ISSN (Online): 2455-3662
EPRA International Journal of Multidisciplinary Research (IJMR) - Peer Reviewed Journal
Volume: 10| Issue: 4| April 2024|| Journal DOI: 10.36713/epra2013 || SJIF Impact Factor 2024: 8.402 || ISI Value: 1.188

WATER QUALITY PREDICTION WITH MACHINE LEARNING ALGORITHMS

Oliver North Rogers III[1], Ambili P S[2]
[1]*School of CSA, REVA University, Bangalore, India*
[2]*School of CSA, REVA University, Bangalore, India*

Article DOI: https://doi.org/10.36713/epra16318
DOI No: 10.36713/epra16318

ABSTRACT

Water quality prediction plays a significant role in safeguarding human health, preserving aquatic ecosystems, supporting sustainable water management practices, and ensuring regulatory compliance in aquatic environments. This study explores the use of machine learning (ML) models to predict water quality in various aquatic environments. By analyzing a comprehensive dataset of water quality indicators like pH, dissolved oxygen, and turbidity, the research employs several ML algorithms including Random Forest, Support Vector Machines, and Gradient Boosting Machines. Through rigorous training, validation, and optimization, the models are evaluated for their accuracy, sensitivity, and error rate. Additionally, the study identifies key factors impacting water quality variations through feature importance analysis. The study provides valuable insights for environmental monitoring, resource management, and regulatory compliance. Integrating advanced ML techniques with water quality assessment, this research aims to contribute to the development of effective early warning systems and decision-support tools that promote sustainable water management practices.

KEYWORDS: *Machine Learning, Water quality prediction, pH, Dissolved oxygen, Random Forest, Support Vector Machines (SVM), Gradient Boosting Machines.*

I. INTRODUCTION

In India, a considerable segment of the population, especially in rural areas, has had limited awareness regarding water quality issues. While there's an increasing focus on water-related problems, especially in urban areas facing acute water scarcity or pollution, rural regions and marginalized communities often have limited awareness of these issues. Water quality prediction models empower individuals to be more informed about the health of their water [1]. These models analyze various data points, including historical water quality records, weather patterns, land use, and pollution sources, to forecast potential changes in water quality. This information can be used to create personalized alerts and warnings, keeping individuals up to date on potential risks.

An important development in environmental science and technology is the use of machine learning algorithms to forecast water quality. Machine learning presents itself as a game-changing technology, able to handle large datasets in an effective manner and change our comprehension and forecasting of the dynamics of water quality. The intrinsic shortcomings of conventional monitoring are the source of this technological revolution. Population growth and industrialization overwhelm current approaches, and thorough analysis and real-time data are still scarce.

Regression models and complex neural networks are only two examples of machine learning algorithms that provide a comprehensive approach [3]. Beyond discrete variables like pH or temperature, they incorporate a variety of data, such as physical, chemical, and biological aspects. This thorough research opens the door for precise prediction models by illuminating the complex interactions between variables impacting water quality. Researchers and experts in water management can benefit greatly from machine learning, which offers insights beyond the constraints of conventional statistical techniques[4]. Through the identification of hidden correlations and patterns in the data, these models enable a more sophisticated understanding of the dynamics of water quality. Furthermore, predictions may be continually refined and learned from because of their inherent flexibility, which helps them remain relevant even in the face of changing environmental conditions. When dealing with dynamic components like shifting pollution sources and changing meteorological circumstances, this flexibility becomes even more crucial.

The ultimate goal remains to implement robust and sustainable water management strategies that can effectively navigate the many difficulties posed by our rapidly changing environment. To put it briefly, the creation and application of machine learning models for the prediction of water quality is a revolutionary step in the direction of guaranteeing that future generations will have access to clean and safe water. We can obtain a more thorough, effective, and timely understanding of the

ISSN (Online): 2455-3662
EPRA International Journal of Multidisciplinary Research (IJMR) - Peer Reviewed Journal
Volume: 10| Issue: 4| April 2024|| Journal DOI: 10.36713/epra2013 || SJIF Impact Factor 2024: 8.402 || ISI Value: 1.188

dynamics of water quality by utilising artificial intelligence and data analytics[18]. Thus, the door is opened for the wise and sustainable management of water resources in the future, safeguarding the welfare of people and the environment.

In conclusion, the use of water prediction models plays a crucial role in empowering individuals and communities with information crucial to making informed decisions regarding their water usage. By raising awareness, enabling proactive measures, encouraging responsible behavior, and fostering community engagement, these models become essential tools in ensuring access to clean and safe water for all.

2. LITERATURE SURVEY

Given its importance in resource management, regulatory compliance, and environmental monitoring, water quality prediction has emerged as a crucial field of study[2]. In a variety of aquatic situations, machine learning (ML) models have shown promise as tools for forecasting water quality. In their analysis of water quality indicators including pH, dissolved oxygen, and turbidity, Singh et al. [1] showed the effectiveness of machine learning techniques like Random Forest, Support Vector Machines, and Gradient Boosting Machines. Their research made clear how crucial it is to use exacting training, validation, and optimization procedures to assess the sensitivity and accuracy of machine learning models when it comes to forecasting changes in water quality.

Furthermore, Jalagam et al. expanded the use of machine learning approaches to urban streams, highlighting the necessity of customized solutions to deal with the problems presented by urban surroundings [6]. The use of Autoencoder-Long Short-Term Memory (AE-LSTM) models for water quality prediction was investigated by Zhang and Jin [5], who demonstrated how well these models capture temporal dependencies and forecast fluctuations in water quality over time. Furthermore, Chahar et al. demonstrated how deep learning methods, such as recurrent neural networks (RNNs) and convolutional neural networks (CNNs), may be used to identify complex patterns in water quality data and improve prediction accuracy [11,13,14].

In a related endeavor, Kavitha et al. examined the incorporation of ensembled machine learning models for forecasting water quality, exhibiting their capacity to enhance predictive precision and resilience through the amalgamation of numerous machine learning algorithms [8]. Tejaswi et al. underscored the significance of utilizing artificial neural networks (ANNs) to identify nonlinear correlations in water quality data, hence enabling more precise forecasts and anticipatory water management approaches [4]. Ooko et al. emphasized the value of employing machine learning techniques to forecast water quality in real time, demonstrating how ML models may provide prompt interventions and proactive management approaches [9]. Additionally, Negi et al. suggested a method based on AI and ML for forecasting water hardness, highlighting the potential of these approaches to address a range of issues related to water quality [16]. In a thorough analysis of machine learning techniques for predicting water quality, Ahmed et al. brought to light the wide variety of ML techniques and modeling strategies used in this field.[17,19]

All things considered, ML model integration into water quality prediction is a potential direction for improving environmental management and monitoring techniques. Researchers want to create efficient early warning systems and decision-support tools that assist sustainable water management practices and guarantee the supply of clean, safe water for future generations by utilizing machine learning techniques.

3. METHODOLOGY

The data science and machine learning community Kaggle offers a platform called “Kaggle Dataset Repositories” where members can find, share, and work together on datasets. It acts as a single repository for a variety of datasets that are supplied by members of the data science community and Kaggle users. "Water_potability.csv," the file we used, seemed to provide data on water quality. Your data is shaped like (3276, 10), meaning that it consists of 10 columns (features or variables) and 3276 rows (instances). Gaining knowledge of the dataset's size and structure is essential for carrying out efficient data analysis and modeling.

3.1 Method

The process of forecasting water quality entails examining many factors and trends present in water bodies. A six-step procedure may be used to explain the methodology:

- Loading the dataset
- Preprocessing the dataset
- Make use of different algorithms
- Evaluating models
- Selecting the best model
- Implementing model

Flow Chart

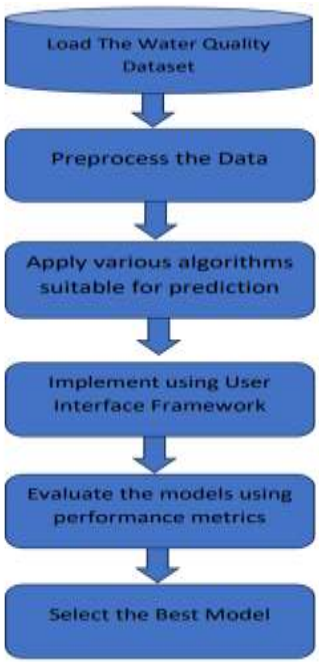

Figure 1. Process Flow

ISSN (Online): 2455-3662
EPRA International Journal of Multidisciplinary Research (IJMR) - Peer Reviewed Journal
Volume: 10| Issue: 4| April 2024|| Journal DOI: 10.36713/epra2013 || SJIF Impact Factor 2024: 8.402 || ISI Value: 1.188

The approach for predicting water quality consists of a methodical six-step procedure. First, the pertinent dataset, which contains historical data and water quality characteristics, is loaded. Preprocessing, which includes duties like resolving missing values, eliminating outliers, and standardizing features to ensure data quality and dependability, is the second stage after loading the information. The third stage then applies a range of prediction-ready algorithms, utilizing machine learning techniques to find patterns and correlations in the information. The resulting models are then thoroughly assessed using performance metrics and error measurements in the fourth stage.

The correctness of the models and their capacity to generalize to new data must be evaluated, and this assessment stage is essential. The best-performing model is chosen in the fifth stage based on its high accuracy and low error rates. To choose the most trustworthy model for predicting water quality, this step is essential. The chosen model is then implemented using a User Interface Framework in the sixth and last stage, which allows end users to interact and use the model with ease. For stakeholders, this intuitive interface provides a useful tool for accessing and interpreting water quality estimates produced by the selected model.

By combining data preprocessing, algorithm selection, model evaluation, and user interface implementation for real-world deployment, this six-step methodology guarantees an exhaustive and methodical approach to water quality prediction.

3.2 Data Models

The optimal model for predicting water quality must be chosen through a thorough review procedure that considers several algorithms. The following algorithms are used to determine which model is most appropriate:

- *Linear Regression*: Predicting a continuous outcome variable from one or more predictor variables is a simple process using the linear regression algorithm.
- *Decision Tree*: A decision tree is a model that resembles a tree in which each node reflects a choice made in response to an attribute. The dataset is partitioned recursively according to characteristics to establish the tree topology. By moving up the tree from the root to a leaf node, the prediction is formed.
- *Random Forest*: This ensemble learning technique builds many decision trees during training and outputs the mean prediction (regression) or mode of the classes (classification) of the individual trees. Using a random subset of characteristics at each split introduces unpredictability.
- *Extreme Gradient Boosting, or XGBoost*: XGBoost is a scalable and effective gradient boosting method. Sequentially, the approach constructs an ensemble of weak learners (usually decision trees), with each tree fixing the mistakes of the preceding ones. A weighted total of the forecasts made by each tree makes up the final forecast. An enhanced gradient boosting method is called XGBoost.
- *K-Nearest Neighbors (KNN)*: This non-parametric technique predicts using the average value (regression) or majority class (classification) of the k-nearest data points in the feature space. To estimate closeness, the distance metric—such as the Euclidean distance—is frequently employed. The average of the target values of the k nearest neighbors determines the forecast value (ρ) for a new data point.
- *Support Vector Machine (SVM) Regressor*: Designed for regression analysis, Support Vector Machine is a supervised learning algorithm that examines data and identifies patterns. Regression support vector machines search for the hyperplane that minimizes the difference between the expected and actual values in order to best represent the data.
- *AdaBoost Regressor*: Also known as Adaptive Boosting, AdaBoost is an ensemble learning technique that builds a strong learner by aggregating the predictions of several weak learners. The performance of each weak learner is used to determine the weights given to the data points as they are trained successively. AdaBoost builds a powerful learner by combining weak learners, usually decision trees. The weighted total of each poor learner makes up the anticipated output.
- *Artificial Neural Network (ANN) Regressor*: Layered networks of linked nodes, or neurons, make up Artificial Neural Networks[7,10]. A neural network usually consists of an input layer, hidden layers, and an output layer for regression problems[12]. The output of every neuron is determined by its activation function and weights. The output layer usually has one neuron in it.

The best method to choose will depend on the particulars of the data and the type of prediction task. Each of these algorithms has advantages and disadvantages. The basis for comprehending how each algorithm generates predictions based on input data is provided by the mathematical formulations.

3.3 System Architecture

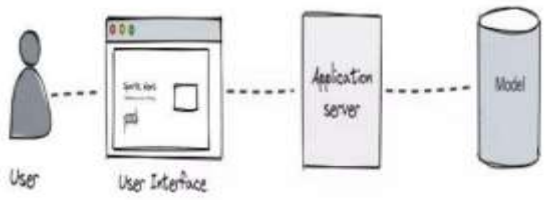

Figure 2. 2: System Architecture

- User: Using a web browser or a mobile application, the end user communicates with the web application. Requests are entered by users, who also engage with the UI and get answers from the program.

- User Interface (UI): The layer that users interact with visually is the UI. It has components that let users browse and interact with the program, such as text fields, forms, and buttons. "Streamlit," an open-source Python toolkit for building web apps for data science and machine learning, was utilized to construct the user interface.

ISSN (Online): 2455-3662
EPRA International Journal of Multidisciplinary Research (IJMR) - Peer Reviewed Journal
Volume: 10| Issue: 4| April 2024|| Journal DOI: 10.36713/epra2013 || SJIF Impact Factor 2024: 8.402 || ISI Value: 1.188

- Application Server: The application server handles the generation of dynamic content, application logic execution, and user request processing. It serves as a go-between for the database and the user interface. Python was used to write the logic on the server side.

- Machine Learning (ML) Model: Predictions based on input data are made by this module through the loading, training, and application of the ML model. Based on fresh input data, the machine learning model, which was trained on past water quality data, can forecast future water quality. The application server houses the ML model, and APIs are used to facilitate communication between the application server and the ML model.

A condensed description of how these elements work together:

- The user submits data or makes requests while interacting with the UI.
- Using the API, the UI makes queries to the application server.
- The application server handles requests and carries out required tasks, such as business logic.
- and changes or obtains information from the model.
- The processed data is returned to the user interface by the application server.
- The user sees the information shown in the UI.

4. RESULTS AND DISCUSSION

The performance parameters (MSE, RMSE, and MAE) of each regression model for water quality prediction are shown in Figure 3.

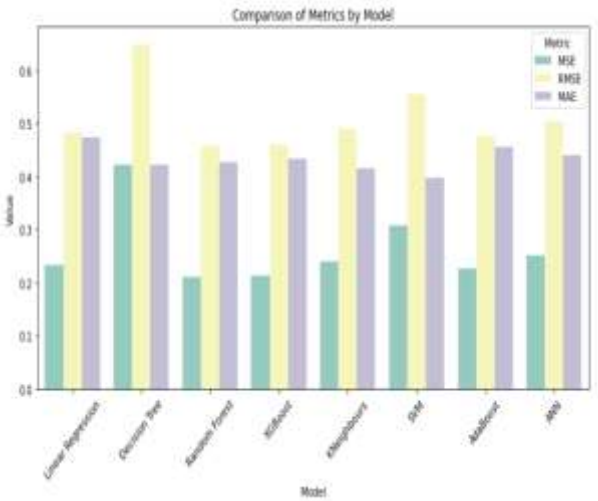

Figure 3: Performance comparison of various models

The Figure explanation:

- Model: The regression models that were trained and assessed are listed in this column.
- Mean Squared Error, or MSE, calculates the average squared difference between the values that were predicted and those that occurred. Better model performance is indicated by lower values. The average of the squared discrepancies between the expected and actual values is used to compute it. The Random Forest model is the best in minimizing squared errors because it has the lowest MSE (0.210705). With an MSE of 0.421442, the Decision Tree has the highest.
- Root Mean Squared Error (RMSE): The MSE's square root is the RMSE. It offers a meaningful measurement in the same units as the intended variable. Better model performance is indicated by lower values. Random Forest has the lowest RMSE (0.459026), similar to MSE. With an RMSE of 0.649185, Decision Tree has the highest.
- The average absolute difference between the expected and actual values is measured by the Mean Absolute Error or MAE. In contrast to MSE, it is less susceptible to outliers. Better model performance is indicated by lower values. With the lowest MAE (0.397225), SVM appears to have the least average absolute errors. The MAE of the Decision Tree is the highest (0.421442).

The precise objectives of your regression work will determine which method performs the best. As a result, our top-performing algorithm is Random Forest. SVM seems to function better if models with smaller absolute errors (MAE) are prioritized. When choosing the optimal method, it's critical to consider the characteristics of your data as well as the real-world effects of prediction mistakes. It's important to keep in mind that these metrics offer several viewpoints on model performance and that your application's particular needs may influence whether the "best" technique is selected. To make a better-informed choice, it is customary to consider a variety of metrics and maybe carry out further analysis, such as cross-validation.

To evaluate the practicality of our idea, we implemented an extensive testing plan. Two previously unpublished data sets were used: a "bad water sample" that simulated suboptimal settings and a "good water sample" that represented ideal conditions. This made it possible for us to thoroughly assess the accuracy, resilience, and responsiveness of the model in a variety of possible scenarios. This rigorous testing validates the model's applicability for real water quality prediction and guarantees its efficacy in a variety of settings.

5. CONCLUSION

The application of machine learning (ML) models to forecast water quality in aquatic situations is investigated in this work. A collection of water quality indicators, including pH, turbidity, and dissolved oxygen, is analyzed by the researchers using a variety of machine learning methods, including Random Forest, Support Vector Machines, and Gradient Boosting Machines. The models undergo extensive training, validation, and optimization processes to assess their accuracy, sensitivity, and error rate. Through feature significance analysis, the research also pinpoints important variables influencing variances in water quality.

The study emphasizes the value of machine learning in environmental science and technology as it can manage massive information efficiently and alter our perceptions of the dynamics of water quality. Conventional approaches, including population expansion and industrialization, are constrained by laborious and time-consuming laboratory experiments.

ISSN (Online): 2455-3662
EPRA International Journal of Multidisciplinary Research (IJMR) - Peer Reviewed Journal
Volume: 10| Issue: 4| April 2024|| Journal DOI: 10.36713/epra2013 || SJIF Impact Factor 2024: 8.402 || ISI Value: 1.188

Regression models and sophisticated neural networks are two examples of machine learning algorithms that offer a thorough method that considers biological, chemical, and physical factors.

To import and preprocess datasets, apply prediction-ready algorithms, assess models, choose the best model, and implement the model using a UI Framework, the study makes use of Kaggle Dataset Repositories. Regression methods are evaluated for performance using matrices like Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and Mean Square Error (MSE). Better model performance is shown by lower MSE, RMSE, and MAE values, which minimize variances between predicted and actual values.

The user interface (UI) is the layer that allows users to interact graphically in the two-tiered system architecture. Random Forest is the best method; fewer absolute mistakes are given priority. It is advised to do more analysis, such as cross-validation, before selecting the best approach.

6. REFERENCES

1. *V. Singh, N. K. Wallia, A. Kudake and A. Raj, "Water Potability Prediction Model Based on Machine Learning Techniques," 2023 World Conference on Communication & Computing (WCONF), RAIPUR, India, 2023, pp. 1-7, doi: 10.1109/WCONF58270.2023.10235096.*
2. *Ambily, P.S., Rebello, S., Jayachandran, K., Jisha, M.S., "A novel three-stage bioreactor for the effective detoxification of sodium dodecyl sulphate from wastewater", (2017) Water Science and Technology, 76 (8), pp. 2167-2176, doi: 10.2166/wst.2017.389*
3. *Ambili, P. S., and Biku Abraham. 2022. "A Predictive Model for Student Employability Using Deep Learning Techniques." ECS Transactions 107 (1): 10149, https://doi.org/10.1149/10701.10149ecst*
4. *Ambili, P. S., and Varghese Paul. "User Span Pattern: A Sequential Pattern Mining Approach for Personalization." International Journal of Applied Engineering Research 11.1 (2016): 621-624.*
5. *H. Zhang and K. Jin, "Research on water quality prediction method based on AE-LSTM," 2020 5th International Conference on Automation, Control and Robotics Engineering (CACRE), Dalian, China, 2020, pp. 602-606, doi: 10.1109/CACRE50138.2020.9230316.*
6. *L. Jalagam, N. Shepherd, J. Qi, N. Barclay and M. Smith, "Water Quality Predictions for Urban Streams Using Machine Learning," SoutheastCon 2023, Orlando, FL, USA, 2023, pp. 217-223, doi: 10.1109/SoutheastCon51012.2023.10115154.*
7. *T. Tejaswi, C. Manoj, P. Venkata Daivakeshwar Naidu, T. Santhosh, P. Venkata Sai Akhil and V. Ganesan, "Nexus of Water Quality prediction by ANN," 2022 International Conference on Innovative Computing, Intelligent Communication and Smart Electrical Systems (ICSES), Chennai, India, 2022, pp. 1-5, doi: 10.1109/ICSES55317.2022.9914054.*
8. *N. S. Kavitha, M. Sakthivel, B. Sreedevi and G. Revathy, "Water Quality Prediction Using Ensembled Machine Learning," 2023 International Conference on Sustainable Communication Networks and Application (ICSCNA), Theni, India, 2023, pp. 975-979, doi: 10.1109/ICSCNA58489.2023.10370362.*
9. *S. O. Ooko, E. K. Pamela and G. Kwagalakwe, "Use of Machine Learning for Realtime Water Quality Prediction," 2023 IEEE AFRICON, Nairobi, Kenya, 2023, pp. 1-6, doi: 10.1109/AFRICON55910.2023.10293701.*
10. *L. Guo and D. Fu, "River Water Quality Prediction Model Based on PCA-APSO-ELM Neural Network," 2023 6th International Conference on Artificial Intelligence and Big Data (ICAIBD), Chengdu, China, 2023, pp. 512-517, doi: 10.1109/ICAIBD57115.2023.10206249.*
11. *Chahar, A. Chowdhury, B. K. Thulasidoss, P. V. Reddy, H. Patel and N. Patil, "Water Quality Analysis Using Deep Learning," 2022 8th International Conference on Advanced Computing and Communication Systems (ICACCS), Coimbatore, India, 2022, pp. 423-426, doi: 10.1109/ICACCS54159.2022.9785189.*
12. *Ambili P S, Agnesh L, & Arun K V. (2023), "Siamese Neural Network Model for Recognizing Optically Processed Devanagari Hindi Script", International Journal of Computational Learning & Intelligence, 2(3), 107–113, https://doi.org/10.5281/zenodo.8210372.*
13. *Mittal, S. Patwal, M. Adhikari and M. Manu, "A Review of Various Water Quality Prediction Models and Techniques," 2023 5th International Conference on Inventive Research in Computing Applications (ICIRCA), Coimbatore, India, 2023, pp. 614-620, doi: 10.1109/ICIRCA57980.2023.10220687.*
14. *Pillai, A.P.S. (2023), "AIoMT-Assisted Telemedicine A Case Study of eSanjeevani Telemedicine Service in India", Handbook of Security and Privacy of AI-Enabled Healthcare Systems and Internet of Medical Things, pp. 445–464, DOI: 10.1201/9781003370321-19.*
15. *S. Babu, B. B. Nagaleela, C. G. Karthik and L. N. Yepuri, "Water Quality Prediction using Neural Networks," 2023 International Conference on Artificial Intelligence and Knowledge Discovery in Concurrent Engineering (ICECONF), Chennai, India, 2023, pp. 1-6, doi: 10.1109/ICECONF57129.2023.10084120.*
16. *P. B. Negi et al., "AI and ML based Prediction of Water Hardness," 2022 2nd International Conference on Intelligent Technologies (CONIT), Hubli, India, 2022, pp. 1-5, doi: 10.1109/CONIT55038.2022.9848161.*
17. *Ali Najah Ahmed, Faridah Binti Othman, Haitham Abdulmohsin Afan, Rusul Khaleel Ibrahim, Chow Ming Fai, Md Shabbir Hossain, Mohammad Ehteram, Ahmed Elshafie, Machine learning methods for better water quality prediction, Journal of Hydrology, Volume 578, 2019, 124084, ISSN 0022-1694, https://doi.org/10.1016/j.jhydrol.2019.124084.*
18. *Mahapatra, S.S., Sahu, M., Patel, R.K. et al., "Prediction of Water Quality Using Principal Component Analysis", Water Qual Expo Health 4, 93–104 (2012).*
19. *https://doi.org/10.1007/s12403-012-0068-9*
20. *Liming Zhang, Haowen Yan, "Implementation of a GIS-based water quality standards syntaxis and basin water quality prediction system," 2012 International Symposium on Geomatics for Integrated Water Resource Management, Lanzhou, 2012, pp. 1-4, doi: 10.1109/GIWRM.2012.6349656.*