JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Predictive Modelling of Urban Water Quality

Using Machine Learning

[1]Dr. Suma Swamy,[2] Siddhi Narayan,[3]Dharmitha V, [4]Y. Yasaswini

[1]Professor, Department of CSE, Sir M Visvesvaraya Institute of Technology, VTU, Bengaluru, India
[2,3,4] B.E Student, Department of CSE, Sir M Visvesvaraya Institute of Technology, VTU, Bengaluru, India

Abstract: Urban water bodies face ongoing threats from pollution, necessitating advanced monitoring and prediction strategies to ensure timely intervention and resource preservation. This study explores the application of machine learning techniques for predictive modeling of water quality parameters in urban environments. Leveraging sensor data, satellite imagery, and historical records, the proposed models aim to forecast crucial indicators such as pH levels, dissolved oxygen, turbidity, and pollutant concentrations. The research emphasizes the development of accurate and robust machine learning algorithms capable of real-time monitoring, enabling early detection of potential contamination events. By integrating predictive analytics into water quality management systems, this approach facilitates proactive decision-making, contributing to the sustainable preservation of urban water resources. The findings hold significant implications for environmental monitoring, policy formulation, and the creation of smart city infrastructures aimed at safeguarding water quality in the face of growing urbanization and environmental challenges.

***IndexTerms* - Urban water bodies, pollution, machine learning, predictive modelling, sensor data, satellite imagery, real-time monitoring, water quality management, smart city, environmental challenges.**

I. INTRODUCTION:

Water quality is a critical aspect of environmental health and human well-being, as it directly impacts ecosystems, agriculture, and human consumption. With the increasing pressures on water resources due to population growth, industrialization, and climate change, ensuring water quality has become a paramount concern. Traditional methods of water quality monitoring involve manual sampling and laboratory analysis, which are often time-consuming and costly.

In recent years, there has been a growing interest in leveraging machine learning (ML) techniques to predict and monitor water quality more efficiently. ML algorithms can analyse large datasets containing diverse water quality parameters, environmental variables, and historical records to make accurate predictions about water quality conditions. This approach offers the advantage of real-time or near-real-time monitoring, allowing for timely interventions and proactive management of water resources.

The utilization of machine learning (ML) for water quality prediction offers a plethora of benefits that address the challenges inherent in traditional monitoring methods. Firstly, ML algorithms enable the early detection of contamination events, providing a crucial advantage in swiftly implementing preventive measures to safeguard public health and environmental integrity. Moreover, the predictive modeling capabilities of ML empower stakeholders to anticipate changes in water quality by analyzing historical data alongside relevant environmental variables. This not only aids in proactive management but also facilitates optimized resource allocation, ensuring that interventions are strategically deployed where they are most needed.

II. RESEARCH OBJECTIVE

Development of an Integrated Machine Learning Model: The primary goal is to engineer a sophisticated machine learning model capable of integrating and processing diverse environmental parameters such as pH, dissolved oxygen, biochemical oxygen demand, total suspended solids, and nitrate-nitrogen. This model will be designed to handle real-time data variations, aiming to deliver robust and accurate predictions essential for efficient water resource management.

Implementation of a Water Quality Classification System: The research will also focus on establishing a reliable classification system that can discern whether water is potable or not based on its environmental and chemical characteristics. This system will incorporate advanced data preprocessing techniques to enhance the accuracy of the classification, providing a dependable tool for environmental agencies and policymakers.

Development of Data Visualization Modules: Another critical objective involves creating intuitive data visualization modules to effectively communicate the complexities of water quality data. These visualizations will assist in making the data comprehensible to various stakeholders, thereby supporting improved decision-making and increasing public awareness about water quality challenges.

III. LITERATURE REVIEW

In 2021, Ali Omram Al-Sulttani [1], studied and developed ensemble machine learning models to predict BOD values in the Euphrates River, Iraq, using feature extraction techniques like Genetic Algorithm and PCA. It focused on Rohri Canal, highlighting the need for broader geographical representation. The study's limitations include neglecting temporal/spatial variability and a static dataset. Real-time data integration is suggested for proactive water management in smart cities, showing promise for future research in water quality prediction.

In 2023, Sarang Karim [2], proposed an IoT and ML framework for water quality analysis in Rohri Canal, Pakistan, using MLP for regression and Random Forest for classification to predict WQI and WQC. Results favor MLP for regression and Random Forest for classification, showing better performance with smaller datasets. Limitations include a short two-year dataset and the absence of a climate change variable. Future research should consider broader datasets, additional metrics, and advanced models for enhanced analysis.

In 2022, Bilal Aslam [3], aimed to enhance Water Quality Index predictions using machine learning on well water samples from North Pakistan. Hybrid algorithms outperformed standalone ones. Future work could explore algorithm performance over extended periods, include important parameters like COD and BOD over multiple years, consider deep learning algorithms like CNNs, conduct PCA tests, and analyse additional water quality variables for a robust study.

In 2023, Mahmoud Y Shams [4] study, focused on water quality prediction using machine learning, optimizing models like RF, Xgboost, GB, etc., through grid search. GB achieves 99.50% accuracy in classification (WQC), and MLP excels with a 99.8% R2 value in regression (WQI). Limitations include dataset specificity, potential parameter sensitivity oversight, limited assessment metrics, and absence of uncertainty analysis. Future research could explore broader datasets, additional metrics, and thorough sensitivity analysis for enhanced model performance evaluation.

In 2022, Mohamed Abbas [5], introduced AdCSO-sELM, enhancing ELM performance with dynamic parameter adjustment. Achieved 96.54% accuracy in classifying water potability with Kaggle dataset. Limitations include data dependency, dataset reliance, and generalizability issues. Novel dynamic adaptation may increase complexity and computational load. Performance may vary across environments, requiring extensive parameter tuning.

In 2020, Salwani Abdullah [6], inspired by the human body's remarkable waste filtration system, researchers have introduced a new optimization algorithm called Kidney-Inspired (KA). This algorithm mimics the way healthy kidneys work, iteratively evaluating potential solutions and discarding less favorable options, much like how kidneys filter waste products from the blood. The study demonstrates KA's potential in real-world applications like water quality prediction and cancer detection, suggesting its competency in handling complex tasks. However, the research also acknowledges limitations in a simplified version of KA known as Dual-KA. While Dual-KA streamlines the optimization process, it faces challenges. It may not fully capture the intricate biological processes of real kidneys, limiting its direct biological relevance. Additionally, Dual-KA might be more effective for specific types of problems, hindering its broader applicability. Furthermore, understanding how Dual-KA arrives at its decisions can be complex, and its ability to handle large-scale problems remains to be fully explored.

In 2013, Xuan Zou [7], proposed a new method for predicting water quality changes over time. It combines multiple prediction models (regressors) with a decision-maker (classifier). The data is divided into sections based on which model works best for that specific part. The classifier then directs new data to the most effective model, allowing the system to handle complex variations in water quality. While promising, the approach has limitations. The model's accuracy relies heavily on the quality and completeness of the training data. Additionally, the combination of multiple models increases training time and makes it more difficult to understand how the model arrives at its predictions.

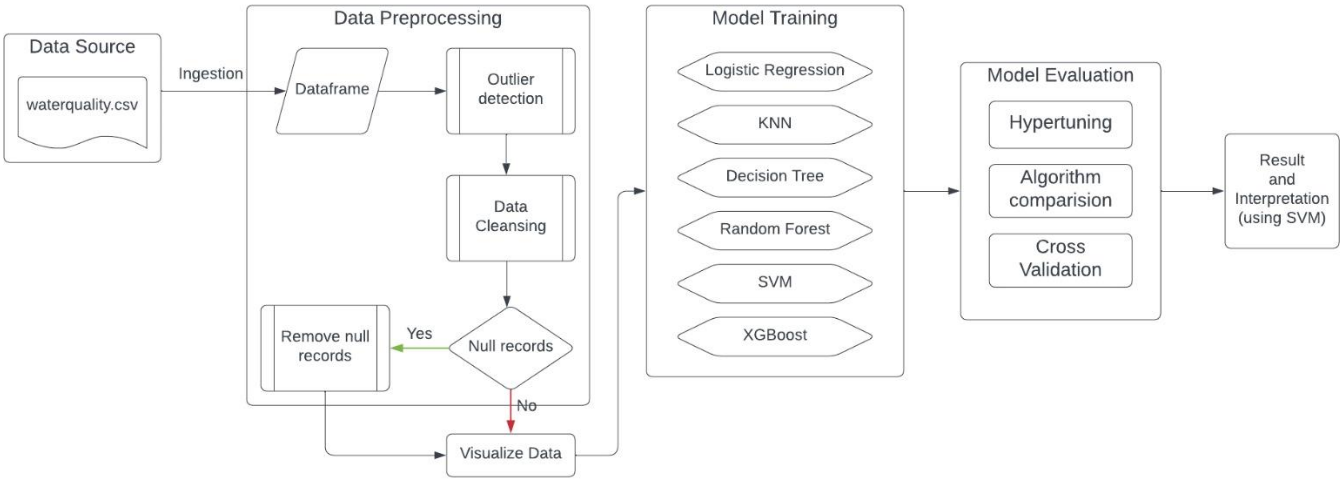
In 2020, Bharat B. Gulyani [8], traditionally, measured Biochemical Oxygen Demand (BOD) in water is time-consuming and expensive. This research proposes a data-driven machine learning model for faster and cheaper BOD prediction. The model utilizes dimensionality reduction techniques to analyze only the most relevant data, improving efficiency. However, the authors acknowledge that the effectiveness of specific techniques might vary and call for further investigation into potential challenges like overfitting and outlier sensitivity. Additionally, they emphasize the need for more comprehensive evaluation metrics to assess the model's generalizability and real-world performance.

In 2009, Tianyou Cha [9], research introduced a new method using a mechanism model and hierarchical neural networks to predict water quality in wastewater treatment plants. This soft measurement approach aims to be more cost-effective than current online sensors. The hierarchical structure mimics the cascaded stages of the treatment process, potentially improving prediction accuracy within the reactors. While the study shows promise, the authors acknowledge limitations. Implementing the method could be complex, and its effectiveness relies on high-quality data. Additionally, adapting the model to real-time changes and different plant configurations might be challenging. Further research is needed to address these limitations before widespread adoption in wastewater treatment facilities.

In 2022, K. P. Rasheed [10], research introduced new models for predicting water quality in aquaculture settings. The models combine the strengths of Convolutional Neural Networks (CNNs) for capturing water quality characteristics and Long Short-Term Memory (LSTM) or Gated Recurrent Unit (GRU) networks for handling long-term trends in water data. These models achieved good results on test datasets. However, the study acknowledges limitations. The models might not adapt easily to different aquaculture systems due to variations in water quality factors. Additionally, they rely on specific settings (hyperparameters) that may need adjustments for new data or environments. Finally, while the most successful model (CNN-LSTM) performs well, its complex calculations could be challenging for real-time use in aquaculture management.

.

. IV. PROPOSED SYSTEM:



Data Preprocessing: It is a crucial step in preparing raw data for analysis and modeling. It involves creating a data frame from a CSV file and performing data cleansing, which includes handling outliers and null records.

- Outlier Detection: The data is scanned for outliers, which are data points that fall outside the expected range.
- Data Cleansing: This step involves removing or correcting any errors in the data, such as null values. Here, the process checks for null records and removes them if there are any.

Data Visualization: The distributions of each dimension in the dataset were visualized using histograms with KDE plots, while outliers were identified through box plots. Additionally, a heat map was employed to illustrate the correlations among all dimensions, providing a comprehensive overview of the data's relationships and anomalies.

Model Training: Various classification models, including Logistic Regression, Decision Tree, Random Forest, Support Vector Machine, XGBoost and K-Nearest Neighbors, were trained on the water quality dataset. Their accuracies were evaluated, and confusion matrices were used to summarize the performance of each model, providing insights into their predictive accuracy and classification performance.

Model evaluation: It is the process of using different evaluation metrics to understand a machine learning model's performance, as well as its strengths and weaknesses.

- Hyper tuning: Training your model sequentially with different sets of hyperparameters.
- Cross-validation: Evaluating models by training several models on subsets of the available input data and evaluating them on the complementary subset of the data.

Result: Post prediction analysis is discussed to finalize our results and review any recommendations with our final conclusions with intent to deploy a robust water quality classifier.

V. RESULTS AND CONCLUSIONS:

In the first iteration of modeling, the algorithms were all run just the default parameters of their respective functions. The confusion matrices of the algorithms are as follows:

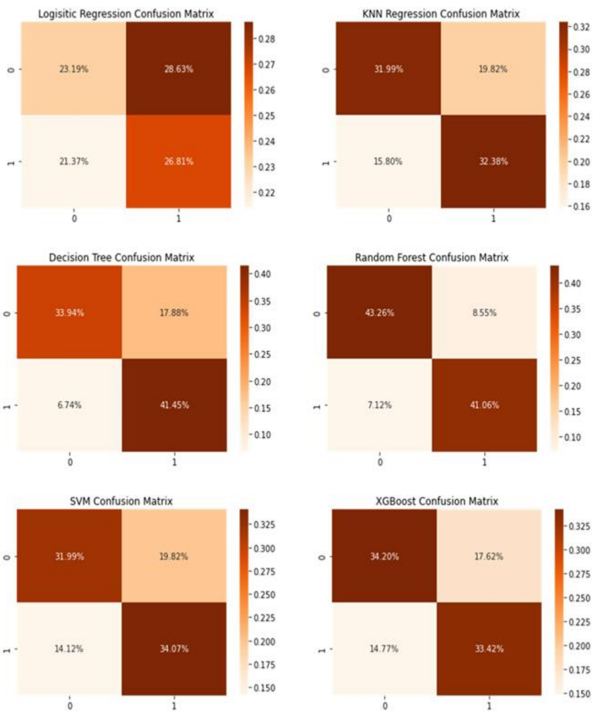


Fig A: Confusion Matrix of each algorithm after the first iteration of modeling

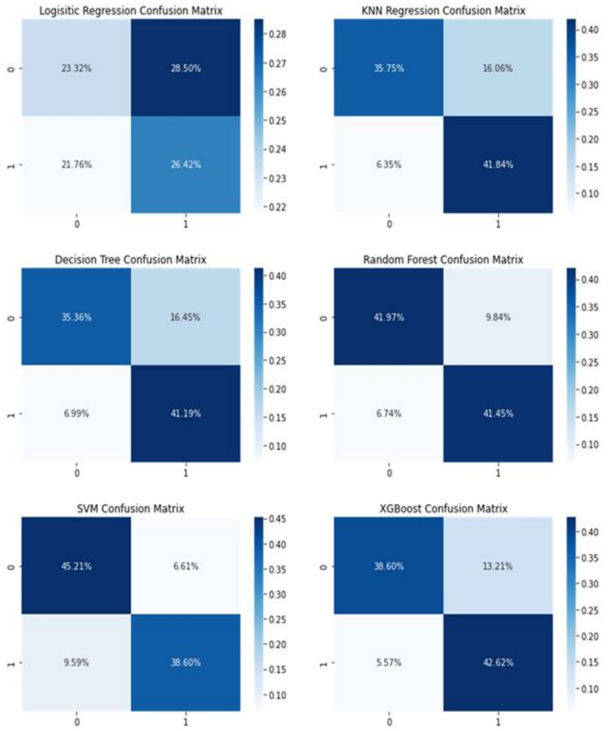
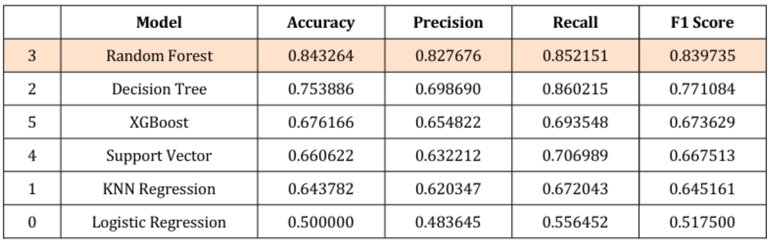


Fig B: Confusion Matrix of each algorithm after the second iteration of modeling.

Statistically, type II errors would be considered more hazardous in particular for assessing water quality or potability. Returning a classification of a false positive would be detrimental to a community consuming unclean water. In reviewing the confusion matrices, the highest type II error occurs in Logistic Regression at 21.37% while the lowest type II error occurs in the Decision Tree algorithm at 7.12%.



	Model	Accuracy	Precision	Recall	F1 Score
3	Random Forest	0.843264	0.827676	0.852151	0.839735
2	Decision Tree	0.753886	0.698690	0.860215	0.771084
5	XGBoost	0.676166	0.654822	0.693548	0.673629
4	Support Vector	0.660622	0.632212	0.706989	0.667513
1	KNN Regression	0.643782	0.620347	0.672043	0.645161
0	Logistic Regression	0.500000	0.483645	0.556452	0.517500

Table 1:Evaluation metrics from the first iteration of modeling.

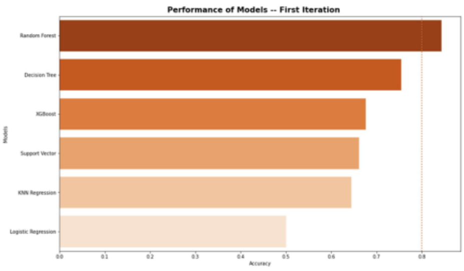
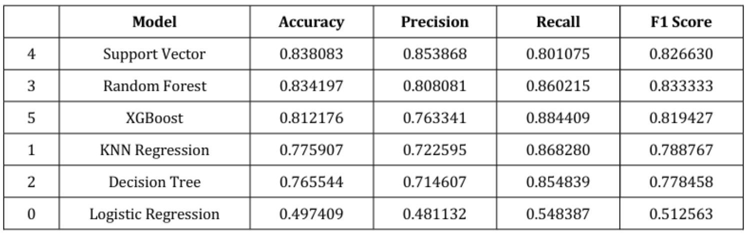


Fig C: Performance of each algorithm after the first iteration of modeling by accuracy.

After reviewing the evaluation metrics of each algorithm for first iteration of modeling, we observed that Random Forest algorithm performed the best with an accuracy of 84.33% while Logistic Regression performed the worst at 50.00%.



	Model	Accuracy	Precision	Recall	F1 Score
4	Support Vector	0.838083	0.853868	0.801075	0.826630
3	Random Forest	0.834197	0.808081	0.860215	0.833333
5	XGBoost	0.812176	0.763341	0.884409	0.819427
1	KNN Regression	0.775907	0.722595	0.868280	0.788767
2	Decision Tree	0.765544	0.714607	0.854839	0.778458
0	Logistic Regression	0.497409	0.481132	0.548387	0.512563

Table 2:Evaluation metrics from the Second iteration of modeling.

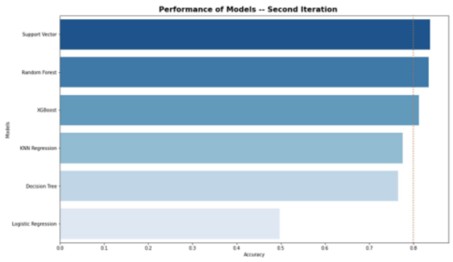
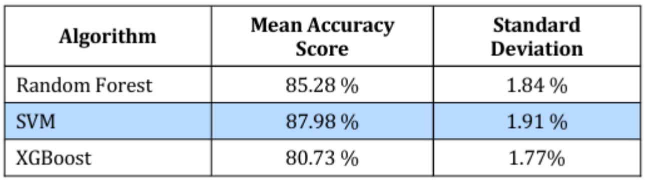


Fig D: Performance of each algorithm after the second iteration of modeling by accuracy.

Looking at the evaluation metrics after the second iteration we can see that hyper-tunning greatly increased the performance of nearly all the algorithms. SVM and Random Forest performed the best with the highest accuracy of 83.81% and 83.42% respectively. Logistic Regression remained the lowest performing algorithm with minute changes in overall performance.



Algorithm	Mean Accuracy Score	Standard Deviation
Random Forest	85.28 %	1.84 %
SVM	87.98 %	1.91 %
XGBoost	80.73 %	1.77%

Table 3: Results of K-Fold Cross Validation.

After the second iteration of model training, we selected the top three algorithms to apply cross validation to. Using the K-Fold Cross-Validation method, the consistent dataset (the dataset before train-test split) was used to be split into k number of subsets, where k-1 subsets are used to train the models and the last subset is kept for validation to test the models. The scores of each fold are then averaged to evaluate the overall performance of each model. Cross-validation using 10-folds, where 9 folds were used for training and 1 used for testing, returned higher accuracy results in all three algorithms: Random Forest, SVM, and XGBoost.

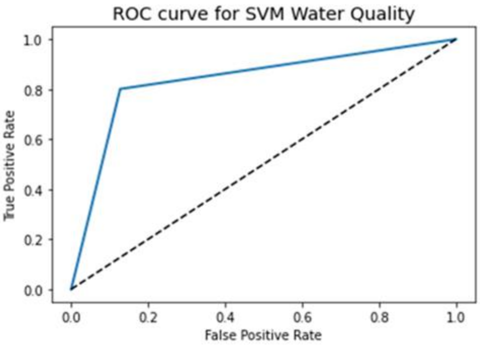


Fig E: ROC Curve of SVM Model after second iteration.

Since SVM has the highest mean accuracy score after cross validation, we returned to the second iteration of model to produce a Receiver Operating Characteristic Curve (ROC Curve) graph to visualize the SVM model's performance with respects to their classification threshold levels. The ROC Curve plots the True Positive Rate (recall) against the False Positive Rate (type II error). We can also calculate the area under the curve (ROC AUC) which will allow us to understand the classifier's performance numerically as a perfect classifier is equal to 1.0. The ROC AUC for SVM after the second iteration was 0.8368 and the cross validated ROC AUC was 0.8674 which is consistent with the rest of our evaluation metrics.

VI. REFERENCES

1. Proposition of New Ensemble Data-Intelligence Models for Surface Water Quality Prediction
   [IEEE, Volume:9], 26 July 2021

2. Toward Design of Internet of Things and Machine Learning-Enabled Frameworks for Analysis and Prediction of Water Quality
   [IEEE Volume 11], 14 September 2023

3. Water Quality Management Using Hybrid Machine Learning and Data Mining Algorithms: An Indexing Approach
   [IEEE Volume 10], 10 November 2022

4. Water quality prediction using machine learning models based on grid search method
   [Multimedia tools and Applications Journal], 29 September 2023

5. Prediction of Water Quality Using SoftMax-ELM Optimized Using Adaptive Crow-Search Algorithm
   [IEEE Volume 11], 2022

6. Dual Kidney- Inspired Algorithm for Water Quality Prediction and Cancer Detection
   [IEEE Volume 8], 2020

7. A Complete Proposed Framework for Coastal Water Quality Monitoring System With Algae Predictive Model
   [IEEE Volume 9], 2021

8. Machine learning algorithms for efficient water quality prediction
   [Modelling Earth Systems and Environments algorithm], 2021

9. Towards design of IoT and Machine Learning Embedded Frameworks for analysis and prediction of water quality

[IEEE Volume 11], 2023

10. A Divide-and-Conquer Method Based Ensemble Regression Model for Water Quality Prediction
2013, [Springer-Verlag Berlin Heidelberg]

11. Dimensionality Reduction for Water Quality Prediction from a Data Mining Perspective
2020, [Springer Nature Singapore Pte Ltd]

12. Hierarchical Neural Network Model for Water Quality Prediction in Wastewater Treatment Plants
2009, [Springer-Verlag Berlin Heidelberg]

13. Water Quality Prediction for Smart Aquaculture Using Hybrid Deep Learning Models
[IEEE Volume 10], 2022

14. Water quality prediction model utilizing integrated wavelet-ANFIS model with cross-validation
2010, [Springer-Verlag London Limited]