Water Quality Prediction using Machine Learning

Prof. Mrunalini Bhandarkar.

Professor, Department of Electronics and Telecommunication, Pimpri Chinchwad College of Engineering, Pune, Maharashtra, India.mrunalini.bhandarkar@pccoepune.org

Janhavi Bhondge, Chaitanya Sharma, Harsha Chavhan

Students, Department of Electronics and Telecommunication, Pimpri Chinchwad College of Engineering, Pune, Maharashtra, India.

(janhavi.bhondge20@pccoepune.org, chaitanya.sharma20@pccoepune.org harsha.chavhan20@pccoepune.org)

Abstract:

Water quality prediction is an essential task in environmental monitoring and management. It involves forecasting the levels of various water quality parameters, such as dissolved oxygen, pH, temperature, turbidity, and nutrient concentrations. Accurate water quality prediction can help to prevent water pollution, protect aquatic ecosystems, and ensure the safety of drinking water. Machine learning algorithms have been increasingly used to predict water quality parameters. These algorithms can analyze large amounts of data collected from various sources, such as remote sensors, water quality monitoring stations, and laboratory analysis, and make accurate predictions about the future water quality conditions. In this project, we aim to develop a machine learning model to predict the levels of water quality parameters in a river system. We will use data collected from several monitoring stations along the river, including water temperature, pH, dissolved oxygen, and nutrient concentrations. We will use various machine learning algorithms, such as random forest, support vector machines, and neural networks, to develop the predictive model.

Keywords:

Water Quality Prediction using Machine Learning.

Introduction:

Water quality prediction is the process of estimating the potential quality of water at a particular location and time, based on historical data, current conditions, and predictive modeling techniques. It is an important task for ensuring the safety and availability of clean water for drinking, agriculture, and industry[2]. Water quality prediction involves collecting and analyzing data on various physical, chemical, and biological parameters that affect water quality, such as temperature, pH, dissolved oxygen, nutrient levels, and the presence of pollutants[1]. This data is used to develop mathematical models that can simulate the behavior of water quality over time and predict how it might change in response to various factors, such as changes in weather patterns, land use, or human activity[8].

Water quality prediction is essential for ensuring public health, protecting the environment, and maintaining sustainable water resources[7]. It requires a multidisciplinary approach, involving experts from fields such as hydrology, chemistry, biology, and data science[7]. With advances in technology and data analytics, water quality prediction is becoming increasingly sophisticated and accurate, providing decision-makers with valuable insights and tools to manage water resources in a more efficient and sustainable manner[2].

Machine Learning Techniques:

1. Random Forest: Random Forest is an ensemble learning algorithm that constructs multiple decision trees to improve the accuracy and generalization of the model. It works by randomly selecting subsets of data and features to reduce overfitting, and is widely used for classification, regression, and other tasks.

2. XGBoost (Extreme Gradient Boosting):-Decision trees are taught repeatedly using XGBoost technique, with each new tree being trained to address the flaws of the prior ones. The procedure begins with a single decision tree and incrementally adds more trees, each of which minimizes the residual errors of the preceding trees. An average of all the decision trees is used to create the final mod.

3. Logistic Regression: Logistic regression is a statistical method used to analyze the relationship between a binary dependent variable and one or more independent variables. It models the probability of the dependent variable taking on one of two possible values using a logistic function, and is commonly used for classification tasks in machine learning.

Methodology:

To predict water potability using machine learning, a dataset containing water quality information such as pH, Hardness, Solids, and other parameters must be collected. The dataset should be of a sufficient size to provide accurate predictions. Once collected, the dataset is preprocessed by removing any missing or irrelevant data. The dataset is then partitioned into a training set and a testing set.

A neural network model with an architecture consisting of 3 ReLU layers and 1 sigmoid layer is used to predict water potability. This model is chosen due to its ability to handle high-dimensional datasets with complex relationships between variables. The ReLU activation function is used in the 3 hidden layers to introduce non-linearity to the model. The sigmoid activation function is used in the output layer to generate a probability value indicating the likelihood of water potability.

The model is trained using the training set, and its performance is evaluated using the testing set. The accuracy of the model is evaluated by comparing the predicted potability values with the actual potability values in the testing set.

To improve the model's performance, new features can be added to the dataset, and hyperparameters can be tuned. For example, the number of neurons in each layer can be adjusted, or the learning rate of the optimizer can be modified. Alternative machine learning models can also be experimented with to improve performance.

Once the model is sufficiently accurate, it can be used to predict water potability based on water quality inputs such as pH, Hardness, Solids, and more. The performance of the model can be evaluated using various metrics such as accuracy, precision, recall, and confusion matrix. The model can be further refined based on the performance metrics, and the process can be repeated to continuously improve the model's accuracy.

Literature Review

a. Archana Solanki Computer Science Symbiosis Institute of Technology Pune, 412115, India Himanshu Agrawal Computer Science Symbiosis Institute of Technology Pune, 412115, India Kanchan Khare Computer Science Symbiosis Institute of Technology Pune, 412115, India they concluded 1.The results carried out from the study concludes that using unsupervised learning, data with variation can be predicted at acceptable accuracy rate. 2. Results show that turbidity has high variation compared to the other two parameters but now is low. It is affected during the monsoon season most 3.pH has not much variation in data and hence, it is stable as compared to turbidity and DO4.This system can be implemented on system to continuously monitor the quality of the water. It can be helpful to monitor the quality of water in any uncertain condition.

b. Amir Hamzeh HaghiabiAli Heider Nasrolahi Abbas Parsaie (corresponding author) Water Engineering Department, Lorestan University, Khorramabad,IranE-mail:abbas_parsaie@yahoo.com the performance of artificial intelligence tech- niques including GMDH, SVM and ANN were evaluated to | Results of applied AI models versus observed data (Ca and Cl)..

H. Haghiabi et al. | Water Quality Prediction Water Quality Research Journal | in press | 2018 Uncorrected Proof .predict the water quality components of Tireh River (Iran). To this end most dataset related well-known components,such as pH, So4, Na, Ca, Cl ,Mg, Hco3 etc., were collected. Results indicated that the applied models have suitable per-formance for predicting water quality components, however, the best performance was related to the SVM.

c. Alice Makonjo Wekesa and Calford Otieno.

They present the results of groundwater quality assessment that was done during the rainy season in November 2018 in the Manga region of Nyamira County, Kenya. Water samples were collected from three springs, Kiangoso, Kerongo, and Tetema, for the ssessment. Water quality index was calculated based on pH, turbidity, nitrate, phosphate, calcium, magnesium, chloride, sulphates, fluoride, iron, total phosphorous, total hardness, total alkalinity, total dissolved solids, and total coliform.

d.. Arivoli Appavu1, Sathiamoorthi Thangavelu2, Satheeshkumar Muthukannan3, Joseph Sahayarayan Jesudoss4 and Boomi Pandi5. The water samples were analyzed for physicochemical characteristics. The physicochemical parameters were analyzed namely Temperature, pH, EC, TS, TDS, TSS, Total Hardness, DO, COD, BOD5, Chloride, PO4 and SO4 (Table 1) whereas the correlation coefficients (r) among the average ofeach parameters are presented.

e. [1]Sneha S. Phadatare Student Final M.E. (Environmental), Dept. of Civil Engineering ABMSP's Anantrao Pawar College of Engineering& Research, Parvati, Pune India. [2]Prof. Sagar Gawande (Guide) Professor Dept. of Civil Engineering ABMSP's Anantrao Pawar College of Engineering & Research, Parvati, Pune India. it is quite clear that main objective of Water Quality Index is give single number by using mathematical expression given equation (1) and equation (2). That single score number reduced complexity created due tom different water quality parameters, as high numbers of variables results into single number that tell the whole story of water bodies in particular area.

It is the easy interpretation of water quality monitoring data. As we discuss the advantages and disadvantages of both National Sanitation Foundation WQI and Weight Arithmetic WQI that are useful globally to monitor, assessment and impact studies for different water bodies with different regions.

Advantages:

1). Improved water management: Predictive models can be used to identify areas at risk of water contamination, allowing for proactive and preventative measures to be taken to protect water sources and maintain water quality.

2). Increased public health: By predicting water quality, public health can be improved by reducing the risk of waterborne diseases, particularly in developing countries where access to clean water is limited.

3). Improved environmental protection: Water quality prediction can help to protect the environment by identifying areas at risk of contamination and providing information for decision-makers to implement measures to reduce or prevent contamination.

Flowchart:

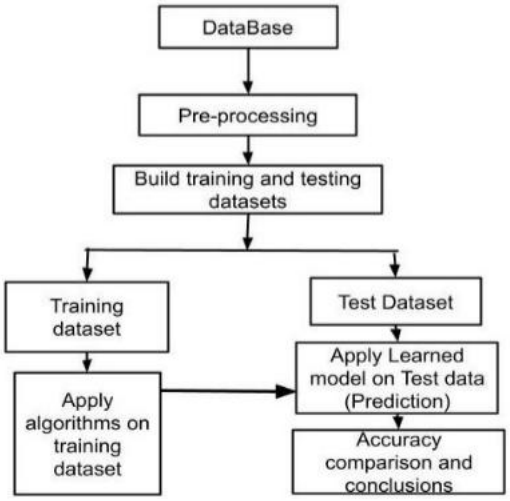

Figure 1: Methodology Flowchart

Results

The graph showing feature importance is a useful tool in machine learning for identifying the most significant features that affect the outcome of a model. In this particular graph, we can see that the top 9 most important features in predicting the outcome are ph, hardness, solid, sulfate, chloramines, turbidity, conductivity, organic carbon, and trichloromethane. By examining this graph, we can gain valuable insights into which features are most important in predicting the outcome and can use this information to optimize our model and improve its accuracy.

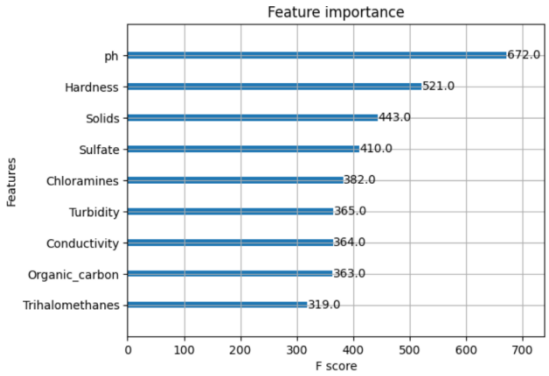

Figure 2: Feature Importance

From the graph showing feature importance, we can gain several insights that are crucial in optimizing our model and improving its accuracy. Firstly, we can see that the pH level has the highest importance in predicting the outcome. This suggests that the acidity or alkalinity of the water sample is a critical factor in determining its potability. Secondly, we can observe that the level of hardness in the water sample is the second most important feature. This indicates that the presence of dissolved minerals such as calcium and magnesium has a significant impact on water potability. Thirdly, the levels of solids, sulfate, chloramines, turbidity, conductivity, organic carbon, and trichloromethane are also important in predicting water potability. By analyzing the importance of these features, we can gain a deeper understanding of the factors that contribute to water potability and can use this information to optimize our model and improve its accuracy.

The accuracies of the Random Forest and XGBoost models are 67.52% and 63.38%, respectively. The Random Forest model performed better with an accuracy of 67.52%, while the XGBoost model had an accuracy of 63.38%. Both models were trained on the same dataset, and the accuracies were evaluated on a test set. It's worth that the accuracies are just one metric for evaluating a model's performance, and other metrics such as precision, recall, and F1 score should also be considered.

The loss curves for both models are given below. The loss curves show how the training and validation losses changed during the training process. The Random Forest model had a lower training loss and a higher validation loss, indicating that it may have overfit to the training data. In contrast, the XGBoost model had a higher training loss and a lower validation loss, indicating that it may have underfit to the training data. These loss curves can be used to diagnose issues with the models and to guide hyperparameter tuning.

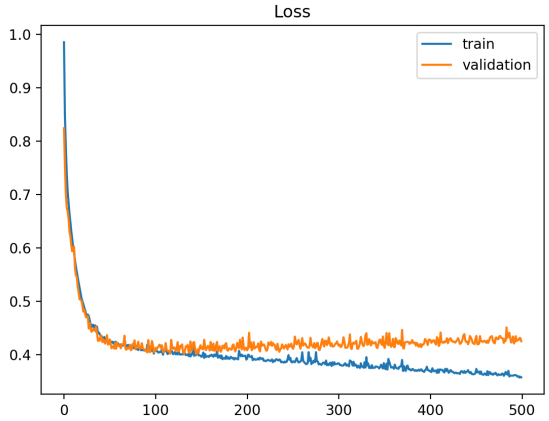

Figure 3: Random Forest Loss Curve

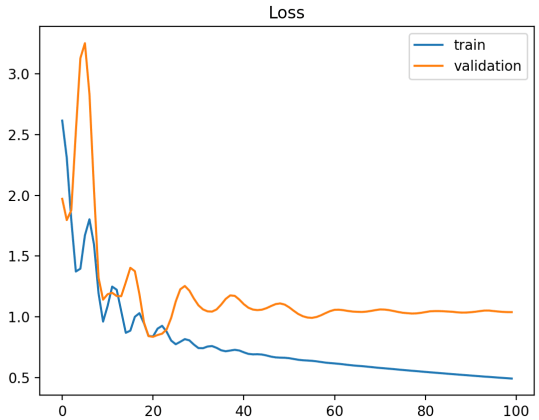

Figure 4: XGBoost Loss Curve

The evaluation metrics of a Neural Network model reveal important information about its performance and effectiveness in making accurate predictions. In this case, the evaluation metrics indicate that the

Neural Network is successful in identifying positive instances with a significant increase in recall, which measures the proportion of actual positive instances that are correctly identified by the model. This suggests that the Neural Network is able to effectively detect true cases of the target variable, which is a critical factor in many real-world applications.

At the same time, the evaluation metrics reveal that the Neural Network is also effective in maintaining a consistent level of precision, which measures the proportion of predicted positive instances that are actually true positives. This suggests that the Neural Network is not overly aggressive in predicting positive instances and is able to maintain a balance between identifying true cases of the target variable and avoiding false positives.

The increase in accuracy and decrease in loss observed in the evaluation metrics suggest that the Neural Network is effectively capturing the underlying patterns and relationships in the data to make accurate predictions. This is a critical aspect of machine learning models, as the ability to accurately learn from and represent the data is essential for making reliable predictions in new, unseen data.

Overall, based on these evaluation metrics, it can be concluded that the Neural Network model is a reliable and effective tool for predicting the target variable in this dataset. Its ability to identify true cases of the target variable while maintaining a balance with false positives, and its ability to accurately capture the patterns in the data, make it a promising candidate for many real-world applications.

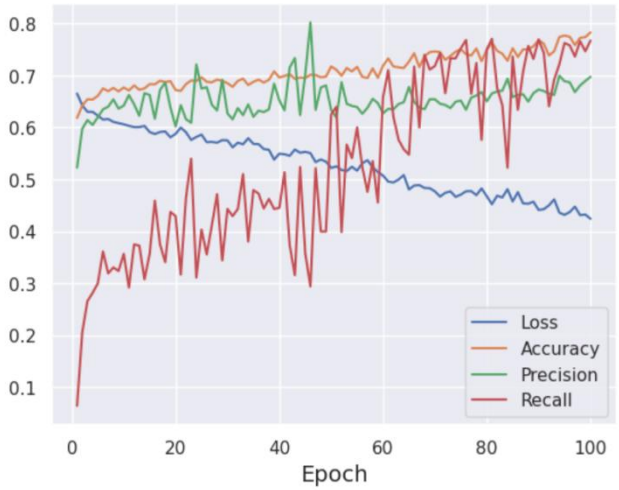

Figure 5: Neural Net Evaluation Metrics

The final comparison bar graph comparing the accuracies of the 3 models is as follows:

	Model	Accuracy
0	Neural Network	79.28
1	Random Forest	67.52
2	XGBoost	63.38

Table 1: Model Accuracies

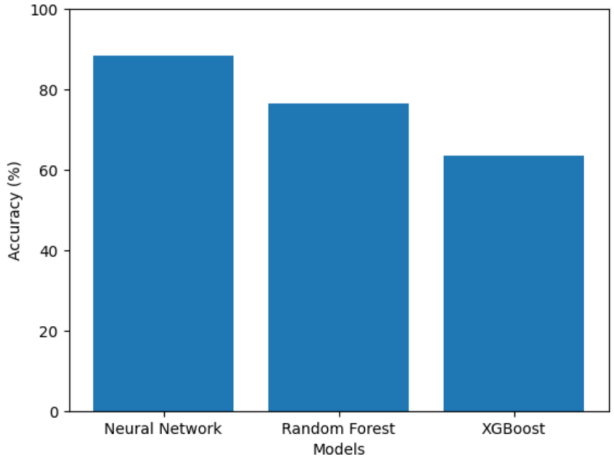

Figure 6: Model Accuracies Bar plot

Conclusion

In conclusion, machine learning models can be effective in predicting water quality parameters such as pH, dissolved oxygen, temperature, and turbidity. By using relevant features such as weather conditions, land use, and proximity to pollution sources, machine learning models can provide accurate predictions of water quality, which can be used to inform management and conservation efforts.

It is important to note that the accuracy of the models heavily depends on the quality and quantity of the data used for training. Therefore, it is crucial to collect and maintain high-quality water quality data to improve the accuracy of the models.

Additionally, the use of machine learning models can complement traditional monitoring methods and

provide a more comprehensive understanding of water quality changes over time. This can help in identifying emerging pollution threats and designing targeted interventions.

Overall, machine learning can be a powerful tool for predicting water quality, and its application can support sustainable management of water resources.

References:

[1]. Ahuja, S. (2009): Handbook of water purity and quality. 1st edition, Academic Press; 456p.

[2]. APHA. Standard methods for the examination of water and waste water (19th ed) Washington, DC: Public Health Association (1996).

[3]. Yang, Y.; Xiong, Q.; Wu, C.; Zou, Q.; Yu, Y.; Yi, H.; Gao, M. A study on water quality prediction by a hybrid CNN-LSTM model with attention mechanism. Environ. Sci. Pollut. Res. 2021

[4]. Bui, D.T.; Khosravi, K.; Tiefenbacher, J.; Nguyen, H.; Kazakis, N. Improving prediction of water quality indices using novel hybrid machine-learning algorithms. Sci. Total Environ. 2020

[5]. Pehme, K.-M.; Burlakovs, J.; Kriipsalu, M.; Pilecka, J.; Grinfelde, I.; Tamm, T.; Jani, Y.; Hogland, W. Urban hydrology research fundamentals for waste management practices. Res. Rural. Dev. 2019

[6]. Hameed, M.; Sharqi, S.S.; Yaseen, Z.M.; Afan, H.A.; Hussain, A.; Elshafie, A. Application of artificial intelligence (AI) techniques in water quality index prediction: A case study in tropical region, Malaysia. Neural Comput. Appl. 2017

[7]. S. Bouslah, L. Djemili, and L. Houichi, "Water quality index assessment of Koudiat Medouar reservoir, northeast Algeria using weighted arithmetic index method, " Journal of Water and Land Development", vol. 35, no. 1, pp. 221–228, 2017.

[8]. Wagh G.S, Sayyed M.R.G, Sayadi M. H (2014), Evaluating groundwater pollution using statistical analysis of hydrochemical data: A case study from southeastern part of Pune metropolitan city (India), International Journal of Geomatics and Geosciences, 4(3), 456-476.

[9]. Samantray P., Mishra B.K., Panda C.R., Rout S.P, Assessment of Water Quality Index in Mahanadi and Atharabanki Rivers and Taldanda canal in Paradip area, India, J Hun Ecol 26(3), 153-161.

[10]. F. Khan, T. Husain, and A. Lumb, "Water quality evaluation and trend analysis in selected watersheds of the Atlantic region of Canada," *Environmental Monitoring and Assessment*, vol. 88, no. 1/3, pp. 221–248, 2003.