Machine Learning-based Water Potability Prediction: Model Evaluation, and Hyperparameter Optimization

Anoushka Mondal, Sudhanshu Sudhakar Dubey

Electronics and Communication Department, RV College of Engineering
Mysore Rd, RV Vidyaniketan, Post, Bengaluru, Karnataka-560059, India

Corresponding author: Sudhanshu Sudhakar Dubey, Email: dubeysudhanshusd.ec20@rvce.edu.in

This research aims to predict water potability, which is of utmost importance for community safety. A comprehensive analysis of a machine learning model is presented here, considering various quality parameters, to achieve this prediction. The model incorporates decision tree, KNN, Random Forest, SGD, SVM, logistic regression, and other algorithms. All steps of the study, including pre-processing, exploratory data analysis, feature scaling, model construction, assessment, and hyperparameter tuning, are thoroughly covered. Performance indicators like accuracy, confusion matrix, and classification report are used to evaluate the effectiveness of each model. Hyperparameter tweaking is implemented in decision trees, random forest algorithms, and K-nearest neighbors through grid search to optimize accuracy. The suggested model demonstrates its capability to forecast water potability accurately. It provides stakeholders with a systematic approach to model construction, evaluation, and optimization, thus ensuring water safety.

Keywords: Machine Learning, Water Potability, Classification, Feature Scaling, Model Evaluation, Hyperparameter Tuning.

in: *Advancements in Communication and Systems*. Ed. by Ashish Kumar Tripathi and Vivek Shrivastava. Computing and Intelligent Systems, SCRS, India., 2023, pp. 37–54. DOI: https://doi.org/10.56155/978-81-955020-7-3-4

1 Introduction

Water is an essential element to sustaining life. The availability of safe and drinkable water is important for public health and environmental conservation. The quality of available water supplies has become even more important matter in today's world which is a consequence of rising human-caused events. Pollutants, chemicals, and microbes impact the safety of drinking water which leads to irreversible damage to people and their environment.These hazards not only impact our well-being but also jeopardize the delicate equilibrium of ecosystems. In order to address these challenges and assess the level of water quality, we can resort to advanced technologies such as machine learning.

Machine learning represents an elaborate technique that employs computers to comprehend intricate information. It facilitates the analysis of substantial volumes of data and enables us to make highly accurate predictions. Through the identification of patterns and correlations within water quality data, we can develop models capable of determining which water samples may possess harmful attributes.

The central aim of our research lies in the creation and examination of an exceptionally ingenious machine learning model. Said model will possess the capacity to forecast whether water is suitable for human consumption or not. By utilizing sophisticated algorithms, the model can scrutinize diverse facets of water quality and establish its potability. This discriminative process will enable us to make well-informed choices and efficaciously manage concerns related to water quality.

Our study abides by distinct crucial steps. Firstly, we undertake the meticulous task of cleaning and analyzing the data in order to ensure its reliability. Subsequently, we explore sundry machine learning algorithms such as LR, SGD,SVM, DT, KNN, RF, and a neural network based on BP. We diligently assess the performance of each algorithm, employing well-established metrics, to gauge their degree of accuracy in predicting water potability.

Furthermore, the paper delves deeply into the realm of hyperparameter tuning, a meticulous process aimed at optimizing the performance of the model by identifying the optimal configuration of algorithm-specific parameters. This endeavor highlights the utmost importance of parameter selection and its profound impact on the predictive accuracy of the developed models.

The contributions of this paper transcend beyond the mere methodology and implementation of machine learning models. By disseminating a comprehensive framework for predicting water potability, the paper equips various stakehold-

ers, ranging from water authorities to researchers and policymakers, with an exceptionally powerful tool to enhance the management of water quality. The proposed model's potential to deliver accurate and timely predictions further underscores its immense significance in ensuring the provision of clean and safe water for a plethora of applications.

In conclusion, this paper accentuates the indispensable role played by machine learning in advancing the assessment of water quality. It exemplifies the remarkable potential of predictive models to revolutionize the field of water safety management, offering an avenue for well-informed decision-making and the safeguarding of public health and environmental sustainability.

2 Literature Review

This paper [5] aims to enhance surface water quality prediction through the evaluation of big data's impact on machine learning models. Using extensive data from major Chinese rivers and lakes (2012-2018), the study compares seven traditional and three ensemble learning models, revealing significant performance improvements with the inclusion of substantial datasets. Notably, DT, RF, and DCF outperform others across all six defined water quality levels. Two key water parameter sets (DO, CODMn, and NH3eN; CODMn, and NH3eN) are identified and validated, reinforcing the specificity of DT, RF, and DCF. The study recommends these models for future water quality monitoring due to their ability to provide timely and precise environmental warnings, enhancing prediction efficiency and reducing costs. The paper emphasizes the novel comparison of 10 learning models using extensive big data, addressing both model selection and dataset parameters. The paper highlights the practical significance of predicting water quality with fewer but indicative fundamental parameters in a unified manner.

This paper [12] introduces a novel approach to predicting surface water quality, focusing on the crucial parameters of DO and TDS. To overcome hyperparameter challenges in traditional prediction methods, the study employs particle swarm optimization (PSO) to optimize the structure, FFNN and GEP models. Using a substantial 30-year dataset from the upper Indus River, PCA identifies influential input parameters for DO and TDS prediction. The proposed hybrid PSO-FFNN and PSO-GEP models demonstrate superior accuracy, with PSO-GEP outperforming PSO-FFNN. The study's contributes to use PSO for hyper-parameter tuning, PCA in input selection, and using extensive historical data for training and testing. The results highlighted the efficacy of the models in predicting wa-

ter quality, with applications with respect to environment and its protection. The study concluded with recommendations in further developing AI and ANN techniques, also utilizing ensemble modeling, and the inclusion of spatio-temporal analysis for a more comprehensive understanding of water quality dynamics.

This paper [13] focuses on water quality prediction utilizing machine learning models and the grid search method for hyperparameter optimization. Recognizing the increasing impact of contamination on water quality, the study aims to predict WQI and WQC, crucial for assessing water validity. Employing data preprocessing, including mean imputation and normalization, the study evaluates the performance of four classification models (Random Forest, XGBoost, AdaBoost, Gradient Boosting) for WQC and four regression models (K-nearest neighbor, Decision Tree, Support Vector Regressor, Multi-Layer Perceptron) for WQI. The grid search method is applied for parameter optimization in both classification and regression models. Experimental results reveal that the Gradient Boosting model achieves the best accuracy of 99.50% for WQC, while the Multi-Layer Perceptron regressor model excels in regression with an R2 value of 99.8% for WQI prediction. The paper suggests potential future research incorporating recurrent neural networks with LSTM for time series analysis of WQI and WQC in the context of climate change variables.

This paper [14] introduces a novel method for water quality prediction using LSTM NN, addressing the practical significance of accurate water quality predictions for resource management and pollution prevention. Utilizing a dataset from Taihu Lake measured monthly from 2000 to 2006, the LSTM NN model is established and trained to improve predictive accuracy through simulations and parameter selection. Comparative analyses with BP NN and OS-ELM reveal that LSTM NN consistently outperforms the other two methods, demonstrating higher accuracy and better generalization. The study emphasizes LSTM NN's suitability for sequential prediction problems, particularly in dealing with time series data, and suggests future work to enhance memory block efficiency to address long training cycles.

With data this paper [8] focused on the Gales Creek location of the Tualatin River, Oregon, USA, the research presented two novel hybrid decision tree-based models, which are CEEMDAN-XGBoost and CEEMDAN-RF, for water quality prediction. These models also improves the performance of Hybrid Random Forest (RF) and XGBoost by preprocessing data with higher fluctuations using CEEMDAN. Six water quality indicators were used in the study, which also had weightage to Standard Deviation of Error (SDE) to assess stability and compared suggested models with the already available benchmark models. The outcome were

clearly showing that CEEMDAN-XGBoost is better with predicting water quality parameters such as pH, whereas CEEMDAN-RF performs better with predicting temperature, dissolved oxygen, and specific conductance. Moreover, The results also showed that CEEMDAN-XGBoost and CEEMDAN-RF have the best predictive performance with smaller Mean Absolute Percentage Errors (MAPEs).

This paper [11] focuses on predicting water quality in different locations across India using ML techniques, including RF, NN, MLR, SVM, and BTM. WQI that was used as a vital indicator, along with data avaliable on DO, total coliform, BOD, nitrate, pH, and electric conductivity. The study used data pre-processing, feature correlation analysis, and ML, revealing that MLR has the highest accuracy, with key water quality contributors were nitrate, pH, conductivity, dissolved oxygen, total coliform, and biological oxygen demand. Further, the model was basically used to develop a software application that can be used for real-time prediction, with promising results for better and accurate prediction on how safe the water is. They also found that MLR and RF models perform exceptionally well, more attention and further research will explore on how to efficiently combine these approaches with other models to improve prediction accuracy.

The reviewed paper [4] systematically evaluates the efficacy of standalone (RF, M5P, RT, and REPT) and hybrid data-mining algorithms for predicting the WQI in the Talar River of northern Iran. By analyzing four years of monthly data, the study identifies fecal coliform concentration as the primary determinant of WQI, followed by BOD, NO3, DO, EC, COD, (PO2)4, Turbidity, TS, and pH. Variable combinations significantly impact model performance, with higher correlation coefficients correlating with enhanced predictions. Hybrid algorithms generally outperform standalone models, with the BA-RT model exhibiting the highest accuracy. However, the paper cautions that these findings may not universally apply to other regions or datasets. The proposed BA-RT algorithm is highlighted as a reliable and cost-effective tool for groundwater quality management, especially in developing regions with limited gauging networks. The study emphasizes the importance of considering algorithm performance across diverse scenarios and encourages further research on variable combinations for improved prediction accuracy. Furthermore, insights into the influence of seasonality, streamflow, and precipitation on water quality underscore the necessity for robust and flexible WQI models, particularly in humid climates like northern Iran.

3 Data Acquisition

The following parameters are usually monitored by BWSSB (Table 1) [3,10]. The acceptable limits are also mentioned. The dataset utilized in this study comprises 3,276 samples sourced from water quality assessments. It encompasses ten distinct parameters: pH, Hardness, Solids, Chloramines, Sulfate, Conductivity, Organic Carbon, Trihalomethanes, Turbidity, and Potability(Table 2). The dataset, procured from Kaggle, is a comprehensive resource for investigating water conditions. Each sample's attributes contribute to an intricate understanding of water quality, facilitating the prediction of potability

Table 1: Parameters Monitored

S. No.	Parameters	Units	Acceptable Limit
1	pH	-	6.5-8.5
2	Turbidity	NTU	≤ 5
3	Color	Hazen units	≤ 5
4	Total Hardness	mg/L	≤ 200
5	Alkalinity	mg/L	≤ 200
6	TDS	mg/L	≤ 500
7	Conductivity	μS/cm	
8	Calcium	mg/L	≤ 75
9	Magnesium	mg/L	≤ 30
10	Chloride	mg/L	≤ 250
11	Residual Free Chlorine	mg/L	
12	Residual Aluminium	mg/L	≤ 0.03
13	Total Coliform	MPN/100ml	≤ 2
14	Fecal Coliform	MPN/100ml	0/100
15	Iron	mg/L	0.3
16	Residual Chlorine @ Tataguni	mg/L	

4 Data Preparation

4.1 Missing Values Handling

Dealing with null data or missing values is a frequently encountered obstacle when working with datasets [2,5,7]. Failure to address these missing values can result in inaccuracies during data analysis or modeling. [13,14] Consequently, it is of paramount significance to proficiently handle null data by either replacing them with specified values or entirely removing them. In this specific scenario, the fillna() method provided by the pandas library is employed to manage null

Table 2: Data Set Parameters

Variable	**Description**
pH value	pH is crucial for assessing the acid-base equilibrium of water. It serves as an indicator of whether water is acidic or alkaline.
Hardness	Hardness is primarily attributed presence of calcium and magnesium salts obtained from geologic deposits.
Solids	Water can dissolve a wide range of inorganic and some organic salts such as K, Ca, Na, Mg, chlorides, sulfates etc.
Chloramines	Chlorine and chloramine are primarily used as disinfectants for public water systems.
Sulfate	Sulfates are naturally occurring compounds commonly found in minerals, soil, and rocks.
Conductivity	Pure water exhibits low electric conductivity and primarily serves as a good insulator. Presence of ions can significantly increase conductivity.
Organic carbon	Total Organic Carbon (TOC) in source waters is obtained from decaying natural organic matter (NOM) and synthetic sources.
Trihalomethanes	THMs are chemicals that can potentially be present in water treated with chlorine.
Turbidity	The turbidity of water is determined by the amount of solid matter present in the suspended state.
Potability	Indicates if water is safe for human consumption where 1 means Potable and 0 means Not potable.

values. More precisely, in this instance, the null values are substituted with the mean of each respective column. (Refer to TABLE 3)

Table 3: Null Values Handling

ph	491
Hardness	0
Solids	0
Chloramines	0
Sulfate	781
Conductivity	0
Organic carbon	0
Trihalomethanes	162
Turbidity	0
Potability	0
dtype	int64

ph	0
Hardness	0
Solids	0
Chloramines	0
Sulfate	0
Conductivity	0
Organic carbon	0
Trihalomethanes	0
Turbidity	0
Potability	0
dtype	int64

4.2 Dimensionality Reduction

Dimensionality reduction involves decreasing the count of variables or features within a dataset [1, 9]. This proves valuable when the dataset contains an exten-

sive array of variables, and certain variables may hold little relevance for analysis or modeling. A heat map was employed in our code implementation to assess the correlations among distinct input attributes. The outcome revealed minimal correlations between features. Consequently, all input attributes were retained for model training due to the absence of strong correlations. (Refer to Fig 1)

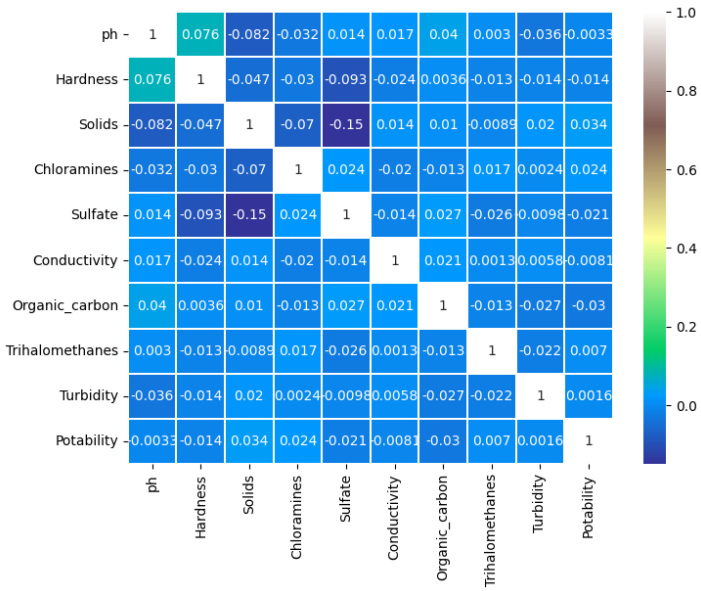

Figure 1: Dimensionality Reduction

4.3 Outlier Checking

Outliers denote data points that markedly deviate from the rest of the dataset [6]. Their presence can detrimentally affect analytical outcomes and modeling results. Hence, it's imperative to detect and address outliers prudently. In the implemented code, outliers are identified through box plots. Specifically, within the "Solids" input attribute, outliers are detected. However, these outliers are retained rather than removed, considering their potential significance and influence on the final output. (Refer to Fig 2)

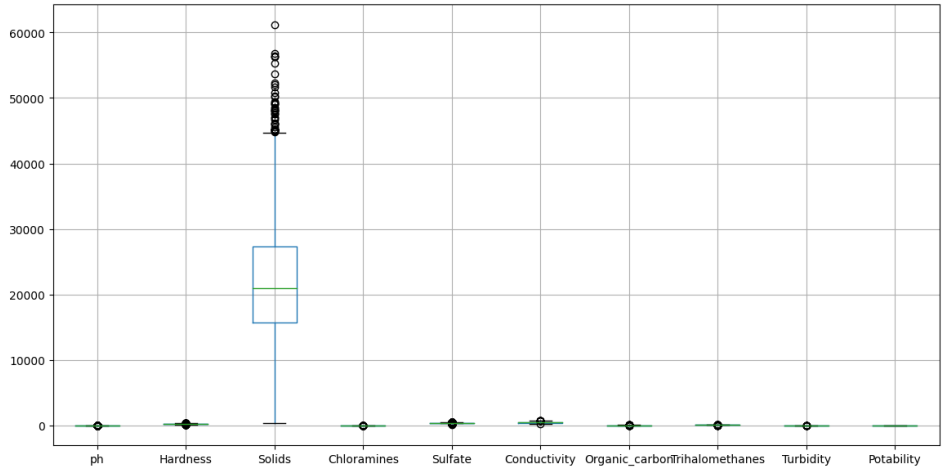

Figure 2: Outlier Checking

4.4 Normalization

Normalization stands as a pivotal stride in data preprocessing for machine learning [11,12]. This process entails rescaling numerical data to a standardized interval, typically from 0 to 1. The aim is to circumvent any undue partiality towards specific attributes within the model. Moreover, normalization bolsters the efficacy and precision of numerous machine-learning algorithms. In our pursuit, we gauged the dataset's normalization status through histogram plots encompassing all input attributes. Gratifyingly, the histograms portrayed normalized distributions across input attributes. Consequently, we opted not to employ additional normalization techniques, as the data was already appropriately scaled. (Refer to Fig 3)

5 Model Training

The train test split function within the sklearn library facilitates the division of data into distinct training and testing sets. By segmenting the data, we are able to train the model using a subset of the data, whilst evaluating its performance with an entirely separate subset. This methodology serves to safeguard against overfitting, a situation in which the model merely memorizes the training data rather than comprehending underlying patterns that can be applied to novel data. We have employed the subsequent machine learning models to train our model:

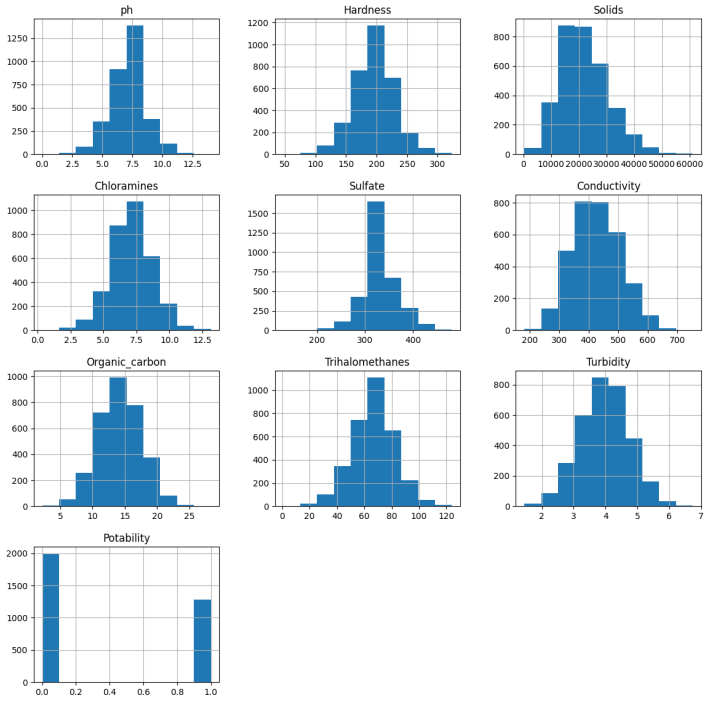

Figure 3: Normalization

5.1 Decision Tree Classifier

Decision Trees represent a supervised learning algorithm applicable to regression and classification tasks. This methodology constructs a tree-shaped model illustrating decisions and their corresponding potential outcomes. The tree's composition entails internal nodes signifying decisions, branches embodying decision outcomes, and terminal leaf nodes encapsulating ultimate predictions.(Refer to Table 4)

Table 4: Decision Tree Classifier

Accuracy using DT	59.45
Accuracy On Training Set Using DT	0.680
Accuracy On Test Set Using DT	0.681
Confusion Matrix	{{379, 186},{23, 68}}

5.2 K-Nearest Neighbors

KNN, a supervised learning algorithm adaptable to regression and classification scenarios. Its methodology involves identifying the K closest training data points to a novel data point, subsequently leveraging their labels to anticipate the label of the new point. (Refer to Table 5)

Table 5: KNN

Accuracy using KNN	68.14
Accuracy On Training Set Using KNN	0.680
Accuracy On Test Set Using KNN	0.681
Confusion Matrix	{{379, 186},{23, 68}}

5.3 Random Forest Classifier

RF enploys an assemblage of decision trees to enhance prediction accuracy. This approach involves generating numerous decision trees using distinct random subsets of the training data and amalgamating their predictions to yield a definitive outcome.(Refer to Table 6)

Table 6: Random Forest Classifier

Accuracy using RF	61.28048780487805
Accuracy On Training Set Using RF	0.609
Accuracy On Test Set Using RF	0.613
Confusion Matrix	{{402, 254},{0, 0}}

5.4 Logistic Regression

Logistic regression is a statistical approach that depicts the connection between a categorical or binary dependent variable and one or more independent variables. Unlike linear regression, which characterizes the link between a continuous dependent variable and independent variables, logistic regression gauges the likelihood of an event's occurrence. The objective is to determine the optimal fitting boundary that distinguishes between the two classes of the dependent variable. Within our constructed model, we have implemented the L1 regularization technique.(Refer to Table 7)

Table 7: Logistics Regression Classifier

Accuracy using LR with regularization	61.280
Confusion Matrix	{{402, 0},{254, 0}}
Prediction ”0” {Precision, Recall, F1-Score, Support}	{0.61, 1.00, 0.76, 402}
Prediction ”1” {Precision, Recall, F1-Score, Support}	{0.00, 0.00, 0.00, 254}
Accuracy {Precision, Recall, F1-Score, Support}	{–, –, 0.61, 656}
Macro Avg {Precision, Recall, F1-Score, Support}	{0.31, 0.50, 0.38, 656}
Weighted Avg {Precision, Recall, F1-Score, Support}	{0.38, 0.61, 0.47, 656}

5.5 Stochastic Gradient

Stochastic gradient constitutes a methodology harnessed in machine learning and optimization, facilitating the parameter adjustments of a model grounded on a subset of the training data instead of the complete dataset. Notably, SGD's potency lies in its swift efficiency, mainly when dealing with expansive datasets, as it only computes gradients for a fraction of the data. Nonetheless, its stability can be compromised, potentially necessitating additional iterations for convergence in contrast to conventional gradient descent.

Following data imputation and train-test partitioning, we instantiated an SGDClassifier(). This instantiation embraced parameters such as loss='log' to employ logistic regression, random_state=42 to ensure reproducibility, max_iter= 1000 for training iterations, and tol=1e-3 for convergence tolerance. Model training on the training set was facilitated using fit(). Subsequently, predictions on the test set were generated via predict(). Lastly, the model's performance was quantified employing accuracy_score().(Refer to Table 8)

Table 8: Stochastic Gradient

Accuracy using SGD	61.59
Confusion Matrix	{{394, 8},{244, 10}}

5.6 Support Vector Machine

SVM ranks among the most widely embraced Supervised Learning methodologies, adept in tackling both Classification and Regression quandaries. SVM's primary objective entails forging an optimal line or decision boundary capable of partitioning n-dimensional space into distinct classes, streamlining future categorization of novel data points. This preeminent decision boundary, denominated as a hyperplane, is meticulously devised by SVM, which selects pivotal extremities/vectors contributing to its formulation. These critical instances, recognized

as support vectors, underscore the moniker of the algorithm—Support Vector Machine. (Refer to Table 9)

Table 9: Support Vector Machine

Accuracy using SVM	68.75
Confusion Matrix	{{375, 27},{178, 76}}

5.7 Back Propagation Algorithm

Backpropagation encompasses an algorithm that retrogresses errors from output nodes to input nodes. Our model adopts a streamlined neural network configuration featuring dual hidden layers and an output layer enhanced by a sigmoid activation function. The model's refinement revolves around training via the binary cross-entropy loss function and the Adam optimization algorithm. A batch size of 32 is employed during training, spanning 50 epochs. Post-training, model appraisal unfolds on the test set to furnish an approximate gauge of its efficacy. The model's test set accuracy is extractable via the test_acc variable. (Refer to Table 10)

Table 10: Support Vector Machine

Accuracy using Back Propagation	66.92

6 Model Optimization and Evaluation

Hyperparameter tuning entails the pursuit of the optimal amalgamation of hyperparameters within a machine learning algorithm, aiming to maximize its effectiveness in a specified task. These hyperparameters, distinct from learned parameters, are often predetermined prior to the training process. [2] GridSearchCV stands as a prevalent technique for hyperparameter tuning. Its methodology involves a thorough exploration across a predefined hyperparameter spectrum, evaluating model performance across all permutations. This assessment is commonly executed through cross-validation, partitioning data into multiple folds for iterative model training and evaluation. GridSearchCV ultimately furnishes the optimal hyperparameter amalgamation, optimizing the model performance on the validation set. [3] Stratified cross-validation emerges as a variant of cross-validation that upholds class balance across folds, akin to the original dataset's composition. This stratification proves vital for imbalanced datasets where one class may be underrepresented. By mitigating bias towards the dominant class,

stratified cross-validation furnishes more robust performance estimates for the model. (Refer to Table 11, 12, 13)

Table 11: Hyper Parameter Tuning for DT

Accuracy using DT after HPT	60.0609756097561
Best Paramaters	{'criterion': 'gini', 'min_samples_split': 29, 'splitter': 'random'}
Confusion Matrix	{{293, 153},{109, 101}}

Table 12: Hyper Parameter Tuning for KNN

Accuracy using KNN after HPT	67.07317073170732
Best Paramaters	{'metric': 'manhattan', 'n_neighbors': 16, 'weights': 'distance'}
Confusion Matrix	{{358, 172},{ 44, 82}}

Table 13: Hyper Parameter Tuning for RF

Accuracy using RF after HPT	66.3109756097561
Best Paramaters	{'criterion':'gini', 'max_depth': 7, 'max_features':'sqrt', 'min_samples_-split': 2, 'n_estimators': 100}
Confusion Matrix	{{384, 203},{18, 51}}

7 Prediction

We utilized various machine learning algorithms, including DT, KNN, RF, LR, SVM, SGD, and BP, to predict the output. Following the collection of predictions from each model, we employed a majority voting approach to determine the final prediction for portability. If four or more out of the seven models indicate a potability of 1, the final prediction is set as 1; otherwise, it is set as 0. (Refer to Table 14)

Table 14: Prediction

DT prediction	[0]
KNN prediction	[1]
RF prediction	[0]
LR prediction	[0]
SVM prediction	[0]
SGD prediction	[0]
BP prediction	[0]
Protablity	[1]

8 Conclusion

Based on the analysis and modeling undertaken, it can be deduced that machine learning methodologies can be effectively employed to forecast water quality using parameters like pH, dissolved oxygen, and biochemical oxygen demand. Among the models evaluated, the Support Vector Machine approach demonstrated superior performance, yielding an impressive accuracy rate of 68.75% and minimal error rates. Consequently, the Support Vector Machine (SVM) algorithm emerges as a viable option for predicting water quality with considerable efficacy. Enhancing the model could involve the integration of supplementary attributes such as weather data, location-specific influencers, and seasonal fluctuations.

These models hold substantial implications for the management of water resources, delivering a cost-efficient and proficient means of monitoring water quality and pinpointing potential sources of pollution. This, in turn, could facilitate the formulating of proactive strategies for safeguarding water conservation and averting pollution. The ultimate result would be the assurance of a secure and sustainable supply of clean water resources for communities.

9 Limitations

There are several potential limitations in our presented code. Firstly, the efficacy of the models depends upon the quality of the input data; any inaccuracies, outliers, or biases may impact performance. The absence of explicit feature engineering steps is another consideration, as careful feature manipulation can significantly enhance a model's ability to discern complex patterns. Imbalanced data could pose a challenge, and addressing this issue using techniques like oversampling or different evaluation metrics may be necessary. The potential for overfitting, especially in complex models, and the interpretability of certain models, particularly neural networks and ensembles, are also noteworthy limitations. Finally, the generalization of the models to new, unseen data and the computational resources required for training certain models should be carefully considered. Addressing these limitations is crucial for building reliable and robust models in practical applications.

10 Future Scope

This paper suggests several promising directions for future research in machine learning-based water potability prediction. These include integrating additional data sources, exploring advanced feature engineering techniques, experimenting with ensemble methods and model stacking, and assessing the model's generalizability across diverse regions. Practical deployment, examination of long-term trends, incorporation of explainable AI techniques, collaboration with stakeholders, real-time monitoring using IoT devices, examination of uncertainties in model predictions, and integration of supplementary attributes like weather data are also important areas for future research. This comprehensive approach aims to refine and apply machine learning models in safeguarding water quality and supporting sustainable water resource management.

References

[1] Ali Najah Ahmed, Faridah Binti Othman, Haitham Abdulmohsin Afan, Rusul Khaleel Ibrahim, Chow Ming Fai, Md Shabbir Hossain, Mohammad Ehteram, and Ahmed Elshafie. "machine learning methods for better water quality prediction". *Journal of Hydrology*, 578, 11 2019.

[2] Ali Omran Al-Sulttani, Mustafa Al-Mukhtar, Ali B. Roomi, Aitazaz Ahsan Farooque, Khaled Mohamed Khedher, and Zaher Mundher Yaseen. Proposition of new ensemble data-intelligence models for surface water quality prediction. *IEEE Access*, 9:108527–108541, 2021.

[3] Bilal Aslam, Ahsen Maqsoom, Ali Hassan Cheema, Fahim Ullah, Abdullah Alharbi, and Muhammad Imran. "water quality management using hybrid machine learning and data mining algorithms: An indexing approach". *IEEE Access*, 10:119692–119705, 2022.

[4] Duie Tien Bui, Khabat Khosravi, John Tiefenbacher, Hoang Nguyen, and Nerantzis Kazakis. Improving prediction of water quality indices using novel hybrid machine-learning algorithms. *Science of the Total Environment*, 721, 6 2020.

[5] Kangyang Chen, Ruqin Shen, Fengrui Liu, Min Zuo, Xinyi Zou, Jinfeng Wang, Yan Zhang, Da Chen, Xingguo Chen, Yongfeng Deng, and Hongqiang Ren. "comparative analysis of surface water quality prediction performance and identification of key water parameters using different machine learning models based on big data". *Water Research*, 171, 3 2020.

[6] Tianan Deng, Kwok Wing Chau, and Huan Feng Duan. Machine learning based marine water quality prediction for coastal hydro-environment management. *Journal of Environmental Management*, 284, 4 2021.

[7] Chunmei Ding. Principal component analysis of water quality monitoring data in xiasha region. pp. 2321–2324. IEEE, 6 2011.

in: *Advancements in Communication and Systems*. Ed. by Ashish Kumar Tripathi and Vivek Shrivastava. Computing and Intelligent Systems, SCRS, India., 2023, pp. 37–54. DOI: https://doi.org/10.56155/978-81-955020-7-3-4

[8] Md. Mehedi Hassan, Md. Mahedi Hassan, Laboni Akter, Md. Mushfiqur Rahman, Sadika Zaman, Khan Md. Hasib, Nusrat Jahan, Raisun Nasa Smrity, Jerin Farhana, M. Raihan, and Swarnali Mollick. Efficient prediction of water quality index (wqi) using machine learning algorithms. *Human-Centric Intelligent Systems*, 1:86, 2021.

[9] Ezz El-Din Hemdan, Youssef M. Essa, Ayman El-Sayed, Marwa Shouman, and Abdullah N. Moustafa. Smart water quality analysis using iot and big data analytics: A review. pp. 1–5. IEEE, 7 2021.

[10] Harish H. Kenchannavar, Prasad M. Pujar, Raviraj M. Kulkarni, and Umakant P. Kulkarni. "evaluation and analysis of goodness of fit for water quality parameters using linear regression through the internet-of-things-based water quality monitoring system". *IEEE Internet of Things Journal*, 9:14400–14407, 8 2022.

[11] Hongfang Lu and Xin Ma. Hybrid decision tree-based machine learning models for short-term water quality prediction. *Chemosphere*, 249, 6 2020.

[12] Muhammad Izhar Shah, Muhammad Faisal Javed, Abdulaziz Alqahtani, and Ali Aldrees. Environmental assessment based surface water quality prediction using hyper-parameter optimized machine learning models based on consistent big data. *Process Safety and Environmental Protection*, 151:324–340, 7 2021.

[13] Mahmoud Y. Shams, Ahmed M. Elshewey, El Sayed M. El-kenawy, Abdelhameed Ibrahim, Fatma M. Talaat, and Zahraa Tarek. "water quality prediction using machine learning models based on grid search method". *Multimedia Tools and Applications*, 2023.

[14] Yuanyuan Wang, Jian Zhou, Kejia Chen, Yunyun Wang, and Linfeng Liu. Water quality prediction method based on LSTM neural network. pp. 1–5. IEEE, 11 2017.