e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:04/Issue:09/September-2022 Impact Factor- 6.752 www.irjmets.com

WATER QUALITY INDEX PREDICTION

G. Richa Shalom*1

*1Student, Department Of Computer Science And Engineering, Mahatma Gandhi Institute Of Technology, Hyderabad, Telangana, India.

DOI : https://www.doi.org/10.56726/IRJMETS292664

ABSTRACT

One of the most valuable resources is water. About 70% of the earth's surface is covered by water, which is one of the most critical resources for maintaining life. Water must be suitable before usage because it is used for a variety of uses. Rapid industrialization and urbanisation have caused an alarming rate of water quality degradation, which has resulted in terrible diseases. Water sources need to be routinely checked to see if they are still safe to use. Since water quality has traditionally been determined by costly and time-consuming statistical and laboratory investigations, real-time monitoring systems are now required. This research investigates a number of supervised machine learning techniques to calculate the water quality index using this as its inspiration. With the use of Machine Learning Model, there will be no limitation of the complexity increasing the number of variables.

Keywords: Statistical Analysis, Monitoring System, Accuracy, Water Quality Index.

I. INTRODUCTION

After air, water is arguably the most valuable natural resource. Water makes up the majority of the Earth's surface, yet because so little of it is useful, it is a scarce resource. Therefore, it is important to use this precious and finite resource carefully. Water must be suitable before usage because it is used for a variety of uses. Additionally, water sources must be routinely checked to see if they are sound or not. Water bodies in poor condition pose a threat to the ecology as well as being a sign of environmental degradation. In industries, poor water quality can result in risks and significant financial loss. Thus, the quality of water is very important in both environmental and economic aspects. Thus, water quality analysis is essential for using it for any purpose. After years of research, water quality analysis now consists of some standard protocols. There are rules for sample collection, storage, and analysis. Here, the typical chain of events is briefly outlined for the benefit of researchers and analysts. Analysis of water quality is necessary primarily for monitoring purposes. Among the significance of such an assessment are:

1. To determine whether the water quality complies with the criteria and is, thus, appropriate for the intended usage.

2. To check a system's effectiveness while maintaining water quality

3. To determine what modifications should be made and to determine whether an existing system needs to be upgraded or changed.

4. To keep an eye on whether the water quality complies with laws and regulations.

Traditional methods for estimating water quality involve costly and time-consuming statistical and laboratory tests, making the idea of real-time monitoring irrelevant today. There must be a quicker, more practical solution because of the dire effects of bad water quality. To estimate the water quality index (WQI), a single measure that describes the overall quality of water, this research investigates a number of supervised machine learning algorithms.

With the use of Machine Learning Model, there will be no limitation of the complexity increasing number of variables. This Models ,trains and test the given factors which Predicts water quality index and with the best performing machine learning model it can effortlessly predict the WQI of water with much higher accuracy than traditional methods. In our project, we have used the Indian Water quality dataset. The data that is used in this project originally comes from the kaggle machine learning dataset. We got to know all the required parameters to predict WQI. We have used Random Forest Classifier and linear regression to make predictions and compared their performance. Random Forest has highest accuracy and is a good choice for this problem. Random Forest trains the model with subsets of data sampled from the training data; this will make our model

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:04/Issue:09/September-2022 Impact Factor- 6.752 www.irjmets.com

more accurate. Python, Data Pre-processing Techniques, Machine Learning, Regression Algorithms are used. The benefits of this model are: No human interference is required, Easy interface, accurate calculations, Faster Results.

II. METHODOLOGY

The objective is to predict the Water Quality index. For achieving this, first we get the water data from the dataset and we monitor regarding the number of rows, columns of the given data followed by the other steps. Our approach comprises a sequence of steps for the prediction of the water quality index by collecting the required data, analyze the relationship between the different columns of the data and applying pre-processing techniques , analyzing the data and finally predicting the water quality index of the water sample given certain values like the pH, the dissolved oxygen, the bio-chemical oxygen demand, total coliform, electric conductivity and the nitrates. The general training work-flow can be seen below.

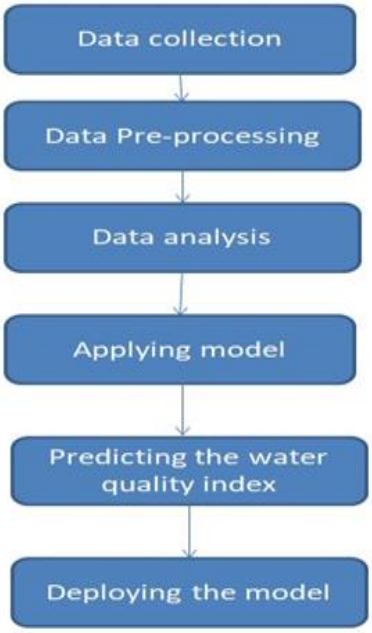

Figure 1: General Work-flow.

The data set used here is water_dataX.csv which consists of around 1992 rows and 12 columns. The dataset is taken from Kaggle. The different columns it contains include the station code, location, state, and the different input parameters like the temperature, the ph, the dissolved oxygen, the conductivity, the bio chemical oxygen demand, fecal coliform, total coliform ,the nitrates and the year.

Calculation Name	Component Range	Calculated Value	Calculation Name	Component Range	Calculated Value
Calculation of npH	PH Range	npH Value	Calculation of ndo	DO Range	ndo Value
	7 to 8.5	100		6+	100
	8.5 to 8.6 or 6.8 to 6.9	80		5.1 to 6	80
	8.6 to 8.8 or 6.7 to 6.8	60		4.1 to 5	60
	8.8 to 9 or 6.5 to 6.7	40		3 to 4	40
	Else	0		Else	0
Calculation of nco	TC Range	Nco Value	Calculation of nbdo	BOD Range	Nbdo Value
	0 to 5	100		0 to 3	100
	5 to 50	80		3 to 6	80
	50 to 500	60		6 to 80	60
	500 to 1000	40		80 to 125	40
	Else	0		Else	0
Calculation of nec	CO Range	Nec Value	Calculation of nna	NA Range	Nna Value
	0 to 75	100		0 to 20	100
	75 to 150	80		20 to 50	80
	150 to 225	60		50 to 100	60
	225 to 300	40		100 to 200	40
	Else	0		Else	0

Figure 2: Calculation of the water parameters

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:04/Issue:09/September-2022 Impact Factor- 6.752 www.irjmets.com

Since the parameter values are stored as fractions, Float64 has been chosen. A variable called WQI, which is built using data from npH, NBDO, NEC, NNA, WPH, WDO, WBDO, WEC, WNA, and WCO, determines the quality of the water.

III. MODELING AND ANALYSIS

Linear Regression:

A quiet and straightforward statistical regression technique called linear regression is utilized to highlight the link between continuous variables in predictive analysis. The term "linear regression" refers to a statistical method that displays a linear relationship between the independent variable (X-axis) and the dependent variable (Y-axis). Such linear regression is known as simple linear regression if there is only one input variable (x). Additionally, this type of linear regression is known as multiple linear regression if there are many input variables. The link between the variables is depicted by a sloping straight line in the linear regression model.

Random Forest:

The precision of the Random Forest Regressor was considered important. The Brieman-invented Random Forest classifier/regressor is a supervised machine learning technique that uses a number of decision trees as its fundamental classifiers. In addition to providing excellent estimation of missing values in a dataset when a significant number of values are missing, Random Forest Classifier also exhibits good accuracy with huge datasets. Sub sampling the training data and selecting node sets introduces randomness and replacement. The algorithm to be followed has the following steps. Select N samples from the dataset through row sampling and feature sampling. These samples are training datasets to draw N decision trees from the samples, each capable of producing a prediction. Aggregate the results produced by each decision tree. Determine the final prediction by considering the majority of results produced by N TREES.

The Random Forest Regressor can be imported from the SCI-KIT learn package in python. WE have chosen 10 as our n-estimators value (number of decision trees). The training data was given and the model predicts the Water Quality Class.

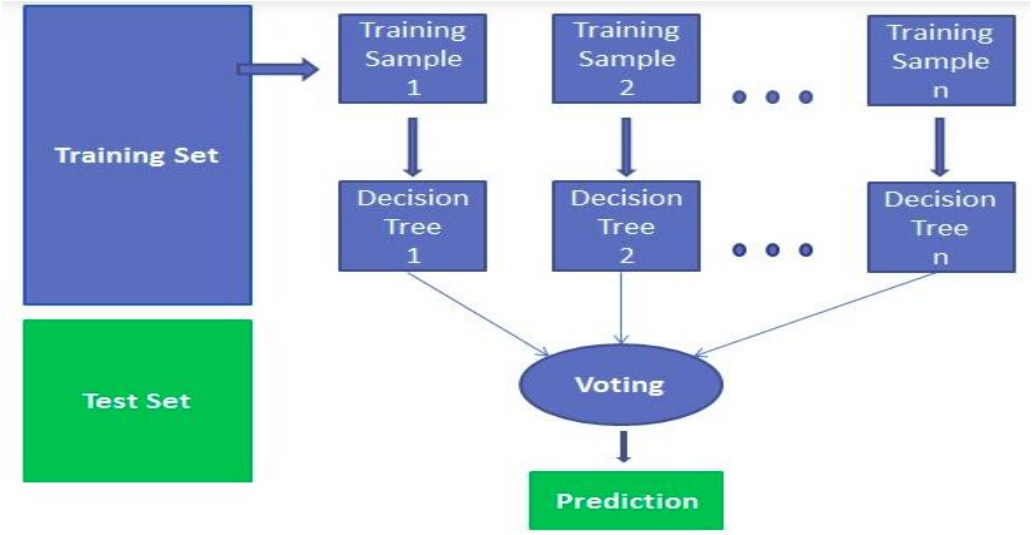

Figure 3: Working of Random Forest

Application Building

The website has a form that users may fill out to input their water parameters and send them to the model. The values are included into the machine learning model, which produces a prediction that is shown on the screen. The flask web framework is utilised to deploy the website.

Build a HTML page

The HTML page that we created includes a form element with a number of input tags to receive the user's input parameters. The form element consists of two attributes. The first is the action attribute, which specifies how data is submitted when the submit button is pressed. The method attribute describes the kind of HTTP request our form sends out. In this instance, the action URL would be "/login," and the method would be "post," which

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:04/Issue:09/September-2022 Impact Factor- 6.752 www.irjmets.com

offers more secure data transmission than other ways. Our HTML is styled with CSS to provide an aesthetically pleasing experience.

Flask

Python-based web framework Flask is simple to use and has an integrated development server that enables rapid application deployment. The static folder, which contains CSS and graphics, and the templates folder, which contains HTML pages, are required components of the Flask structure. Using the pickle.load() function, the model is loaded into the website and stored in the model variable. Our key coding components are @app.route, render template(), and request. form (). The HTML page we created is returned using the render template() function. The returned template is transformed into an HTTP request for the browser to display using the @app.route decorator. To view the page, the flask app is launched using its built-in development server.

IV. RESULTS AND DISCUSSION

Python is the programming language used to create the "Water Quality Index Prediction" function. It displays the environment in which the functioning code will be implemented, as seen below. The path to the directory is given after opening the Anaconda prompt. The name of the file containing the Python code that needs to be implemented ends in.py. When we click the link it provides after execution, our results are shown.

After running the python code and opening the link which was given as a result, we get the result page. This result takes inputs of values and outputs the predicted Water Quality Index. The below are the pictures of the output.

Figure 4: The accuracy of Random Forest Model

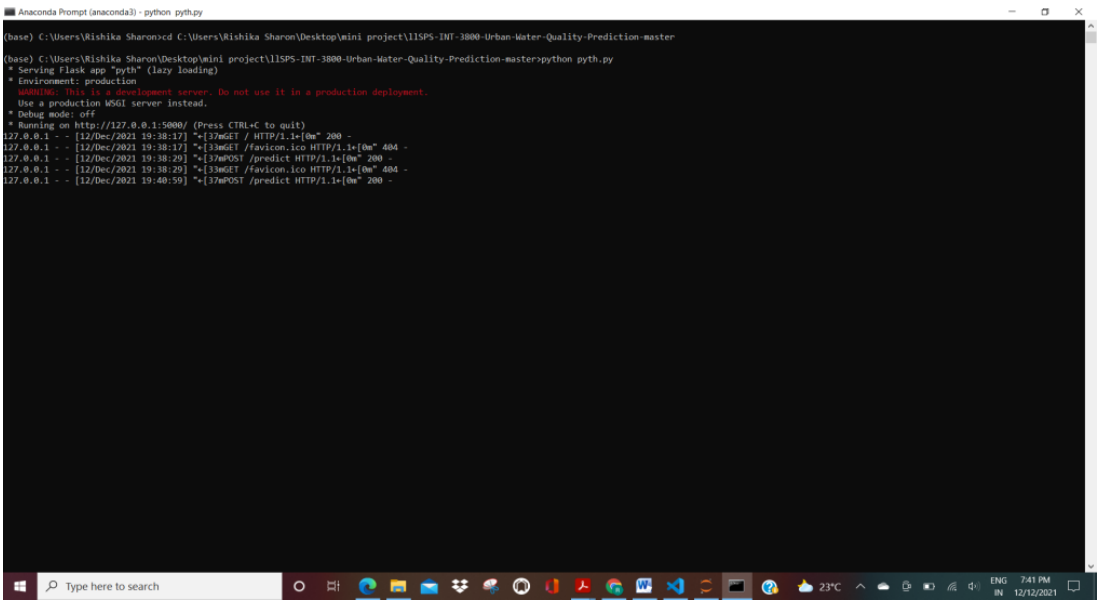

Figure 5: Anaconda prompt

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:04/Issue:09/September-2022 Impact Factor- 6.752 www.irjmets.com

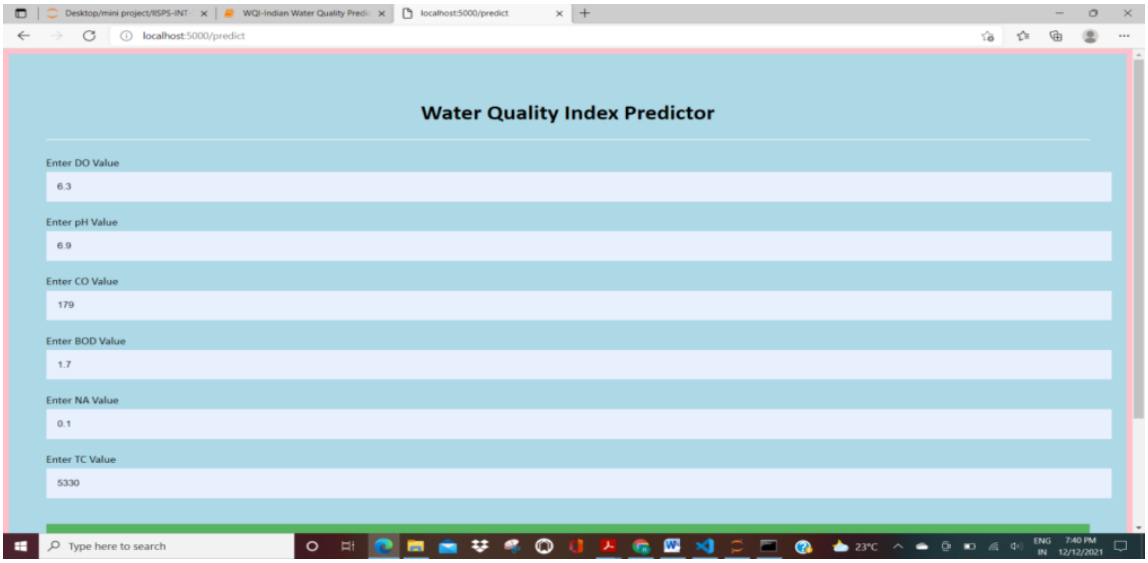

Figure 6: Result Output-1

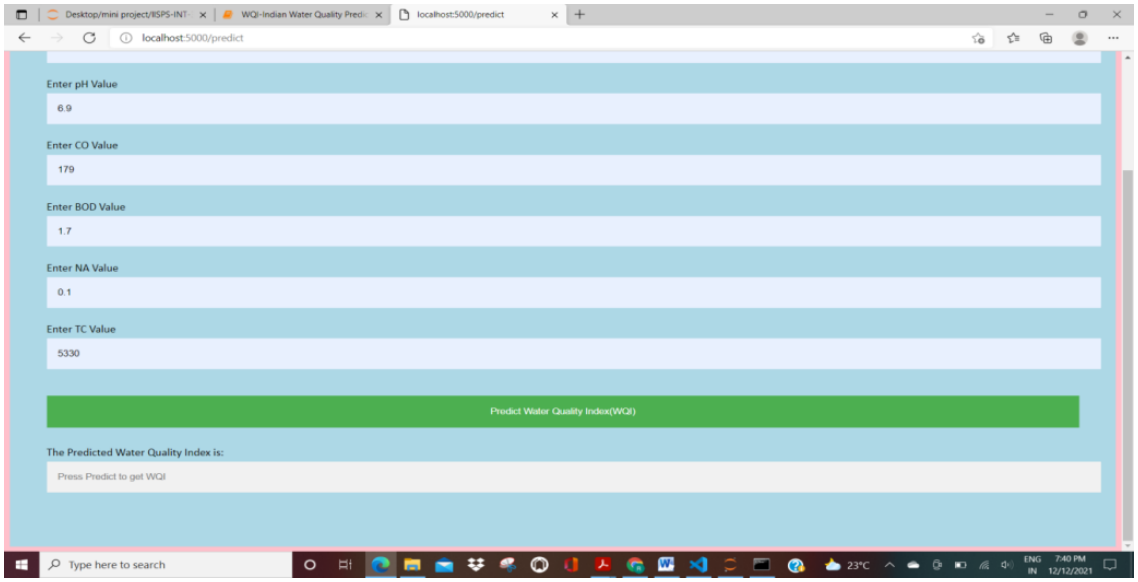

Figure 7: Result Output-2

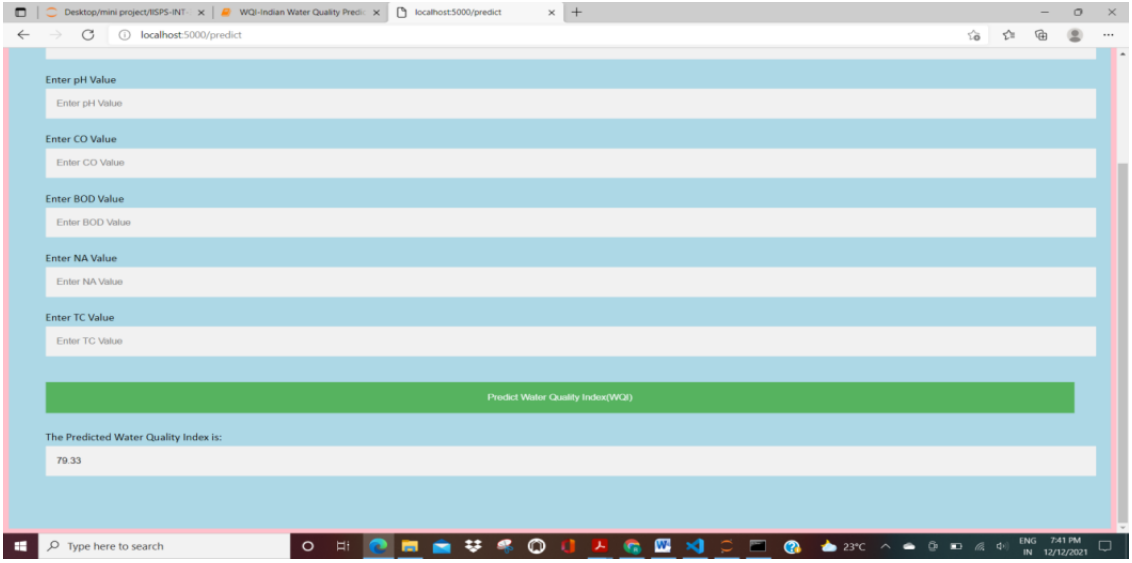

Figure 8: Result Output-3

V. CONCLUSION

This study investigated a different approach to machine learning that uses straightforward, readily accessible water quality characteristics to forecast water quality. We only need a few simple values to enter as input, and when we click on "predict," it accurately predicts the water quality index without requiring manual calculation or human participation. When we provide the values as input, it produces the output immediately because this only needs fundamental values like the ph, the do, the bod, the nitrates, and the total coliform. And this project directly translates and provides the output, reducing tremendous effort and stress, as opposed to putting up the effort of employing equations to convert the ph, nirates, do, etc. into a number according to their value. By entering a few simple values, this can be utilised by a variety of persons or researchers to forecast the WQI quickly and accurately.

In the future, we can incorporate the research's findings into a sizable IoT-based online monitoring system that uses only the sensors for the necessary metrics. Based on the real-time data supplied from the IoT system, the tested algorithms would make an immediate prediction of the water quality.

VI. REFERENCES

[1] S. Shakhari, A. K. Verma and I. Banerjee, "Remote Location Water Quality Prediction of the Indian River Ganga: Regression and Error Analysis," 2019 17th International Conference on ICT and Knowledge Engineering (ICT&KE), 2019, pp. 1-5.

[2] Bharath Singh J, Nirmitha S and Kaviya S, “Smart Urban Water Quality Prediction System Using Machine Learning,” International Conference on Recent Trends in Computing (ICRTCE-2021) 20-22 May 2021.

[3] Al-Akhir Nayan, Ahamad Nokib Mozumder, Joyeta Saha, Khan Raqib Mahmud, Abul Kalam Al Azad, “Analyzing Water Quality Using Machine Learning Algorithm”, International Journal of Advanced Science and Technology, 29(05), 14346 – 14358.

[4] Hongfang Lu, Xin Ma,Hybrid “decision tree-based machine learning models for short-term water quality prediction”,Chemosphere, Volume 249,2020,126169,ISSN 0045-6535.

[5] Asadollah SBHS, Sharafati A, Motta D, Yaseen ZM, “River Water Quality Index prediction and uncertainty analysis: a comparative study of machine learning models”, Journal of Environmental Chemical Engineering (2020).