ISSN: 2278-4632
Vol-14, Issue-8, August: 2024

WATER QUALITY CLASSIFICATION USING SVM AND XGBOOST METHOD USING MACHINE LEARNING

Banoth Naveen
Email: naveenraina006@gmail.com
M. Tech, Department of Computer Science and Engineering.
Annamacharya Institute of Technology and Science, Hyderabad, Telangana, India.

Ramesh Babu Varugu
Email: rameshvarugu82@gmail.com
Assistant Professor
& HOD, Department of CSE.
Annamacharya Institute of Technology and Science, Hyderabad, Telangana, India.

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 Accuracy of 98%. 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.

Keywords – SVM, XGBoost, water quality, machine learning, classification.

I. INTRODUCTION

The prediction of water quality through the application of machine learning techniques has recently emerged as a potentially fruitful and cutting-edge strategy for addressing the issues connected with the effective monitoring and management of water resources. Because of the ever-increasing need for pure and risk-free water, precise and effective forecasting of the criteria that determine water quality has become an absolute necessity. Machine learning is a subset of artificial intelligence that offers powerful tools and techniques to analyses enormous volumes of data and make exact predictions. Because of this, it is a perfect option for improving the process of

Copyright @ 2024 Author

ISSN: 2278-4632
Vol-14, Issue-8, August: 2024

water quality evaluation and management.

Traditional techniques of predicting water quality frequently rely on intricate mathematical models and arduous data processing, both of which can be time-consuming and resource-intensive. Traditional water quality predictions sometimes ignore crucial variables. Machine learning approaches are more efficient and automated. Machine learning algorithms need historical water quality data to find patterns and connections. PH, dissolved oxygen, turbidity, nutritional levels, and pollutants must be included. Because of this, they can accurately predict future water quality conditions

This paper examines the enormous potential of machine learning for water quality forecasting. This cutting-edge technology helps us comprehend water quality dynamics in reservoirs, rivers, lakes, and municipal water sources. More precise water quality predictions allow proactive water resource management and contaminant mitigation to protect human health and the environment. This safeguards human and environmental health. Machine learning's ability to analyses massive amounts of data is one of its biggest benefits in water quality forecasting. Past methods have struggled to interpret and understand such big datasets, which can influence water quality. However, machine learning algorithms excel at finding meaningful patterns from large amounts of data, which helps us understand water quality trends and their causes.

This essay will examine water quality prediction machine learning algorithms. These include decision trees, random forests, support vector machines, neural networks, and ensemble techniques. By knowing the pros and downsides of each method, one may choose the best one for the situation and dataset. Machine learning offers potential but also challenges. Data quality and quantity are essential for powerful and accurate models. Incomplete or erroneous data may make accurate forecasts difficult and cast doubt on the inquiry. Thus, we shall emphasize data pretreatment and quality assurance in model construction.

II. EXISTING SYSTEM

In the existing system the WQC is classified based on the water quality index (WQI) from 7 parameters in a dataset using Support Vector Machine (SVM) and Extreme Gradient Boosting (XGBoost). The research outcome demonstrated that the XGBoost model performed better, with an accuracy of 94%, compared to the SVM model, with only a 67% accuracy.

In the existing system SVM and XGBoost have been applied to water quality classification; however, their execution is non-trivial. Issues such as the constraint of limited water quality parameters and the algorithm's robustness when dealing with noises are still actively investigated.

SVMs use hyperplanes to define decision boundaries between data points of different classes, originally developed for binary classification problems. The hyperplanes are the decision functions that distinguish between positive and negative data and have marked the maximum margins. Because of its low sensitivity to feature space dimensions, SVM is considered a reliable classifier for the Hughes

Copyright @ 2024 Author

ISSN: 2278-4632
Vol-14, Issue-8, August: 2024

effect; classification using SVM has minimal impact on the outcome.

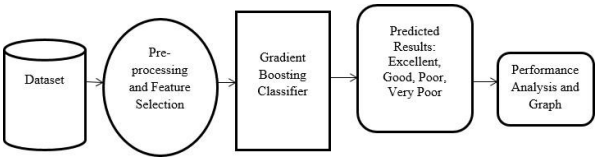
III. PROPOSED SYSTEM

In the proposed system we develop the Water Quality Classification Using Machine Learning with Gradient Boosting Classifier. The dataset used for this research was obtained from the Kaggle website, sourced from an Indian Government Website. The data is appropriate for the current research project since the characteristics required to construct the water quality index are available in this dataset. A water quality classification can be derived from the water quality index.

Before using the data for training, data need to go through data pre-processing, which refers to identifying and correcting errors in the dataset that may negatively impact a predictive model. The water quality index (WQI) was calculated using the dataset's most important parameters, which are Dissolved oxygen (DO), pH, conductivity, biological oxygen demand (BOD), nitrate, fecal coliform, and total coliform. The WQI scores were then used to categorize the water samples. This research uses the weighted arithmetic water quality index method to calculate the WQI. The water quality has been classified into four groups.

The next step is to train the Gradient Boosting Classifier model using the selected features and the calculated WQI. The model is trained using a portion of the water quality data, and the remaining data is used for testing.

IV. SYSTEM ARCHITECTURE:



V. IMPLEMANTATION

MODULES:

- Data Collection
- Dataset
- Data Preparation
- Water Quality Index Calculation
- Model Selection
- Analyze and Prediction
- Accuracy on test set
- Saving the Trained Model

MODULES DESCSRIPTION:

Dataset:

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.

Data Preparation:

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

Copyright @ 2024 Author

ISSN: 2278-4632
Vol-14, Issue-8, August: 2024

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

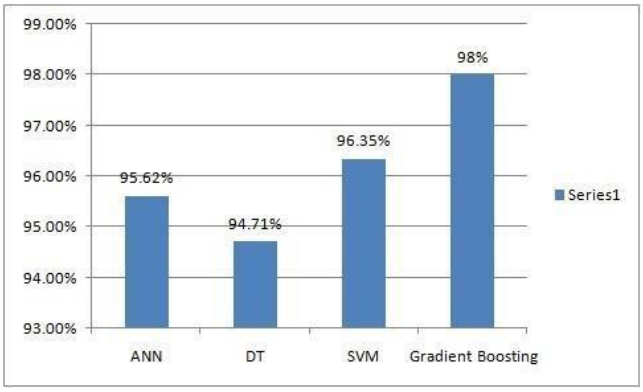
INPUT DESIGN

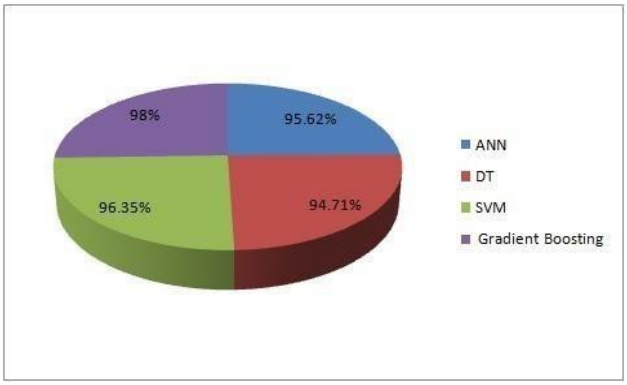
The input design is the link between the information system and the user. It comprises the developing specification and procedures for data preparation and those steps are necessary to put transaction data in to a usable form for processing can be achieved by inspecting the computer to read data from a written or printed document or it can occur by having people keying the data directly into the system. The design of input focuses on controlling the amount of input required, controlling the errors, avoiding delay, avoiding extra steps and keeping the process simple. The input is designed in such a way so that it provides security and ease of use with retaining the privacy. Input Design considered the following things:

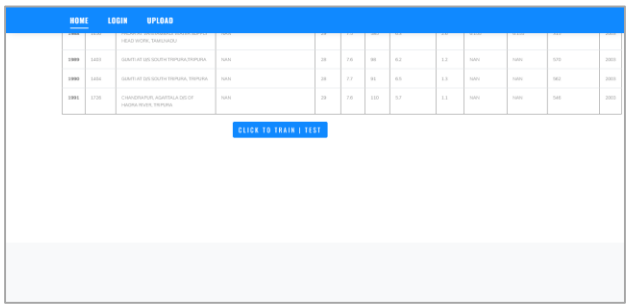
OUTPUT DESIGN

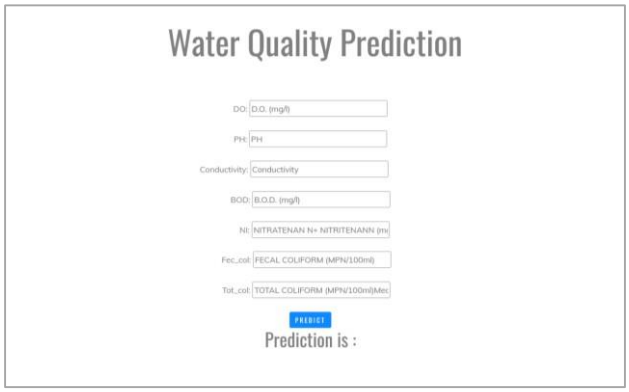
A quality output is one, which meets the requirements of the end user and presents the information clearly. In any system results of processing are communicated to the users and to other system through outputs. In output design it is determined how the information is to be displaced for immediate need and also the hard copy output. It is the most important and direct source information to the user. Efficient and intelligent output design improves the system's relationship to help user decision-making.

VI. RESULT



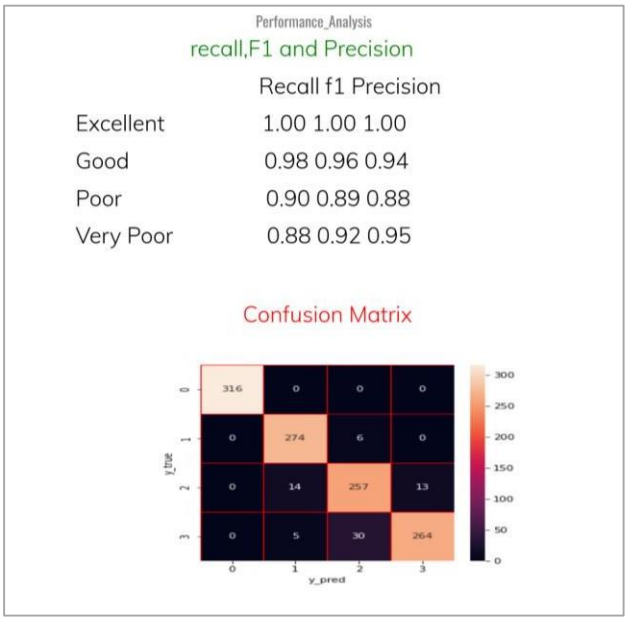






Copyright @ 2024 Author

ISSN: 2278-4632
Vol-14, Issue-8, August: 2024



VII. CONCLUSION

Water quality is important in determining whether the water source is qualified for consumption. WQI is essential to classify whether the water is safe for consumption. Rather than requiring expensive and complex analysis to test the water quality, this research uses Gradient Boosting Classifier, to predict water quality using readily available water quality parameters. The parameters employed for the classification algorithm are dissolved oxygen, pH, conductivity, biological oxygen demand, nitrate, fecal coliform, and total coliform. The outcome showed that Gradient Boosting Classifier outperformed the existing system even after the parameters had been tuned. In conclusion, this project highlights the significance of water quality and the need for an efficient and economical solution to monitor and manage it.

The proposed approach, utilizing the advantages of machine learning techniques, provides an accurate and effective solution for predicting the water quality index and water quality class. The approach achieves a high Accuracy of 98%, indicating its potential for real-time monitoring and management of water quality. The model developed in this study can predict water quality as Excellent, Good, Poor, and Very Poor, enabling various applications such as water treatment, environmental monitoring, and aquatic life management. Overall, this project demonstrates the potential of machine learning techniques in the field of water quality monitoring and management, and it can be further improved and expanded to meet the increasing demand for efficient and reliable water quality management systems.

VIII. FUTURE SCOPE

There are several potential avenues for future work on this project. First, the proposed approach could be extended to include more water quality parameters and features to improve the accuracy and robustness of the model. Additionally, the model could be refined to incorporate more advanced machine learning algorithms to achieve even higher accuracy and to account for complex interactions between different water quality parameters.

Another potential area for future work is to integrate the proposed approach with real-time

Copyright @ 2024 Author

ISSN: 2278-4632
Vol-14, Issue-8, August: 2024

sensor data to develop a fully automated and continuous water quality monitoring system. This would involve deploying sensors at various locations throughout a water system, collecting data, and feeding it into the model to generate accurate predictions of water quality in real-time.

Moreover, it would be beneficial to explore the applicability of this approach to different types of water systems, such as lakes, rivers, and coastal waters, and to investigate the potential impact of climate change and human activities on water quality. Additionally, further research could be conducted to explore the economic feasibility and practical implementation of the proposed approach in various settings and applications.

Overall, there is a vast scope for future work on this project, and it has the potential to make a significant contribution to the field of water quality monitoring and management.

IX. REFERENCES

1) Illa Iza Suhana Shamsuddin, Zalinda Othman * and Nor Samsiah Sani * Water Quality Index Classification Based on Machine Learning: A Case from the Langat River Basin Model. Water 2022, 14, 1552.
2) Ahmed, M.F.; Lim, C.K.; Mokhtar, M.B.; Khirotdin, R.P.K. Predicting Arsenic (As) Exposure on Human Health for Better Management of Drinking Water Sources. Int. J. Environ. Res. Public Health. **2021**, 18, 7997.
3) Kulisz, M.; Kujawska, J.; Przysucha, B.; Cel, W. Forecasting water quality index in groundwater using artificial neural network. Energies **2021**, 14, 5875.
4) Seyed Babak Haji Seyed Asadollah1 , Ahmad Sharafati2,3,4*, Davide Motta 5 , and Zaher Mundher Yaseen6. River Water Quality Index prediction and uncertainty analysis: a comparative study of machine learning models. Journal of Environmental Chemical Engineering. **2020**.
5) Abba, S.I.; Hadi, S.J.; Sammen, S.S.; Salih, S.Q.; Abdulkadir, R.A.; Pham, Q.B.; Yaseen, Z.M. Evolutionary computational intelligence algorithm coupled with self-tuning predictive model for water quality index determination. J. Hydrol. **2020**, 587, 124974.
6) Ho, J.Y.; Afan, H.A.; El-Shafie, A.H.; Koting, S.; Mohd, N.S.; Jaafar, W.Z.; Sai, H.L.; Abdul Malek, M.; Ahmed, A.N.; Wan Mohtar, W.H.M.; et al. Towards a time and cost effective approach to water quality index class prediction. J. Hydrol. **2019**, 575, 148–165.
7) Ahmed, U.; Mumtaz, R.; Anwar, H.; Shah, A.A.; Irfan, R.; García-Nieto, J. Efficient water quality prediction using supervised machine learning. Water **2019**, 11, 2210.
8) Chou, J.S.; Ho, C.C.; Hoang, H.S. Determining quality of water in reservoir using machine learning. Ecol. Inform. **2018**, 44, 57–75.
9) Haghiabi, A.H.; Nasrolahi, A.H.; Parsaie, A. Water quality prediction using machine learning methods. Water Qual. Res. J. **2018**, 53, 3–12.
10) Li, X.; Cheng, Z.; Yu, Q.; Bai, Y.; Li, C. Water-quality prediction using multimodal support vector regression: Case study of Jialing River, China. J. Environ. Eng. **2017**, 143, 04017070.

Copyright @ 2024 Author