Prediction of Water Portability using Machine Learning Methods

Mohak Verma, Sukriti Jaitly, Jaisakthi S M

Vellore Institute of Technology, Vellore, India

Corresponding author: Sukriti Jaitly, Email: sukritijaitly29@gmail.com

Water covers around 3/4th of our planet's surface and is one of the most significant sources of energy for the continuation of life on the planet. In the wake of rapid urbanization and industrialization, water quality has declined at an alarming rate, leading to the spread of life-threatening illnesses and diseases. The consequences of polluted water are far-reaching, affecting every area of human existence. As a result, effective management of water is critical to ensuring that the water's quality is optimized. When data is evaluated and water quality predictions are made in advance, the consequences of water pollution may be dealt with more effectively. There have been many prior studies that have addressed this problem; nevertheless, there is still more work that needs to be done to improve the efficacy, dependability, accuracy, and usefulness of the existing water quality management methods. The goal of this research is to predict water portability by comparing the accuracy of six different machine learning models on a dataset containing water quality metrics for 3276 different water bodies and 10 features.

Keywords: Water quality prediction, Machine learning, XGBoost, SVM, Random Forest, Bagging Classifier.

2021. In Rahul Srivastava & Aditya Kumar Singh Pundir (eds.), *New Frontiers in Communication and Intelligent Systems*, 415–424. Computing & Intelligent Systems, SCRS, India. https://doi.org/10.52458/978-81-95502-00-4-44

1 Introduction

Water is the most important resource on the planet, and it is essential for the survival of the vast majority of living things, including humans. Water that is of sufficient purity is required by living creatures in order to survive. Water creatures can endure only a certain amount of pollution before they become ill. Exceeding these boundaries has a negative impact on the existence of these organisms and poses a threat to their survival. The quality of most ambient water bodies, such as rivers, lakes, and streams, is determined by precise quality criteria that are used to determine their quality. Furthermore, water specifications for various applications and/or uses have their own set of criteria. For example, irrigation water must not be overly saline, nor must it contain poisonous compounds that can be passed to plants or soil and cause ecosystems to be destroyed. Inadequate water quality has long been recognized as one of the most significant contributors to the spread of terrible diseases. According to reports, 80 percent of infections in poor nations are caused by diseases spread through contaminated water, which has resulted in 5 million deaths and 2.5 billion cases of illness to date Millions of people in the United States are infected each year with diseases such as typhoid fever, gastroenteritis, cryptosporidium infections, various forms of hepatitis, and intestinal worms such as giardiasis.

Depending on the individual industrial processes, water quality for industrial applications must have a variety of various attributes. Natural water resources, such as groundwater and surface water, are some of the most affordable sources of freshwater available. Such resources, on the other hand, might be contaminated by human/industrial activity as well as by other natural processes. As a result, significant industrial expansion has resulted in the degradation of water quality at an alarmingly quick rate. Furthermore, facilities that are poorly maintained, lack public knowledge, and have poor sanitary qualities have a substantial impact on the quality of drinking water. It is true that the repercussions of drinking water contamination can be deadly, with negative consequences for human health, the environment, and infrastructure. According to a United Nations (UN) estimate, around 1.5 million people are compromised each year because of illnesses brought on by contaminated drinking water. Approximately 80% of health issues in underdeveloped countries are attributed to dirty water, according to official statistics. The number of deaths and illnesses reported each year is expected to be close to five million per year, with 2.5 billion reported illnesses. Let's shift our focus on water quality, the phrase refers to the chemical, physical, and biological qualities of water, which are often expressed in terms of suitability for a specific use. The geology of the watershed has an impact on the quality of the water as well.

Extremely mineralized soils and rock might, for example, result in extremely mineralized water. Water-quality monitoring is the process of collecting samples and analyzing them to determine the conditions and features of the water. This chapter will explain the numerous water properties that have an impact on the designated uses of water bodies, as well as the use of volunteer monitoring to measure these parameters. Environmental factors such as physical, chemical, and biological characteristics can be utilized to determine the water quality of a specific area or a specific source of drinking water. If the values of these parameters occur more frequently than the set limitations, the results are detrimental to human health. The acceptable quality of water sources for human consumption has been measured using the Water Quality Index (WQI), one of the most effective tools for describing water quality. The Water Quality Institute makes use of water quality data and assists in the adjustment of policies that are developed by a variety of environmental monitoring agencies. It has been discovered that the usage of individual water quality variables to explain the water quality for the public is not easily comprehensible by the public. As a result, WQI has the power of condensing a large amount of information into a single number, allowing the data to be expressed in a more straightforward and logical manner. A water system's overall state can be determined by gathering information from a variety of sources and putting it all together. They improve the ability of policymakers and a collective group of peoplewho use water resources to comprehend the challenges around water quality that have been brought to their attention. The current paper examines several the most prominent water quality indicators and presents their mathematical structure, parameter set, and computations for water quality assessment purposes., as well as their advantages and disadvantages, those are currently in use around the world, as well as their advantages and disadvantages.

2 Related Work

Considering water is such an essential element in human existence, it is imperative that there are new methods when it comes to analyzing water quality and making predictions about where it's going. Many researchers have looked at the water quality issue.

Abyaneh [1] used two common Artificial neural networks and multivariate linear regression are machine learning techniques for estimating chemical and biological oxygen demands. The researchers used four variables to calculate the chemical and biological oxygen demands: total suspended solids and pH, temperature, suspended solids, and total, respectively. When Ali and Qamar [2] attempted to categorize samples into different water quality classes, they used the hierarchical clustering unsupervised approach using average linkage (among groupings). In contrast, they did not include the key factors related to WQI that were used throughout the learning process, but no standardized water quality testing was used. indicator to assess the accuracy of their predictions.

An anomaly identification method for water quality data is proposed by ZHang et al. [3] The technique is based on dual time-moving windows and can identify anomalous Real-time analysis of historical pattern data. in accordance with statistical models, such as the combination model of autoregressive linear, the method was developed. This method has been evaluated with water quality data collected over a three-month period from a river quality monitoring station in the real world. According to the findings of the experiments, their algorithms can considerably reduce the percentage of false positives while also providing superior anomaly detection performance compared to the AD and ADAM algorithms.

In a study conducted by Xiang and Jiang particle swarm optimization methods were used in conjunction with least squares support vector machines to provide accurate predictions about water quality while overcoming the limitations of classic back propagation techniques, which included to be exceedingly difficult to meet and to obtain the absolute minimum. During the simulation testing, they found that, when it comes to predicting the water quality of the Liuxi River, the model performs very well [4].

The Water Quality Index (WQI) was used by Tyagi, Shweta to describe the total water quality status in a single phrase. Although it is difficult to satisfy and even more difficult to obtain the extreme lowest value, the WQI indicates the combined influence of numerous water quality indices informs the public and legislators [5]. Because there is no universally acknowledged composite water quality index, various to develop indices, governments have used and continue to use aggregated data The WQI criteria for drinking water sources have been reviewed. Moreover, this essay calls for the creation of a new, globally recognized “Water Quality Index” in a straightforward style that might be widely adopted and provide a trustworthy picture of water quality [7] The Karoon River: it was found that river dissolved oxygen levels were calculated using multi-layer perceptron, radial basis network, and adaptive neuro fuzzy inference system (ANFIS) models (Iran). The models considered nine different water quality factors, all of which were found in river water, including EC, PH, Ca, Mg, Na, Turbidity, PO4, NO3, and NO2. These models' accuracy was assessed using the coefficient of determination R2, root mean square error, and mean absolute error. Using the artificial neural network and ANFIS models, the researchers were able to estimate DO, BOD, and COD levels in river water that were remarkably close to actual measurements taken from the water itself. MLP also outperformed other models when it came to forecasting water quality characteristics. Finally, a sensitivity analysis was performed to identify the input factors' relative importance and contribution. The phosphate was found to be perhaps the most essential factor for DO, BOD, and COD.

Hence, machine learning may provide excellent results for detecting These previous investigations [6] have prompted us to do this research. Algorithms for machine learning can substantially reduce the number of incorrect predictions.

3 Dataset Description

The dataset used is called There are 3276 separate water bodies represented in the water potability.csv dataset. Our data was split 80/20 into training and test datasets, and six different machine learning strategies were employed to make predictions. Water portability. The different features in the dataset are described below:

A. pH scale and water potency

Water's pH is a critical metric for figuring out the water's acid–base balance. The WHO has established a pH range of 6.5 to 8.5 as the highest permissible value. According to WHO recommendations, the current study's pH value varied from 6.52 to 6.83.

B. Hardness:

Mineral salts such as calcium and magnesium are responsible for hardness. These salts are dissolved in water by geologic deposits. The time water spends in contact with hardness-producing substance helps define its hardness.

C. Total Dissolved Solids:

Minerals and chlorides and bicarbonates (salts with a bicarbonate of carbon) can be dissolved in water. High TDS water is extremely mineralized. Drinking water must have a total dissolved solid (TDS) of no more than 500mg/l and no more than 1000mg/l in order to be safe

D. Chloramines:

Chlorine and chloramine are the primary disinfectants in municipal water. Chloramines form when ammonia and chlorine are combined to cleanse water. Chlorine levels up to 4 mg/L (4 ppm) are considered safe in drinking water.

E. Conductivity:

The dissolved solids content affects water's electrical conductivity. Electrical conductivity (EC) measures how well a solution can conduct electricity due to its ionic mechanism. EC value of 400 S/cm is recommended by the World Health Organization (WHO).

F. Sulphates:

Sulfates can be found in a variety of natural materials, including soil, rocks, and minerals. They can be found in a variety of places, including the atmosphere, water, plants, and food. It is widely employed in the chemical industry because of its high solubility. Seawater has a sulphate concentration of 2,700 mg/L.

G. Organic Carbons:

Natural and anthropogenic sources contribute to the organic carbon (TOC) in source waters. The total carbon content (TOC) of pure water organic molecules is known as TOC. Treatment of drinking water is recommended by the US Environmental Protection Agency (EPA) at 2 mg/L, while source water is recommended at 4 mg/L.

H. Trihalomethanes:

Chlorinated water contains chemicals known as THMs. In drinking water, THM concentrations vary with organic content, chlorine disinfection level, and water temperature. THM concentrations as high as 80 ppm are considered safe.

I. Turbidity:

The amount of suspended solid particles determines turbidity. The test measures the number of colloidal particles present in waste output. The mean turbidity of Wando Genet Campus is 0.98 NTU, which is lower than the WHO- recommended turbidity level of 5.00 NTU.

J. Potability:

Potable water is acceptable for human consumption and is often referred to as "Drinking Water."

4 Data Preprocessing

Data cleansing is one of the most important and time- consuming components. Maintaining the accuracy and cleanliness of the data that is being examined will result in better findings and more reliable procedures that can be duplicated/replicated by others who want to verify or disprove the authenticity of the results provided.

Cleaning up data involves identifying corrupt or erroneous records, eliminating incorrect or unnecessary data elements, and creating an overall standardization of the data so that computers may provide reliable results.

The following were the specifics of the missing values in our data:

Table 1. Table captions should be placed above the tables.

Feature	Missing Values
pH	491
Hardness	0
Solids	0
Chloramines	0
Sulfate	781
Conductivity	0
Organic carbon	0
Trihalomethanes	162

Turbidity	0
Portability	0

We need to Drop missing values because water quality is a sensitive data, we cannot tamper with the data by imputing mean, median, and mode.

5 Evaluation metrics

We utilized the following metrics for classification:

A. Accuracy:

The model's accuracy is measured by how many correct predictions it can generate from all the data it has access to. In Eq.1, precision is defined as the difference between the False positives and false negatives, as well as real positives and true negatives, as well as the false positive and the false negative.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

B.Precision:

Precision can be defined as the percentage of instances of a positive class that have been accurately identified out of all the instances of that class that have been correctly categorized. Percentages are used to gauge precision. We employ Eq. 2's approach to calculate precision, where TP stands for true positive and FP for false positive.

$$Precision = \frac{TP}{TP + FP}$$

C. Recall:

Positive class recall measures the proportion of instances of that class that were properly classified. The recall rate is determined using the method given in Eq. 3, in which TP denotes true positive and FN denotes false negative, respectively.

$$Recall = \frac{TP}{TP + FN}$$

D. F1 Score:

Because precision and memory, taken separately, weutilized their harmonic mean to represent the F1 score in Eq. 4, which covers both features and more precisely reflects the overall accuracy measure despite the fact that it does not cover all aspects of precision than either precision or recall. It has a value between 0 and 1. Greater the score, the higher the accuracy.

$$F1\ Score = \frac{2 * Precision * Recall}{Precision + Recall}$$

6 Water Portability Prediction model

Multiple prediction models were created to compare them and choose the best performing one for future usage. We employed six distinct machine learning models to predict water portability, with 80 percent of the data separated into training sets and 20 percent into testing sets. In our research, we used the following algorithms:

A. K Nearest Neighbor:

To classify, the K nearest neighbor method locates the supplied points that are the closest to N neighbors and assigns them to the class that has the greatest number of neighbors. It is possible to settle a tie using a variety of methods, such as increasing n or adding bias towards one class, depending on the circumstances. When dealing with big datasets, K closest neighbor is not recommended since it does all the processing through all the training data, determining the closest neighbors each time.

$$P(y = j|X = x) = \frac{1}{K} \sum_{i \in \mathcal{A}} I(y^{(i)} = j)$$

B. Decision Tree:

In statistics, an easy-to-understand method for solving classification and regression problems is the decision tree. After training, the decision tree uses all the necessary input factors to make decisions in the decision tree. Entropy is used to choose the root variable, and then it searches for values for each of the other parameters depending on that selection. Every parameter choice is organized in from the top to the bottom of a decision tree, and the decision is projected depending on the values of various parameters

$$\text{Info(D)= - } \sum_{i=1}^{m} pi \log_2 pi$$

C. Random Forest:

When making decisions, the random forest model considers the outcomes from all the base models applied to different parts of the input data. A random forest's core model is a decision tree, which has all the advantages of a decision tree while also mixing many models for greater efficiency.

$$MSE = \frac{1}{N} \sum_{i=1}^{N} (fi - yi)^2$$

D. Bagging Classifier:

In a bagging classifier, random subsets of data are used to fit numerous base classifiers, and the predictions then used of each base classifier are averaged to produce the final prediction. It makes a significant difference in terms of variance.

E. XG Boost:

XGBoost employs a framework for gradient boosting. Regularization is better in XGBoost than in gradient boosting. As a result, it reduces overfitting. Since it supports multiprocessing due to its speed being far superior to standard gradient boosting It's already set up to deal with omitted information. XG Boost splits the node up to the maximum depth specified and has a constructed cross- validation function, making it simpler to determine the number of boosting rounds at every iteration than gradient boosting, which is a greedy method. The XGBoost algorithm has a plethora of hyper - parameters which must be tuned for optimal performance.

F. Support Vector Machines:

SVMs, or support vector machines, are classification machines that can produce the best results with a small quantity of data while being fast and dependable. The SVM algorithm achieves classification by locating the hyper- point in n-dimensional space that distinguishes each data item displayed as a point, and each feature is a distinct coordinate value. The hyperplane in multidimensional space is built iteratively, which decreases the possibility of inaccuracy. Support Vectors are the coordinates of each individual vector. The Classifier fits the input data and returns the best-fit hyperplane that categorizes the data. After obtaining the hyperplane, some features can be input to the classifier to obtain the anticipated class.

Table 2. Classification results

Machine Learning Model	Accuracy	F1-Score	Recall	Precision
K- Nearest Neighbors	0.76	0.58	0.83	0.74
Decision Tree	0.80	0.82	0.88	0.77
Random Forest	0.84	0.84	0.82	0.85
Bagging Classifier	0.84	0.84	0.83	0.86
XGBoost	0.79	0.80	0.81	0.79
SVM	0.71	0.70	0.65	0.75

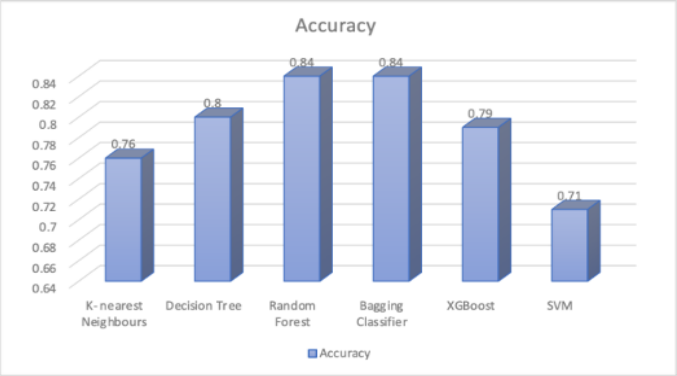

Fig. 1. Comparison chart for accuracy of different classifiers used

After the dataset is pre-processed, it goes through different predictive machine learning classifier models. The different models which we have employed in our system use different predictive analysis based on the mathematics involved in their respective processing. For comparative study of the results of the models, we have deployed a table with attributes corresponding to each model's precision, accuracy, recall and F1 score. Bagging classifier and random forest model gives the highest accuracy, while decision tree and XG boost gives respectable accuracy results, SVM and KNN majorly under performs and gives a very mediocre accuracy level. For precision and recall Bagging classifier edges out random forest with slightly better results. Decision tree gives the highest recall but lacks precision, XG Boost gives decent results with both recall and precision having respectable results. SVM and KNN give very varied results and therefore we can conclude that for the given dataset which we have used, KNN and SVM do not perform well. Similar trends can be noticed for F1 score, Thus, From Table 2 and Fig. 1 we can conclude that for testing the water portability Random Forest classifier and bagging classifier model gives the highest accuracy and precision

Bagging classifier works so well for our dataset as it has the advantage of allowing a group of weak learners to pool their resources to outperform a single strong data group. It also aids in the decrease of variance, so preventing. The bagging classifier edges out random forest as it is a technique for reducing decision tree variance by training several unpruned decision trees on different random subsets of the training data. Since the goal of Bagging classifier is to combine the predictions of several different base learners to provide a more accurate result. Random forests add a random variation to the bagging technique to increase model variety.

The random forest is one of the subordinates of the bagging classifier model. Large datasets with hundreds of variables can be handled with ease using the random forest method of classification. The rarer a class is compared to the others, the more easily it can help balance the data sets. Because it's fast and effective with variables, this method is well-suited for more difficult projects. It grows as many trees as possible on a subset of the data and then merges the results of all the trees. As a result, the overfitting problem in decision trees is reduced, as is the variance, which increases accuracy.

7 Conclusion

Having access to clean water is critical for human survival. Water is essential for human existence and hence its quality is monitored by a variety of water quality measurements. Traditionally, testing water quality required an expensive and time-consuming lab analysis which requires a lot of investment in infrastructure and requires a lot of pre-requisites. This study investigated a machine learning approach for predicting water quality utilizing minimum and easily accessible water quality metrics. The analysis of water quality takes in 10 different aspects to measure the quality of water. The study's data came from samples taken from 3276 distinct bodies of water. To forecast water portability, or whether the water is suitable for human consumption, a collection of algorithms for machine learning are used. Our research found that random forest and bagging classifiers beat other algorithms when it came to predicting portability.

The study effectively underlines the use of machine learning algorithms to predict and determine whether a sample of water is healthy for consumption or not. Out of the 6 machine learning algorithms deployed, bagging classifier gives the best results with random forest method having similar results with slightly lesser accuracy. The study suggests employment of a cost efficient and time saving water quality testing medium, where samples from various sources are tested through a system consisting of machine learning models; Bagging classifier and Random Forest simultaneously to give accurate and precise results on whether a sample of water is potable or not

References

[1] Abyaneh, H. Z. (2014). Evaluation of multivariate linear regression and artificial neural networks in prediction of water quality parameters. *J. Environ. Health Sci. Eng.,* 12: 40.

[2] Ali, M. and Qamar, A. M. (2013). Data analysis, quality indexing and prediction of water quality for the management of Rawal watershed in Pakistan. In *Proceedings of the Eighth International Conference on Digital Information Management*, 10–12.

[3] Zhang, J. et al. (2017). A real-time anomaly detection algorithm/or water quality data using dual time-moving windows. In *Seventh international conference on innovative computing technology*, 36–41.

[4] Xiang, Y. and Jiang, L. (2009). Water quality prediction using ls-svm and particle swarm optimization. In *Second International Workshop on Knowledge Discovery and Data Mining,* 900–904

[5] Tyagi, S. et al. (2013). Water Quality Assessment in Terms of Water Quality Index. *American Journal of Water Resources*, 1(3): 34-38.

[6] Haghiabi, A. M. (2018). Water quality prediction using machine learning methods. *Water Quality Research Journal*, 53 (1): 3–13.

[7] Emamgholizadeh, S. et al. (2014). Prediction of water quality parameters of Karoon River (Iran) by artificial intelligence-based models. *Int. J. Environ. Sci. Technol.* 11: 645–656.

[8] Dirisu, C. et al. (2016). LEVEL OF pH IN DRINKING WATER OF AN OIL AND GAS PRODUCING COMMUNITY AND PERCEIVED BIOLOGICAL AND HEALTH IMPLICATIONS. *European Journal of Basic and Applied Sciences*, 3: 7-12.

[9] EBRAHIMPOUR, M. et al. (2010). Influence of water hardness on acute toxicity of copper and zinc on fish. *Toxicology and Industrial Health*, 26(6): 361-365.

[10]Bowman, R. W., Gramms, L. C., and Craycraft, R. R. (1997). Water Softening of High TDS Produced Water. In the *International Thermal Operations and Heavy Oil Symposium*.

[11] Said K. et al. (2012). Analysis of inorganic chloramines in water. *TrAC Trends in Analytical Chemistry*, 33: 55-67.

[12]Mopper, K. and Jianguo, Q. (2006). Water Analysis: Organic Carbon Determinations. *Environment: Water and Waste*, https://onlinelibrary.wiley.com/doi/abs/10.1002/9780470027318.a0884.

[13]Ganguly, S. et al. (2009). Experimental investigation of the effective electrical conductivity of aluminum oxide nanofluids. *Powder Technology*, 196(3): 326-330.