iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 11 **Issue:** IX **Month of publication:** September 2023

DOI: https://doi.org/10.22214/ijraset.2023.55862

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue IX Sep 2023- Available at www.ijraset.com

Exploring Machine Learning Algorithms for Reliable Water Quality Prediction

Deepak Thakur[1], A J Singh[2]
Department of Computer Science, Himachal Pradesh University

Abstract: ***Ensuring safe drinking water is a vital worldwide task. Accurate water quality prediction is crucial for protecting public health and the environment. Machine learning provides promising solutions for this objective. The study investigates the issue of precisely forecasting water quality with machine learning models. It examines different models and their efficacy in forecasting water quality features utilizing a given dataset. We conducted a comprehensive analysis of multiple machine learning models, including Bagging(REPTree), Multilayer Perceptron, M5P, Additive Regression, Stacking, Random Forest, and Decision Table. Firstly, ourselves imported the dataset into Weka, selected and configured the models, trained them on the dataset, and evaluated their performance using various metrics. Bagging (REPTree) outperformed compared to other models, showing its effectiveness in predicting water quality. Model selection depends on goals and constraints. Future research opportunities include feature engineering, ensemble methods, and data quality issues. The study concludes that Bagging (REPTree) classifier is a strong candidate for properly predicting water quality attributes. Future research should focus on improving feature engineering, exploring ensemble methods, expanding the dataset, and enhancing model explain ability. Deploying selected models for continuous monitoring and early detection can contribute to safer water supplies and sustainable water management practices. Compliance with water quality regulations can be better ensured through the application of these models. Overall, this study offers valuable insights regarding the application of machine learning for water quality prediction and highlights future directions for research and application in this important area.***
Keywords: Water quality classification (WQC), Machine learning, Data mining, Classification algorithms.

I. INTRODUCTION

Water quality assessment and prediction have received significant attention recently due to their critical implications for human health and environmental sustainability. Ensuring the safety and drinkability of water sources is of utmost importance, especially considering growing concerns about water pollution and associated risks. This study focuses on Prediction of water quality using advanced machine learning models and state-of-the-art techniques to address these pressing challenges. The objective of this research is to develop strong predictive models that can accurately classify water quality based on key features such as pH levels, hardness, solids concentration, chloramine content, sulfate concentration, conductivity, organic carbon content, trihalomethanes, turbidity, and drinkability. By harnessing the power of machine learning algorithms, we aim to effectively determine the drinkability of water, providing a valuable tool for monitoring and managing water quality. Throughout this investigation, we assess the performance of various machine learning models, including M5P regression trees, Multilayer Perceptron neural networks, Bagging, Additive Regression with Decision Stump, and Stacking with Zero, Random Forest, and Decision Table with Best First feature selection.

Each model is thoroughly evaluated using essential Metrics like Correlation Coefficient, Root Mean Squared Error, Mean Absolute Error, Relative Absolute Error, and Root Relative Squared Error are used for analysis. This research not only aims to identify the most effective model for water quality prediction but also explores the potential implications of model selection, hyper parameter tuning, and feature selection on predictive accuracy. Furthermore, we anticipate shedding light on the suitability of each model for addressing water quality issues and envision potential avenues for future research and improvements with relation to water quality assessment and prediction.

By leveraging the capabilities of machine learning and data-driven insights, we strive to contribute to ongoing efforts to safeguard water resources and ensure access to clean and drinkable water, which is essential for human well-being and environmental sustainability.

A structured presentation format was adopted for the topics in this study. First, the Background information about the issue were discussed in the 'Literature Review' section. Then, the relevant literature was presented in the 'Materials and Methods' section.

Finally, the data was classified, and machine learning algorithms were used to determine the features that affect water quality in the results analysis and discussion section.

II. LITERATURE REVIEW

A study by A. Najah et al. examines the effectiveness of various AI techniques in predicting water quality in Johor River, Malaysia. The MLP-ANN technique outperforms optimized hidden-layer neurons. SVM technique requires optimal hyperparameters determined through cross-validation[1]. Salisu Yusuf Muhammad et al. provide a literature review on water quality evaluation using machine learning algorithms.

The review discusses diverse scientific methods and studies for water quality classification, including neural networks, remote sensing technology, and support vector machine[2]. Yafra Khan et al. propose a water quality prediction model utilizing artificial neural network (ANN) and time-series analysis. The performance is evaluated using MSE, RMSE, and Regression Analysis. An IoT-based solution is proposed to monitor water quality in real time, using SVM, KNN, single-layer neural network, and deep neural network for classification[3]. Neha Radhakrishnan et al. compare machine learning models for classifying water quality[4]. Gasmin hayder et al. explore machine learning to monitor and predict water quality parameters, finding pH prediction to be most accurate[5].

Batta Mahesh provides a literature review on machine learning algorithms, noting the influence of problem type, variables, and suitable models on algorithm choice[6]. Shweta Agrawal et al. propose a model for water quality assessment using machine learning, achieving high prediction accuracy through a voting classifier with hard voting[7]. Nur afyfah suwadi et al. propose a feature selection method and compare machine learning models for water quality prediction, with XG Boost performing the best[8]. Umair ahmed et al. utilize gradient boosting, MLP, and polynomial regression to forecast water quality, with MLP exhibiting the highest classification accuracy[9].

Theyazn .h .h aldhyani et al use SVM, K-NN, and Naïve Bayes to predict water quality classification, with SVM achieving the highest accuracy[10]. In a study by Chalisa veesommai Sillberg et al., the SVM algorithm and AR were used to categorise the water quality of the Chao Phraya River. (2021). Their findings showed an accuracy of 0.86-0.95 when using three to six features for classification[11].

In a study by Armin azad karami et al. (2017), machine learning methods were used to select quality features for the Gorganroud River water.

Three algorithms - ACOR, GA, and ANFIS - were employed for feature evaluation, with the ANFIS model being the most effective in predicting EC, SAR, and TH during training[12]. Jefferson L. Lerios et al. (2019) conducted a study using data mining techniques to predict water quality in reservoirs. Various features, including the WQI, were utilized and the results indicated that the water quality was mostly fair and marginal, suggesting the presence of pollutants[13]. Md Saikat Islam Khan, et al. (2021) used PCR to identify the dominant WQI features and employed regression algorithms and gradient boosting classifiers for water quality classification[14].

Eli Dritsas et al. focused on the use of Machine learning is used to predict water quality. Gradient Boosting and MLP showed promise for accuracy, while WDT-ANFIS and Hybrid Bagging-Random Forest performed well in different scenarios. Extreme Gradient Boosting, Random Forest, and CEEMDAN provided stability in short-term predictions[15].

III. RESEARCH METHODOLOGY

Machine learning algorithms were used to categorize water quality using a model with multiple algorithms. The following paragraphs describe the appropriate methodology for implementing our approach**.**

A. Data Collection

The main objective of manipulating machine learning is to gather data in digital format. Our machine learning algorithms enable us to make predictions and interpretations by linking the data features. It is common practice to replace missing values in a dataset with the modes and means from the training data when preparing the data.

B. Normalization

In this method, all numbers in the dataset were normalized.

C. Flowchart

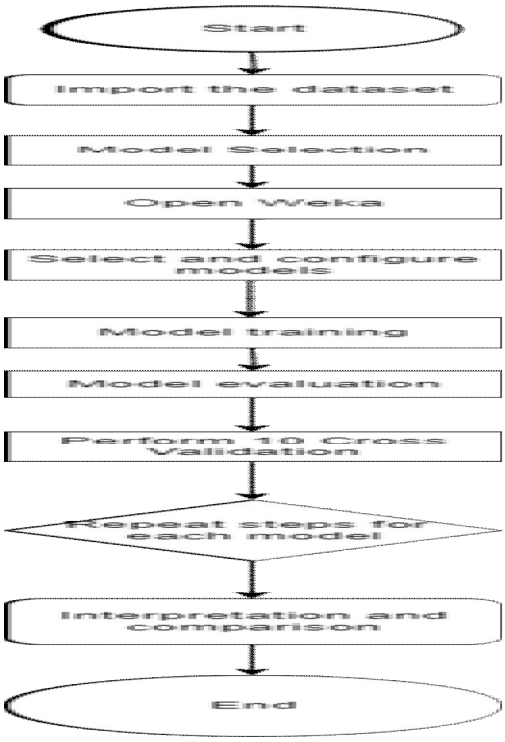


Fig 1: Framework of the proposed approach

D. Data Split:

The dataset is divided into two sets: one for training and one for testing. This division is done using a 10-fold cross-validation method.

E. Classification algorithms:

In the present study, various machine learning algorithm used to include M5P, MLP, Bagging, Additive regression, Stacking, Random forest, and Decision table.

F. Evaluation:

Several significant measurement methods are utilized for the assessment of our artificial intelligence algorithms. In our investigation, the following algorithms were used including a M5P and MLP[16], Bagging(REPTree), Additive regression and Stacking[17], Random forest and decision table[18].

G. M5P:

The M5P tree algorithm was developed to handle enumerated attributes and attribute missing values. Before constructing the tree in the M5P algorithm, all enumerated attributes are transformed into binary variables.

H. MLP:

A MLP neural network consists of an input layer, one or more hidden layers, and an output layer. The output layer can contain multiple nodes based on the problem. Signals flow forward through the network, while error signals propagate backward. Weight adjustments are made to minimize error.

I. Bagging(REPTree):

The process involves combining the outcomes of multiple models created from bootstrap sets. Sets are formed by selecting features or randomly selecting with replacement from the original dataset. The result is obtained by averaging the outputs of each model built over each bootstrap set.

J. Additive regression:

The algorithm begins with an empty ensemble and gradually adds new members. Each stage adds the model that optimizes the performance of the entire ensemble, without modifying the existing members.

K. Stacking:

In stacking, a metal earner combines the results of different base learners. The metal earner determines the optimal method to merge the results of the base learners is needed. The base learners used were LRM, M5P, M5R, MLP, RBF, and SVM used in each run, with one of them serving as the metal earner.

L. Random Forest:

The algorithm is a classification trees algorithm that enhances tree classifiers using the idea of a forest. The referenced research showcased accurate random forest classifiers that can handle noisy dataset values. There are no modifications made during classification. The forest's number of trees must be determined, and the output with the most votes is selected based on each tree's prediction.

M. Decision Table:

The decision table classifier produces important rules that assist in predicting new inputs. The decision table's lookup table can also be utilized in other domains, such as presenting significant rules for complex fuzzy systems with a limited expert knowledge base.

N. Dataset:

To construct a machine learning model, we employed the water_potability.csv file, which contains water quality metrics for 3,276 water bodies. High-quality water datasets are imperative for the undertaking of machine learning research. The dataset was acquired from Kaggle and encompasses 10 attributes, such as pH value, hardness, TDS, chloramines, sulphate, EC, TOC, turbidity, trihalomethanes, and portability. The portability attribute is categorized by values {0, 1}.

O. Data normalization:

Normalization is the first step in preparing data for machine learning. It ensures that numeric features are on a similar scale without distorting value ranges or losing information.

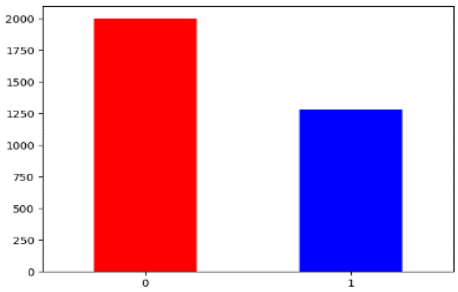


Fig2: Dataset normalization

IV. EXPERIMENTAL SETUP

The study aims to determine water quality using Kaggle's historical data. Various machine learning algorithms such as M5P, MLP, Bagging, Additive regression, Stacking, Random forest, and Decision table were utilized to achieve the results. The following section examines the experiments, findings, and assessment of our approach. WEKA 3.9, a data mining tool, was utilized to implement our model.

V. RESULTS

In this study, we utilized a water quality dataset and applied 10-fold cross-validation to showcase the machine learning models. The training set comprised of 3,276 water bodies, with 1,278 identified as non-potable and the remaining 1,998 as potable. Various metrics including RMSE, correlation coefficient, mean absolute error, relative absolute error, and root relative squared error were employed to assess the models' performance. These metrics can be computed using specific formulas.

A. Correlation coefficient

Before generating a model in machine learning, it is necessary to comprehend the connection between independent variables and the target variable. The correlation coefficient serves as a metric for association.

B. Mean absolute error

MAE is a metric that measures the average size of prediction errors. It calculates the absolute differences between predicted and true values. Negative differences are not considered.

MAE = True values – Predicted values

MAE calculates the mean error of each sample in a dataset and produces the result.

C. Root mean squared error

RMSE is a widely used metric for comparing predicted and observed values. It is derived from MSE and quantifies the error in the same units as the target variable**.**

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

D. Relative absolute error:

$$RAE = \frac{1}{n}\sum_{i=1}^{n}\frac{|y_i - \hat{y}_i|}{|y_i - \overline{y}|}$$

E. Root relative squared error:

$$RRSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{y}_i}{y_i - \overline{y}}\right)^2}$$

Based on the methods find the diff regression metrics for each model, table 1 shows the performance of each model.

TABLE1: MODELS PERFORMANCE EVALUATION METRICS

Models	Correlation Coefficient	Mean Absolute Error	Root Mean Squared Error	Relative Absolute Error	Root Relative Squared Error
M5P	0.3484	0.422	0.4573	88.6781	93.4774
MLP	0.2904	0.4063	0.5543	85.388	113.6426
Bagging	0.7054	0.3694	0.3947	77.6262	80.922
Additive Regression	0.2292	0.4505	0.4748	94.6702	97.3415
Stacking	0.057	6968.284	8770.5676	100	100
Random Forest	0.2189	6787.91	8555.056	97.4110	97
Decision Table	0.1747	6833.10	8649	98.0601	98.6181

From above table results show that, “Bagging (REPTree)" seems to be the top-performing model. It achieves the highest correlation coefficient, lowest mean absolute error, and lowest root mean squared error. However, the best model choice depends on application requirements and constraints. Factors like computational complexity, interpretability, and trade-offs between evaluations metrics should be considered when selecting the most suitable model.

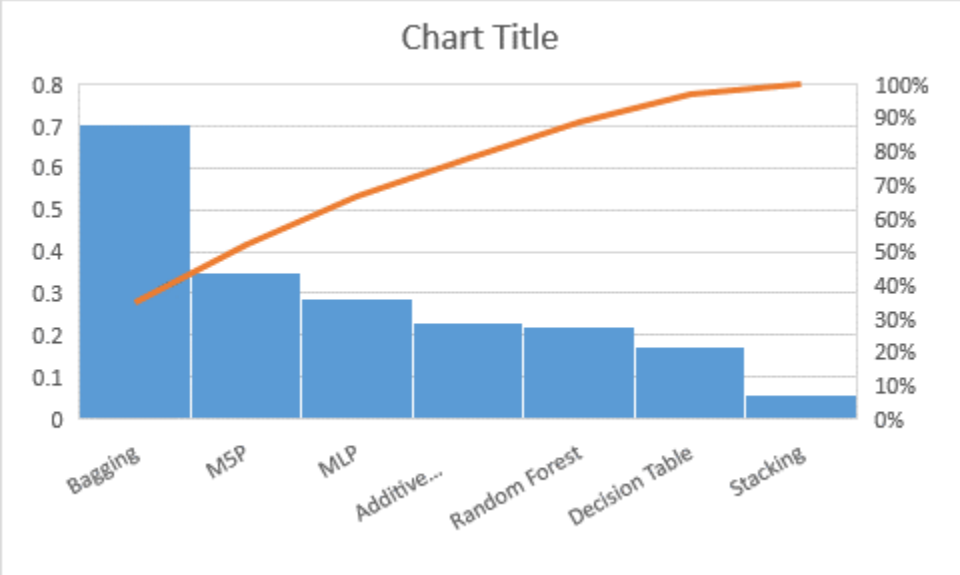


Fig2. Graphical representations of regression metrics using WEKA

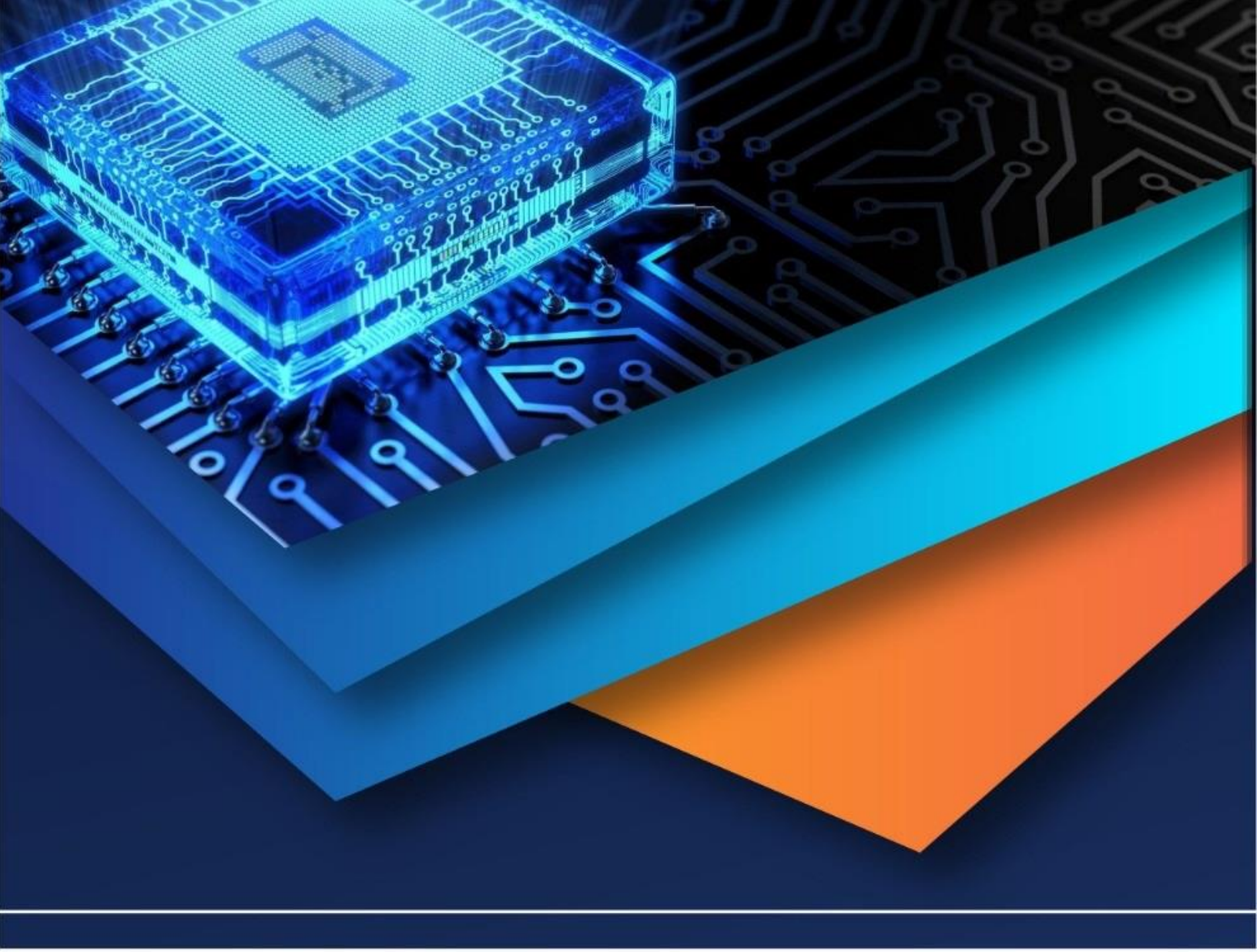
VI. CONCLUSION & FUTURE SCOPE

Analysis of various machine learning models for water quality prediction based on the provided dataset reveals key findings. Bagging (REPTree) consistently outperforms others, demonstrating its effectiveness in accurately predicting water quality. Model selection depends on specific goals and constraints. Future research could explore feature engineering and ensemble methods to improve model performance. Expanding the dataset and addressing missing values can lead to more robust models. Developing methods to explain model predictions is vital for real-world applications. Consider deploying the selected model for continuous monitoring and early detection. The developed models can assist in ensuring compliance with regulations and contribute to sustainable water management practices. Ongoing research should focus on data quality, model interpretability, and real-world applications to address challenges in water quality prediction.

REFERENCES

[1] A. Najah, O. A. Karim, O. Jaafar, and A. H. El-shafie, “An application of different artificial intelligences techniques for water quality prediction,” vol. 6, no. 22, pp. 5298–5308, 2011, doi: 10.5897/IJPS11.1180.
[2] S. Y. Muhammad, M. Makhtar, A. Rozaimee, A. A. Aziz, and A. A. Jamal, “Classification model for water quality using machine learning techniques,” Int. J. Softw. Eng. its Appl., vol. 9, no. 6, pp. 45–52, 2015, doi: 10.14257/ijseia.2015.9.6.05.
[3] C. S. See, “Predicting and Analyzing Water Quality using Machine Learning : A Comprehensive Model,” pp. 1–6.
[4] N. Radhakrishnan and A. S. Pillai, “Comparison of Water Quality Classification Models using Machine Learning,” no. June, pp. 1183–1188, 2020, doi: 10.1109/icces48766.2020.9137903.
[5] G. Hayder, I. Kurniawan, and H. M. Mustafa, “Implementation of machine learning methods for monitoring and predicting water quality parameters,” Biointerface Res. Appl. Chem., vol. 11, no. 2, pp. 9285–9295, 2021, doi: 10.33263/BRIAC112.92859295.
[6] B. Mahesh, “Machine Learning Algorithms - A Review,” vol. 9, no. 1, pp. 381–386, 2020, doi: 10.21275/ART20203995.
[7] S. Agrawal, S. K. Jain, A. Khatri, M. Agarwal, A. Tripathi, and Y. C. Hu, “Novel PSO Optimized Voting Classifier Approach for Predicting Water Quality,” Math. Probl. Eng., vol. 2022, 2022, doi: 10.1155/2022/6445580.
[8] N. A. Suwadi et al., “An Optimized Approach for Predicting Water Quality Features Based on Machine Learning,” Wirel. Commun. Mob. Comput., vol. 2022, 2022, doi: 10.1155/2022/3397972.
[9] U. Ahmed, R. Mumtaz, H. Anwar, A. A. Shah, and R. Irfan, “E ffi cient Water Quality Prediction Using Supervised,” pp. 1–14, 2019.
[10] T. H. H. Aldhyani, M. Al-Yaari, H. Alkahtani, and M. Maashi, “Water Quality Prediction Using Artificial Intelligence Algorithms,” Appl. Bionics Biomech., vol. 2020, 2020, doi: 10.1155/2020/6659314.
[11] C. V. Sillberg, P. Kullavanijaya, and O. Chavalparit, “Water Quality Classification by Integration of Attribute-Realization and Support Vector Machine for the Chao Phraya River,” J. Ecol. Eng., vol. 22, no. 9, pp. 70–86, 2021, doi: 10.12911/22998993/141364.
[12] A. Azad, H. Karami, S. Farzin, A. Saeedian, H. Kashi, and F. Sayyahi, “Prediction of water quality parameters using ANFIS optimized by intelligence algorithms (case study: Gorganrood river),” KSCE J. Civ. Eng., vol. 22, no. 7, pp. 2206–2213, 2018, doi: 10.1007/s12205-017-1703-6.
[13] J. L. Lerios and M. V. Villarica, “Pattern extraction of water quality prediction using machine learning algorithms of water reservoir,” Int. J. Mech. Eng. Robot. Res., vol. 8, no. 6, pp. 992–997, 2019, doi: 10.18178/IJMERR.8.6.992-997.
[14] M. S. Islam Khan, N. Islam, J. Uddin, S. Islam, and M. K. Nasir, “Water quality prediction and classification based on principal component regression and gradient boosting classifier approach,” J. King Saud Univ. - Comput. Inf. Sci., vol. 34, no. 8, pp. 4773–4781, 2022, doi: 10.1016/j.jksuci.2021.06.003.

[15] E. Dritsas and M. Trigka, "Efficient Data-Driven Machine Learning Models for Water Quality Prediction," Computation, vol. 11, no. 2, 2023, doi: 10.3390/computation11020016.
[16] P. Singh and S. Agrawal, "Node localization in wireless sensor networks using the M5P tree and SMOreg algorithms," Proc. - 5th Int. Conf. Comput. Intell. Commun. Networks, CICN 2013, pp. 104–108, 2013, doi: 10.1109/CICN.2013.32.
[17] M. Graczyk, T. Lasota, B. Trawiński, and K. Trawiński, "Comparison of bagging, boosting and stacking ensembles applied to real estate appraisal," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 5991 LNAI, no. PART 2, pp. 340–350, 2010, doi: 10.1007/978-3-642-12101-2_35.
[18] M. S. Alsahli, M. M. Almasri, M. Al-Akhras, A. I. Al-Issa, and M. Alawairdhi, "Evaluation of Machine Learning Algorithms for Intrusion Detection System in WSN," Int. J. Adv. Comput. Sci. Appl., vol. 12, no. 5, pp. 617–626, 2021, doi: 10.14569/IJACSA.2021.0120574.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY