IJITCE

ISSN 2347- 3657

International Journal of Information Technology & Computer Engineering

Email : ijitce.editor@gmail.com or editor@ijitce.com

International Journal of Information Technology & Computer Engineering

ISSN 2347–3657

Volume 13, Issue 2, 2025

Predicting Water Quality with Machine Learning

[1] Ms. Naila Fathima, [2] MAZIN SHAREEF, [3] SAARIM ABDUL RAHMAN, [4] Mohammed Muaz, [5] MOHAMMED AYAZ,

[1`] Assistant Professor, Department of AIML, Lords Institute of Engineering & Technology.

[2345] Student, Department of AIML, Lords Institute of Engineering & Technology.

ABSTRACT

Measuring water quality using machine learning methods is the main objective of this study. One way to measure the purity of water is by looking at its potability, which is a numerical expression. In order to determine the water's overall potability, this research used the following water quality criteria. The variables included turbidity, organic carbon, hardness, solids, chloromines, sulfate, conductivity, and trihalomethanes. These characteristics serve as a feature vector that represents the water quality. The study used Decision Tree (DT) and K-Nearest Neighbor (KNN) classification methods to assess the water quality class. A real dataset including data from several places in Andhra Pradesh and a parameter-generated synthetic dataset were both used in the experiments. The KNN classifier was shown to perform better than other classifiers based on the findings of two other kinds of classifiers. The results show that machine learning methods can reliably forecast the potability. Index keywords include topics such as classification, data mining, potability, and water quality parameters. Machine Learning, Supervised Learning, Decision Tree, Hyper Parameter Tuning, and Python are some of the terms used in this context. I.

INTRODUCTION

There are a lot of moving parts when it comes to water quality analysis. The many uses of water are intrinsically related to this idea. Various requirements call for distinct benchmarks. The topic of water quality prediction is now the subject of much research. In most cases, the physical and chemical characteristics of water that are most relevant to its use are the ones that decide its quality. The next step is to determine what values are considered acceptable and unacceptable for each variable. Water is deemed suitable for a certain use if it satisfies the specified characteristics. It is necessary to treat the water before using it if it does not meet these standards. Numerous physical and chemical characteristics may be used to evaluate water

METHODOLOGY

quality.Consequently, it is not feasible to provide an adequate geographical or temporal description of water quality in practice by investigating the behavior of each individual variable separately. Getting a single value from a set of physical and chemical factors is the more difficult approach. To show how each variable was equivalent to its quality level, the index contained a quality value function, which was typically linear. The concentration of a drug or the value of a physical variable was used to develop these functions in water sample research. Predicting water quality using machine learning algorithms is the primary focus of this study. Chapter Two: Methods Determining potability is the goal of the suggested system. Training and testing are the two distinct parts of it. The two parts carry out the same processes. Test results for hardness and pH used for training A variety of terminology may be used to characterize different substances, including solids, chloramines, sulphate, conductivity, organic carbon, trihalomethanes, turbidity, and potability. Here is the selection process for the data set: As a precondition to building a model, choosing a data set for water quality involves gathering key parameters that impact water quality, determining the amount of data samples, and defining the class labels for each data sample. This study's data sets consist of ten indicator parameters. These characteristics include things like hardness and pH value. The characteristics of a material may be described using a variety of terminology, including solids, chloramines, sulfate, conductivity, organic carbon, trihalomethanes, turbidity, and potability. Nevertheless, the suggested method is unconstrained by either the quantity or choice of parameters. The learning and testing framework in this research is built up using a k-fold cross validation approach, which corresponds to each data sample in the collection. Using this method, the

dataset is partitioned into k-disjointed sets of uniform size, where each set has a distribution of classes that is similar to the other. The subsets from this split are used as test sets one after the other, while the remaining subsets are used as training sets. These are the K-Nearest Neighbor (KNN) and Decision Tree (DT) algorithms. Different from one another, these methods focus on the underlying relational structure of the class label and the indicator parameters. That is why you might expect different results from each method when applied to the same dataset. Verifying how well several classifiers work on a dataset that is not known: In order to verify how well various classifiers work on an unknown dataset, data mining offers a number of indicators. The learning and testing environment was built using a repeated cross-validation technique in the Matlab caret package. To implement the classification algorithm, the following steps were taken: 1. Training and testing groups made up 80% and 20% of the dataset, respectively. (20 percent). 2. A predetermined number of iterations was used to submit the training set to recurrent cross-validation. Such was the process for training the classifiers. 3. We maximized the model's accuracy by selecting its best parameter configuration. 4. The prototype was examined closely. Classification Two data mining techniques, Decision Tree (DT) and K-Nearest Neighbor (KNN), were used to determine the river water quality class. The purpose of these parametric and nonparametric classifier approaches is to create a function that uses a training dataset to convert input variables into output variables. Since the shape of the function is unknown, several algorithms employ training data to generate output based on their assumptions about the shape of the function. Assumptions made by the parametric learning classifier are based on more solid evidence. If the data set assumptions hold, these classifiers will decide how to fix the problem. But the same classifier fails miserably if the assumptions are wrong. These classifiers learn classification tasks based on their assumptions rather than the size of the sample data set. Because of its parametric nature, this classifier is also prone to bias and other prediction errors. The Decision Tree produces highly biased results when the model uses several assumptions. In contrast to parametric learning classifiers, nonparametric classifiers are more accurate since they do not assume anything about the mapping function's shape. From the data used for training, these classifiers may derive any function. Classifiers like DT and KNN fall under this umbrella. Different from KNN, which makes use of learning methods, DT makes use of the similarity concept. Alternately stated, DT Conversely, these classifiers work just as well with small data sets that include exhaustive domain knowledge. The KNN classifier doesn't learn from data but rather sorts the training set into groups of k objects that are most like the test object. In contrast to competing classifiers, DT is domain-agnostic. It uses distance calculations between two attributes to create classification judgments. For the same water quality datasets used for training and testing, it is important to compare all of the algorithms to find the one that best approximates the underlying function. This is because each method operates in a slightly different way.

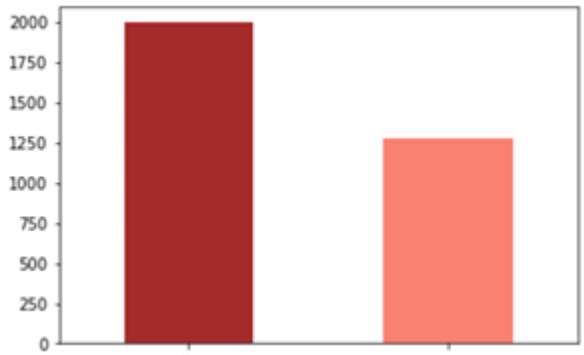


Fig: Potability Counts of Dataset

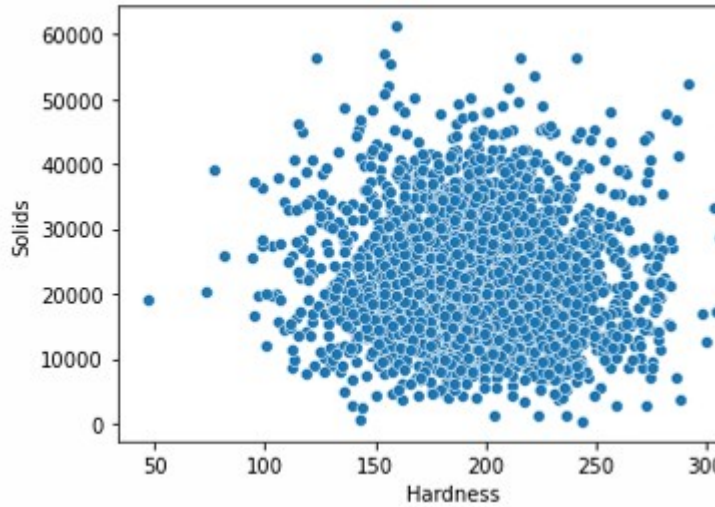
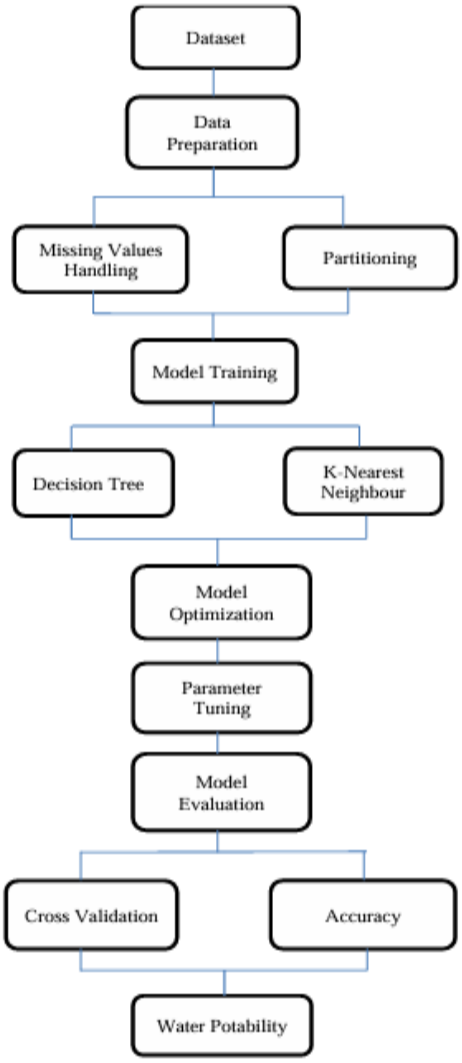


Fig: Scatter Plot of Hardness and Solids

International Journal of Information Technology & Computer Engineering

ISSN 2347–3657

Volume 13, Issue 2, 2025



Data collection and creation

MODELING AND ANALYSIS

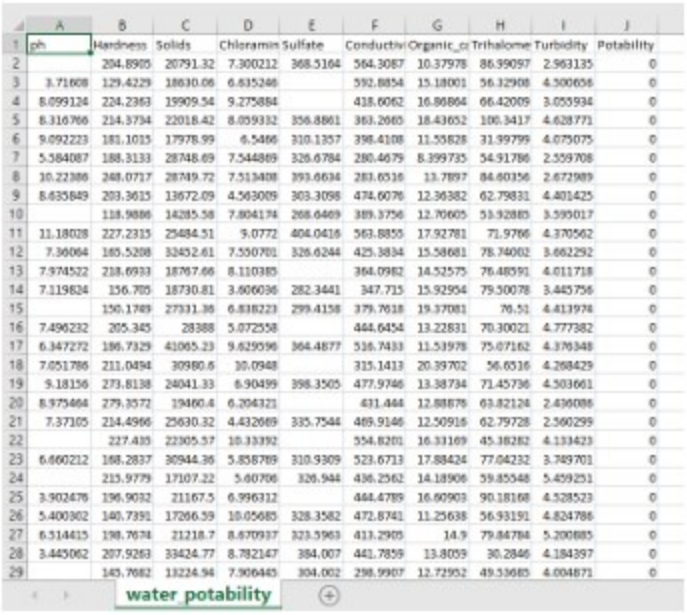
Predictions using data mining methods can only be made with subject expertise. To make informed decisions on water quality applications, it is critical to grasp the relationship between the many water quality metrics. Experts in the field or archives of previously collected data may provide this knowledge. A large synthetic data collection that was painstakingly constructed and an existing actual data set were both used for the forecasting purpose. The most important commonality between the two sets of data is that they are both tested on the same amount of indicator parameters. Data sets vary in the number of samples used for analysis. There weren't a tonne of observations in the actual dataset. An artificial data collection was created since large authentic data sets were not available. Alternatively, the synthetic data set that was created captures the same relational structures and the distribution of water quality metrics is similar to the genuine example. In order to determine the overall potability of each data set, ten water quality metrics were used. Hardness and pH are the two factors at play here. A variety of terminology may be used to characterize different substances, including solids, chloramines, sulphate, conductivity, organic carbon, trihalomethanes, turbidity, and potability. They are all key metrics with well-defined water quality criteria that are regularly monitored, which affected the choice of parameters. In contrast, the paper's detailed predictive modeling can work with any parameter values. Information compiled in an artificially In order to use data mining techniques, you need a target data set. In most cases, a big enough dataset may include patterns that are intended to be discovered by data mining. To provide a realistic way to gather this massive data set, a synthetic data collection was constructed. Careful consideration of potential ranges for water quality parameters went into the production of this synthetic data collection. One advantage of these concentration ranges is that they were created after considering water quality standards set by national and international organizations like the EU, WHO, CPCB, and others. Scientific data was also taken into account during their development. For each of the ten parameter concentration levels that were considered, one sample was taken. In order to build a prediction model using the classification approach, the dataset that will be used must be supervised. The next thing to do was to create a supervised setting for the numerical data set that was made by labeling each case to predict the amount of water pollution. This was accomplished by finding the potability for each instance of the 10 parameters' concentration levels.

International Journal of Information Technology & Computer Engineering

ISSN 2347–3657

Volume 13, Issue 2, 2025

SN.	Algorithm Type	Accuracy score	Precision	Rec
1	Decision Tree	58.5	0.42	0.3
2	K-Nearest Neighbour	61.7	0.43	0.1



	A	B	C	D	E	F	G	H	I	J
1	ph	Hardness	Solids	Chloramin	Sulfate	Conductiv	Organic_cz	Trihalome	Turbidity	Potability
2		204.8905	20791.32	7.300212	368.5164	564.3087	10.37978	86.99097	2.963135	0
3	3.71608	129.4229	18630.06	6.635246		592.8854	15.18001	56.32908	4.500656	0
4	8.099124	224.2363	19909.54	9.275884		418.6062	16.86864	66.42009	3.055934	0
5	8.316766	214.3734	22018.42	8.059332	356.8861	363.2665	18.43652	100.3417	4.628771	0
6	9.092223	181.1015	17978.99	6.5466	310.1357	398.4108	11.55828	31.99799	4.075075	0
7	5.584087	188.3133	28748.69	7.544869	326.6784	280.4679	8.399735	54.91786	2.559708	0
8	10.22386	248.0717	28749.72	7.513408	393.6634	283.6516	13.7897	84.60356	2.672989	0
9	8.635849	203.3615	13672.09	4.563009	303.3098	474.6076	12.36382	62.79831	4.401425	0
10		118.9886	14285.58	7.804174	268.6469	389.3756	12.70605	53.92885	3.595017	0
11	11.18028	227.2315	25484.51	9.0772	404.0416	563.8855	17.92781	71.9766	4.370562	0
12	7.36064	165.5208	32452.61	7.550701	326.6244	425.3834	15.58681	78.74002	3.662292	0
13	7.974522	218.6933	18767.66	8.110385		364.0982	14.52575	76.48591	4.011718	0
14	7.119824	156.705	18730.81	3.606036	282.3441	347.715	15.92954	79.50078	3.445756	0
15		150.1749	27331.36	6.838223	299.4158	379.7618	19.37081	76.51	4.413974	0
16	7.496232	205.345	28388	5.072558		444.6454	13.22831	70.30021	4.777382	0
17	6.347272	186.7329	41065.23	9.629596	364.4877	516.7433	11.53978	75.07162	4.376348	0
18	7.051786	211.0494	30980.6	10.0948		315.1413	20.39702	56.6516	4.268429	0
19	9.18156	273.8138	24041.33	6.90499	398.3505	477.9746	13.38734	71.45736	4.503661	0
20	8.975464	279.3572	19460.4	6.204321		431.444	12.88878	63.82124	2.436088	0
21	7.37105	214.4966	25630.32	4.432669	335.7544	469.9146	12.50916	62.79728	2.560299	0
22		227.435	22305.57	10.33392		554.8201	16.33169	45.38282	4.133423	0
23	6.660212	168.2837	30944.36	5.858789	310.9309	523.6713	17.88424	77.04232	3.749701	0
24		213.9779	17107.22	5.60706	326.944	436.2562	14.18906	59.85548	5.459251	0
25	3.902476	196.9032	21167.5	6.996312		444.4789	16.60903	90.18168	4.528523	0
26	5.400302	140.7391	17266.59	10.05685	328.3582	472.8741	11.25638	56.93191	4.824786	0
27	6.514415	198.7674	21218.7	8.670937	323.5963	413.2905	14.9	79.84784	5.200685	0
28	3.445062	207.9263	33424.77	8.782147	384.007	441.7859	13.8059	80.2846	4.184397	0
29		145.7682	13224.94	7.906445	304.002	298.9907	12.72952	49.53685	4.004871	0

water_potability

Collecting actual data Each dataset was subjected to an analysis using 10 water quality variables to determine its overall potability. Among the metrics that were examined were hardness, turbidity, organic carbon, conductivity, ph, chloramines, sulfate, and trihalomethanes. They are all key metrics with well-defined water quality criteria that are regularly monitored, which affected the choice of parameters. In contrast, the paper's detailed predictive models can work with any parameter numbers. Section IV. Outcomes of Performance Measures

RESULTS AND DISCUSSION

As soon as the model correctly predicts the positive class, we say that the prediction is true (TP). To illustrate the operation of classification algorithms, a confusion matrix includes True Negatives (TN). False positives (FP) are positive results that the model misjudged. When the model incorrectly predicts a negative result, it is called a False Negative (FN). The most fundamental and understandable measure of performance is accuracy, which is defined as the proportion of observations that were correctly predicted relative to the total number of observations.

Accuracy = TP+TN/(TP+FP+FN+TN)

Comparision table

CONCLUSION

Water is one of the most essential resources for life, and its potability indicates its quality. Analyzing water samples in a lab was a laborious and costly process in the past. This research investigated a different machine learning approach to water quality prediction utilizing only a handful of basic water quality variables. A collection of typical supervised machine learning methods was used for the estimation. It would alert the proper authorities if it detected water of poor quality prior to its release for consumption. Hopefully, fewer people will drink water that isn't up to par, which would cut down on cases of typhoid and diarrhea. Decision and policy makers in the future would benefit from future capabilities that are the product of a prescriptive analysis based on anticipated values.

REFERENCES

[1]. PCRWR. National Water Quality Monitoring Programme, Fifth Monitoring Report (2005–2006); Pakistan Council of Research in Water Resources Islamabad: Islamabad, Pakistan, 2007. Available online:http://www.pcrwr.gov.pk/Publications/Water%20Quality%20Reports/Water%20Quality%20Monitoring%20Report%202005-06.pdf (accessed on 23 August 2019).

ISSN 2347–3657

Volume 13, Issue 2, 2025

[2]. Kangabam, R.D.; Bhoominathan, S.D.; Kanagaraj, S.; Govindaraju, M. Development of a water quality index (WQI) for the Loktak Lake in India. Appl. Water Sci. 2017, 7, 2907–2918. [CrossRef]

[3]. Thukral, A.; Bhardwaj, R.; Kaur, R. Water quality indices. Sat 2005, 1, 99. Srivastava,

[4]. G.; Kumar, P. Water quality index with missing parameters. Int.

[5]. J. Res. Eng. Technol. 2013, 2, 609–614. The Environmental and Protection Agency, "Parameters of water quality," Environ. Prot., p. 133, 2001.