Vivekanand Education Society's

Institute of Technology

(Affiliated to University of Mumbai, Approved by AICTE & Recognized by Govt. of Maharashtra)

Water Quality Predictor and Usage Suggestor

Submitted in partial fulfillment of the requirements

of the degree of

Bachelor of Engineering in

Artificial Intelligence and Data Science

by

Parth Suryavanshi

Shreya Singh

Shruti Devlekar

Yash Sarang

Under the guidance of

Dr. Anjali Yeole

Name of Guide

Department of Artificial Intelligence and Data Science

Vivekanand Education Society's Institute of Technology

2022-2023

Vivekanand Education Society's

Institute of Technology

(Affiliated to University of Mumbai, Approved by AICTE & Recognized by Govt. of Maharashtra)

Department of Artificial Intelligence and Data Science

CERTIFICATE

This is to certify that **Mr/Ms** _____ of Second Year of Artificial Intelligence and Data Science studying under the University of Mumbai have satisfactorily presented the Mini Project entitled **Water Quality Predictor and Usage Suggestor** as a part of the MINI-PROJECT for Semester-V under the guidance of **Dr. Anjali Yeole** in the year 2022-2023.

Date:

(Name and sign)
Head of Department

(Name and sign)
Supervisor/Guide

Vivekanand Education Society's

Institute of Technology

(Affiliated to University of Mumbai, Approved by AICTE & Recognized by Govt. of Maharashtra)

Department of Artificial Intelligence and Data Science

DECLARATION

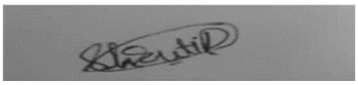
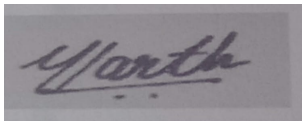
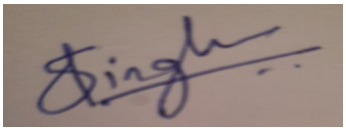
We, ***Parth Suryavanshi, Shreya Singh, Yash Sarang, Shruti Devlekar*** from ***D11AD***, declare that this project represents our ideas in our own words without plagiarism and wherever others' ideas or words have been included, we have adequately cited and referenced the original sources.

We also declare that we have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in our project work.

We declare that we have maintained a minimum 75% attendance, as per the University of Mumbai norms.

We understand that any violation of the above will be cause for disciplinary action by the Institute.

Yours Faithfully

1. ***Shruti Devlekar.*** 
2. ***Parth Suryavanshi.*** 
3. ***Shreya Singh.*** 
4. ***Yash Sarang.*** 

Date: 28/10/2022

Acknowledgement

We are thankful to our college **Vivekanand Education Society's Institute of Technology** for considering our project **Water Quality Predictor and Usage Suggestor** and extending help at all stages needed during our work of collecting information regarding the project.

It gives us immense pleasure to express our deep and sincere gratitude to our Guide, **Dr. Anjali Yeole** for her kind help and valuable advice during the development of the project synopsis and for her guidance and suggestions.

We are deeply indebted to the Vice-Principal and Head of the Artificial Intelligence and Data Science Department **Dr. M. Vijayalakshmi** and our Principal **Dr. J.M. Nair** for giving us this valuable opportunity to do this project.

We express our hearty gratitude to them for their assistance without which it would have been difficult in finishing this project synopsis and project review successfully.

We convey our deep sense of gratitude to all teaching and non-teaching staff for their constant encouragement, support, and selfless help throughout the project work. It is a great pleasure to acknowledge the help and suggestion, which we received from the Department of Artificial Intelligence and Data Science.

Table of Contents

- Abstract 6

1. Introduction 7
   1.1 Introduction
   1.2 Problem Statement 8
   1.3 Objectives
   1.4. Scope

2. Literature Survey 9
   2.1 Literature/Techniques studied
   2.2 Papers/Findings 10

3. Dataset 11
   3.1 Description of the dataset
   3.2 Data collection methodology
   3.3 Exploratory Data Analysis 12
   3.4 Feature extraction 16

4. Analysis and Design 20
   4.1 Analysis of the system
   4.2 Proposed Solutions
   4.3 Prototype design of the proposed system 21

5. Results and Discussion 22

6. Conclusion and Future Work 22

- References 22

Abstract

Water quality becomes one of the important quality factors for the quality of life in smart cities. Recently, water quality has been degraded due to diverse forms of pollution caused by disposal of human wastes, industrial wastes, automobile wastes. The increasing pollution affects water quality and the quality of people's life. Hence, water quality evaluation, monitoring, and prediction become an important and hot research subject. In the past, many environmental researchers have dedicated their research efforts on this subject using conventional approaches. Recently, many researchers began to use the big data analytics approach to studying, evaluating, and predicting water quality due to the advances of big data applications. Following their example, we build a model that predicts the quality of water using its characteristics.

1. Introduction

1.1 Introduction

Water is one of the most essential natural resources for the existence and survival of the entire life on this planet. We use water for drinking, cooking, personal hygiene, agricultural practices, and recreational purposes almost every day.

However, the increasing population, its automobiles and industries are polluting all the water-bodies at an alarming rate. The effluents added by these pollution sources affect the pH value of water. This gives rise to many water-related problems such as water-borne diseases in organisms and deaths of aquatic animals like fish, crab, and so on. This pollution eventually disrupts the food-chain and damages the ecosystem in the long run.

Hence, water pollution is one of the most alarming concerns for us today. Addressing this concern, people have spent lots of research efforts in water quality evaluation and monitoring. In the past decades, many researchers have spent lots of time on studying and developing different models and methods in water quality analysis and evaluation. In recent times, many researchers have developed or used big data analytics models and machine learning based models to conduct water quality evaluation in order to achieve better accuracy in evaluation and perdition.

This project also plans to deal with this problem by preparing a Model to predict the quality of water using different machine learning algorithms like decision tree, artificial neural networks and improved decision tree. It also tells us the various use cases of the water based on its quality.

1.2 Problem Statement

Access to safe drinking-water is essential to health, a basic human right and a component of effective policy for health protection. This is important as a health and development issue at a national, regional and local level. In some regions, it has been shown that investments in water supply and sanitation can yield a net economic benefit, since the reductions in adverse health effects and health care costs outweigh the costs of undertaking the interventions.

The project aims to prepare a Model to predict the quality of water using factors like hardness, pH value, dissolved solids, chloramines, sulfate, carbon, etc and tell us whether the water is safe to drink or not.

Integrating this model in a Mobile Application.

1.3 Objectives

This project is made with the objective to achieve the following sustainable development goals :

- Clean Water and Sanitation - This model can predict the water quality of a water body, thus determining whether the water is safe to drink or not.
- Good Health and Well Being - Since this model determines the safety of water, it will lead to less people drinking unsafe, contaminated water.

1.4 Scope

This application has a wide scope of applications. It can be used by a citizen to check the quality of water coming to his home. It can also be useful to the government officials who can monitor the water quality in their respective regions. This can further be of use to environment enthusiasts and researchers who can track the water quality all over India.

2. Literature Survey

2.1 Literature/Techniques studied

Water quality model : Water quality models can be applied to many different types of water systems, including streams, rivers, lakes, reservoirs, estuaries, coastal waters and oceans. These models describe the main water quality processes and typically require the hydrological and constituent inputs.

Water quality standards : The first step in water quality evaluation is to decide quality standards for a water body based on the desired uses of the water body (lake, stream, estuary etc). The quality standards will be different based on the use cases.

Big data water quality analysis : These models use "Big Data" to optimally model water systems. The systems that are considered are highly dynamic, spatially expansive, and behaviorally heterogeneous. Nowadays, big data solutions have become efficient and receive more attention. Key to making big data actionable is harnessing, standardizing, and integrating the enormous amount of data.

Technology Stack :
ML Models -
SKLearn and
additional Python libraries. (pandas, matplotlib, seaborn)

Web Application:

JavaScript, Node.js, React, Express

Different Machine Learning Algorithms :

1. Decision Tree Model - Water quality prediction is done using the Decision Tree Model. It is a supervised learning technique which uses a predictive model to map observations about an item (represented in the branches) to conclusions about the item's target value (represented in the leaves).

2. Artificial Neural Networks - Artificial Neural Networks model tries to simulate the structures and networks within the human brain. The architecture of neural networks consists of nodes which generate a signal or remain silent as per a sigmoid activation function in most cases. The ANN model has the ability to efficiently describe the non-linear relationship of the complex water quality datasets.

3. Improved Decision Tree Model - This model combines artificial neural networks and decision tree algorithms. The main advantage of this approach lies in the clustering of data when processing the data with stress on inter-dependency between parameters and an aim at reducing rough disadvantages at the same time. Therefore, more accurate predictions can be made by using this method

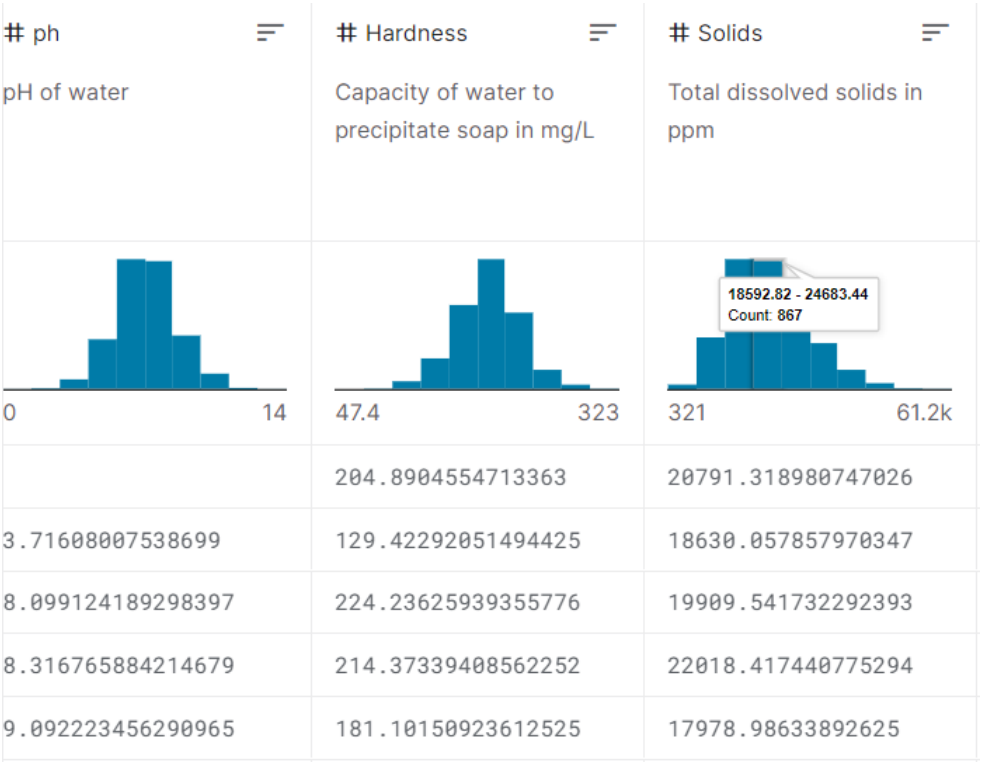
2.2 Papers/Findings

During the literature survey, we learned the basis of the water quality model, water quality standards, big data water quality analysis and different Machine Learning algorithms including Decision Tree Model, Artificial Neural Networks and Improved Decision Tree Model. We also gained insights on the categorical approach to Exploratory Data Analysis.

3. Data Set

3.1 Description of data set

Our dataset would include the water properties of
1. pH value
2. Hardness
3. Solids (TDS-Total dissolved solids)
4. Chloramines
5. Sulfate
6. Conductivity
7. Organic_carbon
8. Trihalomethanes
9. Turbidity
10. Potability



	204.8904554713363	20791.318980747026
3.71608007538699	129.42292051494425	18630.057857970347
8.099124189298397	224.23625939355776	19909.541732292393
8.316765884214679	214.37339408562252	22018.417440775294
9.092223456290965	181.10150923612525	17978.98633892625

3.2 Data collection methodology

Data is collected from various sources. Most of which is from the Central Pollution Control Board which has public datasets about water.

3.3 Exploratory data analysis

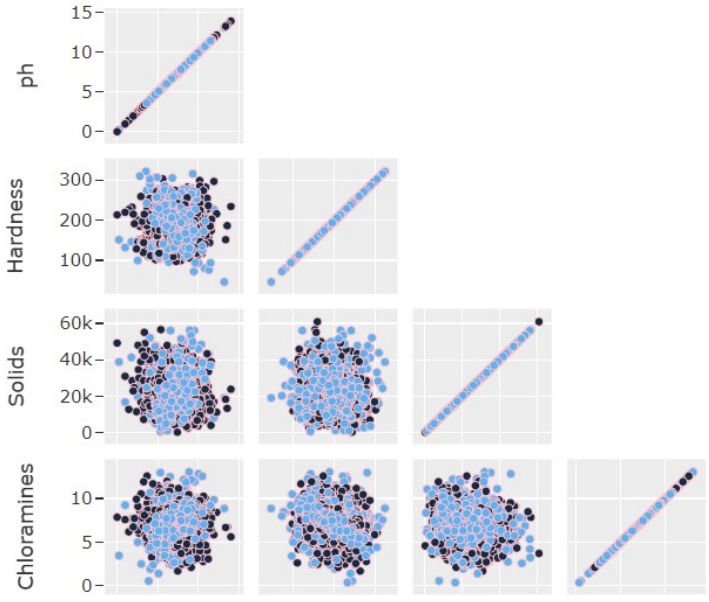
Libraries used for EDA : Matplotlib.pyplot, Seaborn and plotly.offline.

Steps performed :

1. Scaling down for distribution plots
2. Scatter Matrix
3. Correlation Matrix
4. Checking for Missing values
5. Distribution Plots (with Histograms)
6. Feature extraction

1. Scaling down for distribution plots : We observed that the data for different features are widely ranged and their distributions cannot be compared on a single scale. So we will be Scaling down the data (Standardizing) so that the distributions can be compared. Some features are in the range of 100s some in 10ks. This operation will bring down the scale in the range of ~ -3 to +3.

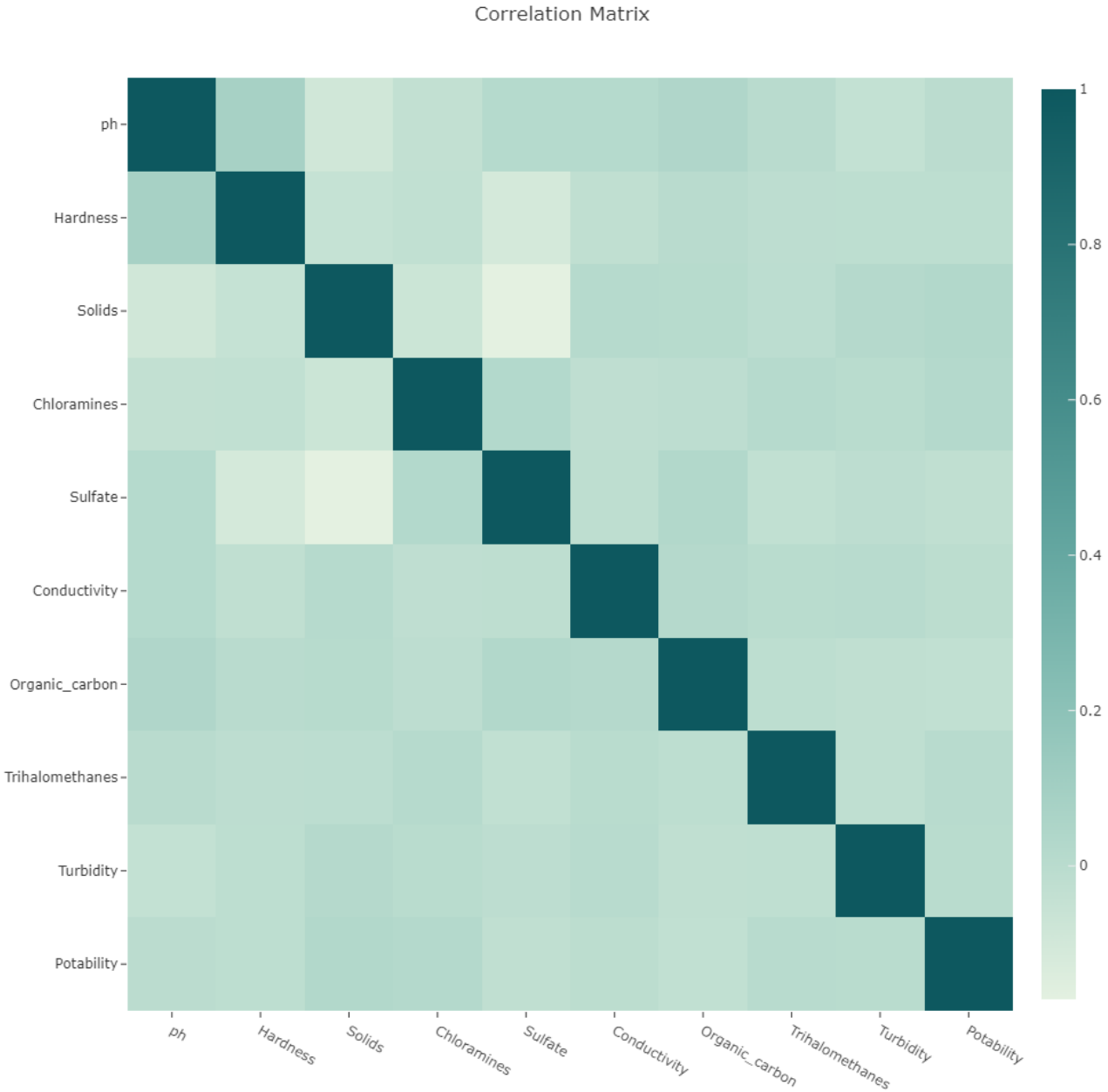
2. Scatter Matrix : To see the correlation between the features we will plot a Scatter matrix. This matrix consists of several graphs (all scatter plots) taking any 2 features as the axes. We can observe the behavior of the features and how it affects the result (Potability).



From the scatter plot it looks like there is no visible relation between the features. The data points are mostly randomly scattered in space. This might be the result of noisy data, or data that isn't much correlated with the result. This might lead to bad performing models.

3. Correlation Matrix :

From the matrix, it looks like there is no correlation. So the features individually do not affect the result as much.



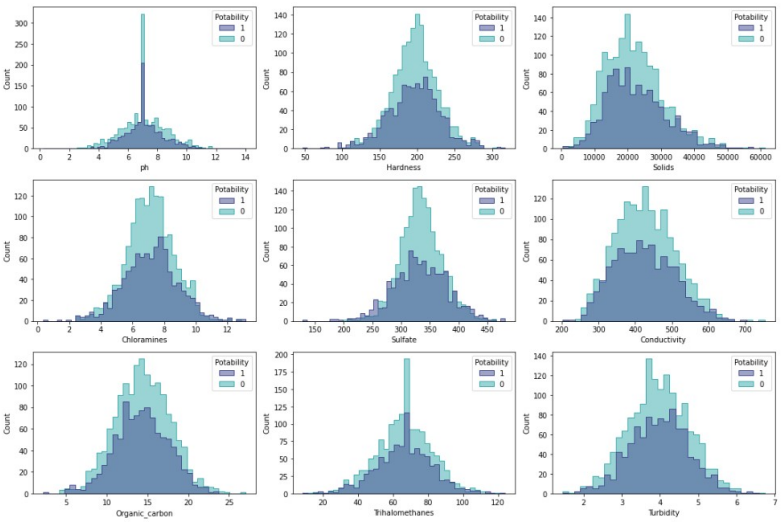
4. Checking for Missing Values :

Now we check for the missing values using Heatmap.

So there are a significant number of missing values for **Trihalomethanes**, **Sulfate** and **pH**. These missing values can be filled using the k-nearest neighbors method.

5. Distribution plots (with histograms) :

We plot the distribution graphs for all the features using histogram.



The above histogram shows the distribution according to the potability.

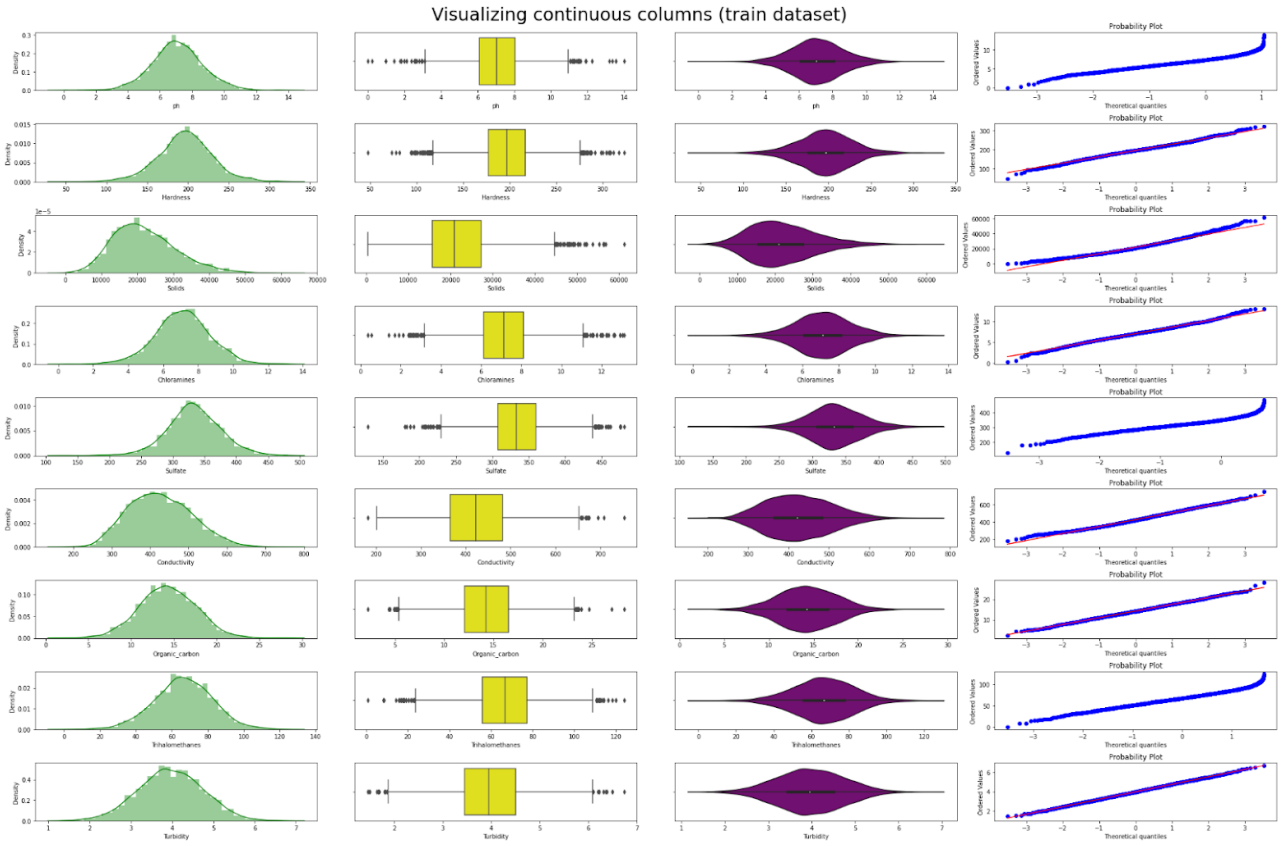
3.2 Feature extraction

First we check the dataset for categorical features

```
Feature 'ph' has '2786' unique values
Feature 'Hardness' has '3276' unique values
Feature 'Solids' has '3276' unique values
Feature 'Chloramines' has '3276' unique values
Feature 'Sulfate' has '2496' unique values
Feature 'Conductivity' has '3276' unique values
Feature 'Organic_carbon' has '3276' unique values
Feature 'Trihalomethanes' has '3115' unique values
Feature 'Turbidity' has '3276' unique values
Feature 'Potability' has '2' unique values
```

We observe that one feature (Potability) is categorical while the rest are continuous numeric.

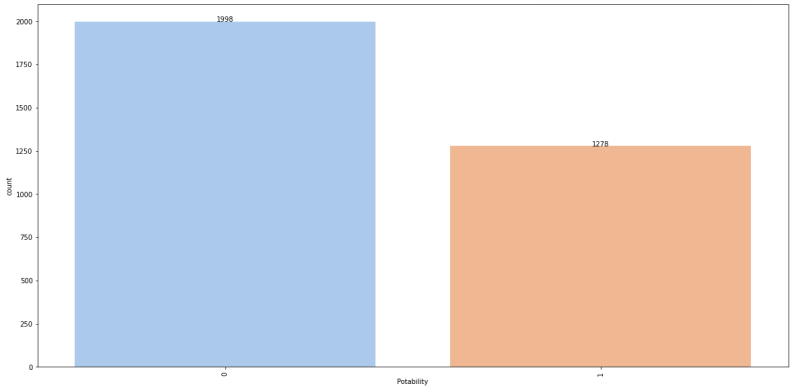
Visualizing the continuous features :



After analyzing these graphs, the following hypotheses are made:

1. Most features are distributed according to the normal distribution law.
2. There are minor outliers for some features.

Plotting the distribution of the target feature :



The plot shows that there is a clear class imbalance which needs to be eliminated we do so by creating artificial duplicates to eliminate this effect

```
from imblearn.over_sampling import SMOTE
```

```
oversample = SMOTE()
features, labels=  oversample.fit_resample(train2.drop(["Potability"],axis=1),train2["Potability"])
```

STATISTICAL TESTS

We then performed some statistical tests to check if the features are normally distributed.

```
Statistics=0.989, p=0.000
Sample does not look Gaussian with ph (reject H0)
Statistics=0.996, p=0.000
Sample does not look Gaussian with Hardness (reject H0)
Statistics=0.978, p=0.000
Sample does not look Gaussian with Solids (reject H0)
Statistics=0.997, p=0.000
Sample does not look Gaussian with Chloramines (reject H0)
Statistics=0.984, p=0.000
Sample does not look Gaussian with Sulfate (reject H0)
Statistics=0.993, p=0.000
Sample does not look Gaussian with Conductivity (reject H0)
Statistics=1.000, p=0.620
Sample looks Gaussian with Organic_carbon (fail to reject H0)
Statistics=0.998, p=0.000
Sample does not look Gaussian with Trihalomethanes (reject H0)
Statistics=1.000, p=0.931
Sample looks Gaussian with Turbidity (fail to reject H0)
```

Since most of the features were not distributed according to the normal distribution law, we used the Mann-Whitney U-test.

```
Statistics=6111.000, p=0.000
The sample distributions are not equal (reject H0)
Statistics=0.000, p=0.000
The sample distributions are not equal (reject H0)
Statistics=0.000, p=0.000
The sample distributions are not equal (reject H0)
Statistics=2556.000, p=0.000
The sample distributions are not equal (reject H0)
Statistics=0.000, p=0.000
The sample distributions are not equal (reject H0)
Statistics=0.000, p=0.000
The sample distributions are not equal (reject H0)
Statistics=0.000, p=0.000
The sample distributions are not equal (reject H0)
Statistics=1278.000, p=0.000
The sample distributions are not equal (reject H0)
Statistics=0.000, p=0.000
The sample distributions are not equal (reject H0)
```

Since the sample distributions did not turn out to be equal, we reject the null hypothesis and decided that our final model will include all the features. We also removed outliers using sklearn's LocalOutlierFactor.

To increase the accuracy of forecasting, we normalized the data between [0,1]

Finally, data looks like :

	ph	Hardness	Solids	Chloramines	Sulfate	Conductivity	Organic_carbon	Trihalomethanes	Turbidity
0	0.529808	0.526368	0.362007	0.543891	0.680385	0.762704	0.313402	0.699753	0.286091
1	0.265434	0.224053	0.323786	0.491839	0.604478	0.824275	0.497319	0.450999	0.576793
2	0.578509	0.603866	0.346413	0.698543	0.568806	0.448774	0.562017	0.532866	0.303637
3	0.594055	0.564356	0.383707	0.603314	0.647347	0.329540	0.622089	0.808065	0.601015
4	0.649445	0.431072	0.312272	0.484900	0.514545	0.405261	0.358555	0.253606	0.496327
...	...	...	...	...	...	...	...	...	...
3933	0.685406	0.414368	0.873724	0.529319	0.546504	0.334393	0.501605	0.745095	0.447228
3934	0.428018	0.441784	0.835895	0.584604	0.531356	0.715047	0.445523	0.486784	0.464903
3935	0.528777	0.506414	0.398796	0.531529	0.507246	0.350857	0.456512	0.551890	0.409597
3936	0.206598	0.937462	0.470568	0.525196	0.499752	0.671474	0.461157	0.337713	0.423405
3937	0.613057	0.556488	0.216796	0.643770	0.609699	0.362815	0.435106	0.508991	0.406724

3938 rows × 9 columns

4. Analysis and Design

4.1 Analysis of the system

As seen from the system architecture (on later pages), the data comes from various sources - lakes, ponds, rivers, canals, sea water and drains and groundwater. This data is stored in the database which further undergoes EDA and feature extraction. After the data is ready, it is fed to the ML model which uses the Improved decision tree algorithm to predict the output (water quality).

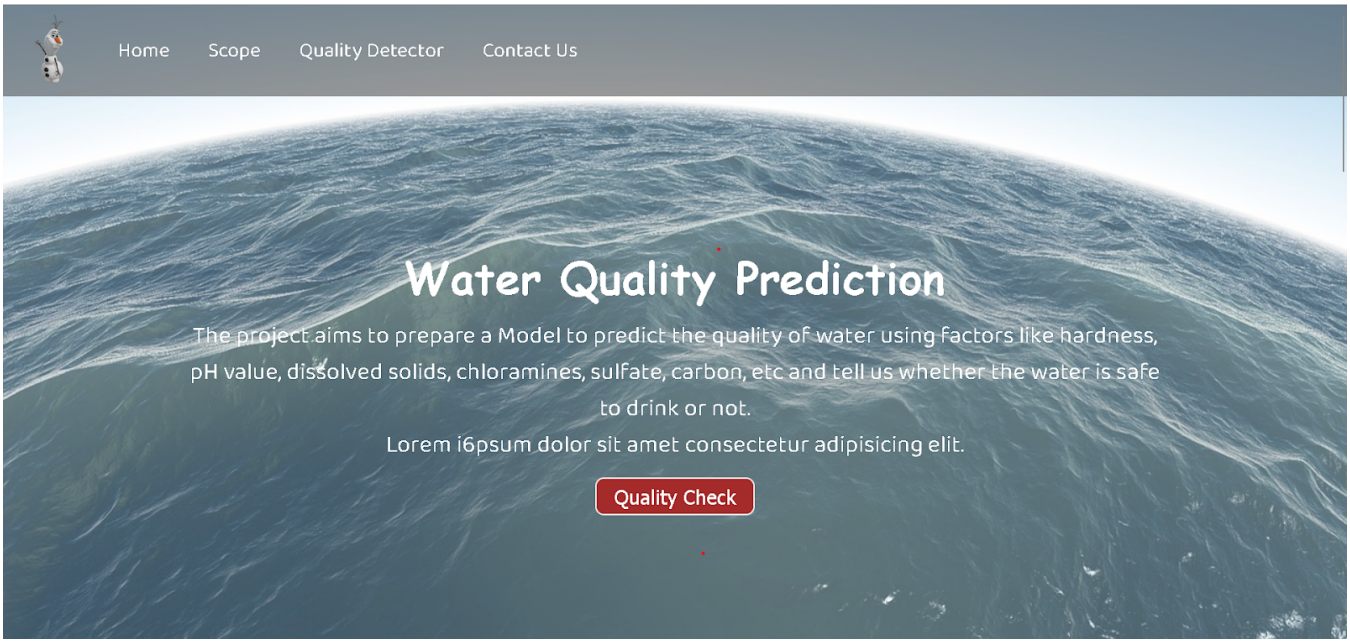
The database is also connected to the web application. So, the user inputs the data according to the water source and location. This data is then worked upon by the ML model which produces the predicted output. This output is then shown in the web application to the user.

4.2 Proposed Solutions

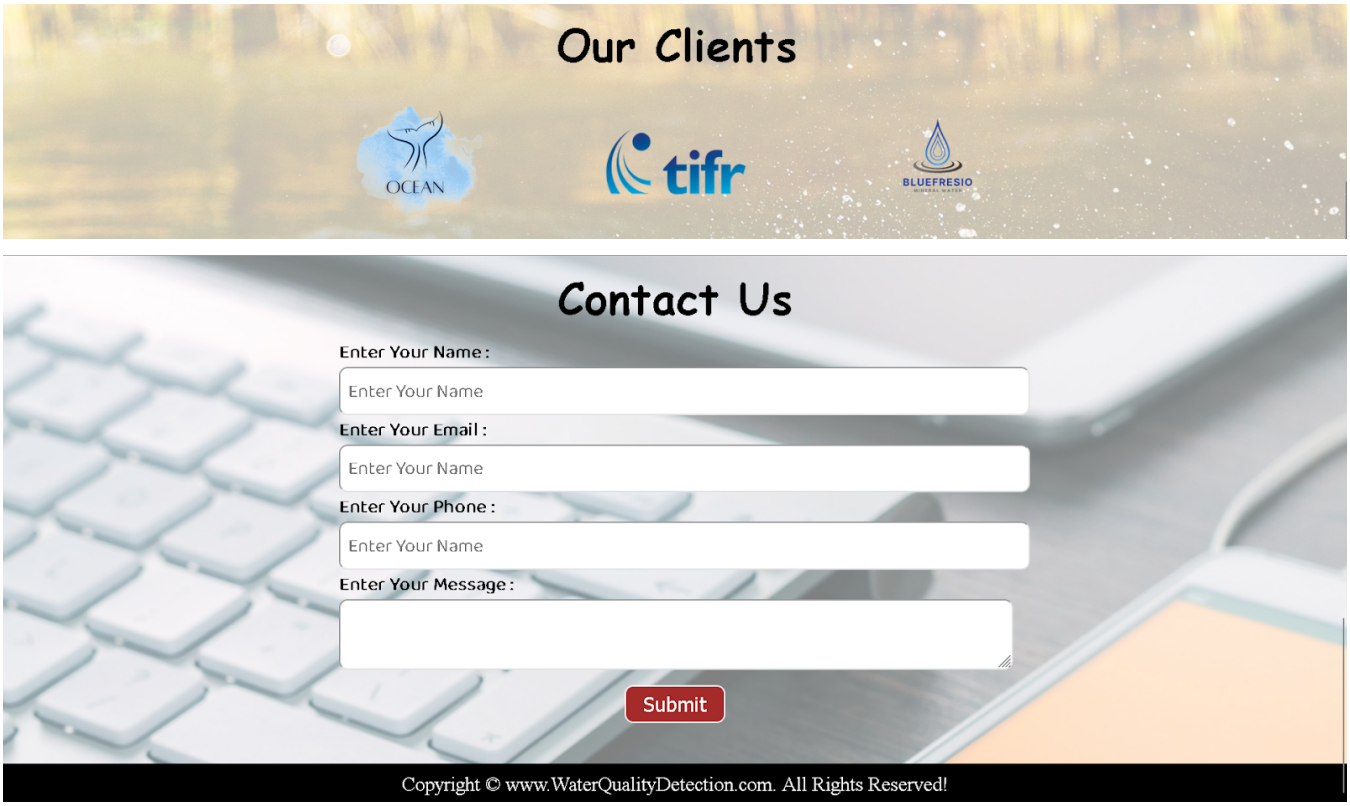
The project aims to prepare a Model to predict the quality of water using factors like hardness, pH value, dissolved solids, chloramines, sulfate, carbon, etc and tell us whether the water is safe to drink or not. This model would be then integrated in a web application using React.

Consumers like the Government officials, Farmers, Private landowners, Industry specialists would be availed of an option to upload the data of the water they have received through various sources. The prediction model would then predict the quality of the water and also its suitability through a user friendly interface.

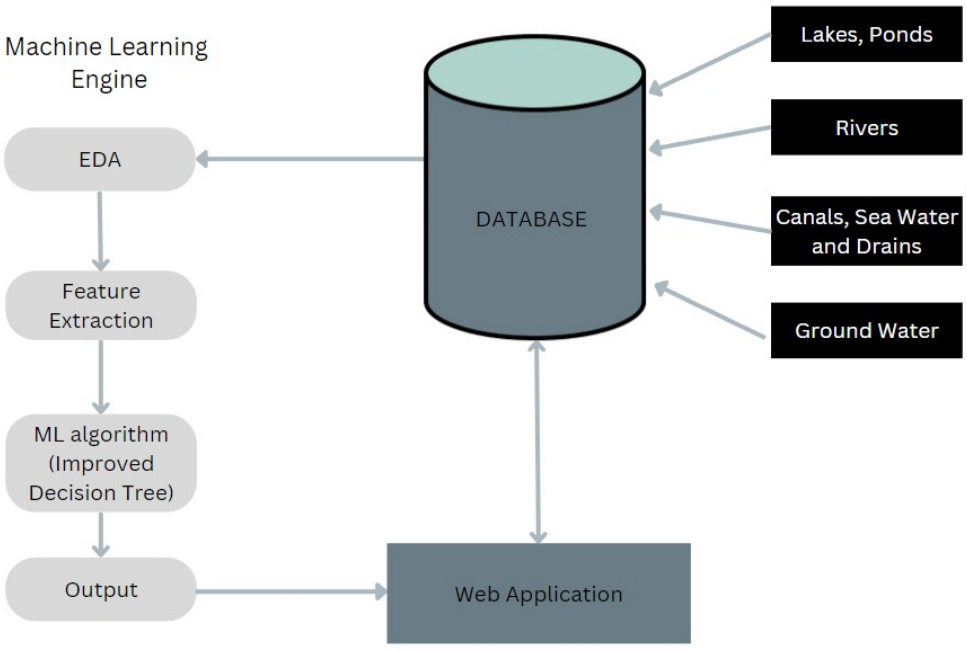
4.3 Prototype design of the proposed system


Home
Scope
Quality Detector
Contact Us
Water Quality Prediction
The project aims to prepare a Model to predict the quality of water using factors like hardness, pH value, dissolved solids, chloramines, sulfate, carbon, etc and tell us whether the water is safe to drink or not.
Lorem i6psum dolor sit amet consectetur adipisicing elit.
Quality Check


Home
Scope
Quality Detector
Contact Us
Scope
For The Citizens
Lorem, ipsum dolor sit amet consectetur adipisicing elit. Unde, porro minus, quam cum magnam ut quidem beatae autem amet eius natus dolorum
Government Officials
Lorem, ipsum dolor sit amet consectetur adipisicing elit. Unde, porro minus, quam cum magnam ut quidem beatae autem amet eius natus
Researchers
Lorem, ipsum dolor sit amet consectetur adipisicing elit. Unde, porro minus, quam cum magnam ut quidem beatae autem amet eius natus



System Architecture :



5. Results and Discussion

After thorough research and discussion, we have arrived at the decision to make a web based application that will predict the quality of water based on the user's input data. This data will be matched with the data on the database and then will be fed to the model which will then predict the quality of water. The water quality is evaluated by its portability.

ML algorithm to be used : We explored three different ML algorithms namely, Decision Tree, Artificial Neural Networks and Improved Decision Tree. We came to the conclusion of using Improved Decision Tree because of the following reasons

- Transparent : gives a clear understanding of which decision is better.
- Efficient : Requires little time to create.
- Flexible : New ideas or factors can be easily added later.
- It is the combination of Decision Tree and Artificial Neural Networks. So theoretically, it will have more depth and accuracy.

6. Conclusion and Future Work

The future work would consist of developing a real time Machine learning model as discussed above with optimal accuracy and integrating it with a web application created from scratch in React and other supporting technical stack.

References

1. https://ieeexplore.ieee.org/abstract/document/7944943
2. https://ieeexplore.ieee.org/document/9497111
3. Data-driven Water Quality Analysis and Prediction: A Survey
4. Proposition of New Ensemble Data-Intelligence Models for Surface Water Quality Prediction
5. Towards data Science - Exploratory Data Analysis

*****