Efficient Water Quality Prediction for Indian Rivers Using Machine Learning

Yogalakshmi S.¹ & Mahalakshmi A.²

¹PG Student, Dept. of Computer Science and Engineering, Sri Shakthi Institute of Engineering and Technology, Coimbatore.
²Assistant Professor, Dept. of Computer Science and Engineering, Sri Shakthi Institute of Engineering and Technology, Coimbatore.

ABSTRACT

Water is perhaps the most fundamental component for the presence of life. The wellbeing and openness of drinking-water are significant worries all around the globe. Well-being dangers may emerge from the utilization of water sullied with irresistible specialists, poisonous synthetics, and so forth. In this paper, a framework is proposed to check the water quality and caution the client before the water gets tainted. There are various boundaries that can debase the water. These boundaries are considered and utilized for foreseeing when to clean the water. The framework utilizes innovations, for example, AI, Web advancement. Here we planned to utilize the accompanying boundaries, for example, pH, turbidity, DO, conductivity, and so forth. The information got from the Kaggle store for investigation. The AI calculation is utilized for foreseeing the outcome. The examination plans to give the best model forecast of water quality in river water utilizing various boundaries and water quality index. A notable AI calculation, for example, Gradient Boost, Naive Bayes, Random Forest, Decision Tree, and Deep learning algorithms were used for data interpretation and analysis. The outcome showed that the water quality record was generally in a reasonable and minor position that demonstrates of water quality was being compromised by various water poisons. A few investigations were directed to decide the ecological states of the lake, rivers that zeroed in on its actual qualities. To lessen the impact of tainted water, it is fundamental to evaluate various parts of water quality. The principal objective of this investigation is to give genuinely exact expectations to variable information. The proposed strategy accomplishes sensible precision utilizing an insignificant number of boundaries to approve the chance of its utilization of continuously water quality discovery frameworks.

Keywords: Machine learning, Water quality prediction, Regression algorithms.

1. Introduction

The current world is found natural issues like environmental change, outrageous degrees of air contamination, wild waste age ashore and in seas because of quick industrialization. With the Fourth Modern Upset social event pace, arising innovations like the Web of Things (IoT) augmented reality (VR), and Computerized reasoning (man-made intelligence) have acted the hero of these issues. The advances have empowered cultural moves by affecting economies and distinguishing various opportunities for people in the future. Water is the most significant of sources, crucial for supporting a wide range of life; in any case, it is in steady danger of contamination by life itself. Water is probably the most transferable and far-reach medium. It is assessed by The Worldwide Association for Preservation of Nature that, by 2050, requests for water, energy, and food will increment by 55%, 80%, and 60%, individually. It is normal that by 2050, the normal hole between worldwide water market interest would be around 40%. Productive water for the executives is huge, as it is a limited asset among numerous contending clients with expanding requests. Adjusting between the market interest of water requires a comprehension of the interest regarding water accessibility through numerous sources and water quality. Quick industrialization has thusly prompted the weakening of water quality at a disturbing rate. Helpless water quality outcomes have been known to be one of the main considerations of the heightening of frightening sicknesses. As announced, in agricultural nations, 80% of the sicknesses are water-borne infections, which have prompted 5 million passings and 2.5 billion ailments. The principle inspiration in this investigation is to propose and assess an elective technique dependent on...
managed AI for the efficient expectation of water quality progressively. An agent set of administered AI calculations were utilized on the said dataset for anticipating the water quality record (WQI) and water quality class (WQC).

The principal commitments of this examination are summed up as follows:

✓ A first investigation was led on the accessible information to clean, standardize and perform include choice on the water quality measures, and hence, to get the base important subset that permits high accuracy with minimal effort. Along these lines, costly and awkward lab investigations with specific sensors can be evaded in additional comparative examinations.

✓ A series of representative supervised prediction calculations were tried on the dataset worked here. The total procedure is proposed with regards to water quality mathematical investigation.

The following is just how the rest of the article is coordinated: Section 2 is dedicated to a summary of knowledge in this field. In Section 3, we look at the material and specific machine learning techniques for determining water quality with minimal parameters. Section 4 explains the impacts of classification or regression algorithms, and Section 5 ties up the paper and suggests directions for future research.

2. Literature Review

This exploration investigates the strategies that have been utilized to help take care of issues identified with water quality. In [1], Amir Hamzeh Haghibi, et al. proposed a Water quality expectation framework utilizing AI. This examination researches the presentation of man-made consciousness procedures including fake neural organizations, bunch techniques for information taking care of, and uphold vector machines for anticipating water quality segments of the Tireh river situated in the southwest of Iran. To build up the ANN and SVM, various kinds of move and piece capacities were tried, individually. Inspecting the aftereffects of ANN and SVM demonstrated that the two models have a reasonable presentation for anticipating water quality segments. In this, the accompanying seven boundaries are thought of. There are pH, SO₄, Na, Ca, Cl, Mg, HCO₃.

In [2], Umair Ahmed, et al. proposed Efficient Water Quality Prediction framework Using Supervised Machine Learning. This framework proposed the approach utilizes four information boundaries, to be specific, temperature, turbidity, pH, and complete broke up solids. The information used to direct the investigation came from PCRWR, which included 663 samples from 12 different wellsprings in Rawal Lake, Pakistan. A bunch of agents-directed AI calculations was utilized to assess WQI. They utilized both relapse and classification calculations. They utilized the relapse calculations to assess the WQI and the classification calculations to order tests into the already defined WQC. They utilized eight relapse calculations and 10 classification calculations.

In [3], C.Ashwini, et al. proposed a Water Quality Monitoring Using Machine Learning And IoT. This paper presents the affordable answer for evading the defilement of water in private overhead tanks. The nature of water is observed utilizing IoT gadgets and the future expectation of water pollution is accomplished utilizing AI calculations. The proposed framework comprises multi-sensors associated with NodeMcU to gather the water boundaries. What's more, the alarm message is shipped off the client before the water gets tainted. The framework
assists with saving the water from defilement and is likewise financially savvy. In this paper following seven boundaries are thought of. There are pH, Temperature, Turbidity, Dissolved Oxygen, Conductivity, Color, and Total natural carbon.

In [5], Shuangyin Liu, et al. proposed A half and half methodology of help vector relapse with hereditary calculation advancement for hydroponics water quality expectation. A forecast model dependent on help vector relapse (SVR) is proposed in this paper to tackle the hydroponics water quality expectation issue. They assemble a successful SVR model, the SVR boundaries should be set cautiously. This investigation presents a half and half methodology, known as genuine worth hereditary calculation uphold vector regression(RGA–SVR), which looks for the ideal SVR boundaries utilizing genuine esteemed hereditary calculations, and afterward receives the ideal boundaries to build the SVR models. The methodology is applied to foresee the hydroponics water quality information gathered from the oceanic plants of YiXing, in China. The test results exhibit that RGA–SVR outflanks the customary SVR and back-engendering (BP) neural organization models dependent on the root mean square mistake (RMSE) and mean outright rate blunder (MAPE). This RGA–SVR model is demonstrated to be a viable way to deal with anticipate hydroponics water quality.

In [6], Salisu Yusuf Muhammad, et al. proposed a Classification Model for Water Quality utilizing Machine Learning Techniques. This article proposes a reasonable grouping model for ordering water quality dependent on AI calculations. The paper dissected and analyzed the exhibition of different arrangement models and calculations to distinguish the critical highlights that added to ordering the water nature of Kinta River, Perak Malaysia. Five models with particular calculations were tried and contrasted and their exhibition. In evaluating the outcome, the Lazy model utilizing the K Star calculation was the best grouping model among the five models had the most remarkable exactness of 86.67%. By and large, wastewater is hurtful to our lives, and getting logical models taking care of this issue is mandatory.

In [7], Jefferson L. Larios, et al. proposed a framework for Pattern Extraction of Water Quality Prediction Using Machine Learning Algorithms of Water Reservoir. The point of this examination is to beg information-digging strategies for design extraction and model expectation of water quality in water supplies utilizing various boundaries and water quality records. A notable AI calculation, for example, Naive Bayes, Decision Trees, Random Forest, Gradient Boost, and Deep learning calculations were utilized for information investigation and translation. The outcome demonstrated that the water quality list was generally in a reasonable and peripheral position that shows water quality was being compromised by various water contaminations. The investigation begged information mining procedure to broke down the water quality checking consequences of Laguna Lake utilizing seven boundaries specifically capability of hydrogen, Ammonia, Nitrate, Inorganic Phosphate, Biochemical Oxygen Demand, Dissolved Oxygen and Fecal Coliform from 9 distinct stations. The water quality record was determined utilizing the Canadian Council of Ministers of the Environment mini-computer with the comparing depiction of the natural condition. The examination used five models for design extraction and forecast. Among the classifiers, the Decision Tree acquired the most noteworthy exactness and accuracy. This shows the degree of trust regarding the model expectation.
In [8], Consolata Gakii, et al. proposed A Classification Model For Water Quality Analysis Using a Decision Tree. This examination presents a characterization model utilizing a choice tree to investigate water quality information from various districts in Kenya. The water quality is vital in guaranteeing residents will drink clean water. The use of a choice tree as an information mining strategy to foresee clean water dependent on the water quality boundaries can facilitate crafted by the research center technologist by anticipating which water tests ought to continue to the following stage of the investigation. The optional information from the Kenya Water Institute was utilized for the production of this model. The information model was executed in WEKA programming. Grouping utilizing a choice tree was applied to order/foresee the spotless and not perfect water. The examination of water Alkalinity, pH level, and conductivity can assume a significant part in surveying water quality. Five choice tree classifiers which are J48, LMT, Random timberland, Hoeffding tree, and Decision Stump were utilized to construct the model and the exactness looked at. J48 choice tree had the most elevated precision of 94% with Decision Stump having the least exactness of 83%.

3. Materials and Methods

3.1. Data Collection

The Kaggle repository dataset included 1990 samples of various river water collected between 2004 and 2014. There were seven criteria in all.

3.2. Data Pre-processing

The data for this analysis was taken from the kaggle repository. After the data was cleaned, it was translated to a range of 0–100 using q-value normalisation in order to measure the WQI using the six available parameters. All original values were standardised using z-score after the WQI was measured, so they were on the same scale. Following that is a rundown of the whole process.

<table>
<thead>
<tr>
<th>Serial No.</th>
<th>Parameters</th>
<th>Standard</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Ph</td>
<td>6.5-8.5</td>
</tr>
<tr>
<td>2</td>
<td>Dissolved Oxygen</td>
<td>5.0-6.0mg/l</td>
</tr>
<tr>
<td>3</td>
<td>Chemical Oxygen Demand</td>
<td>250mg/l</td>
</tr>
<tr>
<td>4</td>
<td>Total dissolved solids</td>
<td>500mg/l</td>
</tr>
<tr>
<td>5</td>
<td>Chlorides</td>
<td>250mg/l</td>
</tr>
<tr>
<td>6</td>
<td>Sulphate</td>
<td>200mg/l</td>
</tr>
<tr>
<td>7</td>
<td>Nitrate-nitrogen</td>
<td>45mg/l</td>
</tr>
<tr>
<td>8</td>
<td>Fluoride</td>
<td>0.6-1.2mg/l</td>
</tr>
<tr>
<td>9</td>
<td>Phenol</td>
<td>0.001mg/l</td>
</tr>
<tr>
<td>10</td>
<td>Iron</td>
<td>0.3mg/l</td>
</tr>
<tr>
<td>11</td>
<td>Conductivity</td>
<td>0.5-1.5mS/cm</td>
</tr>
<tr>
<td>12</td>
<td>Magnesium</td>
<td>30mg/l</td>
</tr>
<tr>
<td>13</td>
<td>Calcium</td>
<td>75mg/l</td>
</tr>
</tbody>
</table>

Fig.1. Standard water quality parameters

3.3. Water quality index (WQI)

The water quality index is a specific metric that depicts the condition of water, and it is calculated using a variety of parameters that are truly intelligent of the water's quality. To ordinarily figure the WQI, nine water quality boundaries are utilized, yet in the event that we didn't have every one of them, we could in any case appraise the water quality list with at any rate six characterized boundaries. We had six boundaries, in particular fecal coliform,
pH, temperature, turbidity, and absolute broke down solids, conductivity in our dataset. We have used nitrites as the seventh boundary because the weight and relative importance of nitrites in the WQI computation is equal to that of nitrates in various WQI examinations. Utilizing these boundaries and their doled out loads, we determined the WQI of each example.

3.4. Water quality class (WQC)

We described the water quality class (WQC) of each example using the WQI in classification calculations whenever we assessed the WQI.

3.5. Data Analysis

After all the information preparing, for information investigation, a few AI calculations were utilized to anticipate the WQI and WQC utilizing the insignificant number of boundaries.

3.6. Machine Learning Algorithms

Both regression and classification calculations were utilized. The regression calculations were used to assess the WQI, and the classification calculations were used to group tests into the previously specified WQC.

The accompanying calculations were utilized in our investigation:

(1) **Linear Regression**

One of the most common and commonly used Machine Learning calculations is linear regression. It is an observable technique that is used in predictive research. It is utilized when we need to predict the estimation of a variable dependent on the estimation of another variable.

Direct Regression is an administered AI calculation where the predicted yield is ceaseless and has a steady slant. It's utilized to anticipate values inside a persistent territory.

(2) **Polynomial Regression**

The objective of polynomial regression is to display a non-straight connection between the independent and dependent factors. In the event that we apply a straight model on a direct dataset, at that point it gives us a decent outcome as we have found in Simple Linear Regression, yet on the off chance that we apply a similar model with no change on a non-straight dataset, at that point it will create an uncommon yield.

Because of which misfortune capacity will build, the blunder rate will be high, and exactness will be diminished. So for such cases, where information focuses are organized in a non-direct design, we need the Polynomial Regression model.

(3) **Random Forest Random**

Random forest (RF) is a gathering classifier that utilizes different models of a few DTs to acquire a superior forecast execution. It makes numerous arrangement trees and a bootstrap test procedure is utilized to prepare each tree from the arrangement of preparing information. Since the Random forest consolidates various trees to
anticipate the class of the dataset, it is conceivable that some choice trees may foresee the right yield, while others may not. Be that as it may, together, all the trees foresee the right output.

(4) Gradient Boosting Algorithm

Gradient boosting classifiers are a set of AI calculations that combine a number of weak learning models to create a strong predictive model. When it comes to angle boosting, decision trees are widely used. GBM consolidates the expectations from different choice trees to create the last forecasts. The hubs in each choice tree take an alternate subset of highlights for choosing the best split. This implies that the individual trees aren't all very similar and subsequently they can catch various signs from the information.

Moreover, each new tree considers the blunders or slip-ups made by the past trees. Thus, every progressive choice tree is based on the blunders of the past trees. This is the means by which the trees in a slope boosting machine calculation are assembled successively.

(5) Support Vector Machines

Support Vector Machine is an administered AI calculation that can be utilized for both grouping or regression difficulties. The aim of the SVM calculation is to find the best line or option limit for isolating n-dimensional space into classes so that we can confidently classify the new information point later. A hyperplane is the name for this best-choice limit. SVM selects the exceptional focuses/vectors that assist in the creation of the hyperplane.

These outlandish scenarios are known as aid vectors, and the resulting equation is known as a Support Vector Machine.

(6) Ridge Regression

Ridge regression is a model tuning technique that is utilized to break down any information that experiences multicollinearity. This strategy performs L2 regularization. At the point when the issue of multicollinearity happens, least-squares are impartial, and fluctuations are enormous, this outcomes in anticipated qualities to be far away from the real qualities. An overall straight or polynomial relapse will fall flat if there is high collinearity between the free factors, so to take care of such issues, Ridge relapse can be utilized.

Edge relapse is a regularization procedure, which is utilized to decrease the unpredictability of the model. It is likewise called L2 regularization. It assists with taking care of the issues in the event that we have a larger number of boundaries than tests.

(7) Naïve Bayes Classifier

The Nave Bayes algorithm is a supervised learning algorithm that is used to solve characterization problems. It is based on the Bayes hypothesis. It's mostly used in a material grouping that includes a high-dimensional preparation dataset. The Bayes Classifier is a simple and effective classification calculation that aids in the development of fast AI models capable of making accurate forecasts.

It's a probabilistic classifier, which means it makes predictions based on an object's probability.
Logistic regression is used to calculate the chances of proportioning in the presence of several logical variables. The technique is very like numerous direct relapses, with the special case that the reaction variable is binomial. The outcome is the effect of every factor on the proportion of the chances of the noticed occasion of interest. Logistic regression predicts the yield of a clear-cut ward variable. As a consequence, the end result should be an undiluted or discrete value. It may be Yes or No, 0 or 1, true or False, and so on, but instead of having a particular reward as 0 or 1, it offers probabilistic attributes that fall somewhere between 0 and 1.

**K Nearest Neighbor**

The k-nearest neighbours (KNN) algorithm is a straightforward, easy-to-implement AI calculation that can be used to address both grouping and relapse issues. The K-NN method saves all available data and arranges another data point based on proximity. This means that as new knowledge becomes accessible, it can be quickly grouped into a useful suite classification using K-NN calculations. The K-NN equation can be used for both Regression and Classification, but it is most commonly used for Classification problems. K-NN is a non-parametric measure, which means it makes no assumptions about simple data.

**Decision Tree**

The estimation of a decision tree is classified as managed learning. They can be used to deal with both relapse and agreement problems. The tree representation is used in the decision tree to resolve the problem where each leaf hub compares to a class mark and attributes are addressed on the tree's internal hub. It can tackle issues for both downright and mathematical information. Option Tree relapse creates a tree-like configuration in which each inward hub represents a quality's "test," each branch represents the test's aftereffect, and each leaf hub represents the ultimate option or outcome.

**4. Results**

In spite of the fact that we have proposed an enormous framework when it came to execution, we so far completed just a piece of it. We implemented few algorithms such as linear regression, polynomial regression and logistic regression. Following figures shows the results of regression algorithms and data analysis.
5. Conclusion

Water is quite possibly the most fundamental assets for endurance and its quality is resolved through WQI. Traditionally, to test water quality, one needs to experience costly and bulky lab examination. This examination investigated an elective technique for AI to foresee water quality utilizing negligible and effectively accessible water quality boundaries. A bunch of agent regulated AI calculations were utilized to appraise WQI. In future works, we propose incorporating the findings of this exploration in an enormous scope IoT-based internet observing framework utilizing just the sensors of the necessary boundaries. The tried calculations would foresee the water quality promptly dependent on the ongoing information took care of from the IoT framework. For boundary readings, the proposed IoT system will use pH, turbidity, temperature, and TDS boundary sensors, which would be transmitted through an Arduino microcontroller and ZigBee handset. It would detect low-quality water until it was distributed for use and warn concerned professionals. It will ideally bring about diminishing of individuals burning-through low quality water and thusly de-raise nerve racking sicknesses like typhoid and the
runs. In such manner, the utilization of a prescriptive examination from the normal qualities would prompt future offices to help choices and policymakers.

**Declarations**

**Source of Funding**

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

**Competing Interests Statement**

The authors declare no competing financial, professional and personal interests.

**Consent for publication**

We declare that we consented for the publication of this research work.

**Code availability**

The programming code that we have used for this research is available and authors are willing to share when it is required.

**References**


