

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

BY

**KHADIZA AKTER RANU
183-15-1008**

This Report Presented in Partial Fulfillment of the Requirements for the
Degree of Bachelor of Science 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

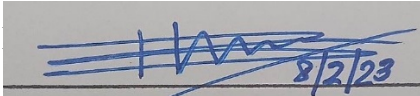
DHAKA, BANGLADESH

FEBRUARY 2023

APPROVAL

This Project titled “**A machine learning approach to predict the quality of drinkable water from different sources**”, submitted by Khadiza Akter Ranu, Id: 183-15-1008 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 05 February 2023.

BOARD OF EXAMINERS

A rectangular box containing a blue ink signature and the date '8/2/23' written in blue ink.

Professor and Head

Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Chairman

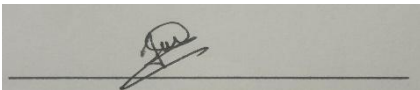
A rectangular box containing a black ink signature.

Nazmun Nessa Moon

Associate Professor

Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner

A rectangular box containing a black ink signature.

Raja Tariqul Hasan Tusher

Assistant Professor

Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner

A rectangular box containing a black ink signature.

Dr. Ahmed Wasif Reza

Professor

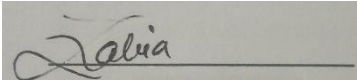
Department of Computer Science and Engineering
East West University

External Examiner

DECLARATION

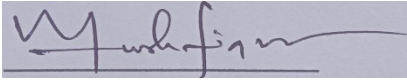
We hereby declare that, this project has been done by us under the supervision of **Zakia Sultana**, Senior Lecturer, Department of CSE Daffodil International University. We also declare that neither this project nor any part of this project has been submitted elsewhere for award of any degree or diploma.

Supervised by:



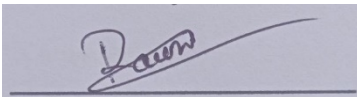
Zakia Sultana
Senior Lecturer
Department of CSE
Daffodil International University

Co-Supervised by:



Mushfiqur Rahman
Senior Lecturer
Department of CSE
Daffodil International University

Submitted by:



Khadiza Akter Ranu
183-15-1008
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

First, we express our heartiest thanks and gratefulness to almighty God for His divine blessing makes us possible to complete the final year Research paper successfully.

We really grateful and wish our profound our indebtedness to **Zakia Sultana**, Senior Lecturer, Department of CSE Daffodil International University, Dhaka. Deep Knowledge & keen interest of our supervisor in the field of “*Data mining*” to carry out this project. His endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior drafts and correcting them at all stage have made it possible to complete this project.

We would like to express our heartiest gratitude to **Dr. Touhid Bhuiyan**, Professor and Head, Department of CSE and **Mushfiqur Rahman**, Senior Lecturer, Department of CSE, for her kind help to finish our project and also to other faculty members and the staff of Daffodil International University.

We would like to thank our entire course mate in Daffodil International University, who took part in this discuss while completing the course work.

Finally, we must acknowledge with due respect the constant support and patients of our parents.

ABSTRACT

Machine learning is now a crucial technique for data analysis, classification, and prediction due to the exponential growth in the amount of data available on the aquatic environment. Data-driven models based on machine learning have the ability to effectively tackle more complicated nonlinear problems, in contrast to conventional models utilized in water-related research. Models and findings from machine learning have been used in water environment research to build, monitor, simulate, evaluate, and optimize various water treatment and management systems. Machine learning can also offer solutions for reducing water pollution, enhancing water quality, and managing the security of the watershed environment. In this paper, we explain the use of machine learning algorithms to assess the water quality in various water contexts, including surface water, groundwater, drinking water, sewage, and others. We also suggest potential future uses of machine learning techniques in aquatic contexts. For forecasting the potability of water, we employ the KNN, SVM, Random Forest, Decision Tree, and XGBoost algorithms. Pre-trained KNN algorithms were employed.

TABLE OF CONTENTS

CONTENTS	PAGE
Board of examiners	ii
Declaration	iii
Acknowledgements	iv
Abstract	v
CHAPTER	
Chapter 1: Introduction	1-4
1.1 Introduction	1
1.2 Motivation	2
1.3 Rationale of study	2-3
1.4 Research question	3
1.5 Expected Output	3
1.6 Project Management and Finance	4
1.7 Report Layout	4
Chapter 2: Background Study	5-8
2.1 Preliminaries	5
2.2 Related Works	5-6
2.3 Comparative Analysis	6-7
2.4 Scope of the problem	7-8
2.5 Challenges	8
Chapter 3: Research Methodology	9-20

3.1 Research Subject and Instrumentation	9
3.2 Data Collection Procedure	9-10
3.3 Statistical Analysis	11-17
3.4 Proposed Methodology	17-19
3.5 Implementation Requirements	20
Chapter 4: Experimental Result and Discussion	21-25
4.1 Experimental Setup	21
4.2 Experimental Results & Analysis	21-25
Chapter 5: Impact on Society, Environment and Sustainability	26-27
5.1 Impact on society	26
5.2 Impact on Environment	26
5.3 Ethical Aspects	26-27
5.4 Sustainability plan	27
Chapter 6: Summary, Conclusion, Recommendation, and Implication for Future Research	28
6.1 Summary of study	28
6.2 Conclusion	28
6.3 Implication for Further Study	28
REFERENCES	29

List of Figures**PAGE**

2.3.1 Comparison of phyco-chemical properties	7
3.2.1 Dataset	10
3.2.2 Preprocessing data design	10
3.3.1 Potability distribution graph of the dataset	11
3.3.2 Distribution graphs of the features of the dataset (1)	11
3.3.3 Distribution graphs of the features of the dataset (2)	12
3.3.4 Confusion matrix for K- Nearest Neighbors	14
3.3.5 Confusion matrix for Support Vector Machine	15
3.3.6 Confusion matrix for Random Forest	15
3.3.7 Confusion matrix for Decission Tree	16
3.3.8 Confusion matrix for XGBoost	16
3.12 Random Forest	19
4.2.1 Accuracy chart	22
4.2.2 Misclassification Chart	22
4.2.3 Cross Validation Chart	23
4.2.4 ROC – AUC Chart	23
4.2.5 ROC-AUC Score	24
4.2.6 Correlations matrix	25

List of Tables**PAGE**

3.3.1 Mean Absolute Error & Mean Squared Error	14
4.2.1 Confusion matrix content	25

CHAPTER 1

INTRODUCTION

1.1 Introduction

As part of our endeavor to see if anything might be recognized by a programming language, I evaluated a small class of diseases. We discovered that utilizing the computer language known as machine learning to identify is rather simple. The yield will increase as I deal with a lot more varieties of potable water in the future. I'll utilize these courses to determine our proportion of detectable water other factors that contribute to water quality decline include human waste (plastics), unwanted items in nearby ponds, lakes, and seas, as well as plastics and other unwanted debris that can lead to harmful events. As a result of all these factors, water quality is declining now. It is particularly true in healthcare facilities, where a lack of water, poor hygiene, and inadequate cleaning expose patients and employees to viruses and bacteria. Only in lower places do the percentages go significantly higher, although 15% of patients worldwide contract a virus during their hospital stay. In India, however, 70% of the water that is accessible has been contaminated by domestic and industrial contaminants. The most important source of life, water is essential to the survival of most living things, including humans. To continue to exist, living things require water that is of sufficient quality. As part of our investigation into whether anything may be identified by a computer language, I evaluated a small class of diseases. We discovered that utilizing machine learning, a type of computer language, to identify objects is rather simple. The yield will rise as a result of my future work with numerous additional varieties of potable water. My calculations for our proportion of measurable water will be based on these courses. For example, irrigation water must be neither too saline nor contain toxic materials that can be transferred to plants or soil and thus destroying the ecosystems. Water quality for industrial uses also requires different properties based on the specific industrial processes. Some of the low-priced resources of fresh water, such as ground and surface water, are natural water resources. However, such resources can be polluted by human/industrial activities and other natural processes.

1.2 Motivation

The major quality goal of this study is to evaluate water quality using machine learning techniques. To measure the quality of water potability concepts is applied. The overall potability of the water was assessed in this study using the following water quality metrics. The factors were turbidity, trihalomethanes, organic carbon, conductivity, pH, hardness, solids, chloramines and solids. These parameters are utilized as a feature vector to depict the water quality. If an IoT-based device could be developed to check potability, it would be really useful to people.

1.3 Rationale of study

The ecosystem and the general public's health are directly impacted by water quality. Water is utilized for many purposes, including drinking, farming, and industrial. They claimed that while the MLP model was only marginally more precise, all applied models had adequate performance for predicting water quality components. It controlled a water supply system's water quality.

They saw this as an optimization challenge, and they used modern optimization methods to solve it. A review of the literature indicates that in order to develop water conservation initiatives, it is essential to forecast and evaluate the quality of the water. Thus, artificial intelligence-based solutions have been offered. As a result, the Tireh River, one of the main rivers in the Dez basin, and its water quality components were researched (one of the major catchments in Iran). Water is one of the most important ingredients for life. Everywhere there are pressing issues with drinking water accessibility and safety. It may not be healthy to drink water that has been contaminated with pathogenic organisms, hazardous chemicals, and other pollutants. Over time, concerns about water quality and protection have spread around the globe. Fresh water resources make up about 2% of the world's total water resources, and even these are becoming contaminated by human activities. This could potentially minimize the effect that tainted water has on individuals. Users can link a wide range of sensors and equipment to the Internet thanks to a well-liked technology known as the Internet of Things (IoT). The WHO (World Health Organization) (World Health Organ

ization). The actual findings were compared to these reference levels, and the user was alerted when any parameter reached its limit before the water became tainted.

1.4 Research Questions

- 1 What Is the U.S. Water Quality Like Compared to the Rest of the World?
2. Who Controls the Water We Drink? 2
3. What Kinds of Substances Are in Water?
4. Which Contaminants Could Be Present in Ground Water?
5. What Is In Municipal Water, Number?
6. Exactly how can lead enter drinking water?

1.5 Expected Output

We're looking for signs of water. Predicting water quality accurately is essential for managing the water environment and protecting the water environment. Multivariate time-series datasets are used to provide information on water quality. Due to population growth, changing lifestyles, development, and agricultural activities, there will be an increase in water demand during the next 20 years. The industrial and domestic sectors are anticipated to use water at rates 20 to 50% greater than current levels by 2050. If current trends continue, the difference between the world's water supply and demand is predicted to increase to 40% by 2030. Demand in many areas already exceeds sustainable supply, while water scarcity in other areas is impeding economic progress.

I tested a tiny class of illnesses as part of our effort to determine whether anything might be recognized by a programming language. We found that identifying using the computer language known as machine learning is rather straightforward. My future work with many more kinds of drinkable water will increase the yield. These classes will be used by me to calculate our percentage of detectable water.

1.6 Project Management and Finance

Water management is the planning, production, distribution, and control of the use of water resources using a range of IoT technologies. Transparency is encouraged by these technologies, which also make it possible for more palatable and environmentally friendly resource use. Water resources management is the planning, development, and management of water resources in terms of water quantity and quality for all water applications (WRM). It is made up of the institutions, infrastructure, financial assistance schemes, and information management systems that support and guide water management.

Aspects of finance include setting a budget, managing money, administering grants, paying bills, and balancing accounts. A top-notch financial model of the water treatment plant that enables the financial team to predict how the project will respond to changing circumstances is the cornerstone of the project's future success.

1.7 Report Layout

- The main ideas of "A machine learning model for predicting water sickness" were presented in chapter 1 together with the objective, target, and anticipated outcomes of our work.
- The synopsis, the scope of the problem, and the challenges are the main topics of the related works section of Chapter 2.
- In Chapter 3, the research approach is covered.
- Details on the outcomes of the experiments are provided in Chapter 4.
- Chapter 5 covers a variety of topics, including how society affects the environment.
- The evaluation results are summarized in Chapter 6 along with a few additional details that may assist future publications more accurately reflect my research efforts.

CHAPTER 2 BACKGROUND

2.1 Preliminaries

This provided the motivation for creating a simple-to-understand descriptive language. The fundamental problem with descriptive language is that it can lead to extremely laborious and drawn-out descriptions of complex systems that could be stated more succinctly via metaphors. In the three months preceding the Haicheng earthquake, we examined the spatial and temporal distributions of roughly 570 reports of changes in ground water and 670 reports of anomalous animal behavior. The reported range of these changes and abnormalities was more than 150 kilometers away from the epicenter, with no concentration nearby. There are indications that (1) there is a spatial and temporal correlation between the two types of anomalies, with ground water changes occurring one to two days prior to abnormal animal behavior, (2) there is a higher concentration of reports close to significant active faults than elsewhere, and (3) the region where frequent observations were made may change over time. On February 1, 1975, the first known foreshock occurred, and that day there was a sharp increase in the number of observed modifications in ground water.

Since the initial foreshock that was significant enough to be felt on February 3rd, there have been several accounts of unusual animal behavior. These results suggest that some animals may have responded to foreshock-induced ground tremor, while others may have noticed changes in the ground water (level, composition, or other properties).

2.2 Related works

It describes an algorithm's output after it has finished training on a batch of old data. It is feasible to anticipate the outcome of a new data set using prediction after studying the historical data set. For unknown variables, the algorithm generates likely values. KNN, Support Vector Machines, Random Forest, Decision Tree, and XGBoost are five algorithms used in investigations. The results showed that the required result may be obtained using a simple classifier, such as KNN.

A supervised machine learning model is the KNN algorithm. In other words, it makes predictions about a target variable based on one or more independent factors. Read *K-Means Clustering in Python: A Practical Guide* to find out more about unsupervised machine learning methods. Good, easily comprehensible forecasts are produced by random forest. Large datasets can be handled effectively. In comparison to the decision tree method, the random forest algorithm offers a higher level of accuracy in outcome prediction.

An approach for supervised machine learning called the Support Vector Machine (SVM) is utilized for both classification and regression. Even if we also refer to regression issues, classification is the best fit. The SVM algorithm's goal is to locate a hyper plane in an N-dimensional space that clearly categorizes the data points.

2.3 Comparative Analysis

The "biochemical oxygen demand" is the quantity of oxygen required for the aerobic biochemical breakdown and transformation of organic molecules in waste water by bacteria and other microorganisms (Dara 2002). One of the most important elements for a body of water is BOD. 109 to 163 mg/L for the Buriganga River and 102 to 149 mg/L for the Balu River were the BOD values discovered in the current study (Table 1). Average levels of this parameter in the research region's Buriganga and Balu Rivers were higher than the DoE norm (50 mg/L), at 135 mg/L and 118 mg/L, respectively (DoE, 2003). The findings in Table 1 demonstrate that BOD value is a reflection of COD value. It should be noted that the COD value reveals the quantity of both biodegradable and non-biodegradable contaminants in a body of water in the Buriganga and Balu River research areas, the average COD values were 275 mg/L and 199 mg/L, respectively. In the case of the Buriganga River, COD readings at four spots gradually fell downstream, exposing the most culpable sources' locations upstream. River surface water typically has a BOD value between 1 mg/L and 8 mg/L. (Retrieved 2016). According to Tasfina et al. (2016), the average BOD value of the water from the Buriganga River was 82.8 mg/L.

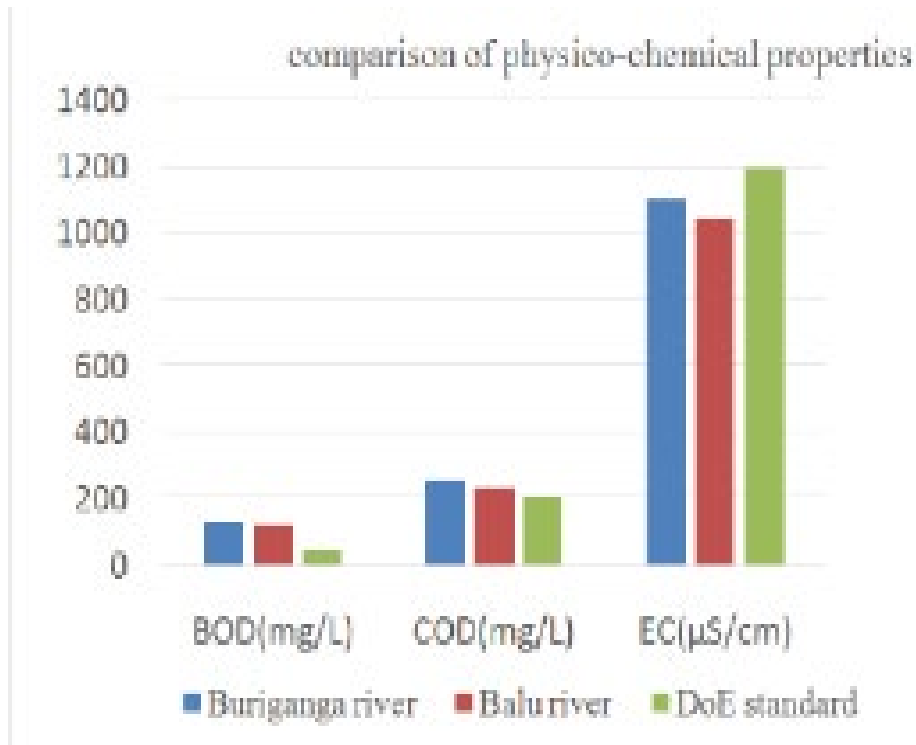


Figure 2.3.1: Comparison of physico-chemical properties

The aforementioned correlation index revealed the BOD and COD to have excellent relationships. Sewage flows from Dhaka and tannery wastes harm the Buriganga River. Organic pollutant load dominated in this aspect. Given that both COD and BOD are indicators of organic compounds, Table 3 shows that BOD and COD have a significant correlation (0.973).

2.4 Score of the problem

Human civilization is built on the principle of water. For the purposes of agriculture, sanitation, and drinking as well as water purification for the protection of the environment and of public health, communities have grown throughout history through increasing access to clean water. The observations may alter with time.

Major Causes of Water Crisis

Pollution of the water. Due to poor sanitation and a lack of waste treatment facilities, the majority of water sources in rural areas are extremely filthy.

- i. Over drafting of groundwater
- ii. Abuse and overuse of water

- iii. Disease
- iv. Climate change
- v. Mismanagement
- vi. Human habitations.
- vii. Corruption

2.5 Challenges

Globally, wetlands have shrunk by more than 50%. Agriculture uses more water than any other business, but a lot of it is wasted due to inefficiencies. Weather and water patterns are changing as a result of climate change, and some regions are experiencing shortages and droughts while others are experiencing floods. 90% of the sewage in underdeveloped nations is dumped directly into water sources. 2 million tonnes of sewage and other effluents are dumped into the ocean every day. 3. The industry annually releases 300-400 megatonnes of waste into rivers.

CHAPTER 3

RESEARCH METHODOLOGY

3.1 Research Subject and Instrument

Our study is focused on predicting the potability of potable water using machine learning, as I already mentioned. We make use of Google Colab software for programming. In Google Colab, we applied machine learning to the DenseNet201 model. We found seven categories of water potability DenseNet201 model to be fairly accurate. [23] We can precisely assess whether the sample water is fit for human consumption. Several processes are taken into consideration in this session, starting with data pre-processing, noise reduction, and transfer to the proper format for training Tensor Flow After the acquisition phase, look over the validated images and send them for testing. And we intermittently show the test image that we employ. Python was the language we used in our research. We employ KNN, SVM, Random Forest, Decision Tree, and XGBoost. Pre-trained KNN algorithms were employed.

For coding, we used Google Colab. For the depiction of the confusion matrix and graph, we used the histogram together with the Matplotlib and pandas libraries.

3.2 Data Collection Procedure

A Kaggle dataset was used in this investigation. There were 8127 samples in all. The dataset contains a number of significant measures, including trihalomethanes, conductivity, organic carbon, pH, hardness, solids, chloramines, and sulfate. The International Water Association's standard data rate ensures the purity of the drinking water in Bangladesh.

	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	9.275863603		418.6062131	16.86663693	66.42009251	3.05593375	0
5	8.316765884	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	5.584086636	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	181.2596172	16540.97905	7.022499179	309.2388651	487.6927878	13.228441		4.333952698	1
13		134.7368557	9000.025591	9.026292723		428.213987	8.666672182	74.77339241	3.699558048	1
14	3.629922065	244.1873915	24856.63321	6.618071066	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	15.48302674	51.53586708	4.013338801	1
17	5.893103408	239.2694815	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 3.2.1: Dataset

For improving data quality, the calculation phase of data processing is essential. The most important properties of the dataset are explored in this step to determine data exploration and feature scaling. Afterward, the water samples were categorized into categories based on their WQI ratings. My dataset contains 8127 water samples. I might show you some test water in figure 2 for your viewing pleasure. It includes water, whether it is drinkable or not.

The number of trainings, testing data for each classification:

Train set: (6501, 9) (6501)

Test set: (1626, 9) (1626)

Pre-processing data:

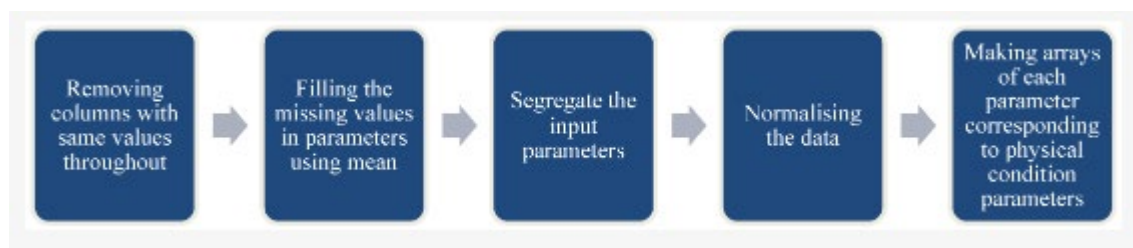


Figure 3.2.2: Preprocessing data design

3.3 Statistical Analysis

We employed pre-processing to predict pH and Hardness. Solids Chloramines Conductivity of Sulfates Organic carbon Trihalomethanes Turbidity Potency, etc. Physical parameters—those that cannot be changed—are distinguished from modifiable parameters. Then show the potability (where 0 is not drinkable, 1 is drinkable)

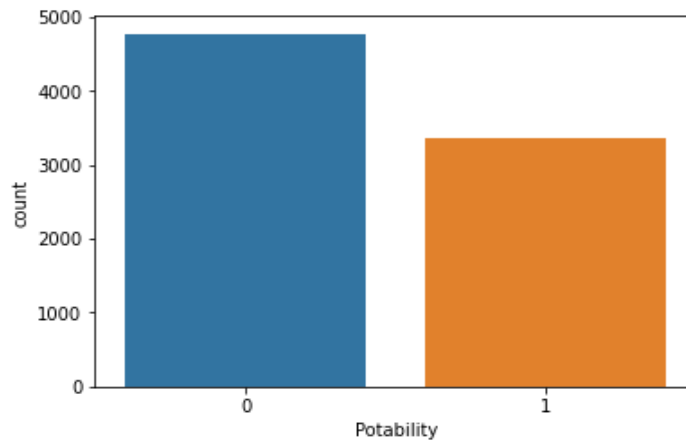


Figure 3.3.1: Potability distribution graph of the dataset

This graph displays the dataset's ratio of potable to non-drinkable water, with 0.0 designating non-potable water and 1.0 designating potable water.

This step's goal is to examine the data's distribution, as demonstrated below.

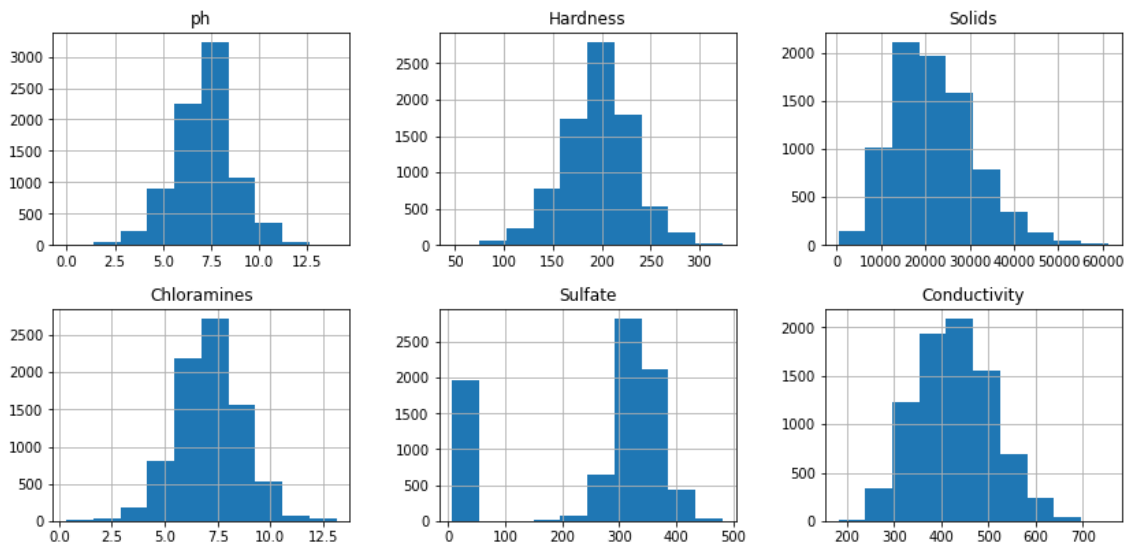


Figure 3.3.2: Distribution graphs of the features dataset (1)

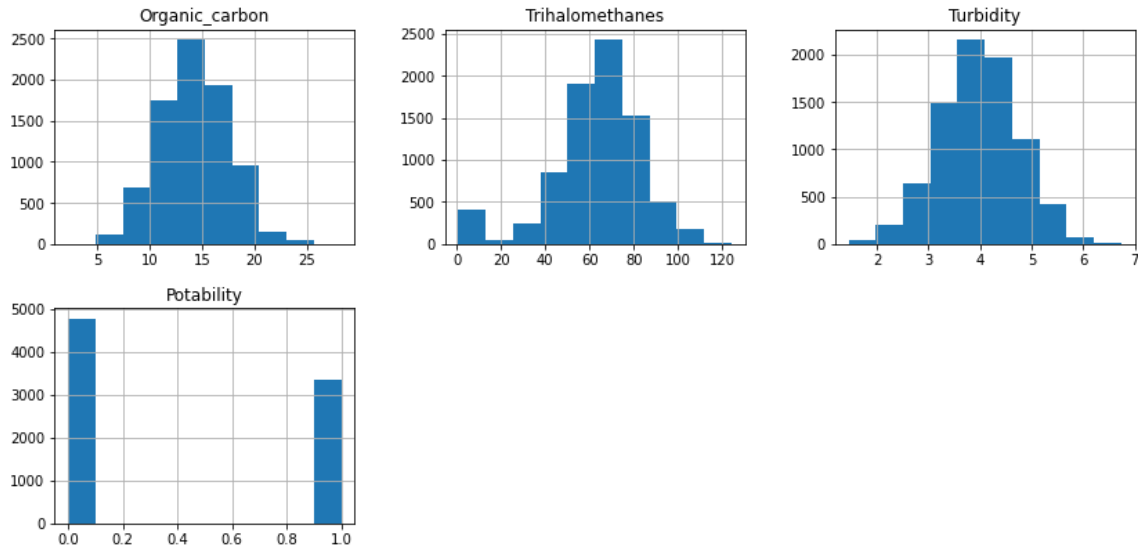


Figure 3.3.3: Distribution graphs of the features dataset (2)

pH Distribution Graph: This shows the pH levels of the various sources of water that were included in the dataset, with the highest distributions having pH values between 6 and 8, which are within the range of WHO standards. The upper allowed threshold has been determined by the WHO to be a pH range of 6.5 to 8.5.

Hardness Distribution graph: The degree of hardness in water depends on how long calcium and magnesium salts are in contact with the liquid. These elements are the main contributors to hardness. The graph shows that there is a substantial concentration of these chemicals in the 150–250 range, which leads to hard water.

Solid Distribution graph: Water that has a high solids content is probably extensively mineralized. The optimum level for solids is 500 mg/l, whereas the maximum allowed is 1000 mg/l for drinking purposes. The dataset comprises water with high solid content values between 1000 and 4000, as shown by the graph.

Chloramine Distribution graph: In comparison to the standard range of 4 mg/l, the dataset comprises water sources with high chloramine concentrations, with the maximum distribution falling between 6 and 8.

Sulfate Distribution graph: There is a significant amount of sulfate between 250 and 350 in the distribution graph. Sulfates typically range in concentration from 3 to 30 mg/l in fresh water supplies, while some locations may have significantly higher amounts (up to 1000 mg/l). Sulfates are organic substances that are naturally present in rocks, soil, and minerals.

Conductivity Distribution graph: The ability of a solution to convey current—its ionic process—is really measured by conductivity. According to WHO guidelines, the conductivity value shouldn't be more than 400 S/cm. The dataset has high conductivity water, which is a result of the high solid concentration, as can be seen from the graph.

Organic Carbon Distribution graph: All types of water naturally include organic carbon, with a maximum permitted limit of 2 mg/L for drinking water. This is the organic content contained in water. Over 99% of the samples, according to the distribution graph, are above the permitted limit. Increased microbial growth in the water caused by a high organic carbon concentration causes the water's oxygen content to decrease.

Trihalomethanes Distribution graph: When organic materials found naturally in water and chlorine from water filtration mix, trihalomethanes are created. When present in high concentrations, they have been shown to cause cancer and harm natural reproduction. The dataset's distribution shows that most samples are between 0 and 80 ppm, with the highest distribution happening between 63 and 67 ppm, whereas the maximum permitted limit is 80 ppm (parts per million).

Turbidity Distribution graph: The WHO defines turbidity as the degree of cloudiness in water. The quantity of individual particles in the water, such as dust, sand, biological debris, etc., is what causes it. A crucial indicator of water quality and potability is turbidity. Water becomes clearer as turbidity decreases and vice versa. The graph demonstrates that most water sources are within the allowable range, with the largest distributions falling between 3.5 and 4.5 NTU, which is typically a sign of clean water.

Potability Distribution graph: Water is either potable or it isn't. The potability of water affects whether it is safe to drink. For non-potable (unfit for drinking) water, this is shown on the graph as "0.0," and for potable (fit for drinking) water, it is represented as "1.0." Since 2000 of them (or around 61% of the samples) had values of 0.0, which indicated that they were unfit for consumption, it is obvious that the majority of the water sources are unfit for consumption.

Mean Absolute Error & Mean Squared Error: The total of the absolute differences between actual and anticipated values is known as the mean absolute error (MAE). The average of the error squares is measured statistically by an estimator's mean squared error (MSE) or mean squared deviation (MSD).

	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

Table 3.3.1: Mean Absolute Error & Mean Squared Error

Confusion matrix: The efficiency of classifier models can be assessed using a confusion matrix. We can easily determine how much data we correctly and erroneously identify. Use Seaborn's heat map tool to examine the relationships between each feature. The heatmap below shows that there is no correlation between any of the features, hence we are unable to lower the dimension. Following is a depiction of the confusion matrices for each of our three models:

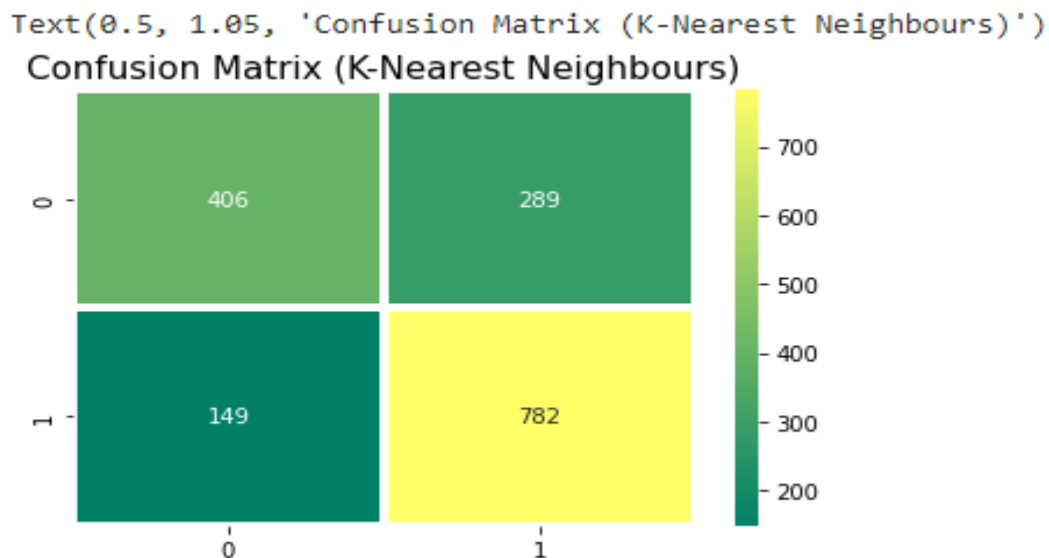


Figure 3.3.4: Confusion matrix for K- Nearest Neighbors

Text(0.5, 1.05, 'Confusion Matrix (Support Vector Machine)')
Confusion Matrix (Support Vector Machine)

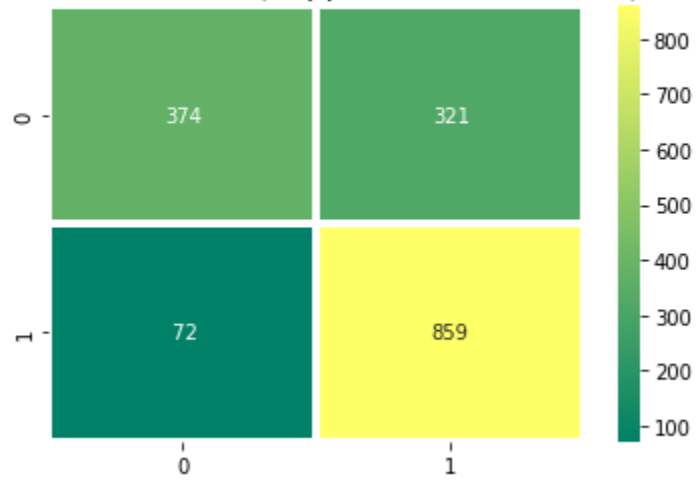


Figure 3.3.5: Confusion matrix for Support Vector Machine

Text(0.5, 1.05, 'Confusion Matrix (Random Forest)')
Confusion Matrix (Random Forest)

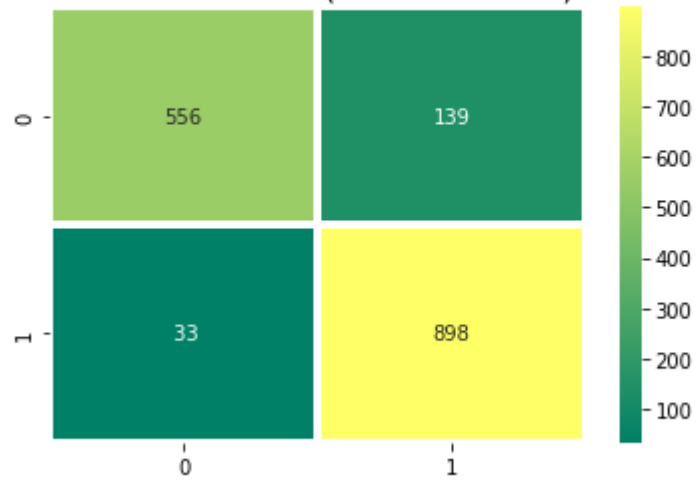


Figure 3.3.6: Confusion matrix for Random Forest


```
Text(0.5, 1.05, 'Confusion Matrix (Decision Tree)')
```

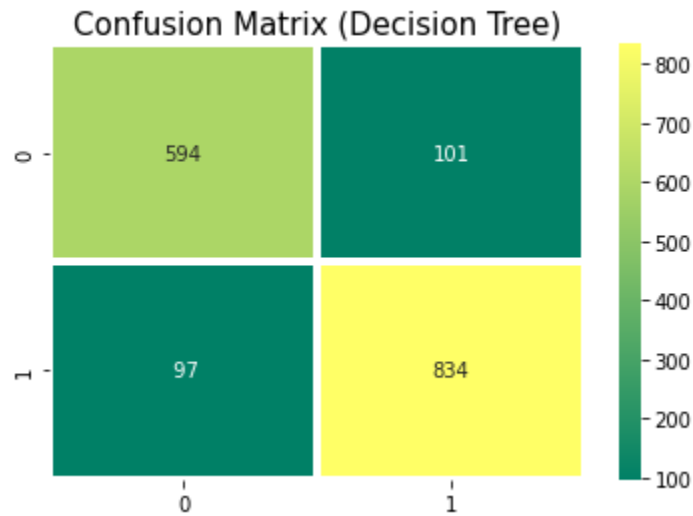


Figure 3.3.7: Confusion matrix for Decision Tree

```
Text(0.5, 1.05, 'Confusion Matrix (XGBoost)')
```

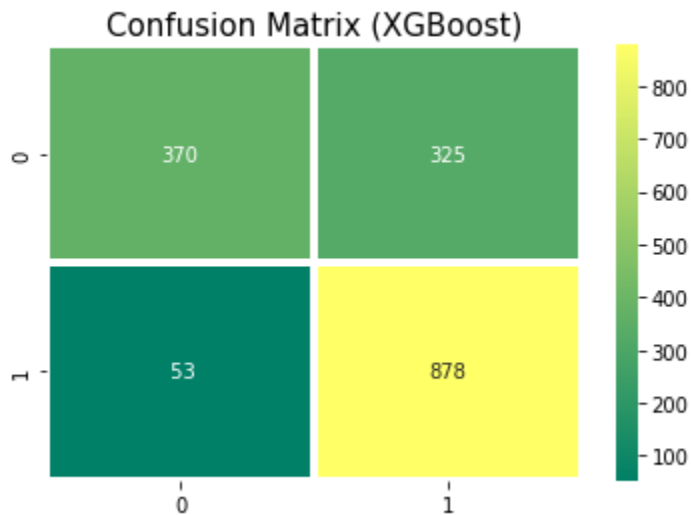


Figure 3.3.8: Confusion matrix for XGBoost

Classification report: A classification report is used to assess the precision of predictions provided by a classification algorithm. How many predictions are accurate and how many are more accurate. More specifically, True Positives, False Positives, True Negatives, and False Negatives are used to predict the metrics of a categorization report as shown below: Here is a description of the order in which f1-score, recall, and precision are determined earlier.

Precision: Precision is the percentage of positive classes that we correctly predicted out of all positive classes combined. The proportion of true positives to false positives is how precision is measured.

Recall: It shows how much of all the positive classifications we properly predicted. The better the model's quality, the higher the values. Recall is defined as True Positive / (False Positive + True Positive).

F1-score: The F-score assists in simultaneously assessing recall and precision. Instead of using Arithmetic Mean, it employs Harmonic Mean. F1-score is equal to (recall + precision)/(2*recall*precision).

Report on Classification: Precision, recall, f1-score, and accuracy are displayed for the water potability classes in the classification report. We were able to attain an accuracy rate of 88.23% with this model.

3.4 Proposed methodology

Algorithms used, KNN, Support Vector Machines, Decision tree regression, Random Forest and XGBoost.

KNN Algorithm

Data are gathered from a variety of sources in the modern world and used for analysis, theory validation, and other objectives. Dealing with these missing values therefore becomes an important stage in the preparation of data.

$$D_{xy} = \sqrt{\text{weight} * \text{squared distance from present coordinates}}$$

where

$$\text{weight} = \frac{\text{total number of coordinates}}{\text{number of present coordinates}}$$

Support Vector Machines: This regression technique is often used to categorize a water-based regression model. Classification is a simple process, and it is recognized for accuracy. By drastically changing the parameter assignment, the procedure involves building a hyper plane between the classes, which optimizes the distinction. Thus, low mismatch ratios. The SVR is trained to find solutions to issues.

$$\text{minimize } \frac{1}{2} \|w\|^2$$

$$\text{subject to } |y_i - \langle w, x_i \rangle - b| \leq \varepsilon$$

$$y = wx + b$$

Where y_i is the target and x_i is a training sample. $(w, x_i) + b$ is the sample prediction. A free parameter named ε acts as a cutoff point for all forecasts to fall inside a certain range of an accurate prediction.

Decision Tree: Decision trees are used as a tree structure in the construction of regression or arrangement models. It divides a dataset into significant subsets while continuously improving a related decision tree. A tree with leaf hubs and decision hubs is the end result. This method is used for both classification and regression.

If the numerical model is perfectly homogenous, its standard deviation is 0.

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

p_i is the probability of happening of event p_i .

Here, an actual number rather than a class serves as the expected result. The level-wise classification of the data is necessary for the creation of the tree and is accomplished using the concepts of Gini impurity or information gain.

Random Forest: We recommend that institutional researchers use random forest as their primary technique for prediction tasks rather than conventional regression and single decision tree analytics tools random forests offer the highest accuracy of all the categorization techniques now in use. The random forest method is also capable of handling huge datasets with a huge variety of factors.

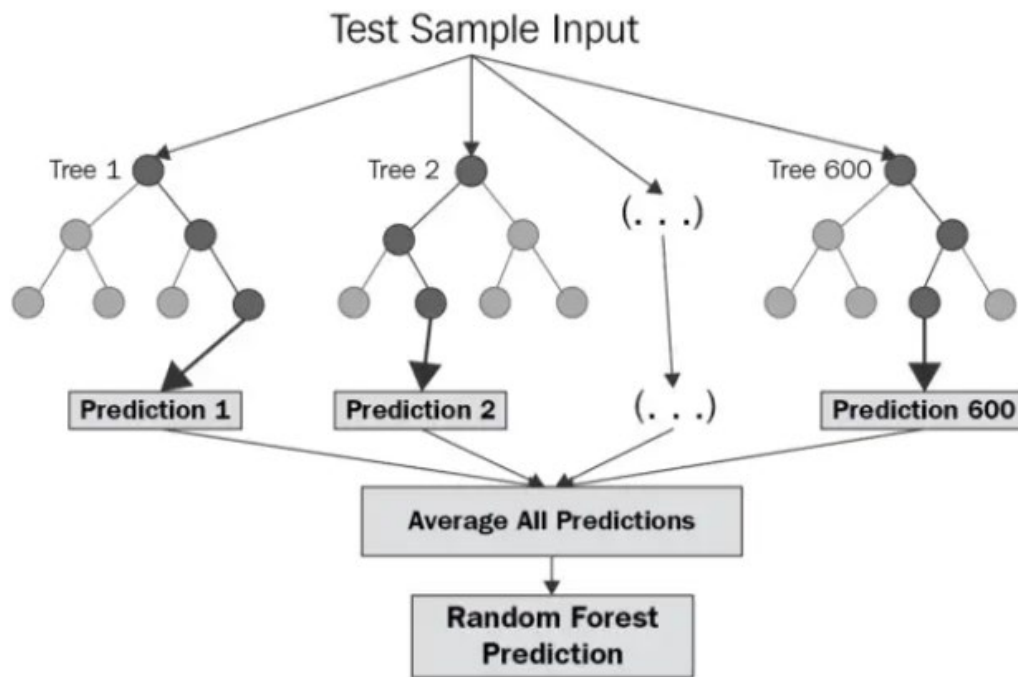


Figure 3.12: Random Forest

XGBoost

The open-source program XGBoost effectively implements the gradient boosted trees method. Gradient boosting is a supervised learning technique that attempts to accurately predict a target variable by combining the predictions of a series of weaker, simpler models. The construction of the residual trees is the primary distinction between Gradient Boost and XGBoost. XGBoost constructs residual trees by selecting the variables to use as the roots and nodes and computing similarity scores between leaves and the nodes preceding them

```
[ ] xgb = XGBClassifier()
xgb.fit(X_train, y_train)
Y_pred_xgb = xgb.predict(X_test)
XGB_score = round(xgb.score(X_test,y_test) * 100, 2)
print("XGBoost Accuracy: ", XGB_score, "%")
```

XGBoost Accuracy: 76.75 %

```
[108] cvs_xgb = round((cross_val_score(xgb, X,y,cv=10,scoring='accuracy')).mean()*100,2)
print('Cross Validated Score:', cvs_xgb)
```

Cross Validated Score: 75.14

3.5 Implementation requirement

This is what you see when you go to the search now. You are supposed to copy the no of row and paste it into the input box.

```
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.

When you paste the news into the input box and select "predict," the model will give you the result. The output will read "Provide is Drinkable" when the water quality seems to be trustworthy. The warning "Provide is not Drinkable" will show up if not. This is how you can use the internet to tell if a water interface is real or not.

An ML model is trained by providing it with training data, which the learning algorithm uses as a learning resource. The "ML model" refers to the model artifact created during training. An ML model is trained by providing it with training data, which the learning algorithm uses as a learning resource. The "ML model" refers to the model artifact created during training.

Six steps can be taken to build a machine learning model.

1. Consider machine learning in the context of your business.
2. Examine the data and choose the appropriate kind of algorithm.
3. Setup and tidy the dataset.
4. Division the dataset that has been prepared and cross validation.
5. Carry out a machine learning optimization.
6. Implement the model.

Since the study and development of machine learning is expanding quickly, there is an increasing need for machine learning specialists. And as more individuals develop a curiosity for understanding computer algorithms and how they operate, demand for these services will only grow in the future.

CHAPTER 4

EXPERIMENTAL RESULT AND DISCUSSION

4.1 Experimental Setup

We go over the outcomes of the final stage of work using actual data in this chapter. With our models, we got accuracy that was nearly as excellent.

Model with the Higher Accuracy

```
[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()
```

The accuracy of Random forest model is 88.23%.Most of the time ,it accurately detects the water when we type the text for the water on the border.

4.2 Experimental Results & Analysis

The characteristics of water were discovered and were ready to be employed as parameters for machine learning algorithms after gathering the findings of the tested samples. It was shown that using the physical characteristics as input parameters, machine learning may be utilized to identify the contaminating parameters.

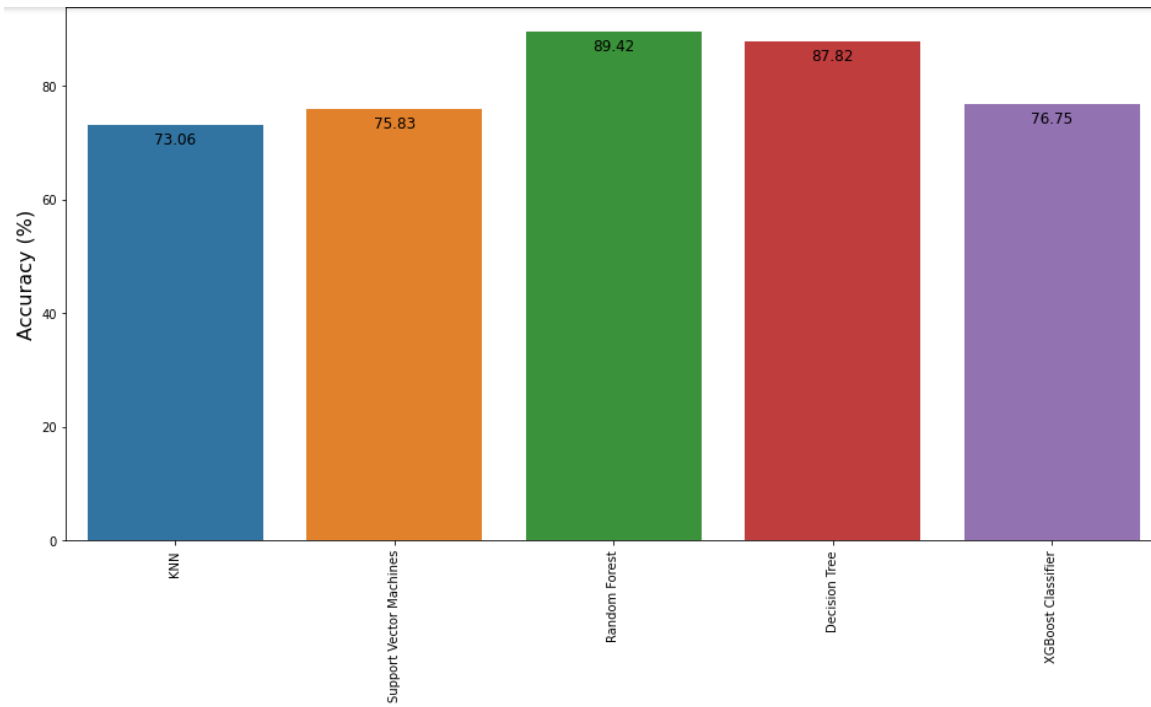


Figure 4.2.1: Accuracy chart

In the below misclassification chart show the percentage of water missing value

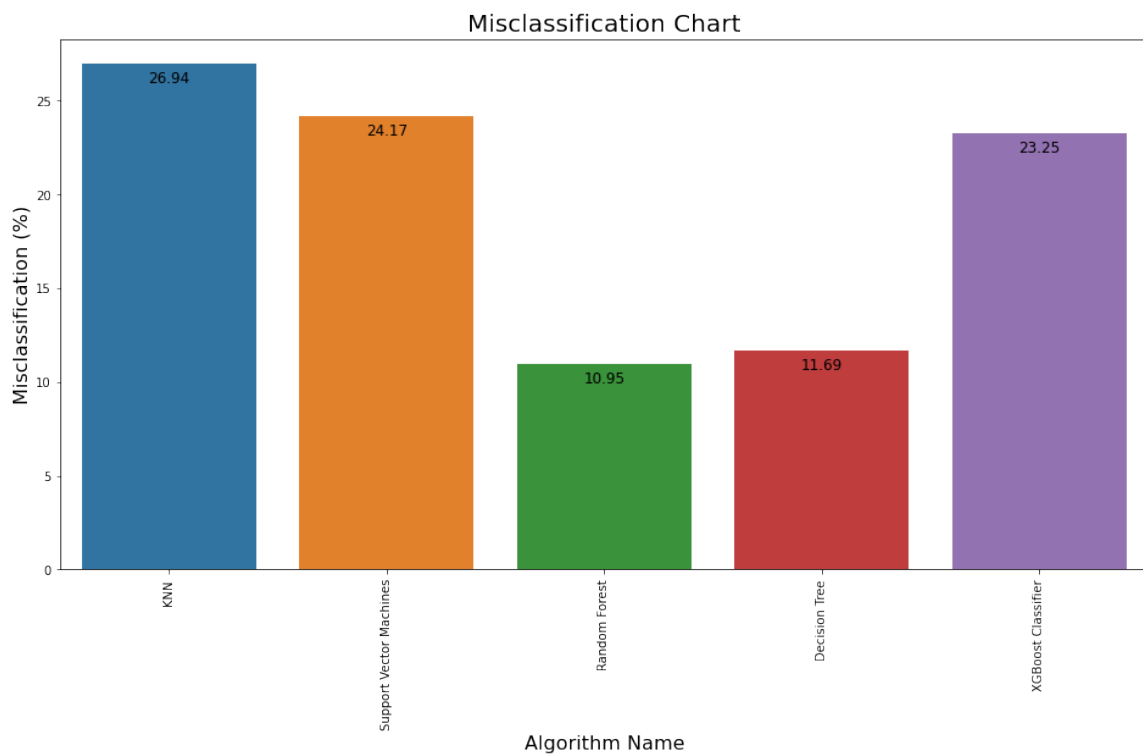


Figure 4.2.2: Misclassification Chart

In the cross validated chart calculates the accuracy of the model by separating the data into two different populations, a training set and a testing set.

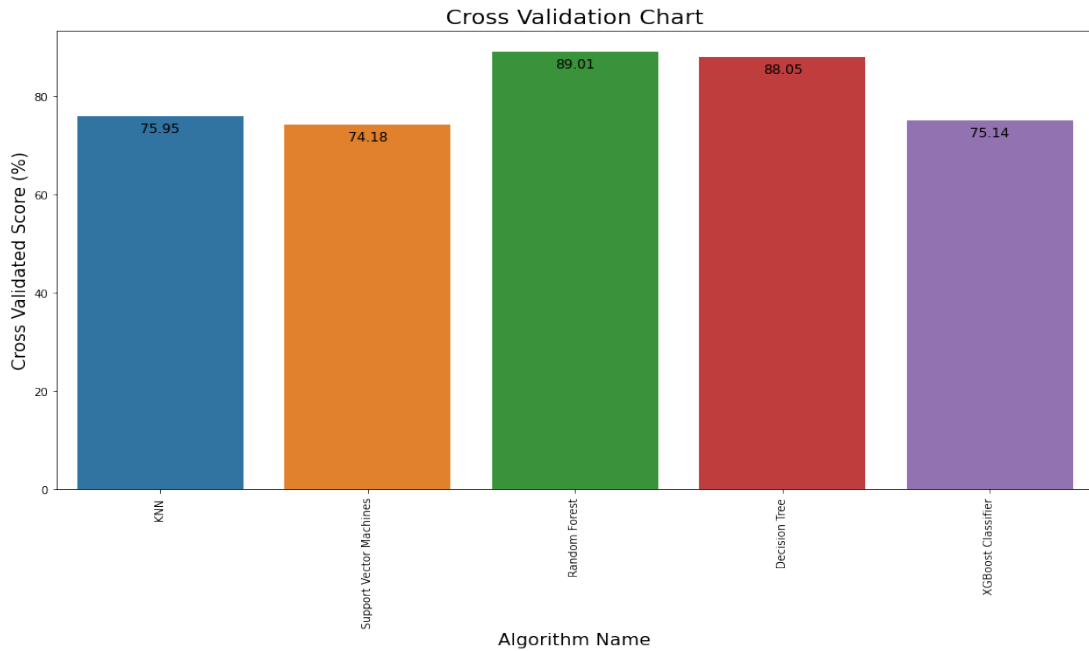


Figure 4.2.3: Cross Validation Chart

AUC stands for the level or measurement of separability, and ROC is a probability curve. It reveals how well the model can differentiate across classes.

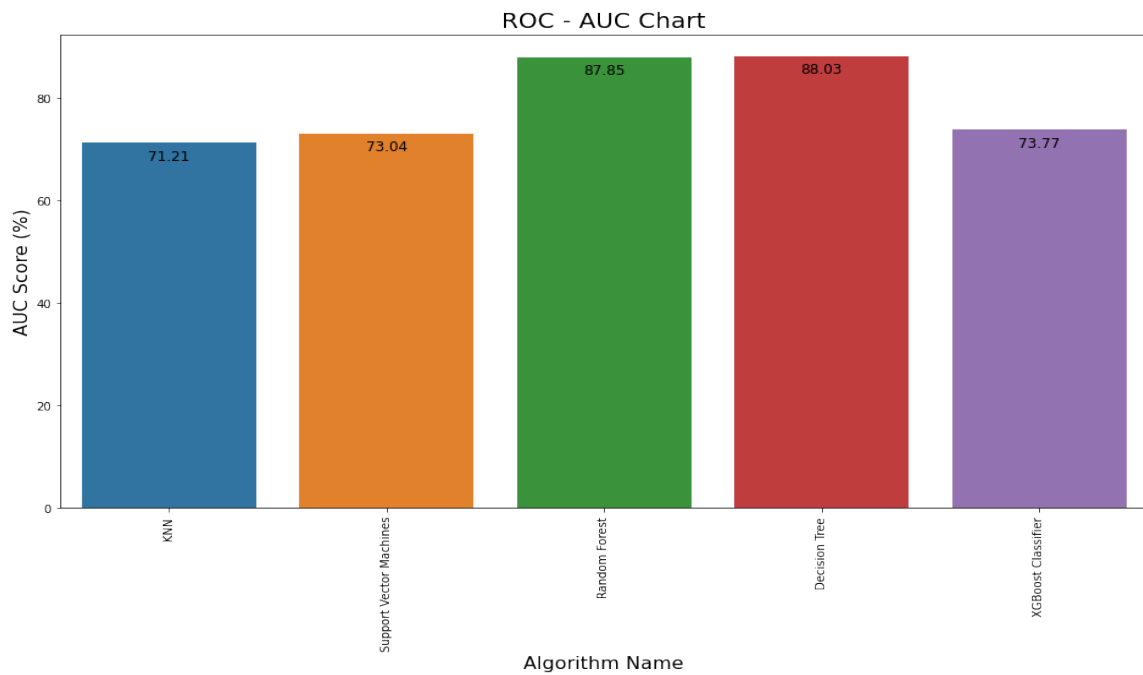


Figure 4.2.4: ROC – AUC Chart

The ROC-AUC measures how well the predictions from the two classes can be separated, distinguished, or combined. The distinction between the two classes is greater and there is less crossover of predictions with higher scores.

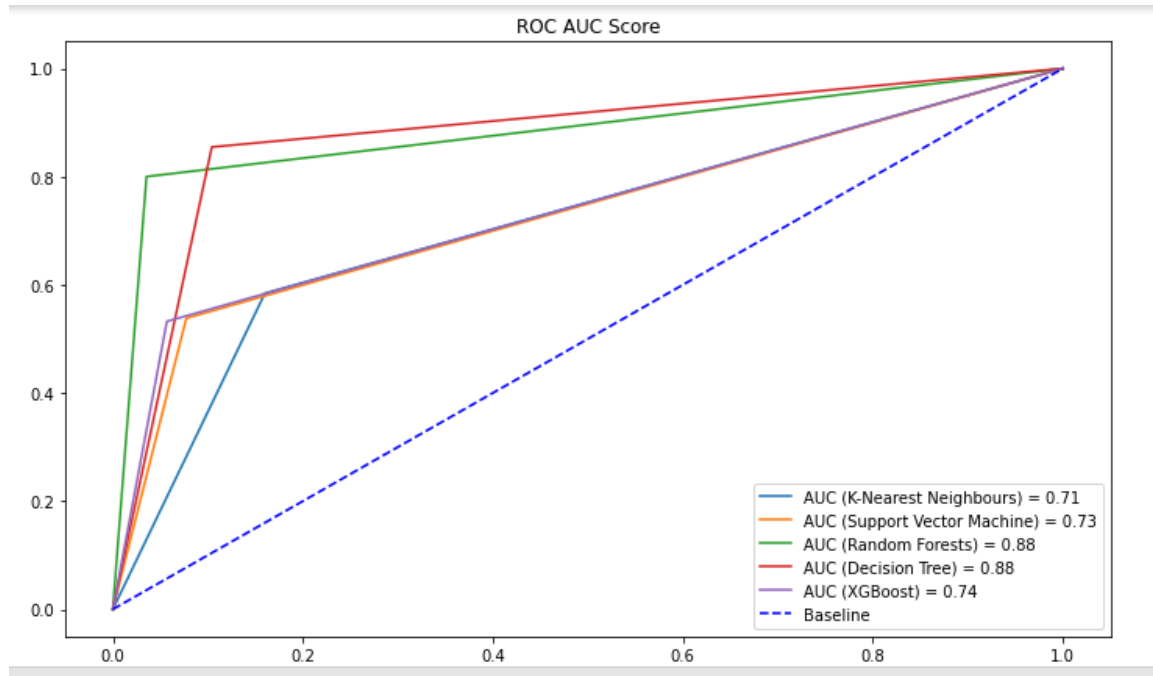


Figure 4.2.5: ROC-AUC Score

A table of correlation coefficients between variables is called a correlation matrix. The correlation between two variables is displayed in each cell of the table. Data are summarized using correlation matrices, which are also utilized as inputs for more sophisticated studies and as diagnostics for such analyses.

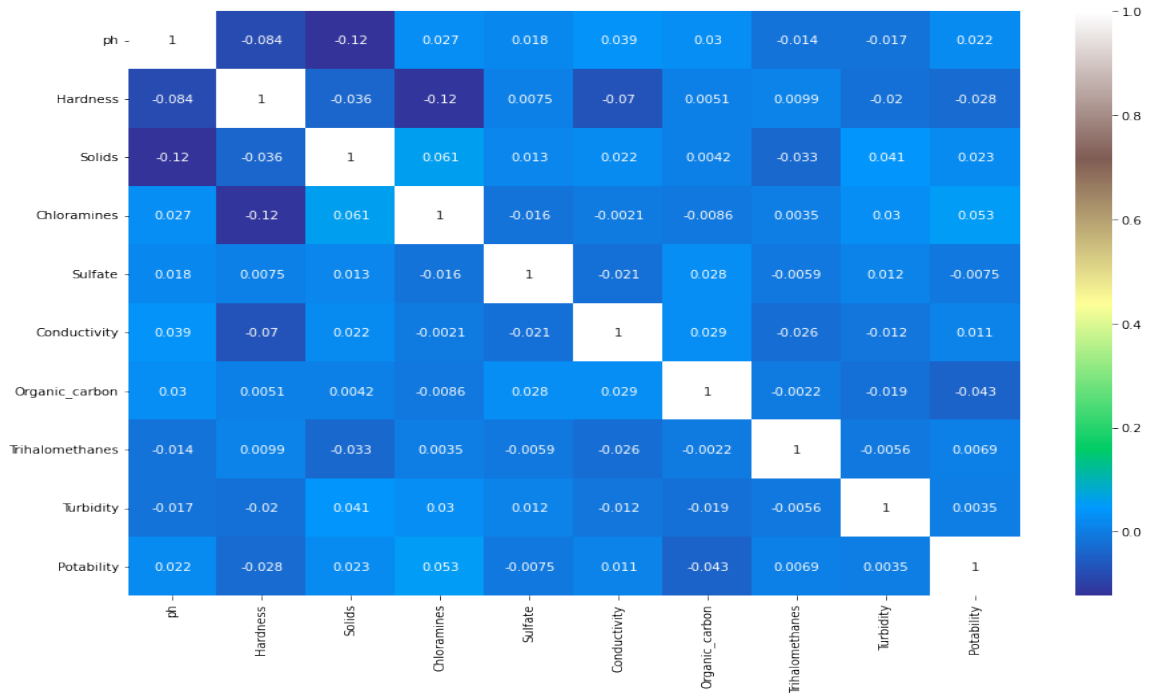


Figure 4.2.6: Correlation matrix

Algorithm	Confusion Matrix		
	Drinkable	Not Drinkable	
KNN	Drinkable	406	289
	Not Drinkable	149	782
Support Vector Machine	Drinkable	374	321
	Not Drinkable	72	859
Random Forest	Drinkable	552	143
	Not Drinkable	38	893
Decision Tree	Drinkable	595	100
	Not Drinkable	94	837
XGBoost	Drinkable	370	325
	Not Drinkable	53	878

Table 4.2.1: Confusion matrix content

CHAPTER 5

IMPACT ON SOCIETY AND ENVIRONMENT

5.1 Impact on society

Nearly 2 billion people are compelled to consume contaminated water, placing them at risk for diseases including cholera, hepatitis A, and dysentery, according to the WHO. deaths of infants. According to the UN, diarrheal illnesses that are caused by bad hygiene claim to kill almost 1,000 children worldwide every day.

Anthropogenic sources, such as improperly disposed-of household trash, runoff from agriculture, and untreated industrial effluents, are the main causes of water pollution. It is critical to comprehend both the extent of water contamination countrywide and the root causes of this major issue in order to evaluate the risk to the public's health.

5.2 Impact on Environment

Pollutants are absorbed by water, transported there, and then reassemble in lakes and oceans. As a result of human activity, many rivers have undergone physical change or influence, which has an impact on fish migration upstream or silt flow downstream. Higher sea levels, a warmer and more acidic ocean, a warmer temperature, and more significant changes in precipitation patterns are all predicted for the future. How much climate change there will be in the future will rely on what we do to cut greenhouse gas emissions right away.

People need water to maintain life, especially access to drinking water. It is critical to know whether there will be enough drinking water now and in the future for everyone. Water resources, however, are not distributed uniformly on Earth. While the availability of water is plentiful in some countries and locations, it is insufficient in others.

5.3 Ethical Aspects

Accurate water quality forecasting is essential to managing the water environment and is a crucial part of protecting the water ecosystem. One source of information about water quality is multivariate time-series datasets.

Students' critical thinking skills are developed by having them apply their prior knowledge, experiences, and observations to formulate predictions. The ability to make logical predictions helps to encourage the development of the capacity to create hypotheses.

5.4 Sustainability plan

Water sustainability plans are an essential component of provincial law that can enable and enhance adaptive water management, boost water sustainability, and forge fresh, innovative linkages in governance. To preserve water and the environment, turn the faucets off. Make sure that you don't use too much water. Shorten your showers. The amount of water used by a power shower per minute might reach 17 liters. Hold onto your dirty garments. As opposed to two half-loads, a complete load of laundry uses less water and energy.

The following are the Sustainable Development Goals: no poverty, no hunger, good health and wellbeing, quality education, gender equality, clean water and sanitation, affordable and clean energy, decent work and economic growth, industry, innovation and infrastructure, reduced inequalities, sustainable cities and communities, and responsible consumption.

CHAPTER 6

SUMMARY, CONCLUSION, RECOMMENDATION, AND IMPLICATION FOR FUTURE RESEARCH

6.1 Summary of the study

Forecasting usable water is crucial for protecting the environment and avoiding pollution. For the population to continue to enjoy excellent health, clean drinking water must be made available. Water that comes from reliable sources can be guaranteed to be potable. Forecasting for drinking water grows harder. To avoid creating inaccurate predictions, it is essential to employ the optimal learning algorithm. An intelligent model based on five various machine learning algorithms may be used to forecast the potability of drinking water based on ten different characteristics. pH, hardness, organic carbon, and other prevalent properties. In the current task, Random Forest obtained 88.23% accuracy with a 10.523% error rate.. Together with an IoT-based quality detection model, the proposed approach will be utilized in the future to evaluate and forecast the drinking water quality in diverse regions.

6.2 Conclusions

On Earth, freshwater is a limited and scarce resource, and an increasing percentage of it is contaminated by chemical pollution and dangerous microbes. Because of the need to irrigate crops more efficiently in order to feed the world's population, which is growing at an alarming rate, more freshwater is being used by people.

6.3. Implications for Upcoming Studies

With the use of water quality monitoring, it is possible to identify human influences on an ecosystem as well as foresee and learn from the environment's natural processes. In addition to helping restoration initiatives or ensuring that environmental criteria are being met, these measurement activities may also assist them. The demand for water will rise during the next 20 years as a result of population expansion, changing lifestyles, development, and agricultural activities. By 2050, it is predicted that domestic and industrial water demand would have increased by 20–50% from current levels.

REFERENCES

- [1] Taskaya-Temizel, T. and Casey, M.C., 2005. A comparative study of autoregressive neural network hybrids. *Neural Networks*, 18(5-6), pp.781-789.
- [2] Zhang, X.P., Hu, N.Q., Cheng, Z. and Zhong, H., 2014. Vibration data recovery based on compressed sensing.
- [3] Cabral Pinto, M.M., Ordens, C.M., Condesso de Melo, M.T., Inácio, M., Almeida, A., Pinto, E. and Ferreira da Silva, E.A., 2020. An inter-disciplinary approach to evaluate human health risks due to long-term exposure to contaminated groundwater near a chemical complex. *Exposure and Health*, 12, pp.199-214.
- [4] Cabral Pinto, M.M., Marinho-Reis, A.P., Almeida, A., Ordens, C.M., Silva, M.M., Freitas, S., Simões, M.R., Moreira, P.I., Dinis, P.A., Diniz, M.L. and Ferreira da Silva, E.A., 2018. Human predisposition to cognitive impairment and its relation with environmental exposure to potentially toxic elements. *Environmental Geochemistry and Health*, 40, pp.1767-1784.
- [5] Lai, Y.C., Yang, C.P., Hsieh, C.Y., Wu, C.Y. and Kao, C.M., 2011. Evaluation of non-point source pollution and river water quality using a multimedia two-model system. *Journal of Hydrology*, 409(3-4), pp.583-595.
- [6] Huang, J., Liu, N., Wang, M. and Yan, K., 2010, October. Application WASP model on validation of reservoir-drinking water source protection areas delineation. In 2010 3rd International Conference on Biomedical Engineering and Informatics (Vol. 7, pp. 3031-3035). IEEE.
- [7] Warren, I.R. and Bach, H., 1992. MIKE 21: a modelling system for estuaries, coastal waters and seas. *Environmental Software*, 7(4), pp.229-240
- [8] Hayes, D.F., Labadie, J.W., Sanders, T.G. and Brown, J.K., 1998. Enhancing water quality in hydropower system operations. *Water Resources Research*, 34(3), pp.471-483

ORIGINALITY REPORT

SIMILARITY INDEX **19** % **13** INTERNET SOURCES% **7** PUBLICATIONS% **13** %
STUDENT PAPERS

PRIMARY SOURCES

1 Internet Source	dspace.daffodilvarsity.edu.bd:8080	4 %
2 Student Paper	Submitted to Daffodil International University	3 %
3 Internet Source	journalcra.com	2 %
4 Internet Source	www.ncbi.nlm.nih.gov	1 %
5 Internet Source	github.com	1 %
