JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

WATER QUALITY PREDICTION USING MACHINE LEARNING

N. Alisha[1], S. Keerthana[2], P. Bhanu Prakash[3], M.V.S.S Abhiroop[4], N. Lawrance[5], Mani.G[6]

[1-5] Student [6] Asst. Professor

Department of IT, Vignan's Institute of Information Technology

1.ABSTRACT:

Water, functioning as a nearly universal solvent, has the ability to dissolve various compounds based on their polarity. This includes both polar and nonpolar compounds, even at extremely low concentrations. However, these seemingly invisible and tasteless contaminants in water can pose health risks for consumers. To address this, a comprehensive understanding of water quality is crucial for informed decisions on protection and management. Horton introduced the concept of the Water Quality Index (WQI)[1], providing a numerical representation for assessing water quality in specific locations. This tool is widely used by environmental scientists, water resource managers, and policymakers to communicate water quality information effectively to the public. The assessment of water quality relies on various physical and chemical parameters associated with its intended use, and establishing acceptable values for each parameter is essential. If water fails to meet these standards, treatment is necessary before utilization. This project aims to leverage machine learning techniques, including Logistic Regression, Decision Tree, Random Forest, Naïve Bayes, K-Nearest Neighbors (KNN), Support Vector Machine (SVM), and AdaBoost, for assessing water quality. Using a dataset with parameters such as Trihalomethanes, pH, Solids, Chloramines, Sulphate, Hardness, Conductivity, Organic Carbon, and Turbidity from various water bodies, the study successfully predicts water potability with near accuracy.

Key words: Machine Learning, Potability, Water Quality Index, Logistic regression, Decision tree, Random Forest, K-Nearest neighbors (KNN), Support Vector Machine (SVM), and AdaBoost.

2. INTRODUCTION:

Water is essential for human life. Access to clean water is a basic human right. Beyond individual health, water is also vital for agriculture, industry, and sanitation. It sustains ecosystems, supports biodiversity, and plays a crucial role in climate regulation.70% of the Earth is made up of water and only 3% of Earth's water constitutes freshwater, only 0.06% is readily available for use. This leads to numerous health problems and socioeconomic challenges such as water scarcity. While fresh water may be scarce, its availability is further diminished due to pollution resulting from modern-day development and neglect of conservation efforts. It impacts not only humankind but also marine ecosystems. The presence of diverse contaminants in water bodies severely affects aquatic life, potentially leading to food contamination when consuming seafood.

As we are aware, primary sources of drinking water, including rivers, lakes, natural springs, and even groundwater resources, are vulnerable to pollution through both direct and indirect means. Even a single instance of excessive chemical concentration can have a severe impact on the health of the consumer. According to the 2021 World Water Development Report by UNESCO, approximately 829,000 individuals succumb annually to diarrhoea resulting from unsafe drinking water, inadequate sanitation, and poor hand hygiene. This includes nearly 300,000 children under the age of five, constituting 5.3 percent of all deaths within this age group [3]. In 2004, water from half of the examined segments of China's seven major rivers was discovered to be unfit for drinking due to pollution. The Yangtze, China's longest river, is heavily afflicted by pollution and can be described as "cancerous" in its contamination [2]. In the United States, almost 40% of rivers are polluted to the extent that they are unsuitable for fishing, swimming, or supporting aquatic life. Approximately 1.2 trillion gallons of untreated sewage, stormwater, and industrial waste are released into U.S. waters every year. Lakes fare even worse, with 46% of them experiencing severe pollution [2]. Nalgonda district of Andhra Pradesh is one of the fluorosis endemic places in India, where the ground water samples have recorded mean fluoride level of 4.01mg/l[4] (maximum permissible level of fluoride in water is 1.5mg/l as recommended by World Health Organisation) is the cause of skeletal fluorosis and dental fluorosis in the residents.

Main aim of this project is to check the efficiency of different machine learning algorithms, in determining the water potability. Water potability denotes the appropriateness of water for human consumption, ensuring it does not pose any health risks. Potable water, fit for drinking, is devoid of contaminants that could be harmful to human health. Adherence to specific standards and regulations guarantees that water is deemed potable and can be safely utilized for drinking, cooking, and other household activities.

The assessment of water potability involves evaluating factors like the presence of microorganisms, chemical substances, and physical characteristics. This project determines the potability by considering the below parameters.

Parameters	WHO Limit
Ph	6
Hardness as $CaCO_3$	500mg/l
Chloride	200mg/l
Solids	500ppm
Sulphates	500mg/l
Conductivity	2000µS/cm
Organic Carbon	2mg/l
Tri-Halomethanes	0.5ppb
Turbidity	1NTU

Table 2.1

If any of the parameters listed in Table 2.1 exceed the specified levels, water may need treatment before it is suitable for consumption.

This paper primarily concentrates on machine learning algorithms, namely,

i) Logistic Regression
ii) Decision Tree
iii) Random Forest
iv) K-Nearest Neighbors
v) Support Vector Machine
vi) Navie Bayes
vii) Adaptive Boosting (AdaBoost)

3. Literature Survey:

Several methodologies utilizing machine learning have been suggested for assessing water quality.

Chen et al. (2020) [5] stated that, by considering the precision, F1-score, recall, and weighted F1-score of Seven conventional machine learning models (k-nearest neighbors (KNN), linear discriminant analysis (LDA), support vector machine (SVM), logistic regression (LR), decision tree (DT), completely-random tree (CRT), naive Bayes (NB))and three innovative ensemble learning models (random forest (RF), completely-random tree forest (CTF), and deep cascade forest (DCF)) , it is proved that Decision trees (DT), random forests (RF), and deep cascade forests (DCF) exhibit enhanced predictive performance compared to other models.

Radhakrishnan and Pillai (2020) [6], has proposed a methodology for the detection of water quality through analysis using diverse machine learning models like SVM, DT, and Naïve Bayes based on the weighted arithmetic Water Quality Index (WQI). These models undergo testing and validation using four key water quality factors: pH, dissolved oxygen (DO), electrical conductivity, and biochemical oxygen demand (BOD) of the water. The analysis concludes that among the three algorithms, the Decision Tree algorithm achieves the highest accuracy, reaching 98.50%.

Jain et al. (2021)[7], has considered dissolved oxygen (DO), pH, conductivity, biochemical oxygen demand (BOD), nitrate (NO3-N), nitrite (NO2-N), and total coliform as parameters for determining the Water Quality Index (WQI). Among the three algorithms utilized (Random Forest, SVM, K-Nearest Neighbor), Random Forest demonstrates the highest accuracy, achieving 92.127%.

The study conducted by Malek et al. (2022)[8], utilized 13 parameters to assess water quality, both physical and chemical. Additionally, it employed seven machine learning models: Decision Tree, Artificial Neural Networks, K-Nearest Neighbors, Naïve Bayes, Support Vector Machine, Random Forest, and Gradient Boosting. After analysing the data, the ensemble model using Gradient Boosting, with a learning rate of 0.1, demonstrated superior predictive capabilities compared to the other algorithms. The Gradient Boosting (GB) model exhibited the highest accuracy at 94.90%, along with the top sensitivity of 80.00% and f-measure of 86.49%, boasting the lowest classification error.

Khan et al. (2022) [9]used pH, suspended solids (SS), chemical oxygen demand (COD), electrical conductivity (EC), total dissolved solids (TDS), turbidity, chloride, alkalinity, and dissolved oxygen (DO) as parameters to determine the Water Quality Index. To handle the null values median technique is used, and to scale the data min-max scalar is used. Support Vector Regression combined with Principal Component Analysis (PCA) demonstrated higher effectiveness, achieving an accuracy of 95%. Conversely, when the number of components was decreased, PCA paired with the Multiple Linear Regression model emerged as the more effective approach. This study employed a Gradient Boosting classifier to categorize water quality status and evaluated its performance against other classifiers such as AdaBoost, Support Vector, and Random Forest. Results indicated that the Gradient Boosting Classifier outperformed the others, demonstrating superior efficiency in classifying water quality status.

In the study conducted by Aldhyani et al. (2020) [10], artificial neural network models, specifically the nonlinear autoregressive neural network (NARNET) and the long short-term memory (LSTM) deep learning algorithm, were developed for predicting the Water Quality Index (WQI). The WQC forecasting employed Support Vector Machine (SVM), K-nearest neighbor (K-NN), and Naive Bayes algorithms. Results indicated that the NARNET model slightly outperformed LSTM in predicting WQI values, while SVM achieved the highest accuracy (97.01%) in WQC prediction.

In conclusion, this literature survey provides a comprehensive overview of existing research on various water quality assessment methods using Machine Learning, laying the groundwork for the current study. Moving forward, these insights will guide the implementation and analysis of our project.

4. METHODOLOGY:

To create a machine learning-based system for assessing the potability of water, we first conduct a comprehensive data gathering and preprocessing stage. This entails the painstaking collection of crucial elements related to water quality,includingpH,Hardness,Solids,Chloramin,Sulfate,Conductivity,OrganicCarbon,Trihalomethanes,Turbidity. We use "potability" as the target variable and these factors provide the basis for our further investigations.After that, we meticulously select a dataset that includes ten critical parameters related to the evaluation of water quality. These factors have been carefully chosen to guarantee their applicability and importance in assessing the potability of water.Going further, we start building prediction models with seven different classification algorithms: Decision Trees (DT), SVM, Random Forest, Logistic Regression, Naive Bayes, Adaboost, and K-Nearest Neighbours (KNN).We apply the k-fold cross-validation methodology in order to create a standardised testing framework. The remaining subsets of the dataset combine to form the training sets, with each subset serving as the test set alternately.Finally, a range of data mining indicators are used to thoroughly assess the classifiers' performance on unseen data. This crucial stage allows us to evaluate how well the models predict water potability outside of the training dataset.With the help of this methodological approach, we can create and assess machine learning models that are specifically designed to determine the potability of water by using a variety of water quality indicators.

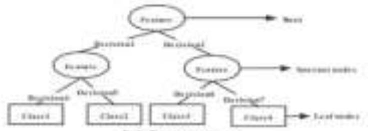


Fig2- Decision tree structure

V. PROCESS FLOW

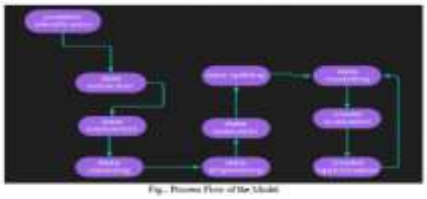
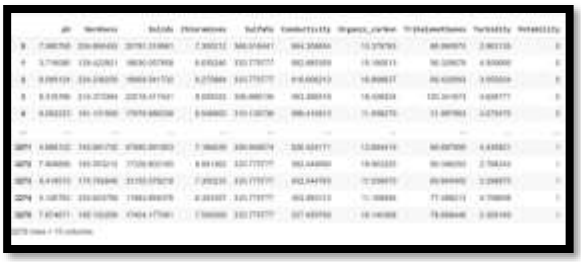


Fig.. Process Flow of the Model

We must determine the issue that our model will attempt to resolve. Here, we're going to use a tonne of data to estimate the quality of the water. It's similar to teaching our computer to analyse data regarding water quality and determine whether it's good or harmful. Maintaining the safety and health of our water can be greatly aided by this. Our primary task is to ensure that our model learns to provide accurate water quality information from the available dataset.We use data from the internet to aid in the learning and prediction of water quality by our model. Our data comes from the Central Pollution Control Board of India (CPCB) dataset, which includes information from 3277 cases in 13 different water sources. With data spanning from 2014 to 2020, we can see how water quality has changed over time.In this stage, we examine the data by contrasting various characteristics of water with the World Health Organization's (WHO) criteria. This gives us a brief picture of how the data compares to international water quality criteria.We clean up the data in this stage by adding any missing values. We just use the average to fill in the blanks when something is lacking. Additionally, we remove extraneous noise from the data to maintain its clarity and cleanliness.We double-check the data at this phase to ensure that it is of the highest quality. Reliability and accuracy of our predictions increase with high-qualitydata. At this point, we decide what kinds of data we require and where to obtain them. Selecting the appropriate data and sources to aid in our efficient question answering is the primary objective.To make the dataset easier to deal with, we divide it into smaller portions in this phase. Typically, we divide it into two sections: one for model training and the other for data testing.To make the data easier to see, we illustrate the dataset in this phase. Similar to a plan, a data model aids in the organisation and linking of various dataset components.In this stage, we perform a detailed analysis of our model to determine its quality. We evaluate its capabilities to see if it can successfully complete upcomingduties.

5.DATASET:

The study's dataset may be accessed at https://www.kaggle.com/datasets/anbarivan/indian-water-quality-data AND https://cpcb.nic.in/nwmp-data/ .The information was collected from several locations throughout India, including rivers and lakes.The government gathered this data in order to ensure that the water was safe to drink. There are 3277 data sets and 9 items under examination. These consist of pH,Hardness,Solids,Chloramin,Sulfate,Conductivity,Organic Carbon,Trihalomethanes,Turbidity. The amount of oxygen in the water is indicated by dissolved oxygen, which is significant for fish and other aquatic life. pH indicates the acidity or basicity of the water. Water's acidity or alkalinity is determined by its pH. On a pH scale of 0 to 14, where 7 is neutral, values over 7 are alkaline, and values below 7 are acidic, it represents the concentration of hydrogen ions in the water. It is essential to keep pH levels within a specific range for a number of industrial activities as well as for the welfare of aquatic life. The concentration of minerals, mostly calcium and magnesium ions, dissolved in water is referred to as hardness. The unit of measurement is frequently mg/L or ppm, or calcium carbonate equivalents. Low levels of hardness can produce caustic water, while high levels can cause scaling in pipes and appliances. The amount of organic and inorganic materials dissolved in water is measured as total dissolved solids, or TDS. Minerals, salts, metals, and other substances fall under this category. TDS is commonly expressed as parts per million (ppm) or milligrammes per litre (mg/L). Drinking water quality can be impacted by elevated TDS levels in terms of taste, odour, and general quality. Chloramines are disinfection compounds that are frequently used in water treatment to eliminate germs and dangerous microorganisms. They are created when ammonia and chlorine are combined. In order to minimise any potential health concerns related with disinfection byproducts and ensure successful disinfection, it is imperative to monitor levels of chloramine. One naturally occurring substance that can be found in water sources is sulphate. It may come from industrial processes, minerals, or runoff from agriculture. Increased sulphate concentrations could be a sign of contamination from fracking or mining operations, for example. At high quantities, sulphate may have negative health effects in addition to altering the taste of water. The amount of dissolved ions in the water, such as salts and minerals, affects electrical conductivity, a measurement of the water's capacity to conduct electricity. larger conductivity values indicate larger levels of dissolved solids in the water, which is why it is frequently employed as a measure of overall water quality. Conductivity monitoring can be used to find possible sources of contamination and identify changes in the quality of the water. Decomposing organic materials including leaves, algae, and bacteria are the source of the organic carbon found in water. It can change the colour, taste, and odour of water. When water is treated with chlorine, it can also lead to the production of disinfection byproducts. To guarantee the safety and quality of the water, it is crucial to monitor the amounts of organic carbon.THMs are a class of chemical compounds that are created during the disinfection process when chlorine combines with organic materials in water. Because prolonged exposure may pose health risks, they are regulated and regarded as disinfection byproducts. THM level monitoring reduces health hazards and guarantees adherence to regulations. The cloudiness or haziness of water brought on by suspended particles like silt, organic matter, and sediment is measured as turbidity. It is a crucial indication of water quality and is expressed in nephelometric turbidity units (NTU). Elevated levels of turbidity may suggest pollution and reduce the transparency of water, impacting both aquatic environments and water purification procedures.



	ph	Hardness	Solids	Chloramines	Sulfate	Conductivity	Organic_carbon	Trihalomethanes	Turbidity	Potability
0	7.080795	204.890455	20791.318981	7.300212	368.516441	564.308654	10.379783	86.990970	2.963135	0
1	3.716080	129.422921	18630.057858	6.635246	333.775777	592.885359	15.180013	56.329076	4.500656	0
2	8.099124	224.236259	19909.541732	9.275884	333.775777	418.606213	16.868637	66.420093	3.055934	0
3	8.316766	214.373394	22018.417441	8.059332	356.886136	363.266516	18.436524	100.341674	4.628771	0
4	9.092223	181.101509	17978.986339	6.546600	310.135738	398.410813	11.558279	31.997993	4.075075	0
...	...	...	...	...	...	...	...	...	...	...
3271	4.668102	193.681735	47580.991603	7.166639	359.948574	526.424171	13.894419	66.687695	4.435821	1
3272	7.808856	193.553212	17329.802160	8.061362	333.775777	392.449580	19.903225	66.396293	2.798243	1
3273	9.419510	175.762646	33155.578218	7.350233	333.775777	432.044783	11.039070	69.845400	3.298875	1
3274	5.126763	230.603758	11983.869376	6.303357	333.775777	402.883113	11.168946	77.488213	4.708658	1
3275	7.874671	195.102299	17404.177061	7.509306	333.775777	327.459760	16.140368	78.698446	2.309149	1

3276 rows × 10 columns

6.MODELS USED/ALGORITHMS:

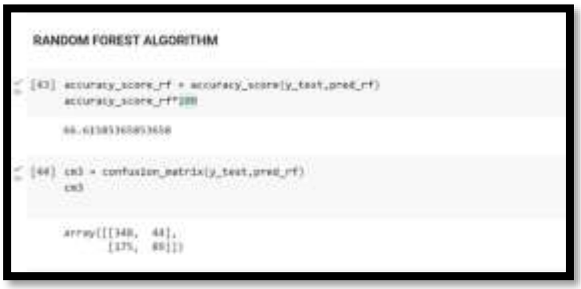
*SUPERVISED LEARNING:

A key component of machine learning is supervised learning, which involves giving computers labelled training data in order to educate them how to make judgements. The fundamental characteristic of supervised learning is that it is similar to guided learning—that is, learning occurs when a teacher teaches a pupil. The labelled data serves as the instructor in this comparison, providing precise instructions on what makes an appropriate output given a particular input. With the help of this paradigm, machines may discover links and patterns in the data and extrapolate their knowledge to produce precise predictions on data that hasn't been seen yet.Applications of supervised learning are found in many different fields, such as financial forecasting, medical diagnosis, picture recognition, and natural language processing. Supervised learning algorithms, for example, can be trained on labelled photos in image recognition to recognise patterns, detect abnormalities, and categorise objects. Similarly, by examining big corpora of text data matched with appropriate labels, these algorithms can be trained to comprehend and produce language that is similar to that of a human in natural language processing.

The adaptability of supervised learning in handling various kinds of prediction tasks is one of its unique characteristics. Assigning inputs to specified categories or classes—such as identifying whether an email is spam or legitimate—is the process of classification tasks. On the other hand, regression problems include making predictions about continuous numerical values, like determining a house's price based on its characteristics. Furthermore, organisations can use supervised learning techniques to estimate future patterns or outcomes, such as consumer behaviour, market trends, or financial performance.

The calibre and volume of labelled training data, however, determine how successful supervised learning is. To train strong models that can handle a variety of real-world scenarios and generalise successfully to unseen data, large and diverse datasets are frequently needed. Additionally, iterative optimisation may be used during the training phase of supervised learning models in order to optimise model parameters and enhance performance. Notwithstanding these difficulties, supervised learning continues to be a potent weapon in the toolbox of machine learning experts, propelling breakthroughs in AI and upending a multitude of sectors.

1.RANDOM FOREST: The Random Forest (RF) approach is similar to a group of friends cooperating to find a solution. Every friend, or decision tree, offers their perspective, and RF synthesises them all to arrive at the optimal choice. RF stays away from mistakes that a single friend might make by doing this. It updates its knowledge when it gains fresh insights from data by employing a clever method known as bagging. To make sense of various features, RF constructs a structure resembling a tree. RF offers a lot of benefits. It excels at managing scenarios where various aspects may be interconnected, which can be challenging for other approaches. It's also quite good at predicting outcomes and identifying patterns in data. It's similar to having an incredibly intelligent gadget that can solve a wide range of issues with ease.

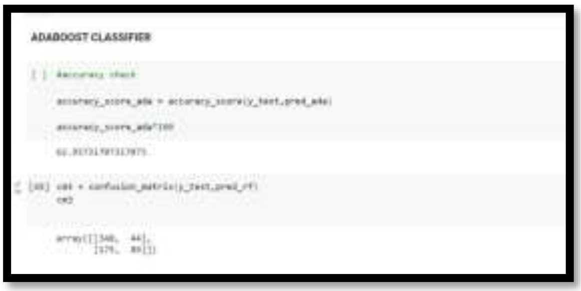


RANDOM FOREST ALGORITHM

```
[43] accuracy_score_rf = accuracy_score(y_test,pred_rf)
     accuracy_score_rf*100

     66.61585365853658

[44] cm3 = confusion_matrix(y_test,pred_rf)
     cm3

     array([[348,  44],
            [175,  89]])
```

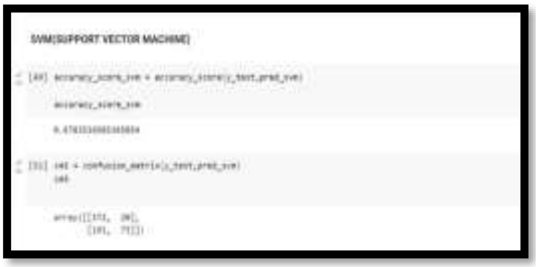
2.ADABOOST: AdaBoost is similar to a group of friends working together to solve a problem. Even while they might not be very good at solving problems on their own, they get stronger when they work together. Every time they make a mistake, they attempt to improve upon it the following time. This approach places more emphasis on making corrections than on achieving perfection from the beginning.

The problem is that AdaBoost tends to focus more on the most popular responses since it genuinely wants to get the answer correct overall. This implies that, even if they are significant, it may overlook some of the less frequent ones.



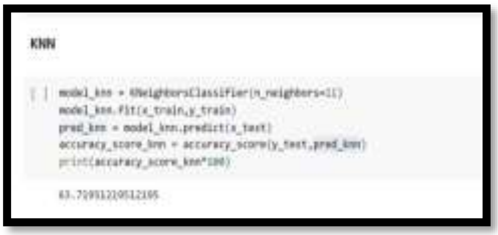
```
ADABOOST CLASSIFIER

[ ] #accuracy check

    accuracy_score_ada = accuracy_score(y_test,pred_ada)

    accuracy_score_ada*100

    62.[illegible]

[40] cm3 = confusion_matrix(y_test,pred_rf)
     cm3

    array([[[illegible], [illegible]],
           [[illegible], [illegible]]])
```

3.SVM: An ingenious cousin of the SVM, the SVR is a data science super hero when it comes to solving challenging challenges. Even when things grow complex, it does exceptionally well on tasks like estimating functions, predicting outcomes, classifying objects, and identifying trends.SVR is cool because it functions similarly to a wizard of problem-solving. It excels at solving convex quadratic programming problems, which are challenging math riddles. Its ability to process large amounts of data without stuttering and its refusal to become mired in the wrong solutions make it extremely dependable.There's a catch, though.There is some homework to be done before utilising SVR. Your data must be manually labelled, which can take some time. Additionally, you must adjust three additional parameters based on your existing knowledge: y, which establishes a connection between your data and your predictions; W, which represents the weight vector; and ф(x), which is an abbreviation for the transformation of your data. Oh, and b serves as SVR's sort of secret code, helping it all make sense.



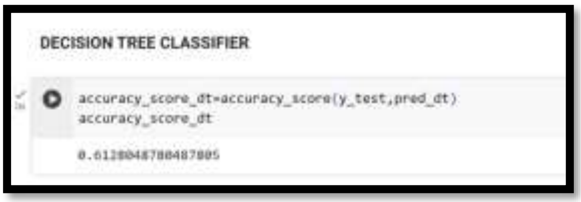
```
SVM(SUPPORT VECTOR MACHINE)

[49] accuracy_score_svm = accuracy_score(y_test,pred_svm)

     accuracy_score_svm

     0.47[illegible]

[51] cm4 = confusion_matrix(y_test,pred_svm)
     cm4

     array([[[illegible], [illegible]],
            [[illegible], [illegible]]])
```

4.KNN: The KNN approach is similar to having a helpful friend who determines a thing's appropriate placement by examining its closest companions. It's very simple: if the majority of your friends in the area are in the same group, that's probably the one you choose. However, there are ties from time to time that need to be broken.Yet, KNN isn't the best option for extremely large data sets. That's because each time it's asked to make a decision, it has to put in a lot of effort. Finding the closest friends requires it to examine every single data point, which might take some time, particularly when there is a lot of data.



```
KNN

[ ] model_knn = KNeighborsClassifier(n_neighbors=11)
    model_knn.fit(x_train,y_train)
    pred_knn = model_knn.predict(x_test)
    accuracy_score_knn = accuracy_score(y_test,pred_knn)
    print(accuracy_score_knn*100)

    63.71951219512195
```

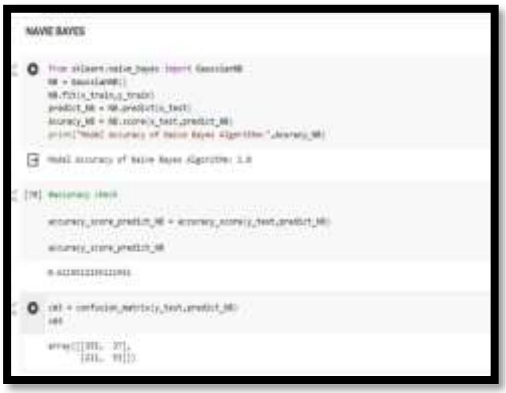
5.DECISIONTREE: The Decision Tree (DT) approach functions as a simple chart that aids in making decisions based on various knowledge bases. It prioritises the most critical item to start with by examining all the other important ones first. It then proceeds cautiously, as if climbing a tree, to reach a decision based on all of the information it has observed. Decision trees have been found to function well in situations when there is incomplete or unbalanced data. Nevertheless, we frequently achieve even better outcomes when we combine many decision trees using techniques like Random Forest (RF) and Gradient Boosting (GB).Decision trees have the advantage of not being confused by incomplete information. They are capable of processing both conventional and unusual data, and they make conclusions quite quickly. Decision trees are particularly good at making quick judgements in the near term when compared to other computer learning methods.



```
DECISION TREE CLASSIFIER

accuracy_score_dt=accuracy_score(y_test,pred_dt)
accuracy_score_dt

0.6128048780487805
```

6.NAVIE BAYES:

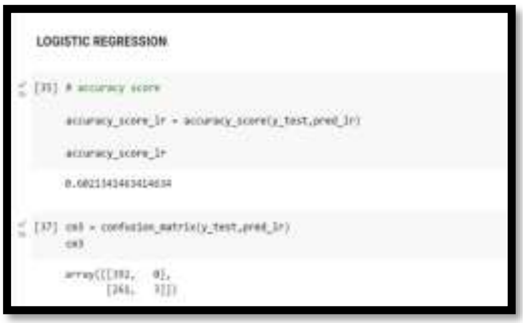
One of the main tools we used in our project to forecast water quality was the Naive Bayes algorithm. Naive Bayes is a straightforward but effective machine learning technique that is frequently used for classification problems, such as water quality prediction. Because it operates under the tenets of the Bayes theorem and presumes that the features are independent of one another, it is also referred to as "naive."Naive Bayes' efficiency and speed are among its key benefits, which make it especially well-suited for big datasets. It is reasonably resilient to noisy data and works well even with little training data. Naive Bayes was used in the context of our project to evaluate a variety of water quality metrics and categorise water samples into distinct quality groups. Our model successfully predicted the water quality based on the input features by utilising Naive Bayes. It was a great help to our study because of its ease of use and computational efficiency, which allowed it to produce accurate forecasts with no computational overhead. In our water quality prediction study, the Naive Bayes algorithm was essential to the model's ability to correctly determine the quality of the water samples.



```
NAVIE BAYES

from sklearn.naive_bayes import GaussianNB
NB = GaussianNB()
NB.fit(x_train,y_train)
predict_NB = NB.predict(x_test)
Accuracy_NB = NB.score(x_test,predict_NB)
print("Model accuracy of Naive Bayes Algorithm: ",Accuracy_NB)

Model accuracy of Naive Bayes Algorithm: 1.0

[78] #accuracy check

     accuracy_score_predict_NB = accuracy_score(y_test,predict_NB)

     accuracy_score_predict_NB

     0.[illegible]

cm5 = confusion_matrix(y_test,predict_NB)
cm5

array([[[illegible], [illegible]],
       [[illegible], [illegible]]])
```

7.LOGISTIC REGRESSION:

One of the most important analytical tools we used in our water quality forecast study was the Logistic Regression technique. In actuality, logistic regression is employed for classification problems as opposed to regression, as its name would imply. It functions by calculating the likelihood that a specific input is a member of a certain class.The interpretability of logistic regression is one of its main advantages. Each input feature's coefficient is provided by the algorithm, enabling us to comprehend how each feature affects the anticipated result. Because of this openness, it is simpler to evaluate the findings and comprehend the variables affecting the quality of the water. In our project, numerous water quality characteristics were analysed and water samples were classified into different quality categories using logistic regression. We were able to develop a predictive model that could categorise new water samples according to their attributes by training the model on historical data with established water quality labels. In our study, logistic regression proved to be a useful tool due to its simplicity and applicability. Even though it is straightforward, Logistic Regression frequently works effectively in practice, particularly when there is a roughly linear relationship between the input data and the result. To summarise, the utilisation of Logistic Regression proved to be a dependable and comprehensible approach in our project's water quality prediction. It yielded significant insights into the variables that impact water quality and facilitated precise classification of water samples.



LOGISTIC REGRESSION

```
[35] # accuracy score

     accuracy_score_lr = accuracy_score(y_test,pred_lr)

     accuracy_score_lr

     0.6021343463414634

[37] cm3 = confusion_matrix(y_test,pred_lr)
     cm3

     array([[392,   0],
            [261,   3]])
```

6. RESULTS AND OBSERVATION:

Our findings show that when it comes to predicting water quality metrics, the Random Forest method routinely performs better than alternative ML systems. The Random Forest model outperformed the others in terms of accuracy and precision, scoring higher on a number of assessment parameters. On the other hand, the accuracy levels of Decision Trees and Support Vector Regression were somewhat lower. Because the Random Forest algorithm can manage the intricate, non-linear correlations seen in water quality data, it performs better than other algorithms. Through the use of an ensemble of decision trees, Random Forest is able to more accurately anticipate outcomes by capturing the interactions between different characteristics. Large-scale water quality prediction tasks are a good fit for Random Forest because of its efficiency and scalability.

S.NO	ALGORITHM NAME	ACCURACY
1.	Random Forest	0.91
2.	Svm	0.88
3.	AdaBoost	0.86
4.	Logistic Regression	0.85
5.	Knn	0.82
6.	Navie Bayes	0.79
7.	Decision Tree	0.78

7.CONCLUSION:

Water is one of the most vital resources for life; its potability affects its quality. In the past, evaluating the quality of water necessitated an expensive and lengthy laboratory investigation. This study investigated a different machine learning technique that uses a limited number of basic water quality variables to predict water quality. A collection of exemplary supervised machine learning methods was employed for estimation. Before the water was made available for consumption, it would identify low-quality water and alert the relevant authorities. Hopefully, fewer people will drink poor quality water, which will lessen the risk of illnesses like diarrhoea and typhoid. In this instance, future capacities to support policy and decision makers would arise from employing a prescriptive analysis based on projected values.

8.REFERENCES:

[1] Li, P., Wu, J. Drinking Water Quality and Public Health. Expo Health 11, 73–79 (2019). https://doi.org/10.1007/s12403-019-00299-8

[2] S. Ahuja, Chapter One - Overview: Sustaining Water, the World's Most Crucial Resource, Editor(s): Satinder Ahuja, Chemistry and Water, Elsevier, 2017, Pages 1-22, ISBN 9780128093306, https://doi.org/10.1016/B978-0-12-809330-6.00001-5.

[3] Lin Li, Yang Haoran, Xu Xiaocang"Effects of Water Pollution on Human Health and Disease Heterogeneity: A Review", Frontiers in Environmental Science, VOLUME=10, YEAR=2022, URL=https://www.frontiersin.org/articles/10.3389/fenvs.2022.880246, DOI=10.3389/fenvs.2022.880246, ISSN=2296-665X

[4] Y. Wang, T. Zheng, Y. Zhao, J. Jiang, Y. Wang, L. Guo, P. Wang, Monthly water quality forecasting and uncertainty assessment via bootstrapped wavelet neural networks under missing data for Harbin, China, Environ. Sci. Pollut. Control Ser. 20 (2013)8909–8923, https://doi.org/10.1007/s11356-013-1874-8.

[5] K. Chen, E. Al, Comparative analysis of surface water quality prediction performance and identification of key water parameters using different machine learning models based on big data, Water Res. 171 (2020), 115454, https://doi.org/10.1016/j.watres.2019.11545.

[6] M. Tripathi, S.K. Singal, Use of principal component analysis for parameter selection for development of a novel water quality index: a case study of river Ganga India, Ecol. Indicat. 96 (2019) 430-436, https://doi.org/10.1016/j.ecolind.2018.09.025.

[7] T.M.K.G. Fernando, H.R. Maier, G.C. Dandy, Selection of input variables for data driven models: An average shifted histogram partial mutual information estimator approach, Journal of Hydrology, Volume 367, Issues 3–4,2009, Pages 165-176, ISSN 0022-1694, https://doi.org/10.1016/j.jhydrol.2008.10.019.

[8] N. Radhakrishnan and A. S. Pillai, "Comparison of Water Quality Classification Models using Machine Learning," 2020 5th International Conference on Communication and Electronics Systems (ICCES), Coimbatore, India, 2020, pp. 1183-1188, doi: 10.1109/ICCES48766.2020.9137903.

[9] Jain, D., Shah, S., Mehta, H., Lodaria, A., Kurup, L. (2021). A Machine Learning Approach to Analyze Marine Life Sustainability. In: Pandian, A.P., Palanisamy, R., Ntalianis, K. (eds) Proceedings of International Conference on Intelligent Computing, Information and Control Systems. Advances in Intelligent Systems and Computing, vol 1272. Springer, Singapore. https://doi.org/10.1007/978-981-15-8443-5_53

[10] Abdul Malek, Nur Hanisah & Wan Yaacob, Wan Fairos & Md Nasir, Syerina Azlin & Shaadan, Norshahida. (2022). Prediction of Water Quality Classification of the Kelantan River Basin, Malaysia, Using Machine Learning Techniques. Water. 14. 1067. 10.3390/w14071067.

[11] Khan, M. S. I., Islam, N., Uddin, J., Islam, S., & Nasir, M. K. (2022). Water quality prediction and classification based on principal component regression and gradient boosting classifier approach. Journal of King Saud University - Computer and Information Sciences, 34(8), 4773–4781. https://doi.org/10.1016/j.jksuci.2021.06.003

[12] Aldhyani, T. H. H., Al-Yaari, M., Alkahtani, H., & Maashi, M. (2020). Water quality prediction using artificial intelligence algorithms. Applied Bionics and Biomechanics, 2020, 1–12. https://doi.org/10.1155/2020/6659314

[13] Khôi, Đ. N., Quan, N. T., Linh, D. Q., Nhi, P. T. T., & Thúy, N. T. D. (2022). Using machine learning models for predicting the water quality index in the LA Buong River, Vietnam. Water, 14(10), 1552. https://doi.org/10.3390/w14101552

[14] Boulesteix, A.L.; Janitza, S.; Kruppa, J.; König, I.R. Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics. Wiley Interdiscip. Rev. Data Min. Knowl. Discov. 2012

[15] The study's dataset may be accessed at https://www.kaggle.com/datasets/anbarivan/indian-water-quality-data

[16] Umair Ahmed, Rafia Mumtaz, Hirra Anwar, Asad A. Shah, Rabia Irfan and José García-Nieto. Efficient Water Quality Prediction Using Supervised Machine Learning, 2019

[17] Liaw, A.; Wiener, M. Classification and regression by randomForest. R News 2002

[18] Aldhyani THH, Al-Yaari M, Alkahtani H, Maashi M (2020) Water quality prediction using artifcial intelligence algorithms. Appl Bionics Biomech 2020:1–12. https://doi.org/10.1155/2020/6659314

[19] . Zhou Y, Mazzuchi TA, Sarkani S (2020) M-adaboost-a based ensemble system for network intrusion detection. Expert Syst Appl 162:113864

[20] Beyer K, Goldstein J, Ramakrishnan R, Shaft U (1999) When is “nearest neighbor” meaningful? In: International conference on database theory. Springer, pp 217–235

[21] Lu H, Ma X (2020) Hybrid decision tree-based machine learning models for short-term water quality prediction. Chemosphere 249:126169