Water Quality Analysis using Machine Learning

K Rajesh[†] , K Tharun Kumar Reddy[†] , Karthik Kemidi[†]
Chaitanya Bharathi Institute of Technology[†]
Department of Artificial Intelligence and Machine Learning[†]
Email: kurvarajesh72@gmail.com*, tharunkumar5297@gmail.com[†], kemidikarthik2004@gmail.com[‡]

Abstract

Water quality prediction was conducted to ascertain whether water is safe for consumption. This study also aimed to evaluate the performance of various machine learning models, including Decision Tree, Random Forest, XGBoost, KNN, SVM, and Gaussian Naive Bayes, in order to identify the most effective method for predicting water quality. Water is an essential resource for life, vital for the survival of all organisms, including humans. The sustainability of both business and agriculture relies heavily on the availability of freshwater. A critical aspect of managing freshwater resources is the assessment of water quality. Prior to utilizing water for any purpose—be it drinking, applying chemicals such as pesticides, or providing hydration for animals—it is imperative to evaluate its purity. Water quality has a direct influence on both the ecosystem and public health. Consequently, the analysis and prediction of water quality are essential for safeguarding environmental and human health. Machine learning techniques can be employed to analyze and predict water quality based on various parameters, including pH value, turbidity, hardness, conductivity, and dissolved solids. In this study, water quality is predicted by inputting the concentrations of these parameters into machine learning algorithms, classifying the water as either safe or unsafe for domestic use.

***Keywords*—Decision Tree, Random Forest, KNN, SVM, XGBoost, Gaussian Naive Bayes, Performance Metrics.**

Introduction

The domain of machine learning examines the ways in which computers acquire knowledge through experience. Given that the ability to learn is a core characteristic of what is considered intelligent, the terms "Machine Learning" and "Artificial Intelligence" are often used interchangeably by researchers. The primary objective of machine learning is to create systems that can adapt based on their experiences. Recent advancements in machine learning techniques have made it feasible to address this challenge. We have devised a method that employs data mining to determine the potability of water. The vast array of data concerning water quality can be analyzed to uncover valuable insights. Consequently, this area of study has gained increased importance. This research seeks to establish a system capable of predicting water quality with greater accuracy.

Literature Survey

[1] One comprehensive model analyzed water quality parameters using machine learning algorithms to predict conditions with high accuracy and reliability. This study highlighted the importance of data preprocessing and feature selection in improving prediction outcomes.[2] Another study compared multiple machine learning techniques to identify the most effective approach for water quality prediction. It was found that methods such as random forests and support vector machines demonstrated superior predictive performance due to their ability to handle non-linear relationships among water quality parameters.[3] Real-time monitoring systems integrated with machine learning algorithms have been developed to provide continuous water quality information. These systems enable prompt responses to potential health risks by detecting contaminants and other changes in water composition in real time. [4] Hybrid models, which combine different machine learning approaches, have shown improved performance in predicting specific water quality indicators such as salinity. These hybrid methods are particularly effective in capturing complex interactions between variables that traditional models may miss. [5] Some studies focus on analyzing and predicting water quality using datasets from field sensors and public records. By applying exploratory data analysis techniques, researchers have identified key patterns in the data, enabling them to predict long-

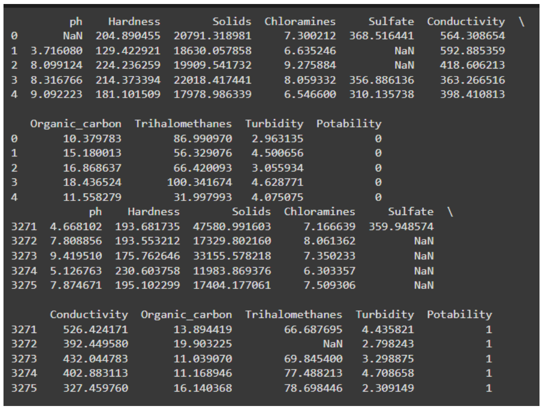
term trends and seasonal variations in water quality.[6] Recently, there has been a rise in research leveraging machine learning to assess and predict water potability, training models to evaluate water's suitability for drinking based on a variety of quality metrics. [7] Studies have evaluated machine learning algorithms for sustainable monitoring of drinking water quality, emphasizing these models' potential to make water management more efficient. This research often discusses the implications for sustainability and public health.[8] Comparative analyses of various machine learning algorithms on datasets with statistically imputed missing values have deepened the understanding of each model's strengths and limitations, highlighting how data preprocessing affects model accuracy. [9] The integration of machine learning with comprehensive weighting techniques has been explored to enhance prediction reliability by combining the outputs of multiple models. This method reduces errors from individual models, yielding more robust predictions.[10] Several studies have aimed to predict water potability by classifying water samples based on various quality indicators. This research is particularly valuable for public health monitoring.[11] Recent research has focused on optimizing machine learning models for water quality prediction by refining data preprocessing and tuning model parameters. These tailored models improve prediction accuracy significantly, especially when adapted to local environmental conditions and specific water quality attributes.

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.

Methodologies

The machine learning model is employed to ascertain whether water is safe for consumption or not. It is essential to import the necessary libraries to facilitate the testing and training of our dataset, along with installing specific packages related to nature-inspired algorithms. The dataset should be divided into training and testing subsets in an 80:20 ratio, followed by the selection of the appropriate model. The classifiers under consideration include the Support Vector Classifier (SVC), Decision Tree, GaussianNB, Random Forest, and XGBoost.



```
         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
```

Figure 1

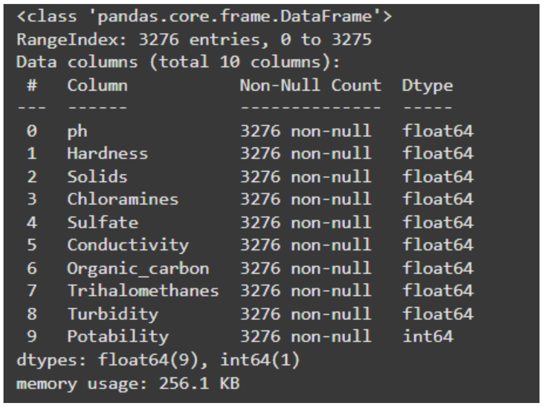
The dataset consists of observations of water quality for 3276 different sources of water:

pH - This parameter measures the acidity or alkalinity of water on a scale from 0 to 14. According to the Environmental Protection Agency (EPA) guidelines, the pH of tap water should ideally fall between 6.5 and 8.5. The pH level is a significant determinant of the water's acid-base characteristics. The current study recorded pH values ranging from 6.52 to 6.83, which aligns with World Health Organization (WHO) standards.

Hardness - This refers to the quantity of soap that can dissolve in one liter of water. The primary contributors to water hardness are salts composed of calcium and magnesium. The duration of water's exposure to hardness-inducing substances affects its hardness in its natural state. Traditionally, hardness is defined by the water's capacity to form soap due to the precipitation of calcium and magnesium.

Total Dissolved Solids (TDS) - Water has the ability to dissolve a variety of chemicals and certain organic minerals or salts, including sodium, calcium, iron, zinc, bicarbonate ions, chloride ions, magnesium, and sulfates. These dissolved minerals can alter the water's appearance and contribute to unpleasant odors. A high TDS level indicates a significant presence of minerals in the water. For safe drinking, the recommended maximum TDS limit is 500 mg/l, with a desired limit of 100 mg/l.

Sulfates - Sulfates are organic compounds that occur naturally in various environments, including minerals, soil, rocks, air, groundwater, plants, and food within a given area. In the chemical industry, sulfates are primarily employed for commercial applications. Seawater typically contains approximately 2,700 mg/L



```
<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
```

Figure 2: data description

of sulfate, while certain regions may exhibit much higher concentrations, reaching up to 1,000 mg/L. In contrast, most freshwater sources generally have sulfate levels ranging from 3 to 30 mg/L.

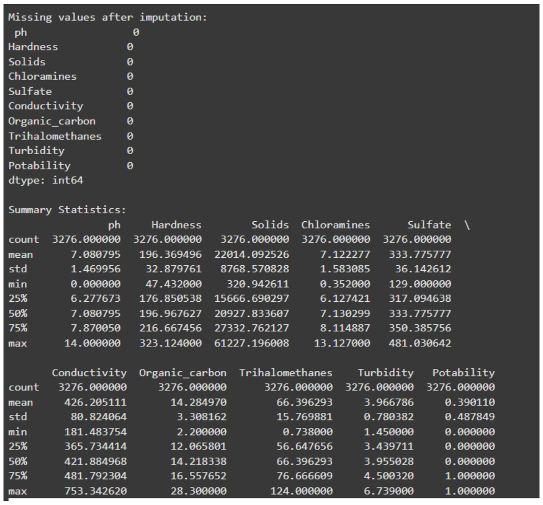
Conductivity - Pure water serves as an excellent insulator of electrical current. The level of dissolved particles within the liquid typically dictates its conductivity. Electrical conductivity (EC) assesses the efficiency with which these particles transmit electricity through their ionic interactions. According to the recommendations of the World Health Organization (WHO), the EC value should not exceed 400 S/cm.

Chloramines - The two main disinfectants employed in public water systems are chloride and chlorine. Ammonia is utilized alongside chlorine to purify drinking water. It is considered safe for drinking water to contain up to 4 mg/L of chlorine.

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

A. Data Pre-processing

1. A. Data Pre-processing Enhancing data quality is essential during the processing phase of data analysis. In this stage, the Water Quality Index is established based on key parameters from the dataset. Data preparation refers to the transformation of collected data into a format suitable for machine learning algorithms. This step is crucial and serves as the foundational phase in the development of a machine learning algorithm. It is necessary to eliminate all instances where the value is zero, as zero is not a valid value; thus, such instances are discarded. The process of selecting feature subsets, which reduces data dimensionality and facilitates faster processing, involves the removal of irrelevant features 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
```

Figure 3: data pre-processing

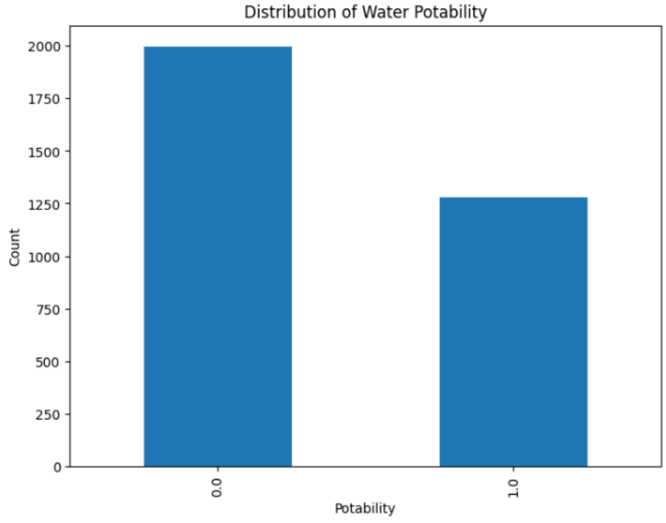


Figure 4: Water Potability Distribution

B. Correlation Matrix

The correlation among all features is visualized using a heat map function. However, the heat map presented below indicates that there is no correlation among any of the features, suggesting that dimensionality reduction is not feasible.

*C. **Training and Testing of Data***

In the field of machine learning, a model is directed to execute various tasks by utilizing a training dataset. This model undergoes training based on specific features extracted from the training data. For sentiment analysis, words or clusters of words are sourced from tweets. The model establishes relationships, comprehends concepts, formulates conclusions, and evaluates its confidence levels through the training data. The effectiveness of our data project is influenced by both

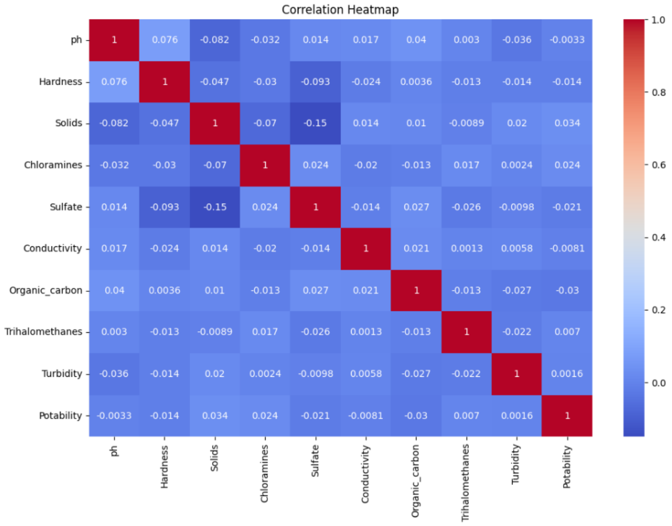


Figure 5: correlation matrix

the quality and quantity of the machine learning training data, as much as by the algorithms employed. Consequently, if the training set is accurately labeled, the model will be capable of learning the relevant features.

D. Decision Tree

The decision tree is a machine learning algorithm primarily concerned with classification tasks. It features a structured classification system where the nodes represent various elements of a specific dataset. Within decision trees, there are two main types of nodes: decision nodes and leaf nodes.

E. Support Vector Machine

Support Vector Machine (SVM) is an algorithm employed in machine learning for the purpose of task categorization. It is commonly utilized in classification challenges. SVM distinguishes between two classes by transforming the data points into a high-dimensional space and subsequently identifying the optimal hyperplane.

F. Random Forest Classifier

Random Forest is an integral component of the supervised learning framework. It can be utilized for machine learning tasks that encompass both classification and regression. This technique is founded on the principle of ensemble learning, which involves the combination of multiple classifiers to tackle complex problems and improve the overall performance of the model. As suggested by its name, Random Forest operates as a classifier that employs numerous decision trees, each trained on different subsets of the input data, and aggregates their outputs to enhance the accuracy of predictions. Rather than relying on a single decision tree, the random forest aggregates the predictions from all trees and determines the final outcome based on the majority vote of these predictions.

G. 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.

H. KNN

K-Nearest Neighbors (KNN) is a simple, non-parametric machine learning algorithm used for classification and regression tasks. It works by finding the "k" closest data points (neighbors) to a given input and making predictions based on the majority class or average value of these neighbors. KNN is easy to implement but can be computationally intensive, especially with large datasets, since it requires calculating distances between points.

I. 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.

F1-score - The harmonic mean performance metric parameters precision with recall known as f1-score.

Algorithm

1. Import Libraries: Import necessary libraries like pandas, numpy, matplotlib, seaborn, and machine learning modules from sklearn.

2.Load Dataset: Load the datset from kaggle $water_potability.csv$ into a DataFrame.

3. Handle Missing Values: Impute missing values using the mean strategy with SimpleImputer.

4. Data Exploration and Visualization: Compute the correlation matrix and visualize it using a heatmap. Analyze the distribution of the target variable (Potability) with a bar chart.

5. Data Preprocessing: Separate features (X) and the target variable (y). Split the dataset into training and testing sets using an 80-20 split. Scale features using StandardScaler for uniform scaling.

6. Define and Train Models: Define multiple classifiers: Random Forest, Gradient Boosting, AdaBoost, and SVM. Train each model on the scaled training data.

7. Evaluate Models: Make predictions on the test data for each model. Calculate evaluation metrics: Accuracy, Classification Report, and Confusion Matrix.Identify the model with the highest accuracy.

8. Cross-Validation: Perform 5-fold cross-validation on the best model and compute mean cross-validation scores.

9. Feature Importance (for Random Forest): If Random Forest is the best model, extract and visualize feature importances.

10. Save the Best Model: Save the trained best model to a .pkl file using joblib.

11. Load and Use the Model: Load the saved model for future predictions. Ensure input samples are scaled before making predictions.

12. Prediction Function: Define a function to predict water potability for a given sample. Use the best model to classify whether the water is potable or not.

13. Test Prediction: Provide sample inputs to the prediction function for testing. Interpret the model's output as "Potable" or "Not Potable".

Hardware Requirements

Hardware	Description
System	13th Gen Intel(R) Core(TM),1.90GHz
Hard Disk	512 GB
Monitor	HP P204v
Process Intel	i5-1340P

Table I: hardware requirements

Software Requirements

Software	Description
Operating System	Windows 11
Programming Language	Python 3.2
Database	Firebase
Tools	Jupyter Notebook, Google Colab, Python IDE

Table II: Software Requirements

Results and Discussion

- The figure 6 shows the accuracy of six different machine learning models: SVM, KNN, Decision Tree, Gaussian Naive Bayes, Random Forest, and XGBoost. The accuracy is fairly consistent with SVM, KNN, Random Forest, and XGBoost achieving higher values compared to Decision Tree.

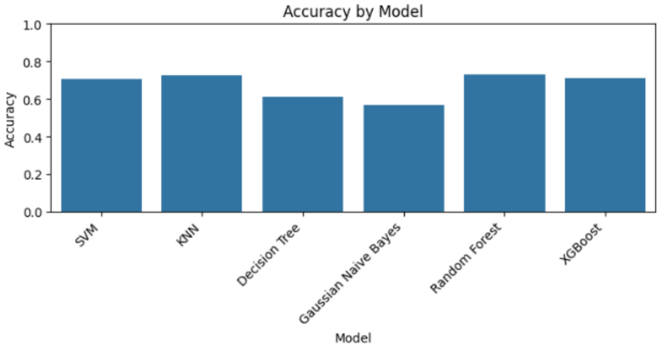


Figure 6: Accuracy

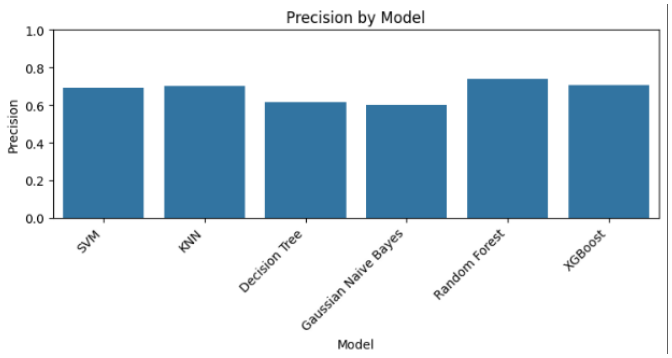


Figure 7: Precision

- Figure 7 displays the precision for each model. Precision is highest for the Random Forest model, followed closely by KNN and XGBoost, while Gaussian Naive Bayes has slightly lower precision. Higher precision means fewer false positives for these models.

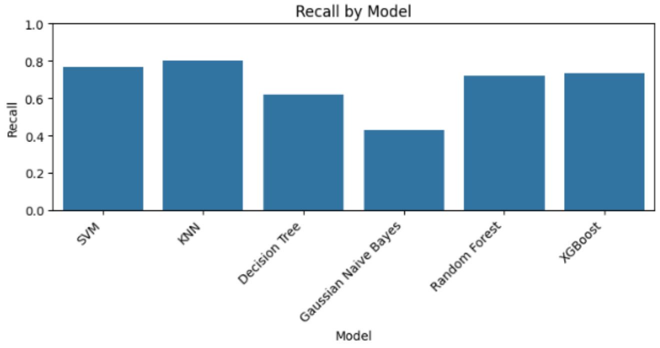


Figure 8: Recall

- Figure 8 illustrates recall values, where SVM and KNN show the highest recall, meaning these models are better at identifying true positives. Gaussian Naive Bayes has the lowest recall, indicating it misses more true positives.
- Figure 9 illustrates the F1-score, which is the harmonic mean of precision and recall. KNN has

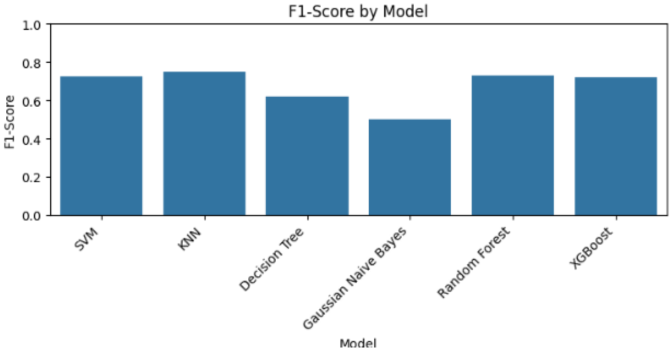
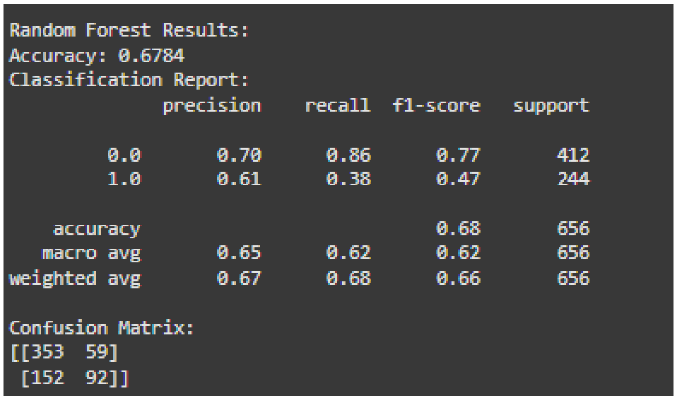


Figure 9: F1-score

the highest F1-score, indicating a good balance between precision and recall. Gaussian Naive Bayes has the lowest F1-score, suggesting it is less effective in balancing precision and recall.

In summary, the Random Forest Classifier proved to be the most effective model, attaining an ideal equilibrium among accuracy, precision, recall, and F1-score. This model's exceptional balance suggests its suitability for applications that demand reliable accuracy across various evaluation metrics. The Random Forest Classifier demonstrated the highest performance in training the model, yielding an accuracy of approximately 70 percent when balanced with precision and recall.



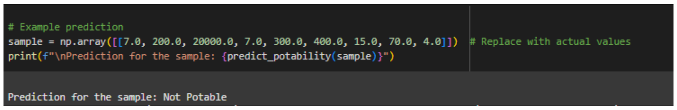
```
Random Forest Results:
Accuracy: 0.6784
Classification Report:
              precision    recall  f1-score   support

         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]]
```

Figure 10: Best Model



```
# 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
```

Figure 11: Sample prediction

Conclusions and Future Scope

Future urban environments stand to gain significantly from the implementation of real-time monitoring and evaluation of water quality, facilitated by advancements in machine learning techniques. This study presents the findings of our latest literature review and comparative analysis of recent research focused on assessing water quality through big data analytics and machine learning methodologies. Additionally, it provides insights into the challenges, requirements, and priorities for future research endeavors. The modeling and forecasting of water quality play a crucial role in environmental protection. The algorithm developed in this study enhances the efficacy of water quality classifiers. Our previous investigations into the performance metrics of machine learning algorithms revealed that the combination of Hyperparameter Tuning with the Random Forest Classifier yielded substantial improvements in various performance metrics of the models. This approach has resulted in enhanced performance metrics overall. Such a strategy can be further refined and applied to automated water quality monitoring systems.

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.