K. Vigneshwar, 2026, 14:3
ISSN (Online): 2348-4098
ISSN (Print): 2395-4752

International Journal of Science, Engineering and Technology
An Open Access Journal

Machine Learningapproaches for Estimating Drinking Water Safety: Assessing Human Consumption Suitability

K. Vigneshwar[1], A. Vedhika[2], B. Sai Teja[3], B. Josmitha[4],
[1]Assistant Professor, Guru Nanak Institute of Technology, CSE Department, Hyderabad
[2,3,4]Student, Guru Nanak Institute of Technology, CSE Department, Hyderabad

Abstract- Drinking Water Supply (DWS) systems are among the most essential and sensitive infrastructures required for maintaining urban life and public health across the world. In Europe, rapid population growth combined with aging and obsolete water supply infrastructure has created significant challenges in ensuring safe and continuous water distribution. Maintaining high water quality standards is critical not only for providing clean water for daily consumption but also for preventing health hazards caused by contamination. Traditional water quality monitoring methods mainly rely on periodic laboratory testing of parameters such as pH, turbidity, dissolved oxygen, and bacterial content. However, these testing procedures generally require 24–48 hours to produce results, creating a delay in identifying contamination and increasing the risk of bacterial spread within the water distribution network. To address these issues, this study proposes an Exploratory Data Analysis (EDA) based model for water quality assessment and prediction. The proposed model considers two major dimensions: water quality parameters and water quality score. Furthermore, machine learning techniques are applied to predict water quality changes within the DWS system. In this research, the Random Forest algorithm is implemented using PyCaret for efficient model development and analysis. A case study was conducted on an industrial water supply system to evaluate the model's effectiveness. The preliminary results demonstrate that the proposed approach can successfully analyze and predict water quality conditions, helping authorities improve monitoring efficiency and reduce response time to contamination risks.

Keywords: Drinking Water Supply System (DWS), Water Quality Monitoring, Exploratory Data Analysis (EDA), Water Quality Prediction, Machine Learning, Random Forest Algorithm.

I. INTRODUCTION

Water plays a vital role in everyone's life and is observed everywhere and in every form . In Today's world, due to climatic changes and pollution the water quality is been affected in areas and various experiments are done to test the quality of water . Due to poor water quality, risk occurs in the industrial areas which damage the whole environment and causes an economical loss .The root cause for many diseases such as typhoid, diarrhea, cholera is due to usage of contaminated water caused by increased industrialization and urbanization in India. . According to reports form WHO, it is estimated that about 77 million people affected by contaminated water in India and 21% of diseases are caused due to it.

Due to insufficient rainfall and drying up of main reservoirs that supplies water, India faces water crisis frequently, hence making water one of the most precious and limited land resources. Many Organizations including WHO and BIS has framed standards for water parameters that can be used to efficiently analyze the quality of water. For checking the quality of water, conventionally it required to collect water samples and send it to the lab for testing which is a tedious process. With IoT and Machine Learning algorithms it is easy to obtain the sensor values form a water sample, monitor and predict the quality of water at the comfort of our home. IOT is a buzzing technology that allows sensors to transfer data between them or to the cloud without the intervention of humans.

Water quality index of the water, which helps in determining the quality of water, can be predicted

by the extensive use of machine learning regression algorithm.

Most important infrastructures in our daily lives. According to the report from the World Health Organization (WHO) [1], there are still over 681 million people on earth struggling to receive sufficient clean water. In DWS systems, water quality is a key factor across the whole process, from the water source, treatment, and distributed pipelines. Prevalent water quality is controlled using a series of parameters. They are different from countries or regions based on geographical and development conditions. Typical water quality parameters can be divided into three groups, as physical, chemical and biological parameters. To test the parameters in practice can take from several minutes to 24 hours. The outbreaks of contagious bacteria can be much faster than the testing time and therefore

II. LITERATURE REVIEW

.Y. Amit, D. Geman, and K. Wilder proposed a model using Genetic Algorithm–Least Squares Support Vector Regression (GA-LSSVR) and Genetic Programming (GP) for predicting water quality parameters such as pH, EC, and TDS. Their study showed that the GA-LSSVR model provided higher prediction accuracy and better performance compared to traditional GP methods.

N. Mahmoudi, H. Orouji, and E. Fallah-Mehdipour introduced a hybrid approach by integrating the Shuffled Frog Leaping Algorithm (SFLA) with Support Vector Regression (SVR). The proposed SFLA-SVR model successfully predicted various water quality parameters and achieved lower RMSE values with improved efficiency when compared to Genetic Programming methods.

F.-J. Chang, Y.-H. Tsai, P.-A. Chen, A. Coynel, and G. Vachaud developed a water quality prediction model using Artificial Neural Networks (ANNs) combined with hydrological data. Their study focused on predicting NH3–N concentration in urban rivers and demonstrated that hydrological factors such as rainfall, discharge, and temperature can effectively estimate water quality conditions.

H. Rowley, S. Baluja, and T. Kanade proposed the Random Decision Forest algorithm, which improves prediction accuracy by combining multiple decision trees. Their research highlighted the advantages of Random Forest models, including reduced overfitting, better generalization, and higher robustness in classification problems.

F.A. Aziz, M. Sarosa, and E. Rohadi designed a real-time water monitoring system using sensors to measure pH, turbidity, and temperature. The system used Node MCU and Android applications for continuous monitoring and demonstrated accurate and efficient water quality observation

III. METHODOLOGY

The proposed water quality prediction system consists of several important modules that work together to monitor and predict water quality conditions effectively. Initially, water quality data such as pH, turbidity, temperature, and TDS values are collected using sensors connected to the NodeMCU microcontroller. The collected data is then transmitted wirelessly through Wi-Fi to a cloud server or database for storage and further processing. In the preprocessing stage, the dataset is cleaned by removing missing values, duplicate records, and unwanted noise to improve data quality.

After preprocessing, Exploratory Data Analysis (EDA) is performed using graphs, statistical methods, and correlation analysis to identify relationships between different water quality parameters. The Random Forest algorithm is then implemented using PyCaret to develop the prediction model. The dataset is divided into training and testing datasets, where the model is trained and evaluated to measure prediction accuracy. Based on the trained model, the system predicts water quality conditions using sensor inputs and learned patterns. Finally, the collected and predicted water quality values are displayed in real time through an Android application or web interface for continuous monitoring and analysis.

Disadvantages of Existing Systems:

- Manual Monitoring Process
- High Operational Cost
- Delay in Data Processing
- Difficulty in Handling Large Data

Proposedsystem

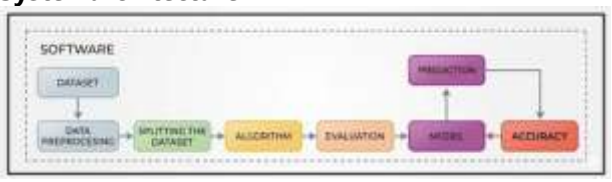
The proposed system utilizes the Random Forest algorithm implemented using PyCaret for efficient and accurate water quality prediction and assessment. Random Forest is an advanced machine learning technique based on ensemble learning, where multiple decision trees are created and combined to improve prediction accuracy and reduce overfitting problems commonly found in single decision tree models. PyCaret, a low-code machine learning library in Python, simplifies the model development process by providing automated data preprocessing, model training, parameter tuning, and performance evaluation.In this system, important water quality parameters such as pH, turbidity, temperature, total dissolved solids (TDS), electrical conductivity (EC), and other chemical components are analyzed to predict water quality conditions effectively.

The Random Forest model can handle large datasets, noisy data, and nonlinear relationships between parameters with high efficiency. Compared to traditional methods such as Support Vector Regression (SVR), the proposed system provides better prediction stability, faster processing, and improved generalization capability.The proposed technique also supports real-time monitoring and decision-making by integrating sensor-based data collection systems with wireless communication technologies. The generated predictions can help authorities identify contamination risks early and take preventive measures quickly.

Advantages Proposed System Advantages:

- Reduced Overfitting
- Automatic Data Processing
- Dependence on Human Intervention
- Better Prediction Stability

Systemarchitecture



The system architecture of the water quality prediction system begins with collecting water quality data such as pH, turbidity, temperature, and TDS values. The collected data is preprocessed to remove missing values, duplicate records, and unwanted noise for better accuracy. After preprocessing, the dataset is divided into training and testing datasets to build and validate the machine learning model. The Random Forest algorithm with PyCaret is then used to train the prediction model and improve prediction performance. Finally, the system predicts water quality conditions and evaluates the prediction accuracy to determine the effectiveness of the model.

MODULES:

1. **Data Collection :** Water quality data is collected using sensors such as pH sensor, turbidity sensor, temperature sensor, and TDS sensor. The sensors are connected to the NodeMCU microcontroller, which gathers real-time water quality readings.
2. **Data Transmission :** The collected sensor data is transmitted wirelessly through the WiFienabled NodeMCU to a database server or cloud platform for storage and processing.
3. **Data Preprocessing :** The collected dataset is cleaned and processed by removing missing values, noise, and unnecessary information. The data is then organized for machine learning analysis.
4. **Exploratory Data Analysis (EDA) :** EDA techniques are applied to understand the relationships between different water quality parameters. Graphs, statistical analysis, and correlation analysis are performed to identify patterns in the data.
5. **Model Development** : The Random Forest algorithm is implemented using PyCaret for

water quality prediction. The model is trained using historical and real-time water quality data.

6. **Model Training and Testing :** The dataset is divided into training and testing data. The Random Forest model is trained on the training dataset and evaluated using testing data to measure prediction accuracy.
7. **Quality Prediction :** The trained model predicts water quality conditions based on sensor inputs and generated patterns from the dataset.
8. **Real-Time Monitoring :** The predicted and collected water quality values are displayed in real time through an Android application or web interface for user monitoring.

IV. IMPLEMENTATION

Random Forest method using PyCaret:

PyCaret is an open-source, low-code machine learning library in Python that simplifies the complete machine learning workflow, including preprocessing, training, testing, and deployment. It is built on Scikit-learn and supports multiple machine learning algorithms with minimal coding effort. In the proposed system, the Random Forest algorithm is used through PyCaret for water quality prediction and classification. Random Forest is an ensemble learning technique that combines multiple decision trees to improve prediction accuracy and reduce overfitting. Groundwater samples are collected and analyzed to measure important physicochemical parameters such as pH, TDS, and turbidity. Based on these parameters, Water Quality Index (WQI) values are calculated to classify water as safe or unsafe for drinking. Finally, the trained Random Forest model predicts water quality conditions using the input data and generated patterns from the dataset.

V. EXPERIMENTAL RESULTS

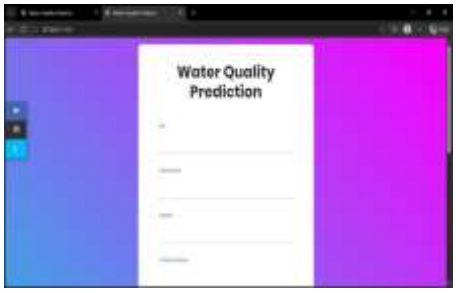


Figure. Home Page

This screenshot displays the output when the system predicts that the water is safe for human consumption. A green checkmark icon is used to visually indicate a positive result, enhancing user understanding at a glance. The pop-up notification provides clear and immediate feedback, ensuring users can quickly interpret the result. This feature improves usability by combining machine learning prediction with an intuitive visual confirmation.

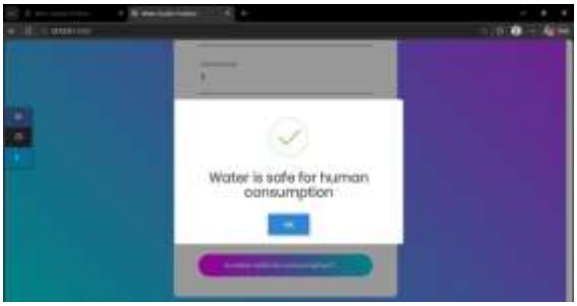


Figure. Water quality report 1

This screen shows the result when the water is predicted to be unsafe for human consumption. A red cross icon is displayed to clearly communicate a negative outcome. The alert-style popup ensures that users are immediately aware of potential risks. This helps users take necessary precautions and reinforces the system's role in promoting safe water usage through accurate prediction and clear messaging.

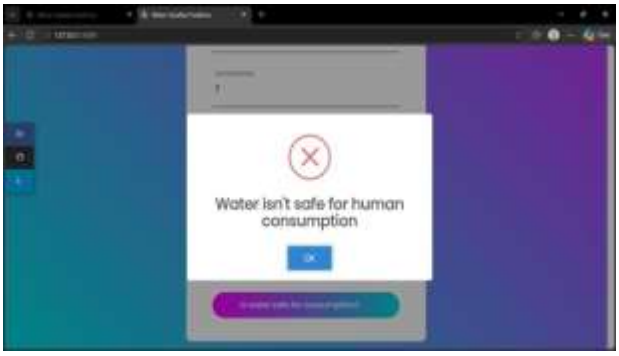


Figure. Water Quality report 2

The image presents a water quality monitoring interface displaying several important parameters and their corresponding values. The listed parameters include Solids with a value of 89, Chloramines with a value of 23, Sulfate with a value of 312, and Conductivity with a value of 124589. Each parameter is separated by horizontal divider lines, making the layout clean and organized. The interface uses a white background with black text for clarity, while a purple gradient sidebar on the right side adds a modern visual design to the display.

VII. FUTURE ENHANCEMENT

The future enhancement of the water quality prediction system can significantly improve its efficiency, accuracy, and usability. The system can be integrated with advanced IoT sensors to enable automatic real-time monitoring and collection of water quality parameters such as pH, temperature, turbidity, and TDS levels. Cloud-based technology can also be incorporated to securely store, manage, and access large volumes of water quality data from anywhere at any time. In addition, a dedicated mobile application for Android or iOS platforms can be developed to provide users with easy monitoring, instant updates, and remote access to water quality information. Furthermore, a real-time alert and notification system can be implemented to automatically warn users whenever the water quality exceeds safe or permissible limits, helping to ensure timely action and improved water safety management.

REFERENCES

1. Shamsul Haq, Shoukat Ara, Syed Maqbool Geelani and Asma Absar Bhatt, 2016, Assessment of surface and ground water for irrigational purposes, International Journal of Applied And Pure Science and Agriculture, vol.2, no. 2.
2. Manish Kumar, Gurmeet Singh, Tushara Chaminda, Pham Van Quan and Keisuke Kuroda, 2014, Emerging Water Quality Problems in Developing Countries, The Scientific World Journal . vol.2014, no.158796.
3. Roy and Ritabrata, 2018, An Introduction to water quality analysis, International Journal for Environmental Rehabilitation and Conservation, vol.2018, no. 3, pp. 94-100.
4. Subodh Kumar, Hari Mohan Meena and Kavita Verma, 2017, Water Pollution in India: Its Impact on the Human Health: Causes and Remedies, International Journal of Applied Environmental Sciences, vol.12, no. 2, pp. 275-279.
5. K Gopavanitha and S Nagaraju, 2017, A low cost system for real time water quality monitoring and controlling using IOT, International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS), DOI.10.1109/ICECDS.2017.8390054.
6. Brinda Das and P C Jain, 2017, Real-time water quality monitoring system using Internet of Things, 2017 International Conference on Computer, Communications and Electronics (Comptelix), DOI 10.1109/COMPTELIX.2017.8003942
7. qVrushali Y Kulkarni and Pradeep K Sinha, 2014, Effective Learning and Classification using Random Forest Algorithm, International Journal of Engineering and Innovative Technology (IJEIT), vol.3, no. 11, pp. 45-53.
8. Anna Bosch, Andrew Zisserman and Xavier Munoz, 2007, Image Classification using Random Forests and Ferns, IEEE International Conference on Computer Vision, DOI 10.1109/ICCV.2007.4409066
9. Hossam M Zawbaa, Maryam Hazman, Mona Abbass and Aboul Ella Hassanien, 2014, Automatic fruit classification using random forest algorithm, International Conference on

Hybrid Intelligent Systems, DOI.10.1109/HIS.2014.7086191