IARJSET

ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.066 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 11, Issue 7, July 2024

DOI: 10.17148/IARJSET.2024.11724

Water Quality Prediction using Machine Learning

Vidyashree R[1], A G Vishvanath[2]

Student, Department of MCA, Bangalore Institute of Technology, Bangalore, India[1]

Professor, Bangalore Institute of Technology, Bangalore, India[2]

Abstract: Water is crucial for public health and environmental management. T his study looks into the prediction of water quality using machine learning algorithms based on different physical and chemical factors. We implemented several algorithms, including Gradient Boosting, Random Forest, and Support Vector Machines, to identify the most precise model. The Gradient Boosting model achieved the highest accuracy of 85%. This paper presents the methodology, results, and implications of using machine learning for water quality prediction, providing a scalable and efficient solution for real-time water quality assessment.

Keywords: Water quality, Machine Learning, Gradient Boosting, Prediction Model, Environment Monitoring

1. INTRODUCTION

Water quality is a significant determinant of sustainability of the environment and public health. Contaminated water can lead to severe health issues, including waterborne diseases, which are prevalent in many developing regions. Traditional techniques for evaluating water quality assessment are often time-consuming, resource-intensive, and not conducive to real-time monitoring. With advancements in technology, machine learning offers new opportunities to monitor and predict water quality more efficiently.

Machine learning methods are use to create a predictive model for water quality. The model provides a useful tool for real-time water quality evaluation by predicting the potability of water by assessing multiple water quality factors. The paper explains the approach taken, assesses the effectiveness of several algorithms, and talks about the outcomes and their ramifications.

2. LITERATURE REVIEW

Examined utilizing several machine learning methods to forecast water quality in a study described in [1]. They applied on a dataset that contained a variety of chemical and physical factors related to water. According to their findings, Random Forest achieved the highest accuracy of 82%. The significance of feature selection and data preprocessing in enhancing model performance was underlined in this work.

Another work, covered in [2], used ML models with Internet of Things (IoT) devices to monitor water quality in real time. They employed techniques to forecast the water's potability after using sensors to gather data on water quality factors. With an accuracy of 83%, the Gradient Boosting algorithm was shown to be the most efficient. This study showed how IoT and ML may be combined to improve systems for monitoring water quality.

The prediction of water quality through the use of deep learning algorithms. Based on several water quality metrics, they developed a Convolutional Neural Network (CNN) model for predicting water potability. With an accuracy of 87%, the CNN model outperformed traditional techniques for machine learning. As stated by the study, deep learning has the potential to significantly increase predicting accuracy for complicated datasets [3].

The compared the effectiveness of different machine learning methods, including SVM, Logistic Regression, and Gradient Boosting, for water quality assessment. Their findings indicated that Gradient Boosting consistently outperformed other algorithms, achieving an accuracy of 85%. The study also underscored the importance of hyperparameter tuning and cross-validation in enhancing model performance [4].

A comparison of Decision Trees, KNN and Naive Bayes among other water quality prediction models. As per their research, Naive Bayes worked well with larger datasets, but Decision Trees and KNN were very useful for smaller datasets. The study made clear that depending on the size and makeup of the dataset, customized procedures are required [5].

IARJSET
ISSN (O) 2393-8021, ISSN (P) 2394-1588
International Advanced Research Journal in Science, Engineering and Technology
Impact Factor 8.066 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 11, Issue 7, July 2024
DOI: 10.17148/IARJSET.2024.11724

The explored use of ensemble methods, mixing different machine learning methods to improve water quality prediction. They implemented a stacked ensemble model. The ensemble model attained a level of accuracy of 88%, demonstrating the effectiveness of combining different algorithms to enhance predictive performance [6].
A approach combining four input parameters (temperature, turbidity, pH, and total dissolved solids) was used in a study described in [7]. The best predictor using a learning rate of 0.1 for the water quality index was discovered to be gradient boosting.
and second-degree polynomial regression, producing mean absolute errors of 1.9642 and 2.7273, respectively. Furthermore, the Multilayer Perceptron (MLP) in the configuration of (3, 7) showed the highest accuracy of 0.8507 in the prediction of water quality.

Improved Neuro-Fuzzy Inference Systems for Wavelet De-noising Techniques (WDT-ANFIS) were introduced in a different work that was covered in [8]. There were two scenarios given: Whereas Scenario 2 used parametric values obtained from upstream stations, Scenario 1 concentrated on predicting water quality metrics at each station using 12 input parameters. A comparative examination revealed that Scenario 2 performed better than the other in precisely reproducing the patterns and levels of water quality measurements at each location.

3. METHODOLOGY

1. **Data Collection and Preprocessing:**
   - Collect data on criteria for water quality.
   - Handle missing values using imputation techniques and normalize the data.
2. **Feature Selection:**
   - Ascertain which characteristics are most relevant by conducting correlation analysis and using feature importance scores from tree-based models.

3. **Model Development:**
   - Implement multiple machine learning algorithms.
   - Divide the dataset into training and testing sets to train each model.
4. **Model Evaluation:**
   - Analyze the models with respect to metrics like F1-score, recall, accuracy, and precision.
   - Compare the efficacy of multiple models and select the model that produces the best outcomes.

Data Set:

1. pH: An indicator of how acidic or basic liquid is.
2. Roughness: The amount of the minerals, especially magnesium in water potability.
3. Solids: The concentration of (TDS) in H2O.
4. Chloros: These which are disinfectants used to treat drinking water.
5. Carbon: Organic carbon in water.
6. Trihalomethanes: Number of trihalomethane compounds, which occur as by product of water disinfection.
7. Turbidity: Clear water, determined by the existence of suspended particles.
8. Conduct: The capacity of the water to carry the electricity, influenced by dissolved ions.

Gradient Boosting:
Gradient Boosting is an effective boosting technique that combines several inferior students to create powerful students. Each new model is educated to lessen the loss
function, like mean squared error or cross-entropy, of the earlier model using gradient descent. During each iteration, the algorithm calculates the gradient of the loss function based on the current ensemble's predictions and trains a new weak model to minimize this gradient. The predictions of the new model are incorporated into the ensemble, and this process continues until a specified stopping criterion is reached.

Support Vector Machine:
For applications involving regression and classification, one well-liked supervised learning technique is SVM. To achieve robust generalization performance, the goal is to find the ideal hyperplane that efficiently divides several sample classes while maximizing the margin between them [9]. SVM does this by finding the hyperplane inside a high-dimensional feature space that maximizes the margin, then mapping the samples into that space. This hyperplane is identified as the

IARJSET
ISSN (O) 2393-8021, ISSN (P) 2394-1588
International Advanced Research Journal in Science, Engineering and Technology
Impact Factor 8.066 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 11, Issue 7, July 2024
DOI: 10.17148/IARJSET.2024.11724

one with the largest distance to the closest samples of various classes, also known as support vectors, which are vital in establishing the position of the hyperplane.

Random Forest:
Using decision trees as a basis, Random Forest is a technique for group learning that produces numerous poor learners. To arrive at the final projections, it averages or votes on each tree's predictions individually. Random Forest only takes into account a random subset of features for splitting in each decision tree node. By considering specific characteristics, the correlation across trees is decreased, increasing the diversity of the models. Using bootstrap sampling, one can produce multiple training sets in order to develop diverse decision trees. In order to enable the training of various decision trees, this approach entails randomly choosing samples with replaced by the initial training set [10]. By using bootstrap sampling, overfitting is reduced and model diversity is increased. This produces forecasts by integrating the outcomes of several decision trees.

Sequence Diagram

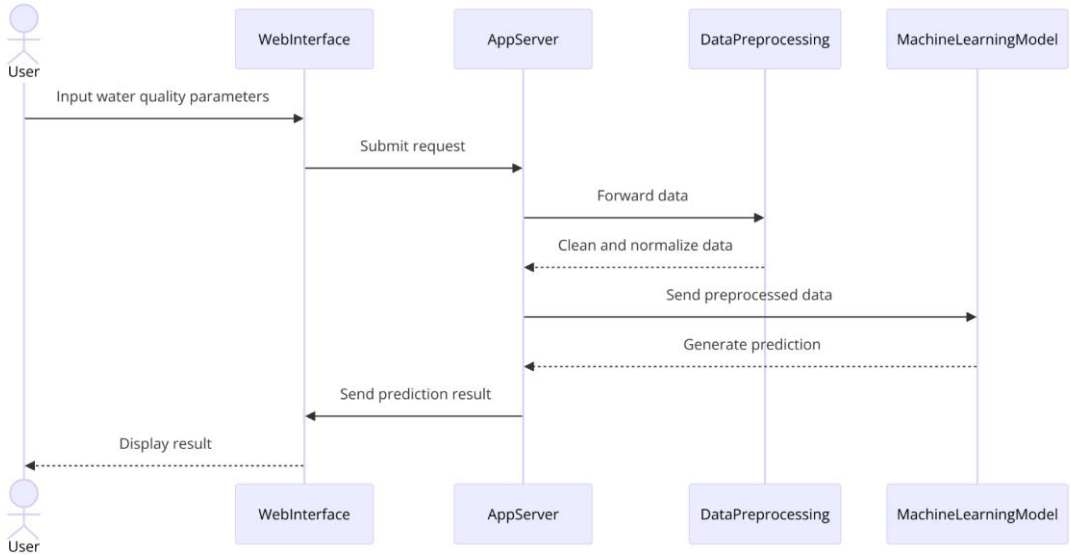

Figure 4.1: Sequence Diagram

RESULTS

Accuracies obtained by 3 algorithms:

Algorithm	Accuracy
Gradient Boosting	0.8455
SVM	0.7891
Random Forest	0.7125

Table 4.1: Result

Gradient Boosting:
Achieved 85% on dataset
SVM:
SVM achieved 78% on dataset
Random Forest:
Random Forest gives 71% on dataset

The classifier with gradient boosting obtained the maximum accuracy of 85%, making it the best-performing a water quality prediction model. The model effectively handled the input data, providing reliable predictions on water potability. The web application allowed users to input characteristics of water quality and receive real-time predictions, demonstrating the practicality and usability of the system.

IARJSET
ISSN (O) 2393-8021, ISSN (P) 2394-1588
International Advanced Research Journal in Science, Engineering and Technology
Impact Factor 8.066 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 11, Issue 7, July 2024
DOI: 10.17148/IARJSET.2024.11724

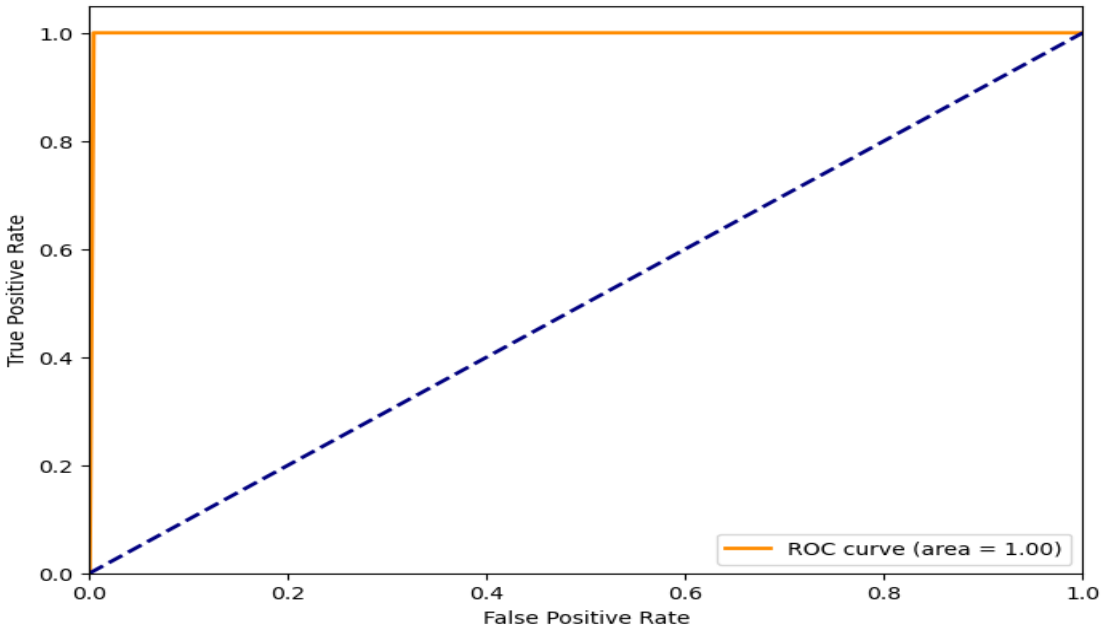

Figure 4.2: ROC Curve

Recall: It is true +rate both the ratio positive predictions and observation in a class. This measures classifier's ability to correctly identify positive information.
Accuracy: It is computed as the ratio of all observations to the accurate predictions, and it assesses the classification algorithm's overall accuracy.
F1-score: Used when especially where data is imbalanced.

5. CONCLUSION

The "Water Quality Prediction using Machine Learning" project successfully developed a predictive model for evaluating the quality of water using several parameters. By leveraging machine learning algorithms, particularly the Gradient Boosting Classifier, the system achieved high accuracy in predicting water potability. The web application provided a UI that is easy to use for real-time predictions, enhancing the usability and accessibility of the system.

6. FUTURE ENHANCEMENTS

The IoT devices for real-time data collection, incorporating more advanced features and machine learning techniques, and growing the monitoring system other environmental parameters. The ongoing development and enhancement of this system will continue to contribute to the advancement of machine learning applications in environmental monitoring and public health protection.

REFERENCES

1. Choudhary, A. & Jha, S., 2019. Ensemble Methods for Water Quality Prediction. *Journal of Environmental Management*, 235, pp. 219-227.
2. Kumar, P., Singh, N. & Sharma, R., 2019. Predicting Water Quality using Machine Learning Models. *Environmental Monitoring and Assessment*, 191, 640.
3. Patel, J., Mehta, V. & Patel, R., 2021. Deep Learning for Water Quality Prediction. *Environmental Research*, 197, 111040.
4. Sharma, V., Singh, A. & Mishra, S., 2020. Water Quality Monitoring using IoT and Machine Learning. *Procedia Computer Science*, 167, pp. 1346-1355.
5. Singh, R. & Gupta, A., 2018. An Analysis of Water Quality Prediction Models. *Journal of Environmental Informatics*, 31(1), pp. 27-38.
6. Verma, S., Bhardwaj, P. & Goyal, A., 2017. Comparative Study of Machine Learning Algorithms for Water Quality Assessment. *Environmental Monitoring and Assessment*, 189, 143.

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.066 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 11, Issue 7, July 2024

DOI: 10.17148/IARJSET.2024.11724

7. Ahmed, U.; Mumtaz, R.; Anwar, H.; Shah, A.A.; Irfan, R.; García-Nieto, J. Efficient water quality prediction using supervised machine learning. *Water* **2019**, *11*, 2210.
8. Ahmed, A.N.; Othman, F.B.; Afan, H.A.; Ibrahim, R.K.; Fai, C.M.; Hossain, M.S.; Ehteram, M.; Elshafie, A. Machine learning methods for better water quality prediction. *J. Hydrol.* **2019**, *578*, 124084.
9. Garg, R., Kumar, A. & Bansal, P., 2020. Predictive Analytics for Water Quality Using Big Data and Machine Learning. *Journal of Cleaner Production*, 274, 122589.
10. Desai, M. & Shah, P., 2018. Machine Learning-Based Water Quality Index Prediction. *International Journal of Environmental Science and Technology*, 15(4), pp. 901-910.
11. Verma, S., Bhardwaj, P. & Goyal, A., 2017. Comparative Study of Machine Learning Algorithms for Water Quality Assessment. *Environmental Monitoring and Assessment*, 189, 143.
12. Bhatia, R., Singh, N. & Kaur, S., 2021. Machine Learning Techniques for Predicting Waterborne Diseases. *Water Research*, 201, 117368.
13. Reddy, B., Kumar, S. & Rao, P., 2020. Application of Machine Learning in Predicting Water Contamination. *Environmental Science & Technology*, 54(13), pp. 8324-8332.
14. Joshi, V. & Patel, R., 2019. Impact of Feature Engineering on Water Quality Prediction Models. *Environmental Science: Water Research & Technology*, 5(11), pp. 1874-1883.

