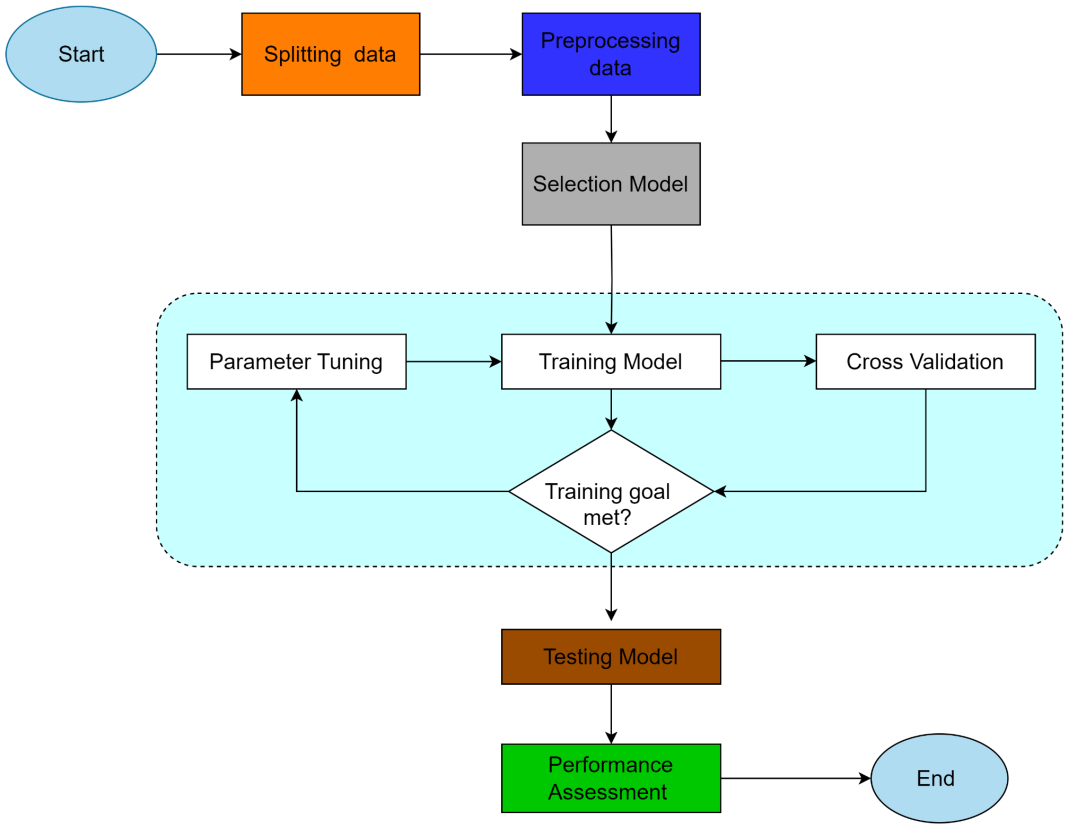


Chapter 4 Module descriptions

4.1 Materials & Methods

Following the primary data preprocessing, a particular ML approach is chosen to be trained and verified using the training and validation sets. Before being tested, the corresponding hyper variables will be fine-tuned until the predetermined training target is satisfied. The test dataset will eventually be applied to evaluate the trained approach and assess its enhancement. For clarity, the ML modeling flow chart is given in Fig. 1. The general block diagram of ML models begins with data splitting and preprocessing, followed by model selection. The selected model then undergoes training, testing, and validation. Cross-validation is used to evaluate whether the training model has met its goals. If so, the model can proceed to testing and performance assessment. If not, the model parameters need further fine-tuning during training.

FIG 1.



4.2 DataSet

The dataset that was utilized for this research was taken from Kollerudataset.csv.. The dataset comprises various features that affect the environment of Lake Kolleru. The dataset contains 10 features/ samples and more than 3200+ instances.

About dataset Content The water_potability.csv file contains water quality metrics for 3276 different water bodies.

1. **pH value:** PH is an important parameter in evaluating the acid–base balance of water. It is also the indicator of acidic or alkaline condition of water status. WHO has recommended a maximum permissible limit of pH from 6.5 to 8.5. The current investigation ranges were 6.52–6.83 which are in the range of WHO standards.

2. **Hardness:** Hardness is mainly caused by calcium and magnesium salts. These salts are dissolved from geologic deposits through which water travels. The length of time water is in contact with hardness-producing material helps determine how much hardness there is in raw

water. Hardness was originally defined as the capacity of water to precipitate soap caused by Calcium and Magnesium.

3. **Solids (Total dissolved solids - TDS):** Water can dissolve a wide range of inorganic and organic minerals or salts such as potassium, calcium, sodium, bicarbonates, chlorides, magnesium, sulfates, etc. These minerals produced unwanted taste and diluted color in the appearance of water. This is an important parameter for the use of water. The water with a high TDS value indicates that water is highly mineralized. The desirable limit for TDS is 500 mg/l and the maximum limit is 1000 mg/l which is prescribed for drinking purposes.

4. **Chloramines**: Chlorine and chloramine are the major disinfectants used in public water systems. Chloramines are most commonly formed when ammonia is added to chlorine to treat drinking water. Chlorine levels up to 4 milligrams per liter (mg/L or 4 parts per million (ppm)) are considered safe in drinking water.

5. **Sulfate:** Sulfates are naturally occurring substances that are found in minerals, soil, and rocks. They are present in ambient air, groundwater, plants, and food. The principal commercial use of sulfate is in the chemical industry. Sulfate concentration in seawater is about 2,700 milligrams per liter (mg/L). It ranges from 3 to 30 mg/L in most freshwater supplies, although much higher concentrations (1000 mg/L) are found in some geographic locations.

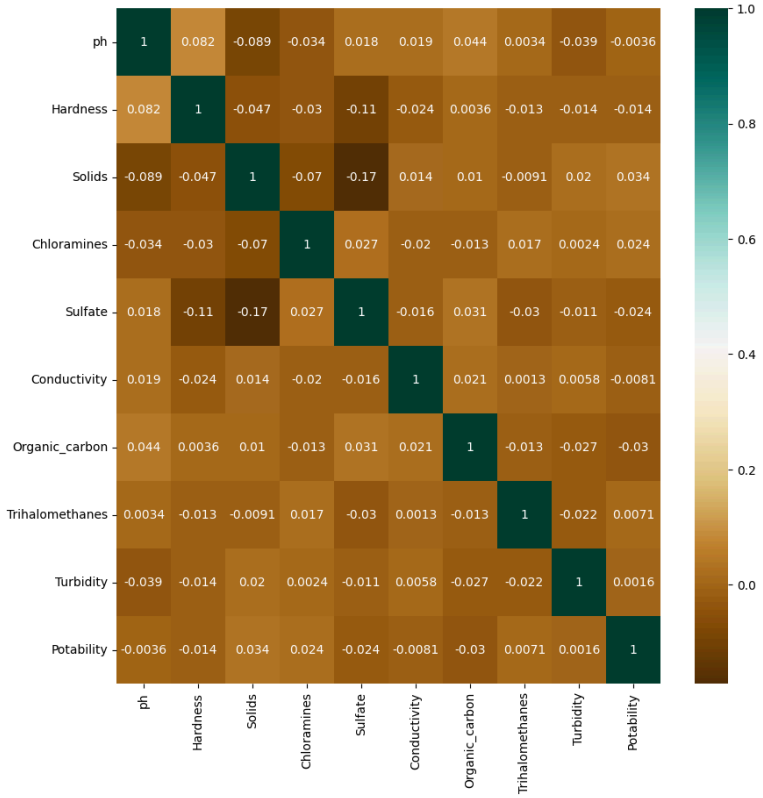
6. **Conductivity:** Pure water is not a good conductor of electric current rather's a good insulator. An increase in ion concentration enhances the electrical conductivity of water. Generally, the amount of dissolved solids in water determines the electrical conductivity. Electrical conductivity (EC) measures the ionic process of a solution that enables it to transmit current. According to WHO standards, EC value should not exceed 400 μS/cm.

7. **Organic_carbon:** Total Organic Carbon (TOC) in source waters comes from decaying natural organic matter (NOM) as well as synthetic sources. TOC is a measure of the total amount of carbon in organic compounds in pure water. According to US EPA < 2 mg/L as TOC in treated / drinking water, and < 4 mg/Lit in source water which is used for treatment.

8. **Trihalomethanes:** THMs are chemicals that may be found in water treated with chlorine. The concentration of THMs in drinking water varies according to the level of organic material in the water, the amount of chlorine required to treat the water, and the temperature of the water that is being treated. THM levels up to 80 ppm are considered safe in drinking water.

9. **Turbidity:** The turbidity of water depends on the quantity of solid matter present in the suspended state. It is a measure of light emitting properties of water and the test is used to indicate the quality of waste discharge concerning colloidal matter. The mean turbidity value obtained for Wondo Genet Campus (0.98 NTU) is lower than the WHO recommended value of 5.00 NTU.

10. **Potability:** Indicates if water is safe for human consumption where 1 means Potable and 0 means Not potable.



4.3 Data Loading and Preprocessing

Initially, the dataset (Kollerudataset.csv) containing water quality parameters is loaded into a Pandas DataFrame. This dataset undergoes preprocessing steps, such as removing duplicates and handling missing values, ensuring data cleanliness and integrity.

Data Loading: Read the dataset file into a DataFrame, ensuring all relevant information is captured for subsequent analysis.

Data Cleaning: Identify and remove duplicate records to maintain data integrity. Handle missing values appropriately, either by imputation (filling missing values with mean, median, or mode) or removal.

Data Visualization: Generate histograms to visualize the distribution of each feature, aiding in understanding their spread and identifying potential outliers. Create a heatmap to visualize the correlation between features, helping in feature selection and understanding feature importance.

Feature Engineering: Fill missing values with mean values of corresponding features to retain data completeness. Perform feature scaling if necessary to standardize the range of features, ensuring no single feature dominates the model training process.

4.4 Model Building

Various machine learning models including Logistic Regression, K-Nearest Neighbors, Decision Tree, Random Forest, AdaBoost, Bagging, and XGBoost are constructed. These models are trained and optimized using techniques like GridSearchCV or RandomizedSearchCV for hyperparameter tuning, enhancing their predictive accuracy.

4.4.1 Random Forest (RF) - with accuracy of 88%

RF method is an ensemble technique used for categorization. It is a supervised machine learning method composed of numerous decision trees. Because it is an ensemble technique, it uses the best outcome given by the many decision trees, mitigating and limiting generalization mistakes as the volume of the tree architecture in the forest grows. The classification and regression tree (CART) algorithm is used by the decision tree to categorize the tuples depending on the target parameter. This approach is applied in conjunction with bagging for resampling goals, updating the training data as new tree forms. Based on the parameters and equations listed below, a tree structure is built to categorize the features. The Gini Index may be used to create the decision tree for any tuple S and is determined using the formula:

$$G_i^n(y,s) = 1 - \left(\sum_{c_j \epsilon \sigma_y} c_j \cdot \frac{|\sigma_y = c_j \cdot S|}{|S|} \right)$$

The entropy and information gain are also important when creating a decision tree and determining its outcome. It may be computed using the following formulas:

$$\text{"Entropy}(S) = \sum_i -p(i)\log_2 p(i)\text{"}$$

where p is the fraction of S that belongs to class ‘i’, for each given set S.

$$Gain(S, A) = Entropy(S) - \sum \frac{|S_v|}{|S|} Entropy(S_v)$$

where Sv denotes the subset of S for which parameter A has value v.

RF presents numerous benefits. It avoids the issue of multivariate collinearity, which is a disadvantage of ordinary regression analysis. It excels in regression and classification and has a solid grasp of multi-dimensional data**.**

4.4.2 Extreme Gradient Boosting (XGBoost) - with accuracy of 88%

The XGBoost is a decision tree enhancement approach that is distinct from the classic gradient boosting decision tree methodology. Based on the optimization issue, the standard GBDT solely employs first-order derivative information. The loss function is then subjected to the second Taylor extension, which employs the first and second-order derivatives. The loss function includes a regularization term to manage the technique's intricacy and prevent overfitting. The XGBoost technique is derived as follows:

$$\widehat{y_i} = \phi(X_i) = \sum_{k=1}^{K} f_k(X_i), f_k \epsilon F$$

where $F = \{f(x) = w_{q(x)}\} (q : R^m \to T, w \in R^T)$ indicates a function space that defines a decision tree and **T** is the leaf node number of a decision tree. The following is the loss function:

$$L(\phi) = \sum_i l(y_i y_i) + \sum_k \Omega(f_k)$$

$$\Omega(f_k) = \Upsilon T + \frac{1}{2}\lambda \|w\|^2$$

The first component in Eq. (5) presents the number of leaves, while the second component is the size of the outcome. XGBoost calculates Gain for every node in the tree to assess whether the generated branch is relevant.

$$Gain = \frac{1}{2}(Gain_L + Gain_R - Gain_O) - \Upsilon$$

Where $Gain_O$ denotes the authentic gain before splitting and $-\Upsilon$ is the number of the new leaves.

4.4.3 Adaptive Boosting (Adaboost) model - with accuracy of 87%

The AdaBoost method enhances the performance of the classifier by integrating numerous weak learners into a single strong one. It repeatedly adjusts sample weights depending on classification mistakes, raising the weights of misclassified samples while reducing the weights of well-classified samples. As a result, classification methods that focus on miscategorized data rather than minority class examples are used. Because AdaBoost concentrates on prediction performance, the method is biased toward the majority class, which provides more to total prediction performance**.**

4.4.4 K-Nearest Neighbors (KNN) model - with accuracy of 77%

The KNN technique distinguishes samples by locating the nearest neighboring provided points and assigning the majority of n neighbors to a class. If there is a tie, many ways may be employed to settle it. Nevertheless, KNN is not recommended for big datasets because it does all computation throughout testing and converges during all trained data, calculating the closest neighbor each time. To locate the nearest neighbor in the features vector, the Euclidean distance function (Di) was used as follows:

$$D_i = \sqrt{(x_1 - x_2) + (y_1 - y_2)^2}$$

where x1,x2,y1,and y2 are parameters for data input.

4.4.5 Decision Tree (DT -) - with accuracy of 86%

The DT is a straightforward, basic approach that generates judgments depending on the values of all relevant input variables. DT chooses the root parameter based on entropy before analyzing the weights of the other variables. DT gathers all variable decisions grouped in a top-down tree and prepares the choice based on various values from special attributes. Previous research has revealed that decision tree models work well on unbalanced data. Nevertheless, ensemble techniques based on decision trees, such as Random Forest (RF), virtually usually surpass single decision trees. The benefits of decision-tree-based models are their insensitivity to missing values, ability to maintain both regular qualities and data, and high efficiency. Decision-tree-based techniques, as compared to other ML algorithms, are better for short-term forecasting and may have a faster computation speed.

4.4.6 Logistic Regression - with accuracy of 76%

Logistic Regression is a linear model used for binary classification tasks. In this project, Logistic Regression was applied to predict the safety of water based on features such as pH, hardness, solids, chloramines, sulfate, conductivity, organic carbon, trihalomethanes, and turbidity. The model achieved

an accuracy of approximately 61.3%. The classification report indicates that the precision for predicting unsafe water (Class 1) is particularly low, suggesting that the model struggles to correctly identify unsafe instances.

4.5 Model Evaluation

Cross-Validation: Implement k-fold cross-validation to evaluate model performance robustly. This involves partitioning the dataset into k subsets and iteratively training and testing the model on different subsets, ensuring all data points are used for both training and validation.

Evaluation Metrics: Calculate relevant evaluation metrics such as accuracy, precision, recall, F1-score, or area under the ROC curve (AUC-ROC) to assess the model's performance. These metrics provide insights into the model's ability to make correct predictions and its performance across different classes or thresholds.

4.6 Model Deployment

Model Serialization: Save the best-performing trained model to a file format compatible with deployment platforms (e.g., pickle). Serialization ensures the model's parameters are preserved, allowing for easy retrieval and reuse for future predictions.

Prediction Demonstration: Showcase how the saved model can be used to make predictions on new, unseen data. Provide sample input data and demonstrate how the model generates predictions, illustrating its practical usability in real-world scenarios.

4.7 Flask Integration In the Water Quality Prediction project, we seamlessly connected a frontend crafted in HTML with backend Python code using Flask. Acting as a mediator, Flask facilitates communication between the frontend and backend. Requests from the frontend are directed to specific Python functions defined in app.py, where machine learning models and processing logic are executed. Results are then returned to Flask, which renders the appropriate HTML templates to display the outcomes on the frontend. By running app.py, Flask initializes a local server, granting access to the application via a dynamically generated IP address and port, thereby enabling interaction with the project through a web browser.

Chapter 5.IMPLEMENTATION

5.1 Technologies used

The technologies used in this project include Python, Flask, Pandas, and various machine learning libraries such as scikit-learn. Several machine learning models like Logistic Regression, K-Nearest Neighbors, Decision Tree, Random Forest, AdaBoost, Bagging, and XGBoost are trained and tuned using GridSearchCV or RandomizedSearchCV for hyperparameter optimization. Additionally, the project involves the utilization of data visualization libraries such as Matplotlib and Seaborn for data exploration and visualization. Furthermore, HTML and CSS are employed for building the web application interface.

5.2 Sample code

5.3 Screenshots of webpages

Chapter 6. Testing

6.1 Testing strategies

Cross-Validation: Employed k-fold cross-validation to robustly evaluate model performance across different subsets of the water quality dataset. This technique ensured that our models generalize well to unseen data and helped identify potential overfitting issues.

Holdout Validation: Split the water quality dataset into training and testing sets, reserving a portion exclusively for testing. This allowed us to independently assess the performance of our models on unseen data, providing valuable insights into their real-world applicability.

Performance Metrics: Utilized standard evaluation metrics such as accuracy, precision, recall, and F1-score to measure the performance of our water quality prediction models.

Hyperparameter Tuning Validation: Validated the effectiveness of hyperparameter tuning strategies using techniques like nested cross-validation and separate validation sets. This ensured that the selected hyperparameters generalized well and led to improved model performance in predicting water quality.

6.2 Test case reports

Data Loading and Preprocessing

- **Description:** Verify that the water quality dataset is loaded correctly into the system and undergoes necessary preprocessing steps such as handling missing values and removing duplicates.
- **Expected Outcome:** The dataset is successfully loaded, and preprocessing steps are completed without errors.

Model Training and Evaluation

- **Description:** Train the water quality prediction models using various algorithms such as Random Forest, Decision Tree, and XGBoost. Evaluate the models' performance using appropriate metrics such as accuracy, precision, recall, and F1-score.
- **Expected Outcome:** Models are trained successfully, and performance metrics indicate satisfactory performance in predicting water quality.

Hyperparameter Tuning

- **Description:** Perform hyperparameter tuning using techniques like Grid Search or Random Search to optimize model performance. Validate the effectiveness of hyperparameter tuning strategies.
- **Expected Outcome:** Optimal hyperparameters are selected, leading to improved model performance compared to baseline models.

User Interface Integration

- **Description:** Integrate the machine learning pipeline with a user-friendly interface using Flask. Test the interface's functionality for inputting water attributes and obtaining predictions regarding water potability.
- **Expected Outcome:** The user interface is intuitive and functional, allowing users to easily input water attributes and receive accurate predictions regarding water safety.

End-to-End Testing

- **Description:** Conduct end-to-end testing to validate the entire water quality prediction system, including data loading, preprocessing, model training, evaluation, and user interface integration.

- **Expected Outcome:** The complete system operates smoothly, providing accurate predictions of water potability while ensuring a seamless user experience.

Chapter 7. Conclusion

In conclusion, our water quality prediction project extensively evaluated various machine learning models, with XGBoost emerging as the top performer with an accuracy of 88%. This comprehensive comparison highlighted the effectiveness of XGBoost, alongside notable contenders like Random Forest. The choice of model depends on specific project requirements and trade-offs between performance metrics. Future iterations could focus on further feature engineering or exploration to enhance accuracy. Ultimately, our project demonstrates the utility of machine learning in predicting water quality, offering reliable insights for real-world applications.

Chapter 8 Bibliography

1. Jain D, Shah S, Mehta H et al (2021) A Machine Learning Approach to Analyze Marine Life Sustainability. In: Proceedings of International Conference on Intelligent Computing, Information and Control Systems. Springer, pp 619–632

2. Clark RM, Hakim S, Ostfeld A (2011) Handbook of water and wastewater systems protection. In: Protecting Critical Infrastructure. Springer, pp 1–29. https://doi.org/10.1007/978-1-4614-0189-6

3. Hu Z, Zhang Y, Zhao Y et al (2019) A water quality prediction method based on the deep LSTM network considering correlation in smart mariculture. Sensors 19:1420

4. Zhou J, Wang Y, Xiao F et al (2018) Water quality prediction method based on IGRA and LSTM. Water 10:1148

5. Waqas M, Tu S, Halim Z et al (2022) The role of artificial intelligence and machine learning in wireless networks security: principle, practice and challenges. Artif Intell Rev 55:5215–5261. https://doi.org/10.1007/s10462-022-10143-2

6. Halim Z, Waqar M, Tahir M (2020) A machine learning-based investigation utilizing the in-text features for the identification of dominant emotion in an email. Knowl Based Syst 208:106443. https://doi.org/10.1016/j.knosys.2020.106443

7. Wu J, Wang Z (2022) A Hybrid Model for Water Quality Prediction Based on an Artificial Neural Network, Wavelet Transform, and Long Short-Term Memory. Water 14:610

8. Lee S, Lee D (2018) Improved prediction of harmful algal blooms in four Major South Korea's Rivers using deep learning models. Int J Environ Res Public Health 15:1322