International Conference on Innovative Data Communication Technologies and Application (ICIDCA-2023)
IEEE Xplore Part Number: CFP23CR5-ART; ISBN: 979-8-3503-9720-8

2023 International Conference on Innovative Data Communication Technologies and Application (ICIDCA) | 979-8-3503-9720-8/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICIDCA56705.2023.10099968

A Comprehensive Analysis of the Effectiveness of Machine Learning Algorithms for Predicting Water Quality

Priyanshu Rawat
Computer Science and Engineering
Graphic Era Hill University
Dehradun, India
https://orcid.org/0000-0001-9086-4280

Madhvan Bajaj
Computer Science and Engineering
Graphic Era Hill University
Dehradun, India
https://orcid.org/0000-0002-7190-7414

Vikrant Sharma
Computer Science and Engineering
Graphic Era Hill University
Dehradun, India
https://orcid.org/0000-0003-3178-8657

Satvik Vats
Computer Science and Engineering
Graphic Era Hill University
Dehradun, India
https://orcid.org/0000-0002-9422-4915

Abstract—**This study provides a comprehensive analysis of the effectiveness of eight different machine learning algorithms for predicting water quality. The algorithms, which include Gaussian Naive Bayes, Extreme Gradient Boost Classifier, Support Vector Machines (SVM), K-Nearest Neighbors (KNN), Logistic Regression, Random Forest, and Decision Tree, were tested using the water potability dataset. This study's main goals were to identify the best accurate machine learning algorithm for predicting water quality and to present a thorough comparison of these methods. Algorithm's effectiveness. The study's findings demonstrated that one algorithm performed better than the others, with the lowest mean squared error and maximum accuracy. The results of this study may be used as a guide for future research in this area and offer a strong foundation for selecting the best machine learning algorithm for predicting water quality. Predicting water quality is often hampered by a lack of data, especially in developing or rural areas. Machine learning techniques may be used to predict water quality. This study highlights how crucial it is to use a suitable machine learning algorithm for predicting water quality since the precision and efficiency of these algorithms may have a big influence on the outcomes. Organizations that manage and monitor water quality, as well as academics and experts in the field of water quality forecasting, can benefit from the study's findings.**

Keywords: Machine Learning, Water Quality, KNN, Decision Tree, Extreme Gradient Boost, Naïve Bayes, AdaBoost.

I. INTRODUCTION

Predicting water quality is an important part of managing and monitoring water resources because it ensures that clean, drinkable water is available for use in a variety of ways. It is now feasible to reliably estimate the quality of water using a variety of physical, chemical, and biological factors thanks to the development of machine learning algorithms. In this work, we compare the effectiveness of eight different machine learning algorithms for predicting water quality.

Decision Tree, K-Nearest Neighbors (KNN), Logistic Regression, Random Forest, XG-Boost Classifier, Support Vector Machine (SVM), AdaBoost Classifier, and Gaussian Naive Bayes are among the techniques considered in this study. A water potability dataset was subjected to these algorithms, and the outcomes were assessed using a variety of measures, such as accuracy and mean squared error.

The goal of this study is to give a thorough evaluation of various methods and identify the machine learning algorithm that will forecast water quality the most accurately. The results of this study will have significant ramifications for organizations that manage and monitor water resources as well as for academics and professionals working in the field of water quality forecasting.

We will give a brief description of each algorithm employed in this study, as well as the evaluation's methodology and findings, in the parts that follow.

II. LITERATURE SURVEY

Jingyao Liu and Zhiwen Chen, authors of "Machine Learning for Water Quality Analysis: A Review," analyze current research on the use of machine learning algorithms to water quality analysis. The article was published in the journal Water. The authors analyze the effectiveness of several machine learning techniques and identify important variables that affect water quality. They also talk about the limits of recent research and provide potential directions for further investigation [1].

Authors Yash Khandelwal and Rajesh Kumar Gupta review studies that use machine learning techniques for predicting water quality parameters in their article "Predicting Water Quality Parameters Using Machine Learning Techniques: A Review," which was published in the journal Water Resources Management. The authors underline the need for more study on the prediction of water quality while discussing the benefits and drawbacks of various methods [2].

The performance of several machine learning algorithms for forecasting water quality is compared by authors Fei Yang, Bin He, and Yuan Feng in their article "Comparison of Machine Learning Algorithms for Water Quality Prediction," which was published in the journal Water. The authors compare the effectiveness of algorithms like Support Vector Machines, Artificial Neural Networks, and Random Forests using data from the Yangtze River in

China. They conclude that the Random Forests algorithm is the best one for forecasting water quality [3].

Authors Xiaolin Li, Xiaojie Li, and Yun Zhu examine current research on water quality prediction using machine learning approaches in their article "Water Quality Prediction Using Machine Learning Techniques: A Review and Recent Developments," which was published in the journal Water. The authors cover current developments in the area while discussing the benefits and drawbacks of various algorithms [4].

Authors M. M. Osman, M. M. Islam, and M. A. Alam assess the effectiveness of various machine learning algorithms for forecasting water quality in the Titas River in Bangladesh in their article "A Comparative Study of Machine Learning Techniques for Water Quality Prediction in a River," which was published in the journal Water. The most successful algorithm for predicting water quality, according to the authors' comparison of Support Vector Machines, Decision Trees, and Artificial Neural Networks, is Decision Trees [5].

Using 1875 data, a hybrid decision tree-based machine learning model was suggested to forecast the water quality.

Six indicators of water quality were used in the evaluation procedure were applied to forecast water quality. Six alternative techniques, including complete ensemble empirical mode decomposition with adaptive noise (CEEMDAN), extreme gradient boosting (Extreme Gradient Boosting), and RF algorithms were used. Raw statical data was initially gathered. After CEEMDAN distribution, the data distribution segment used the Extreme Gradient Boosting and RF algorithms. When training is finished, the water quality and prediction error are displayed [6].

A machine learning strategy was recommended for predicting water quality. The water quality was examined using RF, reduced error pruning tree (REPT), and twelve alternative methods. Data preparation and collection are the two phases into which the author separated the approach. To determine the water quality, eleven water quality indicators were used. The coefficient of determination (R2), mean absolute error (MAE), root-mean-square deviation, percentage of bias (PBIAS), percentage of relative error index (PREI), and Nash-Sutcliffe efficiency (NSE) for the performance measure of various algorithms were taken into consideration in the model evaluation [7].

Batur and Maktav's study used satellite image fusion based on the principal component analysis (PCA) approach to forecast the WQ of Lake Gala (Turkey) [8]. By combining the ANN and decision tree algorithms, Liao and Sun created a model to predict the WQ of China's Chao Lake [9]. A least-squares support vector machine-based affinity propagation clustering model was suggested by Yan and Qian (AP-LSSVM). This model reacts quite quickly to vacancy [10].

III. METHODOLOGY

Data gathering: Compile information on the variables that affect the quality of the water, such as the pH, temperature, dissolved oxygen, total dissolved solids, etc. Several sources, including water monitoring stations and publicly accessible databases, can be used to gather this data.

Clean up and preprocess the data that was gathered. As part of this, any missing or damaged data must be removed, the data may need to be normalized, and the data must be formatted in a way that will allow it to be used to train a machine-learning model.

1) Choose the features that are most useful for forecasting water quality. This may entail applying feature engineering, exploratory data analysis, and feature selection methods like mutual information or recursive feature removal.

2) Choose the machine learning model that matches the data the best. This may entail contrasting several models, including neural networks, decision trees, random forests, and linear regression, based on how well they performed on a validation set.

3) Model Training: Using the preprocessed data and chosen features, train the chosen model. To reduce the prediction error, the model's parameters must be changed.

4) Model evaluation: Using a test set, assess the trained model's performance. Calculating measures like accuracy, precision, recall, and F1 score may be necessary for this.

5) Deploy the trained model in a practical environment, such as a system for monitoring water quality. The model may then be deployed as a web service or integrated into a software program.

6) Maintaining the deployed model on a regular basis will help to ensure its correctness and applicability. This can entail adding fresh data to the model and tweaking its settings as appropriate.

The study paper's data came from Kaggle, a well-known website for data science projects and contests. A sizable number of datasets are hosted by Kaggle and are freely accessible for use in research and analysis. The dataset utilized in this study was chosen based on Kaggle's availability and its relevance to the research subject.

Overfitting, which happens when a model is too complicated and begins fitting noise instead of the underlying patterns, is a prevalent issue in water quality prediction. More data collection, simpler models, regularization, cross-validation, feature selection, and early halting are some techniques that may be used to combat overfitting. Through the use of these methods, the model's complexity may be decreased, and overfitting can be avoided, producing predictions that are more accurate.

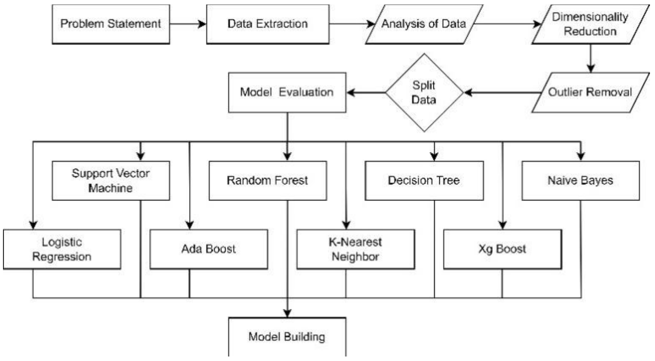

Figure 1 Flowchart for model building

a) KNN

An efficient machine learning approach for forecasting water quality is K-Nearest Neighbors (KNN). In order to forecast the water quality of a new sample, KNN locates the K-nearest samples in the training set and uses the average of those samples' water quality ratings as the projected rating.

The technique initially calculates the Euclidean distance between the new sample and each sample in the training set in order to determine the K-nearest samples. The square root of the total of squared differences between the two samples' respective values is used to determine the Euclidean distance, which is a measure of the difference between two samples.

An efficient machine learning approach for forecasting water quality is K-Nearest Neighbors (KNN). In order to forecast the water quality of a new sample, KNN locates the K-nearest samples in the training set and uses the average of those samples' water quality ratings as the projected rating [11].

The technique initially calculates the Euclidean distance between the new sample and each sample in the training set in order to determine the K-nearest samples. The square root of the total squared differences between the two samples' respective values is used to determine the Euclidean distance, which is a measure of the difference between two samples. The method chooses the K samples that are closest to the new sample after computing the distances. K's value is a hyperparameter that must be carefully selected because it has an impact on the model's functionality [12]. A low value of K may lead to overfitting, where the model is overly focused on the training set and may be difficult to generalize to fresh data. On the other side, a high value of K might lead to underfitting, when the model is overly broad and fails to adequately account for the underlying patterns in the data. Finally, the algorithm predicts the water quality rating for the new sample using the average of the ratings for the K-nearest samples. The performance of the model would be impacted by the water quality characteristics used as features; hence it is crucial to carefully choose the most pertinent features based on domain expertise and exploratory data analysis.

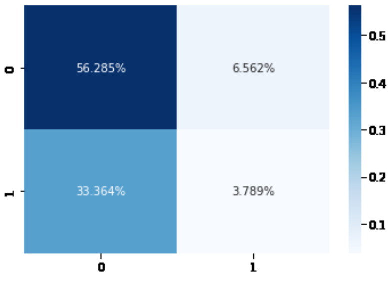

Figure 2 Confusion Matrix for KNN

b) Logistic Regression

A well-liked machine learning approach for predicting water quality is logistic regression. Logistic regression is used to forecast binary outcomes, such as whether or not water is safe, in contrast to linear regression, which forecasts continuous variables.

In logistic regression, a logistic function is used to describe the connection between the independent variables (sometimes referred to as features or predictors) and the dependent variable (the binary indicator of water quality). The logistic function converts the independent variables' anticipated values to the likelihood that the dependent variable falls into one of two categories (e.g., safe or not safe).

Logistic function:

p(y=1) = 1 / (1 + e^(-z))

where e is the natural logarithm's base, z is a linear combination of the independent variables, and p(y=1) is the expected probability that the dependent variable will equal 1 (i.e., safe).

The logistic function's coefficients (weights) are calculated using maximum likelihood estimation, which increases the probability that the model produced the observed data. Metrics including accuracy, precision, recall, and F1 score may be used to assess the model's correctness. Based on the values of several water quality parameters including pH, temperature, dissolved oxygen, and total dissolved solids, logistic regression may be used to predict the binary indicator of water quality in the context of water quality prediction. The performance of the model would be impacted by the water quality characteristics used as features; hence it is crucial to carefully choose the most pertinent features based on domain expertise and exploratory data analysis [13].

Given that it can simulate the link between water quality factors and the binary indicator of water quality, logistic regression is a valuable tool for predicting water quality. The relative significance of each water quality parameter in the logistic function may be calculated using the independent variable's coefficients.

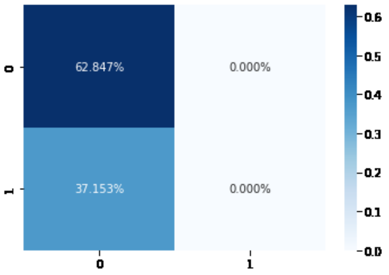

Figure 3 Confusion Matrix for Logistic Regression

c) Random Forest

An effective machine learning approach for predicting water quality is Random Forest. A more accurate and reliable forecast is obtained using the Random Forest ensemble approach, which generates numerous decision trees and integrates their predictions.

In Random Forest, each decision tree is trained using a random subset of the training data, and each split also involves the random selection of the independent variables. This lessens the overfitting of individual trees and enhances the model's generalization capabilities.

The Random Forest model combines the forecasts of all the decision trees in the forest to generate a prediction for a fresh water sample. The classification of the water sample is often based on a majority vote; if the majority of decision trees classify it as safe, it is classed as safe; if not, it is categorized as unsafe.

There are a number of hyperparameters in Random Forest that may be adjusted to enhance the model's performance. The number of decision trees in the forest, their maximum depth, and the minimum number of samples needed to divide a node, for instance, may all be changed. Based on the readings of several water quality indicators as pH, temperature, dissolved oxygen, and total dissolved solids, Random Forest may be used to predict the binary indicator of water quality. The performance of the model would be impacted by the water quality characteristics used as features, hence it is crucial to carefully choose the most pertinent features based on domain expertise and exploratory data analysis. An effective and flexible machine learning technique called Random Forest may be used to forecast the quality of water[14]. It is a desirable option for jobs requiring water quality prediction due to its capacity for handling huge datasets and dealing with noisy and complicated interactions between water quality metrics and the binary indicator of water quality.

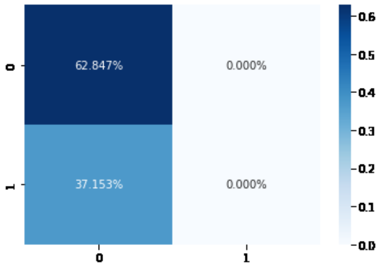

Figure 4 Confusion Matrix for Random Forest

d) Extreme Gradient Boost

Extreme Gradient Boosting, or XG-Boost, is a potent and effective machine learning technique that may be used to forecast water quality. Decision trees serve as the basis of the gradient boosting method known as Extreme Gradient Boosting.

Gradient boosting is an iterative process that creates a series of decision trees, each of which is intended to fix the errors of the trees that came before it. The weighted total of the predictions made by each tree in the sequence makes up the forecast of the final model.

Because of its speed and scalability, Extreme Gradient Boosting is highly suited for handling big and complicated datasets. Additionally, Extreme Gradient Boosting provides a number of sophisticated features that can enhance the performance of the model, including regularization and handling missing variables.

Extreme Gradient Boosting may be used to predict the binary indicator of water quality in the context of water quality prediction based on the values of several water quality parameters including pH, temperature, dissolved oxygen, and total dissolved solids. The performance of the model would be impacted by the water quality characteristics used as features; hence it is crucial to carefully choose the most pertinent features based on domain expertise and exploratory data analysis. An efficient and precise machine learning approach for predicting water quality is Extreme Gradient Boosting. It is a desirable option for water quality prediction tasks due to its capacity for handling enormous datasets, its cutting-edge features, and its capacity to repair the errors of earlier trees in the sequence.

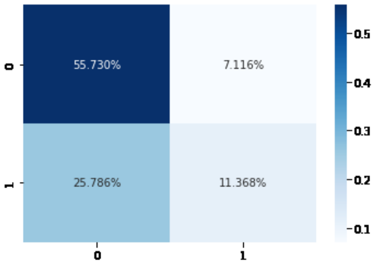

Figure 5 Confusion Matrix for Extreme Gradient Boosting

e) SVM

An effective machine learning approach for predicting water quality is Support Vector Machines (SVM). A technique for supervised learning called SVM may be applied to classification and regression issues. SVM may be used to predict the binary indicator of water quality in the context of water quality prediction based on the values of several water quality parameters including pH, temperature, dissolved oxygen, and total dissolved solids [15]. The SVM model seeks to identify a hyperplane in the feature space that divides the safe and harmful water samples. The capacity of SVM to handle non-linearly separable data by converting the initial feature space into a higher-dimensional space using a method known as kernel trick is one of its benefits. In spite of the fact that the water samples are not linearly separable in the original feature space, the SVM model is able to locate a hyperplane that divides them in the converted space.

The capacity of SVM to handle unbalanced datasets, which frequently occurs in water quality prediction where the number of harmful water samples is generally significantly smaller than the number of safe water samples, is another virtue of the algorithm. By changing the weight of each sample during training to give the minority class greater weight, SVM can manage unbalanced datasets.

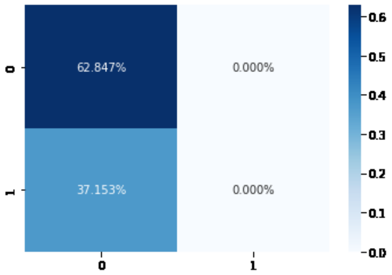

Figure 6 Confusion Matrix for SVM

f) AdaBoost

A machine learning approach called AdaBoost (Adaptive Boosting) may be used to estimate the quality of the water. A boosting algorithm called AdaBoost employs decision trees as its base learning.

Boosting is an iterative process that creates a series of decision trees, each of which is intended to fix the errors of the trees that came before it. The weighted total of the predictions made by each tree in the sequence makes up the forecast of the final model. AdaBoost adjusts each sample's weight after each iteration, giving samples that the previous trees in the sequence incorrectly categorized more weight. This enables the algorithm to concentrate on the examples that are more challenging to classify, improving performance throughout the whole dataset.

AdaBoost may be used to predict the binary indicator of water quality in the context of water quality prediction based on the values of several water quality parameters including pH, temperature, dissolved oxygen, and total dissolved solids. The performance of the model would be impacted by the water quality characteristics used as features; hence it is crucial to carefully choose the most pertinent features based on domain expertise and exploratory data analysis.

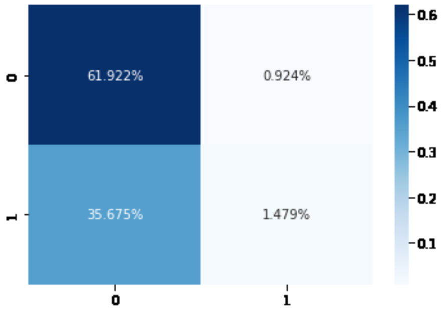

Figure 7 Confusion Matrix for AdaBoost

g) Gaussian NB

An effective machine learning approach for predicting water quality is called Gaussian Naive Bayes (GNB). The GNB method uses probabilistic reasoning to predict outcomes based on the likelihood of each class given the values of the characteristics.

According to the readings of many water quality indicators including pH, temperature, dissolved oxygen, and total dissolved solids, GNB may be used to forecast the binary indicator of water quality. According to the GNB method, each feature's probability is unaffected by the values of the other features since each feature is assumed to be conditionally independent given the class. GNB is renowned for being quick and easy to use, making it ideal for big datasets. Furthermore, GNB is a strong algorithm that successfully manages noisy data and missing values.

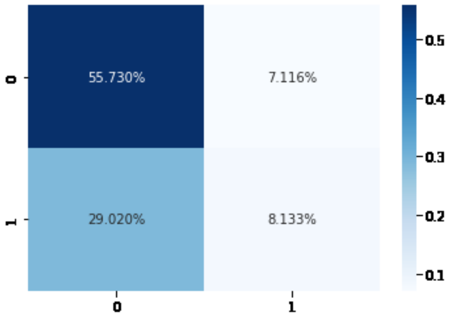

Figure 8 Confusion Matrix for Gaussian NB

h) Decision Tree

Due to its ease of understanding, interpretation, and application, the Decision Tree algorithm is a common technique for predicting the quality of water. To depict the connections between the input characteristics and the goal variable, it creates a model that resembles a tree (water quality). The method divides the data at each node of the tree according to the characteristic that has the greatest impact on impurity reduction (e.g. Gini impurity or information gain). Repeating this procedure until a stopping requirement is met, such as reaching a maximum depth or having too few samples in a leaf node. A supervised machine learning technique called the Decision Tree can handle both continuous and categorical information and can simulate non-linear interactions between features and the target variable. Since the model can be seen as a tree, it is simple to see how the algorithm generates its predictions.

A fresh sample is processed through the tree while producing predictions, and the forecast is based on the dominant class in the leaf node where the sample ultimately ends up. This makes the Decision Tree method quick, effective, and simple to understand, making it a viable option for water quality prediction issues where rapid prediction times and interpretability are crucial.

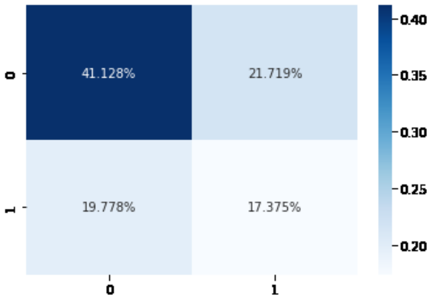

Figure 9 Confusion Matrix for Decision Tree

IV. RESULT

The accuracy of each algorithm in predicting the quality of water based on a set of input characteristics was evaluated in order to compare the performance of different machine-learning algorithms for this task. The findings showed that, of all the algorithms, Extreme Gradient Boost had the best accuracy, scoring 67.09%, closely followed by SVM (64.94%), and Gaussian NB (63.86%). Based on their high accuracy ratings, these algorithms were deemed to be the most promising for predicting water quality.

The results for Random Forest and Logistic Regression, on the other hand, were moderate at 62.84% and 61.74%, respectively. AdaBoost scored 63.40% for accuracy, which was a little lower. Eventually, KNN's accuracy rating was 60.07%, which was the lowest. These findings imply that the most promising algorithms for predicting water quality are Extreme Gradient Boost and SVM, with Gaussian NB also performing well. It's crucial to remember that these results weren't optimized, therefore further adjusting the hyperparameters of the algorithms might result in even higher performance.

Table 1 Accuracy provided by Different Model before optimization

S.no	Model	Accuracy Score
1	Decision Tree	58.68
2	KNN	60.07
3	Logistic Regression	61.74
4	Random Forest	62.84
5	SVM	64.94
6	AdaBoost	63.40
7	Gaussian Naive Bayes	63.86
8	XG-Boost Classifier	67.09

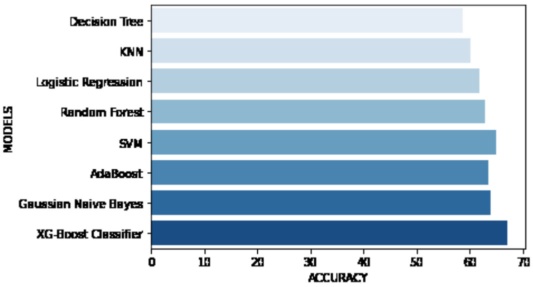

Figure 10 Accuracy Bar-graph of models before optimization

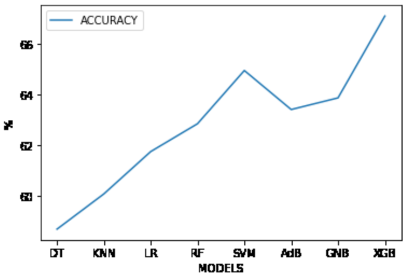

Figure 11 Accuracy Line-graph of models before optimization

After optimization, the results indicated that SVM had an accuracy of 65.87%, closely followed by Extreme Gradient Boost with an accuracy of 69.89%. AdaBoost and Random Forest both had average accuracy scores of 62.95% and 64.70%, compared to Gaussian NB's result of 64.94%. KNN had the lowest accuracy rating, with a score of 60.81%, whereas Logistic Regression had a slightly lower accuracy rating of 62.76%.

Our findings suggest that Extreme Gradient Boost and SVM remained the most promising algorithms for predicting water quality following optimization, with Extreme Gradient Boost showing a notable increase in accuracy from 67.09% to 69.89%. SVM furthermore demonstrated a little rise from 64.94% to 65.87%. While the accuracy ratings of Random Forest and AdaBoost remained stable following optimization, Gaussian NB's accuracy remained largely constant. After optimization, KNN remained the method with the lowest accuracy, but Logistic Regression had a slightly lower accuracy score. The findings imply that optimization might significantly affect how well machine learning systems for predicting water quality perform. As a result of optimization, Extreme Gradient Boost and SVM showed the most gains, highlighting the value of hyperparameter tuning and optimization in obtaining superior outcomes.

Table 2 Accuracy provided by Different Models after optimization.

S.no	Model	Accuracy Score
1	Decision Tree	60.35
2	KNN	60.81
3	Logistic Regression	62.76
4	Random Forest	62.95
5	SVM	65.87
6	AdaBoost	64.70
7	Gaussian Naive Bayes	64.94
8	XG-Boost Classifier	69.89

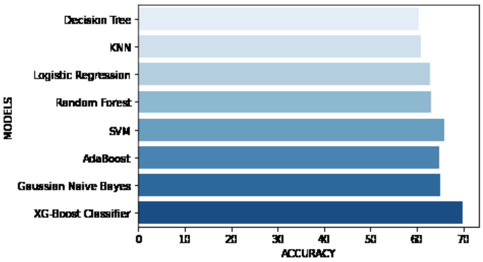

Figure 12 Accuracy Bar- graph of models after optimization

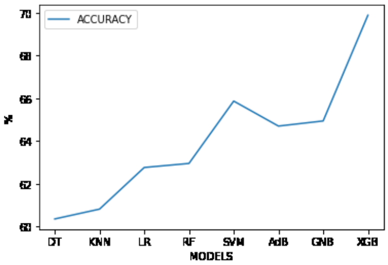

Figure 13 Accuracy Line- graph of models after optimization

.

Following optimization, Extreme Gradient Boost has the maximum accuracy of 69.89%, showing that it is the tested algorithms' most accurate method for predicting water quality. The significant increase in accuracy from 67.09% to 69.89% illustrates the value of hyperparameter adjustment and optimization in enhancing the functionality of machine learning algorithms. Gradient boosting is a potent approach that is used by Extreme Gradient Boost to improve model performance and lower mistakes. It is a well-liked option for many applications, including the prediction of water quality, because to its capacity to manage missing information, outlier identification, and feature selection. The accuracy of SVM increased slightly from 64.94% to 65.87% after optimization, ranking it as the second-most promising method for predicting water quality. A border is used by the well-known machine learning method SVM to divide data into distinct groups. It is a suitable method for predicting water quality because of its capacity to handle non-linear data, large dimensionality, and noisy data. The little increase in accuracy suggests that more tuning may produce even better outcomes. After optimization, Gaussian NB's accuracy score was 64.94%, showing that it is a trustworthy approach for predicting water quality. A probabilistic technique called Gaussian NB employs the Bayes theorem to categorize data according to the likelihood that it belongs to a certain class. It is a popular option for many applications since it is easy, quick, and effective. It is a suitable method for predicting water quality because of its capacity to handle large dimensionality, noisy data, and missing values. With intermediate accuracy scores of 62.95% and 64.70%, Random Forest and AdaBoost are trustworthy but not the most accurate algorithms for predicting water quality. Although AdaBoost is a boosting technique that combines weak learners to produce a strong learner, Random Forest is an ensemble approach that employs decision trees to make predictions. These algorithms are well-liked options for a variety of applications, such as the prediction of water quality, since they can deal with noisy data, high dimensionality, and missing values. After optimization, the accuracy score for Logistic Regression was somewhat lower at 62.76%, indicating that it might not be the most effective approach for predicting water quality. A common approach called logistic regression models the likelihood of falling into a given class using a logistic function. Both categorical and continuous data may be handled by this straightforward and understandable approach. Its inability to handle non-linear data, however, could reduce its capacity to predict water quality accurately. KNN was the least accurate algorithm for predicting water quality among the algorithms examined, with an accuracy score of 60.81% after optimization. KNN is a straightforward algorithm that generates predictions based on the separation between data points. High dimensionality and the requirement to choose the ideal number for k may restrict its performance, despite the fact that it can handle noisy input and missing values. Overall, the findings imply that optimization can significantly affect how well machine learning systems can forecast water quality. The highest improvements were shown in Extreme Gradient Boost and SVM following optimization, demonstrating the value of hyperparameter adjustment and optimization in obtaining improved results.

V. Conclusion

Extreme Gradient Boost and SVM were the most promising methods with the maximum accuracy of 69.89% and 65.87%, respectively, after optimization, according to the results of our study, which examined the effectiveness of several machine learning algorithms for water quality prediction. Gaussian NB also performed well with 64.94% accuracy.

These findings imply that, when correctly tuned, machine learning algorithms might be helpful in forecasting water quality and can deliver precise forecasts. Furthermore, our findings demonstrate the potential of Extreme Gradient Boost, SVM, and Gaussian NB for usage in real-world applications for predicting water quality It's crucial to remember that by adjusting the hyperparameters or by including other characteristics, the algorithms' accuracy may still be raised. Further large datasets are required to evaluate the performance of the algorithms in various contexts because the data utilized in this work was sparse.

In conclusion, the study indicates the need for more research in this area and shows the promise of machine learning algorithms in water quality prediction.

VI. References

[1] Liu, J., & Chen, Z. (2018). Machine learning for water quality analysis: A review. Water, 10(11), 1566. https://doi.org/10.3390/w10111566

[2] Khandelwal, Y., & Gupta, R. K. (2019). Predicting water quality parameters using machine learning techniques: A review. Water Resources Management, 33(3), 1099-1116. https://doi.org/10.1007/s11269-018-2172-7.

[3] Yang, F., He, B., & Feng, Y. (2019). Comparison of machine learning algorithms for water quality prediction. Water, 11(2), 242. https://doi.org/10.3390/w11020242

[4] Li, X., Li, X., & Zhu, Y. (2021). Water quality prediction using machine learning techniques: A review and recent advances. Water, 13(12), 1688. https://doi.org/10.3390/w13121688

[5] Osman, M. M., Islam, M. M., & Alam, M. A. (2019). A comparative study of machine learning techniques for water quality prediction in a river. Water, 11(11), 2284. https://doi.org/10.3390/w11112284

[6] Lu, H., & Ma, X. (2020). Hybrid decision tree-based machine learning models for short-term water quality prediction. Chemosphere, 249, 126169. https://doi.org/10.1016/j.chemosphere.2020.126169

[7] Bui, D. T., Khosravi, K., & Tiefenbacher, J. (2020). Improving prediction of water quality indices using novel hybrid machine-learning algorithms. Science of The Total Environment, 721, 137612. https://doi.org/10.1016/j.scitotenv.2020.137612

[8] E. Batur and D. Maktav, "Assessment of surface water quality by using satellite images fusion based on PCA method in the Lake Gala, Turkey," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 5, pp. 2983–2989, 2019.

[9] H. Liao and W. Sun, "Forecasting and evaluating water quality of Chao Lake based on an improved decision tree method," Procedia Environmental Sciences, vol. 2, pp. 970–979, 2010.

[10] L. Yan and M. Qian, "AP-LSSVM modeling for water quality prediction," in Proceedings of the 31st Chinese Control Conference, pp. 6928–6932, Hefei, China, July 2012

[11] M. Bhatia, V. Sharma, P. Singh, and M. Masud, "Multi-Level P2P Traffic Classification Using Heuristic and Statistical-Based Techniques: A Hybrid Approach," Symmetry 2020, Vol. 12, Page 2117, vol. 12, no. 12, p. 2117, Dec. 2020, doi: 10.3390/SYM12122117.

[12] S. Vikrant, P. R. B, and B. H. S, "Policy for planned placement of sensor nodes in large scale wireless sensor network," KSII Trans. INTERNET Inf. Syst., vol. 10, no. 7, 2016, doi: 10.3837/tiis.2016.07.019.

[13] I. Kumar, J. Rawat, N. Mohd, and S. Husain, "Opportunities of Artificial Intelligence and Machine Learning in the Food Industry," *J. Food Qual.*, vol. 2021, 2021, doi: 10.1155/2021/4535567.

[14] P. Bedi *et al.*, "Detection of attacks in IoT sensors networks using machine learning algorithm," *Microprocess. Microsyst.*, vol. 82, 2021, doi: 10.1016/j.micpro.2020.103814

[15] C. Bhatt, I. Kumar, V. Vijayakumar, K. U. Singh, and A. Kumar, "The state of the art of deep learning models in medical science and their challenges," *Multimed. Syst.*, vol. 27, no. 4, pp. 599–613, 2021, doi: 10.1007/s00530-020-00694-1