

Water Quality Prediction

Prashanth. T¹, Keerthi. T², Lasya Priya. D³, Sowmya. S⁴, Sree Ramya. K⁵

¹Professor, Dept. of computer science and Engineering, SET-Jain University, Bengaluru-Karnataka, India

²⁻⁵Student, Dept. of computer science and Engineering, SET-Jain University, Bengaluru-Karnataka, India

Abstract - Water is used for various purposes and it has a strong impact on public health and the environment. Drinking contaminated water can cause many diseases. Even some of the packaged water that is available does not have the appropriate mineral content which in turn leads to an adverse health effect. Many people rely upon the water from river, lake which is much prone to pollution. They contain more nitrate contaminant which causes more BOD i.e. there will be less oxygen content in the water, which is not healthy for drinking. The proposed system is to check whether the given water sample is eligible for drinking by creating an application for which Machine Learning is used by taking some of the basic parameters of the water sample of tested data.

Key Words: Water quality prediction¹, Random Forest², KNN³, Logistic Regression⁴.

1. INTRODUCTION

In India, many people depend on rivers and lake water for drinking and various purposes. Considering this condition, it is required to know whether the source of water is eligible for use or is contaminated. Due to many factors contaminated water causes many problems such as diseases like dysentery, cholera, diarrhoea, etc. In this paper, an old data set is taken to predict current water quality to check whether it is eligible for drinking and other purposes.

Natural water resources like groundwater and surface water have always been the cheapest and most widely available resources of freshwater. These resources are also most likely to become contaminated due to various factors including human, industrial and commercial activities as well as natural processes. The effects of water quality deterioration are far-reaching, impacting health, environment and infrastructure in a very adverse manner. According to the United Nations

(UN), waterborne diseases cause the death of more than 1.5 million people each year, much greater than deaths caused by accidents, crimes and terrorism. Therefore, it is very crucial to devise novel approaches and methodologies for analyzing water quality and to forecast future water quality trends. Different methodologies have been proposed and applied for analysis and monitoring of water quality as well as time series analysis. The methodologies range from statistical techniques, visual modelling, analysis algorithms and prediction algorithms and decision making. Water quality index (WQI) plays a key role in conveying information on the quality of any water sample to the decision-makers. It is a mathematical equation that is used to convert complex numbers into a single number. It enhances the understanding of water quality status by integrating complex data and generating a more simplified result for the users. In recent years, researchers have used various machine learning algorithms like SVR, LS-SVM, gradient boosting and many such algorithms using the pre-recorded information. In this study, an application is created where it focuses more on generating dynamic results with the highest possible accuracy. We hereby have proposed a system that creates a Graphical user interface (GUI) which takes the inputs from the users to generate the status of the water quality instantaneously.

2. RELATED WORK

Drinking potable water is important to our health and well-being. We use water daily throughout our homes for various purposes. While only 3% of the world's water is suitable for drinking, 2% of this is found in glaciers and ice. In matter of fact, only 1% of the earth's water is accessible and potable. Hence detecting the

water quality plays a prominent role before using the water for various purposes.

In various parts of India, groundwater resources are the principal source of drinking water. Since the municipal water supplied through the surface water is not reliable, there is an increasing demand for groundwater in urban areas. In recent work, a groundwater monitoring system was designed which is based on wirelessly linked pressure sensors connected through a network. They have enabled groundwater researches to have quick access to the groundwater data with less effort and cost for their various important decision making. The prime focus of the paper was on the architectural model of an Internet-enabled WSN system for water resource management which includes network deployment, setup, discovery and maintenance. The hardware infrastructure is partitioned into three domains namely sensing domain, coordination domain and supervisory domain.

Dividing the framework into three vital important segments namely Data Acquisition Unit (DAU), Data Transfer Unit (DTU) and Data Processing Unit (DPU).

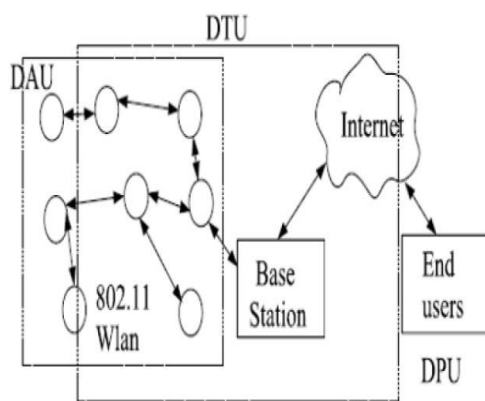


Fig -1: Framework diagram

DAU comprises a pressure sensor that is associated with a Microcontroller (FPGA) through an analogue to digital converter (ADC). ADC changes the value to digital and senses the information to the microcontroller using a pre-characterised protocol. DTU conveys the information from the DAU to DTU.

Their objective for the two technologies is to transfer information progressively to the DPU. DPU is the place where information can be controlled by experts and is made accessible to the public. The data must be measured on an hourly basis to monitor the groundwater as it is an initial step to economic success. To avoid the continuous monitoring of the groundwater a future technology called artificial intelligence and specifically machine learning is used. Hence, reducing the need for the manpower (managers). Their system has a low cost of deployment; sensors are compact and are flexible. There are few limitations as well, deploying Smart Sensor on the field is a challenging task and requires a lot of time and resources to set everything up and sensors might get damaged due to excess of water which might fail sensor network.

In [2], supervised learning was used to water quality prediction and fault detection. To predict the water quality parameters such as dissolved oxygen and turbidity play a significant role in policy decisions regarding the maintenance and use of the nation's major water bodies. The measurements are regularly corrupted due to sensor faults, fouling, and de-calibration. Thus, the United States Geological Survey (USGS) are forced to spend their time and resources to check for anomalies. Supervised learning algorithms such as support vector regression and gradient boosting is the XGBoost implementation is used in this study [2]. In this study [2] a total of 52,563 examples were collected over three years thus allowing for a more accurate prediction and a thorough understanding of the relationships between the parameters. Their aim of the study is to show how machine learning plays a key role in this area of the sciences using a combination of accessible algorithms and a detail inspection of the last data set. Thus, proving that supervised learning has the potential to significantly streamline the data validation process at the USGS, which saves valuable resources for the organization and improves the data fidelity.

Using supervised learning, they have performed a 90% /10% test data split to predict the accuracy. Comparing the two algorithms, temperature and dissolved oxygen

performed pretty well with the two algorithms resulting in better accuracy. The other parameters resulted in poor performance, and the reason for some of their performance was revealed through a plot but the reason for few parameters was not evident. To determine if the change inaccuracy is detected, the prediction score for the unscaled data is compared to the prediction score for the scaled data. The scalar value of $\alpha = 1.15$ was applied after the first 1000 samples of the testing data to detect a fault in the sensor which resulted in a reduction of performance. Therefore, concluding that supervised regression has strong potential for use in fault detection for water quality measurement system.

In [3] this paper, the fuzzy information granulation method is combined with least square support regression (LS-SVR) to set up a water quality time series model which can predict the trend of changes in water quality data with more accuracy in three days. This paper [3] concentrates on contamination of water quality. Hence, building a water quality assessment model to analyze the water quality level and makes an objective further prediction of the trend of its factors. This paper [3] sets up a partial minimum LS-SVR water quality prediction model based on the historical data of the water quality factors, taking the historical data as a dependent variable and water quality factor is used as an independent variable to predict the further trend of the water quality. The prediction intervals and the water quality factors of all water quality factors were consistent with their actual ranges. Compared with BP Neural network their prediction algorithm and the assessment model were faster in training speed and higher inaccuracy. This water quality prediction realizes the prediction of specific data and future interval changes of important water quality factors. Therefore, achieving the forecasting of future water quality conditions.

3. PROPOSED WORK

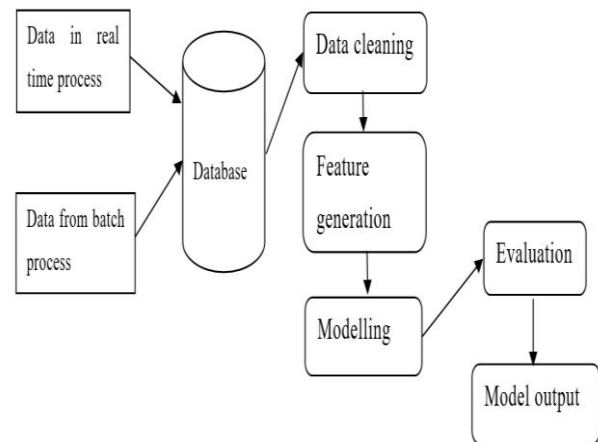


Fig -2: System architecture

Data in the real-time process: Real-time data processing involves continuous input, process and output of data which is processed in a short period.

Data from batch processes: Batch process means the data is collected in a large volume all at once. It can consist of millions of records for a day and can be stored in a variety of ways.

Database: It is a collection of information that is organised so that it can be easily accessed, managed and updated.

Data cleaning: It is the process of preparing data for analysis by removing or modifying data that is incorrect, incomplete, irrelevant, duplicated and improperly formatted.

Feature generation: It is the process of taking raw, unstructured data and defining features in your analysis.

Evaluation: It is the process of applying various algorithms on a dataset to find which gives the best accuracy for the taken dataset.

Modelling: Splitting the data into training and testing.

Model output: It is the process of displaying the accuracy of the applied algorithms.

4. METHODOLOGY

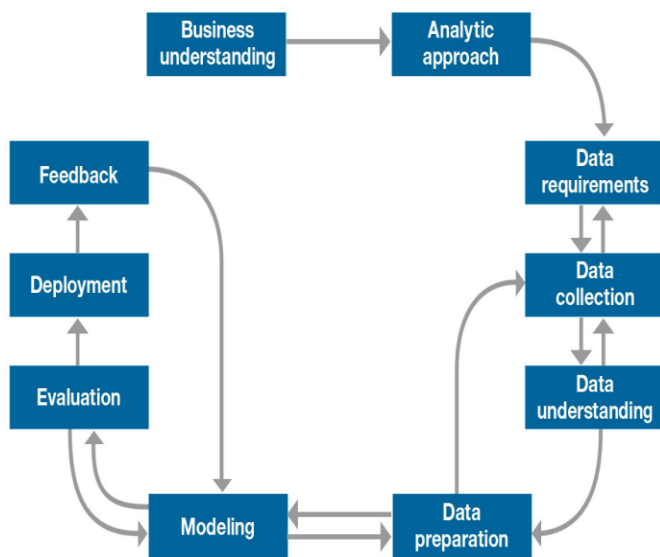


Fig-3: Methodology

5. Formula:

From [4] the calculation of WQI was done using weighed arithmetic Water Quality Index which was originally proposed by Horton (1955) and developed by Brown et al (1972). The weighed arithmetic Water Quality Index (WQI_A) is in the following form:

$$WQI_A = \frac{\sum_{i=1}^n w_i q_i}{\sum_{i=1}^n w_i}$$

Where n is the number of variables or parameters, w_i is the relative weight of the i^{th} parameter and q_i is the water quality rating of the i^{th} parameter. The unit weight w_i of the various water quality parameters are inversely proportional to the recommended standards for the corresponding parameters. According to Brown et al (1972), the value of q_i is calculated using the following equation:

$$q_i = 100 \frac{[V_i - V_{id}]}{[S_i - V_{id}]}$$

Where V_i is the observed value of the i^{th} parameter, S_i is the standard permissible value of the i^{th} parameter and V_{id} is the ideal value of the i^{th} parameter in pure water. All the ideal values (V_{id}) are taken as zero for drinking water except pH and

dissolved oxygen. The quality rating is calculated from the following equation:

$$qDO = 100 \frac{[VDO - 14.6]}{[5.0 - 14.6]} \quad (1.4)$$

where VDO = observed value of dissolved oxygen.

Table-1: Classification of water quality based on weighted arithmetic WQI method

WQI	STATUS
0-25	Excellent
26-50	Good
51-75	Poor
76-100	Very poor
Above 100	Unsuitable for drinking

Table 2: Calculation of Water Quality Index (WQI).

Parameter	Observed values (V_i)	Standard values (S_i)	Unit weights (w_i)	Quality rating (q_i)	$w_i q_i$
pH	7.7	6.5 - 8.5	0.2190	46.67	10.2207
electrical conductivity	96.3	250	0.3710	38.52	14.2909
BOD	1.1	5	0.3723	22.00	8.1906
NitrateNan	1.7	50	0.0412	3.40	0.1401
DO	3.2	5	0.3723	118.75	44.2106

6. CONCLUSION

This paper suggests that water is used for various purposes and it has a strong impact on public health and the environment. Drinking contaminated water can cause many diseases. In India, many industries rely on water bodies. Besides, more than 5 million people die each year from diseases caused by unsafe drinking water, lack of sanitation and insufficient water for hygiene. Around 2 million deaths occur each year from

water-related diarrhoea alone. So, detecting whether the water is drinkable or not is a very prominent part. In this paper, we have used three supervised algorithms (KNN, Logistic Regression and Random Forest) which are used to predict water quality factors generating a high degree of accuracy. Results indicated that the applied models have a suitable performance for predicting the water quality. The best accuracy was related to Random forest and the lowest accuracy was related to KNN model. This paper also builds a Graphical User Interface (GUI) which allows the users to enter inputs of the water quality parameters and displays the dynamic results thus indicating the eligibility of water quality instantaneously. Hence using this application makes the users know the quality of water immediately which in turn improves health because in India many people rely on water bodies for drinking. The current implementation does not give you the dynamic result instead uses the already stored data in the database to display the result. Using our application gives dynamic results along with a high degree of accuracy since Machine Learning gives more accurate results which lead to better decisions and potable water. Considering the results being encouraging, there is still an abundant opportunity for research in this area.

7. RESULTS AND DISCUSSION

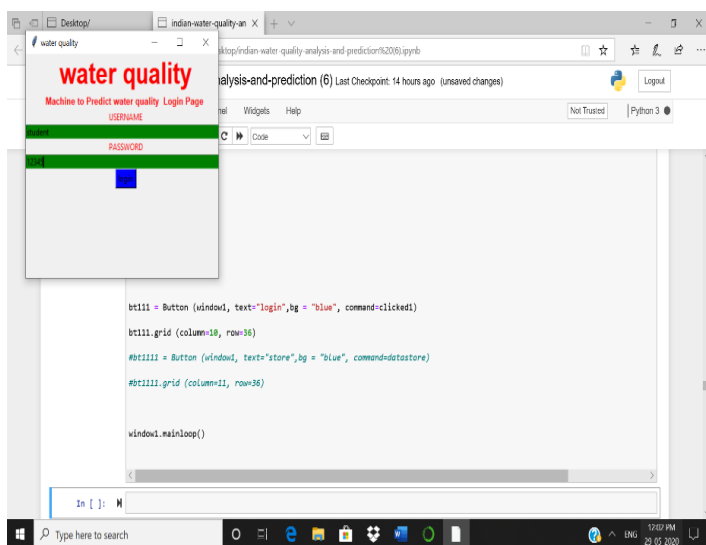


Fig-4: User login page

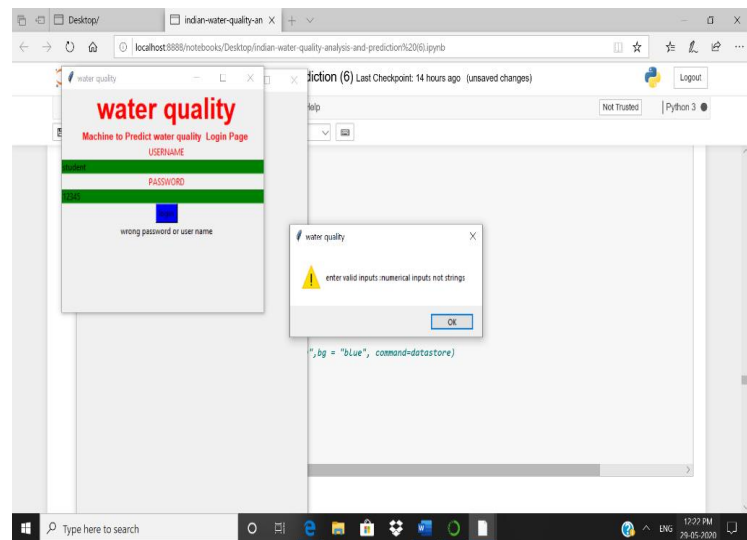


Fig-5: displaying a warning message to enter only numeric values

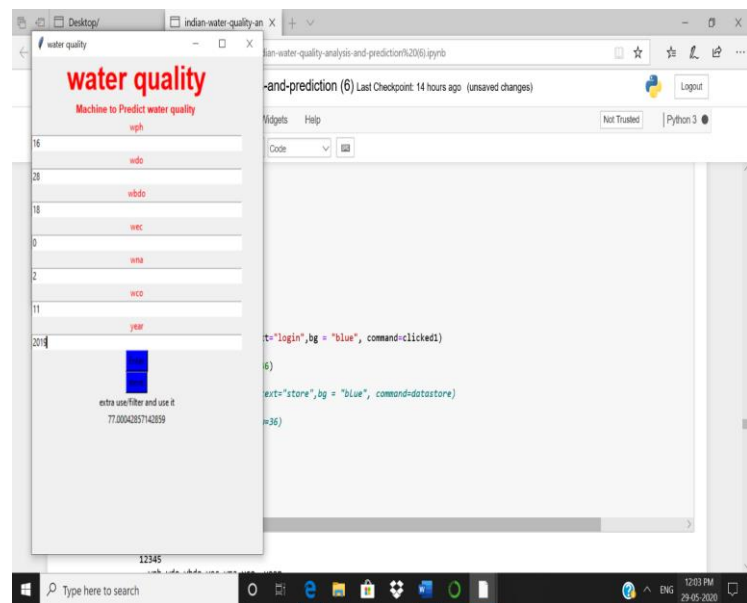


Fig-6: predicting the water quality and displaying its index value (WQI)

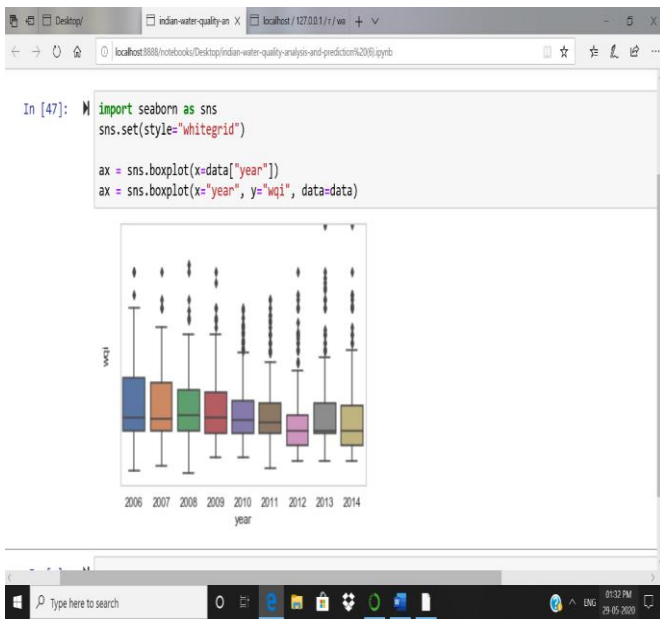


Fig-7: a boxplot representation of parameters year and wqi

REFERENCES

- [1] Aaditya Gupta, Chesta Bansal, and Agha Imran Husain, "Ground Water Quality Monitoring Using Wireless Sensors and Machine Learning", IEEE, Vol.#, No.#, pp.121-125,2018.
- [2] Kathleen Joshlyn and John Lipor, "Supervised Learning Approach to Water Quality Parameter Prediction and Fault detection", IEEE, Vol.#, No.#, pp.2511-2514,2018
- [3] Sheng Cao, Shucheng Wong and Yan Zhang "Design of River Water Quality Assessment and Prediction Algorithm", IEEE, Vol.#, No. #, pp.901-906, 2018.
- [4] Touglas kwasi boah, Scethen Boakye twum, and Kenneth V. Pelig-ba "Mathematical Computation of Water Quality Index of via Dam in Upper East Region of Ghana", Hikari Ltd, Vol:3, No.1, pp.11-16, 2015.
- [5] F. J Thakor, D. K. Bhoi, H.R. Dabhi, S.N. Pandya and Nikitaraj B. Chauhan, "Water Quality Index (W.Q.I) of Pariyej Lake Dist. Kheda-Gujarat, Vol. 6, No.2, pp. 225-231, 2011