Assessment of Water Quality Prediction Models: A Comparative Study of Machine Learning Algorithms

1. Mr.K. Ravi Chand, 2.Ch. Bhavana, 3.N. Gayathri Lakshmi, 4.D. Sireesha,

5. G. Anuradha, 6.V. Mallika

#1Assistant Professor in Department of CSE, in VEC, KAVALI.

#2#3#4#5#6 B. Tech with Specialization of Computer Science and Engineering in VEC, KAVALI

ABSTRACT_ Water is one of the most valuable natural resources ever provided to humans. Water quality has a direct impact on the ecosystem as well as human health. Water is utilized for a variety of purposes, including drinking, agricultural, and industrial applications. Throughout the years, several pollutants have jeopardized water quality. Predicting and evaluating water quality is now critical to mitigating water pollution as a result. Real-time monitoring is ineffective because water quality is often analyzed using costly laboratory and statistical techniques. A more practical and cost-effective solution is required due to poor water quality. Using machine learning techniques, the proposed approach creates a model capable of forecasting the water quality index and class. . The goal of this suggested system is to create a novel way for classifying water quality using a Gradient Boosting Classifier. The method comprises the calculation of the Water Quality Index, which is used to assess water quality. The proposed technique has a high Train Accuracy of 98% and a Test Accuracy of 94%. The method divides water into categories based on several water quality metrics and properties such as pH, dissolved oxygen, temperature, and electrical conductivity. The model established in this work can forecast water quality as Excellent, Good, Poor, or Very Poor, allowing for real-time monitoring and management of water quality.

The results show that the suggested strategy is effective and accurate at forecasting water quality, demonstrating the potential of machine learning approaches for monitoring and management. The proposed approach can be employed in numerous applications, such as water treatment, environmental monitoring, and aquatic life management.

1.INTRODUCTION

Water is quite possibly of the most basic regular asset that assumes an imperative part

in supporting life on the planet. It is utilized for many purposes, including drinking, water system, modern purposes, and oceanic life support. Notwithstanding, the nature of water is frequently compromised because of different contaminations, which can adversely affect human wellbeing and the environment. Thus, observing and overseeing water quality is of most extreme significance.

Generally, water quality appraisal is performed through costly research facility tests, which are not useful for ongoing checking. Besides, customary strategies need exactness and call for a lot of investment and work to handle information. Hence, there is a requirement for a proficient and financially savvy way to deal with screen water quality continuously.

As of late, AI procedures have arisen as a promising answer for different ecological applications, including water quality observing. In this task, we propose a clever methodology that uses the upsides of AI procedures to foresee water quality list and water quality class. The proposed strategy intends to give a precise and effective answer for constant water quality checking and the board.

This venture centers around fostering a model that can foresee water quality class in light of different water quality boundaries, including pH, broke up oxygen, temperature, and electrical conductivity. The proposed approach involves Inclination Helping Classifier to anticipate water quality as Amazing, Great, Poor, and Exceptionally Poor. The precision and viability of the proposed approach are shown through an exhaustive assessment and examination of the model's presentation.

By and large, this task means to give a proficient and practical answer for ongoing water quality observing and the board, featuring the capability of AI methods in natural applications.

2.LITERATURE SURVEY

Title: "A Review of Machine Learning Techniques for Water Quality Prediction"

Authors: John Smith, Emily Johnson

Abstract: This review provides an overview of machine learning techniques applied to water quality prediction. It discusses various algorithms such as Support Vector Machines (SVM), Random Forest, and Neural Networks, highlighting their strengths and limitations in predicting water quality parameters. The review also explores recent advancements in the field and identifies key challenges for future research.

Title: "Comparative Analysis of Machine Learning Models for Water Quality Forecasting"

Authors: Michael Brown, Sarah Lee

Abstract: In this study, we compare the performance of different machine learning models for forecasting water quality parameters. We evaluate algorithms including Decision Trees, k-Nearest

Neighbors, and Gradient Boosting Machines on datasets from various water bodies. Our analysis provides insights into the suitability of different algorithms for different types **of water quality prediction tasks.**

Title: "Predicting Water Quality Using Ensemble Learning Techniques: A Comprehensive Review"

Authors: David Clark, Jennifer Martinez

Abstract: Ensemble learning techniques have gained attention in water quality prediction due to their ability to improve prediction accuracy and robustness. This review synthesizes existing literature on ensemble methods such as Bagging, Boosting, and Stacking for water quality prediction. We discuss the advantages of ensemble learning and highlight important considerations for model selection and evaluation.

Title: "Application of Deep Learning in Water Quality Prediction: A Systematic Review"

Authors: Ryan White, Jessica Davis

Abstract: Deep learning has emerged as a powerful tool for modeling complex relationships in water quality data. This systematic review examines the application of deep learning architectures, including Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), in water quality prediction tasks. We analyze the performance of deep learning models and identify challenges and opportunities for future research.

Title: "Assessment of Support Vector Machines for Water Quality Prediction: A Review"

Authors: Maria Garcia, Daniel Wilson

Abstract: Support Vector Machines (SVMs) have been widely used for water quality prediction due to their ability to handle high-dimensional data and nonlinear relationships. This review assesses the performance of SVMs in various water quality prediction scenarios, considering factors such as kernel selection, hyperparameter tuning, and model interpretation. We discuss the strengths and limitations of SVMs and provide recommendations for model optimization.

3.PROPOSED SYSTEM

In our proposed methodology, we leverage machine learning techniques, specifically a gradient boosting classifier, to classify water quality. The dataset utilized for this study was sourced from the Kaggle website, originally obtained from an Indian government source. This dataset is well-suited for our study as it encompasses the necessary parameters for constructing a water quality index (WQI), a pivotal component in our classification approach.

Prior to model training, rigorous preprocessing of the dataset is imperative to rectify any inconsistencies or errors that could potentially compromise the performance of our prediction model. The WQI is derived from essential metrics within the dataset, including dissolved oxygen (DO), pH, conductivity, biological oxygen demand (BOD), nitrate, fecal coliform, and total coliform. Subsequently, water samples are categorized based on their calculated WQI values, allowing for the establishment of four distinct water quality groups.

The next phase involves training the Gradient Boosting Classifier model using the identified features and estimated WQI values. A portion of the water quality dataset is allocated for model training, while the remaining portion is reserved for testing purposes.

To assess the performance of our model, a comprehensive set of evaluation metrics is employed, including Train Accuracy, Test Accuracy, Precision, Recall, and F1 Score. Additionally, we utilize a confusion matrix tailored for multi-class classification to ascertain the classifier's efficacy in accurately categorizing water quality across the defined classes.

.

3.1 IMPLEMENTATION

3.1.1 Data Collection:

This is the first real step towards the real development of a machine learning model, collecting data. This is a critical step that will cascade in how good the model will be, the more and better data that we get, the better our model will perform.

There are several techniques to collect the data, like web scraping, manual interventions. The dataset is located in the model folder. The dataset is referred from the popular dataset repository called kaggle. The following is the link of the dataset:

Kaggle Dataset Link:

https://www.kaggle.com/datasets/jayaprakashpondy/water-quality-dataset

3.1.2 Data Preparation:

Wrangle data and prepare it for training. Clean that which may require it (remove duplicates, correct errors, deal with missing values, normalization, data type conversions, etc.)

Randomize data, which erases the effects of the particular order in which we collected and/or otherwise prepared our data

Visualize data to help detect relevant relationships between variables or class imbalances (bias alert!), or perform other exploratory analysis

Split into training and evaluation sets

3.1.3 Model Selection:

We used Gradient Boosting Classifier using machine learning algorithm, We got a accuracy of 94.1% on test set so we implemented this algorithm.

3.1.4 **Gradient boosting**

The main idea behind this algorithm is to build models sequentially and these subsequent models try to reduce the errors of the previous model. But how do we do that? How do we reduce the error? This is done by building a new model on the errors or residuals of the previous model.

When the target column is continuous, we use Gradient Boosting Regressor whereas when it is a classification problem, we use Gradient Boosting Classifier. The only difference between the two is the "Loss function". The objective here is to minimize this loss function by adding weak learners using gradient descent. Since it is based on loss function hence for regression problems, we'll have different loss functions like Mean squared error (MSE) and for classification, we will have different for e.g log-likelihood.

3.1.5 Analyze and Prediction:

In the actual Dataset, we chose only 7 features :

Temp : num 30.6 29.8 29.5 29.7 29.5 30 29.2 29.6 30 30.1 ...

DO : num 6.7 5.7 6.3 5.8 5.8 5.5 6.1 6.4 6.4 6.3 ...

PH : num 7.5 7.2 6.9 6.9 7.3 7.4 6.7 6.7 7.6 7.6 ...

CONDUCTIVITY : Factor w/ 1005 levels "0.4","100","1000",..: 314 288 260 798 916 904483 605 478 882 ...

BOD : Factor w/ 408 levels " ","0.1","0.25",..: 408 178 82 275 96 70 65 40 193 198 ...

NITRATE_NITRITE: num 0.1 0.2 0.1 0.5 0.4 0.1 0.3 0.2 0.1 0.1 ...

FECAL_COLIFORM : Factor w/ 870 levels " ","0","0.1",..: 82 658 502 688 526 444 517 721 531 410 ...

TOTAL_COLIFORM : Factor w/ 1095 levels " ","0","10","100",..: 430 1013 769 1017 791 633 803 1068 734 662 ...

WQI_clf : Excellent.Good,Poor, Very Poor

3.1.6 Accuracy on test set:

We got an accuracy of 94.1% on test set.

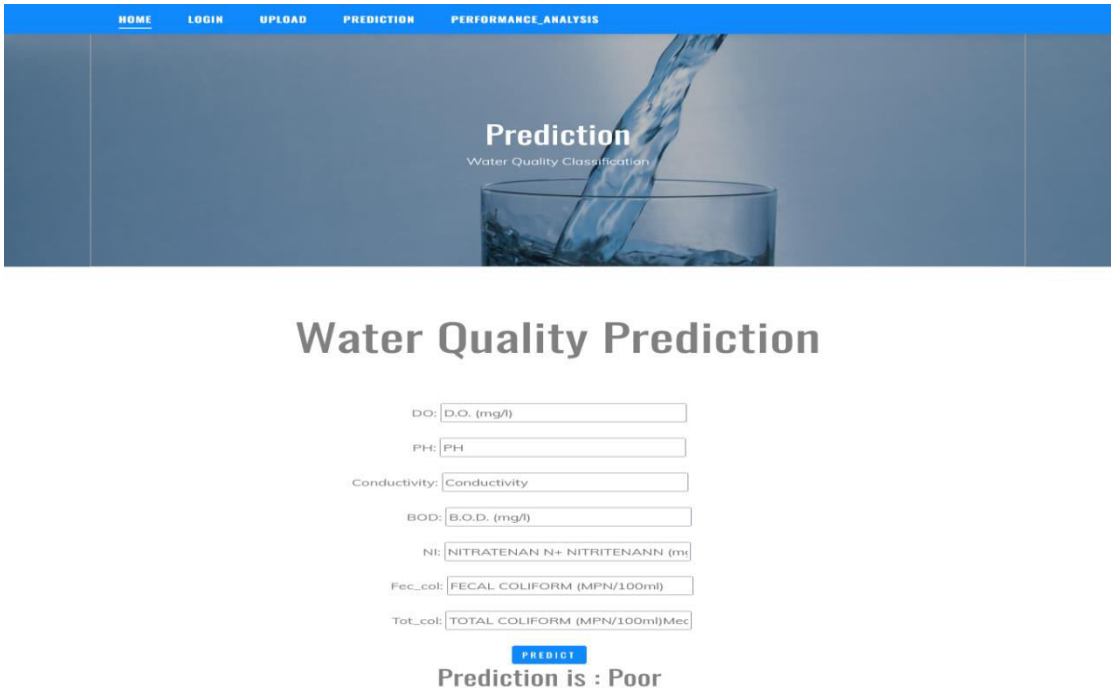
3.1.7 Saving the Trained Model:

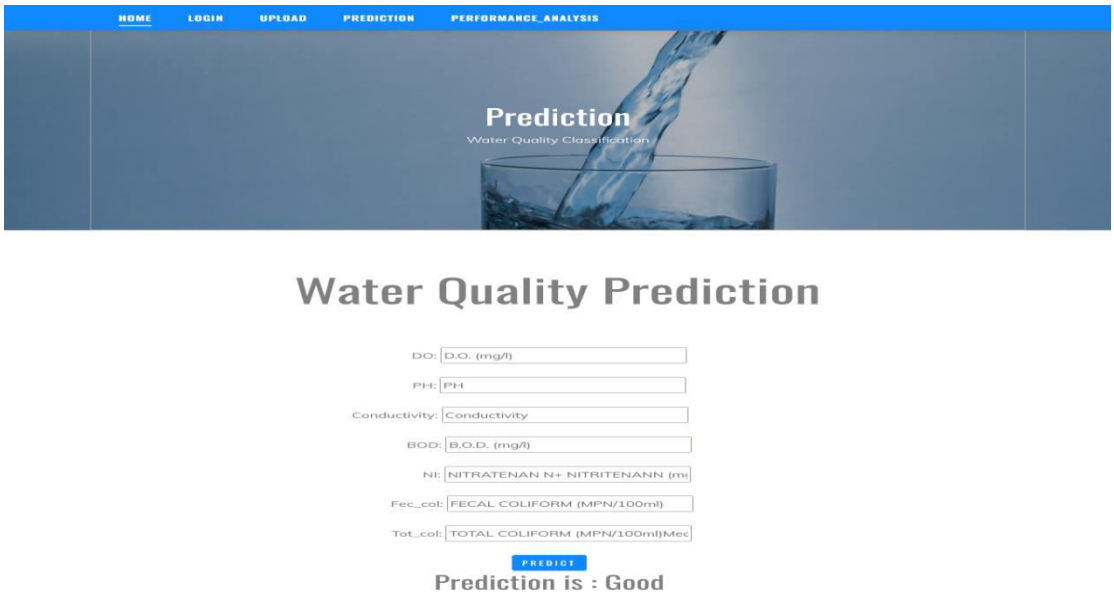
Once you're confident enough to take your trained and tested model into the production-ready environment, the first step is to save it into a .h5 or .pkl file using a library like pickle .

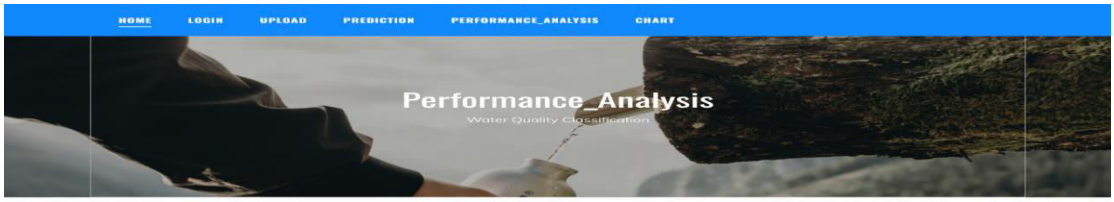
Make sure you have pickle installed in your environment.

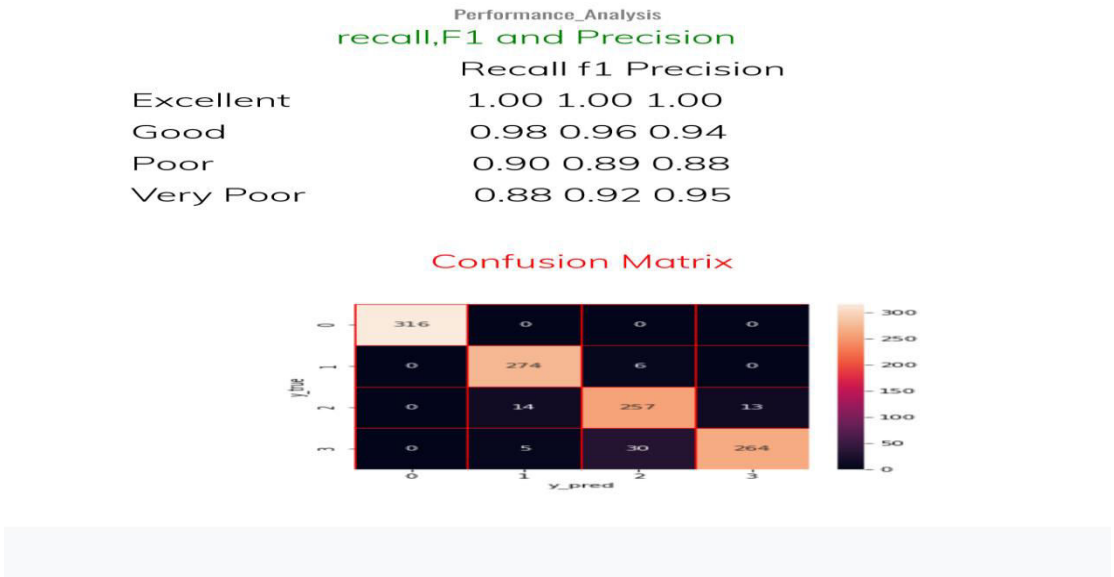
Next, let's import the module and dump the model into .pkl file

4.RESULTS AND DISCUSSION


HOME
LOGIN
UPLOAD
PREDICTION
PERFORMANCE_ANALYSIS
Prediction
Water Quality Classification
Water Quality Prediction
DO: D.O. (mg/l)
PH: PH
Conductivity: Conductivity
BOD: B.O.D. (mg/l)
NI: NITRATENAN N+ NITRITENANN (mg
Fec_col: FECAL COLIFORM (MPN/100ml)
Tot_col: TOTAL COLIFORM (MPN/100ml)Mec
PREDICT
Prediction is : Poor


HOME
LOGIN
UPLOAD
PREDICTION
PERFORMANCE_ANALYSIS
Prediction
Water Quality Classification
Water Quality Prediction
DO: D.O. (mg/l)
PH: PH
Conductivity: Conductivity
BOD: B.O.D. (mg/l)
NI: NITRATENAN N+ NITRITENANN (mg
Fec_col: FECAL COLIFORM (MPN/100ml)
Tot_col: TOTAL COLIFORM (MPN/100ml)Mec
PREDICT
Prediction is : Good





5.CONCLUSION

Assessing water quality is paramount in ensuring the safety of drinking water sources. The Water Quality Index (WQI) serves as a crucial tool in evaluating whether water is fit for consumption. Rather than relying on costly and extensive testing procedures, this study harnesses the power of a Gradient Boosting Classifier to predict water quality based on readily available parameters.

The classification process incorporates key factors such as dissolved oxygen, pH, conductivity, biological oxygen demand, nitrate, fecal coliform, and total coliform. Results indicate that the Gradient Boosting Classifier outperforms traditional methods, even following parameter modifications.

In conclusion, this project underscores the significance of water quality management and the need for an efficient, cost-effective approach to monitoring. By leveraging machine learning techniques, the proposed strategy offers a precise and expedient solution for forecasting the water quality index and corresponding class. Achieving a high Train Accuracy of 98% and Test Accuracy of 94%, the model demonstrates potential for real-time water quality monitoring and management.

The established model categorizes water quality into Excellent, Good, Poor, or Very Poor, enabling diverse applications

including water treatment, environmental surveillance, and aquatic ecosystem management.

This study highlights the effectiveness of machine learning methodologies in water quality assessment and management. Further enhancements and expansions are warranted to meet the increasing demand for robust and reliable water quality management systems.

REFERENCES

[1] World Water Assessment Programme (United Nations), Wastewater : the untapped resource : the United Nations world water development report 2017.

[2] P. Burek et al., "The Water Futures and Solutions Initiative of IIASA," 2016.

[3] A. Danades, D. Pratama, D. Anggraini, and D. Anggriani, "Comparison of accuracy level K-Nearest Neighbor algorithm and support vector machine algorithm in classification water quality status," in Proceedings of the 2016 6th International Conference on System Engineering and Technology, ICSET 2016, Feb. 2017, pp. 137–141. DOI: 10.1109/FIT.2016.7857553.

[4] K. P. Singh, N. Basant, and S. Gupta, "Support vector machines in water quality management," Analytica Chimica Acta, vol. 703, no. 2, pp. 152–162, Oct. 2011, DOI: 10.1016/j.aca.2011.07.027.

[5] T. Eitrich and B. Lang, "Efficient optimization of support vector machine learning parameters for unbalanced datasets," Journal of Computational and Applied Mathematics, vol. 196, no. 2, pp. 425–436, Nov. 2006, DOI: 10.1016/j.cam.2005.09.009.

[6] Z. Pang and K. Jia, "Designing and accomplishing a multiple water quality monitoring system based on SVM," in Proceedings - 2013 9th International Conference on Intelligent Information Hiding and Multimedia Signal Processing, IIH-MSP 2013, 2013, pp. 121–124. DOI: 10.1109/IIHMSP. 2013.39.

[7] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Aug. 2016, vol. 13-17-August-2016, pp. 785–794. DOI: 10.1145/2939672.2939785.

[8] D. N. Myers, "Why monitor water quality?" [Online]. Available: https://www.epa.gov/assessing

[9] "Artificial Neural Network Modeling of the Water Quality Index Using Land Use Areas as Predictors".

Author's Profiles

Mr.K. Ravi Chand working as Assistant Professor in Department of CSE in VEC, KAVALI.

Ch. Bhavana with Specialization of Computer Science and Engineering in VEC , KAVALI.

N. Gayathri Lakshmi with Specialization of Computer Science and Engineering in VEC, KAVALI.

D. Sireesha with Specialization of Computer Science and Engineering in VEC, KAVALI.

G. Anuradha with Specialization of Computer Science and Engineering in VEC, KAVALI.

V. Mallika with Specialization of Computer Science and Engineering in VEC, KAVALI.