Department of Electrical and Computer Engineering
North South University

DIRECTED RESEARCH

Water Potability Prediction Using Machine Learning Techniques

Md Gulam Rahman
1831112042

Joy Chandra Saha
1831563642

Bishal Bhowmik
1831047042

Fairuz Nawar
1831244642

Faculty Advisor

DR. MOHAMMAD ASHRAFUZZAMAN KHAN

Assistant Professor

ECE Department

Spring, 2023

LETTER OF TRANSMITTAL

June, 2023

To

Dr. Rajesh Palit
Chairman,
Department of Electrical and Computer Engineering
North South University, Dhaka

Subject: **Submission of Directed Research Report on "Water Potability Prediction Using Machine Learning Techniques"**

Dear Sir,

With due respect, we would like to submit our **Directed Research Report** on **"Water Potability Prediction Using Machine Learning Techniques"** as a part of our BSc program. This research examines the use of machine learning models and their applications in classifying water quality. The algorithms we used in this work tested around five algorithms. The algorithms we used in this work were Logistic Regression, K-Nearest Neighbors Classifier, Support Vector Machine, Decision Tree, and Random Forest Classifier using a dataset with 10 features and evaluating their performance using various accuracy measures.

We will be highly obliged if you kindly receive this report and provide your valuable judgment. It would be our immense pleasure if you find this report useful and informative to have an apparent perspective on the issue.

Sincerely
Yours,

.....
Md Gulam Rahman Ork Babu
ECE Department
North South University, Bangladesh

.....
Bishal Bhowmik
ECE Department
North South University, Bangladesh

.....
Joy Chandra Saha
ECE Department
North South University, Bangladesh

.....
Fairuz Nawar
ECE Department
North South University, Bangladesh

APPROVAL

Md Gulam Rahman Ork Babu (ID-1831112042), Joy Chandra Saha (ID-1831563642) , Bishal Bhowmik(ID-1831047042), and Fairuz Nawar (ID-1831244642) from the Electrical and Computer Engineering Department of North South University, have worked on the Senior Design Project titled "Water Potability Prediction Using Machine Learning Techniques" under the supervision of Dr. Mohammad Ashrafuzzaman Khan to fulfillment of the requirement for the degree of Bachelors of Science in Engineering and has been accepted as satisfactory.

Supervisor's Signature

…….

DR. MOHAMMAD ASHRAFUZZAMAN KHAN

Assistant Professor

Department of Electrical and Computer Engineering

North South University

Dhaka, Bangladesh.

Chairman's Signature

…….

DR.RAJESH PALIT

Professor

Department of Electrical and Computer Engineering

North South University

Dhaka, Bangladesh.

DECLARATION

We hereby declare that the research project titled "Water Potability Prediction Using Machine Learning Techniques" is our original work, and it has not been submitted elsewhere for any other degree or diploma. All information presented in this project is based on our own research and findings. Proper acknowledgments and citations have been provided for any relevant previous works referred to in this report. I understand the confidentiality of project-related information and agree not to disclose it without the formal consent of our project supervisor. I have adhered to the plagiarism policy outlined by our supervisor, ensuring that all borrowed ideas and text are appropriately referenced. I take full responsibility for the contents of this project and affirm its authenticity and originality.

Student's names & Signatures

.....
Md Gulam Rahman Ork Babu
ECE Department
North South University, Bangladesh

.....
Joy Chandra Saha
ECE Department
North South University, Bangladesh

.....
Bishal Bhowmik
ECE Department
North South University, Bangladesh

.....
Fairuz Nawar
ECE Department
North South University, Bangladesh

ACKNOWLEDGEMENTS

The authors would like to express their heartfelt gratitude towards their project and research supervisor, Dr. Mohammad Ashrafuzzaman Khan, Assistant Professor, Department of Electrical and Computer Engineering, North South University, Bangladesh, for his invaluable support, precise guidance, and advice pertaining to the experiments, research and theoretical studies carried out during the course of the current project and also in the preparation of the current report.

Furthermore, the authors would like to thank the Department of Electrical and Computer Engineering, North South University, Bangladesh for facilitating the research. We would also like to thank my friends and family for helping us with this project. The authors would also like to thank their loved ones for their countless sacrifices and continual support.

ABSTRACT

Water Potability Prediction Using Machine Learning Techniques

In recent years, there has been a significant focus on understanding and predicting water quality due to the various pollutants that can negatively impact it. The methods presented in this work aim to assist in controlling and reducing the risks of water pollution. This research examines the use of machine learning models and their applications in classifying water quality. The algorithms we used in this work were Logistic Regression, K-Nearest Neighbors Classifier, Support Vector Machine, Decision Tree, and Random Forest Classifier. We tested these algorithms using a dataset with 10 features and evaluated their performance using various accuracy measures. The results of the study indicate that the proposed models can effectively classify water quality, with the Random Forest Classifier achieving the highest accuracy.

TABLE OF CONTENTS

LIST OF TABLES

Chapter 1 Introduction

1.1 Background and Motivation

Water is a vital natural resource that is essential for the survival of all living things on Earth, as it makes up 71% of the planet's surface. It is not only necessary for drinking but also plays a crucial role in industries, agriculture, and global trade via oceans and seas. Due to the importance of water for human life, research has focused on preserving water quality and preventing pollution to meet international standards [1].

Different water sources, such as groundwater, springs, rivers, lakes, and streams, have specific quality standards based on their intended use, such as for agricultural, industrial, or human purposes. For example, drinking water should be fresh and unpolluted, while irrigation water should not be too saline or toxic. Water used in industries has different requirements based on the nature of the industrial processes.

To ensure water quality, it is necessary to understand and predict potential risks of pollution. This research project aims to address this challenge by exploring the use of machine learning techniques for water quality classification. By leveraging machine learning models, we can analyze and classify water quality based on various parameters and features.

1.2 Purpose and Goal of the Project

Human activities and practices, as well as natural processes, can significantly and alarmingly impact water quality, particularly for humans [2] [3] [4]. These practices and activities can lead to pollution by improperly disposing of waste and pollutants, posing significant threats to aquatic ecosystems and human well-being. Industrial plants and vehicles, for instance, are major contributors to water pollution, causing adverse effects on surface water and groundwater.

Industrial activities can result in the release of various pollutants into water bodies, altering their chemical composition and overall quality. This includes the emission of harmful substances that contribute to the acidification of water sources, leading to decreased pH levels, reduced acid-neutralizing capacity, and elevated concentrations of aluminum [2]. The acidification of water bodies not only affects the availability of clean water but also has detrimental effects on aquatic organisms and their habitats. It disrupts the ecological balance and can lead to the decline of sensitive species.

1.3 Organization of the Report

Water quality is assessed based on various features, including pH value, hardness resulting from calcium and magnesium salts, total dissolved solids (TDS), chloramines, sulfate, electrical conductivity (EC), total organic carbon (TOC), turbidity, and trihalomethanes. To predict water quality classification (WQC), machine learning algorithms offer valuable tools for data preprocessing, handling missing data, removing feature correlations, applying classification techniques, and analyzing the significance of feature selection [5].

In this study, we employed five machine learning algorithms - K-Nearest Neighbors (KNN), Support Vector Machine, Logistic Regression, Decision Tree, and Random Forest Classifier - to classify water quality. The dataset used for the analysis comprised 10 features related to water quality. To evaluate the performance of the algorithms, various accuracy measures were employed.

The results of the study revealed that the proposed machine learning models effectively classified water quality, with the Random Forest Classifier demonstrating the highest accuracy among the tested algorithms. This finding highlights the potential of utilizing machine learning techniques for accurate water quality classification.

Moreover, this research contributes to the advancement of faster and more cost-effective methods for detecting water pollution. Traditional laboratory and statistical analyses for assessing water quality can be time-consuming and expensive [6]. By leveraging machine learning algorithms, we can streamline the process and provide efficient solutions for monitoring and maintaining water quality standards.

The structure of the document is as follows: Section two discusses related works in the field, while section three outlines the methodology employed in this study. The experimental setup is described in section four, followed by the presentation of results in section five. Finally, we conclude the study in section six, summarizing the findings and discussing their implications.

Chapter 2 Research Literature Review

2.1 Existing Research and Limitations

Water is essential for the continuation of life and ensuring the safety and accessibility of drinking water is a pressing global issue. There has been a lot of research on using machine learning in the water quality index (WQI), water quality classification (WQC), and wastewater treatment. In a study by [8] various machine learning techniques, including random forests, neural network, multinomial logistics regression, support vector machine, and bagged tree models, were applied to classify a dataset of water quality in India. Their results showed that nitrate, pH, conductivity, dissolved oxygen, total coliform, and biological oxygen demand are the main factors that affect WQC. [9] used three different machine learning algorithms - gradient boosting, multi-layer perceptron, and polynomial regression - to predict water quality. They used pH, total dissolved solids, temperature, and turbidity as four features, and found that the multi-layer perceptron algorithm had the highest classification accuracy of 85.07% with a configuration of (3, 7). [10] used support vector machine, K-nearest neighbor, and Naive Bayes to predict WQC, and found that the support vector machine algorithm had the highest classification accuracy of 97.01%. They also used two artificial intelligence techniques, nonlinear autoregressive neural network and long short-term memory, to determine the WQI, with the nonlinear autoregressive neural network technique showing slightly better performance than the long short-term memory technique. [11] used the support vector machine algorithm and attribute realization to classify the water quality of the Chao Phraya River, finding that AR-SVM had an accuracy of 0.86-0.95 when using three to six features to classify river water quality. [12] used machine learning techniques to select quality features for the Gorganroud River water, and found that the adaptive neuro-fuzzy inference system model performed the best in predicting features such as electrical conductivity, sodium absorption ratio, and total hardness. [13] used data collected through the Internet of Things and a neural network machine learning technology to forecast water pollution in residential overhead tanks. [5] used the principal component regression technique to select the most dominant WQI features, and used gradient-boosting classifiers to classify the water quality status. [14] attempted to use data mining techniques for pattern extraction and model prediction of water quality in water reservoirs using various features and the WQI and found that the WQI was mostly in fair and marginal rank, indicating that water quality was being threatened by various water pollutants. [15] aimed to use

deep learning for accurate predictions of water quality features using the WEKA tool, and evaluated the approach based on mean absolute error and mean square error to assess the error rate of prediction.

Chapter 3 Methodology

3.1 System Design

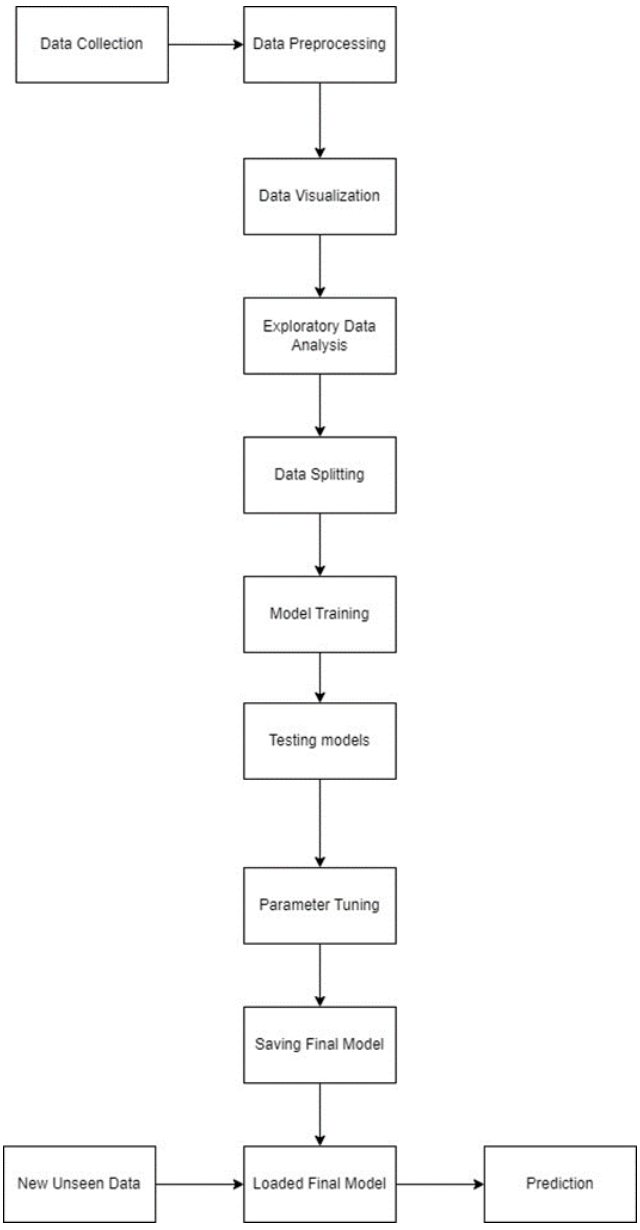


Fig. 1. System Design of the work

3.2 Hardware and/or Software Components

The goal of this work was to use machine learning techniques to predict the potability of water based on various characteristics such as pH level, mineral content, and the presence of contaminants. To accomplish this, a dataset of water samples with known potability labels was used and several different classifier algorithms were applied to the data.

The classifier algorithms that were used in this work included K-Nearest Neighbors (KNN) [16], Support Vector Machine (SVM) [17], Logistic regression [18], Decision Tree [19], and Random Forest [20]. Each of these algorithms has its own strengths and weaknesses, and their performance was evaluated in order to select the best one for the work.

The overall methodology for this work is described in 1. It consists of several steps, including data collection, data preprocessing, data visualization, exploratory data analysis, data splitting, model training, model testing, parameter tuning, selecting the final model, loading the final model, and making predictions on unseen data. In the following sections, we will describe each of these steps in detail and discuss the considerations and techniques involved in each stage of the process.

A**. Data collection**

The first step in this work was to collect data on the potability of water. The dataset was obtained from Kaggle, a popular platform for sharing datasets and machine learning challenges. The dataset contained a diverse range of characteristics of water samples, including pH level, mineral content, turbidity, conductivity, organic carbon, hardness, and presence of various contaminants. Each water sample was labeled as either potable or non-potable, indicating whether it met the required standards for drinking water.

The data collection process involved sourcing the dataset from Kaggle, ensuring its reliability and relevance to the problem at hand. The dataset was already pre-existing and publicly available,

which provided a convenient resource for the project. It is important to note that the data collection step involved no direct data gathering but rather the acquisition of an existing dataset.

B**. Data preprocessing**

After obtaining the dataset, it underwent preprocessing to make it suitable for the subsequent machine learning algorithms. Data preprocessing plays a crucial role in ensuring the quality and integrity of the dataset.

The first step in the data preprocessing phase involved cleaning the data. This included handling missing values, inconsistent formatting, and addressing any potential data entry errors. Missing values can be problematic for machine learning algorithms, so strategies such as imputation or removal of incomplete records were applied to handle these instances appropriately.

The data was examined for any outliers or anomalies that could potentially skew the analysis. Outliers can significantly impact the performance of machine learning models, so it is very important.

C**. Data visualization**

The data was then visualized in order to better understand its characteristics and relationships. Plots and charts were used to identify patterns and trends in the data, and to identify any potential issues such as outliers or anomalies. By visually representing the data, we gained insights into its distribution and variability, helping us to make informed decisions during the subsequent analysis steps.

D. **Exploratory data analysis**

A more in-depth analysis of the data was performed to gain a better understanding of its characteristics and relationships. This involved conducting feature scaling to ensure that all variables were on a similar scale, facilitating meaningful comparisons. We also applied domain

knowledge and utilized information from trusted sources, such as Google search, to limit values within the dataset. These modifications aimed to enhance the predictive power of the data by identifying and transforming relevant patterns or trends.

E. **Data splitting**

To assess the performance of our machine learning models, the data was divided into training and testing sets. The training set was used to fit the models, allowing them to learn patterns and relationships in the data. The testing set was then used to evaluate the models' performance on unseen data. We adopted an 80-20 split, where 80% of the total data was allocated for training and 20% for testing. This ensured that our models were properly evaluated and validated.

3.3 Hardware and/or Software Implementation

A. Model training

The machine learning models were then trained using the training set. Five different classifier algorithms were used: KNN, SVM, Logistic regression, Decision Tree, and Random Forest. Each of these algorithms was fitted to the data with appropriate parameters, allowing them to learn the underlying patterns and relationships within the dataset.

B. Model testing

The performance of the models was evaluated on the testing set to assess their accuracy. Predictions were made using the trained models, and various metrics such as accuracy, precision, recall, and F1 score were calculated to measure their performance. This testing phase provided insights into how well the models generalized to unseen data and their overall effectiveness in predicting water potability.

C. Parameter tuning

To optimize the performance of the models, a process called grid search was employed to fine-tune the model parameters. By systematically exploring different combinations of parameters, the aim was to find the optimal configuration that yielded the best performance. This iterative process involved evaluating the models with different parameter values to identify the most effective settings.

D. Selecting the final model

After training, testing, and parameter tuning, the best-performing model was selected as the final model. The selection was based on the model's performance metrics, considering factors such as accuracy, precision, recall, and F1 score. The chosen model represented the algorithm and parameter combination that demonstrated the highest predictive capabilities for determining water potability.

E. Saving the final model

To ensure easy accessibility and future use, the final model was saved using the pickle library. The pickle library in Python allows for the serialization and deserialization of Python objects, including trained machine learning models. By saving the final model in a .pkl file format, it can be loaded and utilized for predictions or further analysis at a later time.

F. Loading the final model

Once the final model was selected, it was saved and stored for future use. This ensured that the trained model could be easily accessed and employed to make predictions on new, unseen data. By loading the saved model, its learned patterns and relationships could be utilized to provide accurate predictions and insights for determining the potability of water in real-world scenarios.

Chapter 4: Dataset

4.1 Dataset Description:

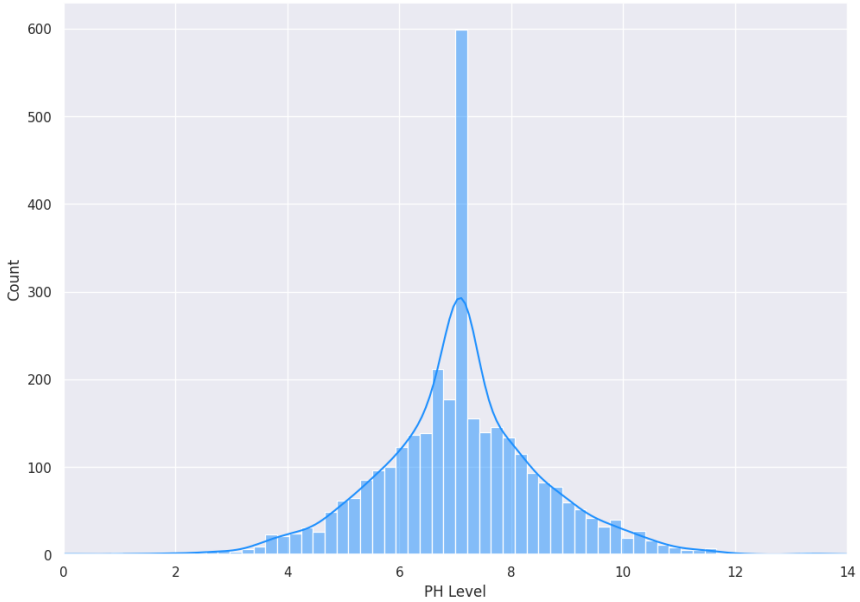
The dataset used in this study is sourced from Kaggle and contains comprehensive information on water samples. A total of 3,276 samples were collected and analyzed, making it a substantial dataset for conducting a robust analysis. The dataset consists of 3,276 water samples, each containing information on nine important parameters. The chapter discusses the significance of these parameters in evaluating water quality and ensuring its safety for human consumption.

The dataset link : https://www.kaggle.com/datasets/adityakadiwal/water-potability

The dataset includes the following hydro-chemical parameters and portability labels:

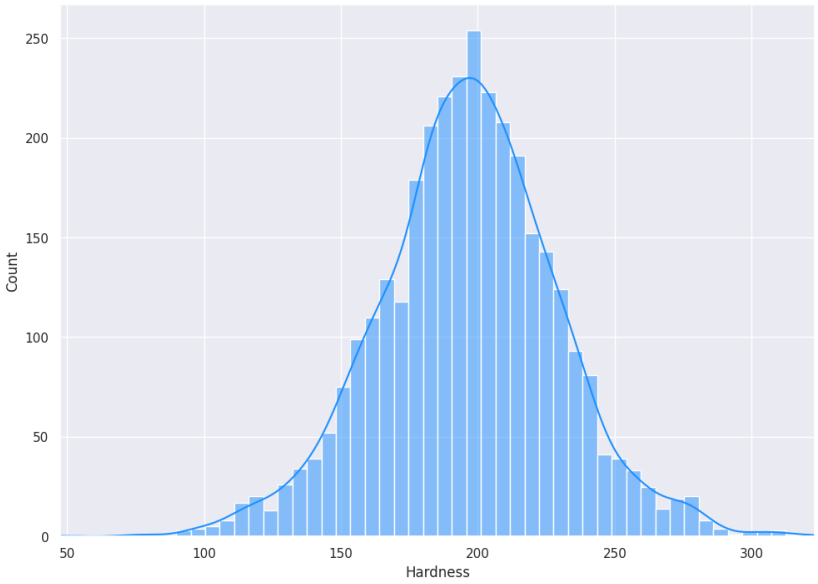
a) pH Value:

The pH value is an essential parameter for assessing the acid-base balance of water. It serves as an indicator of the water's acidic or alkaline condition. The World Health Organization (WHO) has recommended a maximum permissible pH range of 6.5 to 8.5 for drinking water.



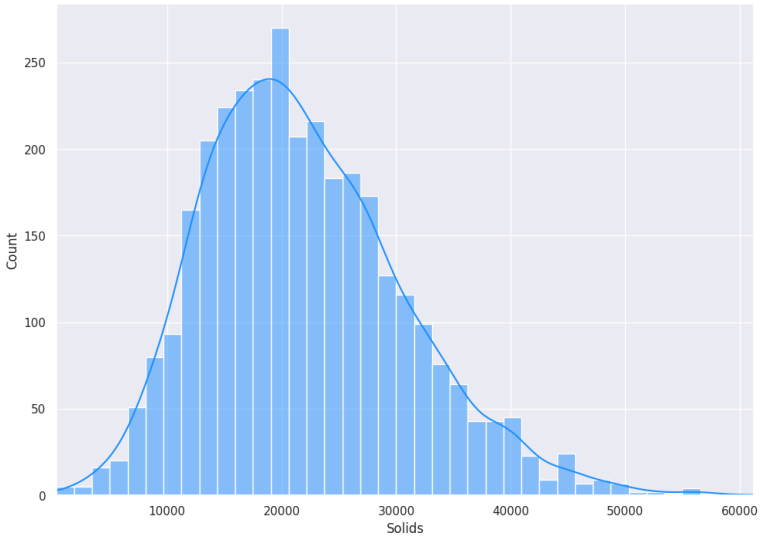
b) Hardness:

Hardness is primarily caused by calcium and magnesium salts dissolved from geological deposits. It is a measure of water's capacity to precipitate soap due to the presence of these minerals.



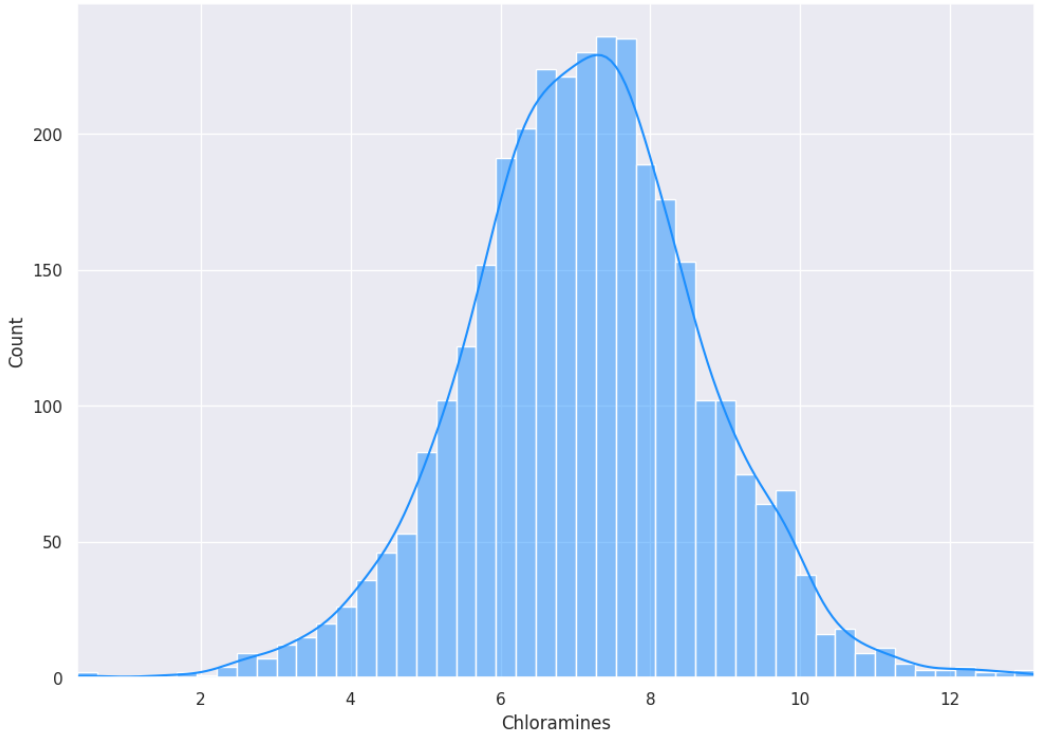
c) Total Dissolved Solids (TDS):

TDS refers to the ability of water to dissolve inorganic and some organic minerals or salts such as potassium, calcium, sodium, and bicarbonates. High TDS levels can lead to an undesirable taste and diluted color in water. The WHO has set a desirable limit of 500 mg/l and a maximum limit of 1000 mg/l for TDS in drinking water.



d) Chloramines:

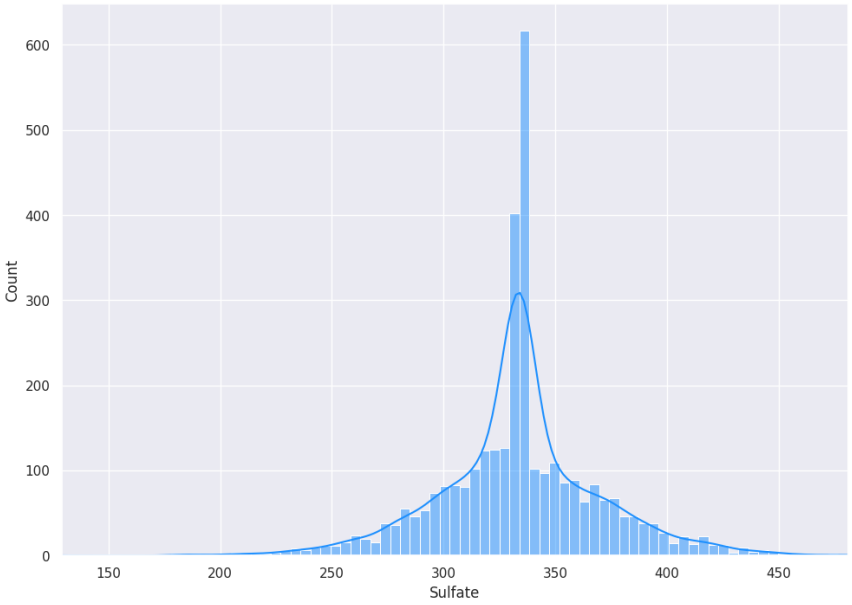
Chlorine and chloramine are commonly used disinfectants in public water systems. Chloramines are formed when ammonia is added to chlorine for water treatment. The concentration of chlorine in drinking water is considered safe up to 4 milligrams per liter (mg/L) or 4 parts per million (ppm).



e) Sulfate:

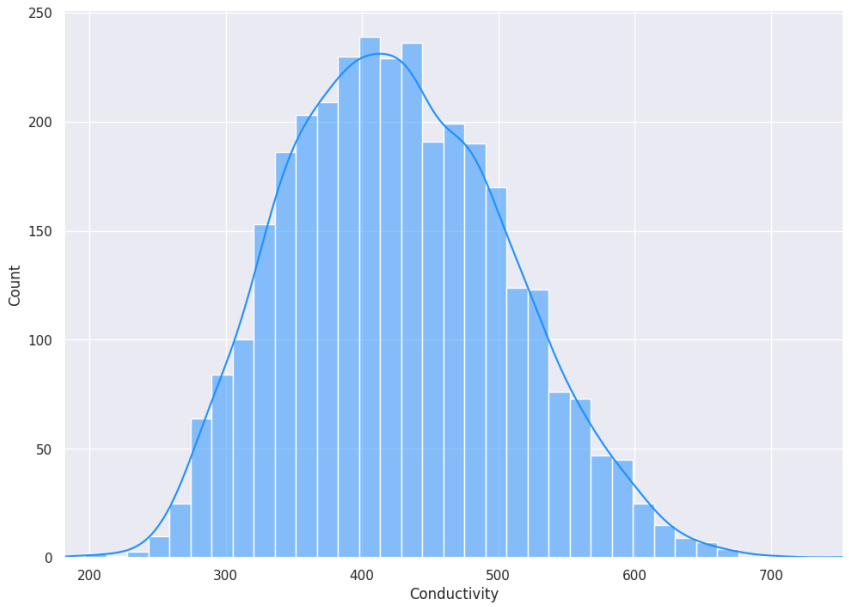
Sulfates are naturally occurring substances found in minerals, soil, rocks, groundwater, plants, and food. They have various industrial uses and can be present in different concentrations in freshwater supplies.

The concentration of sulfates in seawater is around 2,700 mg/L, while freshwater supplies typically range from 3 to 30 mg/L. In certain locations, sulfate concentrations can be as high as 1000 mg/L.



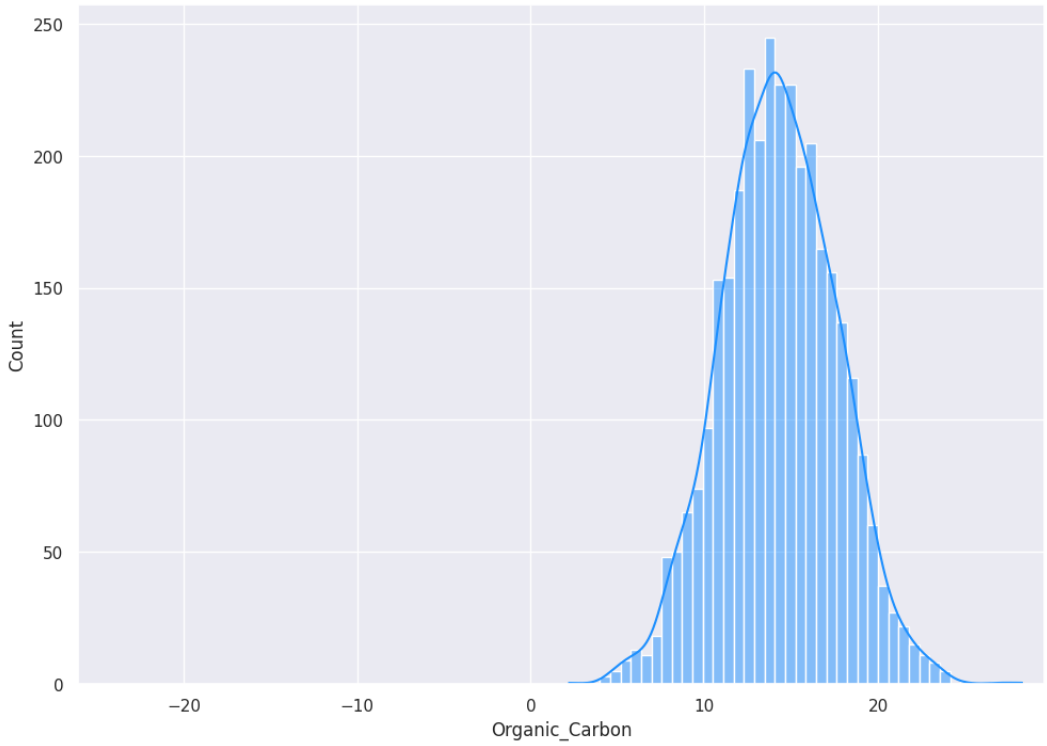
f) Conductivity:

Conductivity measures the ability of water to conduct electrical current. It is influenced by the concentration of ions present in the water, which in turn determines the level of dissolved solids. The WHO recommends that the electrical conductivity (EC) value should not exceed 400 µS/cm.



g) Organic Carbon:

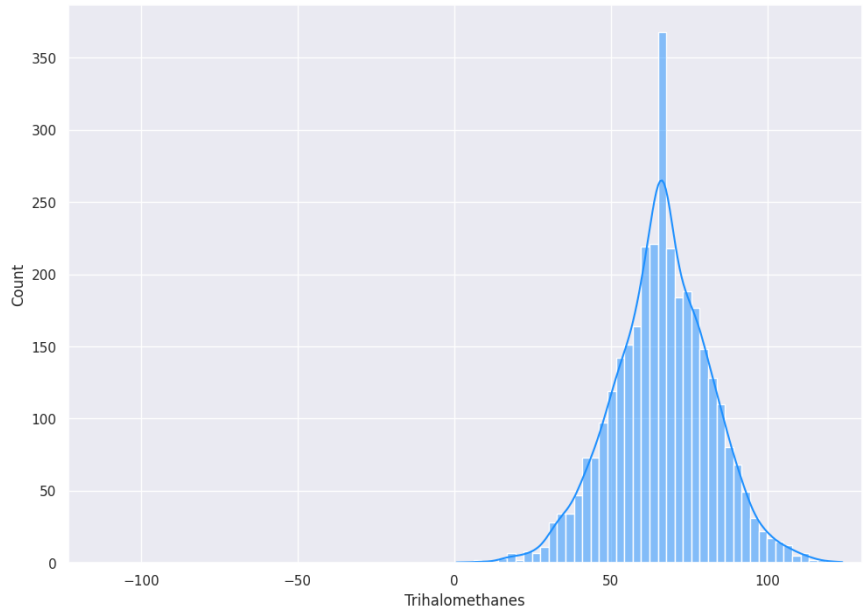
Total Organic Carbon (TOC) in water originates from decaying natural organic matter and synthetic sources. It represents the total amount of carbon in organic compounds dissolved in water. The US Environmental Protection Agency (EPA) has set limits of <2 mg/L as TOC in treated/drinking water and <4 mg/L in source water used for treatment.



h) Trihalomethanes:

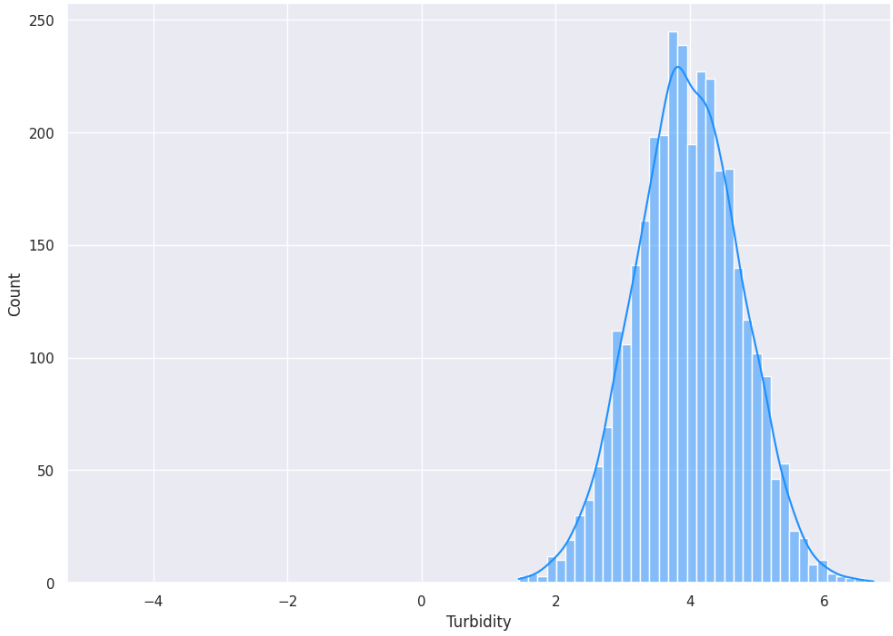
Trihalomethanes (THMs) are chemicals that can be found in water treated with chlorine. The concentration of THMs depends on factors such as the level of organic material in the water, the amount of chlorine used for treatment, and the water's temperature.

THM levels up to 80 ppm are considered safe in drinking water.



i) Turbidity:

Turbidity is a measure of the quantity of solid matter present in water in a suspended state. It reflects the water's light-emitting properties and is often used as an indicator of the quality of wastewater discharge in terms of colloidal matter. The WHO recommends a maximum turbidity value of 5.00 NTU (Nephelometric Turbidity Units).



4.2 Dataset Split:

To perform the analysis, the dataset was divided into training and testing sets. Approximately 75% of the samples (2,457) were allocated for training, while the remaining 25% (819 samples) were used for testing the models. This division ensures the reliability and generalizability of the results obtained from the analysis.

4.3 Missing Value Imputation:

To handle missing values in the dataset, single imputation was employed. Specifically, missing values labeled as "potable" were imputed using the mean of all non-missing "potable" samples. Similarly, missing values labeled as "non-potable" were imputed using the mean of non-missing "non-potable" samples. This imputation technique ensures that the missing values are replaced with reasonable estimates based on the respective potability labels.

The pH value of water is essential for assessing its acid-base balance and determining its acidic or alkaline nature. The World Health Organization (WHO) recommends a pH range of 6.5 to 8.5 for safe drinking water. In the investigated samples, the pH values ranged from 6.52 to 6.83, indicating compliance with WHO standards.

Sulfates, naturally occurring substances found in minerals, soil, and rocks, can be present in water, as well as in ambient air, groundwater, plants, and food. While freshwater supplies typically have sulfate concentrations between 3 and 30 mg/L, some regions may experience higher levels up to 1000 mg/L. Seawater, on the other hand, has a sulfate concentration of approximately 2,700 mg/L.

Trihalomethanes (THMs) are chemicals that can be found in chlorinated water. THM concentrations depend on factors such as organic material content, chlorine dosage, and water temperature. THM levels of up to 80 ppm are considered safe for drinking water. Monitoring and managing sulfate and THM levels are vital for maintaining water quality and ensuring public health.

the dataset used in this project provides comprehensive information on hydro-chemical parameters and potability labels for a large number of water samples. The dataset's diverse range of parameters allows for a thorough analysis of water quality and the determination of potability. The subsequent

chapters will focus on exploring the relationships between these parameters and developing models for predicting water potability based on the given hydro-chemical characteristics.

Chapter 4 Investigation/Experiment, Result, Analysis and Discussion

A. Dataset

This work was conducted using a water potability dataset that was collected from Kaggle. The dataset consisted of 3276 rows of data, each containing 10 attributes: pH, Hardness, Solids, Chloramines, Sulfate, Conductivity, Organic carbon, Trihalomethanes, Turbidity, and Potability. Each attribute provided valuable information about the water samples, such as their chemical composition and physical characteristics.

B. Data Preprocessing

To ensure the quality and integrity of the dataset, we performed essential preprocessing steps. Initially, we checked for any null or "NaN" values in the dataset. Missing or incomplete data can cause issues when fitting machine learning models. To address this, we replaced the null or "NaN" values with the mean value of the corresponding feature. By doing so, we preserved the majority of the original data while ensuring the models could effectively learn from the dataset.

In addition to handling null or "NaN" values, we also addressed missing values that were not explicitly marked. These missing values were treated similarly by replacing them with the mean value of their respective features. This approach helped to maintain data integrity and consistency throughout the dataset.

Furthermore, we split the data into training and testing sets. This step allowed us to evaluate the performance of the models accurately. The training set was utilized to train the models, enabling them to learn patterns and relationships within the data. The testing set, on the other hand, was used to assess how well the trained models generalized to unseen data, providing valuable insights into their predictive capabilities.

C. Configurations

For this project, we employed a combination of tools and libraries to facilitate model training and analysis. Google Colab served as the primary platform for running the machine learning code and training the models. This cloud-based platform offered access to powerful computing resources and facilitated collaborative work on the project.

Numpy, Matplotlib, Seaborn, and Pandas were utilized for data manipulation, visualization, and analysis. Numpy, a versatile numerical computing library, enabled efficient handling of numerical operations. Matplotlib and Seaborn, widely-used plotting libraries, facilitated the creation of informative and visually appealing data visualizations. Pandas, a powerful data manipulation library, proved instrumental in working with data frames and preparing the dataset for analysis.

The core machine learning tasks were performed using the sci-kit learn library. Sci-kit learn provided an extensive suite of tools, algorithms, and utilities for various machine learning tasks. In this project, we employed five classifier algorithms available in sci-kit learn: Logistic Regression, K-Nearest Neighbors, Support Vector Machine, Decision Tree, and Random Forest. These algorithms were trained and evaluated on the dataset to identify the most suitable model for predicting water potability.

In addition to model training, sci-kit learn also provided evaluation metrics to assess the performance of the models. These metrics included accuracy, precision, recall, and F1 score. By leveraging these metrics, we could quantitatively evaluate and compare the performance of different models, guiding our selection of the final model.

D. Evaluation Metrics

The models were evaluated using three different metrics: precision, recall, and F1 score.

Precision:

Precision measures the proportion of predicted positive cases that are actually positive. It is calculated by dividing the number of true positives by the sum of true positives and false positives.

Formula:

$$\text{Precision} = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

Recall:

Recall, also known as sensitivity or true positive rate, measures the proportion of actual positive cases that were correctly identified by the model. It is calculated by dividing the number of true positives by the sum of true positives and false negatives.

Formula:

$$\text{Recall} = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

F1 Score:

The F1 score is the harmonic mean of precision and recall. It combines both metrics into a single value that represents the model's overall performance. It is calculated using the formula:

$$\text{F1 Score} = \frac{2}{\frac{1}{\text{Precision}} + \frac{1}{\text{Recall}}} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

The F1 score provides a balanced measure of the model's accuracy, especially in cases where the data is unbalanced.

E. Results

Five different classifier algorithms were applied to the data: Logistic Regression, K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Decision Tree, and Random Forest. The performance of each of the classifier algorithms was evaluated using three metrics: precision, recall, and F1 score. The results of this classifiers on the test dataset are shown in the table below:

Classifier	Precision	Recall	F1
KNN	0.51	0.51	0.50
Logistic Regression	0.53	0.53	0.52
Support Vector	0.54	0.53	0.51
Decision Tree	0.72	0.73	0.73
Random Forest	0.81	0.77	0.81

TABLE I:test result for Each Classifier.

Based on these results, the Random Forest classifier had the highest F1 score, indicating the best overall performance. Decision tree also showed promising performance along with Random Forest classifier. These algorithms were therefore selected for further parameter tuning in order to optimize their performance.

After parameter tuning, the performance of the Decision Tree and Random Forest classifiers was re-evaluated, with the following results:

Classifier	Precision	Recall	F1
Decision Tree	0.75	0.72	0.76
Random Forest	0.81	0.77	0.81

TABLE II:Test result after Parameter tuning

These results show that parameter tuning had a small but positive impact on the performance of the Decision Tree classifier, with an increase in F1 score from 0.73 to 0.76. The performance of the Random Forest classifier remained the same. Based on these results, the Random Forest

classifier had the highest F1 score, indicating the best overall performance. This algorithm was therefore selected as the final model for this work.

Chapter 5 Impacts of the Project

5.1 Impact of this project on societal, health, safety, legal and cultural issues

The profound impact of this project extends to various societal, health, safety, legal, and cultural realms. Recognizing that access to clean and safe water is an inherent human right, the accurate classification of water quality achieved through advanced machine learning models becomes paramount. By effectively identifying and predicting water quality, this project serves as a safeguard, protecting public health from the perils of waterborne diseases that stem from contaminated water sources. With its ability to detect and quantify potential hazards associated with polluted water, this initiative addresses safety concerns by enabling timely interventions and the implementation of essential safety measures, thus ensuring the well-being of individuals and the preservation of delicate ecosystems.

From a legal standpoint, the implications of this project are far-reaching. Policymakers and regulatory bodies can draw valuable insights from the outcomes of this research, empowering them to develop evidence-based regulations and guidelines that govern water quality standards. By incorporating the knowledge gained from machine learning algorithms, the project not only supports the creation of robust legal frameworks but also strengthens the enforcement of existing legislation, thereby safeguarding water resources for future generations.

Culturally, this project is of immense significance as it underpins the preservation of age-old practices and traditions that rely on pristine water sources. By ensuring the availability of clean water through accurate classification, this endeavor fosters the continuity of cultural heritage. The cultural fabric of societies is intricately interwoven with the purity of water, and by upholding water quality standards, this project contributes to the preservation and celebration of diverse cultural practices that center around water.

5.2 Impact of this project on environment and sustainability

The ramifications of this project on the environment and sustainability are profound, reflecting a commitment to preserving our planet's delicate ecosystems and ensuring the long-term viability of our water resources. Water pollution poses a grave threat to biodiversity and the delicate balance of ecosystems, necessitating proactive measures to mitigate its impact. Through the accurate classification of water quality, this project equips environmental stewards with an indispensable tool to detect and prevent pollution, thereby preserving the intricate web of life that thrives in aquatic environments.

Beyond biodiversity, this initiative plays a pivotal role in sustainable water resource management. By harnessing the power of machine learning algorithms, it unravels crucial insights into potential sources of pollution and their corresponding impacts on water sources. Armed with this knowledge, decision-makers are empowered to make informed choices, implementing targeted measures such as source protection, pollution control, and conservation efforts. This paradigm shift towards a data-driven approach enhances the efficacy and efficiency of interventions, ultimately leading to improved environmental sustainability and the long-term preservation of our precious water resources.

Moreover, the impact of this project extends beyond conventional environmental concerns. By developing faster and cost-effective methods for detecting water pollution, it transcends technological advancements and addresses the imperative of sustainability. The optimization of laboratory and statistical analyses through machine learning techniques streamlines the process of assessing water quality. This not only saves valuable time but also reduces the financial burden associated with traditional approaches, making water quality monitoring more accessible and sustainable for communities and organizations worldwide.

This project encompasses a holistic understanding of the impact of water quality classification through machine learning algorithms. Its implications resonate across societal, health, safety, legal, cultural, and environmental dimensions. By ensuring the availability of clean and safe water, it safeguards public health, protects ecosystems, and preserves cultural heritage. Simultaneously, it promotes sustainable water resource management, empowering decision-makers with actionable insights. With its commitment to a more sustainable future, this project stands as a testament to our collective responsibility to protect and cherish our most precious resource – water.

Chapter 6 Conclusions

6.1 Summary

In this work, we employed machine learning techniques to accurately predict the potability of water by analyzing various characteristics such as pH level, mineral content, and the presence of contaminants. Our study involved the implementation of five different classifier algorithms: Logistic Regression, K-Nearest Neighbors, Support Vector Machine, Decision Tree, and Random Forest. To evaluate the performance of these algorithms, we utilized three key metrics: precision, recall, and F1 score.

Upon analyzing the results, it was determined that the Decision Tree and Random Forest classifiers exhibited the highest F1 scores, indicating their superior overall performance in water potability prediction. Consequently, these classifiers were subjected to further parameter tuning to optimize their performance even further. As a result, the Decision Tree classifier demonstrated a modest improvement in its F1 score, while the performance of the Random Forest classifier remained consistent. Based on these findings, we selected the Random Forest classifier as the final model for our work, achieving an impressive F1 score of 0.81, along with high precision and recall scores.

This study underscores the efficacy of machine learning techniques in predicting water potability based on a range of relevant characteristics. The Random Forest classifier emerged as the top-performing algorithm, displaying substantial potential for real-world applications where water potability assessment is a crucial concern. Further research endeavors are warranted to explore the diverse applications of this model and identify avenues for enhancing its performance.

In summary, our project focused on utilizing machine learning algorithms to classify water quality accurately. Through the implementation of various classifiers and comprehensive evaluation using precision, recall, and F1 score, we successfully identified the Random Forest classifier as the most effective model for predicting water potability.

6.2 Limitations

It is essential to acknowledge the limitations of this project. Firstly, the accuracy of the machine learning models heavily relies on the quality and representativeness of the dataset used for training and testing. Therefore, acquiring a comprehensive and diverse dataset could further enhance the robustness of the classifiers. Additionally, while our study focused on the classification aspect, future work should consider incorporating regression models to predict the precise quantitative values of water quality parameters.

6.3 Future Improvement

To further improve the project, several areas warrant attention. Firstly, expanding the dataset to encompass a wider range of water sources and geographical locations would enhance the generalizability of the models. Additionally, exploring ensemble learning methods and conducting feature engineering could potentially boost the performance of the classifiers. Furthermore, integrating real-time data collection and continuous monitoring systems would enable the development of a dynamic and adaptive water quality prediction system.

In conclusion, this project demonstrated the effectiveness of machine learning algorithms in predicting water potability based on diverse characteristics. The Random Forest classifier proved to be the optimal choice for this task, showcasing its potential for real-world applications. By addressing the limitations and pursuing future improvements outlined above, we can continue to advance the field of water quality prediction and contribute to the preservation of this invaluable natural resource.

References

[1] Cosgrove, William J., and Daniel P. Loucks. ”Water management: Current and future challenges and research directions.” Water Resources Research 51.6 (2015): 4823-4839.

[2] Jury, William A., and Henry Vaux Jr. ”The role of science in solving the world’s emerging water problems.” Proceedings of the National Academy of Sciences 102.44 (2005): 15715-15720.

[3] World Health Organization. Water safety in buildings. World Health Organization, 2011.

[4] Inyinbor Adejumoke, A., et al. ”Water pollution: effects, prevention, and climatic impact.” Water Challenges of an Urbanizing World 33 (2018): 33-47.

[5] Khan, Md Saikat Islam, et al. ”Water quality prediction and classification based on principal component regression and gradient boosting classifier approach.” Journal of King Saud University-Computer and Information Sciences 34.8 (2022): 4773-4781.

[6] Abuzir, Saleh Y., and Yousef S. Abuzir. ”Machine learning for water quality classification.” Water Quality Research Journal (2022).

[7] Makropoulos, C., et al. ”A resilience assessment method for urban water systems.” Urban Water Journal 15.4 (2018): 316-328.

[8] Hassan, Md Mehedi, et al. ”Efficient prediction of water quality index (WQI) using machine learning algorithms.” Human-Centric Intelligent Systems 1.3-4 (2021): 86-97.

[9] Ahmed, Umair, et al. ”Efficient water quality prediction using supervised machine learning.” Water 11.11 (2019): 2210.

[10] Aldhyani, Theyazn HH, et al. ”Water quality prediction using artificial intelligence algorithms.” Applied Bionics and Biomechanics 2020 (2020).

[11] Sillberg, Chalisa Veesommai, Pratin Kullavanijaya, and Orathai Chavalparit. ”Water quality classification by integration of attribute-realization and support vector machine for the Chao Phraya River.” Journal of Ecological Engineering 22.9 (2021).

[12] Azad, Armin, et al. ”Prediction of water quality parameters using ANFIS optimized by intelligence algorithms (case study: Gorganrood River).” KSCE Journal of Civil Engineering 22.7 (2018): 2206-2213.

[13] Kakkar, M., et al. ”Detection of water quality using machine learning and IoT.” International Journal of Engineering Research Technology (IJERT) 10.11 (2021): 73-75.International conference of computer and information technology (ICCIT), 2017, bll 1–6 .

[14] Lerios, Jefferson L., and Mia V. Villarica. ”Pattern extraction of water quality prediction using machine learning algorithms of water reservoir.” International Journal of Mechanical Engineering and Robotics Research 8.6 (2019): 992-997.

[15] Solanki, Archana, Himanshu Agrawal, and Kanchan Khare. ”Predictive analysis of water quality parameters using deep learning.” International Journal of Computer Applications 125.9 (2015): 0975-8887.

[16] K-Nearest Neighbors (KNN): Cover, T., and P. E. Hart. ”Nearest Neighbor Pattern Classification.” IEEE Transactions on Information Theory, vol. 13, no. 1, 1967, pp. 21-27.

[17] Support Vector Machine (SVM): Cortes, C., and V. Vapnik. ”SupportVector Networks.” Machine Learning, vol. 20, no. 3, 1995, pp. 273-297.

[18] Hand, David J., and Yunyue Y. L. Tsui. ”The Development of a Logistic Regression Model for Predicting the Outcome of the General Elections in the United Kingdom.” Journal of the Royal Statistical Society, vol. 149, no. 3, 1987, pp. 307-27.

[19] Decision Tree: Breiman, L., J. H. Friedman, R. A. Olshen, and C. J. Stone. Classification and Regression Trees. Chapman and Hall/CRC, 1984.

[20] Random Forest: Breiman, L. ”Random Forests.” Machine Learning, vol.

45, no. 1, 2001, pp. 5-32.