CHAITANYA BHARATHI INSTITUTE OF TECHNOLOGY
An Autonomous Institute | Affiliated to Osmania University
Kokapet Village, Gandipet Mandal, Hyderabad, Telangana-500075, www.cbit.ac.in

COMMITTED TO RESEARCH, INNOVATION AND EDUCATION 45 years

Department of Artificial Intelligence and Machine Learning

Mini Project

Project Title : WATER QUALITY ANALYSIS USING ML

Guided By:

Sri. M. Vishnu Chaitanya

Assistant Professor

Presented By:

1601-22-748-026 K. Rajesh

1601-22-748-027 K. Tharun Kumar Reddy

1601-22-748-031 K. Karthik

ABSTRACT

A water quality prediction was generated for predicting if the water is safe to drink or not. This experiment was also conducted to compare the machine learning model performance between Decision Tree, Random Forest, and Logistic Regression to determine the most suitable technique for predicting Water Quality.

Water is the most crucial resource of life and it is necessary for the survival of all living creatures including human beings. The survival of business and agriculture depends on freshwater. An essential step in managing freshwater assets is the evaluation of the quality of the water. Before using water for anything, including drinking, chemical spraying (pesticides, etc.), or animal hydration, it is crucial to assess its purity. The ecosystem and the general public's health are directly impacted by water quality. Therefore, analysing and predicting water quality is necessary for both environmental and human protection. Machine learning can be used to analyse and predict the water quality based on the parameters like PH value, turbidity, hardness, conductivity, dissolved solids in water and other parameters. In this work, the water quality is predicted by giving the concentration of various parameters as input to machine learning algorithms and the water is classified as safe or unsafe for the usage of domestic purposes

Base Paper

Akshay, R., Tarun, G., Kiran, P. U., Devi, K. D., & Vidhyalakshmi, M. (2022, December). Water-Quality-Analysis using Machine Learning. In *2022 11th International Conference on System Modeling & Advancement in Research Trends (SMART)* (pp. 13-18). IEEE.

Objective

1.Develop and Implement Machine Learning Models: To create and deploy various machine learning models tailored for analyzing water quality data, aiming to improve predictive accuracy and data interpretation.
2.Evaluate Model Performance: To assess the effectiveness and performance of different ML algorithms in predicting water quality parameters, comparing their accuracy, precision, and computational efficiency.
3. Optimize Data Processing Techniques: To refine data preprocessing and feature extraction methods to enhance the quality of input data for ML models, ensuring better model training and prediction outcomes.
4.Integrate Real-Time Data Analysis: To implement ML techniques that enable real-time water quality monitoring and analysis, facilitating prompt detection of anomalies and enabling timely interventions.

5.Enhance Decision-Making Processes: To leverage ML-generated insights to support and improve decision-making processes in water quality management, contributing to better environmental and public health outcomes.

INTRODUCTION

In the scientific field of machine learning, it is investigated how computers learn via experience. Since the capacity to learn is the fundamental quality of an entity regarded as intelligent in the broadest meaning of the word, the words "Machine Learning" and "Artificial Intelligence" are frequently used synonymously in the minds of scientists. Building adaptable, experience-based computer systems is the goal of machine learning. It is now possible to discover a solution to this problem because of the development of machine learning methods. We have developed a technique that uses data mining to identify whether the water is portable or not. The enormous amount of data related to water quality can be mined for hidden knowledge. As a result, it now has a more significant role in the study. This research aims to develop a system that can predict water quality more precisely.

PROBLEM STATEMENT

Develop a machine learning model to check the water quality of a sample and to classify water samples as potable or non-potable based on physicochemical properties, including pH, hardness, and conductivity etc. to ensure safe drinking water availability.

The Water Quality Analysis project aims to develop a machine learning model to classify water samples as either potable (safe for drinking) or non-potable (unsafe) based on various physicochemical attributes. The dataset includes features such as pH, hardness, total dissolved solids, chloramines, sulphate, conductivity, and nitrates.

OBJECTIVE OF THE WORK

The objective of the Water Quality Analysis project is to:

1. ***Develop a Predictive Model*:** Build a machine learning model that accurately classifies water samples as either potable or non-potable based on water quality analysis of the physicochemical properties.

2. ***Analyse Water Quality*:** Evaluate the impact of various features, such as pH, hardness, and nitrates, on water safety to understand their significance in determining potability.

3. ***Improve Public Health*:** Provide a tool for monitoring and assessing water quality to help in identifying unsafe water sources, thereby aiding in timely interventions and promoting better water safety practices.

4. ***Support Decision-Making*:** Generate insights and recommendations for water quality management and policy-making to enhance the safety and quality of drinking water.

LITERATURE SURVEY

1. Khan, Y., & See, C. S. (2016, April). Predicting and analysing water quality using machine learning: a comprehensive model. In *2016 IEEE Long Island Systems, Applications and Technology Conference (LISAT)* (pp. 1-6). IEEE.

In the paper, the authors provide a detailed examination of the application of machine learning techniques for water quality prediction. The study reviews existing water quality prediction models and highlights their limitations, such as reliance on linear methods and limited data handling capabilities. The authors discuss various machine learning algorithms like Support Vector Machines (SVM), Decision Trees, and Neural Networks that can handle nonlinear relationships and large datasets. Additionally, the paper emphasizes the importance of feature selection and preprocessing to improve prediction accuracy. The authors propose a comprehensive model that integrates multiple machine learning techniques for real-time water quality monitoring and prediction. They also explore the potential benefits of such models in managing water resources and mitigating pollution risks. Overall, the study underscores the evolving role of machine learning in environmental monitoring.

2. Ahmed, A. N., Othman, F. B., Afan, H. A., Ibrahim, R. K., Fai, C. M., Hossain, M. S., ... & Elshafie, A. (2019). Machine learning methods for better water quality prediction. *Journal of Hydrology, 578*, 124084.

The paper provides a comprehensive review of various machine learning methods applied to predict water quality parameters, such as pH, dissolved oxygen, and pollutant concentrations. It highlights the limitations of traditional modeling techniques, such as regression and physical models, which may struggle with nonlinear and complex data structures. The study discusses several machine learning algorithms, including artificial

neural networks (ANNs), support vector machines (SVMs), decision trees (DTs), and ensemble methods, emphasizing their potential for improving prediction accuracy. The authors also review feature selection methods that enhance model performance by identifying relevant water quality indicators. Additionally, the paper explores hybrid models that combine machine learning with conventional approaches to achieve better generalization. The review emphasizes the need for more comprehensive datasets and proper validation techniques to improve model reliability. Ultimately, the study concludes that machine learning holds promise for effective water quality management, though further research is needed to address challenges like overfitting and interpretability.

3. Vergina, S. A., Kayalvizhi, S., Bhavadharini, R., & Kalpana Devi, S. (2020). A real time water quality monitoring using machine learning algorithm. *Eur. J. Mol. Clin. Med*, *7*(8), 2035-2041.

The paper explores the application of machine learning techniques to monitor and assess water quality in real time. The study highlights the growing need for efficient, automated systems to address water pollution and manage water resources effectively. Traditional water quality monitoring methods, which involve manual sampling and laboratory analysis, are time-consuming and expensive. This paper introduces a machine learning-based system that can predict water quality by analyzing sensor data on parameters such as pH, turbidity, and dissolved oxygen. The authors employ supervised learning techniques to train models capable of real-time classification and prediction. The proposed system enhances decision-making for water quality management and provides a cost-effective, scalable solution for continuous monitoring. Various algorithms are compared to identify the most accurate for water quality prediction.

4. Melesse, A. M., Khosravi, K., Tiefenbacher, J. P., Heddam, S., Kim, S., Mosavi, A., & Pham, B. T. (2020). River water salinity prediction using hybrid machine learning models. *Water*, *12*(10), 2951.

The paper focuses on predicting river water salinity using advanced hybrid machine learning models. The authors highlight the importance of accurately forecasting salinity levels due to their significant impact on water quality, agriculture, and ecosystems. The study integrates multiple machine learning algorithms, including support vector machines, decision trees, and neural networks, to improve prediction accuracy. Various environmental factors, such as water discharge, temperature, and dissolved oxygen, are used as inputs to these models. A key contribution of the research is the development of hybrid models that combine individual machine learning approaches to exploit their strengths and compensate for weaknesses. The performance of these models is compared to traditional methods, demonstrating superior accuracy and efficiency. This paper also discusses the challenges of modeling complex hydrological processes and emphasizes the potential of machine learning in addressing these issues. The findings can inform water resource management and policy-making, particularly in regions facing salinity-related challenges.

5. Kuthe, A., Bhake, C., Bhoyar, V., Yenurkar, A., Khandekar, V., & Gawale, K. (2022). Water quality analysis using machine learning. *International Journal for Research in Applied Science and Engineering Technology*, *10*(12), 581-585.

The paper focuses on utilizing machine learning techniques to evaluate and predict water quality. The literature review within the paper discusses various traditional methods of water quality analysis, such as physical and chemical parameter testing, which are time-consuming and often costly. It highlights recent advances in machine learning as an alternative for accurate and efficient prediction models. The authors examine various algorithms, including Decision Trees, Support Vector Machines (SVM), and Neural Networks, which have been applied to environmental monitoring. Prior research shows that ML techniques can efficiently predict the concentrations of harmful substances and overall water quality, with some studies achieving high accuracy. Additionally, the literature acknowledges the integration of Internet of Things (IoT) devices and sensors for real-time data collection, enhancing machine learning's predictive capabilities. Overall, the survey emphasizes the growing role of ML in environmental data analysis.

6. Akshay, R., Tarun, G., Kiran, P. U., Devi, K. D., & Vidhyalakshmi, M. (2022, December). Water-Quality-Analysis using Machine Learning. In *2022 11th International Conference on System Modeling & Advancement in Research Trends (SMART)* (pp. 13-18). IEEE.

The paper presents an extensive review of machine learning techniques applied to water quality analysis. The authors highlight the importance of assessing water quality to ensure public health and environmental safety. Traditional water quality assessment methods are time-consuming and expensive, making machine learning a viable alternative for efficient, real-time analysis. The paper reviews various machine learning algorithms, including regression, classification, and clustering, used in predicting water quality parameters such as pH, turbidity, and dissolved oxygen. Additionally, it discusses the use of sensor-based data collection and emphasizes the need for large, accurate datasets to train models effectively. The paper also touches upon challenges such as overfitting, data preprocessing, and the complexity of water systems. Finally, the authors conclude by showcasing the potential of machine learning in advancing water quality monitoring systems.

7. Kaddoura, S. (2022). Evaluation of machine learning algorithm on drinking water quality for better sustainability. *Sustainability*, *14*(18), 11478.

In the paper the author explores the application of machine learning techniques in assessing drinking water quality. The study focuses on the potential of machine learning algorithms to enhance water quality monitoring and management systems, aiming for improved sustainability in water resources. By using various datasets on water quality parameters, the research investigates the accuracy and efficiency of different machine learning models in predicting contamination levels and ensuring safe drinking water. The paper highlights the importance of real-time monitoring and predictive analytics in preemptively addressing water pollution issues. It also emphasizes how integrating machine learning can reduce the need for frequent manual testing and help in identifying patterns that are critical for maintaining water safety standards. Ultimately, the research contributes to the development of more sustainable water quality management practices through advanced technological interventions.

8. Poudel, D., Shrestha, D., Bhattarai, S., & Ghimire, A. (2022). Comparison of machine learning algorithms in statistically imputed water potability dataset. *Journal of Innovations in Engineering Education*, *5*(1), 38-46.

The paper focuses on comparing machine learning algorithms for predicting water potability using a statistically imputed dataset. The study addresses the challenge of missing data, which is common in environmental datasets, by applying statistical imputation methods to fill in gaps. Various machine learning algorithms, such as Decision Trees, Random Forest, Support Vector Machines (SVM), and k-Nearest Neighbors (k-NN), were evaluated for their performance in classifying water as potable or non-potable. The authors emphasize the importance of data preprocessing, particularly handling missing data, to improve the accuracy of predictions. Their results indicate that ensemble methods like Random Forest outperform other classifiers. The study provides valuable insights into how different algorithms handle imputed data and suggests that machine learning can significantly enhance water quality monitoring efforts. The comparison offers a framework for selecting the best algorithm based on dataset characteristics and specific project goals.

9. Wang, X., Li, Y., Qiao, Q., Tavares, A., & Liang, Y. (2023). Water quality prediction based on machine learning and comprehensive weighting methods. *Entropy*, *25*(8), 1186.

The paper focuses on improving water quality prediction through the application of machine learning techniques combined with comprehensive weighting methods. The authors explore how different factors affect water quality and propose a predictive model that integrates data-driven approaches with domain expertise. They evaluate various machine learning models, including neural networks and decision trees, and apply multi-criteria decision-making techniques to weigh the influence of different water quality indicators. The study leverages entropy-based methods to enhance prediction accuracy. Experimental results demonstrate the effectiveness of their hybrid approach in improving prediction performance compared to traditional models. This research provides valuable insights into environmental monitoring and the sustainable management of water resources.

10. Patel, S., Shah, K., Vaghela, S., Aglodiya, M., & Bhattad, R. (2023). Water Potability Prediction Using Machine Learning.

The paper explores the use of machine learning models for predicting water potability, focusing on the classification of water as either potable or non-potable based on various water quality parameters. In their literature review, the authors highlight previous works that applied traditional statistical methods for water quality analysis, noting their limitations in handling large datasets and complex relationships. They also reference studies that employed machine learning techniques like decision trees, random forests, and support vector machines (SVM), emphasizing their improved accuracy. Additionally, the authors discuss the growing trend of using deep learning models, although they point out the challenges related to data availability and model interpretability. Lastly, they survey the use of real-time monitoring systems and IoT technologies to enhance data collection, which can be integrated with machine learning algorithms for dynamic water quality prediction.

11. Brindha, D., Puli, V., NVSS, B. K. S., Mittakandala, V. S., & Nanneboina, G. D. (2023, February). Water quality analysis and prediction using machine learning. In *2023 7th International Conference on Computing Methodologies and Communication (ICCMC)* (pp. 175-180). IEEE.

The paper focuses on water quality analysis and prediction using machine learning techniques. The authors highlight the growing concern over water pollution and its detrimental effects on health and the environment. They present a systematic approach to analyze water quality parameters, utilizing various machine learning algorithms to predict water quality levels. The study emphasizes the importance of timely and accurate monitoring of water quality for effective management. Additionally, the authors discuss the dataset used for training and validation, which includes parameters like pH, turbidity, and chemical concentrations. Results indicate that machine learning models significantly enhance prediction accuracy compared to traditional methods. The paper also suggests potential applications of these techniques in environmental monitoring and policy-making. Overall, the research contributes to the field by providing insights into innovative methodologies for water quality assessment.

METHODOLOGIES

In our project, we will employ a systematic and data-driven approach to develop the machine learning model for water quality analysis. First, we will collect a comprehensive dataset of water samples, which will include various physicochemical properties such as pH, hardness, conductivity, turbidity, and total dissolved solids (TDS). Once the data is collected, we will preprocess it by handling missing values, normalizing the data, and possibly applying feature selection techniques to identify the most significant indicators of water quality.

Next, we will split the dataset into training and testing subsets to ensure the model's ability to generalize to unseen data. we will experiment with different machine learning algorithms, such as decision trees, random forests, support vector machines (SVM), and neural networks, to identify the best-performing model for this classification task. The performance of these models will be evaluated using metrics such as accuracy, precision, recall, and F1-score.

After selecting the most effective model, we will fine-tune its hyperparameters to optimize its performance further. Finally, we will validate the model using the testing dataset and cross-validation techniques to ensure its robustness and reliability. Throughout this process, we will also consider the interpretability of the model to ensure that the results can be easily understood and applied by water management authorities for real-world decision-making.

- *Data Description*

The dataset consists of observations of water quality for 3276 different sources of water:

```
         ph    Hardness        Solids  Chloramines     Sulfate  Conductivity
0       NaN  204.890455  20791.318981     7.300212  368.516441    564.308654
1  3.716080  129.422921  18630.057858     6.635246         NaN    592.885359
2  8.099124  224.236259  19909.541732     9.275884         NaN    418.606213
3  8.316766  214.373394  22018.417441     8.059332  356.886136    363.266516
4  9.092223  181.101509  17978.986339     6.546600  310.135738    398.410813

   Organic_carbon  Trihalomethanes  Turbidity  Potability
0       10.379783        86.990970   2.963135           0
1       15.180013        56.329076   4.500656           0
2       16.868637        66.420093   3.055934           0
3       18.436524       100.341674   4.628771           0
4       11.558279        31.997993   4.075075           0
            ph    Hardness        Solids  Chloramines     Sulfate  \
3271  4.668102  193.681735  47580.991603     7.166639  359.948574
3272  7.808856  193.553212  17329.802160     8.061362         NaN
3273  9.419510  175.762646  33155.578218     7.350233         NaN
3274  5.126763  230.603758  11983.869376     6.303357         NaN
3275  7.874671  195.102299  17404.177061     7.509306         NaN

      Conductivity  Organic_carbon  Trihalomethanes  Turbidity  Potability
3271    526.424171       13.894419        66.687695   4.435821           1
3272    392.449580       19.903225              NaN   2.798243           1
3273    432.044783       11.039070        69.845400   3.298875           1
3274    402.883113       11.168946        77.488213   4.708658           1
3275    327.459760       16.140368        78.698446   2.309149           1
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3276 entries, 0 to 3275
Data columns (total 10 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   ph               3276 non-null   float64
 1   Hardness         3276 non-null   float64
 2   Solids           3276 non-null   float64
 3   Chloramines      3276 non-null   float64
 4   Sulfate          3276 non-null   float64
 5   Conductivity     3276 non-null   float64
 6   Organic_carbon   3276 non-null   float64
 7   Trihalomethanes  3276 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
```

pH-The water's pH (0 to 14). According to EPA recommendations, tap water's pH should range between (6.5 and 8.5). The pH level is a crucial factor in determining the acid-base nature of water. Additionally, it shows if the water is either alkaline or acidic. The present investigation's range fell between 6.52 to 6.83, which is within WHO guidelines.

Hardness - It is the amount of soap that may dissolve in one litre of water. Salts made of calcium and magnesium are the major causes of hardness. How long water is exposed to a hardness-producing substance influences how hard the water is while it is in the raw state. The ability of water to form soap due to calcium and magnesium precipitation was the original definition of hardness.

Total Dissolved Solids (TDS) - Water can dissolve a wide variety of chemicals and certain organic minerals or salts, including sodium, calcium, iron, zinc, bicarbonate ions, chloride ions, magnesium, and sulphates. These minerals affected the water's appearance and gave it foul smells. This is an important consideration while using water. A highTDS rating indicates that the water contains a lot of minerals. For drinking purposes, the maximum and desired TDS limits are 500 mg/l and 100 mg/l, respectively.

Sulfates - These are the organic substances that are found naturally in minerals, soil, and rocks. They are present in the air, groundwater, plants, and food in the area. Sulfate is mostly utilized for business purposes in the chemical sector. Around 2,700 mg/L of sulphate can be found in seawater. While certain places have significantly higher levels (1000 mg/L), most freshwater sources have values between 3 and 30 mg/L.

Conductivity - Pure water is great insulation of electrical current. By raising the ion concentration, the liquid's electric conductivity has been enhanced. The quantity of dissolved particles in the liquid often determines its conductivity. Electrical Conductivity measures how well they carry electricity through their ionic mechanism (EC). WHO recommendations state that the EC value shouldn't be higher than 400 S/cm.

Chloramines - The two primary disinfectants used in public water systems are chloride and chlorine. Ammonia is used in combination with chlorine to clean potable water. Drinking water can include up to 4 mg/L of chlorine, which is regarded as a safe quantity.

Potability - It is a metric for determining whether water is fit for human consumption. Unpotable equals zero (0), while potable is one (1).

- *Data Pre-processing*

The data quality must be improved at the processing stage of the data analysis process. The Water quality index has been determined in this phase using the important dataset parameters. The act of converting collected data into something an algorithm for machine learning can use is known as data preparation. The most important and first stage in building an algorithm for machine learning is this one. Remove all instances where the value is 0. (zero). Zero is not a possible value. Therefore, this instance is terminated. The process of deciding on feature subsets, which decreases the dimension of the data and helps to work more quickly, involves removing irrelevant characteristics and instances.

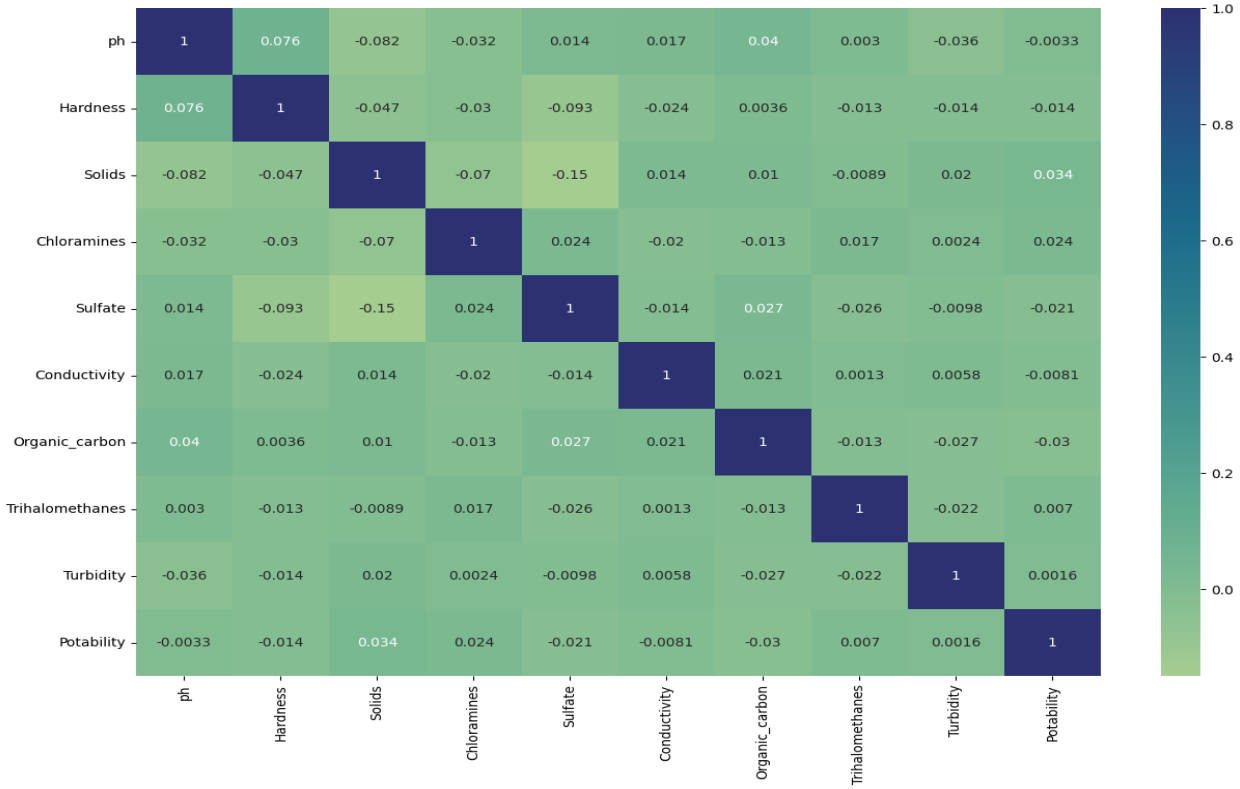
```
Missing values after imputation:
 ph                 0
Hardness            0
Solids              0
Chloramines         0
Sulfate             0
Conductivity        0
Organic_carbon      0
Trihalomethanes     0
Turbidity           0
Potability          0
dtype: int64

Summary Statistics:
                ph     Hardness        Solids  Chloramines      Sulfate  \
count  3276.000000  3276.000000   3276.000000  3276.000000  3276.000000
mean      7.080795   196.369496  22014.092526     7.122277   333.775777
std       1.469956    32.879761   8768.570828     1.583085    36.142612
min       0.000000    47.432000    320.942611     0.352000   129.000000
25%       6.277673   176.850538  15666.690297     6.127421   317.094638
50%       7.080795   196.967627  20927.833607     7.130299   333.775777
75%       7.870050   216.667456  27332.762127     8.114887   350.385756
max      14.000000   323.124000  61227.196008    13.127000   481.030642

       Conductivity  Organic_carbon  Trihalomethanes    Turbidity   Potability
count   3276.000000     3276.000000      3276.000000  3276.000000  3276.000000
mean     426.205111       14.284970        66.396293     3.966786     0.390110
std       80.824064        3.308162        15.769881     0.780382     0.487849
min      181.483754        2.200000         0.738000     1.450000     0.000000
25%      365.734414       12.065801        56.647656     3.439711     0.000000
50%      421.884968       14.218338        66.396293     3.955028     0.000000
75%      481.792304       16.557652        76.666609     4.500320     1.000000
max      753.342620       28.300000       124.000000     6.739000     1.000000
```

- *Correlation Matrix*

By Visualizing the correlation of all characteristics using a thermal foot map function. But you can see from the heat map below that there is no correlation between any characteristic; this means that we cannot reduce the dimension.



- *Training and Testing of Data*

In machine learning, the model is instructed to perform a variety of tasks using a training set of data. The model is trained using certain features from the training set. Therefore, the prototype contains these structures. Words or word clusters are taken from tweets for sentiment analysis. They build connections, understand concepts, come to judgments, and assess their level of confidence using the training data. The quality and quantity of the Machine Learning training data, we use determines how well our data project performs, just as much as the algorithms they do. As a result, provided the training set is correctly labelled, the model will be able to learn about the features.

➢ *Decision Tree*

The decision tree is a Machine Learning algorithm, it is mostly focused on classification-related issues. The decision tree has a structured classifier in which the nodes within display the components of a particular dataset. Decisionnodes and leaf nodes are both types of nodes found in decision trees.

➢ *Support Vector Machine*

The SVM is an algorithm which is used in machine learning to categorize the task. It is frequently used for classification problems. SVM separates the data into two classes by mapping the data points to a high-dimensional space and then locating the best hyperplane.

➢ *Random Forest Classifier*

The popular learning algorithm Random Forest is a part of the supervised learning methodology. It may be applied to ML issues involving both classification and regression. It is built on the idea of ensemble learning, which is a method of integrating many classifiers to address difficult issues and enhance model performance.

Random Forest, as the name implies, is a classifier that uses several decision trees on different subsets of the provided dataset and averages them to increase the dataset's prediction accuracy. Instead, then depending on a single decision tree, the random forest uses forecasts from each tree and predicts the result based on the votes of most predictions.

➢ *XGBoost*

Extreme Gradient Boosting is a framework that can run on multiple languages. It is popular supervised learning which works on large datasets. It is implemented on top of the gradient boost. The way the XGBoost algorithm is designed to work uses the parallelization concept. It uses sequentially generated shallow decision trees and a highly scalable training method to minimize overfitting to deliver accurate results.

- *Performance Metrics*

Accuracy - Accuracy is measured as the total count of actual predictions to the available predictions and it is multiplied by 100.

Precision - The ratio of actual positives to the total available positives is known as precision.

Recall - It mainly focuses on type-2 errors the ratio of true positives to false negatives is called recall.

Fl-score - The harmonic mean performance metric parameters precision with recall known as f1-score.

RESULTS

Random Forest worked the best to train the model, giving us an Accuracy (Balanced with precision & recall) of around 70%.

```
Random Forest Results:
Accuracy: 0.6784
Classification Report:
              precision    recall  f1-score   support

         0.0       0.70      0.86      0.77       412
         1.0       0.61      0.38      0.47       244

    accuracy                           0.68       656
   macro avg       0.65      0.62      0.62       656
weighted avg       0.67      0.68      0.66       656

Confusion Matrix:
[[353  59]
 [152  92]]
```

Sample Prediction:

#1:

```
# Example prediction
sample = np.array([[7.0, 200.0, 20000.0, 7.0, 300.0, 400.0, 15.0, 70.0, 4.0]])  # Replace with actual values
print(f"\nPrediction for the sample: {predict_potability(sample)}")
```

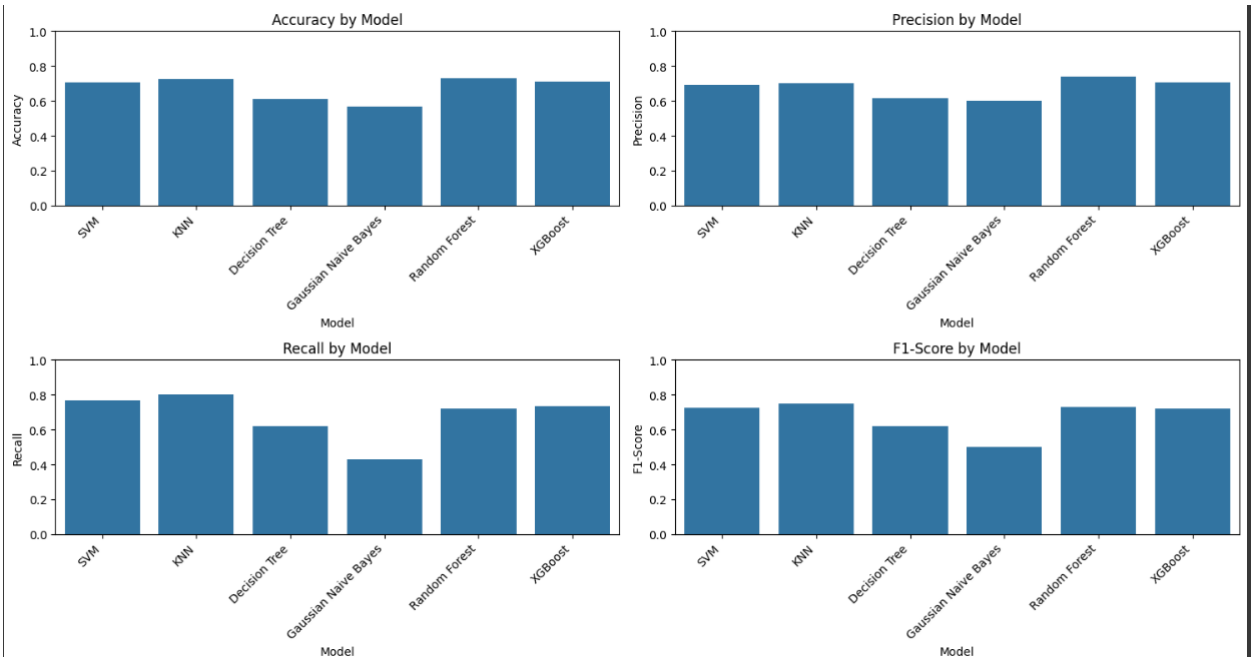
```
Prediction for the sample: Not Potable
```

#2:

```
sample1 = np.array([[9.445130, 145.805402, 13168.529156, 9.444471, 310.583374,
                     592.659021, 8.606397, 77.577460, 3.875165]])
```

```
Prediction for the sample: Potable
```

Performance Metrics for machine learning algorithms :



CONCLUSION

Future cities would benefit from real-time monitoring and evaluation of water quality due to the advancement of machine learning techniques. This work presented the results of our most recent literature analysis and comparative recent studies on the assessment of water quality using big data analytics and machine learning models and methods. Finally, it offers a few insights into theproblems, demands, and needs of future studies. Environmental protection greatly benefits from the modelling and forecasting of water quality. The algorithm implemented in this work improves the performance of water quality classifiers. We previously examined the performance metrics of machine learning algorithms, and we found that by utilizing Hyperparameter Tuning along with Random Forest Classifier, we delivered a better improvement in the execution of different performance metrics of the models using Hyperparameter Tuning. We have got better improvement in performance metrics. This strategy may be applied and improved for automated water quality monitoring.

REFERENCES

1. Khan, Y., & See, C. S. (2016, April). Predicting and analysing water quality using machine learning: a comprehensive model. In *2016 IEEE Long Island Systems, Applications and Technology Conference (LISAT)* (pp. 1-6). IEEE.

2. Ahmed, A. N., Othman, F. B., Afan, H. A., Ibrahim, R. K., Fai, C. M., Hossain, M. S., ... & Elshafie, A. (2019). Machine learning methods for better water quality prediction. *Journal of Hydrology*, *578*, 124084.

3. Vergina, S. A., Kayalvizhi, S., Bhavadharini, R., & Kalpana Devi, S. (2020). A real time water quality monitoring using machine learning algorithm. *Eur. J. Mol. Clin. Med*, *7*(8), 2035-2041.

4. Melesse, A. M., Khosravi, K., Tiefenbacher, J. P., Heddam, S., Kim, S., Mosavi, A., & Pham, B. T. (2020). River water salinity prediction using hybrid machine learning models. *Water*, *12*(10), 2951.

5. Kuthe, A., Bhake, C., Bhoyar, V., Yenurkar, A., Khandekar, V., & Gawale, K. (2022). Water quality analysis using machine learning. *International Journal for Research in Applied Science and Engineering Technology*, *10*(12), 581-585.

6. Akshay, R., Tarun, G., Kiran, P. U., Devi, K. D., & Vidhyalakshmi, M. (2022, December). Water-Quality-Analysis using Machine Learning. In *2022 11th International Conference on System Modeling & Advancement in Research Trends (SMART)* (pp. 13-18). IEEE.

7. Kaddoura, S. (2022). Evaluation of machine learning algorithm on drinking water quality for better sustainability. *Sustainability*, *14*(18), 11478.

8. Poudel, D., Shrestha, D., Bhattarai, S., & Ghimire, A. (2022). Comparison of machine learning algorithms in statistically imputed water potability dataset. *Journal of Innovations in Engineering Education*, *5*(1), 38-46.

9. Wang, X., Li, Y., Qiao, Q., Tavares, A., & Liang, Y. (2023). Water quality prediction based on machine learning and comprehensive weighting methods. *Entropy*, *25*(8), 1186.

10. Patel, S., Shah, K., Vaghela, S., Aglodiya, M., & Bhattad, R. (2023). Water Potability Prediction Using Machine Learning.

11. Brindha, D., Puli, V., NVSS, B. K. S., Mittakandala, V. S., & Nanneboina, G. D. (2023, February). Water quality analysis and prediction using machine learning. In *2023 7th International Conference on Computing Methodologies and Communication (ICCMC)* (pp. 175-180). IEEE.