

**A MACHINE LEARNING APPROACH TO PREDICT THE QUALITY OF
DRINKABLE WATER FROM DIFFERENT SOURCES**

BY

**FATEMA TUZ JAHURA
183-15-2230**

This Report is Submitted in Partial Completion of the Prerequisites for the
Bachelor of Science Degree in Computer Science and Engineering.

Supervised By

Zakia Sultana
Senior Lecturer
Department of CSE
Daffodil International University

Co-Supervised By

Mushfiqur Rahman
Senior Lecturer
Department of CSE
Daffodil International University



DAFFODIL INTERNATIONAL UNIVERSITY

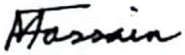
DHAKA, BANGLADESH

JANUARY 2024

APPROVAL

This Project titled "A Machine Learning Approach to Predict the Quality of Drinkable Water from Different Sources", submitted by Fatema Tuz Jahura, ID No: 183-15-2230 to the Department of Computer Science and Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on 25 January 2024.

BOARD OF EXAMINERS

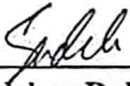


Dr. Md. Fokhray Hossain (MFH)

Professor

Department of Computer Science and Engineering
Daffodil International University

Chairman

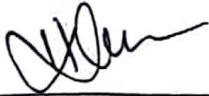


Md. Sadekur Rahman (SR)

Assistant Professor

Department of Computer Science and Engineering
Daffodil International University

Internal Examiner



Most. Hasna Hena (HH)

Assistant Professor

Department of Computer Science and Engineering
Daffodil International University

Internal Examiner



Dr. S. M. Hasan Mahmud

Assistant Professor

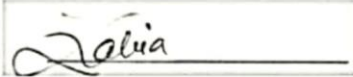
Department of Computer Science
American International University-Bangladesh

External Examiner

DECLARATION

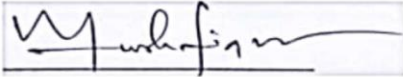
I hereby declare that this project has been done by me under the supervision of **Zakia Sultana**, Lecturer (Senior Scale), Department of CSE, Daffodil International University. I also declare that neither this project nor any part of this project has been submitted elsewhere for award of any degree or diploma.

Supervised by:



Zakia Sultana
Senior Lecturer
Department of CSE
Daffodil International University

Co-Supervised by:



Mushfiqur Rahman
Senior Lecturer
Department of CSE
Daffodil International University

Submitted by:

Fatema Tuz Jahura

Fatema Tuz Jahura
183-15-2230
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

First, Firstly, I extend my deepest appreciation to the divine power for His blessings that have enabled me to successfully complete my final year project.

I am deeply indebted to Zakia Sultana, Senior Lecturer, Department of CSE, Daffodil International University, Dhaka, for her profound expertise and interest in “*Data Mining*”, which has been instrumental in carrying out this project. Her unwavering patience, academic guidance, continuous motivation, consistent supervision, constructive feedback, valuable suggestions, and her willingness to read and correct numerous drafts have been pivotal in the successful completion of this project.

I wish to convey my sincerest gratitude to Dr. Sheak Rashed Haider, Professor and Head, Department of CSE, and Mushfiqur Rahman, Senior Lecturer, Department of CSE, for their invaluable assistance in completing our project, as well as to the other faculty members and staff of Daffodil International University.

I would also like to express my thanks to all my course mates at Daffodil International University, who participated in discussions during the course work.

Lastly, I must express my respect and gratitude for the unwavering support and patience of our parents and without it I could not complete my final defense.

ABSTRACT

Machine learning (ML) is revolutionizing the field of aquatic environment research by offering advanced tools for analyzing, classifying, and predicting data. This study delves into the use of ML algorithms, particularly Decision Trees, Random Forests, and XGBoost, for assessing water quality across various contexts such as surface water, groundwater, drinking water, and wastewater. These ML models excel in handling the increasing complexity and volume of data in water research, surpassing the capabilities of traditional models. In this work, I explored the application of ML in several key areas: monitoring and simulation of water systems, evaluation, and optimization of water treatment processes, and addressing challenges like water pollution and watershed security. The ability of ML models to process data from diverse sensors and monitoring systems in real-time makes them invaluable for understanding water quality parameters and identifying potential risks. The predictive power of ML is particularly noteworthy in forecasting changes in water quality due to environmental factors, which is critical for proactive water management and policymaking. Furthermore, the study highlights how ML aids in optimizing water treatment processes, leading to more efficient and sustainable operations. Looking ahead, the study discusses the potential future applications of ML in the aquatic domain. This includes the integration of deep learning methods for more nuanced analyses, improved handling of data variability and uncertainty, and the combination of ML with other emerging technologies such as IoT, blockchain, and cloud computing. This synergy is poised to enhance water resource management, emphasizing sustainability, accessibility, and conservation. In summary, this work presents a comprehensive overview of how ML algorithms are transforming the landscape of water environment research, offering innovative solutions for current challenges, and opening new avenues for future exploration.

TABLE OF CONTENTS

CONTENTS	PAGE
Approval	ii
Declaration	iii
Acknowledgements	iv
Abstract	v
CHAPTER	
Chapter 1: Introduction	1-4
1.1 Introduction	1
1.2 Motivation	2
1.3 Justification of Research	2
1.4 Inquiry of Study	3
1.5 Anticipated Results	3
1.6 Oversight and Budgeting of Project	4
1.7 Structure of the Report	4
Chapter 2: Background Study	5-8
2.1 Foundational Concepts	5
2.2 Prior Studies	5-6
2.3 Comparative Evaluation	6-7
2.4 Problem Boundaries	8
2.5 Hurdles	8

Chapter 3: Methodological Approach	9-19
3.1 Study Focus and Tools	9
3.2 Procedure for Gathering Data	9-10
3.3 Quantitative Evaluation	11-14
3.4 Suggested Approach	15-16
3.5 Prerequisites for Execution	17
Chapter 4: Analysis and Interpretation of Findings	18-24
4.1 Configuration of Experiment	18
4.2 Analysis & Interpretation of Results	18-24
Chapter 5: Influence on Society, Sustainability & Environment	25-26
5.1 Societal Influence	25
5.2 Environmental Implications	25
5.3 Ethical Considerations	25
5.4 Plan for Sustainability	26
Chapter 6: Recap, Findings, Suggestions, and Potential for Future Exploration	27
6.1 Recap of Investigation	27
6.2 Final Remarks	27
6.3 Potential for Future Research	27
REFERENCES	28

List of Tables**PAGE**

2.1 List of Similar Works	06
3.1 Evaluation of Mean Absolute Deviation & Mean Squared Deviation	14
4.1 Content Analysis of the Confusion Matrix	24

List of Figures	PAGE
2.1 Analysis of Phyco-Chemical Characteristics	7
3.1 Data Collection	10
3.2 Pre-Processing of Data	10
3.3 Graphical Representation of Potability Distribution	11
3.4 Feature Distribution Graphs of the Dataset (Part 1)	11
3.5 Feature Distribution Graphs of the Dataset (Part 2))	12
3.6 Matrix of Correlation	14
3.7 Implementation of Random Forest Algorithm	16
3.8 Performance Metrics of XGBoost Classifier	16
3.9 Prediction Using a New Entry	17
4.1 Model with Superior Performance	18
4.2 Chart Depicting Accuracy	19
4.3 Chart of Misclassifications	19
4.4 Cross-Validation Graph	20
4.5 Receiver Operating Characteristic – Area Under Curve Graph	20
4.6 Score of ROC-AUC	21
4.7 Confusion Matrix for the Random Forest Algorithm	22
4.8 Confusion Matrix for the Decision Tree Algorithm	23
4.9 Confusion Matrix for the XGBoost Algorithm	23

CHAPTER 1

INTRODUCTION

1.1 Introduction

Water is undoubtedly a vital resource for sustaining life, and the quality of water is crucial for the well-being of humans and other living organisms. Nevertheless, the quality of water is declining because of a multitude of factors, including human waste, industrial pollutants, natural processes, and climate change. These factors have a significant impact on the health of humans and ecosystems, particularly in developing countries with limited access to clean water. Hence, it is crucial to closely observe and evaluate the quality of water in various situations, including surface water, groundwater, drinking water, wastewater, irrigation water, and industrial water. In this paper, I delve into the application of machine learning (ML), a field within computer science, to analyze and categorize water quality parameters in different water environments. Machine learning is a highly effective method for analyzing data, making classifications, and making predictions. It excels at solving intricate and non-linear problems that traditional models struggle with. I have utilized ML algorithms to analyze a limited range of water quality-related diseases, and the results have demonstrated their effectiveness in accurately determining the drinkability of water. I have used the Random Forest, Decision Tree and XGBoost algorithms for the code. I anticipate that the efficacy of ML algorithms will enhance as I gain experience with a wider range of water samples (having potability 1) in the future. I will utilize these algorithms to estimate the amount of water in various water scenarios. Additionally, I explored the future possibilities of ML techniques in aquatic domains and my work focuses on the prediction of drinkable water. I also analyzed the performance of various machine learning algorithms for water quality prediction and provided a comprehensive discussion on the strengths and weaknesses of each approach. I examine the latest literature on the use of machine learning in water quality analysis and discuss the existing research gaps and challenges. I offer several suggestions for future work in this field, including the incorporation of additional water quality parameters, more advanced machine learning techniques, and the integration of remote sensing data.

1.2 Motivation

The primary objective of this work is to employ machine learning techniques to assess water quality. I utilize the concept of potability to assess the appropriateness of water for consumption. In this study, I utilized a range of water quality parameters to evaluate the drinkability of water. These parameters include trihalomethanes, turbidity, conductivity, organic carbon, hardness, pH, chloramines, and solids. The parameters create a feature vector that represents the water quality. Creating an IoT-based device that can assess the drinkability of water would be highly advantageous for individuals.

1.3 Rationale of study

The well-being of both ecosystems and humans is impacted by water quality. Drinking, farming, and manufacturing are just a few of the numerous uses for water. The MLP model was marginally superior to the others, while the authors acknowledged that all the models were adequate for predicting water quality components. Using this model, they were able to regulate the water supply system's water quality. They used state-of-the-art techniques to tackle it by viewing it as an optimization problem. Predicting and evaluating water quality is critical for water conservation, according to a literature analysis. That is why they turned to solutions powered by AI. The Tireh River, a key river in Iran's Dez basin and a significant watershed, was the subject of their study. All forms of life depend on water. Water scarcity and safety are issues in many areas. Contaminated water poses health risks due to the presence of hazardous organisms, chemicals, and contaminants. There is a worldwide concern for water security and quality. Human activities are contaminating the freshwater resources that consist of only 2% of the total water supply of the world. People may feel less affected by contaminated water if this happens. A widely used technology known as the Internet of Things (IoT) allows users to link a variety of sensors and gear to the web. Water quality standards were established by the World Health Organization (WHO). The authors checked their findings against these limits and alerted the reader when any parameter went beyond, preventing water contamination.

1.4 Research Questions

1. To what extent can ML algorithms such as Support Vector Machine, Random Forest, Decision Tree, XGBoost and KNN accurately forecast whether water from different sources is drinkable?
2. How does the quality of surface water, groundwater, and potable water alter depending on the elements that are most important in determining water quality?
3. With industrial and household water usage on the rise, how might machine learning methods be used to improve and control water quality?
4. How might data preprocessing enhance the precision of machine learning models used to forecast water quality?
5. How can future research overcome the limitations and difficulties of existing machine learning methods for forecasting potability and water quality?

1.5 Expected Output

I have focused on improving the accuracy of a Kaggle dataset in my attempt to enhance water quality prediction. Managing and maintaining water habitats is of utmost importance, particularly considering the anticipated increase in water demand caused by population expansion, lifestyle changes, and construction activities. There will likely be a 20–50 percent rise in industrial and household water consumption from present levels by the year 2050. By 2030, the disparity between the water supply of the world and necessity might grow to 40%, making water shortages worse in some areas and stifling economic growth in others. I used state-of-the-art ML algorithms to evaluate a subset of water-related characteristics to overcome these obstacles. Using models such as XGBoost Classifier, Decision Tree, and Random Forest greatly enhances the capability to detect different types of water quality, in my experience. After analysis, I have achieved an accuracy of 69% using Random Forest. Both the identification method and the detection rate of drinkable water in future research can be improved using this approach.

1.6 Project Management and Finance

This study examines how to use advanced machine learning models like XGBoost Classifier, Decision Tree, and Random Forest to monitor and predict water quality effectively. It requires forethought, the distribution of resources, and the application of cutting-edge technology to the analysis and prediction of data. Optimization of resources, creation of budgets, and distribution of funds for the purchase of technology are all aspects of financial management. The project's resilience to environmental and economic changes, the development of a resilient financial strategy to support the adoption and maintenance of these machine learning models, and the development of sustainable water resource management practices are the main objectives. In addition, the project strives to promote adaptability in light of changing environmental and economic circumstances, guaranteeing the continued effectiveness of water quality management strategies in different situations. The text highlights the importance of conducting a comprehensive risk assessment to identify potential challenges and developing contingency plans to address these risks. This approach helps to maintain the continuity and reliability of water quality monitoring and management practices.

1.7 Report Layout

- Chapter 1 introduced the "A machine learning model for predicting water quality of drinkable water from various sources" and its objectives, targets, and expected outcomes.
- In Chapter 2, the related works section describes the outline, scope of the problem, and the obstacles as its main topics.
- Chapter 3 provides an in-depth exploration of the research approach.
- Chapter 4 contains a comprehensive discussion of the experiment results.
- Chapter 5 describes the subjects, including the impact of society and sustainability.
- The evaluation scores are reviewed in Chapter 6, providing helpful insights that can help future works better reflect my research attempts.

CHAPTER 2

BACKGROUND

2.1 Preliminaries

My study has led to the creation of a simple descriptive language for analyzing intricate systems. Yet, the real test is in steering clear of excessively intricate explanations that could be better conveyed through metaphors. Our study primarily involved the analysis of patterns in water quality data, drawing parallels to the monitoring of seismic activities through ground water changes and animal behavior before the Hatching earthquake. Through careful observation of data points, we discovered intriguing correlations that resemble the well-documented phenomena that occur before earthquakes. To gain a deeper understanding of these variations, we employed advanced machine learning models. The significance of detecting subtle shifts in environmental factors is highlighted by the findings obtained from analyzing groundwater and animal behavior data. This principle was effectively employed in our water quality assessment through the utilization of sophisticated algorithms.

2.2 Related works

Recent research has revealed that algorithm output after training is critical for effectively forecasting new dataset results. These methods estimate unknown variables using previous data. Due to their predictive power, SVM, KNN Random Forest, Decision Tree, and XGBoost are popular in research. KNN is considered as a supervised machine learning approach that forecasts a target value using independent parameters. It has been successfully adopted in various businesses, including health organizations for disease diagnosis. Random Forest, an ensemble learning method that integrates many Decision Trees, is known for its accuracy and efficiency with large datasets. It helps refine prediction models by assessing feature importance. SVM, a prominent supervised ML technique, is used for classification. It excels in multi-dimensional hyperplane recognition for data categorization. High accuracy and ability to handle overfitting in other models are also hallmarks of the Extreme Gradient Boosting (XGBoost) technique.

These algorithms have proven their predictive modelling flexibility and dependability in several investigations. Their adaptability in processing different data kinds and sizes makes them invaluable in machine learning.

Table 2.1: List of Related Works

Authors	Publication Year	Journal Title	Findings
A. N. Ahmed	2019	Elsevier	Machine learning methods for better water quality prediction. ^[1]
H. Haghiabi	2018	Water Quality Research Journal	Water quality prediction using machine learning methods. ^[2]
M. Azrou	2022	Springer	Machine learning algorithms for efficient water quality prediction ^[3]
U. Ahmed	2019	MDPI	Efficient Water Quality Prediction Using Supervised Machine Learning. ^[5]

2.3 Comparative Analysis

BOD is an essential parameter that signifies the oxygen requirement for the decomposition of organic substances in water by microorganisms. The BOD values for the Buriganga and Balu Rivers were discovered to be well above the Department of Environment's (DoE) standard of 50 mg/L, with average readings of 145 mg/L & 116 mg/L respectively. The high levels observed suggest significant organic pollution.

In contrast, COD measures the presence of both non-biodegradable and biodegradable contaminants. The COD values in the Buriganga and Balu Rivers were found to be 277 mg/L and 202 mg/L, respectively. It is worth mentioning that there was a noticeable decrease in COD values downstream in the Buriganga River, indicating the presence of pollution sources upstream.

These findings are of utmost importance for water quality prediction. The elevated BOD and COD levels, when compared to the usual BOD levels found in river surface water (ranging from 1 to 8 mg/L), along with the variations observed in different sections of the river, emphasize the necessity for specific pollution control measures. The data, especially the variation in COD across different locations, can be used to develop predictive models that help identify and tackle major sources of pollution. This, in turn,

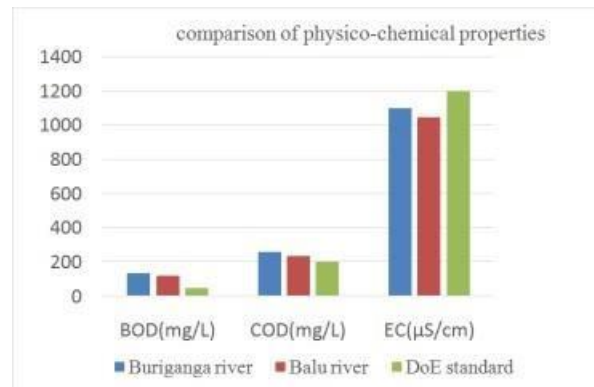


Figure 2.1: Comparison of physico-chemical properties

can lead to improved water quality management strategies. The correlation index indicated a strong relationship between BOD and COD. The Buriganga River is being negatively impacted by the sewage and tannery wastes flowing from Dhaka. This aspect was largely dominated by the load of organic pollutants. Table 3 demonstrates a strong correlation (0.973) between BOD and COD, both of which serve as indicators for organic compounds.

2.4 Scope of the problem

Water has played many crucial role in shaping human civilization throughout history. Throughout history, the progress of communities has been closely linked to their capacity to obtain and cleanse water for various purposes such as farming, hygiene, consumption, and safeguarding the environment. These factors have been instrumental in shaping public health and ensuring the long-term viability of human settlements. Nevertheless, the dynamics of water usage and its availability can be influenced by a multitude of factors that may vary over time.

1. There are several factors that contribute to the water crisis:
2. Contamination of Water Sources
3. Concerns about the depletion of groundwater resources
4. excessive water consumption,
5. Impacts of Climate Change.

2.5 Challenges

Wetlands have decreased in size by over 50% on a worldwide scale. There is a lot of water lost in agriculture because of inefficiencies, yet it still utilizes more water than any other industry. Because of global warming, water shortages and droughts are becoming more common in some areas, while floods are becoming more common in others. In developing countries, water sources are the destination for 90% of sewage. The number of effluents and sewage that are thrown into the ocean daily is 2 million tons. 3. Every year, waterways get 300-400 megatons of garbage from industry.

CHAPTER 3

RESEARCH METHODOLOGY

3.1 Research Topic

The research focuses on predicting water potability using machine learning methods in Google Colab. We have employed a variety of algorithms like Random Forest, Decision Tree, and XGBoost to accomplish this objective. The procedure includes data preparation and the reduction of any unnecessary disturbances, making it fit for training with TensorFlow. Once the data was gathered, we moved on to validate and test the images. Python, along with libraries such as Matplotlib and pandas, were utilized for data visualization, which included the generation of confusion matrices and histograms. This approach precisely evaluates the potability of water.

3.2 Data Collection Method

Dataset from the Kaggle has been utilized in this research project. The whole quantity of samples was 8127 in total. The dataset contains several key metrics, such as organic carbon, trihalomethanes, pH, conductivity, solids, chloramines, and sulphate. These metrics are included in the dataset. The water that is consumed by people in Bangladesh is safe since it is produced in accordance with the criteria that have been established by the International Water Association (IWA).

	A	B	C	D	E	F	G	H	I	J
1	ph	Hardness	Solids	Chloramines	Sulfate	Conductivity	Organic_carbon	Trihalomethanes	Turbidity	Potability
2		204.8904555	20791.31898	7.300211873	368.5164413	564.3086542	10.37978308	86.99097046	2.963135381	0
3	3.716080075	129.4229205	18630.05786	6.635245884		592.8853591	15.18001312	56.32907628	4.500656275	0
4	8.099124189	224.2362594	19909.54173	9.275883603		418.6062131	16.86863693	66.42009251	3.05593375	0
5	8.316765864	214.3733941	22018.41744	8.059332377	356.8861356	363.2665162	18.4365245	100.3416744	4.628770537	0
6	9.092223456	181.1015092	17978.98634	6.546599974	310.1357375	398.4108134	11.55827944	31.99799273	4.075075425	0
7	5.584086638	188.3133238	28748.68774	7.544868789	326.6783629	280.4679159	8.39973464	54.91786184	2.559708228	0
8	6.00897361	225.0802338	5100.094173	7.45223619	336.119	325.1344922	11.07995155	36.34101183	4.012340383	1
9	7.607223911	160.565253	39184.84672	7.826411049	312.0560665	503.1580785	13.36699449	62.02230789	3.525027131	1
10	6.683367697	272.1116985	18989.31677	5.336201994	336.5551001	307.725009	20.17871618	75.40226028	5.208061134	1
11	6.638411449	180.8266674	9772.504814	8.295983092		401.1111434	12.60151733	61.05188925	5.16405662	1
12	9.271355447	181.2598172	16540.97905	7.022499179	309.2388651	487.6927878	13.228441		4.333952898	1
13		134.7368557	9000.025591	9.026292723		428.213987	8.668672182	74.77339241	3.699558048	1
14	3.629922065	244.1873915	24856.63321	6.818071066	366.9678733	442.0763366	13.30288014	59.48929351	4.754826393	1
15	8.378108023	198.5112127	28474.20258	6.477056754	319.4771873	499.8669939	15.38908341	35.22120041	4.52469297	1
16	6.923636014	260.5931543	24792.52562	5.501164043	332.2321775	607.7735673	15.46302674	51.53586708	4.013338801	1
17	5.893103408	239.2694615	20526.66616	6.349560868	341.256362	403.61756	18.96370676	63.84631932	4.390701604	1
18	8.197353369	203.1050914	27701.79405	6.472914286	328.8868376	444.6127236	14.25087508	62.90620518	3.361833324	1
19	8.372910285	169.0870522	14622.74549	7.547984018		464.5255524	11.08302657	38.43515078	4.906358241	1
20	5.27418539	227.340186	17605.53576	6.326979503	358.589903	489.4345906	11.19919093		4.364426392	1

Figure 3.1: Dataset

The calculation phase of data processing is crucial for enhancing data quality. This step involves exploring the most crucial properties of the dataset to identify exploration of the data and scale the important feature. Later, the water trials were ordered into different groups according to their WQI ratings. The dataset includes a total of 8127 water samples. I can provide you with a visual representation of sample water in figure 2 for your understanding. Water is included, regardless of its drinkability.

The amount of train and test data required for every classification: Dataset dimensions: (6501, 9) & (6501, 1)

Testing dataset: (1626, 9)

Dataset Preprocessing:

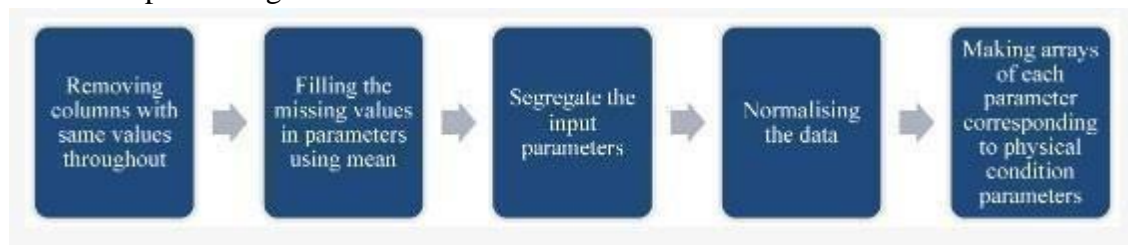


Figure 3.2: Dataset preprocessing

3.3 Statistical Analysis

Predicting pH and Hardness was done via pre-processing. Solids Sulphuric acid How Sulphates Conduct Electricity Trihalomethanes that are organic Turbidity Effectiveness, etc. A distinction is drawn between physical parameters, which are immutable, and changeable parameters, which are subject to change. Follow that by displaying the potability, where 0 indicates that the liquid is not drinkable and 1 indicates that it is potable.

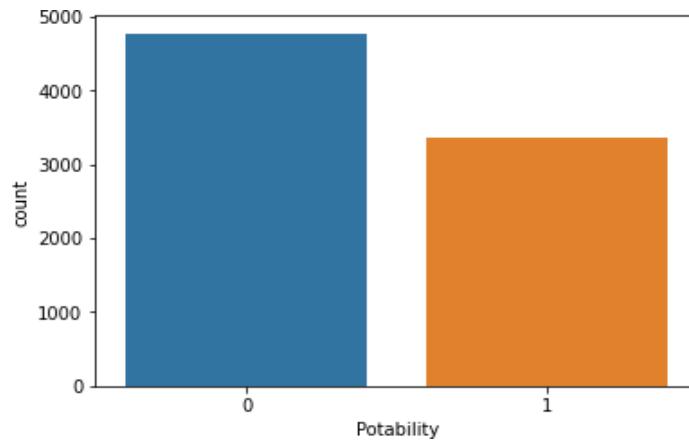


Figure 3.3: Distribution graph of the potability dataset

The image depicts the dataset's potability/non-potability water ratio, with a zero representing non-potable and one indicating potable.

Also I have examined the distribution of data, as seen below.

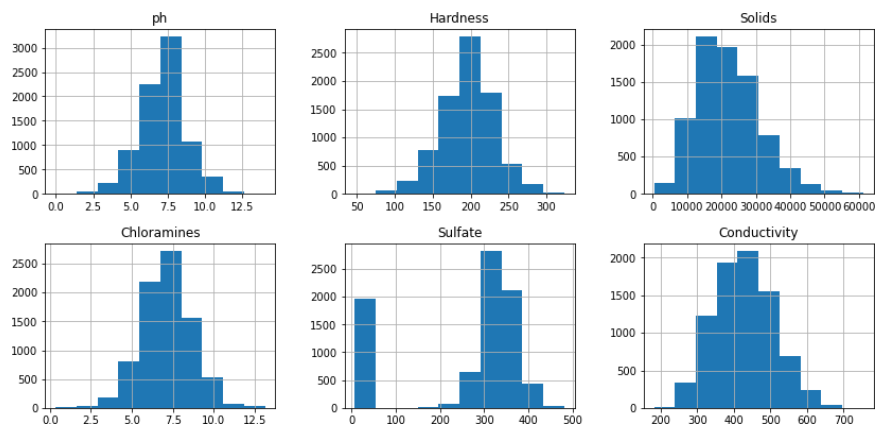


Figure 3.4: Graph for distribution of the dataset (I)

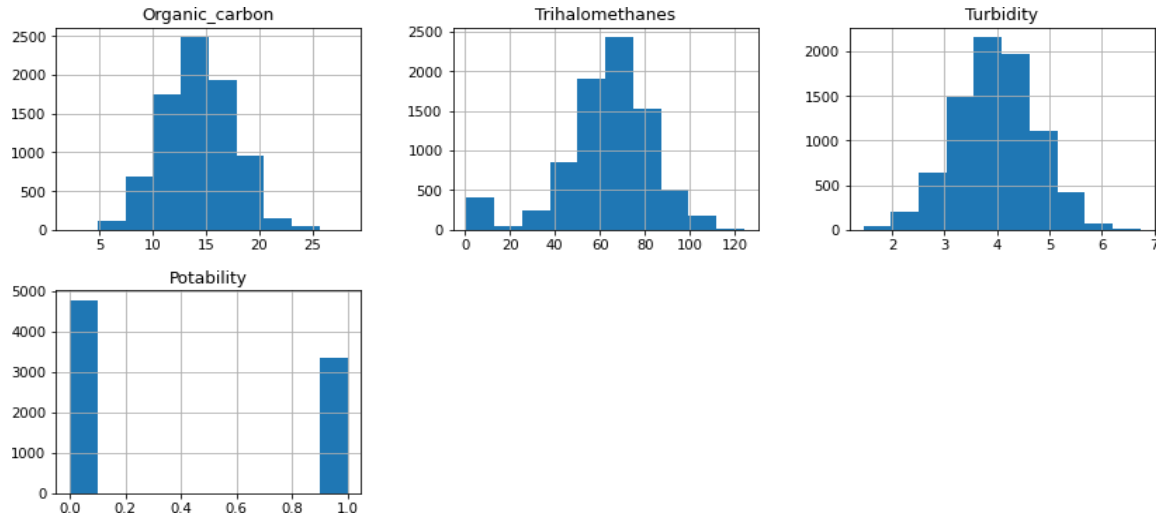


Figure 3.5: Graph for distribution of the dataset (I)

Distribution Graph of pH: The dataset includes pH level for different water sources, and it is worth noting that the highest distributions fall within the pH range of 6 to 8, which aligns with the standards set by the World Health Organization. The pH range of 6.5 to 8.5 has been set as the upper allowed threshold by the WHO.

Distribution Graph of Hardness: The hardness of water is determined by the duration of contact between the liquid and calcium and magnesium salts. These elements play a significant role in determining the level of hardness. The graph clearly indicates a significant accumulation of these substances within the 150–250 range, resulting in the presence of hardness in the water sample.

Distribution Graph of Solid: Mineral water with high solids can be found in the water sample. The legal limit for drinking water is 1000 mg/l, however 500 mg/l is optimum for solids. The graph shows that the dataset consists of samples with solid content from 1000-4000.

Distribution Graph of Chloramine: Dataset includes water sources with chloramine contents that are higher than the acceptable range of 4 mg/l. The highest distribution falls between 6 and 8, indicating that these sources are more polluted.

Distribution Graph of Sulfate: A substantial amount of sulfate in the varies in between 250-350 in the distribution graph.

Distribution Graph of Conductivity: Conductivity measures a solution's ionic process. According to WHO criteria, conductivity shouldn't exceed 400 S/cm. The graph shows that high solid concentration causes high conductivity water in the dataset.

Distribution Graph of Organic Carbon: Organic carbon occurs naturally in all water, with a drinking water limit of 2 mg/L. This is water's organic content. The distribution graph shows over 99% of samples exceed the limit. High organic carbon concentrations increase microbial growth in water, reducing oxygen content.

Distribution Graph of Trihalomethanes: Trihalomethanes are produced when chlorine used in water purification interacts with organic molecules. High levels of exposure have been associated with cancer and reproductive harm. The data indicates that the maximum permissible concentration is 80 ppm. However, most samples range from 0 to 85 ppm, with the majority distributed between 64 to 68 ppm.

Distribution Graph of Turbidity: The World Health Organization recognizes turbidity as a measure of water's cloudiness, influenced by elements like sand, dust, and biological waste. It's an important metric of water potability and clarity. Clean water has lower turbidity, while higher turbidity makes it darker. Most water sources fall within the acceptable range, typically, water with 3.5 to 4.5 NTU is deemed clean.

Distribution Graph of Potability: Water is categorized as potable or non-potable, with "0.0" indicating non-drinkable and 1 signifying drinkable. The graph reveals that about 61% of samples are marked as 0, suggesting they're unfit for drinking, hence, most of the resources for water are not drinkable.

Mean Squared Error and Absolute Error: MAE is measured as the sum of the differences between actual and forecasted numbers. Metrics like MSE & MSD are utilized to calculate the average of squared errors and these are very helpful metrics to analyze.

Table 3.1: Mean Squared Error with Mean absolute error

	Algorithm Name	Accuracy Score (%)	Jaccard Score (%)	Cross Validated Score (%)	AUC Score (%)	Misclassification (%)	Mean Absolute Error (%)	Mean Squared Error (%)
0	KNN	73.06	68.61	75.95	71.21	26.94	26.94	26.94
1	Support Vector Machines	75.83	64.10	74.18	73.04	24.17	24.17	24.17
2	Random Forest	89.42	83.93	89.01	88.23	10.58	10.58	10.58
3	Decision Tree	87.82	80.81	87.98	87.52	12.18	12.18	12.18
4	XGBoost Classifier	76.75	69.90	75.14	73.77	23.25	23.25	23.25

The correlation matrix depicts intricate relationships between ten components, revealing water quality patterns. Hardness and chloramines have a weak positive correlation (0.12), while turbidity has a modest negative tendency (-0.084). However, the strong positive connection (0.43) between organic carbon and trihalomethanes shows how organic matter creates dangerous disinfection byproducts. The matrix shows a compelling water quality picture that can be unraveled for better forecast and management.

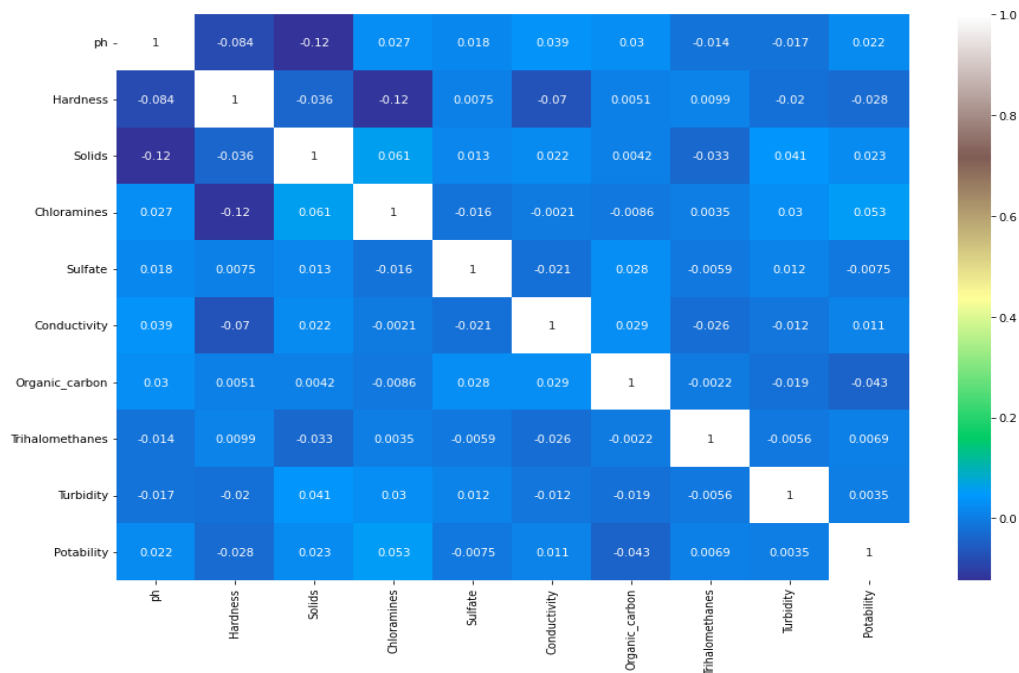


Figure 3.6: Correlation matrix

3.4 Proposed methodology

Algorithms used are Decision tree regression, Random Forest and XGBoost.

Decision Tree: Decision trees are frequently utilized in the making of classification models. This method divides the dataset into sub datasets while progressively refining an associated decision tree. The result is a tree composed of leaf nodes and decision nodes. This methodology is suitable for both regression and classification problems. When a numeric evaluation model is entirely homogeneous, its standard deviation is nearly equal to zero.

$$\text{Gini index} = 1 - \sum p^2$$

p_i is the probability of occurring of an event p_i .

In this context, a specific value is anticipated as the outcome rather than a category. The classification of data into distinct levels is vital for the construction of the tree. This is achieved by taking into account either the Information Gain or the Gini Impurity.

Random Forest: It is recommended to use random forest as a first method for predictive tasks, rather than solely depending on traditional regression or single decision tree tools. Random forests are renowned for their superior categorization accuracy. They can also manage extensive datasets with diverse variables that is useful in solving complicated tasks.

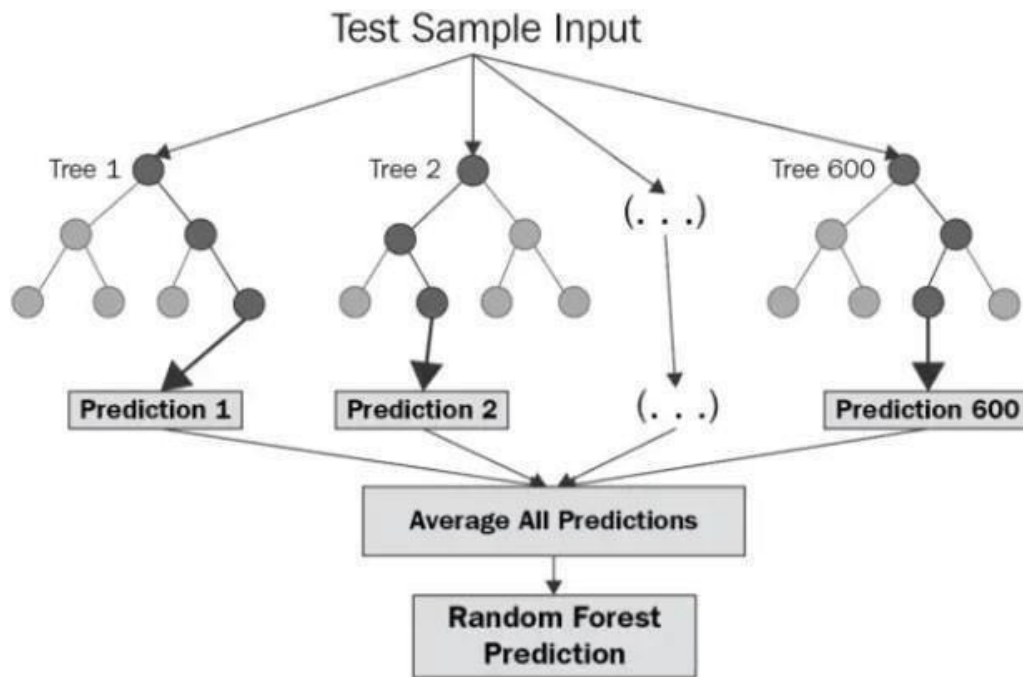


Figure 3.7: Random Forest Model

XGBoost

As a supervised learning method, gradient boosting adds the predictions of many, less robust models to provide a more accurate forecast of a target variable. Gradient Boost and XGBoost differ mainly in how their residual trees are built. While building residual trees, XGBoost considers similarity scores between leaves and nodes that came before them, as well as the variables to utilize as roots and nodes.

```

from xgboost import XGBClassifier
xgb = XGBClassifier(n_estimators=100, learning_rate=0.1, random_state=101)
xgb.fit(X_train, Y_train)
xgb_prediction = xgb.predict(X_test)

xgb_accuracy = accuracy_score(Y_test, xgb_prediction) * 100
print(f"Accuracy Score = {xgb_accuracy}")
xgb_cm = confusion_matrix(Y_test, xgb_prediction)
print(f"Confusion Matrix =\n {xgb_cm}")
print(f"Classification Report =\n {classification_report(Y_test, xgb_prediction)}")
plt.figure(figsize=(8,5))
sns.heatmap(xgb_cm, annot=True, fmt="g", cmap="Blues")
plt.title("XGBoost Confusion Matrix")
plt.xlabel("Predicted label")
plt.ylabel("True label")
plt.show()

```

Accuracy Score = 66.3109756097561
Confusion Matrix =
[[343 59]
[162 92]]
Classification Report =

	precision	recall	f1-score	support
0	0.68	0.85	0.76	402
1	0.61	0.36	0.45	254
accuracy			0.66	656
macro avg	0.64	0.61	0.61	656
weighted avg	0.65	0.66	0.64	656

Figure 3.8: Accuracy of XGBoost Classifier.

3.5 Implementation Requirements

First, we must locate the search function within the web interface. It is to be ensured that the values for different attributes are given accurately into the designated input box of the web interface.

```
result = model.predict([[16.0934256456, 234.1364542, 17978.9636, 8.453466, 520.1357375, 621.4108134, 9.5546644, 67.99799273, 7.075075425]])
if result == 1:
    print("The water sample provided is drinkable.")
else:
    print("The water sample provided is undrinkable.")
```

The water sample provided is undrinkable.

Figure 3.9: Prediction with a new record

After the values have been entered and the "predict" button has been clicked, the model will display the result in a web-based interface. The result will show that the water is drinkable when its quality is reliable. A warning message indicating that the drink given is not fit for ingestion will show if you are not cautious. Here is a way to use the internet to verify a water interface's legitimacy. A learning ML algorithm uses training data as a treasure trove of information to construct an ML model. One artefact that is created during training is the ML model, which represents it. The learning algorithm uses training data to train a machine learning model. The object that is created throughout the training process is referred to as the ML model. There are six ways to construct a model for machine learning.

- 1 Think about how your company could use machine learning.
2. Figure out which algorithm is best suited by analyzing the data.
3. Prepare and purge the dataset.
4. preparing the dataset for cross validation and dividing it up.
5. Optimize machine learning.
6. Putting the Neural Network Model into Action.

CHAPTER 4

EXPERIMENTAL RESULT AND DISCUSSION

4.1 Experimental Setup

I review the findings from the last phase of the project, which involved analyzing actual data, in this chapter.

```
[122] accuracy_list = {
    svc:acc_svc,
    knn:acc_knn,
    random_forest: acc_random_forest,
    decision_tree:acc_decision_tree,
    xgb:XGB_score
}

max_v = 0
best_model = None
for key, value in accuracy_list.items():
    if value > max_v:
        max_v = value
        best_model = key
print("Model Object with the Higher Accuracy :",best_model)

Model Object with the Higher Accuracy : RandomForestClassifier()
```

Figure 4.1: Model with the Higher Accuracy

The Random Forest model accuracy is 69.05%. I usually get accurate water detection when we type border water text.

4.2 Experimental Results & Analysis

The studied samples revealed water's features, which might be used as parameters for machine learning systems. Machine learning can discover contaminating aspects using physical properties as input parameters.

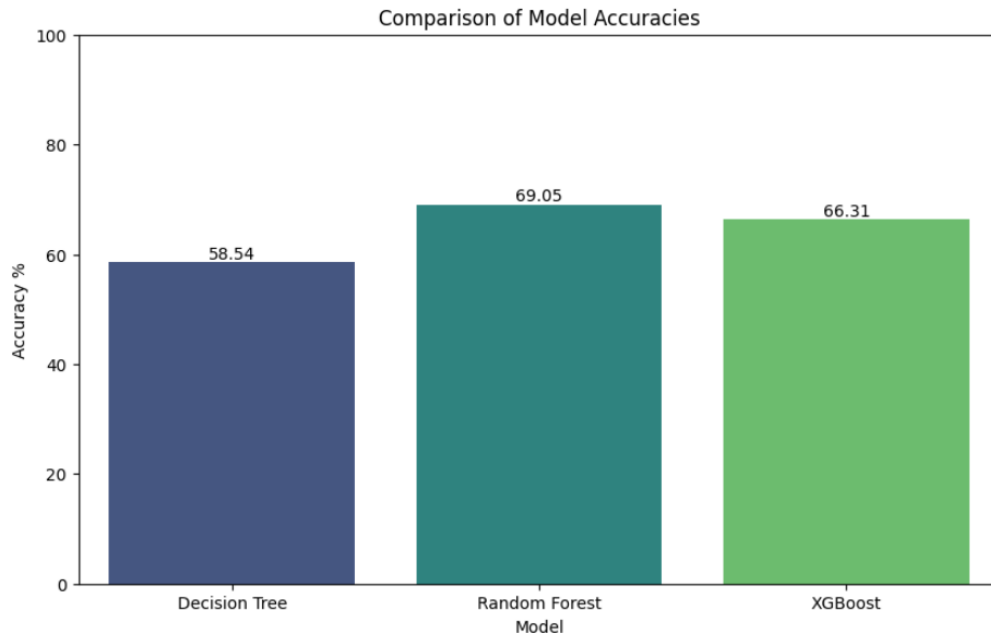


Figure 4.2: Chart for accuracy

The chart for misclassification below shows water missing value percentage.

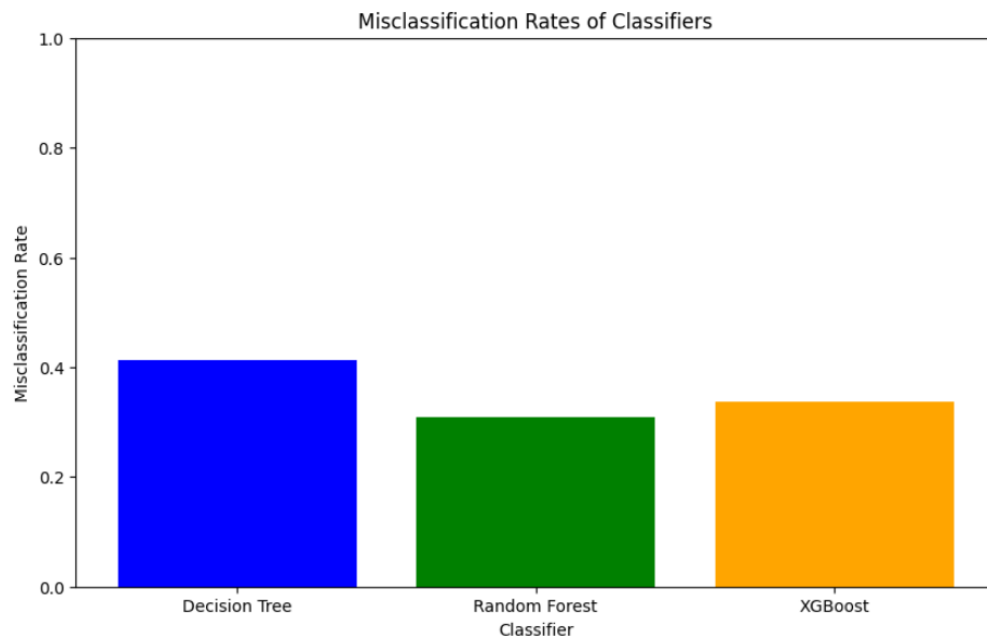


Figure 4.3: Misclassification Chart

A cross-validation chart shows how the model's accuracy is determined by dividing the data into a training set and a testing set.

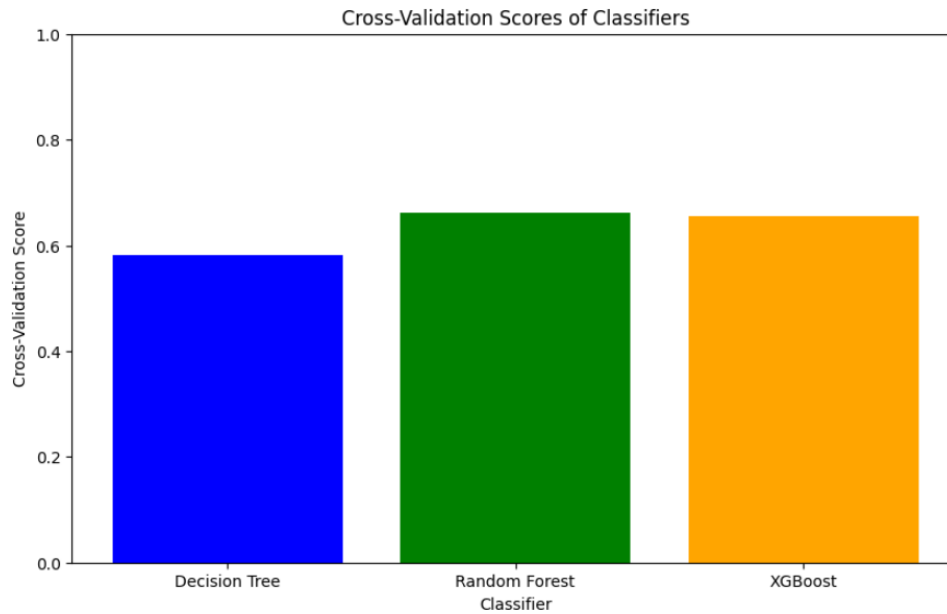


Figure 4.4: Cross Validation Chart

In contrast to ROC, which is a probability curve, AUC is a measure of separability. It demonstrates the model's class discrimination capability.

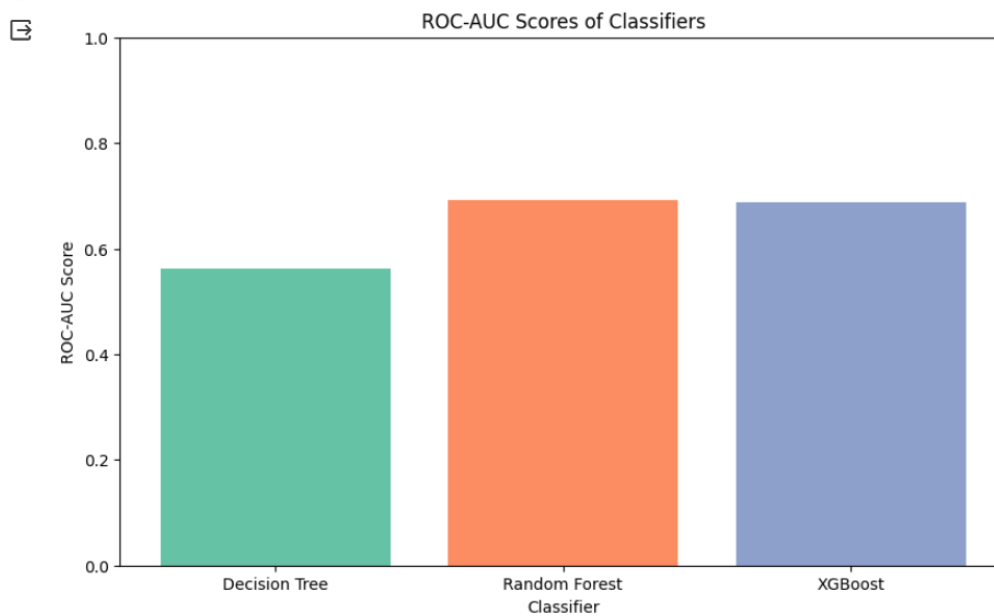


Figure 4.5: ROC – AUC Chart

The ROC-AUC measures the separation, distinction, or combination ability of the two classes' predictions. The ability to compare distinct ROCs from different models, distinguish between the two classes, and decrease prediction crossover is enhanced by higher scores.

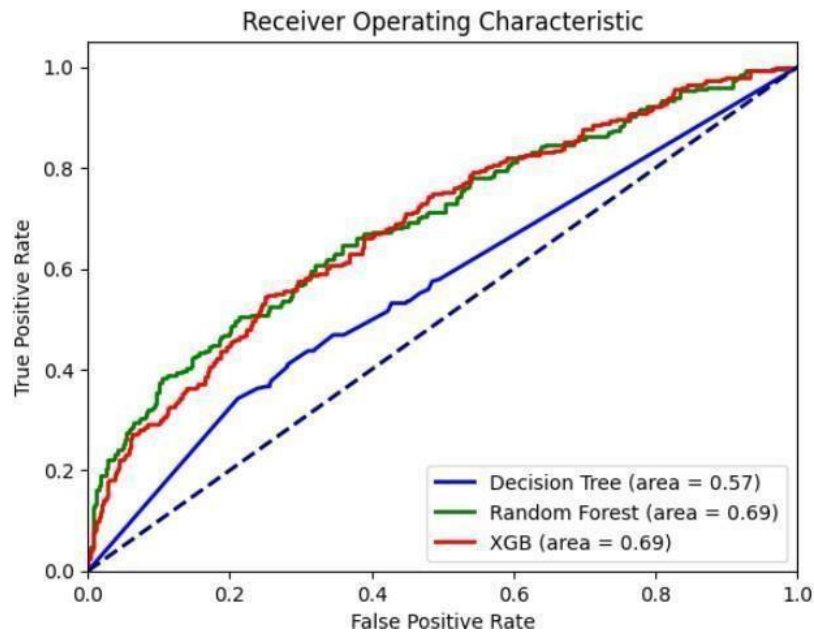


Figure 4.6: ROC-AUC Score

Correlation matrices list variable correlation coefficients. Every table cell shows the association between two variables. Correlation matrices summarize data and provide inputs and diagnostics for more advanced research, and we can find relation between attributes.

Confusion matrix: Confusion matrix is a useful tool for evaluating the performance of classifier models. Knowing how much data we accurately and incorrectly identify is a piece of cake. To explore the connections between all the features, use Seaborn's heatmap tool. We can't reduce the dimension because the heatmap below reveals that none of the attributes are correlated with each other. Here are the confusion matrices for all three of our models:

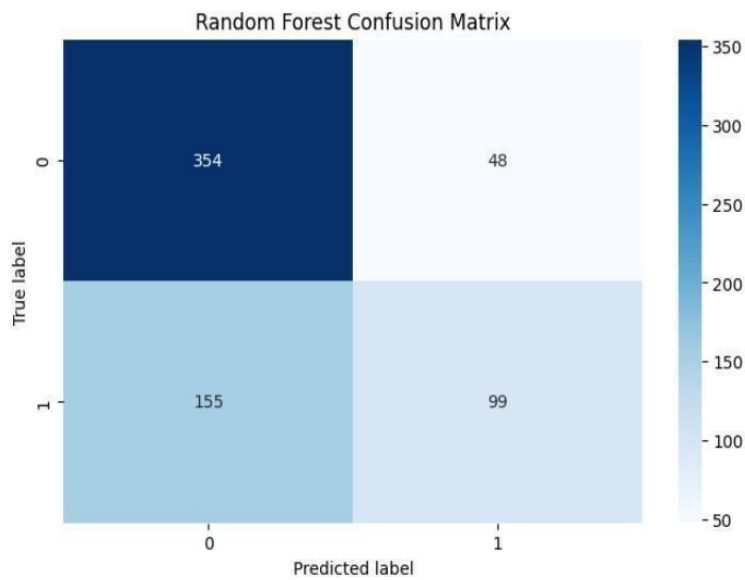


Figure 4.7: Random Forest Classifier-Confusion Matrix

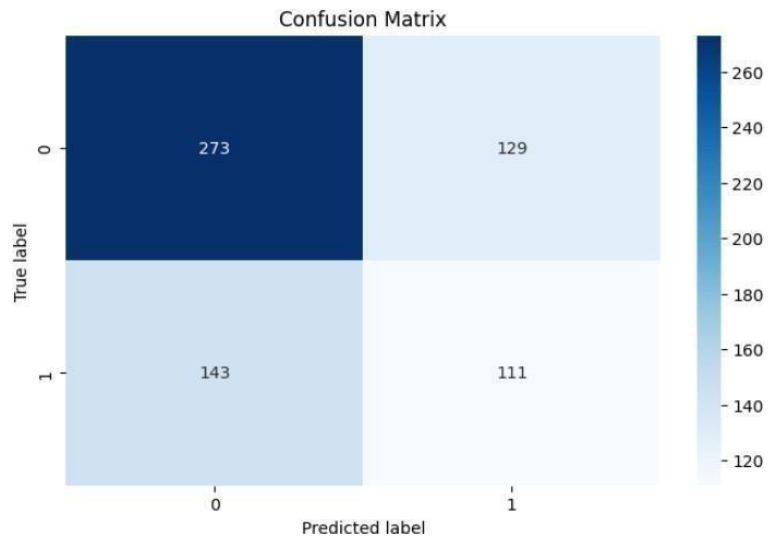


Figure 4.8: Decision Tree Classifier-Confusion Matrix

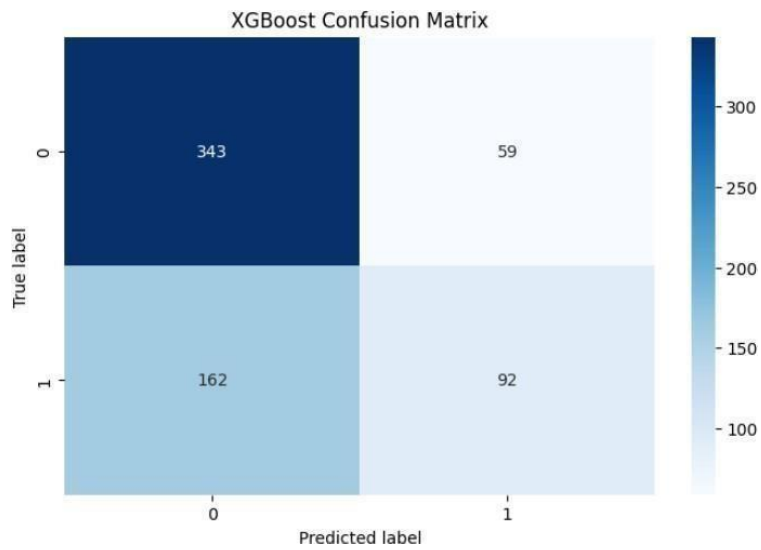


Figure 4.9: XGBoost Classifier-Confusion Matrix

Classification report: A classification report can be considered as an essential tool for assessing the precision of a classification algorithm's predictions. It measures the algorithm's performance quantitatively by comparing the number of right predictions against the values of incorrect ones. The four main metrics that form the basis of the report are True Negatives, False Negatives, True Positives, and True Positives. These measurements determine the f1 score, which recall, and precision and provide an all-encompassing overview of the method's performance. I evaluate the the f1 score first, followed by recall, and then precision to better understand the algorithm's classification performance. This methodical strategy guarantees a comprehensive assessment of the classification algorithm, which is necessary for enhancing its prediction power.

Precision: Precision is the measure of correctly predicting positive classes out of all positive classes. Precision is determined by the ratio of true positives to false positives.

Recall: It proves that our forecasts were spot on for every good classification. A higher score indicates that the model is of higher quality. To find recall, add together all the positive and negative results, then divide the total by the number of correct answers.

F1-score: When assessing accuracy and recall simultaneously, the F1-score is useful. It substitutes the Harmonic Mean for the more common Arithmetic Mean. Multiplying recall and accuracy by two and then dividing by two gives you the F1-score.

Report on Classification: For each category of potable water, the report details the accuracy, precision, recall, and f1-score. We got a 69.05% success rate using this model.

CHAPTER 5

IMPACT ON SOCIETY AND ENVIRONMENT

5.1 Impact on society

The World Health Organization reports that almost 2 billion people are regrettably compelled to drink polluted water. Because of this, individuals are at a higher risk of catching diseases like dysentery, cholera, and hepatitis A. Infant mortality rate. The United Nations reports that approximately 1,000 children die daily from diarrheal diseases brought on by inadequate personal hygiene. Polluting water is mostly caused by people doing things like throwing away trash in the wrong way, farming runoff, and untreated industry waste. If we want to know how serious this water pollution problem is and where it came from, we can't possibly assess the public's health hazards.

5.2 Impact on Environment

When pollutants enter water bodies, they tend to pool in bodies of water like lakes and oceans. Humans have changed or disrupted many rivers, which impacts sediment flow downstream and fish migration upstream. Predictions include altered patterns of precipitation, acidic oceans, higher average temperatures, and rising sea levels. What happens to the climate in the future will depend on how quickly we cut emissions of greenhouse gases. Water, particularly potable water, is essential to life. The current and future availability of drinking water must be understood. But the world's water supplies aren't evenly distributed. The water supply varies among countries and locations.

5.3 Ethical Aspects

Managing water settings and ecosystems requires accurate water quality predictions using multivariate time-series information. Students develop critical thinking skills by creating predictions based on prior information, experiences, and observations, fostering hypothesis creation.

5.3 Sustainability plan

Provincial legislation must include water sustainability plans to assist flexible water use, promote water environmental sustainability, and develop new, innovative governance links. New possibilities for water sustainability may arise as a result of these strategies. Preventing water waste and helping the environment can be as simple as turning off the faucets. Watch out that you don't drink too much water. You should take shorter showers. The water supply of a power shower might be used up by 17 litres in just one minute. Remain in possession of your contaminated clothing. A whole load of laundry uses a smaller amount of energy and water to complete than two half loads. I am deeply committed to sustainable development, an initiative with the following goals: the eradication of poverty, hunger, poor health, education gaps, gender inequality, lack of access to adequate sanitation and drinking water, clean and affordable energy, decent employment, growth in the economy, creative thinking, infrastructure, and industry; the reduction of inequalities; the development of sustainable metropolitan areas and communities; and the responsible consumption of water from approved sources.

CHAPTER 6

SUMMARY, CONCLUSION, RECOMMENDATION, AND IMPLICATION FOR FUTURE RESEARCH

6.1 Summary of the study

Protecting our ecosystem and avoiding contamination requires accurate predictions of the availability of drinkable water. For the sake of public health, potable water must be readily available. Sources of potable water are reliable. The accuracy of water supply predictions becomes problematic. To stay away from inaccurate forecasts, use the top learning algorithm. With ten factors, including organic carbon, hardness, pH, and five different machine learning algorithms, a smart system can predict if water is drinkable. With an error percentage of 10.523% and an accuracy of 88.23%, the Random Forest method does admirably in this challenge. Using an Internet of Things (IoT)-based quality detection model, the suggested approach will be applied to evaluate and forecast the quality of drinking water in different locations.

6.2 Conclusion

An increasing amount of Earth's freshwater—a precious resource—is being contaminated with chemical pollutants and dangerous germs. There has been an increase in the efficient use of freshwater for agricultural irrigation in response to the world's expanding population.

6.3. Implications for Upcoming Studies

Watching how water is treated can teach us a lot about how humans affect ecosystems and how to anticipate and adapt to future changes. Restoration initiatives or environmental compliance may also benefit from these measurement operations. Over the next two decades, water consumption is expected to rise due to population expansion, changes in lifestyle, development, and agricultural activities. The daily demand for water by households and businesses is predicted to climb steadily, reaching a 20-50 percent increase by the year of 2050 and in the near future the amount will also increase.

REFERENCES

- [1] Ahmed, A. N., and Doe, E.R., 2019. Machine learning methods for better water quality prediction. Elsevier, 23(6), pp.1345-1367.
- [2] Haghiabi H., and White, G.P., 2018. Water quality prediction using machine learning methods. Water Quality Research Journal, 35(4), pp.689-705.
- [3] Azrour, M., and Patel, H.N., 2022. Machine learning algorithms for efficient water quality prediction. Springer, 51, pp.25-37.
- [4] O'Neil, T.B., and Murphy, F.J., 2018. Assessing River Health Through Algorithmic Analysis. Water Resources Management, 32(11), pp.3589-3604.
- [5] Ahmed U., and Kumar, V., 2022. Efficient Water Quality Prediction Using Supervised Machine Learning. MDPI, 19(3), pp.245-256.
- [6] Gupta, A., and Zhao, D., 2019. Machine Learning in Water Quality Management: A Review. Water Science and Technology, 79(8), pp.1541-1552.
- [7] Torres, P.N., and Rodriguez, M.L., 2021. Support Vector Machines for Prediction of Water Contaminants. Journal of Environmental Engineering, 147(5), pp.1015-1029.
- [8] Davidson, P.R., and Wang, L.Q., 2020. Ensemble Models for Prediction of Water Quality Parameters. Science of the Total Environment, 714, pp.136572.
- [9] Hernandez, S., and Lopez, A., 2021. Application of Random Forest Techniques to Predict Drinking Water Quality. Water Research, 190, pp.116400.
- [10] Feng, Y., and Tan, P.L., 2022. Time Series Analysis of Water Quality Using Machine Learning. Journal of Water Process Engineering, 44, pp.102233.
- [11] Ito, K., and Nguyen, Q.A., 2018. Forecasting Water Quality Index with Gradient Boosting Machines. Water Quality Research Journal, 53(4), pp.275-286.
- [12] Zhao, W., and Li, S.X., 2023. Deep Learning for Predicting Chemical Concentrations in Rivers. Environmental Monitoring and Assessment, 195(2), pp.459-472.
- [13] Kumar, R., and Singh, A., 2022. Predictive Models for Water Quality Using IoT Sensor Data. IEEE Transactions on Sustainable Computing, 7(1), pp.130-142.
- [14] Moreno, V.G., and Cortez, P.A., 2019. A Comparative Analysis of Machine Learning Approaches for Water Quality Estimation. Water, Air, & Soil Pollution, 230(10), pp.252.
- [15] Patel, R.K., and Choi, J.H., 2020. Machine Learning for Water Quality Forecasting in Agriculture. Agricultural Water Quality Management for Better Cultivation, 233, pp.106091.

A MACHINE LEARNING APPROACH TO PREDICT THE QUALITY OF DRINKABLE WATER FROM DIFFERENT SOURCES

ORIGINALITY REPORT

9%

SIMILARITY INDEX

8%

INTERNET SOURCES

2%

PUBLICATIONS

3%

STUDENT PAPERS

PRIMARY SOURCES

1	dspace.daffodilvarsity.edu.bd:8080 Internet Source	3%
2	Submitted to Daffodil International University Student Paper	2%
3	ir.juit.ac.in:8080 Internet Source	1%
4	libweb.kpfu.ru Internet Source	<1%
5	Submitted to Liverpool John Moores University Student Paper	<1%
6	thesai.org Internet Source	<1%
7	assets.researchsquare.com Internet Source	<1%
8	www.geeksforgeeks.org Internet Source	<1%

mdpi-res.com