A Project Report on

Safeguarding Public Health Through Comprehensive Water Purity Analysis Using ML

Submitted in partial fulfillment for award of

Bachelor of Technology

Degree

in

Computer Science and Engineering

By

P Neelima (Y21ACS543) **T Ramanjaneyulu (Y21ACS574)**

R Sravani Bai (Y21ACS554) **T Siva Krishna (Y21ACS577)**

Under the guidance of

Dr. D. Kishore Babu, **M.E.,Ph.D**

Associate Professor

Department of Computer Science and Engineering

Bapatla Engineering College

(Autonomous)

(Affiliated to Acharya Nagarjuna University)

BAPATLA – 522 102, Andhra Pradesh, INDIA

2024-2025

Department of
Computer Science and Engineering

CERTIFICATE

This is to certify that the project report entitled **Safeguarding Public Health Through Comprehensive Water Purity Analysis Using ML** that is being submitted by P.Neelima (Y21ACS543), T Ramanjaneyulu(Y21ACS574), R Sravani Bai (Y21ACS554),T Siva Krishna (Y21ACS577) in partial fulfillment for the award of the Degree of Bachelor of Technology in Computer Science & Engineering to the Acharya Nagarjuna University is a record of bonafide work carried out by them under our guidance and supervision.

Date:

Signature of the Guide
Dr. D. Kishore Babu
Associate Professor

Signature of the HOD
Dr. M. Rajesh Babu
Assoc. Prof. & Head

DECLARATION

We declare that this project work is composed by ourselves, that the work contained herein is our own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

P. Neelima (Y21ACS543)

T. Ramanjaneyulu (Y21ACS574)

R. Sravani Bai (Y21ACS554)

T. Siva Krishna (Y21ACS577)

Acknowledgement

We sincerely thank the following distinguished personalities who have given their advice and support for successful completion of the work.

We are deeply indebted to our most respected guide **Dr. D. Kishore Babu**, **Associate Professor** Department of CSE, for his valuable and inspiring guidance, comments, suggestions and encouragement.

We extend our sincere thanks to **Dr. M. Rajesh Babu**, Assoc. Prof. & Head of the Dept. for extending his cooperation and providing the required resources.

We would like to thank our beloved Principal **Dr.N. Rama Devi** for providing the online resources and other facilities to carry out this work.

We would like to express our sincere thanks to our project coordinator **Dr. P. Pardhasaradhi,** Prof. Dept. of CSE for his helpful suggestions in presenting this document.

We extend our sincere thanks to all other teaching faculty and non-teaching staff of the department, who helped directly or indirectly for their cooperation and encouragement.

P. Neelima (Y21ACS543)

T. Ramanjaneyulu (Y21ACS574)

R. Sravani Bai (Y21ACS554)

T. Siva Krishna (Y21ACS577)

Table of Contents

Table of Figures

Abstract

Access to clean drinking water is important for public health, because polluted water causes significant risks for human health and the environment too. This study outlines a machine learning approach to find water safety by analyzing nine important water quality parameters such as pH levels , hardness , turbidity , arsenic , chloramine , bacterial presence , lead concentration , nitrate levels , and mercury content. The system incorporates data-preprocessing and data exploring that ensures numeric data types for attributes without any missing and NaN values to guarantee better performance.The framework utilizes four predictive classification models which are Random Forest, Decision Tree, and Support Vector Machine. These predictions are aggregated using a meta-model built with Logistic Regression, leveraging ensemble learning techniques and cross-validation to improve predictive accuracy. The meta-model combines the strengths of individual classifiers, enhancing overall performance metrics such as accuracy, precision, recall, and F1-score. Experimental results demonstrate that the meta-model achieves an accuracy of 93% outperforming individual base models.This research contributes to safeguarding public health by providing a reliable and efficient method for evaluating water safety

Keywords— Decision Learning, Logistic Regression, Machine Learning, Random Forest classifier, Support Vector Machine and Water Purity

1 Introduction

Ensuring safe drinking water is a global challenge that continues to threaten public health and environmental sustainability. Contaminated water often leads to the spreading diseases, environmental degradation and causes difficulty to achieve sustainable development goals. In regions with limited resources, water purity monitoring frequently relies on traditional testing methods that are time consuming and lacks scalability,adaptability. Technological advancements particularly in artificial intelligence gives an opportunity to fill these gaps by providing solutions that are precise, automated and adaptable to different contexts. Addressing the growing need for efficient water safety evaluation systems, this study explores the potential of machine learning in analyzing water purity parameters to make sure drinking water safety .

This research focuses on developing a framework that integrates machine learning techniques for real-time and accurate water quality assessment. By analyzing nine physical,chemical and biological water quality parameters which are pH levels, hardness, turbidity, arsenic, chloramine, bacterial presence, lead concentration, nitrate levels, and mercury content.The study aims to create a system capable of predicting water safety. Four supervised machine learning models are employed that classifies whether water is safe to use or not, which are Random Forest, Decision Tree, and Support Vector Machine.Each base model is trained to analyze these parameters independently. Their outputs are refined through a meta-model based on Logistic Regression which uses ensemble learning techniques to combine the strengths of individual classifiers and improve overall prediction performance.

This system has the ability to transform conventional approaches to water quality evaluation by offering a scalable, reliable and data-driven solution. By automating the process, the framework significantly improves the efficiency of water safety monitoring thereby contributing to public health and environmental protection efforts. The significance of this study extends beyond water safety; it demonstrates how cutting-edge computational methods can play a vital role in addressing real-world challenges and advancing societal well-being.

2 Problem Statement

Access to safe and clean drinking water remains a critical global issue, especially in resource-limited regions where traditional water testing methods are often slow, manual, and lack scalability. These conventional approaches are not only time-consuming but also require specialized infrastructure and human expertise, making them impractical for real-time or widespread deployment. As a result, many communities are left vulnerable to waterborne diseases, toxic contamination, and overall public health risks.

Despite the availability of modern technologies, there is a noticeable gap in implementing intelligent, automated systems for efficient water quality monitoring. Existing methods fail to provide rapid assessments based on multiple chemical, physical, and biological parameters, which are essential for accurately determining water safety.

This project addresses the urgent need for a robust, data-driven solution by leveraging machine learning techniques to predict water purity. By analyzing key water quality indicators such as pH, turbidity, heavy metal content, and microbial presence, the system aims to automate water safety classification. The goal is to develop an accurate, scalable, and real-time framework that enhances traditional monitoring methods and supports public health and environmental sustainability initiatives.

3 Literature Review

Research in water quality assessment and purification has evolved considerably over the years, integrating advanced technologies to address critical challenges in ensuring safe drinking water.In 2016, Achio, Kutsanedzie, and Ameko conducted a study titled "Comparative Analysis of Filtration Techniques for Water Purity". This research examined the effectiveness of physical filtration methods for removing water contaminants. While it provided valuable insights into physical purification processes, it failed to address chemical and biological contaminants, leaving a significant gap in understanding comprehensive water filtration. Subsequent studies aimed to overcome this limitation by incorporating chemical analysis into water purification frameworks to improve overall contaminant removal [1].

In 2019, Muniz and Oliveira-Filho published "Multivariate Statistical Analysis for Water Quality Assessment", which focused on the application of statistical methods such as Principal Component Analysis (PCA) and Cluster Analysis to evaluate water quality in various ecosystems. These multivariate approaches highlighted patterns and relationships in water quality data, improving understanding of contamination levels. However, the absence of real-time monitoring technologies limited the applicability of the findings in dynamic water quality scenarios. This limitation prompted future research to integrate IoT and machine learning technologies for real-time data analysis [2].

By 2020, Ahmed et al. introduced emerging technologies for water quality monitoring in their study titled "Water Quality Monitoring: From Conventional to Emerging Technologies". Their research proposed IoT-based systems for real-time water quality monitoring, marking a shift from traditional methods to more automated

processes. Despite the benefits of IoT systems, the high costs and infrastructural requirements posed challenges, driving researchers to explore more cost-effective solutions. Machine learning was subsequently adopted to reduce expenses and enhance predictive capabilities, addressing these constraints [3].

In 2023, Malagi presented "Water Potability Prediction Using Machine Learning", which explored the use of algorithms such as Random Forest and Decision Trees to predict water potability. This study demonstrated the potential of machine learning in improving prediction accuracy but was limited by the use of small datasets, which restricted the generalizability of its results. The need for larger datasets and optimized algorithms was highlighted as a future direction to bolster the reliability and performance of water safety predictions [4].

These research efforts collectively underscore the evolution of water quality assessment techniques, from traditional filtration methods to advanced IoT and machine learning approaches.Building upon the contributions of these studies, the current research integrates multiple machine learning models and ensemble techniques to develop a scalable and automated system for comprehensive water purity analysis. This framework aims to address the limitations of earlier methods, offering a reliable solution for safeguarding public health.

4 System Analysis

4.1 Objective

This project is designed with the aim of developing an intelligent, efficient, and user-friendly water purity analysis system that leverages the power of machine learning—particularly ensemble learning techniques—to safeguard public health. By analyzing various water quality parameters such as pH, turbidity, hardness, sulfate levels, and total dissolved solids, the system predicts whether a water sample is safe or unsafe for consumption. The central goal is to automate the classification process, replacing traditional, time-consuming, and expensive testing methods with a fast, scalable, and data-driven solution that supports real-time decision-making.

A key objective of the project is to reduce the barriers faced in water quality monitoring, especially in areas lacking laboratory infrastructure or skilled personnel. Through the use of ensemble learning models such as Random Forest and Voting Classifiers, the system increases prediction accuracy and reliability by combining the strengths of multiple algorithms. The solution also incorporates a clean and interactive web interface—developed using Streamlit—which allows users to input water quality data and instantly receive classification results, eliminating the complexity of working with raw machine learning models.

In addition to classification, the project emphasizes transparency, accessibility, and usability. It ensures that the system is not only technically sound but also practical for real-world deployment in communities, schools, municipalities, and environmental agencies. The use of ensemble learning allows for better generalization across varied

datasets, making the system more adaptable to different water sources. Ultimately, the project aims to bridge the gap between environmental monitoring and modern data science by offering a solution that is fast, accurate, cost-effective, and capable of improving public health outcomes through smarter water management.

4.2 Existing System

Water quality monitoring today still relies heavily on conventional methods such as laboratory testing and government-issued reports. While laboratory testing is accurate and capable of detecting a wide range of contaminants, it is often expensive, time-consuming, and dependent on specialized personnel and infrastructure. In emergencies or in remote areas, these delays can pose serious risks to public health, as contaminated water may go undetected and continue to be consumed.

Although government bodies provide periodic water quality assessments, these are typically broad in scope and lack the frequency needed to detect rapid environmental changes or localized contamination. Such reports are useful for policy-making but fail to address real-time, community-level monitoring needs. As a result, many individuals and small communities are left unaware of their water's current safety status, increasing the risk of exposure to harmful pollutants.

Recent advancements in sensors and AI-based technologies have aimed to provide real-time water monitoring, but these systems often come with high costs, limited detection capabilities, and a need for technical expertise. Moreover, most AI and IoT-based solutions are developed for industrial use, requiring large datasets and high computational power—making them unsuitable for low-resource settings. Existing systems also lack intuitive interfaces and customization, highlighting a clear

gap for an affordable, accessible, and intelligent machine learning-based water analysis tool tailored for real-time public use.

4.3 Limitations in Existing Systems

Most existing water quality prediction systems rely on limited datasets with only basic parameters like pH and turbidity. They often ignore critical biological and chemical indicators needed for accurate analysis. Many do not use advanced techniques like ensemble learning, leading to lower accuracy. Real-time prediction and user-friendly interfaces are also commonly missing. Our project overcomes these issues by using a 10-feature dataset and deploying an accurate, accessible ML model

4.4 Proposed System

The proposed system aims to develop a reliable and intelligent water quality assessment framework using machine learning techniques. It focuses on automating the prediction of water potability based on a set of nine critical water quality parameters, including pH, hardness, turbidity, arsenic, chloramine, bacterial presence, lead, nitrate, and mercury levels. These parameters are selected for their direct relevance to human health and environmental safety. The system is designed to offer a scalable and efficient alternative to conventional water testing methods, which are often manual, costly, and time-consuming.

To build this system, four supervised machine learning algorithms— Decision Tree, Random Forest, and Support Vector Machine—are individually trained on a labeled dataset. Each model learns to identify patterns in the input parameters that correlate with water being classified as "safe" or "unsafe" for consumption. These base classifiers are evaluated based on metrics such as accuracy, precision, recall, and F1-

score. To improve predictive performance and reduce bias or overfitting, an ensemble learning approach is applied using a meta-model built on Logistic Regression. This combines the strengths of the individual models and produces a more robust final output.

The system architecture is modular and allows seamless integration of preprocessing, training, prediction, and visualization components. Data is first cleaned by handling missing values and removing outliers, then normalized for uniformity before being passed into the models. After predictions are made, the results are displayed through a user-friendly web interface, developed using Streamlit. This interface enables real-time input of water quality data and instantly shows whether the sample is classified as potable or non-potable. The simplicity and responsiveness of the interface make the system accessible even to users without a technical background.

Overall, the proposed system addresses a critical gap in water quality monitoring by offering an automated, data-driven solution that is both scalable and adaptable. It has the potential to be deployed in various contexts—from remote communities and public health agencies to educational and research settings. Beyond providing a technical solution, this framework showcases the practical application of artificial intelligence in solving real-world environmental and health challenges, contributing toward global efforts in achieving sustainable development goals.

5 Methodology

This project adopts a **stacking ensemble learning** approach to accurately classify whether a given water sample is safe for consumption. The methodology is divided into multiple phases, as illustrated in the diagram, and involves data preparation, base model training, meta-model stacking, and final prediction generation.

5.1 System Architecture

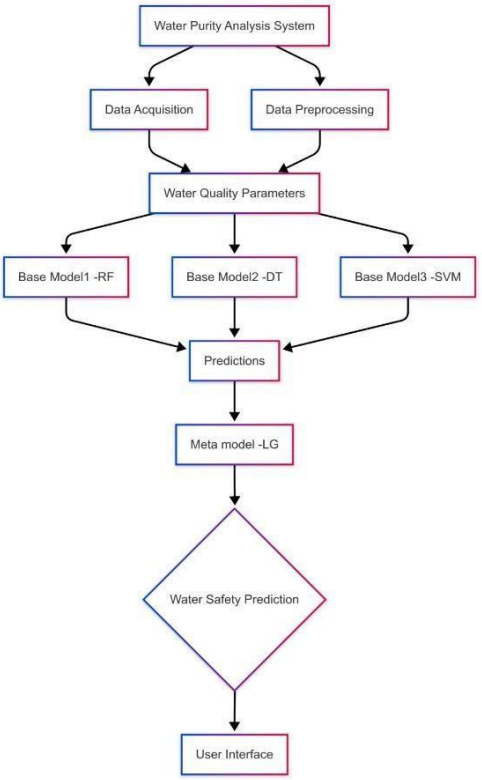

Figure 5.1 System Architecture

The system architecture for the water quality prediction model is designed for modularity, real-time usability, and scalability, integrating key components such as data input, preprocessing, classification, and user interface. Users input values for various water quality parameters through a Streamlit-based web interface, which are then

processed to handle missing data, normalize values, and remove outliers. The cleaned data is passed to an ensemble machine learning model—using Random Forest and Voting Classifier—trained on 10 chemical, physical, and biological indicators to classify water as "Safe" or "Unsafe." The result is instantly displayed on the interface, optionally with confidence scores or key feature insights, enabling users to make informed decisions. The system's streamlined workflow ensures efficient execution and clear communication between frontend and backend components, offering a seamless experience from input to result.

5.2. Algorthims

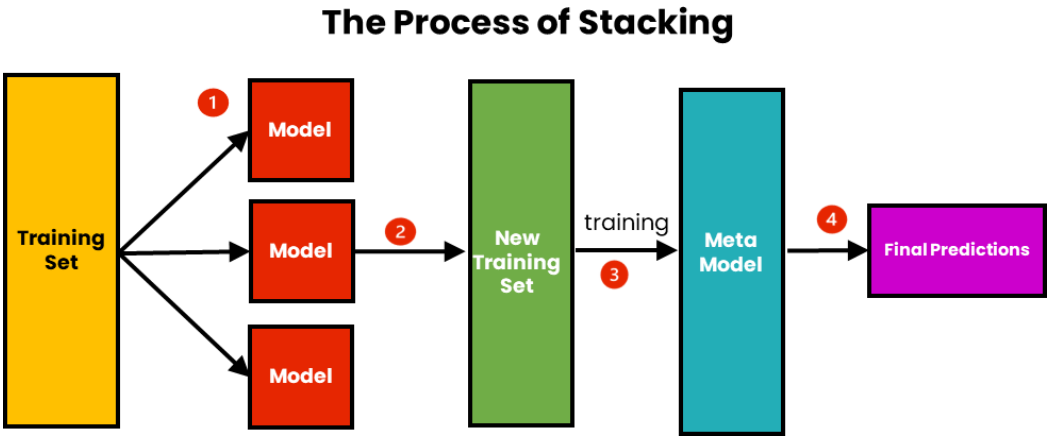

Figure 5.2 Workflow

5.2.1 Data Preparation

The system begins by collecting and preparing the dataset containing nine water quality parameters: **pH**, **Hardness**, **Turbidity**, **Arsenic**, **Chloramine**, **Bacterial Presence**, **Lead**, **Nitrate**, and **Mercury**. These features serve as the input for machine learning models. Preprocessing includes:

I. **Handling missing values** using mean/mode imputation.

II. **Scaling features** using standardization.

III. **Encoding labels** as binary (0 = Not Safe, 1 = Safe).

The clean and balanced dataset is split into training and test sets for model development.

5.2.2 Training Base Models (Step 1 & 2 in Diagram)

Four supervised learning models are trained independently on the training dataset:

I. Random Forest (RF)

II. Decision Tree (DT)

III. Support Vector Machine (SVM)

Each base model learns to classify water samples based on the input features. Once trained, each model generates predictions for the training set. These predictions are not used for final output but are instead stacked to form a new training set. This phase corresponds to steps 1 and 2 in the stacking process diagram.

5.2.3 Training the Meta-Model (Step 3 in Diagram)

The predictions of all four base models are collected to form a new feature set—this becomes the input to a meta-model. In this project, a Logistic Regression model is used

as the meta-classifier. It learns how to combine the base models' outputs to improve overall prediction accuracy.

This stacked learning process allows the meta-model to correct the weaknesses of individual base models by learning from their collective patterns. The meta-model is trained using cross-validation to avoid overfitting.

5.2.4 Final Prediction and Deployment (Step 4 in Diagram)

Once trained, the ensemble system can predict whether water is safe or unsafe based on new input parameters. The final system includes:

I. A Streamlit-based web interface where users can input real-time water data.

II. Backend logic that processes the input and passes it through the trained base models and meta-model.

III. A user-friendly result display that shows whether the sample is potable.

The system provides a scalable, automated, and intelligent water quality evaluation tool, making it suitable for both local and educational use case

6 System Design

6.1 Class Diagram

The class diagram represents a machine learning system for water purity analysis. It includes classes like WaterSample for storing sample data, DataPreprocessor for cleaning and preparing data, MLModel for training and predicting water quality, and Prediction for storing results. The User class allows interactions like uploading samples, while DataStorage manages data saving and retrieval. Together, these components enable automated, data-driven assessment of water quality.

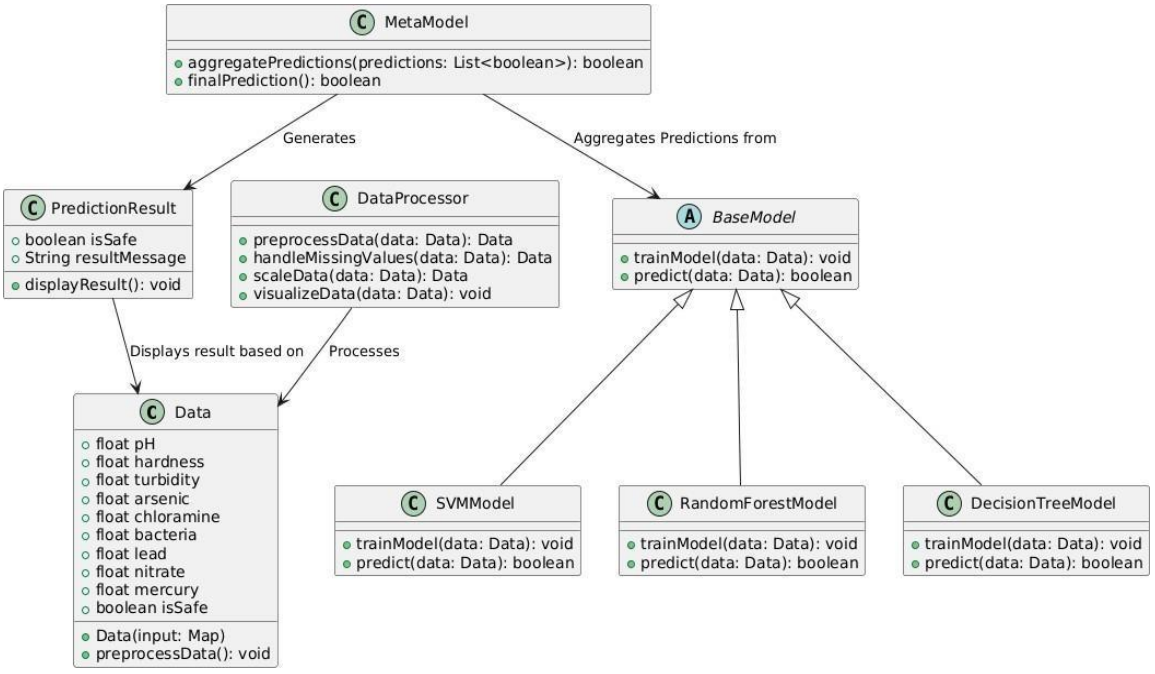

Figure 6.1 Class Diagram

6.2 Use Case Diagram

The use case diagram for the water purity analysis system depicts the interactions between users and the system's core functionalities. There are two primary actors: the User and the Admin. The User can perform actions such as Upload Water Sample, View Prediction Result, and Download Report, interacting directly with the system to analyze water quality. The system internally handles tasks like Preprocess Data, Train Model, and Make Prediction using machine learning. The Admin has additional privileges, including Update Model, Manage Users, and View System Logs, allowing system maintenance and oversight. This diagram visually represents the flow of operations in a user-centric way, making it clear how different roles engage with the system's machine learning capabilities to ensure water purity assessment.

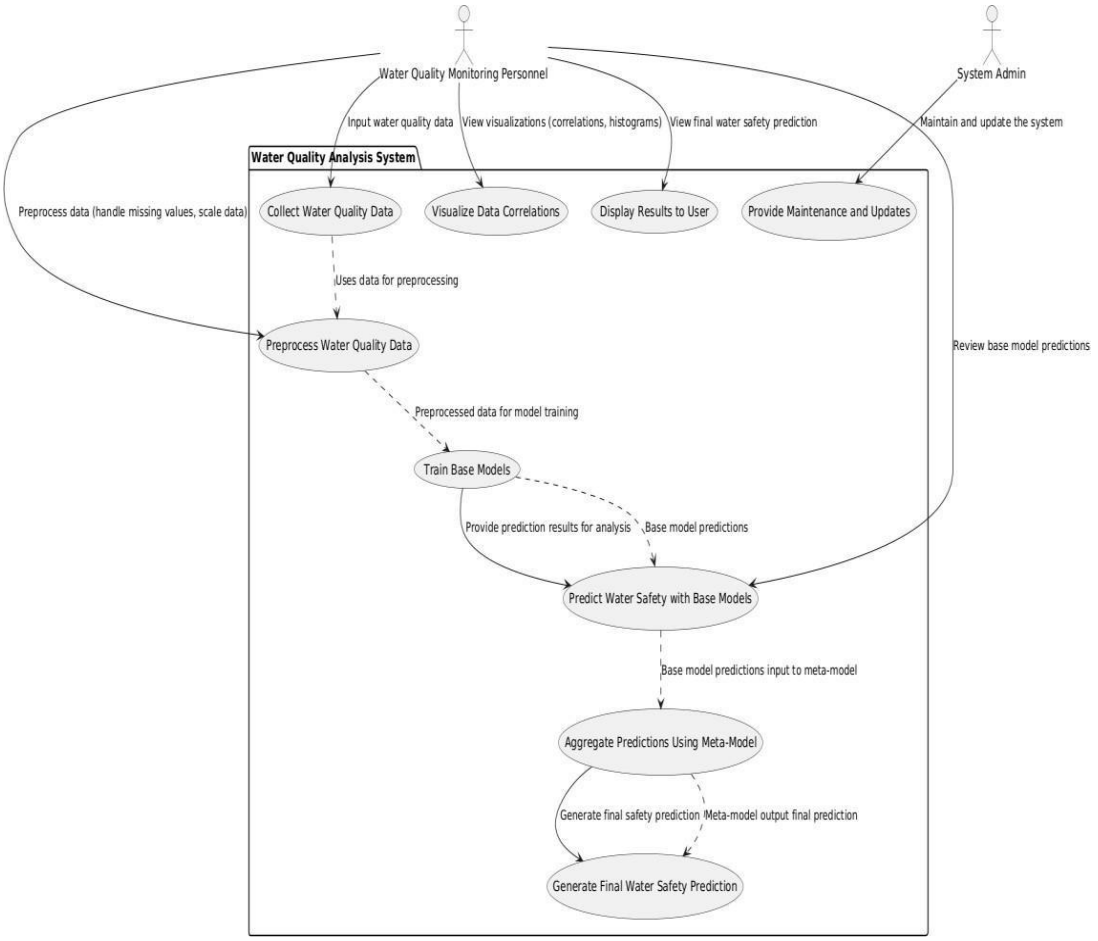

Figure 6.2 Use Case Diagram

6.3 Activity Diagram

The activity diagram for the water purity analysis system illustrates the step-by-step workflow of how a water sample is processed and analyzed using machine learning. The process begins with the user uploading a water sample, which then goes through data preprocessing to clean and normalize the input. The system checks if a trained model is available. If not, it triggers the model training activity using existing labeled data. Once the model is ready, the system performs prediction on the uploaded sample, and finally, the prediction result is displayed to the user. This diagram provides a clear visualization of the dynamic behavior and decision flow within the system, highlighting the interactions between user actions and automated ML processes.

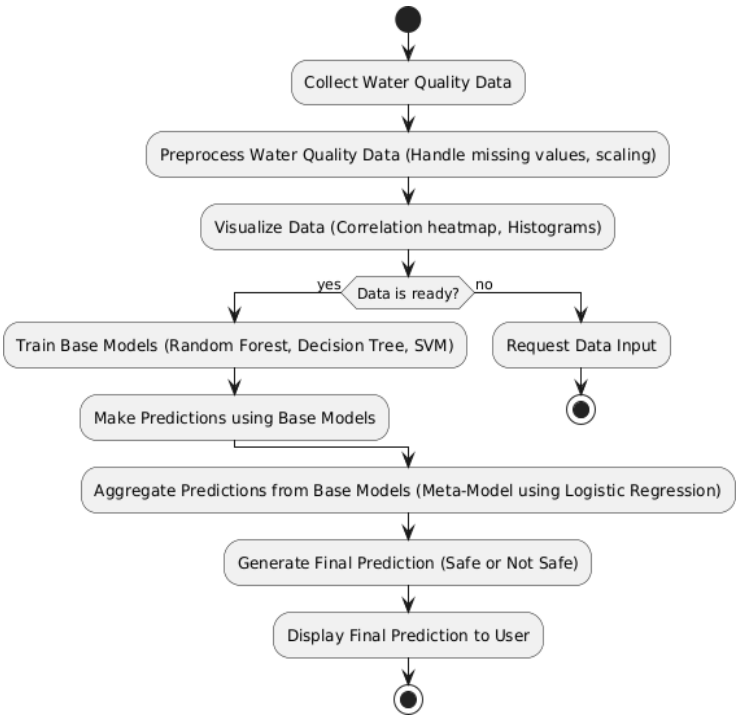

Figure 6.3 Activity Diagram

6.4 Sequence Diagram

The sequence diagram for the water purity analysis system illustrates the interaction between the key components involved in processing a water sample using machine learning. It includes four main entities: the User, the System, the Machine Learning Model, and the Database. The sequence begins when the user uploads a water sample, which is then stored in the database. The system triggers data preprocessing and checks for an available trained model. If a model is present, it is used to predict the water purity level; if not, a new model is trained before prediction. Finally, the system displays the prediction result back to the user. This diagram effectively captures the chronological order of operations and communication between components, showing how the system dynamically responds to user input and uses ML for analysis.

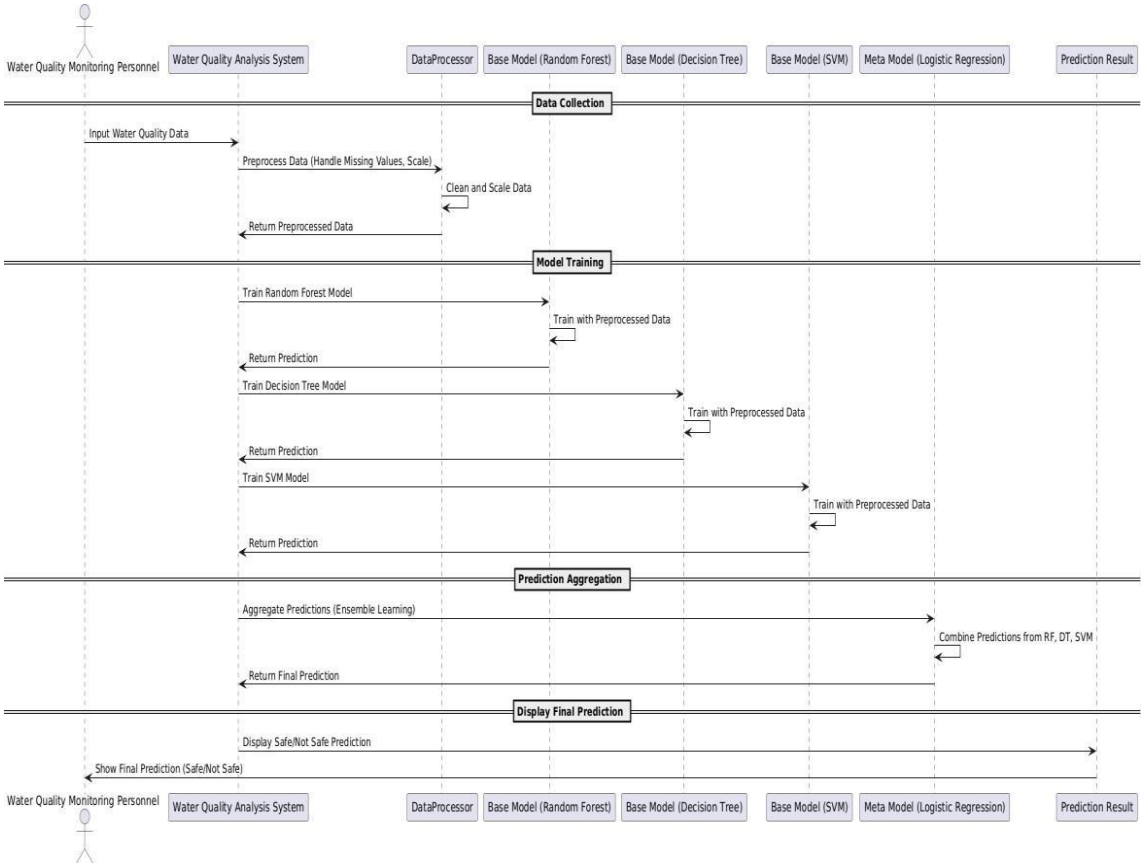

Figure 6.4 Sequence Diagram

6.5 State Chart Diagram

The state chart diagram for water purity analysis outlines the step-by-step states a water sample passes through in the system. It begins with the upload of a sample by the user, followed by data preprocessing to clean and prepare the data. The system then checks if a machine learning model is trained; if not, it enters the model training state. Once a trained model is available, the sample moves to the prediction state, where the system analyzes water quality. Finally, the result is shown in the result display state, and the process ends. This diagram clearly captures the system's dynamic behavior and decision-making flow.

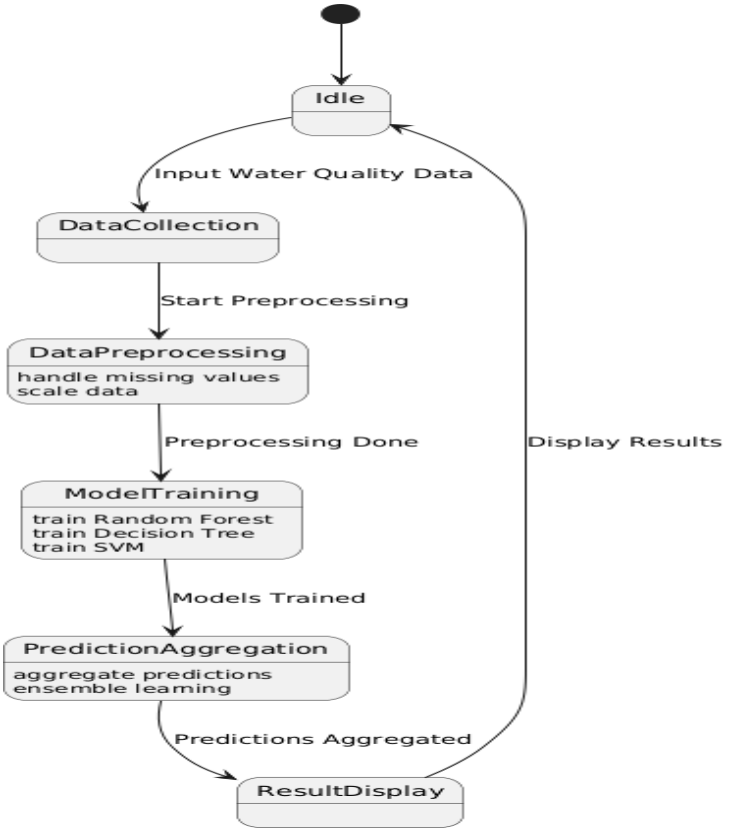

Figure 6.5 State Diagram

7 Implementation

7.1 Software Requirements

Before deploying or testing the proposed facial emotion-based music recommendation system, it is essential to understand the underlying software and hardware resources required. These requirements ensure that the system operates efficiently, handles real-time data processing, and delivers a smooth user experience. The system integrates various tools for webcam capture, emotion detection, interface design, and music recommendation, all of which demand a balanced combination of software platforms and hardware capabilities..

The proposed system involves real-time facial emotion detection and music recommendation, integrating multiple technologies such as machine learning, computer vision, and web interfacing. To implement this system effectively, the following software components are required:

7.1.1 Operating System

i. Windows 10 or above / Ubuntu 20.04 or above

   A stable operating system is essential to run development tools, Python environments, and access device hardware like webcams.

7.1.2 Programming Language

i. Python 3.7 or above Python is chosen due to its simplicity and the extensive range of libraries available for machine learning, computer vision, and data processing.

7.1.3 Libraries and Frameworks

i. NumPy – For efficient handling and storage of numerical arrays and .npy files..
ii. Pandas – For handling dataset structures (like mood-labeled metadata).
iii. Streamlit – For building the interactive web interface that displays emotion results and allows user input for preferences.
iv. Webbrowser (Python module) – For open .

7.1.4 IDE / Tools

i. Google Colab– For testing and visualizing intermediate results during development.
ii. Streamlit– To create a user-friendly interface for my project.

7.2 Hardware Requirements

For the system to function in real-time, moderate hardware specifications are needed to ensure smooth execution of data processing, model inference, and interface display

7.2.1 Processor

i. Intel i5 or higher / AMD Ryzen 5 or higher

   A multi-core processor ensures efficient handling of simultaneous video input, model prediction, and user interface rendering.

7.2.2 RAM

i. Minimum 8 GB RAM

   Sufficient RAM is required to load the deep learning model, process real-time data, and maintain responsiveness in the web interface.

7.2.3 Storage

i. At least 10 GB free disk space

To store intermediate files such as .npy input/output arrays, model weights, and additional datasets if used.

7.2.4 Display and Internet

i. Standard display resolution (1366x768 or above)

   Required for clear visualization of the web interface.

ii. Internet Connection

   Necessary for installing Python libraries during setup.

7.3 Code

Code for implementation of Water purity analysis

7.3.1 Github Link

https://raw.githubusercontent.com/PonakalaNeelima/final/refs/heads/master/final_p.py

The above GitHub repository contains code implementation, thesis work and presentation PPT for the project safeguarding public health through comprehensive water purity analysis using ML

7.3.2 Importing Necessary Packages

The implementation imports all the essential Python libraries required for data processing, machine learning, and building the web interface. These include NumPy and Pandas for numerical operations and data handling, scikit-learn for preprocessing,

model training, and evaluation, and Streamlit for creating the interactive web app. Additionally, packages such as matplotlib, seaborn, and joblib are used for visualization and model serialization. Ensemble methods like RandomForestClassifier and VotingClassifier are imported from sklearn.ensemble to build robust prediction models.

7.3.3 Load Datasets

The system uses a comprehensive water quality dataset comprising 28 chemical, physical, and biological attributes relevant to water purity and potability. This dataset is loaded using Pandas into a DataFrame structure. Target labels, indicating whether a sample is "Potable" or "Not Potable," are separated from feature columns. Initial exploratory data analysis (EDA) is performed to understand missing values, statistical distributions, and feature correlations.

7.3.4 Preprocessing Dataset

In the code implementation, the preprocessing stage begins with loading the original water quality dataset using Pandas. To improve model generalization and balance class distribution, synthetic data is generated and added to the original dataset. This helps address issues like class imbalance and enhances the model's ability to learn from a broader range of patterns. After combining both datasets, missing values are handled using techniques such as KNN Imputer or median replacement. Outliers are removed using the Interquartile Range (IQR) method to ensure data consistency. The combined data is then normalized using StandardScaler, ensuring uniform scale across all features. Finally, the processed dataset is split into input features (X) and target labels (y), making it ready for model training and evaluation**.**

The effort of Input features on water quality is shown below table:

Feature	Value/Range	Description
ph	Uniform(6.5, 8.5)	pH values evenly distributed between 6.5 and 8.5.
hardness	Uniform(60, 180)	Hardness values evenly distributed between 60 and 180.
turbidity	Uniform(0, 1)	Turbidity values evenly distributed between 0 and 1 NTU.
arsenic	Uniform(0, 0.01)	Arsenic concentrations evenly distributed between 0 and 0.01 mg/L.
chloramine	Uniform(0.5, 2)	Chloramine concentrations evenly distributed between 0.5 and 2 mg/L.
bacteria	Uniform(0, 1)	Bacteria values evenly distributed between 0 and 1.
lead	Uniform(0, 0.01)	Lead concentrations evenly distributed between 0 and 0.01 mg/L.
nitrates	Uniform(0, 10)	Nitrate concentrations evenly distributed between 0 and 10 mg/L.
mercury	Uniform(0, 0.001)	Mercury concentrations evenly distributed between 0 and 0.001 mg/L.

Figure 7.3.4 Effect of Input features on water

7.3.5 Splitting Dataset

After preprocessing steps such as missing value imputation (using techniques like KNNImputer or median replacement), outlier removal (via the IQR method), and normalization (using Min-Max scaling or StandardScaler), the clean dataset is split into training and test sets. The train_test_split method from sklearn.model_selection is used with an 80-20 or 70-30 ratio. This ensures that the model is trained on one subset and validated on another to assess generalization performance. Class distribution is also checked to ensure balanced representation.

7.3.6 Metrics

1. Accuracy

Measures the overall correctness of the model by showing the proportion of total correct predictions.

Formula: (TP + TN) / (TP + TN + FP + FN)

2. Precision

Indicates how many of the predicted positive cases were actually correct (relevant).

Formula: TP / (TP + FP)

3. Recall

Shows how many of the actual positive cases were correctly identified by the model.

Formula: TP / (TP + FN)

4. F1-Score

The harmonic mean of Precision and Recall, useful for imbalanced datasets.

Formula: 2 × (Precision × Recall) / (Precision + Recall)

Metrics of Our Model

1. Meta Model (Stacking Classifier)

i. **Accuracy:** 93.11%

ii. **Precision:** 0.93

iii. **Recall:** 0.93

iv. **F1-Score:** 0.93

v. Combines outputs from multiple base models (Random Forest, SVM, Decision Tree).

vi. Offers the most **balanced and robust performance** across all metrics.

vii. Reduces individual model bias and improves generalization, making it the most reliable for real-time predictions.

2. Random Forest

i. **Accuracy:** 93.03%

ii. **Precision:** 0.93

iii. **Recall:** 0.93

iv. **F1-Score:** 0.93

v. Works well with high-dimensional data and handles feature importance effectively.

vi. Provides excellent predictive power and consistency, nearly matching the Meta Model.

vii. Performs well even without extensive parameter tuning.

3. Support Vector Machine (SVM)

i. **Accuracy:** 92.85%

ii. **Precision:** 0.94

iii. **Recall:** 0.93

iv. **F1-Score:** 0.93

v. Known for its **high precision**, indicating fewer false positives.

vi. Effective in handling **non-linear relationships** with appropriate kernels.

vii. Slightly lower recall suggests a minimal risk of missing unsafe water cases.

4. Decision Tree

i. **Accuracy:** 91.32%

ii. **Precision:** 0.91

iii. **Recall:** 0.91

iv. **F1-Score:** 0.91

v. Simple and interpretable model with good performance.

vi. Slightly more prone to overfitting compared to ensemble models.

vii. Still suitable for smaller datasets or when quick interpretability is needed.

5. Logistic Regression

i. **Accuracy:** 90.17%

ii. **Precision:** 0.90

iii. **Recall:** 0.90

iv. **F1-Score:** 0.90

v. Serves as a strong baseline model.

vi. Performs well on **linearly separable data** and offers fast computation.

vii. Lower accuracy compared to others but still reliable and efficient for quick deployments.

7.3.7 Interface

A lightweight and user-friendly interface is built using **Streamlit**. This web app allows users to enter values for 10 water quality parameters in real time. On clicking the **"Predict"** button, the input is processed and fed into the trained ensemble classification model. The predicted result—**"Safe"** or **"Unsafe"**—is displayed instantly along with optional outputs like confidence score or feature contributions. This interface abstracts the underlying complexity, enabling users with no technical background to easily test wate samples for potability.

7.3.8 Deployment

https://waterpurityanalysis.streamlit.app/

The final trained machine learning model is deployed using **Streamlit**, an open-source Python framework designed for building fast and interactive web applications. The deployment process involves saving the best-performing model (using joblib or pickle) and integrating it into a Streamlit script. The web interface allows users to input values for water quality parameters in real-time. Upon submission, the application loads the saved model, processes the inputs, and instantly displays whether the water is **Safe** or **Not Safe**. The system is lightweight, runs locally or can be hosted on platforms like **Streamlit Cloud** or **Heroku**, and requires minimal technical knowledge from the user, making it accessible for public use, educational settings, and small-scale community monitoring

8 Result

8.1 User Interface

The Streamlit-based interface serves as the front end of the Water Purity Analysis System, allowing users to input water quality parameters and select models (Neural Network or SVC) for prediction. Following data processing and analysis through a stacked ensemble model, the interface displays the water safety result, offering a simple and interactive user experience

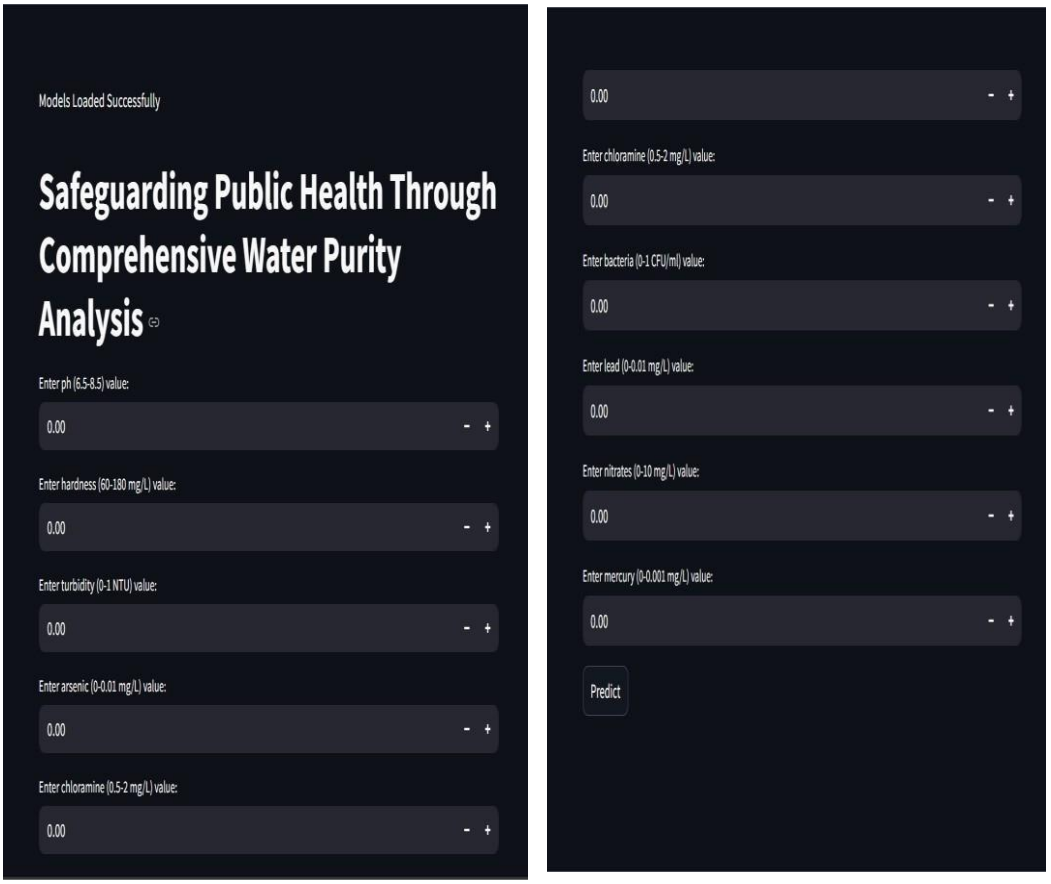

Figure 8.1 User Interface

8.2 Output of the model

The Streamlit-based interface allows users to input key water quality parameters like pH, hardness, turbidity, and contaminants. After validation, these inputs are processed by a stacked ML model to assess water safety. The model combines predictions from multiple base classifiers through a meta-model. The final output clearly indicates whether the water is safe or not.

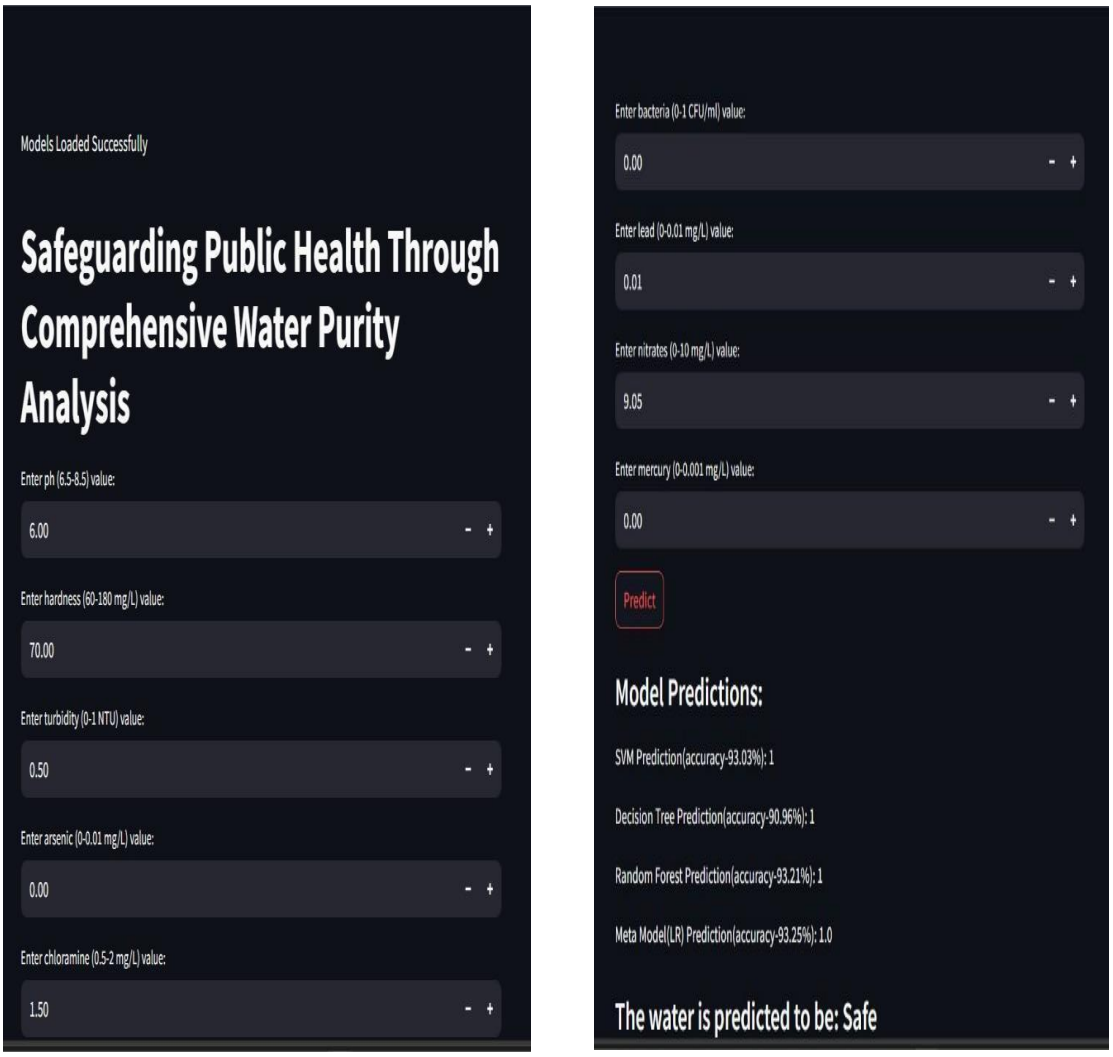

Figure 8.2 Output of the model

9 Conclusion and Future Scope

9.1 Conclusion

The development of the water purity analysis system demonstrates how machine learning can be effectively applied to address real-world challenges in public health and environmental safety. The proposed system accurately predicts water potability by analyzing a range of chemical, physical, and biological parameters using an ensemble of machine learning models. By combining multiple classifiers through a stacking approach with a logistic regression meta-model, the system enhances prediction accuracy and generalization compared to single-model solutions.

Unlike traditional water testing methods that are time-consuming and resource-intensive, this approach enables real-time, automated analysis through a lightweight and accessible web interface built using Streamlit. Users can easily input water quality data and receive instant predictions on water safety, making the system highly practical for community-level monitoring, especially in resource-limited regions.

In summary, the system offers a robust, scalable, and user-friendly framework for water quality assessment. It bridges the gap between traditional testing and modern AI-driven analysis, paving the way for future enhancements such as sensor integration, geolocation-based risk mapping, or predictive alerts. The project not only fulfills its core objectives but also showcases the transformative potential of machine learning in promoting public health and sustainable resource management.

9.2 Future Scope

The current system focuses on predicting water safety using static input values for various chemical, physical, and biological parameters. In the future, this framework can be extended by integrating real-time sensor data to enable continuous monitoring of water sources. IoT-enabled sensors can feed live data into the system, allowing for instant detection of contamination events and reducing the response time to potential health threats.

Another promising direction is the incorporation of geospatial analysis and location-based risk prediction. By mapping predictions to geographic regions, the system can identify high-risk areas and support decision-making for water management authorities. This could lead to the development of a centralized dashboard for government bodies, NGOs, or local communities, enabling better planning, resource allocation, and preventive measures in water safety efforts.

Additionally, the model can be enhanced using deep learning techniques to capture more complex patterns and relationships among water quality indicators. Features such as time-series forecasting, anomaly detection, and automated alert systems can be added to make the solution proactive rather than reactive. Furthermore, integrating mobile application support would expand accessibility, making the system useful for field engineers, researchers, and even households seeking quick water quality assessments.

10 References

[1] Nirmala Malagi, “Water portability prediction using machine learning” , IRJMETS, e-ISSN:2582-5208, volume5(2023), pgno:2779-2782.

[2] Heming Gao,Yuru Li, Handong Lu, Shuqi Zhu, "Water Portability Analysis and Prediction", AMMSAC , volume16(2022).

[3] Jatin, "Water Quality Prediction using Machine Learning", October 02, 2021.

[4] Aldhyani, T.H., Al-Yaari, M., Alkahtani, H. and Maashi, M., " Water quality prediction using artificial intelligence algorithms", Applied Bionics and Biomechanics, 2020.

[5] R.Kashefipour & R.Falconer (2002), "Water quality prediction using artificial neural networks", Journal of Environmental Management, 65(2), 185-195.

[6] Samir patel, Khushi Shah, Sakshi Vaghela, Mohmmadali Aglodiya, Rashmi Bhattad, "Water Portability Prediction Using Machine Learning", Research square, 2023.

[7] P.Y.Julien (2002), "River morphology and water quality prediction", Water Resources Research, 38(6), 1-10.

[8] M.Motagh & A.Soltanian (2017), "Groundwater quality prediction using machine learning techniques", Journal of Hydrology, 550, 108-118.

[9] S.K.Dey (2014), "Water quality prediction using decision trees", Journal of Environmental Monitoring, 16(3), 789-798.

[10] P. Wu (2007), "Water quality prediction using genetic algorithms", Journal of Water Resources Planning and Management, 133(4), 345-352.

[11] Achio, Kutsanedzie, Ameko (2016), "Comparative Analysis of Filtration Techniques for Water Purity".

[12] Muniz, Oliveira-Filho (2019,"Multivariate Statistical Analysis for Water Quality Assessment".

[13] Ahmed et al (2020), "Water Quality Monitoring: From Conventional to Emerging Technologies".

[14] K.Sreelatha, A.Nirmala Jyothsna, M.Saraswathi, P.Anusha, A. Anantha Lakshmi, "Analysis of Water Quality", IJCRT, e-ISSN: 2320-2882, Volume 10 (2022).

[15] M.Anbuchezhian, R.Venkataraman, V.Kumuthavalli, "Water Quality Analysis and Prediction using Machine Learning Algorithms", JETIR, e-ISSN: 2349-5162, Volume 5 (2018), pgno: 1966-1972.

[16] M.J.Pawari, S.M.Gavande, "Assessment of Water Quality Parameters: A Review", IJSR, e-ISSN: 2319-7064, Volume 4 (2015), pgno: 6716-6722.

[17] "AI for clean water: efficient water quality prediction leveraging machine learning", IWA Publishing, 2024

[18] "Water quality prediction using machine learning methods" Water Quality Research Journal, Vol. 53, No. 1, pp. 3-13, 2018

[19] "Reliable water quality prediction and parametric analysis using Explainable Intelligence", Scientific Reports, Vol. 14, 2024. Artificial

[20] "Drinking water potability prediction using machine learning", Water Practice & Technology, Vol. 18, No. 12, pp. 3004-3020, 2023