PREDICTION OF WATER QUALITY USING ML ALGORITHMS

Mr. G. RAMAMOHANA RAO [1], Mr. VENKATESH MADINNI [2]

#1 Associate professor in the Department of IT at DVR & DR. HS MIC College of Technology (Autonomous), Kanchikacherla, NTR District.

#2 MCA student in the Department of Computer Applications (DCA) at DVR & DR. HS MIC COLLEGE OF TECHNOLOGY, Kanchikacherla, NTR District

ABSTRACT_ One of the most valuable natural resources ever given to humans is water. The ecosystem and human health are directly impacted by the water quality. Water is used for many different things, including drinking, farming, and industrial uses. Over the years, numerous pollutants have put water quality in danger. Predicting and estimating water quality are now crucial to reducing water pollution as a result. Real-time monitoring is unsuccessful because conventionally, water quality is assessed using expensive laboratory and statistical processes. Low water quality calls for a more workable and economical solution. The proposed system builds a model that can forecast the water quality index and water quality class by utilizing the advantages of machine learning techniques. This proposed system is to develop a novel approach for water quality classification using Gradient Boosting Classifier. The method includes the calculation of the Water Quality Index, which is used as a measure of water quality. The proposed approach achieves a high Train Accuracy of 98% and Test Accuracy of 94%. The approach uses various water quality parameters and features such as pH, dissolved oxygen, temperature, and electrical conductivity to classify water into different categories. The model developed in this study is capable of predicting the water quality as Excellent, Good, Poor and Very Poor, which can be used for real-time monitoring and management of water quality. The results demonstrate the effectiveness and accuracy of the proposed approach in predicting water quality, highlighting the potential of machine learning techniques for water quality monitoring and management. The proposed approach can be used in various applications such as water treatment, environmental monitoring, and aquatic life management.

1.INTRODUCTION

Water, one of Earth's most essential resources, is necessary for all forms of life. Its many purposes include providing a home for marine life, potable water, modern conveniences, and water systems. However, many pollutants change the way water originally was, putting people and the planet in peril. Due diligence in monitoring the water supply is, hence, essential. It is impractical to employ expensive tests performed at research institutes for continuous monitoring of water quality, despite their widespread usage for assessment. Also, while working with data, traditional approaches demand accuracy and demand a lot of investment and time. Consequently, controlling the quality of the screen water requires a practical and economical approach. Artificial intelligence (AI) has recently shown potential in a variety of ecological applications, including water quality monitoring. Using the advantages of AI approaches, our suggested solution to this issue may accurately forecast the water quality class and water quality list. The objective of the proposed method is to provide a precise and efficient means of continuously monitoring the board's water quality.

This project's overarching goal is to create a model that, given a number of water quality characteristics including pH, dissolved oxygen, temperature, and electrical conductivity, can forecast the water quality level. Using the Inclination Helping Classifier, the proposed method can forecast water quality as Amazing, Great, Poor, or Extremely Poor. The proposed approach is accurate and feasible, as shown by a comprehensive review and analysis of the model's presentation. By delivering a competent and practical solution for continuous water quality monitoring and the board, this project aspires to demonstrate the potential of AI techniques in natural applications.

2.LITERATURE SURVEY

1) Comparison of the accuracy of the K-Nearest Neighbor and support vector machine algorithms for classifying water quality conditions
A. Danades, D. Anggraini, A. Pratama, and A. ANGGRANI wrote the piece. In terms of water purity, there are four levels: very clean, somewhat filthy, quite contaminated, and very dirty. In order to utilize and maintain water properly, one must know how to categorize its quality. Due to the critical nature of quality status categorization, both the K-Closest Neighbor (KNN) and Support Vector Machine (SVM) layout must be used. The water quality status is categorized

according to the limitations. Here we compare and contrast KNN and SVM computations for water quality status characterisation, check at their correlation to find out which one is more accurate for the Water Quality Status Order, and then see how they do when evaluated using 10-overlay Cross Approval. Based on the test findings, SVM has the most normal worth of exactness due to its greater accuracy value of 92.40% at straight bit. The accuracy value of KNN is generally 71.28% when K=7.

2) The use of support vector machines to improve water quality regulation

AUTHORS: K. P. Singh, N. Basant, and S. Gupta

A combination of surface water quality data and support vector order (SVC) and relapse (SVR) models was used to improve the monitoring technique. The data set consisted of 1500 water samples taken from 10 distinct locations that were observed for an extended period of time. Review goals included developing a realistic SVR model for predicting water's biochemical oxygen interest (Body) using several characteristics and streamlining water quality comparisons via the rational grouping of testing locations (spatial) and months (global). During preparation, the spatial SVC model produced a misclassification rate of 12.39%; during approval, it increased to 17.70%; and during test sets, it reached 26.38%. Furthermore, the two models produced a total of twelve test locations and times, with three of those locations and times allocated to each model. The corresponding projected water body values for the preparation, approval, and test sets were 0.952, 0.909, and 0.907, respectively, according to the SVR model. Root mean squared errors of 1.53, 1.44, and 1.32 were minimal, thanks to the planned qualities. In addition to their remarkable prediction capabilities, the suggested display tracks the limits of the built models' adequacy. While the SVR model achieved a data reduction of 92.5% to enhance the future checking program, the SVC model assisted with water body prediction using a limited number of observable parameters. Nonlinear models (SVM, KDA, KPLS) outperformed related straight techniques (DA, PLS) in order and relapse detection, with findings that were comparable.

3) In order to optimize support vector machine learning parameters efficiently for datasets that are not balanced,

AUTHORS: T. Eitrich and B. Lang

Support vector machines excel in assembling and retrying tasks. They

provide beautiful isolating hyperplanes when built correctly. It is challenging to change the supplied preparation data and extra learning boundaries, which determine the nature of the preparation; this is particularly true for asymmetric datasets. Matrix search techniques have been used to find appropriate border features in most circumstances. Our automated method for handling changes to the study's learning boundaries is based on a subsidiary free mathematical enhancer. In order to make the improvement interaction more efficient, one more subtle quality measure is added. We demonstrate how our approaches may train support vector machines to excel at grouping jobs by performing mathematical tests on a prominent dataset.

3.PROPOSED SYSTEM

The proposed technique classifies water quality using a gradient boosting classifier. Using the Kaggle platform, the dataset used in this study was obtained from an Indian government website. This data set is helpful for our current investigation since it includes the attributes required to construct the water quality index. Water quality may be categorized using the water quality index.

Before a dataset can be used to train a prediction model, it must undergo data pre-processing, which involves finding and resolving any flaws. The most significant metrics used to produce the water quality index (WQI) from the dataset were dissolved oxygen (DO), pH, total coliform bacteria, fecal and total coliform bacteria, and nitrogen. The water samples were categorized based on the values of the Water Quality Index (WQI). As a measure of water quality, this study calculates the WQI using the weighted arithmetic technique. There are four separate categories used to categorize the water quality.

Training a Gradient Boosting Classifier model follows feature selection and WQI estimation. We use some of the water quality data for training and some for testing the model.

Several measures are used to measure the model's correctness, including train correctness, test accuracy, precision, recall, and F1 Score. A confusion matrix is used to evaluate algorithms that classify water quality. With four different kinds of water included in the study, the researchers used a confusion matrix for multi-class classification to show how the data sets were really structured.

3.1 IMPLEMENTATION

3.1.1 Data Collection:

To begin developing a machine learning model, data collection must take place. Since more and better data equals a better model, this is an important step that determines the quality of the model. There are a variety of ways to collect this data, including web scraping and human interventions. To find the dataset, open the model folder. The dataset was sourced from the popular dataset repository called kaggle. You may obtain the dataset by following this link:

Kaggle Dataset Link:

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

3.1.2 Data Preparation:

Accumulate data and prepare it for training. Data normalization, error correction, handling of missing values, data type conversion, etc.
Randomize the data so it doesn't matter what order our data was gathered and processed.
There are many applications for data visualization, including exploratory research, identifying class imbalances (caution: bias!), and finding relevant relationships between variables.

Divide the data into individual sets for testing and training purposes.

3.1.3 Model Selection:

After achieving a 94.1% success rate on the test set, we decided to use the Gradient Boosting Classifier method, which is based on machine learning.

3.1.4 **Gradient boosting**

This strategy is based on the core idea of building models sequentially with the purpose of lowering the mistakes of the prior model. But what are we going to do if we fail? Can the error be reduced in any way? In order to build a new model, the residuals or errors of the previous model are used.
In cases when the target column is continuous, we use the Gradient Boosting Regressor; nonetheless, 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:

We limited ourselves to seven characteristics in the real dataset: Temperature: 30.6, 29.8, 29.5, 29.7, 29.5, 30.92, 29.6, 30.1,... Dioxide: 6.7, 5.7, 5.8, 5.5, 6.1, 6.4, 6.4, 6.3,...

Total: 7.5 7.2 6.9 6.9 7.3 7.4 6.7 6.7 7.6 7.6...

3.1428,260,798,916, 904483, 605 478,882; factor with 1005 levels "0.4","100","1000",..... ...

Factor with 408 values of " ", "0.1","0.25",...: 408 178 82 275 96 70 65 40 193 198...

number of NITRATE molecules: 0.1, 0.2, 0.1, 0.5, 0.4, 0.1, 0.3, 0.2, 0.1,.....

Factor including 870 levels of " ", "0","0.1",..: 82,658,502, 688,526 444,517,721,531,440 ...

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

:WQI_clf is great.Decent, Bad, Extremely Bad

4.RESULTS AND DISCUSSION

3.1.6 Accuracy on test set:

On the test set, we achieved a precision of 94.1%.

3.1.7 Saving the Trained Model:

In order to move a trained and tested model to a production-ready environment, a library like pickle is needed to save the model as a.h5 or.pkl file.

Be sure to install pickle on your environment.

Afterwards, you'll need to import the module and save the model as an APK file.

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

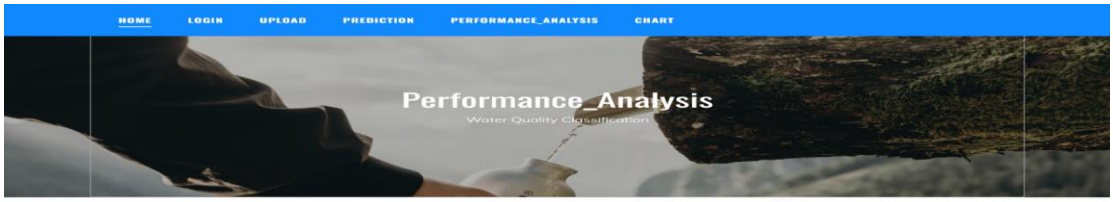
Prediction is : Poor

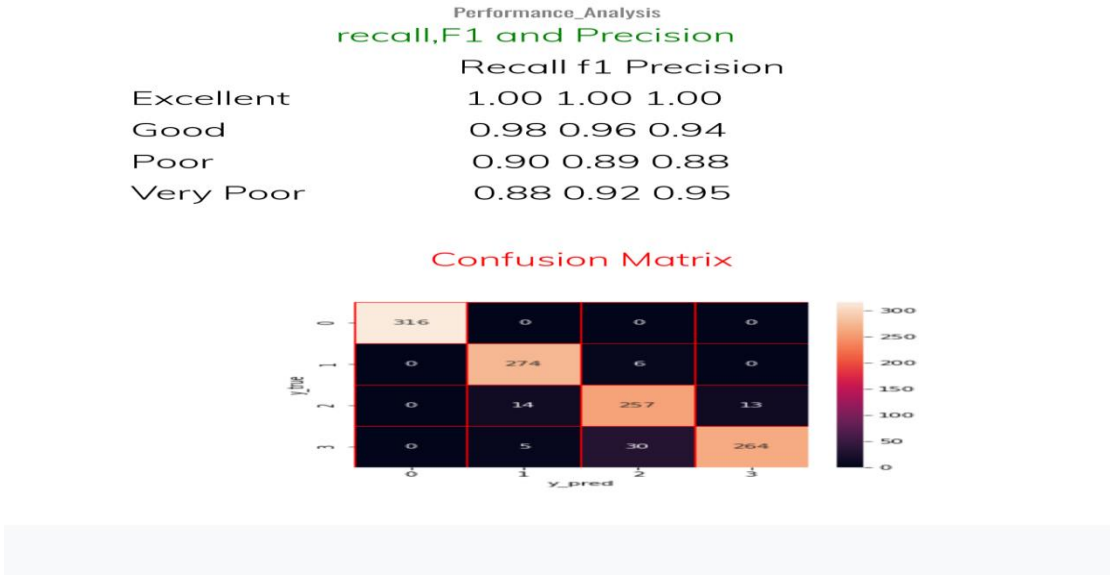
Water Quality Prediction

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

When deciding whether water is safe to drink, quality is a major consideration. Water quality index testing is necessary to determine the water's potability. Avoiding time-consuming and expensive research, this study forecasts water quality using publicly available data using a Gradient Boosting Classifier. The classification process makes use of the following variables: dissolved oxygen, pH, conductivity, biological oxygen demand, nitrate, fecal coliform, and total coliform. Despite tweaking the parameters, the results still demonstrated that the Gradient Boosting Classifier was superior to the prior method. According to the results of this study, there has to be an affordable and practical way to keep an eye on water quality. The suggested approach effectively and efficiently forecasts the water quality index and class using machine learning techniques. There is hope for real-time water quality management and monitoring with this method's 98% Train Accuracy and 94% Test Accuracy. The model established in this work has several potential applications, including water treatment, environmental monitoring, and aquatic life management. It has the ability to forecast water quality as Excellent, Good, Poor, or Very Poor. Taken together, the findings indicate that machine learning approaches might be valuable for water quality monitoring and management. They also provide

suggestions for how this technology could be enhanced and expanded to address the growing need for dependable and efficient 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," AnalyticaChimicaActa, 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".

[10] M. Bouamar and M. Ladjal, "Evaluation of the performances of ANN and SVM techniques used in water quality classification."

Author's Profiles

Mr. G. RAMAMOHANA RAO completed her M.TECH (CBIT) from Osmania University. He has published more than 10 papers in indexing journals. Currently working as an Associate professor in the department of IT at DVR & DR. HS MIC College of Technology (Autonomous), Kanchikacherla, NTR (DT). His areas of interest are Java and Python.

Mr. VENKATESH MADINNI, as MCA student in the department of DCA at DVR & DR. HS MIC COLLEGE OF TECHNOLOGY, Kanchikacherla, NTR (DT). He has completed B.Sc (MPC) in Chaitanya junior college, Nandigama From KRISHNA UNIVERSITY. His areas of interests are C ,java and Web development.