# Water Quality Prediction using Machine Learning

**P Ramu[1], P Suketh Reddy[2], B Anjali Reddy[3], Sriraj Katkuri[4], M Sathyanarayana[5]**

*[1,5] Professor, Dept. of Computer Science and Engineering, SNIST, Hyderabad-501301, India*
*[2,3,4] B.Tech Scholars, Dept. of Computer Science and Engineering Hyderabad-501301, India*

---------------------------------------------------------------------***---------------------------------------------------------------------

***Abstract*** **:- The survival of a nation's wellbeing greatly depends on the availability of freshwater. An essential step in managing freshwater assets is the evaluation of the quality of the water. According to the World Health Organization's annual report, people across all walks of life fall prey to the lack of access to safe drinking water. This occurs as a result of mismanagement and inefficient methodologies to prevent the occurrence of harmful water. Before using water for any purpose, it is crucial to assess its quality to ensure it is potable and hence can be used safely. For examining the safety levels of water sources, analysis of water and its underlying components is essential. A water's readiness for a particular use based on its physical, chemical, and biologicalcharacteristics is referred to as its quality.**

***Keywords****: Machine learning; potability; entropy; Gini index*

## I. INTRODUCTION

Each cell in the body receives its energy mostly from water, which also controls all the body's functions. 80% of the cerebrum is made up of water. Extreme dehydration may result in mental impairments and a loss of the ability to clearly think. One of the most important regular resources for the survival of all species on Earth is water. Water is used for many different things, such as drinking, washing, and water systems, due to its nature. Water is essential for both living things and plants. Simply put, all organic living things require a huge quantity and exceptional quality of water to exist.

Freshwater is a fundamental asset to horticulture and industry for its essential presence. Water quality observation is a key stage in the administration of freshwater assets. As indicated by the yearly report of WHO, many individuals are kicking the bucket because of the absence of unadulterated drinking water parti. It is critical to check the nature of water for its expected reason, whether it be animals watering, compound showering, or drinking water.

A tool called water quality testing can be used to locate pure drinking water. This means that for the protection of pure and clean water, the proper water testing is quite important. Water testing is crucial in determining the proper operation of water sources, evaluating the safety of drinking water and deducing the measures to curb the menace.

We can respond to questions like whether the water is fit for drinking, washing, or water systems, to name a few applications, by testing the nature of a water body. It can use the results of water quality tests to examine the nature of water in a location, a state, or the entire country, starting with one water body and moving on to the next. Since irresistible illnesses caused by pathogenic bacteria, infections, helminths, and other parasites are the most well-known and pervasive health danger associated with drinking water, microbiological quality is typically the most urgent issue to be addressed during this process.

When certain synthetic compounds are present in drinking water in excess, health risks result. These synthetics contain nitrate, fluoride, and arsenic. To the client should be given safe drinking (consumable) water for drinking, meal preparation, personal hygiene, and cleaning. To ensure purity at the point of client supply, the water must adhere to standard quality standards.

## II. BACKGROUND STUDY (LITERATURE)

Iran's Dez Catchment is one of its major watersheds. There are several sporadic and perennial streams in the watershed. One of the primary perennial streams of this basin is the Tireh River, which flows through the two largest cities and is located in the province of Lorestan. tireh River's coordination To determine the stage discharge relation and monitor the water quality components, the regional water authority (RWA) in Lorestan province (Iran) constructed hydrometry stations along this river. Constructed hydrometry stations by RWA are shown by triangular symbols. Measuring the stage discharge relation and water quality components by RWA was conducted monthly.

It is interesting that a number of measurements have been taken almost every month. More than 55 years have passed since the sampling began. The river is still being monitored now after the first measurement was reported in 1960. The components of the water quality measured by RWCA are listed in summary form. measurement parameters include temperature (T), pH, specific C–1), sulfates (SO4–2), chlorides (Cl), total dissolved solids

(TDS), sodium (Na+), magnesium (Mg+2), calcium (Ca+2).

## III. METHODOLOGY

After understanding the data, processing some attributes, andanalyzing the correlations and predictive potential of the attributes, the major goal of any data science project is model construction. like it was explained in the earlier chapters. Creating a model using the decision tree technique is one of the most straightforward and effective ways of predicting information based on test values.

A categorization paradigm called a decision tree, which resembles a flowchart, is frequently employed. Each internal node (non-leaf node) of a decision tree represents a test on an attribute, each branch a test result, and each leaf node (or terminal node) a class label. The root node is the topmost node in a tree.

Tree induction, which is the learning or creation of decision trees from a class-labeled training dataset, is a method for creating decision trees. Deduction is the process of classifying a test dataset using a decision tree that has already been built. The method of deduction involves applying the test condition to a record or data sample starting at the root node of a decision tree, then, depending on the results of the test, the appropriate branch is proceeded to. This step leads to either a leaf node or to another internal node for which a new test condition is applied. The record or data sample is subsequently given the class label associated with the leaf node.

Decision trees facilitate decision-making under certain conditions and enhance communication. The idea that different actions can result in different operational nature of the situation is easier for computational purposes. Making the best choice possible is beneficial. When instances are represented by attribute values and training data contains errors, the method performs well. In cases where the target function contains discrete output values, it is also relevant.

It automatically screens variables, and prepares data with comparatively little user work. Non-linear relations are simple to comprehend and have little impact on the performance of trees. The decision tree is helpful for exploring data and highly suggested when the requirement topredict data is based on expectations

## IV. ALGORITHMS

### 4.1 Decision Tree:

The non-parametric supervised learning approach used for classification and regression applications is the decision tree.It is organized hierarchically and has a root node, branches, internal nodes, and leaf nodes.

A decision tree has a root node at the beginning that has no incoming branches. The internal nodes, sometimes referred to as decision nodes, are fed by the root node's outgoing branches. Both node types undertake assessments based on the available attributes to create homogenous subsets, which are represented by leaf nodes or terminal nodes. All the outcomes within the dataset are represented by the leaf nodes.

Decision tree learning employs a divide and conquer strategy by conducting a greedy search to identify the optimal split points within a tree. This process of splitting is then repeated in a top-down, recursive manner until all, or the majority of records have been classified under specific class labels.

Some of the presumptions when utilizing a decision tree are:

- The entire training set is first regarded as the root.
- Categorical feature values are desired. If the values are continuous, they must first be discretized beforethe model can be constructed.
- Based on attribute values, records are dispersed recursively.
- Using a statistical approach, properties are arranged to serve as the tree's root or internal node.

To construct a decision tree, the flowing parameters aretaken into consideration:

1. Probability

Probability is defined as the possibility of the occurrenceof a value out of the total in the output data set that is considered while constructing a decision tree.

P(playgolf=yes)= 9/14; P(playgolf-no)= 5/14;

| Attributes | | | | Classes |
|---|---|---|---|---|
| Outlook | Temperature | Humidity | Windy | Play Golf |
| Rainy | Hot | High | FALSE | No |
| Rainy | Hot | High | TRUE | No |
| Overcast | Hot | High | FALSE | Yes |
| Sunny | Mild | High | FALSE | Yes |
| Sunny | Cool | Normal | FALSE | Yes |
| Sunny | Cool | Normal | TRUE | No |
| Overcast | Cool | Normal | TRUE | Yes |
| Rainy | Mild | High | FALSE | No |
| Rainy | Cool | Normal | FALSE | Yes |
| Sunny | Mild | Normal | FALSE | Yes |
| Rainy | Mild | Normal | TRUE | Yes |
| Overcast | Mild | High | TRUE | Yes |
| Overcast | Hot | Normal | FALSE | Yes |
| Sunny | Mild | High | TRUE | No |

2. Entropy

Entropy is a metric used in data science to assess how "mixed" a column is. It is specifically used to quantify disorder.

Entropy= - P (class 1) x Log (P(class 1)) - P(class 2) X Log(P(class 2)) where P denotes probability.

E(S) = [(9/14)log(9/14) + (5/14)log(5/14)] = 0.94

3. Information gain

The primary factor used to determine whether a feature should be used to split a node is information gain. The feature that results in the maximum information gain at a decision tree node, or the feature with the best split, is utilized to divide the node.

Information gain=Entropy(s) - [(Weighted average) X (Entropy of each feature)

IG(S, outlook) = 0.94 - 0.693 = 0.247

IG(S, Temperature) = 0.940 - 0.911 = 0.029

IG(S, Humidity) = 0.940 - 0.788 = 0.152

IG(S, Windy) = 0.940 - 0.8932 = 0.048

4. Gain Index

The Gini Index, also known as Impurity, calculates the likelihood that a randomly selected instance will be incorrectly classified. The likelihood of misclassification is based on this parameter.

Gini Impurity= 1 — (Probability of 'Class 1') $^2$ - (Probability of 'Class 2')$^2$

Step 1: According to S, start the tree at the root node, which has the entire dataset.

Step 2: Utilize the Attribute Selection Measure to identify the dataset's top attribute (ASM).

Step 3: Subset the S to include potential values for thebest qualities.

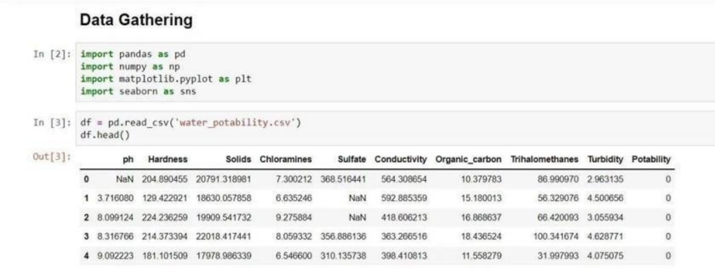
Step 4: Create the decision tree node that has the best attribute.

Step 5: Use the selections of the dataset generated to iteratively develop new decision trees

Step 6: Continue along this path until you reach a point when you can no longer categorize the nodes and you refer to the last node as a leaf node

## V. IMPLEMENTATION

Import all the necessary libraries that are needed for data visualization or to train the model. The top five rows of the data set should then be displayed after loading the data set using the Pandas method read csv().

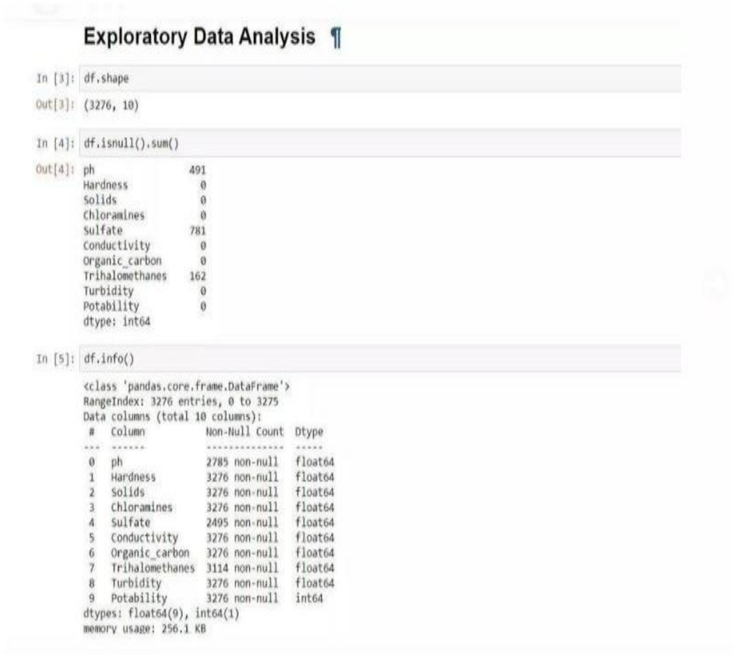

Data Gathering

```
In [2]: import pandas as pd
        import numpy as np
        import matplotlib.pyplot as plt
        import seaborn as sns

In [3]: df = pd.read_csv('water_potability.csv')
        df.head()
```

Out[3]:

| | ph | Hardness | Solids | Chloramines | Sulfate | Conductivity | Organic_carbon | Trihalomethanes | Turbidity | Potability |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | NaN | 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 | NaN | 592.885359 | 15.180013 | 56.329076 | 4.500656 | 0 |
| 2 | 8.099124 | 224.236259 | 19909.541732 | 9.275884 | NaN | 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 |

Exploratory Data Analysis should then be done. Check the data set's shape first in EDA. Check to see if there are any NULL values, as you can see in the image below for ph, Sulfate, and Trihalomethanes. then verify the data set's information.


Exploratory Data Analysis ¶

```
In [3]: df.shape
Out[3]: (3276, 10)

In [4]: df.isnull().sum()
Out[4]: ph                 491
        Hardness             0
        Solids               0
        Chloramines          0
        Sulfate            781
        Conductivity         0
        Organic_carbon       0
        Trihalomethanes    162
        Turbidity            0
        Potability           0
        dtype: int64

In [5]: df.info()
        <class 'pandas.core.frame.DataFrame'>
        RangeIndex: 3276 entries, 0 to 3275
        Data columns (total 10 columns):
         #   Column           Non-Null Count  Dtype
        ---  ------           --------------  -----
         0   ph               2785 non-null   float64
         1   Hardness         3276 non-null   float64
         2   Solids           3276 non-null   float64
         3   Chloramines      3276 non-null   float64
         4   Sulfate          2495 non-null   float64
         5   Conductivity     3276 non-null   float64
         6   Organic_carbon   3276 non-null   float64
         7   Trihalomethanes  3114 non-null   float64
         8   Turbidity        3276 non-null   float64
         9   Potability       3276 non-null   int64
        dtypes: float64(9), int64(1)
        memory usage: 256.1 KB
```

The dataset that displays the lowest value, maximum value, mean value, count, standard deviation, etc. is now described.

Finally, we take care of the missing data. We used the mean value of each feature to fill in the missing values in our features' data, handling missing data by filling in the mean value. Next, confirm whether any null values are present.

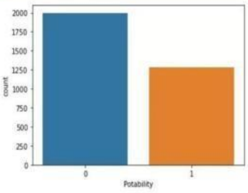
```
In [6]: df.describe()
Out[6]:
```

| | ph | Hardness | Solids | Chloramines | Sulfate | Conductivity | Organic_carbon | Trihalomethanes | Turbidity | Potability |
|---|---|---|---|---|---|---|---|---|---|---|
| count | 2785.000000 | 3276.000000 | 3276.000000 | 3276.000000 | 2495.000000 | 3276.000000 | 3276.000000 | 3114.000000 | 3276.000000 | 3276.000000 |
| mean | 7.080795 | 196.369496 | 22014.092526 | 7.122277 | 333.775777 | 426.205111 | 14.284970 | 66.396293 | 3.966786 | 0.390110 |
| std | 1.594320 | 32.879761 | 8768.570828 | 1.583085 | 41.416840 | 80.824064 | 3.308162 | 16.175008 | 0.780382 | 0.487849 |
| min | 0.000000 | 47.432000 | 320.942611 | 0.352000 | 129.000000 | 181.483754 | 2.200000 | 0.738000 | 1.450000 | 0.000000 |
| 25% | 6.093092 | 176.850538 | 15666.690297 | 6.127421 | 307.699498 | 365.734414 | 12.065801 | 55.844536 | 3.439711 | 0.000000 |
| 50% | 7.036752 | 196.967627 | 20927.833607 | 7.130299 | 333.073546 | 421.884968 | 14.218338 | 66.622485 | 3.955028 | 0.000000 |
| 75% | 8.062066 | 216.667456 | 27332.762127 | 8.114887 | 359.950170 | 481.792304 | 16.557652 | 77.337473 | 4.500320 | 1.000000 |
| max | 14.000000 | 323.124000 | 61227.196008 | 13.127000 | 481.030642 | 753.342620 | 28.300000 | 124.000000 | 6.739000 | 1.000000 |

```
In [8]: df.fillna(df.mean(), inplace=True)
        df.isnull().sum()

Out[8]: ph                 0
        Hardness           0
        Solids             0
        Chloramines        0
        Sulfate            0
        Conductivity       0
        Organic_carbon     0
        Trihalomethanes    0
        Turbidity          0
        Potability         0
        dtype: int64
```

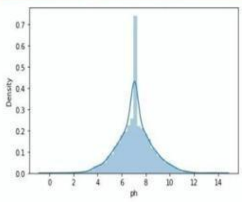
Verify the potability value counts for our target feature. then make use of seaborn's countplot function to illustrate portability.

```
]: df.Potability.value_counts()

]: 0    1998
   1    1278
   Name: Potability, dtype: int64

]: sns.countplot(df['Potability'])
   plt.show()

   /usr/local/lib/python3.7/dist-packages/seaborn/_decorators.py:43: FutureWarning: Pass the following variable as a keyword arg:
   x. From version 0.12, the only valid positional argument will be `data`, and passing other arguments without an explicit keywor
   d will result in an error or misinterpretation.
     FutureWarning
```

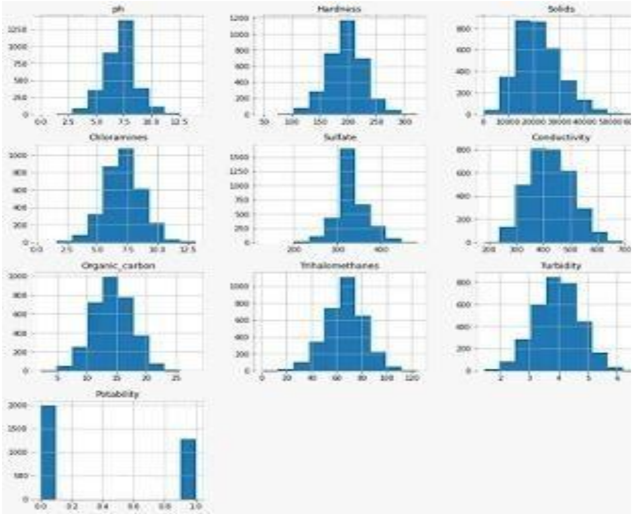



To determine whether the pH value has a normal distribution or not, depict it using the distplot function. Since it is a normal distribution, you can observe that.

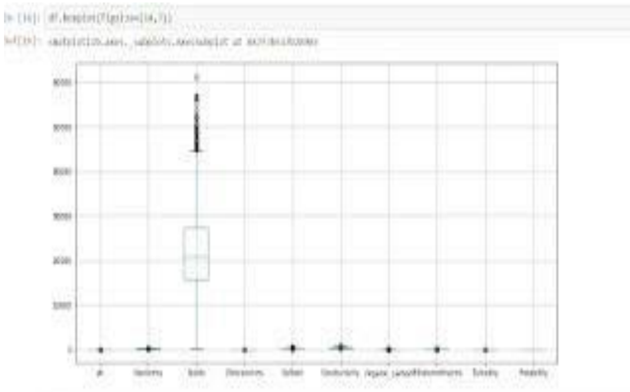
```
]: sns.distplot(df['ph'])
   plt.show()

   /usr/local/lib/python3.7/dist-packages/seaborn/distributions.py:2557: FutureWarning: `distplot` is a deprecated function and wi
   ll be removed in a future version. Please adapt your code to use either `displot` (a figure-level function with similar flexibi
   lity) or `histplot` (an axes-level function for histograms).
     warnings.warn(msg, FutureWarning)
```




Visualize every aspect of the data set as shown below.




Now use a boxplot function to view the outlier. You can see that the Solid feature has outliers, but we are unable to eliminate them because doing so would make the Solid feature unusable. Water will therefore always be safe to drink. We will know whether the water is safe or not since it contains an anomaly that makes the water unclean. Water may be dangerous to drink if the solid content is high.



The data set needs to be prepared now. Separate the features that are independent and dependent from the data. Except for Potability, which is our dependent characteristic, they are all independent features.

Using the train test split function, which yields four data sets, divide the data set into the training and testing sets.

The decision tree classifier model will now be defined, and the data set (X train, Y train) will be used to train the model.

Utilizing the test data set (Xtest,YTest), we further test the model .

International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 09 Issue: 12 | Dec 2022 www.irjet.net p-ISSN: 2395-0072

```
In [18]: X = df.drop('Potability',axis=1)
         Y= df['Potability']
```

```
In [19]: from sklearn.model_selection import train_test_split
         X_train, X_test, Y_train, Y_test = train_test_split(X,Y, test_size= 0.2, random_state=101,shuffle=True)
```

It's time to assess the model using the classification report, confusion matrix, and accuracy score. The actual data and the expected data are the two parameters used in evaluationmethodologies. And you can see that 59% of the time is accurate overall.

**Train Decision Tree Classifier and check accuracy**

```
In [24]: from sklearn.tree import DecisionTreeClassifier
         from sklearn.metrics import accuracy_score,confusion_matrix,classification_report
         dt=DecisionTreeClassifier(criterion= 'gini', min_samples_split= 10, splitter= 'best')
         dt.fit(X_train,Y_train)

Out[24]: DecisionTreeClassifier(ccp_alpha=0.0, class_weight=None, criterion='gini',
                                max_depth=None, max_features=None, max_leaf_nodes=None,
                                min_impurity_decrease=0.0, min_impurity_split=None,
                                min_samples_leaf=1, min_samples_split=10,
                                min_weight_fraction_leaf=0.0, presort='deprecated',
                                random_state=None, splitter='best')

In [25]: prediction=dt.predict(X_test)
         print(f"Accuracy Score = {accuracy_score(Y_test,prediction)*100}")
         print(f"Confusion Matrix =\n {confusion_matrix(Y_test,prediction)}")
         print(f"Classification Report =\n {classification_report(Y_test,prediction)}")

         Accuracy Score = 59.29878048780488
         Confusion Matrix =
          [[274 128]
          [139 115]]
         Classification Report =
                       precision    recall  f1-score   support

                    0       0.66      0.68      0.67       402
                    1       0.47      0.45      0.46       254

             accuracy                           0.59       656
            macro avg       0.57      0.57      0.57       656
         weighted avg       0.59      0.59      0.59       656
```

The model is then tested using a unique data set, and the results are shown in the graphic below.

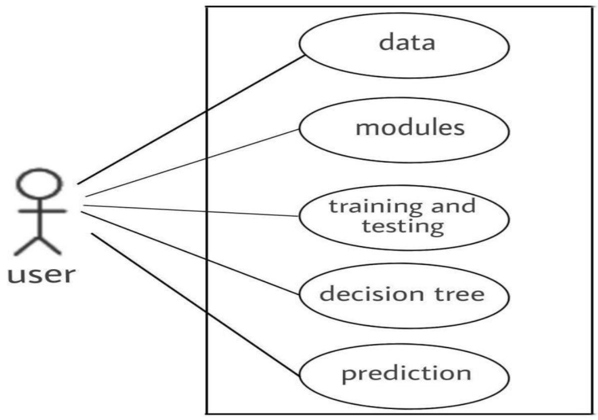
```
[26]: res = dt.predict([[5.735724, 158.318741,25363.016594,7.728601,377.543291,568.304671,13.626624,75.952337,4.732954]])[0]
      res

t[26]: 1
```

The model is then tested based on a given set of values to predict the potability of water. As per the given values, the model predicts the water to be fit for drinking.

## VI. UML DIAGRAMS

The dynamic behavior of a system is represented by a use case diagram. It incorporates use cases, actors, and their interactions to encapsulate the functionality of the system. It simulates the duties, services, and operations needed by asystem or application subsystem. It shows a system's high- level functionality and also describes how a user interacts with a system.

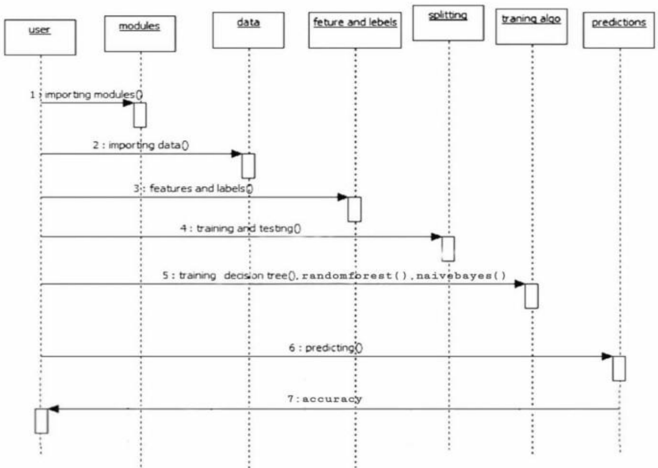
Use-case diagram




The sequence diagram, which is also known as an event diagram, shows how messages move through the system.

It aids in creating a variety of dynamic settings. It depicts communication between any two lifelines as chronologicallyordered series of activities, implying that these lifelines were active at the moment of communication

Sequence diagram




## VII. RESULTS AND ANALYSIS

This research investigated how well machine learning approaches predicted the water quality elements of a water quality dataset. For this, the most well-known dataset variables, including conductivity, ph, tcm, nitrate, and organic salts, were acquired. The results showed that the implemented decision tree model performs well in predicting the parameters of water quality, with an accuracy of 59%. To increase the effectiveness of the selection process, additionalresearch will be conducted to create models that incorporatethe suggested method with other methods and deep learningapproaches.

## VIII. ADVANTAGES OF THE SYSTEM

A decision tree has the important benefit of requiring the consideration of all potential outcomes and tracing each path to a conclusion. It generates a thorough analysis of the outcomes along each branch and pinpoints decision points that require additional research.

They provide each problem, option, and result a particular value. Costs and advantages are made clear when expressedin monetary terms. This method reveals the financial repercussions of various courses of action, lowers confusion, eliminates ambiguity, and highlights the pertinent decision paths. They also employ probability for circumstances to put choices in perspective with one another for straightforward comparisons when factual information is unavailable.

## IX. CONCLUSION

We are all aware of how vital water is to human health. Knowing the water's quality is crucial because if we consume water without first making sure it is safe to do so, we run the risk of getting sick. Numerous illnesses that are transmitted through water exist and if we consume non-drinkable water, we risk contracting hazardous diseases. Consequently, the most crucial factor is understanding the water's quality. But this is where the real issue is. We must test the water at a lab, which is expensive and time-consuming in addition to being necessary for determining the water's quality. In this study, we therefore provide a different strategy for predicting waterquality using artificial intelligence.

## X. FUTURE ENHANCEMENT

Decision trees' relative instability in comparison to other decision predictors is one of their drawbacks. A minor change in the data can have a significant impact on the decision tree's structure, which can express a different outcome than what users would receive in a typical event.Hence better prediction models can replace this algorithmfor more robust result

## XI. ACKNOWLDEMENT

We would like to thank Mr. P Ramu and M Satyanarayana for their wise advice, wise counsel, and moral support throughout the writing of this work.

## XII. REFERENCES

[1] Jayalakshmi, T.; Santhakumaran, A. Statistical normalization and back propagation for classification. Int. J.Comput. Theory Eng. 2011.

[2] Park, J.; Kim, K.T.; Lee, W.H. Recent advances in information and communications technology (ICT) and sensor technology for monitoring water quality.

[3] Kangabam, R.D.; Bhoominathan, S.D.; Kanagaraj, S.; Govindaraju, M. Development of a water quality

[4] The Environmental and Protection Agency, "Parameters of water quality," Environ.Prot., p. 133, 2001.

[5] Kangabam, R.D.; Bhoominathan, S.D.; Kanagaraj, S.; Govindaraju, M. Development of a water quality.

[6] Jiang, J.; Tang, S.; Han, D.; Fu, G.; Solomatine, D.; Zheng, Y. A comprehensive review on the design and optimization of surface water quality monitoring

[7] Manish Kumar Jha,Rajni Kumari Sah, M.S. Rashmitha, Rupam Sinha, B. Sujatha. Smart Water Monitoring System for Real-Time Water Quality and Usage Monitoring,2018

[8] Ashwini K, D. Diviya, J.Janice Vedha, M. Deva Priya. Intelligent Model For Predicting Water Quality

[9] Priya Singh,Pankaj Deep Kaur. Review on Data Mining Techniques for Prediction of Water Quality,2017

[10] Hadi Mohammed, Ibrahim A. Hameed, Razak Seidu. Machine Learning: Based Detection of Water Contamination in Water Distribution systems,2018.