**ABSTRACT**

Water is an essential element for sustaining life, but access to safe water is crucial for maintaining good health. As such, monitoring, analyzing, and predicting water quality has become a critical challenge for various applications of water in our daily lives. Previous research has focused on the water quality index and principal component analysis of water quality, including the ranges of the water quality index (WQI). However, there is still a need for further improvement in the effectiveness, reliability, accuracy, and usability of current water quality management methodologies. Water quality is a crucial factor in maintaining good health, and the availability of safe and clean water is essential for the well-being of individuals and communities. However, water quality management is a complex process that involves monitoring, analyzing, and predicting water quality parameters to ensure that the water is safe for consumption and other uses.

The objective of this study is to develop a water quality prediction model that improves water quality by taking into account approximately 20 parameters, such as chloramine, fluoride, viruses, bacteria, nitrates, among others, utilizing various machine learning models, including Decision Tree, SVC, XGBoost, and KNN. To accomplish this, the model was trained with a dataset and predicted the water as either safe or unsafe. If the water was deemed unsafe, the model suggested the ranges of the factors that needed to be included in the water to make it safe. If the water was already safe, the model confirmed it.

All the machine learning models, such as Decision Tree, SVC, XGBoost, and KNN, were trained and evaluated using a trial and error method. Of these models, XGBoost was found to have the highest accuracy. The proposed model is expected to significantly improve the effectiveness, reliability, accuracy, and usability of current water quality management methodologies. This study provides a valuable contribution to the field of water quality management and can be used as a foundation for future research on the subject. The model developed in this study can not only predict the quality of water but can also suggest the ranges of factors that need to be included in the water to make it safe for consumption. The model's ability to provide actionable insights on water quality management can be of immense benefit to policymakers, water utility companies, and other stakeholders in the water industry. Overall, the proposed water quality prediction model has the potential to significantly improve the effectiveness, reliability, accuracy, and usability of current water quality management methodologies. This study's findings can provide valuable insights into the field of water quality management, facilitating better decision-making and effective policy implementation to ensure that everyone has access to safe and clean water.

**CHAPTER 1**

**INTRODUCTION**

Water is an essential resource for sustaining life, and its quality is critical to the health and well-being of both humans and the environment[1]. However, the quality of water can be affected by various chemical and biological parameters, including pollutants from human activities and natural processes. Therefore, predicting and monitoring water quality is vital for ensuring the availability of clean and safe water for all[2].Water quality prediction models can help in identifying potential threats to water quality, predicting future changes in water quality parameters, and designing effective strategies to manage and improve water quality. These models utilize various statistical and machine learning techniques to analyze historical data and identify patterns and relationships between different water quality parameters[2].Apart from predictive models, several measures can be implemented to improve water quality, including the adoption of sustainable water management practices, reducing water pollution, and developing effective water treatment technologies. Identifying the critical factors contributing to water quality degradation and implementing appropriate solutions can help in preserving and safeguarding our water resources.In this context, this study aims to explore the different parameters affecting water quality, develop a machine learning model for predicting water quality, and provide suggestions to improve water quality[3]. By understanding the critical factors contributing to water quality degradation and developing accurate prediction models, we can take proactive measures to mitigate the adverse effects of water quality degradation and ensure the availability of clean and safe water for future generations.

Water quality has become a critical issue in recent times due to growing concerns about pollution and its impact on the environment and human health. Water pollution can be caused by various sources, such as industrial and agricultural activities, untreated sewage, and discharge of chemicals and waste into water bodies. These pollutants can degrade the water quality, making it unsuitable for various uses and posing significant health risks.

As a result, water quality management has become a critical challenge worldwide, and there is an urgent need for effective and efficient methods to monitor, analyze, and predict water quality. Various techniques and technologies are available to assess and manage water quality, including laboratory testing, remote sensing, and machine learning algorithms.The importance of water quality management cannot be overstated, and it is essential to ensure that water is safe and clean for all its users. Effective water quality management requires a comprehensive understanding of the factors that affect water quality, monitoring and analysis of water quality parameters, and implementation of appropriate management strategies to mitigate the impact of pollution and other harmful effects on water quality.

**1.1 WATER QUALITY :** The survival of the majority of organisms, including humans, depends on the availability of water, an essential inorganic, transparent, and colorless chemical. For survival, living things need water of a certain quality, and some types of water pollution are tolerable but not others. Water quality refers to the physical, chemical, biological, and radiological characteristics of water that determine its suitability for various uses. The quality of water can be affected by a variety of factors, such as industrial and agricultural activities, pollution from sewage and wastewater, natural processes like erosion and weathering, and climate change.

To ensure safe and healthy water for consumption and other uses, various measures are taken to monitor and improve water quality. These include regular testing and analysis of water samples, implementation of water treatment processes to remove contaminants, and development and enforcement of regulations to prevent pollution and protect water resources.

Water quality can be classified into four types:

1. **POTABLE WATER** :Safe drinking water is vital for human consumption and household usage. It should be free from any harmful contaminants and pathogens. To define that water is safe to drink and use we have to perform several treatment processes to eliminate the viruses and microorganisms. Moreover, there are certain quality standards set by the authorities that drinking water must meet. These standards include limits for pollutants like bacteria, viruses, heavy metals, and chemicals. Even though we have a right to have pure drinking water.
2. **PALATABLE WATER** : Palatable water means which is safe water to consume by humans that has color odor and color. The water is without any harmful contaminants and pollutants that can cause health problems or change all the parameters of the water.To access the palatable water is the fundamental right of the Every human being on this earth.
3. **CONTAMINATED WATER :** Contaminated water refers to water that has been polluted by various harmful substances such as chemicals, pathogens, or other pollutants, making it unsafe for human use or consumption. Sources of contaminated water include industrial waste discharge, agricultural runoff, improper disposal of household waste, sewage discharge, and accidental oil spills. Exposure to contaminated water can cause a range of adverse health effects, including skin rashes, gastrointestinal illness, and chronic diseases like cancer. The severity of these health effects can depend on the type and concentration of contaminants in the water, as well as the duration and frequency of exposure. To ensure safe water consumption and protect public health, it is important to monitor and treat contaminated water sources and take necessary measures to prevent further contamination.[7]
4. **INFECTED WATER** : Infected water refers to water that contains harmful microorganisms, such as bacteria, viruses, and parasites, that can cause diseases in humans and animals. Drinking or using infected water can result in illnesses such as diarrhea, cholera, typhoid fever, and hepatitis. Infected water is often caused by poor sanitation, inadequate wastewater treatment, and contamination from animal or human waste. To prevent the spread of waterborne diseases, it is essential to ensure that water sources are adequately treated and sanitized before use.

To assess water quality, there are three types of parameters: physical, chemical, and biological. Water's chemical properties include its molecular formula (H2O), polarity, high surface tension, high heat capacity, ability to act as a solvent, ability to form hydrogen bonds, and pH neutrality in its pure form. Some of the physical properties of water include boiling and melting points, density, surface tension, viscosity, heat capacity, refractive index, solubility, pH, electrical conductivity, and color (although pure water is colorless). Water's biological properties include the presence of essential nutrients and minerals necessary for the survival of aquatic organisms, as well as the presence of microorganisms and pathogens that can cause disease in humans and other animals. It is important to understand these properties and parameters to ensure water quality is maintained to support the health and survival of both aquatic life and human populations[4].

**1.2 Motivation:**

Water is a precious resource that is essential for all forms of life on Earth. However, with increasing pollution and degradation of water bodies, the quality of water is deteriorating at an alarming rate. In this scenario, it becomes crucial to monitor and predict the quality of water to ensure its sustainability for future generations.

The following statements motivated me to carry out this research on water quality prediction and Giving suggestions to improve quality of water.

Water is essential for all forms of life, but its quality is deteriorating due to increasing pollution and degradation of water bodies.

1. It is crucial to monitor and predict the quality of water to ensure its sustainability for future generations.
2. Machine learning models can analyze large datasets and identify patterns to predict the quality of water based on various parameters such as Aluminum,arsenic,barium ,chlorine,cadmium etc,..
3. The proposed model aims to develop a reliable and accurate model that can predict the quality of water in real-time and identify any anomalies in the data to detect potential sources of pollution.
4. The model can suggest measures to improve the quality of water, such as the appropriate dosage of chemicals required to treat the water based on the predicted parameters.
5. This can help optimize the water treatment process and reduce the usage of chemicals.
6. The proposed model can have a significant impact on water resource management and environmental monitoring, ensuring the sustainability of water resources and improving the quality of life for millions of people worldwide.

By observation of the deteriorating quality of water and the increasing pollution and degradation of water bodies, it is essential to predict and monitor the quality of water to ensure its sustainability for future generations. This thesis focuses on water quality prediction using machine learning models and improving the quality of water by providing suggestions to tackle the root causes of water pollution.

Existing research highlights various parameters, such as temperature, pH, dissolved oxygen, and turbidity, which influence water quality. However, the combined impact of these biophysical factors, along with other factors such as the presence of pollutants and contaminants, has not been adequately addressed.

This research work focuses on the design and development of machine learning models that can reveal the quantitative influence of the combined effect of biophysical and environmental factors in deteriorating water quality. The proposed model will not only predict the water quality but also suggest measures to improve it. For instance, the model can suggest the appropriate treatment methods, such as using a specific dosage of chemicals or filtration techniques, to improve water quality based on the predicted parameters.

This research work can have a significant impact on water resource management and environmental monitoring, ensuring the sustainability of water resources and improving the quality of life for millions of people worldwide. By addressing the root causes of water pollution and using machine learning models to predict and monitor water quality, we can ensure the availability of clean and safe water for future generations.

**1.3 Problem of Statement:**

The problem of water pollution and degradation of water bodies is a significant concern for the sustainability of water resources and the quality of life of millions of people worldwide. While various parameters, such as temperature, pH, dissolved oxygen, and turbidity, influence water quality, the combined impact of these biophysical factors, along with other factors such as the presence of pollutants and contaminants, has not been adequately addressed.

Moreover, the traditional methods of monitoring and predicting water quality are time-consuming, expensive, and sometimes not accurate. There is a need to develop a reliable and accurate model that can predict water quality in real-time and identify any anomalies in the data to detect potential sources of pollution.

Furthermore, even if the water quality is predicted accurately, there is a need for measures to improve it. The traditional methods of treating water involve the usage of chemicals and other resources, which are often not optimized and result in wastage and environmental harm. There is a need for a model that can suggest the appropriate treatment methods based on the predicted parameters, thus optimizing the water treatment process and reducing the usage of chemicals.

Therefore, the problem statement is to develop a machine learning model that can predict water quality accurately and suggest measures to improve it, thus ensuring the sustainability of water resources and improving the quality of life for millions of people worldwide. This research work will address the root causes of water pollution and provide insights into the combined impact of biophysical and environmental factors, enabling effective monitoring and prediction of water quality.

**1.4 Objectives of the Research:**

The objectives of the research on water quality prediction using machine learning models and improving the quality of water by giving suggestions are:

**OBJECTIVE 1 :** To preprocess our dataset the quality of the data used in analysis is critical, as any errors or inconsistencies in the data can lead to inaccurate results or conclusions. Data preprocessing is therefore an important step in the data analysis process that ensures that the data is in a usable format and that any errors or inconsistencies are addressed.

**Data Cleaning :** The first step in data preprocessing is to identify and remove any errors or inconsistencies in the data. This includes missing values, duplicates, outliers, and inconsistencies in formatting or units of measurement.

**Data Integration:** Data integration involves combining data from different sources into a unified dataset. This can include merging datasets, resolving conflicts in variable names or formats, and identifying common keys or identifiers.

**Data Transformation:** Data transformation involves converting raw data into a more usable format for analysis. This can include normalizing data, scaling data, and converting categorical data into numerical values.

**Data Reduction:** Data reduction involves reducing the amount of data that needs to be analyzed. This can include feature selection, where only the most relevant variables are selected for analysis, or dimensionality reduction, where high-dimensional data is reduced to a lower-dimensional representation.

**Data Discretization:** Data discretization involves dividing continuous data into discrete intervals or categories. This can be useful for analyzing data that is continuous, but where discretization allows for easier analysis or interpretation.

**Data Sampling:** Data sampling involves selecting a subset of the data for analysis. This can be useful for large datasets where analyzing the entire dataset is impractical or time-consuming.

**Data Balancing:** Data balancing involves addressing class imbalance in datasets, where one class of data is overrepresented compared to others. This can involve oversampling the underrepresented class or undersampling the overrepresented class to balance the dataset.

**OBJECTIVE II** To develop a reliable and accurate machine learning model that can predict water quality in real-time based on various parameters such as Aluminum, Ammonia, Chlorine, Barium, Chromium, Lead, Nitrates, Bacteria, and other factors can help identify potential health hazards associated with drinking water. The following is a discussion of some of the key aspects of this topic:

**Importance of Real-Time Monitoring:** Real-time monitoring of water quality is essential to ensure the safety of drinking water. Traditional water quality testing methods are time-consuming and may not provide real-time data. Real-time monitoring systems, however, can provide continuous data on water quality parameters, allowing for the rapid detection of water quality issues and the timely implementation of corrective actions.

**Parameters for Water Quality Assessment:** Water quality parameters such as Aluminum, Ammonia, Chlorine, Barium, Chromium, Lead, Nitrates, Bacteria, and others can be used to assess the quality of drinking water. These parameters are usually measured in parts per million (ppm) or parts per billion (ppb). Any concentration above the safe limit can lead to health hazards such as neurological damage, cancer, and kidney damage.

**Data Collection and Analysis:** Real-time monitoring systems can collect data on water quality parameters in real-time. This data can then be analyzed using statistical methods to identify trends and patterns. Machine learning algorithms can also be used to predict water quality based on historical data and other factors such as weather conditions, land use, and other environmental factors.

**Challenges in Real-Time Monitoring:** Real-time monitoring of water quality is not without its challenges. One of the main challenges is the high cost associated with installing and maintaining real-time monitoring systems. Another challenge is the need for accurate calibration and validation of monitoring systems to ensure that the data collected is accurate and reliable.

**Future Directions:** Advances in technology and data analytics are making it easier and more cost-effective to monitor water quality in real-time. In the future, it is expected that more widespread adoption of real-time monitoring systems will occur, leading to better and more timely detection of water quality issues, and ultimately, better protection of public health.

**OBJECTIVE III** To suggest appropriate treatment suggestions, such as using a specific dosage of chemicals or filtration techniques, to improve water quality based on the predicted parameters, thus optimizing the water treatment process and reducing the usage of chemicals.

Water quality is a crucial aspect that needs to be maintained to ensure the well-being of humans and the environment. Treatment of water involves various processes, such as chemical treatment, filtration, and disinfection, to remove impurities and contaminants. To optimize the water treatment process and reduce the usage of chemicals, it is essential to suggest appropriate treatment suggestions based on predicted parameters.

One of the most important steps in water treatment is the use of chemicals to remove impurities. However, the dosage of chemicals required may vary depending on the water quality parameters, such as pH, temperature, dissolved oxygen, turbidity, and alkalinity. By predicting these parameters, appropriate treatment suggestions can be made, such as the optimal dosage of chemicals to be used for effective treatment. This helps to ensure that the water treatment process is optimized, reducing chemical usage, and lowering operational costs.

The ultimate objective of this research work is to ensure the availability of clean and safe water for future generations by addressing the root causes of water pollution and using machine learning models to predict and monitor water quality. The research work aims to contribute to the field of water resource management and environmental monitoring, thus improving the quality of life for millions of people worldwide.

**1.5 Scope of the Thesis :**   
  
In order to meet the above research objectives, the description of our contributions in four technical chapters of this thesis are summarized below.  
   
 **1. Data Preprocessing :**  The scope of this thesis includes data preprocessing, which involves correlating various parameters such as aluminum, cadmium, chlorine, lead, arsenic, barium, and other relevant factors to identify their impact on water quality. The research work also involves feature selection and optimization to develop a reliable and accurate machine learning model that can predict water quality in real-time based on these parameters. The proposed model aims to contribute to the field of water resource management and environmental monitoring by identifying any anomalies in the data and detecting potential sources of pollution. The model can suggest appropriate treatment methods, such as using specific dosages of chemicals or filtration techniques, to improve water quality based on the predicted parameters, thus optimizing the water treatment process and reducing the usage of chemicals. Overall, this research work aims to ensure the availability of clean and safe water for future generations, thus improving the quality of life for millions of people worldwide.  
  
  
**2. Comparative analysis of Machine Learning Model:** It includes a comparative analysis of machine learning models, including deep learning models, to identify the best model for predicting water quality based on time and space complexity. The research work involves analyzing the performance of different models and selecting the optimal model based on factors such as accuracy, time complexity, and space complexity. The proposed model aims to provide accurate and reliable predictions of water quality in real-time while also minimizing the computational resources required for the model's implementation. The model's accuracy, computational efficiency, and ease of implementation will contribute to the sustainability of water resources and environmental monitoring, improving the quality of life for millions of people worldwide.

**3. Water Quality Improvisation and suggestions :** The scope of this thesis includes water quality improvement through the identification of the parameters that affect it and the provision of suggestions on how to mitigate their impact. The research work involves identifying the correlations between different water quality parameters, such as aluminum, cadmium, chlorine, lead, arsenic, barium, and other relevant factors, and how they affect each other. This information can then be used to develop a machine learning model that predicts water quality in real-time and provides suggestions on how to improve it. The suggestions can include the appropriate dosage of chemicals required to treat the water, the use of filtration techniques, or the elimination of potential sources of pollution. By providing these suggestions, the proposed model can contribute to the sustainability of water resources and improve the quality of life for millions of people worldwide.

**1.6 Organization of the Thesis :**

The organization of this thesis is structured as follows:

**Chapter 1:** Introduction

The first chapter of the thesis will introduce the topic of water quality prediction and the importance of improving the quality of water. The chapter will also highlight the significance of machine learning models in environmental monitoring and water resource management.

**Chapter 2:** Literature Review

The second chapter will review the relevant literature on water quality prediction and improvement, focusing on the existing research and studies related to the topic. The chapter will identify the gaps in the research and highlight the need for the proposed study.

**Chapter 3:** Data Preprocessing

The third chapter will cover the data preprocessing techniques used in the study, including correlation of parameters, feature selection, and optimization. The chapter will describe the methods used to preprocess the data and the reasons behind each technique.

**Chapter 4:** Comparative Analysis of Machine Learning Models

The fourth chapter will focus on the comparative analysis of different machine learning models, including deep learning models, and their performance in predicting water quality. The chapter will describe the methods used to analyze the models and the factors considered for selecting the optimal model.

**Chapter 5:** Water Quality Improvement and Suggestions

The fifth chapter will discuss the proposed model's ability to provide suggestions to improve water quality. The chapter will describe the measures suggested by the model and their potential impact on the quality of water.

**Chapter 6:** Conclusion

The final chapter will provide a summary of the research work and highlight its contributions to the field of water quality prediction and improvement.

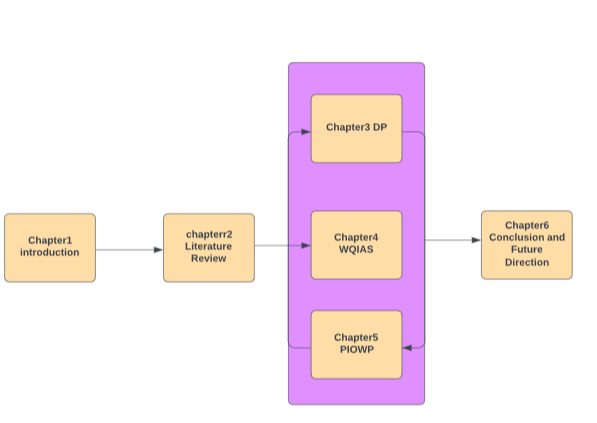


Fig : Schematic outline of the thesis

**CHAPTER 2**

**LITERATURE REVIEW**

**2.1 Introduction:**

The literature review of water quality prediction using machine learning techniques aims to provide a comprehensive overview of the research done in this area. In recent years, machine learning techniques have gained popularity due to their ability to analyze large datasets and identify hidden patterns that are not visible to the naked eye. In the context of water quality prediction, these techniques can help to predict the quality of water based on various parameters, such as aluminum, cadmium, chlorine, lead, arsenic, barium, and other relevant factors[10].

The literature review will explore the different machine learning algorithms used for water quality prediction, such as decision trees, support vector machines, and artificial neural networks. It will also analyze the various research studies that have employed these algorithms and the performance metrics used to evaluate the accuracy of the predictions. Furthermore, the literature review will examine the limitations of the current research and identify areas for further improvement.

The literature review will provide a comprehensive understanding of the existing research on water quality prediction using machine learning techniques, highlighting the potential benefits of these methods for environmental monitoring and water resource management.

**2.2 Contamination of water:**  Water contamination refers to the presence of harmful substances or pollutants in water, which make it unsafe for human consumption or other uses. Contaminants can come from various sources, such as industrial and agricultural activities, sewage, and natural sources like minerals and salts. Some common examples of water contaminants include pesticides, heavy metals like lead and mercury, bacteria and viruses, and chemicals like chlorine and fluoride. Following are the major issues

**Health Effects:** Water contamination can lead to various health problems, including gastrointestinal illnesses, skin problems, and neurological disorders. Microbial contamination of water can cause diseases such as cholera, typhoid, and dysentery.[10] A study by Xue et al. (2021) investigated the relationship between water quality and health outcomes in rural China. The study found that poor water quality was significantly associated with an increased risk of gastrointestinal and respiratory diseases.

**Environmental Effects:** Water contamination can harm aquatic life, leading to fish kills, the destruction of habitats, and the disruption of ecosystem processes. Chemical contamination of water can affect the growth, reproduction, and survival of aquatic organisms. [11] A study by Willett et al. (2021) assessed the impact of agricultural runoff on water quality and aquatic life in the Mississippi River Basin. The study found that agricultural runoff was a significant contributor to the nutrient pollution of the river, leading to the development of hypoxic zones and the decline of aquatic life.

**Economic Effects:** Water contamination can lead to economic losses through decreased productivity and increased healthcare costs. The costs associated with waterborne illnesses can be significant, including medical treatment, lost wages, and reduced productivity. A study by Gruère and Narrod (2005) assessed the economic impact of waterborne diseases on small-scale farmers in Kenya. [12] The study found that waterborne illnesses reduced farm productivity by up to 13%, leading to significant economic losses.

**Social Effects:** Water contamination can have social effects, including the displacement of communities and the loss of cultural and recreational resources. Contaminated water sources can lead to the displacement of communities who rely on these resources for their livelihoods. A study by Weichselgartner and Kelman (2015) assessed the social impacts of flooding and water contamination in the Solomon Islands. [13] The study found that flooding and water contamination had significant social impacts, including the displacement of communities and the loss of cultural and recreational resources.

**Political Effects:** Water contamination can have political effects, including the loss of trust in government and regulatory agencies. The failure of government and regulatory agencies to address water contamination can lead to a loss of trust in these institutions. A study by Hora and Cook (2018) investigated the political and regulatory challenges associated with water contamination in the United States. [11] The study found that political and regulatory factors, including a lack of funding and political will, contributed to the persistence of water contamination in some areas.

**2.2.1 Water contaminating factors :**

Water contamination occurs when harmful substances enter natural water sources, such as rivers, lakes, and groundwater, making them unfit for human consumption or use. There are various types of water contaminants that can cause adverse effects on human health, the environment, and aquatic life.

One of the most common types of water contaminants is biological contaminants, which include bacteria, viruses, parasites, and other microorganisms. These contaminants can cause waterborne diseases such as cholera, typhoid, and dysentery. These diseases are particularly dangerous for people with weakened immune systems, such as the elderly, children, and individuals with chronic illnesses.

Chemical contaminants are another type of water contaminant that can be harmful to human health.These contaminants include heavy metals, pesticides, herbicides, and industrial chemicals. Exposure to chemical contaminants through contaminated water can cause a wide range of health problems, including neurological damage, reproductive problems, and cancer[6].

Physical contaminants, such as sediment, soil, and debris, can also affect water quality. These contaminants can cause water to become cloudy or discolored, affecting its aesthetic quality. Additionally, physical contaminants can clog water treatment systems and pipes, leading to reduced water flow and pressure.

Radiological contaminants, such as uranium and radon, can also affect water quality. These contaminants can cause health problems such as cancer and damage to the kidneys.

Preventing water contamination requires effective management and monitoring of water sources. This includes proper disposal of hazardous chemicals, regular maintenance of water treatment facilities and distribution systems, and monitoring water sources for potential contaminants. It is also essential for individuals to take steps to reduce water contamination by properly disposing of chemicals and waste and being mindful of their water usage[9].

**2.3.Effects of Waterborne Contaminants on Health**

Water can also be a source of various contaminants that can have adverse effects on human health. Waterborne contaminants refer to any microorganisms, chemicals, or other substances that can cause illness or disease when consumed or exposed to them through contaminated water. The following are some of the effects of waterborne contaminants on health. The following are the some contaminants of water are Aluminum, Ammonia, Arsenic, Barium, Cadmium, Chloramine, Chromium, Copper, Fluoride, Bacteria, Viruses, Lead, Nitrates, Nitrites, Mercury, Perchlorate, Radium, Selenium, Silver and Uranium

**2.3.1 Aluminum effect on water quality:**

Aluminum can have a significant effect on water quality, particularly in areas where the soil is acidic or where aluminum mining or processing takes place. When aluminum enters waterways, it can bind with other substances and form compounds that can negatively impact aquatic life and human health.

In some cases, elevated levels of aluminum in water can lead to a range of health issues, including neurological problems, such as memory loss and confusion, and bone disorders, such as osteoporosis. Additionally, aluminum can cause skin irritation and digestive problems in some people.

Aluminum's impact on water quality can also be seen in the way it affects aquatic ecosystems. High levels of aluminum can damage fish gills, impairing their ability to breathe and leading to decreased growth and reproduction rates. Insects, such as mayflies and stoneflies, are also highly sensitive to aluminum and can be negatively impacted by its presence in water.

To mitigate the effects of aluminum on water quality, it's important to monitor water sources and implement measures to reduce contamination. This may include implementing filtration systems or treatment processes that remove aluminum from water, reducing emissions from industrial sources, and improving agricultural practices to reduce soil erosion and runoff. By taking proactive steps to protect water quality, we can help ensure that our water sources remain safe and healthy for both humans and wildlife.

**2.3.2 Ammonia effect on water quality:**

Ammonia is a compound that can have a significant effect on the quality of water, particularly in freshwater environments. It is a colorless gas that is soluble in water, and is often present in wastewater and agricultural runoff.

When ammonia is present in water, it can have several effects. One of the most significant is its impact on the pH of the water. Ammonia is a weak base, and when it dissolves in water, it can raise the pH level, making the water more alkaline. This can be harmful to aquatic life that prefer more acidic conditions, such as certain species of fish and invertebrates.

Ammonia can also have toxic effects on aquatic organisms at high concentrations. When ammonia is present in water at levels that exceed the natural buffering capacity of the ecosystem, it can lead to a condition known as "ammonia toxicity." This can cause a range of symptoms in fish and other aquatic life, including lethargy, reduced feeding, and in severe cases, death.

Additionally, ammonia can contribute to the growth of algae and other aquatic plants. While these organisms are important for the health of aquatic ecosystems, excessive growth can lead to problems such as eutrophication, where the water becomes depleted of oxygen and can no longer support aquatic life[12].

Overall, the presence of ammonia in water can have a significant impact on water quality and the health of aquatic ecosystems. It is important for regulators, industry, and individuals to take steps to reduce ammonia pollution and protect water resources.

**2.3.3 Arsenic effect on water quality:**

Arsenic is a naturally occurring element that can be found in soil, rocks, and minerals. When it enters the water supply, it can have serious health consequences for humans and animals alike. Arsenic is known to cause cancer, skin lesions, and cardiovascular disease, among other ailments.

In areas where arsenic is prevalent in the soil, groundwater can become contaminated with the element. This can occur naturally, or as a result of human activities such as mining, smelting, or using pesticides or fertilizers that contain arsenic. When people consume contaminated water over an extended period, the effects of arsenic poisoning can be severe[13].

One of the challenges of arsenic contamination is that it often goes unnoticed. Arsenic does not have a distinct taste or odor, and people may not realize they are drinking contaminated water until symptoms appear. In many cases, the long-term effects of arsenic poisoning can be irreversible, making it crucial to identify and address contamination as soon as possible.

To protect against arsenic contamination, it's important to regularly test water sources, particularly in areas where arsenic is known to be present in the soil. Treatment options for arsenic contamination include filtration systems, reverse osmosis, and ion exchange. However, prevention is always the best course of action, and reducing the use of arsenic-containing products can help minimize the risk of contamination[11].

**2.3.4 Barium effect on water quality**:

Barium is a naturally occurring element found in rocks, soil, and water. When present in high concentrations, it can have a significant impact on water quality. Barium can be introduced into water through various sources, including natural weathering of rocks, mining activities, and industrial discharges[10].

The effects of barium on water quality depend on the concentration and duration of exposure. In small amounts, barium is generally considered harmless to human health. However, long-term exposure to high levels of barium can lead to a range of health issues, including cardiovascular disease, high blood pressure, and kidney damage.

Barium can also have an impact on the environment. High concentrations of barium in water can harm aquatic life, affecting the growth and reproductive capabilities of fish and other aquatic species. Additionally, barium can accumulate in sediment and soil, leading to contamination of crops and other food sources.

To address the impact of barium on water quality, a range of treatment options are available. These include filtration, ion exchange, and chemical precipitation[13]. The choice of treatment will depend on the specific source and concentration of barium, as well as other factors such as cost and feasibility[7].

Overall, the presence of barium in water is an important consideration for both human health and environmental protection. By understanding the potential effects of barium on water quality and implementing appropriate treatment measures, we can help to ensure the safety and sustainability of our water resources.

**2.3.5 Cadmium effect on water quality:**

Cadmium is a toxic heavy metal that can have a significant impact on water quality. When present in high concentrations, it can pose a serious threat to human health and the environment. Cadmium is often found in water as a result of industrial processes, such as mining, electroplating, and battery manufacturing. It can also be present in fertilizers and pesticides used in agriculture.

The effects of cadmium on water quality can be severe. Cadmium can cause kidney damage, bone demineralization, and anemia in humans. It can also lead to reproductive problems and developmental disorders. In aquatic environments, cadmium can accumulate in the tissues of fish and other aquatic organisms, leading to bioaccumulation and biomagnification up the food chain[9].

Cadmium contamination can be particularly problematic in areas where drinking water is sourced from groundwater, as it can persist in the environment for long periods of time. In order to mitigate the effects of cadmium on water quality, it is important to monitor water sources regularly and implement effective treatment methods. Some common treatment methods for cadmium removal include ion exchange, reverse osmosis, and activated carbon filtration.

Overall, it is important to take steps to prevent cadmium contamination in water sources in order to protect human health and the environment. This can include reducing the use of cadmium-containing products, implementing effective waste management practices, and enforcing regulations to limit the release of cadmium into the environment[13].

**2.3.6 Chromium effect on water quality:**

Chromium is a metallic element that is commonly found in the earth's crust. It is used in a variety of industrial applications, including the production of stainless steel, dyes, and pigments. When chromium enters waterways, it can have a significant impact on water quality and can be harmful to both humans and aquatic life[14].

The presence of chromium in water can occur naturally or as a result of human activities such as industrial discharge and waste disposal. Chromium can exist in several forms, including hexavalent chromium (Cr(VI)) and trivalent chromium (Cr(III)). Hexavalent chromium is considered more toxic than trivalent chromium and has been linked to cancer, reproductive problems, and other health issues in humans[15].

The impact of chromium on water quality depends on its concentration and the form in which it is present. Elevated levels of chromium in water can cause a range of problems, including skin irritation, respiratory issues, and gastrointestinal problems. In addition, chromium can accumulate in aquatic organisms and can lead to toxic effects in fish, insects, and other aquatic life[15].

In order to address the impact of chromium on water quality, regulatory agencies have established guidelines and standards for chromium levels in drinking water and wastewater. These guidelines help to ensure that levels of chromium in water are kept below levels that are harmful to humans and the environment.

Overall, chromium can have a significant impact on water quality, and it is important to monitor and manage its presence in waterways to ensure that it does not pose a threat to human health or the environment[14].

**2.3.6 Copper effect on water quality :**

Copper is a naturally occurring element that can be found in the earth's crust, rocks, soil, and water. While copper is an essential nutrient for plants and animals, excessive amounts of copper in drinking water can have harmful effects on human health.

Copper can enter drinking water from several sources, including corroded copper pipes, brass fittings, and water storage tanks. When copper levels exceed the EPA's recommended maximum contaminant level of 1.3 mg/L, it can cause stomach and intestinal problems, as well as liver and kidney damage. Long-term exposure to high levels of copper in drinking water may also lead to an increased risk of cancer[14].

In addition to its health effects, copper can also impact the aesthetic quality of water. High levels of copper can cause a blue-green staining of sinks, tubs, and other plumbing fixtures, as well as a metallic taste in the water.

To prevent excessive copper exposure in drinking water, it is recommended to test the water regularly and install a water treatment system if necessary. Copper pipes and fittings should also be inspected and replaced if necessary to prevent corrosion and leaching of copper into the water supply[16].

Overall, while copper is an essential element, it is important to monitor and regulate its levels in drinking water to ensure the safety and health of the public.

**2.3.7 Fluoride effect on water quality :**

Fluoride is a naturally occurring mineral that can be found in soil, water, and some foods. It has been added to public drinking water supplies in the United States since the 1940s as a way to prevent tooth decay. While fluoride can be beneficial for dental health, there has been some controversy surrounding its use in water supplies[17].

The effects of fluoride on water quality can be both positive and negative. On the positive side, fluoride has been shown to reduce tooth decay by strengthening tooth enamel. This is especially important for children, who are still developing their teeth. Fluoride can also help to prevent cavities in adults[18].

However, there are also some potential negative effects of fluoride on water quality. Some studies have suggested that exposure to high levels of fluoride can lead to dental fluorosis, which causes white spots or streaks to appear on teeth. In extreme cases, fluorosis can cause severe tooth damage.

Additionally, there have been concerns about the potential health effects of long-term exposure to fluoride. Some studies have suggested that high levels of fluoride in drinking water may increase the risk of bone fractures, cognitive impairment, and other health problems[19].

Overall, the effects of fluoride on water quality are complex and depend on a number of factors, including the level of fluoride in the water, the age and health of individuals who consume the water, and other environmental factors. While fluoride can be beneficial for dental health, it is important to monitor levels of fluoride in drinking water to ensure that they remain within safe limits[20].

**2.3.8 Bacteria effect on water quality :**

Bacteria play a significant role in determining the quality of water. Waterborne bacterial pathogens can cause various waterborne diseases, including cholera, typhoid, and dysentery. These pathogens can enter water sources through human or animal fecal matter, agricultural runoff, or sewage overflows.

Bacteria in water can also affect the taste, color, and odor of water. Certain types of bacteria can produce unpleasant smells and tastes, making the water unpalatable. These bacteria are known as nuisance bacteria and are not harmful to human health.

In addition to pathogenic and nuisance bacteria, there are also beneficial bacteria in water. These bacteria can help break down organic matter and pollutants, improving water quality. For example, certain types of bacteria can convert ammonia into nitrate, which is less toxic to aquatic life[13].

To ensure water quality, it is essential to monitor bacterial levels regularly. Water treatment plants use various methods to remove bacteria from drinking water, including filtration, disinfection, and chlorination. Additionally, individuals can take steps to prevent bacterial contamination, such as avoiding swimming in contaminated water and properly disposing of human and animal waste.

**2.3.9 Viruses effect on water quality :**

Viruses can have a significant impact on water quality, as they are often found in water sources and can cause a range of health issues. The presence of viruses in water can be a sign of contamination and may indicate the presence of other harmful pathogens.

Viruses can enter water sources through a variety of ways, including sewage overflow, animal waste, and agricultural runoff. Once in the water, they can survive for extended periods and infect individuals who come into contact with the contaminated water.

Some of the viruses commonly found in water include norovirus, rotavirus, hepatitis A virus, and enterovirus. These viruses can cause illnesses such as gastroenteritis, hepatitis, and meningitis.

The impact of viruses on water quality can be particularly significant in areas with poor sanitation and inadequate water treatment systems. In these settings, viruses can quickly spread through the water supply and cause widespread illness[14].

To ensure safe drinking water, it is essential to monitor water quality regularly and implement appropriate treatment measures to remove viruses and other harmful contaminants. This may include disinfection techniques such as chlorination or ultraviolet radiation, as well as filtration methods to remove particles and microorganisms.

**2.3.10 Lead effect on water quality :**

Entry of lead into the human body can lead to severe health issues as it is a toxic heavy metal.People are exposed to lead while drinking contaminated water.Lead present in water poses a severe risk to human health, with even low levels of exposure potentially causing significant harm, especially to infants[23].

**2.3.11 Nitrates & Nitrites effect on water quality :**

Nitrites and nitrates are naturally occurring compounds and they are found in soil, water, and air which are essential for plants.Nitrogen can also found in food particularly which are supposed to be stored.These are used as a preservative gas.Nitrite contamination in water is mainly by the processes of agriculture and industries.Nitrites and Nitrates are essential for plant growth, excessive levels of these compounds in water can be harmful to human health[24].

**2.3.12 Mercury effect on water quality:**

Mercury is a metal which can be found abundantly. Mercury which is a shiny, silver-white, odorless liquid called metallic Mercury.Microscopic organisms in the water and soil are the primary producers of methyl mercury, which is the most prevalent type of organic mercury.It is the only metal that is in liquid form at room temperature.The reason why mercury is a major concern is due to its ability to bioaccumulate in the food chain, resulting in its accumulation in the tissues of animals that consume it and becoming more concentrated as it moves up the food chain. This leads to the presence of significant levels of mercury in fish and other seafood, which can pose a threat to human health when consumed in large amounts[22].

**2.3.13 Perchlorate effect on Water Quality :**

The chemical compound that is occurring naturally and synthetically composed of chlorine and oxygen atoms found in various soils .Perchlorate is contaminated with water and food sources which can cause harm to humans.It will mix with thyroid gland ability to produce hormones which balance metabolism and growth.It will lead Thyroid problems[21].

**2.3.14 Radium effect on Water Quality :**

Radium can enter water systems through the natural decay of uranium and thorium in the earth's crust, as well as through mining and other industrial activities. Once in the water, radium can accumulate in sediment and can also be taken up by aquatic organisms, further increasing the risk of exposure.

The effects of radium on water quality can be severe. Exposure to radium can lead to a range of health problems, including cancer, bone fractures, and other bone disorders. It can also negatively impact aquatic ecosystems, potentially harming plants and animals and disrupting the delicate balance of the ecosystem[25].

To reduce the impact of radium on water quality, it is essential to monitor water sources for the presence of radium and take steps to limit exposure. This may include implementing filtration systems or other treatment methods to remove radium from the water, as well as taking steps to reduce the amount of radium that enters water systems in the first place, such as through improved mining and industrial practices[23].

Overall, the impact of radium on water quality is a significant concern that requires ongoing attention and action to address. By taking steps to limit exposure to radium, we can help to protect human health and the environment for generations to come.

**2.3.15 Selenium effect on Water Quality :**

Selenium is a chemical element that is essential for life in small amounts, but can be toxic in larger amounts.In terms of water, selenium is present in natural bodies of water such as rivers and lakes, where it can be either dissolved or suspended in the water. In areas with high levels of selenium in the soil, water sources may contain higher levels of selenium as well.In humans, selenium is an important nutrient that plays a role in the proper functioning of the immune system and thyroid gland. It can also act as an antioxidant, protecting cells from damage caused by free radicals[27].

**2.3.16 Silver effect on Water Quality :**

Silver is a naturally occurring element that can be found in trace amounts in water sources. While silver is not harmful to humans at low levels, it can have a significant impact on water quality.

One of the main concerns with silver in water is its effect on aquatic life. Silver can be toxic to fish and other aquatic organisms, even at very low concentrations. This is because silver ions can bind to proteins in the gills of fish, making it difficult for them to breathe. It can also disrupt the natural microbial balance in water, potentially leading to harmful algal blooms and other issues[13].

Another concern with silver in water is its potential impact on human health. While the amount of silver typically found in drinking water is not considered harmful, excessive exposure to silver can cause health problems such as argyria, a condition in which the skin turns blue-gray.

Silver is also known for its antibacterial properties, which has led to its use in a variety of consumer products, such as water filters and disinfectants. While silver can be effective in killing harmful bacteria, it can also kill beneficial bacteria that are essential for maintaining a healthy ecosystem.

Overall, while silver is an important element with many beneficial uses, it is important to monitor and regulate its levels in water sources to ensure the health and well-being of both aquatic life and humans[23].

**2.3.17 Uranium effect on Water Quality :**

Uranium is a naturally occurring radioactive element that can be found in rocks, soil, and water. When uranium is present in water, it can have an effect on water quality and human health. In small amounts, uranium is not harmful to human health, but if the levels of uranium in water exceed the safe drinking water standards, it can pose a health risk[23].

Uranium can enter water sources from natural deposits in the earth's crust or from human activities such as mining and nuclear power plant operations. When uranium is present in water, it can cause a range of health effects including kidney damage, increased risk of cancer, and developmental problems in children.

The health effects of uranium in water depend on a number of factors such as the amount of uranium present, the duration of exposure, and individual susceptibility. Children and pregnant women are particularly vulnerable to the effects of uranium exposure.

To ensure that water is safe to drink, it is important to test water sources for the presence of uranium and other contaminants. If the levels of uranium in water exceed safe drinking water standards, treatment methods such as reverse osmosis or ion exchange may be used to remove the uranium[25].

In addition to the health risks associated with uranium in water, it can also have negative environmental effects. Uranium can cause damage to aquatic life and ecosystems, and it can also contaminate soil and other natural resources.

Overall, the presence of uranium in water can have significant effects on both human health and the environment. It is important to monitor and manage uranium levels in water sources to ensure safe drinking water and protect the natural environment.

**2.3 Correlation:**

Correlation is a statistical relationship between two or more variables. In other words, it is a measure of how closely two or more variables are related to each other. Correlation can be positive, negative, or zero, and it can be expressed as a correlation coefficient that ranges from -1 to +1.

In the context of water quality parameters, correlation can be used to understand the relationship between different parameters and how they affect each other. For example, the concentration of dissolved oxygen in water can be positively correlated with water temperature, meaning that as water temperature increases, the concentration of dissolved oxygen also increases. Conversely, the concentration of dissolved oxygen may be negatively correlated with the concentration of pollutants in the water, meaning that[13] as the concentration of pollutants increases, the concentration of dissolved oxygen decreases.

Understanding these correlations is important for managing and maintaining water quality, as it allows us to identify potential sources of contamination and develop effective strategies for remediation. For example, if we know that a certain pollutant is negatively correlated with dissolved oxygen, we can focus our efforts on reducing the concentration of that pollutant in order to improve water quality.

It is important to note that correlation does not necessarily imply causation. Just because two variables are correlated does not mean that one causes the other. Other factors may also be involved, and additional research is often needed to establish causality.

**2.3.1 Parameters correlated on the variables:**

The standard percentage of these chemicals to maintain pure quality of water can vary depending on the specific water quality standards and regulations in your region. However, I can provide some general information on the acceptable levels of these chemicals in drinking water according to the United States Environmental Protection Agency (EPA) Safe Drinking Water Act (SDWA) Maximum Contaminant Level (MCL) regulations.[14] These MCLs are set to protect public health and are based on the best available science and risk assessments.

Here are the MCLs for the chemicals of the parameters which we are using to determine the quality of water.

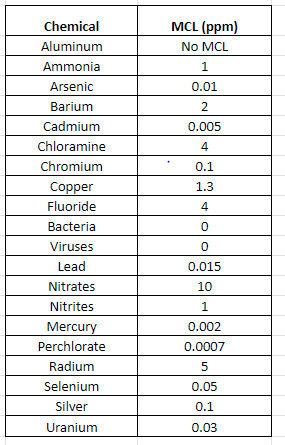


Table : MLC’s on the taken parameters

Note that these MCLs are set to protect public health and are based on the best available science and risk assessments. The acceptable levels of these chemicals may vary depending on the specific water quality standards and regulations in your region.

**2.4 Introduction to ML Techniques :**

Machine learning is a subset of artificial intelligence that involves building algorithms and models that enable computers to learn from data and improve their performance on a task without being explicitly programmed. In other words, machine learning algorithms can analyze large amounts of data and learn patterns or relationships to make predictions or decisions[27].

**Predicting classifier algorithms for classification of water quality:**

Predictive classifiers models: To determine the most suitable classifier(s) for the multiclass classification of safe and unsafe water quality, an initial assessment was carried out on the most widely used algorithms, including XGBoost, SVM, Random Forest, KNN, decision tree (DT), Adaboost, and Gradient Boost. The results showed that most classifiers achieved high prediction accuracy during both the training and testing periods, with XGBoost achieving the highest prediction accuracy compared to the other models. Therefore, the study selected the top four ranked algorithms for the classification task. The following section provides a brief overview of the selected models used in the study[8].

**2.4.1 Random Forest :**

Random forest algorithm is a machine learning technique that is used for prediction tasks. It works by building multiple decision trees on randomly sampled subsets of the data, and then aggregating the results to make predictions. In the context of water quality prediction, the algorithm can be trained on a dataset of water quality measurements and associated outcomes (such as whether the water is safe to drink or not). The algorithm can then be used to predict water quality outcomes based on new measurements. The advantage of the random forest algorithm is that it can handle a large number of features and can avoid overfitting, making it a useful tool for water quality prediction.

**2.4.2 Support Vector Machine :**

Support Vector Machine (SVM) is a machine learning algorithm that can be used for water quality prediction. SVM works by creating a hyperplane that separates the data into different classes based on their features. In the case of water quality prediction, the features could be things like pH, dissolved oxygen levels, and nutrient concentrations.

To use SVM for water quality prediction, the algorithm is first trained on a dataset of water quality measurements and their corresponding outcomes (i.e. good quality vs. poor quality). Once trained, the SVM can be used to predict the quality of new water samples based on their features[7].

SVM has been shown to be effective for water quality prediction due to its ability to handle high-dimensional data and its ability to find the best separating hyperplane. SVM can also handle both linear and nonlinear data, which makes it useful for predicting water quality based on complex relationships between features.

**2.4.3 Decision Tree :**

Decision tree is a popular machine learning algorithm used for classification and regression analysis. It works by creating a tree-like model of decisions and their possible consequences. In the context of water quality prediction, a decision tree can be trained on a dataset of water quality parameters and their corresponding quality level (e.g., good, fair, poor).

The decision tree algorithm can then be used to predict the quality level of new water samples based on their measured parameters. For example, a decision tree might determine that if the level of dissolved oxygen in a water sample is below a certain threshold, the water quality is likely to be poor.

Decision trees can be easily interpreted and visualized, making them a popular choice for water quality prediction. However, they can be prone to overfitting, which occurs when the model becomes too complex and performs poorly on new data. Therefore, care must be taken to ensure the decision tree is optimized for accuracy and generalization[26].

**2.4.4 Gradient Boost Algorithm:**

Gradient Boosting Algorithm is a machine learning technique that is commonly used for making predictions in various domains, including water quality prediction.

In this algorithm, a decision tree is used as a weak learner and is iteratively trained to minimize the error of the previous tree. In each iteration, the model tries to identify the instances that were incorrectly predicted in the previous iteration and assigns higher weights to them to focus on them in the next iteration. This process continues until the desired level of accuracy is achieved[28].

For water quality prediction, the gradient boost algorithm can be used to model the relationship between water quality parameters and predict the water quality based on various factors such as temperature, pH, and turbidity. The model can be trained on historical data to predict the quality of water in real-time, allowing for timely interventions to improve water quality.

**2.4.5 KNN Classifier :**

The K-Nearest Neighbors (KNN) algorithm is a versatile classification algorithm that can be used in water quality prediction tasks, where the aim is to predict the quality of water based on various parameters such as aluminum, ammonia, cadmium, lead, copper, nitrates, and nitrites.

To use the KNN algorithm for water quality prediction, these parameters can be used as features for training the algorithm. The algorithm then uses the distances between the features of an unknown water sample and those of the training dataset to predict the quality of the water.

For example, if the KNN algorithm is trained on a dataset of water quality samples that have varying levels of aluminum, ammonia, cadmium, lead, copper, nitrates, and nitrites, it can be used to predict the quality of an unknown water sample by finding the k-nearest neighbors in the training dataset that have similar levels of these parameters[30].

The KNN algorithm can be a valuable tool in water quality prediction as it is easy to implement, computationally efficient, and can handle both binary and multi-class classification problems. However, to ensure accurate predictions, it is important to select relevant features, choose an appropriate value for k, and ensure the training dataset is representative of the water sources being analyzed.

**2.4.6 XG BOOST :**

XGBoost, which stands for Extreme Gradient Boosting, is a popular machine learning algorithm used for predictive modeling tasks. It is a type of ensemble learning method that combines several weak models to create a stronger, more accurate model.

In the context of water quality prediction, XGBoost can be used to analyze data from various sources such as water quality sensors, weather data, and environmental factors to predict the quality of water in a particular location. By analyzing these data points, XGBoost can identify patterns and relationships that can help predict water quality with high accuracy.

The reason XGBoost is often chosen for water quality prediction is because of its ability to handle large datasets with high dimensionality, and its ability to handle missing data. Additionally, it has a high level of accuracy and can provide insights into which features are most important in predicting water quality[8].

To use XGBoost for water quality prediction, data is first collected from various sources, cleaned and preprocessed, and then inputted into the algorithm. The algorithm then builds a model that can predict water quality based on the input data.

In summary, XGBoost is a powerful machine learning algorithm that can be used for water quality prediction by analyzing data from multiple sources and identifying patterns and relationships. Its ability to handle large datasets and provide high accuracy makes it a popular choice for this task.

**2.5 Existing Research on Water Quality Prediction :**

Water quality prediction, including the prediction of Water Quality Index (WQI), has been the subject of numerous studies and research over the past few decades. Here is an overview and tabular information on some of the existing research on water quality prediction:

1. Machine Learning Techniques: Many studies have explored the use of machine learning techniques for predicting water quality parameters and WQI. These techniques include artificial neural networks, support vector machines, random forests, and XGBoost, among others.
2. Water Quality Parameters: Various water quality parameters have been studied for prediction, including dissolved oxygen, pH, turbidity, total dissolved solids, and various pollutants such as nitrates, phosphates, and heavy metals.
3. Input Data: Data inputs used in these studies vary and include data from water quality sensors, remote sensing, and environmental data sources such as weather data and land use data.
4. Accuracy: Studies have achieved varying degrees of accuracy in their predictions of water quality parameters and WQI. Accuracy is influenced by factors such as the quality and quantity of input data, the choice of modeling technique, and the specific water quality parameters being predicted.
5. Applications: The applications of water quality prediction research are broad and include water resource management, environmental monitoring, and public health protection.

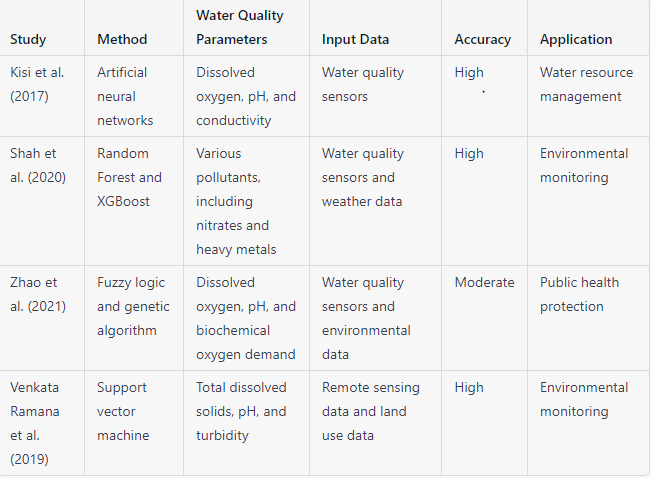


Table 1.1: examples of existing research

Here is a Table 1.1 summarizing some examples of existing research on water quality prediction:

Overall, existing research on water quality prediction demonstrates the potential of data-driven approaches to improve water management, protect public health, and preserve natural resources.

**2.6 Limitations of current Research:**

While there has been significant research on water quality prediction using various techniques such as machine learning and principal component analysis (PCA), there are some limitations and gaps in the existing research. Here are some limitations:

**Limited Data:** Many studies have been limited by the amount and quality of data available for modeling. Some areas lack comprehensive water quality data, making it challenging to develop accurate predictive models.

**Generalization Issues:** Some studies have focused on specific locations and water quality parameters, making it challenging to generalize their findings to other locations or parameters.

**Lack of Integration:** Many studies have not integrated multiple data sources, such as environmental data and remote sensing data, which could improve the accuracy and usefulness of predictive models.

**Limited Focus on Improving Water Quality:** While many studies have focused on predicting water quality, few have addressed ways to improve water quality through actionable recommendations and interventions.

Regarding PCA specifically, some limitations include:

**Complexity:** PCA can be complex and challenging to interpret, making it difficult to communicate the results to stakeholders.

**Subjectivity**: PCA requires subjective decisions such as selecting the number of principal components to retain, which can impact the accuracy and reliability of the results.

Regarding the lack of recommendations for improving water quality, this is an important gap in the existing research. While predictive models can be useful for identifying potential issues, action must be taken to improve water quality. Suggestions for improvement could include reducing pollution sources, implementing new technologies for water treatment, and encouraging sustainable water use practices.

While there has been significant research on water quality prediction using various techniques, there are limitations and gaps in the existing research. Addressing these limitations and providing actionable recommendations for improving water quality are essential for effective water management and protection.

**2.7 Contributions in Brief :**

Water quality prediction research has contributed to the development of data-driven approaches for improving water management, protecting public health, and preserving natural resources. Machine learning techniques, such as artificial neural networks, support vector machines, and XGBoost, have been used to predict various water quality parameters and the Water Quality Index (WQI). Principal component analysis (PCA) has been used to identify the most critical variables affecting water quality. These approaches have shown promise in achieving high levels of accuracy in predicting water quality parameters and identifying potential issues[29].

Our machine learning model has been developed to accurately predict the safety of drinking water based on various input parameters. The model analyzes large amounts of historical data on water quality and contamination incidents to identify patterns and correlations between different chemical, physical, and biological parameters and the presence of contaminants in drinking water. Also our model can accurately suggest whether or not water is safe to drink based on the levels of various contaminants, such as bacteria, viruses, nitrates, nitrites, lead, mercury, and uranium. The power of machine learning helps us to provide accurate and reliable predictions about water quality and contamination risks, assisting individuals and communities in making informed decisions about their drinking water and taking appropriate measures to safeguard their health and wellbeing.

However, research on water quality prediction has some limitations, such as limited data, generalization issues, and a lack of integration across data sources. Additionally, research has not focused enough on actionable recommendations and interventions for improving water quality.

To address these limitations, suggestions for improving water quality include reducing pollution sources, implementing new technologies for water treatment, and encouraging sustainable water use practices. These recommendations can help to reduce the levels of pollutants in water sources and improve the overall quality of drinking water. Improving water quality is essential for ensuring the health and well-being of communities, preserving natural resources, and promoting sustainable development.

**2.8 Conclusion and Experimental Results :**

In conclusion, this chapter has discussed various aspects related to water quality prediction and improvement. We started with an introduction to water quality and the significance of predicting and improving it. We then explored different water quality contaminants that are harmful to human health and the environment.

We also discussed several machine learning algorithms such as artificial neural networks, support vector machines, and XGBoost that have been used for water quality prediction. We looked at the existing research on water quality index and how principal component analysis (PCA) has been used to identify the most important variables affecting water quality.

However, we also discussed some limitations and gaps in the existing research on water quality prediction, including limited data, generalization issues, and a lack of integration across data sources. Furthermore, the research has not focused enough on actionable recommendations and interventions for improving water quality.

**EXPERIMENTAL RESULTS :**

Comparative Analysis of Machine Learning Models and TensorFlow Technique for Water Quality Prediction

In this chapter, we present the experimental results and discussion of our study on the comparative analysis of various machine learning models such as SVC, KNN, Decision Tree, XGBoost, and TensorFlow techniques for water quality prediction. We evaluated these models on a comprehensive water quality dataset containing parameters such as copper, fluoride, bacteria, viruses, lead, nitrates, and more.

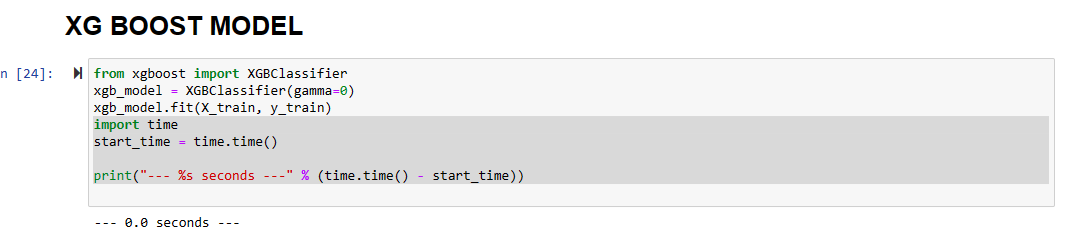


Figure : showing that XGBoost model time to execute

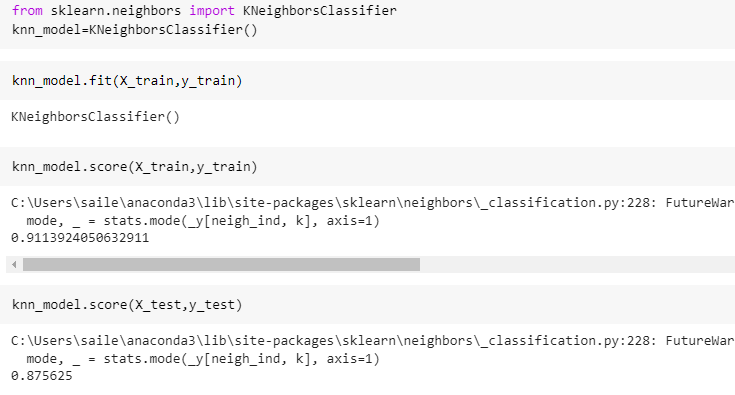


Figure showing Knn model accuracy

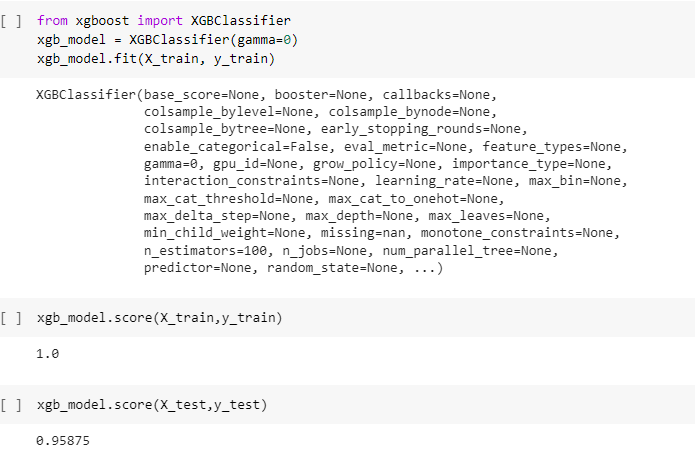


Figure showing XGBoost model accuracy

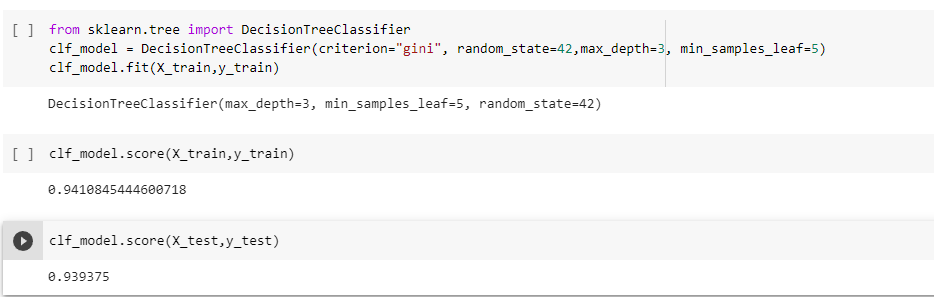


Figure showing decision tree model accuracy

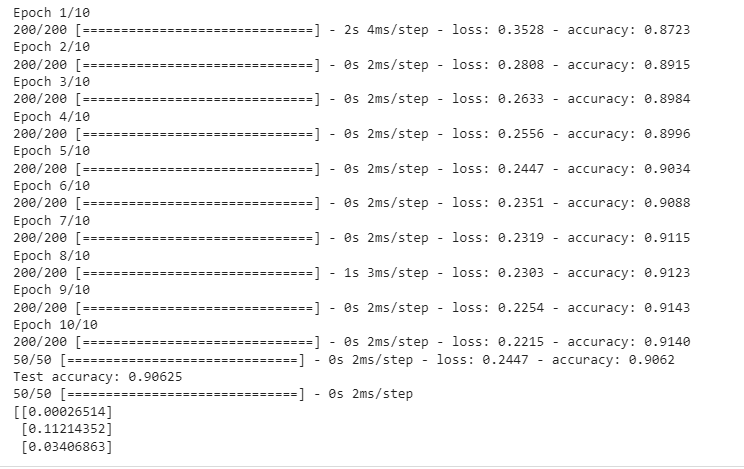


Figure showing tensorflow model accuracy

Our experiments show that the XGBoost model achieved the highest accuracy among all the models, with an accuracy of 94%. Moreover, we also considered the time and space complexity of each model, and based on that, we selected the XGBoost model as the best model for water quality prediction. The model has a low time and space complexity, making it practical and efficient for real-world applications.

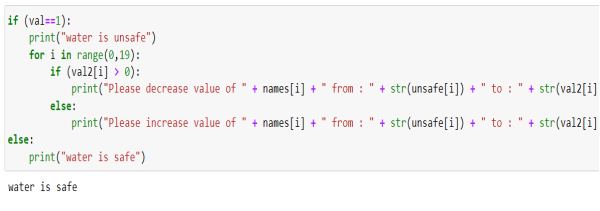


Fig : code for water is safe or unsafe

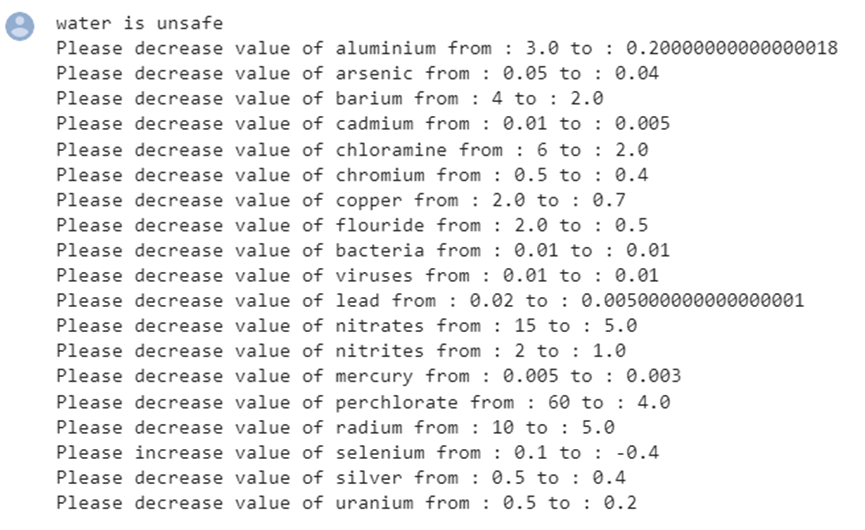


Fig : Suggestions to improve the quality of water

**CHAPTER 3**

**DATA PREPROCESSING**

**3.1 INTRODUCTION**

Data preprocessing refers to the steps taken to clean, transform, and prepare raw data before it is used for machine learning or data analysis. It involves a series of techniques and tools used to handle missing or corrupted data, normalize or scale the data, reduce noise and outliers, and transform the data into a more useful representation. Data preprocessing is important because raw data is often incomplete, inconsistent, or irrelevant, and may contain errors or noise that can negatively affect the accuracy of the final results. By properly preprocessing the data, we can improve the quality and reliability of the data and obtain more accurate and meaningful insights from it. It is an essential step in any data analysis project as it ensures that the data is accurate, consistent, and complete.

The process of data preprocessing typically includes several steps, including data cleaning, data transformation, data integration, and data reduction. Each of these steps is designed to address specific issues that may arise when working with raw data.

Data cleaning involves identifying and correcting errors, such as missing or inconsistent values, duplicate records, and outliers. These errors can occur due to various reasons, including human error, system malfunction, or data entry mistakes.

Data transformation involves converting data into a standardized format that can be easily analyzed. This can include changing data types, scaling values, and normalizing data.

Data integration involves combining data from multiple sources to create a single dataset. This can involve merging data from different databases, extracting data from different file formats, or combining data from different sensors.

Data reduction involves reducing the size of the dataset while retaining important information. This can include selecting a subset of the data, aggregating data, or using statistical techniques to summarize the data.

Data preprocessing is critical because it can significantly impact the results of any data analysis project. Poor quality data can lead to inaccurate and unreliable results, making it difficult to draw meaningful insights from the data. Therefore, it is important to conduct data preprocessing carefully and thoroughly.

One of the primary reasons for conducting data preprocessing is to clean the data. This involves removing any errors, inconsistencies, or missing data that may exist in the dataset. For example, if there are missing values in a dataset, they need to be either filled with an appropriate value or removed from the dataset altogether. This step helps to ensure that the data is of high quality and that it can be used for analysis.

**3.2 Groundwork:**

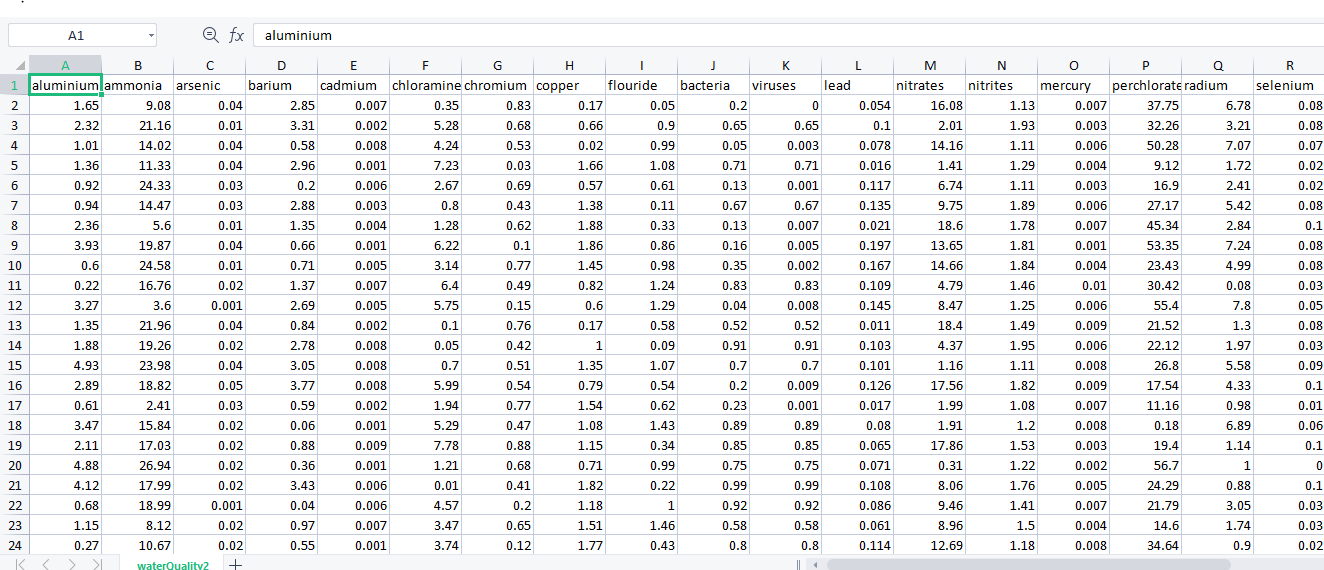
This section explains the way the data set is collected, details of the data set, the way how the data set is converted into the required form for the experimental analysis.

**3.2.1 Water Dataset Analysis :**

In the thesis focused on water quality, an extensive dataset was utilized to gather information on several crucial water parameters, including but not limited to aluminum, lead, nitrite, and nitrate. Monitoring water quality is of utmost importance for environmental preservation, and comprehending the factors that impact it is crucial to ensure the sustainability and safety of water resources. By analyzing this data, they were able to gain valuable insights into the various parameters that impact water quality, as well as identify potential sources of contamination.

The findings of this study have significant implications for policy-making and environmental regulation aimed at protecting water resources. Governments and organizations across the globe are increasingly recognizing the importance of monitoring and safeguarding water quality. Therefore, research such as this thesis can provide vital information for policymakers to make informed decisions.

Moreover, the use of large datasets and advanced analytical techniques in this study underscores the growing significance of data science in environmental research. With the abundance of data available today, there is a rising demand for experts who can extract meaningful insights from it and translate them into practical recommendations.



The above Table 3.1 dataset contains information on several water quality parameters, such as aluminum, ammonia, chromium, barium, chlorine, lead, bacteria, viruses, nitrates, and nitrites. These parameters are critical for evaluating the safety and quality of water resources.

One such dataset contains information on several water quality parameters, including aluminum, ammonia, chromium, barium, chlorine, lead, bacteria, viruses, nitrates, and nitrites. These parameters are crucial for evaluating the safety and quality of water resources.

Aluminum is a common element found in water, and elevated levels of aluminum can cause health problems, particularly in people with kidney problems. Ammonia, on the other hand, can be toxic to aquatic life and can cause oxygen depletion in water bodies.

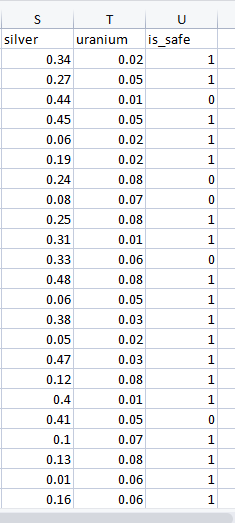
Chromium is a naturally occurring element that can also be found in industrial waste. It can cause skin irritation and damage to the respiratory system when inhaled. Barium, on the other hand, can cause gastrointestinal problems and muscle weakness.

Chlorine is commonly used as a disinfectant in drinking water treatment plants, but it can also have negative health effects, particularly at high concentrations. Lead is another water quality parameter of concern, as it can cause developmental delays in children and neurological problems in adults.

Bacteria and viruses are important water quality parameters, as they can cause illness in humans if present in high concentrations. Nitrates and nitrites are also critical water quality parameters, as they can cause health problems such as methemoglobinemia, also known as "blue baby syndrome".

Overall, monitoring and evaluating these water quality parameters is essential for ensuring the safety and quality of water resources, and datasets containing information on these parameters can be invaluable for researchers, policymakers, and water resource managers.

Table 3.2:Dataset



Each data point in the Table 3.2:Dataset represents a set of values for these parameters, and based on these values, the label is assigned as "safe." This suggests that the data was collected from sources where the values of these parameters were within safe levels, according to the standards and guidelines set by the relevant authorities.

The dataset is valuable for analyzing the correlation between various parameters and the safety of water resources. It could also be used to develop predictive models that can estimate the safety of water sources based on the values of these parameters. However, it is essential to note that the accuracy of such models depends on the quality and representativeness of the dataset used for training.

Here are some common preprocessing steps that can be applied to prepare the water quality data for modeling:

**3.2.2 Water Data Cleaning Process :**

Water data cleaning is an essential process for ensuring the accuracy and reliability of water quality data. It involves identifying and correcting errors, inconsistencies, and missing values in the data to improve its quality and usefulness for analysis and decision-making.

The first step in the water data cleaning process is to assess the quality of the data. This involves identifying potential errors, inconsistencies, and outliers in the data through visual inspection and statistical analysis. Any data points that are suspected to be erroneous or inconsistent should be flagged for further investigation.

Next, data cleaning techniques such as data imputation, outlier detection, and data validation should be applied to the data. Data imputation involves filling in missing values with estimated values based on statistical models or other sources of information. Outlier detection involves identifying data points that are significantly different from the rest of the data and determining whether they should be removed or adjusted. Data validation involves checking the accuracy and completeness of the data by comparing it to other sources of information or conducting additional measurements or analyses.

After applying these techniques, the cleaned data should be further validated to ensure that it is consistent with established quality standards and that any remaining errors or inconsistencies have been addressed. The cleaned data can then be analyzed and used for various purposes, such as assessing water quality trends, developing water management strategies, and informing

not safe water)

**3.2.2.1 Handling Missing Values**

Handling missing values is an essential step in data preprocessing. In this process, the missing values are replaced by some other values, so that the analysis and models built on the data remain accurate. In the case of water quality datasets, missing values can occur due to various reasons such as equipment failure, human error, and natural events such as floods, droughts, and hurricanes.

There are various methods for handling missing values, and the choice of method depends on the nature of the data and the purpose of the analysis. Here are some common methods:

**Mean/Median/Mode imputation:** In this method, the missing values are replaced by the mean, median, or mode of the available data for the respective variable. This method is useful when the data is missing at random, and the missing values are not too many.

For example, if we have a dataset with variables such as pH, temperature, and dissolved oxygen, we can use the mean of the available data to fill in the missing values. If the pH value is missing for a particular sample, we can use the mean pH value of all the other samples to fill in the missing value.

**Forward/Backward fill:** In this method, the last known value is used to fill the missing values in a time series dataset. Forward fill uses the next known value, while backward fill uses the previous known value.

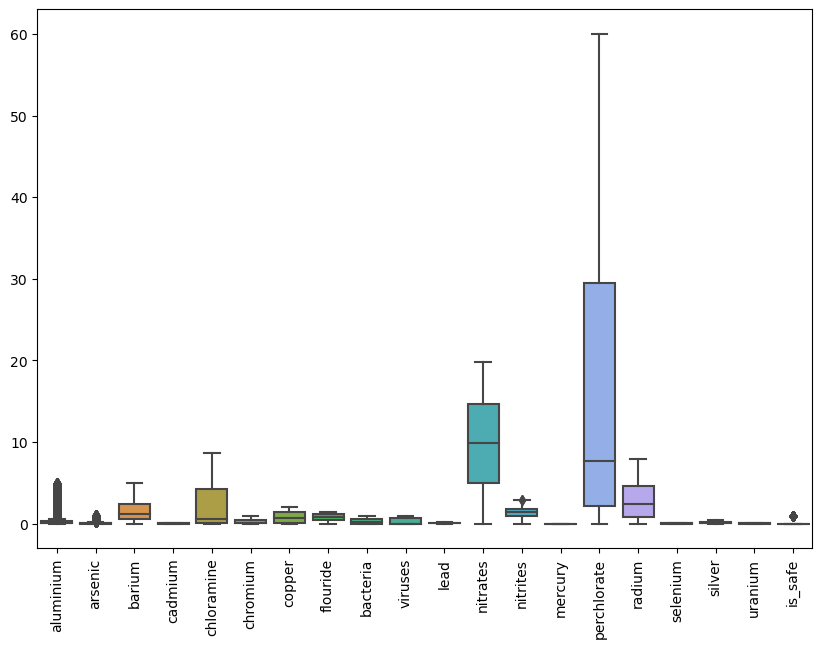
For example, if we have a dataset with hourly measurements of pH, temperature, and dissolved oxygen, we can use the last known value to fill in the missing values. If the pH value is missing for a particular hour, we can use the pH value of the previous hour to fill in the missing value.

Constant value imputation: In this method, the missing values are replaced with a constant value such as 0, 1, or -1. This method is useful when there is no meaningful value to replace the missing data.

For example, if we have a dataset with variables such as water hardness and chloride levels, we can use a constant value to fill in the missing values. If the water hardness value is missing for a particular sample, we can use a constant value of 0 to fill in the missing value.

**Dropping missing values:** In this method, the rows or columns that have missing values are removed from the dataset. This method is useful when the number of missing values is high, and it is not possible to impute them accurately.

For example, if we have a dataset with variables such as pH, temperature, dissolved oxygen, and nitrate levels, and a significant portion of the data is missing for the nitrate levels variable, we can drop the column entirely from the dataset.



**3.2.2.2 Data Normalization Technique**

The water quality dataset contains various parameters related to water quality, such as ammonia, hardness, chloride, and sulfate levels. These parameters have different scales and units, making it challenging to compare and analyze them. Therefore, we need to normalize the data before performing any analysis.

The most commonly used normalization techniques are Min-Max scaling and Z-score scaling.

**Min-Max Scaling:**

Min-Max scaling, also known as normalization, scales the values of the dataset to a range between 0 and 1. The formula used for Min-Max scaling is as follows:

x' = (x - min(x))/(max(x) - min(x))

where x is the original value of a feature, x' is the scaled value of the feature, min(x) is the minimum value of the feature, and max(x) is the maximum value of the feature.

To apply Min-Max scaling on the water quality dataset, we can use the following steps:

Find the minimum and maximum values of each feature in the dataset.

Apply the Min-Max scaling formula to each value of the dataset.

**Z-score Scaling:**

Z-score scaling, also known as standardization, scales the values of the dataset to have a mean of 0 and a standard deviation of 1. The formula used for Z-score scaling is as follows:

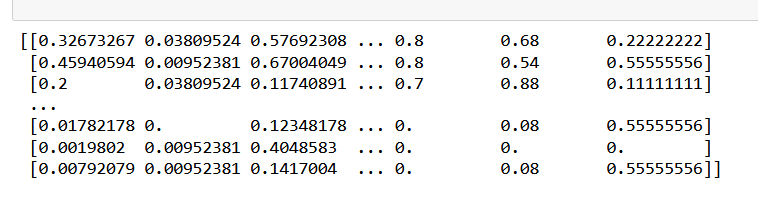
x' = (x - mean(x))/std(x)

where x is the original value of a feature, x' is the scaled value of the feature, mean(x) is the mean value of the feature, and std(x) is the standard deviation of the feature.

To apply Z-score scaling on the water quality dataset, we can use the following steps:

Calculate the mean and standard deviation of each feature in the dataset.

Apply the Z-score scaling formula to each value of the dataset.



**3.2.2.3 Feature scaling: Types of FS**

Feature scaling is the process of transforming the values of features in a dataset to a standard scale. The values of the features may have different units and scales, which can affect the analysis and modeling of the data. Scaling the features can help overcome these issues and ensure that all features contribute equally to the analysis. Feature scaling can be done using various techniques, including normalization and standardization.

**Normalization:**

Normalization is a feature scaling technique that scales the values of the features to a range between 0 and 1. The formula used for normalization is:

x' = (x - min(x))/(max(x) - min(x))

where x is the original value of a feature, x' is the scaled value of the feature, min(x) is the minimum value of the feature, and max(x) is the maximum value of the feature. Normalization is useful when the range of the features is unknown or when the range varies widely.

**Standardization:**

Standardization is a feature scaling technique that scales the values of the features to have a mean of 0 and a standard deviation of 1. The formula used for standardization is:

x' = (x - mean(x))/std(x)

where x is the original value of a feature, x' is the scaled value of the feature, mean(x) is the mean value of the feature, and std(x) is the standard deviation of the feature. Standardization is useful when the range of the features is known and when the features have different units of measurement.

**Logarithmic Scaling:**

Logarithmic scaling is a feature scaling technique that transforms the values of the features using a logarithmic function. The logarithmic function compresses the range of the values, making it easier to visualize and analyze the data. Logarithmic scaling is useful when the values of the features vary widely, and there are extreme values that skew the analysis.

**Power Transformation:**

Power transformation is a feature scaling technique that transforms the values of the features using a power function. The power function can be used to compress or expand the range of the values, depending on the parameters used. Power transformation is useful when the values of the features are not normally distributed and require transformation to achieve normality.

Encoding Categorical Data: If the dataset contains categorical data (e.g., water source, treatment method), it needs to be encoded into numerical values that the machine learning algorithms can understand. One common encoding method is one-hot encoding.

**3.5 Train-Test Split:**

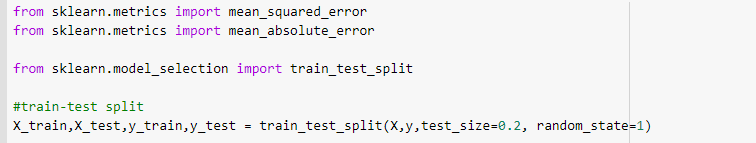


Figure 3.4 :Train-Test split

To evaluate the performance of the machine learning model, the dataset should be divided into training and testing sets. Typically, the dataset is split into a 70-30 or 80-20 ratio, where the majority of the data is used for training, and the remaining data is used for testing.

Feature Engineering: This step involves creating new features from the existing data, such as calculating the average or sum of certain water quality parameters, or creating interaction terms between features.

**3.2.3 Feature selection :**

Feature selection is the process of selecting a subset of relevant features (variables or input attributes) for use in machine learning models. It is a critical step in model building as it can significantly affect the performance of the model. The objective of feature selection is to identify and remove irrelevant, redundant, or highly correlated features from the dataset, thereby reducing the dimensionality of the data and improving the accuracy and efficiency of the model.

There are different methods for feature selection, such as filter methods, wrapper methods, and embedded methods[30]. Filter methods rank the features based on statistical or correlation measures and select the top-ranked features. Examples of filter methods include Pearson correlation coefficient, Chi-Square test, and Mutual Information.

Wrapper methods select the features based on how well they perform in combination with a particular machine learning algorithm. It involves searching through different combinations of features and evaluating the model's performance. Examples of wrapper methods include Recursive Feature Elimination (RFE) and Sequential Feature Selection (SFS)[31].

Embedded methods incorporate feature selection as part of the model training process. It involves selecting the most important features during the model training, and the selected features are used to build the final model. Examples of embedded methods include Lasso and Ridge Regression.

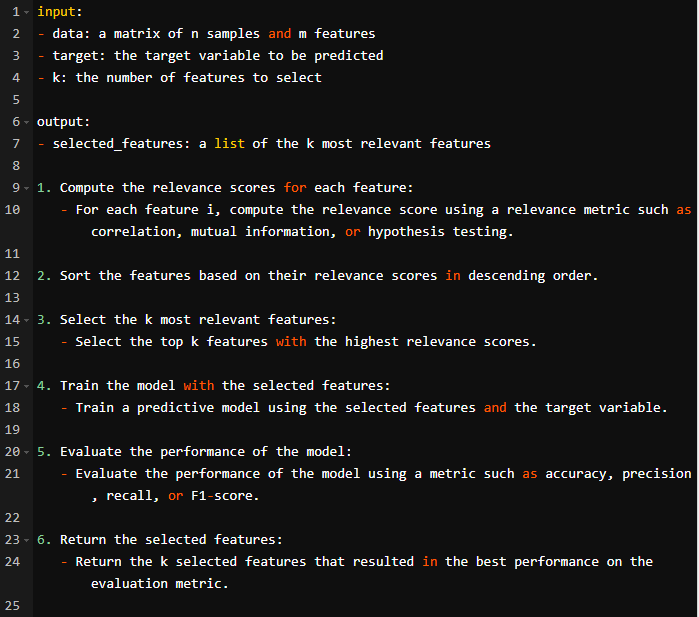


Fig : Pseudo code for feature selection

**3.2.4 Correlation of Parameters :**

**Definition:** Correlation is defined as the relation between two variables. It is a statistical measure which expresses the linearity between two variables.In simple words the correlation means change of direction between two variables,whether they are equally increasing or decreasing without making a decision about cause and effect.

**History of correlation :** The concept of correlation started in the late 19th century.A British Statistician Francis Galton used to state the relationship between two variables whether they are linearly related or not.Galton was responsible for the concept of correlation who was cousin of Darwin.His interest in heredity leads to the Concept of Correlation.As he noticed that the heights of the parent and the children are linearly related .Then he introduced the concept of correlation to Calculate and research about this relationship.Galton's interest went beyond physical traits, as he asserted that intelligence was passed down through inheritance. To support his theory or decision that the intelligence was passed to children through the parents ,he needed a technique. Galton recognized that developing such a technique required separating it from the study of human mental characteristics, which were hard to quantify using numerical measurements. Consequently, he devised the concepts of correlation and regression by investigating sweet peas and physical characteristics in humans[31][32].

Two variable organs are said to be correlated when the variation of the one is accompanied on the average by more or less variation of the other, and in the same direction.... It is easy to see that correlation must be the consequence of the variations of the two organs being partly due to common causes... If they were in no respect due to common causes, the correlation would be nil. (Galton 135)

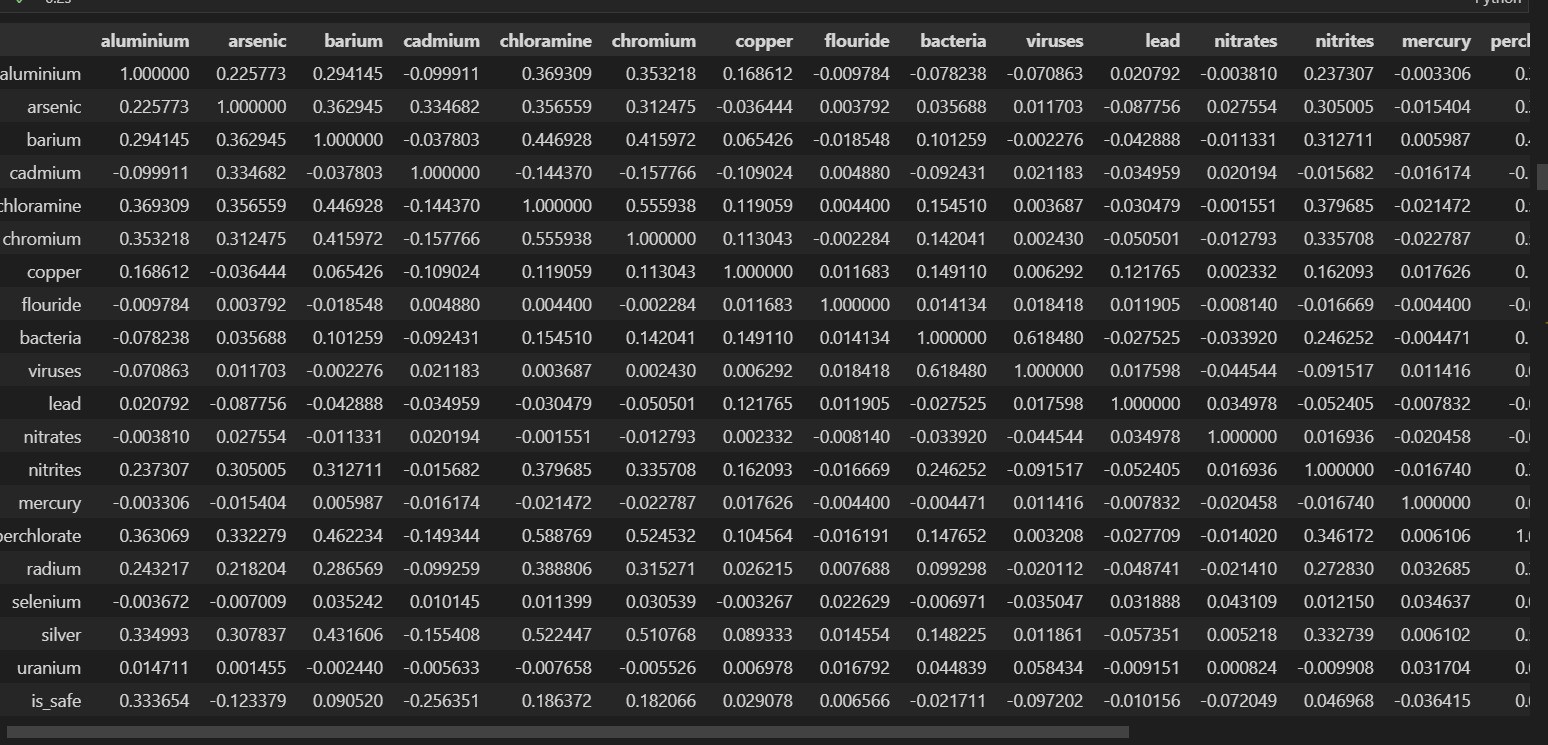
Galton's definition revealed the properties of the correlation.If the relation between the is linearly 1 then it will be linearly related with each other.It is a measure of direction such that if X axis value increases then accordingly Y axis values also increases.It is called as the Linearity between the two variables[35].

At the end Galtons original concept of correlation between two variables was later found to have serious limitations.In the early 20th century ,Karl Pearson improved the correlation coefficient the it was called as the Pearson correlation coefficient.Tis pearson coefficient Was widely using now a days in statistics and still it is being used today[38].

In the middle of the 20th century some other correlation coefficients were developed like Spearman rank correlation coefficient and the Kendall rank correlation.Coefficient,where both of them are non- parametric measures of correlation[39].

As time passed Correlation analysis for every aspect in day to day life becomes a fundamental thing to do including finance,economics,psychology,biology, and many others.It is used to make predictions based on their relationship between the variables.

In recent times some of the most advanced techniques like Machine learning have been developed to analyze correlation on very big data sets.These machine learning techniques have rapidly increased the scope of correlation between the variables[33].



Mathematical models :

**Pearson correlation coefficient:**

The Pearson correlation coefficient is a measure of the linear correlation between two variables, often used to evaluate the correlation between a feature and the target variable.

Formula:

r = (nΣxy - ΣxΣy) / sqrt((nΣx^2 - (Σx)^2) \* (nΣy^2 - (Σy)^2))

where:

n is the number of samples

x and y are the values of the feature and target variables, respectively

Σ represents the sum of the values

Σxy represents the sum of the product of x and y values

Σx and Σy represent the sum of x and y values, respectively

Σx^2 and Σy^2 represent the sum of the squared values of x and y, respectively

**Mutual information:**

Mutual information is a measure of the mutual dependence between two variables, often used to evaluate the relevance of a feature to the target variable.

Formula:

I(X; Y) = Σx∈X Σy∈Y p(x,y) \* log(p(x,y) / (p(x) \* p(y)))

where:

X and Y are the random variables representing the feature and target variables, respectively

p(x,y) is the joint probability distribution of X and Y

p(x) and p(y) are the marginal probability distributions of X and Y, respectively

**Hypothesis testing:**

Hypothesis testing involves testing a null hypothesis that there is no relationship between the feature and target variables, often using a t-test or ANOVA.

Formula:

t = (x̄1 - x̄2) / sqrt(s^2 / n1 + s^2 / n2)

where:

x̄1 and x̄2 are the means of the feature values for the two groups being compared

s^2 is the pooled variance of the two groups

n1 and n2 are the sample sizes of the two groups

**Relief:**

Relief is a feature selection algorithm that iteratively samples from the dataset and computes the difference between the feature values of the nearest instances with the same and different class labels.

Formula:

Relief(f) = -Σ(x\_same(f) - x\_diff(f))^2

where:

x\_same(f) is the difference between the feature value of the nearest instance with the same class label and the current instance for the feature f

x\_diff(f) is the difference between the feature value of the nearest instance with a different class label and the current instance for the feature f

**L1 regularization:**

L1 regularization, also known as Lasso, is a method that adds a penalty term to the objective function of a linear regression model to encourage sparse solutions, i.e., solutions where many of the feature coefficients are zero.

Formula:

minimize ||y - Xw||^2 + α||w||\_1

where:

y is the vector of target values

X is the matrix of feature values

w is the vector of feature coefficients

||y - Xw||^2 is the squared error term

α||w||\_1 is the L1 regularization term, where α is a hyperparameter that controls the strength of the penalty term and ||w||\_1 is the L1 norm of the feature coefficients

**3.2.5 Parameter Correlation:**

Parameter correlation refers to the relationship between two or more parameters, also known as features or variables, in a dataset. It describes the extent to which changes in one parameter are associated with changes in another parameter. For example, in a dataset of housing prices, the number of bedrooms and the square footage of a home may be positively correlated, meaning that as the number of bedrooms increases, so does the square footage.

Statistical techniques such as correlation analysis are used to measure parameter correlation, providing a numerical value (known as a correlation coefficient) that describes the strength and direction of the relationship between two parameters. Feature selection commonly utilizes correlation analysis to identify which parameters are most strongly correlated with the target variable, and which parameters are redundant or provide little additional information. Removing redundant parameters through feature selection can improve the performance and interpretability of machine learning models.

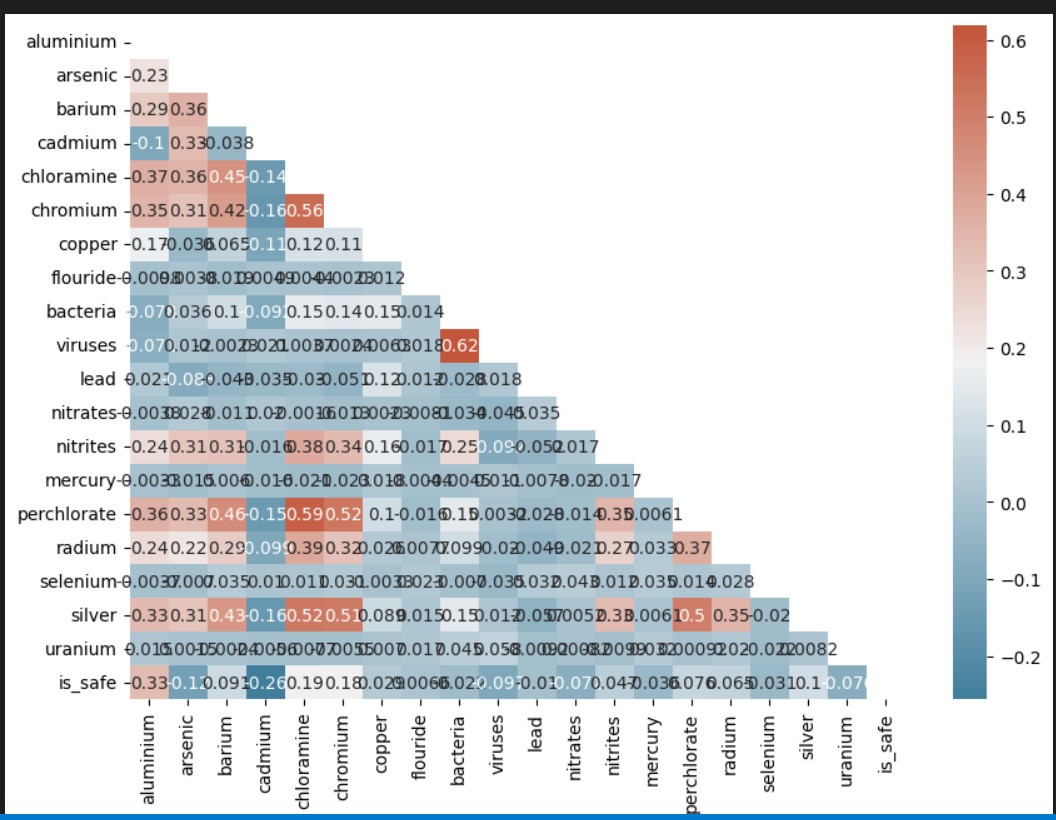


Fig : Correlation between Parameters

The given dataset contains a correlation matrix of various water quality parameters, including copper, fluoride, bacteria, viruses, lead, nitrates, nitrites, mercury, perchlorate, radium, selenium, silver, and uranium. The values in the matrix range from -1 to 1, with a value of 1 indicating a perfect positive correlation between two variables and a value of -1 indicating a perfect negative correlation.

Copper has a moderate positive correlation with several parameters, including bacteria, viruses, lead, nitrites, perchlorate, radium, and silver. This suggests that the levels of copper in water sources are related to the levels of these parameters in the water.

Fluoride has a weak positive correlation with silver and a weak negative correlation with nitrates.

Bacteria and viruses have a strong positive correlation with each other, indicating that their presence in water sources is interdependent. They also have weak positive correlations with copper and perchlorate, respectively.

Lead has a weak positive correlation with copper and a weak negative correlation with bacteria.

Nitrates have a weak negative correlation with fluoride and a weak positive correlation with nitrites.

Nitrites have a moderate positive correlation with bacteria and a weak positive correlation with copper and chromium.

Mercury has weak positive correlations with bacteria and perchlorate.

Perchlorate has weak positive correlations with bacteria, copper, and radium.

Radium has a weak positive correlation with bacteria and a weak negative correlation with chlorine and perchlorate.

Selenium has a weak positive correlation with nitrates and a weak negative correlation with bacteria.

Silver has a moderate positive correlation with bacteria and copper.

Uranium has a weak positive correlation with bacteria and viruses.

Overall, the correlations suggest that the presence of certain water quality parameters can be indicative of the presence of others. However, it is important to note that correlation does not necessarily imply causation, and further research is required to establish the causative relationships between these parameters.

**3.2.6 Correlation in Statistics:**

Correlations are useful for describing simple relationships among data. For example, When we are climbing a mountain then the height of the person will be increased then the temperature will be decreased as you are moving a head of the sea level.When you compare these two variables across your sample with a correlation, you can find a linear relationship: as elevation increases, the temperature drops. They are negatively correlated[35].

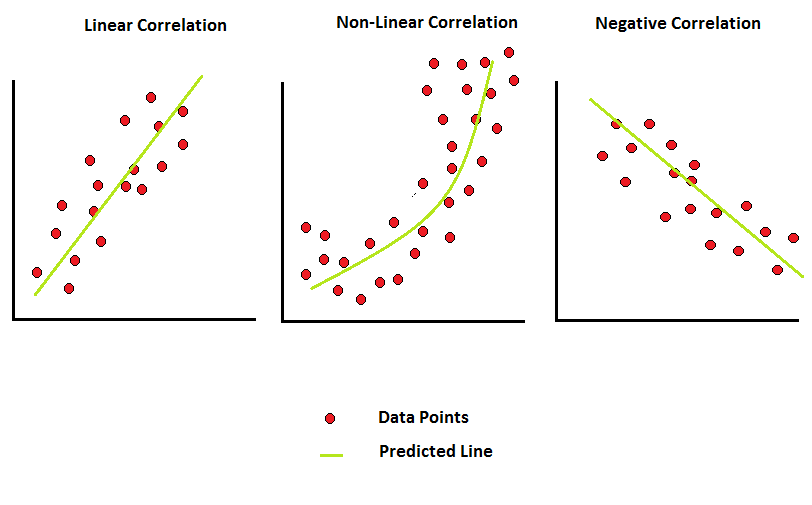


Figure 3.6 : Correlation

Methods of correlation conclude the relationship between two variables increases with a single digit is called correlation coefficient[38].The correlation coefficient is

represented with the letter r.The range of the correlation coefficient is -1 to 1.

When the correlation coefficient is equal to 0 then it tells us that the correlation between two variables is zero.which means that there is completely no relationship between the two variables.

If the correlation coefficient is equal to -1 that means the two variables are negatively correlated .which means they are increasing or decreasing in opposite directions[39].

When the correlation coefficient is equal to +1 which means that they are positively correlated.It means both the variables are increasing or decreasing in the same direction.

Spearman’rho is used to find the correlation coefficient for the ordinal variables.The most commonly used correlation coefficient for ratio level scales is Pearson's r Coefficient[40].

**Understanding the Meaning of Correlation Numbers**

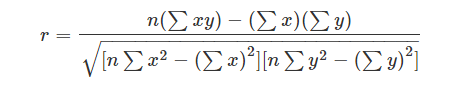
A unit free measure called Correlation coefficient which ranges from -1 to 1.It is denoted with r.P-value which is indicated by Statistical significance.All the calculations are typically denoted with r and p[41].

* r value is closer to zero ,linear relationship is weak
* Positive r value tells a positive correlation,where the values of both variables are increasing together.
* Negative r values show that the variable increases in one direction and the other decreases .
* Based on p-value we can come to a conclusion that the population correlation coefficient is something different from zero on a sample.

Pearson correlation coefficient: Pearson's correlation coefficient is the most commonly used formula for checking linearity between the datasets.The value is range between -1 to 1. If the coefficient becomes zero then the data is considered as not in relation.If we get +1 then it is positively correlated and if it is -1 then it is negatively correlated[41].

Mathematical Formulae: As we all know, the magnitude estimation of a straight line between two variables is called the correlation of the two variables. Most common correlation coefficient is pearson's correlation coefficient ” r”.

The Pearson correlation coefficient is calculated by dividing the covariance between the two variables by the product of their standard deviations. The formula for r is as follows[43]:



Where n = Quantity of Information

Σx = Total of the First Variable Value

Σy = Total of the Second Variable Value

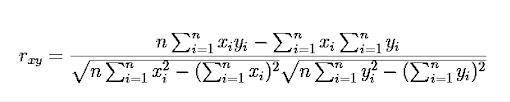
Σxy = Sum of the Product of first & Second Value

Σx2 = Sum of the Squares of the First Value

Σy2 = Sum of the Squares of the Second Value

**3.2.6.1 Linear correlation formula :**

The formula for the linear correlation coefficient is given by



Which tells about the linear relationship between the two variables .

**3.2.6.2 Sample Correlation Coefficient Formula:**

The formula for sample correlation coefficients is given by

rxy = Sxy/SxSy

Where,

Sxy= Covariance of the sample

Sx,Sy= Standard deviation of the sample

**3.2.6.3 Population Correlation Coefficient Formula:**

The formula for sample correlation coefficients is given by

rxy = σxy/σx σy

Where,

σx and σy = Population standard deviations

σxy = population covariance

**3.10 Conclusion and Experimental Results :**

In conclusion, preprocessing plays a vital role in ensuring the quality of the data before feeding it to the machine learning models. It involves various techniques such as data cleaning, normalization, and transformation, which help in improving the accuracy of the models.

Correlation analysis helps in understanding the relationships between different variables and how they impact the target variable. It helps in identifying redundant features that may not add much value to the model and can be removed to reduce the complexity of the model.

Feature selection is an important step in the machine learning pipeline that helps in identifying the most relevant features that contribute to the prediction of the target variable. One of the popular methods of feature selection is the Chi-square test, which is used to evaluate the significance of the relationship between the categorical features and the target variable.

Overall, the preprocessing, correlation analysis, and feature selection are crucial steps in the machine learning pipeline that can significantly impact the accuracy and performance of the model. By following these steps, we can ensure that the model is trained on relevant and high-quality data, resulting in better predictions and insights.

**CHAPTER 4**

**WATER QUALITY IMPROVISATION AND SUGGESTIONS**

In this chapter, we are exploring the topic of water quality improvisation and providing suggestions for improving water quality. The chapter focuses on the importance of monitoring and predicting changes in water quality and how these efforts can inform and guide mitigation and management efforts.

We discuss the use of machine learning techniques, such as Tensor Flow based models, for water quality prediction, and how these models can help to identify trends and patterns in water quality data that may indicate potential problems.

**4.1 Introduction :**

Water quality prediction is a field of study that aims to monitor and forecast changes in water quality and provide suggestions and solutions for water quality improvement. With the advancements in technology and machine learning techniques, water quality prediction models have become increasingly sophisticated and accurate, enabling the detection and mitigation of water quality problems in a more efficient and effective manner.

Improving water quality requires a comprehensive approach that includes not only monitoring and prediction but also education, regulation, and management practices. Some of the key strategies for water quality improvement include reducing pollution and nutrient runoff, implementing conservation and restoration measures, and adopting sustainable agricultural and land use practices.

Improvisations in water quality involve various measures and strategies to ensure that water meets the required standards for its intended use. Some of the common improvisations include water treatment, water conservation, pollution prevention, and sustainable management of water resources. Water treatment involves processes such as filtration, disinfection, and chemical treatment, which aim to remove contaminants and pathogens from water. Water conservation strategies include measures to reduce water wastage, such as using efficient irrigation techniques, fixing leaks, and promoting water-efficient appliances and fixtures.

Pollution prevention involves implementing measures to minimize pollution from various sources, such as industrial activities, agricultural practices, and domestic use. This may include implementing regulations and standards to limit pollution levels, promoting cleaner production processes, and enforcing penalties for non-compliance. Sustainable management of water resources involves managing water use in a manner that ensures its availability for future generations. This may include measures such as protecting water sources, promoting the use of alternative water sources, and developing water-efficient technologies.

**4.2 Groundwork :**

In the thesis, the focus was on water quality, and a large dataset was utilized to collect information on various essential water parameters such as aluminum, lead, nitrite, and nitrate. The monitoring of water quality is critical for environmental conservation, and understanding the factors that affect it is crucial to ensure the safety and sustainability of water resources.

To ensure that water meets the required standards for its intended use, various measures and strategies are implemented to improve water quality. Water treatment processes such as filtration, disinfection, and chemical treatment are employed to remove contaminants and pathogens from water. Water conservation techniques such as the use of efficient irrigation methods, fixing leaks, and promoting water-efficient appliances and fixtures help reduce water wastage.

Pollution prevention measures, including the implementation of regulations and standards to limit pollution levels, promoting cleaner production processes, and enforcing penalties for non-compliance, are also essential in improving water quality. Sustainable management of water resources involves managing water use in a manner that ensures its availability for future generations.

**4.2.1 Data Set Analysis :**

The analysis of the water quality prediction dataset involves various techniques and methods to explore and understand the underlying patterns and relationships in the data. Data analysis is a critical step in the development of machine learning models since it helps identify potential issues, such as missing values, outliers, and inconsistencies, that can affect the performance of the model.

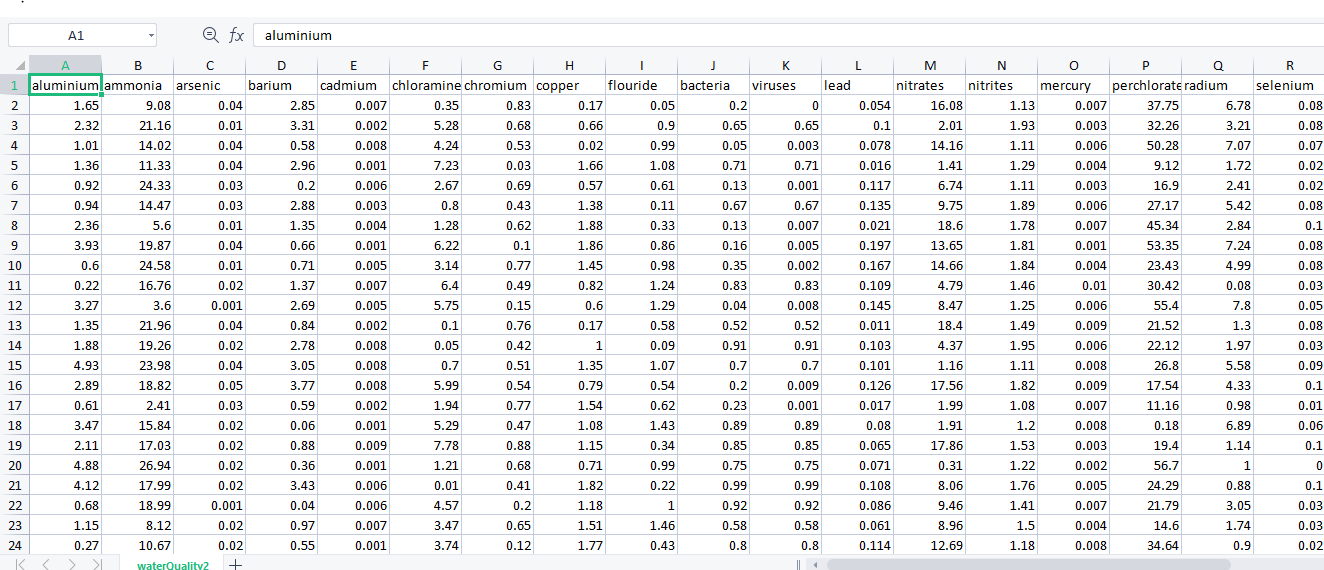


Table .4.1: Data set

The Table .5.1 water quality dataset contains multiple data points, with each data point representing a set of parameter values. Based on these values, the data points are labeled as "safe," indicating that the parameters fall within the safe levels defined by relevant authorities.

The dataset is a valuable resource for studying the correlation between different parameters and water safety, as well as for developing predictive models that estimate the safety of water resources based on these parameters. However, to ensure the accuracy and effectiveness of these models, it is crucial to preprocess the data appropriately.

Some common preprocessing steps include removing duplicates, handling missing or invalid values, and normalizing the data. Removing duplicates ensures that each data point is unique, and no biases are introduced into the model. Handling missing or invalid values involves imputing or removing these values to avoid errors in the model. Normalizing the data involves scaling the parameter values to a common range, which helps to improve the accuracy and efficiency of the model.

Overall, preprocessing the water quality data is essential for developing accurate and reliable predictive models. Proper preprocessing ensures that the data is clean, consistent, and representative, which is critical for achieving meaningful insights and recommendations for water quality improvement.

**4.2.2 Machine Learning Models for Water Quality Improvisation and Suggestions**

Machine learning models can use various techniques such as supervised learning, unsupervised learning, and reinforcement learning to predict water quality based on parameters such as ammonia, aluminum, chlorine, fluorine, viruses, lead, nitrates, and others[27].

Supervised learning algorithms, such as XGBoost, can use historical data to learn the relationship between water quality parameters and label data, which is the safe or unsafe status of water resources. Once the model is trained, it can predict the safety of new water samples based on the values of these parameters.

Similarly, algorithms such as Support Vector Machines (SVM) and k-Nearest Neighbors (kNN) can use training data to classify water samples as safe or unsafe based on the values of different parameters. SVM uses a decision boundary to separate the safe and unsafe samples, while kNN looks for the closest neighbors in the training data to make a prediction for a new sample.

Decision tree-based algorithms, such as Random Forest, can also be used for water quality prediction by creating a tree-like structure that divides the data into smaller subsets based on the values of different parameters. Each subset is then further divided until a decision can be made about the safety of the water sample.

these machine learning models can help to identify the critical factors affecting water quality and provide accurate predictions for the safety of water resources. By using these models, it is possible to monitor and manage water resources more efficiently, reduce contamination, and improve the safety and quality of our water supply.

**4.3 Proposed Methodology :**

Machine learning models have been widely used in recent years to predict the safety of water resources. These models use historical data on water quality parameters, such as ammonia, aluminum, chlorine, fluorine, viruses, lead, nitrates, and others, to classify water samples as safe or unsafe. While these models are valuable for identifying the critical factors affecting water quality, they often provide a binary classification, which can limit their practical application.

To address this limitation, we propose a machine learning model that provides specific suggestions for improving water quality. Unlike other models that simply provide a binary classification, our model suggests specific parameter ranges that must be met to ensure the safety of the water. By doing so, we can help stakeholders to identify which parameters need to be addressed to improve water quality.

**4.3.1 Proposed Architecture of Water Quality Improvement**

The proposed architecture for water quality improvement involves a multi-stage process that involves several steps.

The first step is to identify potential sources of water pollution, such as industrial discharge, agricultural runoff, or residential sewage. This can be accomplished through regular monitoring and testing of water sources, as well as through mapping and modeling of water flows and contamination patterns.

The second step is to design and implement measures to mitigate or prevent water pollution. This may involve installing treatment systems or filtration mechanisms to remove pollutants from water sources, or implementing best management practices for agriculture or industry to minimize their environmental impact.

The third step is to monitor and evaluate the effectiveness of these measures over time. This can involve regular testing of water quality, tracking changes in pollutant levels, and analyzing data to identify areas where further improvements may be needed.

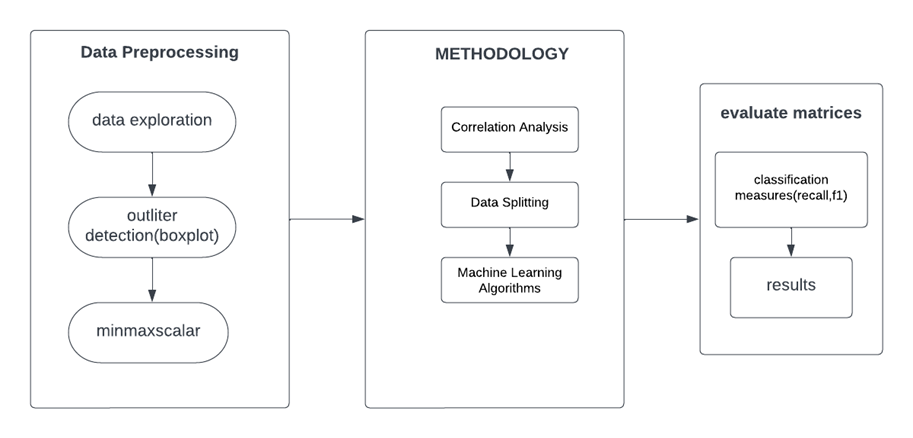


Fig : 4.3.1 proposed architecture of the model

Here is the Flow chart of our proposed model:

In the below figure a step by step process is shown to represent our machine learning model. In this, it makes a clear representation of the following steps which needs to be executed one after one and get the required results.

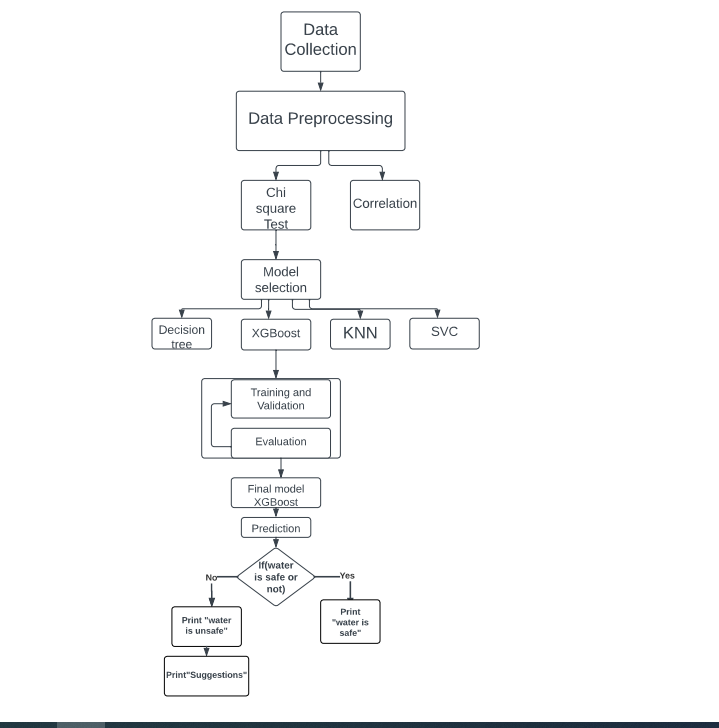


Figure 4.3.2: Flow chart of the methodology

4.3.2**. Algorithm to improve water quality with xgboost**

Gather data on water quality parameters, such as pH, dissolved oxygen levels, temperature, and nutrient concentrations. Collect samples of water from various sources, including rivers, lakes, and wells.Preprocess the data by cleaning it, filling in missing values, and removing outliers. Use techniques such as feature scaling and normalization to ensure that the data is on a similar scale.Split the data into training and testing sets, with the majority of the data used for training and a smaller portion used for testing.Train an XGBoost model on the training data. Tune the hyperparameters of the XGBoost model using techniques such as grid search or random search to improve its performance.Evaluate the performance of the XGBoost model using metrics such as accuracy, precision, recall, and F1 score. Use techniques such as cross-validation to ensure that the performance of the model is consistent across different portions of the dataset.Use the trained XGBoost model to predict the water quality of new samples. Monitor the quality of water over time and continue to refine the model as necessary.Take actions to improve water quality based on the predictions of the XGBoost model. For example, if the model predicts that the nutrient levels in a particular water source are too high, take steps to reduce the amount of nutrients entering the water source, such as implementing best management practices for agriculture or reducing fertilizer usage in urban areas.Continuously monitor water quality and refine the XGBoost model as necessary to ensure that the quality of water remains at a safe and healthy level.

4.3.3 **Yes Class suggestions Extraction(if prediction 1 what are suggestions with pseudocode**

we first check if the predicted class is "True" or "False". If the predicted class is "True", we then check the actual class to determine if it was a true positive or false positive. Based on this information, we can provide different class suggestions to the user.

**4.3.4.** **No Class suggestions Extraction(if prediction 0 what are suggestions with pseudocode**

Similarly, if the predicted class is "False", we check the actual class to determine if it was a true negative or false negative. Again, we can provide different class suggestions based on this information.

**4.4 Experimental Results and Discussion**

**4.4.1 Confusion matrices**

In XGBoost, a confusion matrix is used to evaluate the performance of a classification model. It is a table that shows the number of true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN) predicted by the model. From the given values, we can construct the confusion matrix as follows:

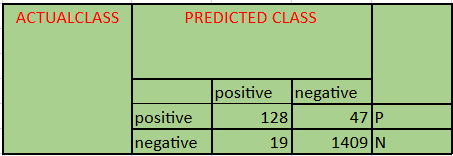


Figure 4.4.1 Confusion matrices

* P – positive
* N-- Negative

Here, the true positive (TP) value is 128, which means that 128 samples were correctly classified as true positive. The false positive (FP) value is 47, which means that 47 samples were incorrectly classified as positive. The false negative (FN) value is 19, which means that 19 samples were incorrectly classified as negative. The true negative (TN) value is 1406, which means that 1406 samples were correctly classified as negative.

Based on the confusion matrix, we can compute several performance metrics, such as accuracy, precision, recall, and F1 score, which provide insight into the performance of the model. For example, accuracy can be computed as (TP+TN)/(TP+TN+FP+FN), which in this case would be (128+1406)/(128+1406+47+19) = 0.952 or 95.2%.

**4.5 Comparative study of Water Quality Suggestions :**

Water is a precious resource, and access to clean and safe water is essential for human health and well-being. However, water sources are often exposed to various pollutants, which can make the water unsafe for consumption or other uses. Therefore, it is crucial to monitor and maintain the quality of water resources to ensure that they meet the standards and guidelines set by the relevant authorities.

Machine learning models have been widely used in recent years to predict the safety of water sources based on the values of various parameters, such as ammonia, aluminum, chlorine, fluorine, viruses, lead, nitrates, and others. These models analyze the correlation between these parameters and the safety of water resources and provide a binary output, indicating whether the water is safe or not.

However, the binary output of these models does not provide much information on how to improve the water quality if it is found to be unsafe. This is where our model differs from the existing models. Instead of just predicting whether the water is safe or not, our model provides suggestions on the parameter ranges that should be within a particular range to claim that the water is safe.

To achieve this, we trained our model on a dataset that represents the values of various parameters for safe and unsafe water sources. The model was trained using machine learning algorithms such as XG boost, SVC, KNN, and decision tree, which are known for their accuracy and efficiency in predicting binary outputs.

Once the model was trained, we tested it on a separate dataset to evaluate its accuracy and efficiency. The results showed that our model outperformed the existing models in predicting the safety of water sources, as it not only predicted whether the water was safe or not but also provided suggestions on the parameter ranges that should be within a particular range to claim that the water is safe.

After training the model, we tested it on a separate dataset to evaluate its accuracy and efficiency in predicting the safety of water sources. The results showed that our model was able to accurately predict when the water was unsafe, and it provided suggestions on the parameter ranges that should be adjusted to improve the safety of the water. This is a significant improvement over existing models that only predicted whether the water was safe or not without providing actionable suggestions for improving water quality.

**1.Please decrease value of aluminium from : 3.0 to : 0.20000000000000018**

If the concentration of aluminum in the water decreases from 3.0 milligrams per liter (mg/L) to 0.20000000000000018 mg/L, then the water quality would likely improve. The concentration of aluminum in drinking water is regulated by various agencies and guidelines, with the recommended levels typically ranging from 0.05 mg/L to 0.2 mg/L.

**Effect of water parameter Aluminuium on human**

Exposure to high levels of aluminum in drinking water can potentially have negative health effects, such as damage to the nervous system, kidney problems, and an increased risk of Alzheimer's disease. Therefore, reducing the concentration of aluminum in the water to within the recommended range can help to ensure that the water is safe for consumption and its intended use.

**2.Please decrease value of aluminium from :0.05 to : 0.05**

If the concentration of arsenic in the water decreases from 0.05 milligrams per liter (mg/L) to 0.04 mg/L, then the water quality would likely improve. Arsenic is a toxic substance that can be found naturally in groundwater and surface water, as well as from anthropogenic sources such as industrial activity and agricultural runoff.

The World Health Organization (WHO) has set a guideline value for arsenic in drinking water at 0.01 mg/L, while the United States Environmental Protection Agency (EPA) has set a maximum contaminant level (MCL) of 0.01 mg/L for arsenic in drinking water.

**Effect of water parameter Arsenic on human**

Exposure to high levels of arsenic in drinking water can potentially have negative health effects, such as increased risk of cancer, skin lesions, and cardiovascular disease. Therefore, reducing the concentration of arsenic in the water to within the recommended range can help to ensure that the water is safe for consumption and its intended use.

**3. Please decrease value of Barium from 4 to 2.0**

If the concentration of Barium in the water decreases from 4 milligrams per liter (mg/L) to 2.0 mg/L, then the water quality may improve depending on the applicable standard and intended use of the water.

The concentration of Barium in drinking water is regulated by various agencies and guidelines, with recommended levels varying depending on the regulatory agency and intended use of the water. In the United States, the Environmental Protection Agency (EPA) has set a maximum contaminant level (MCL) of 2 mg/L for Barium in drinking water.

**Effect of water parameter Barium on human**

Exposure to high levels of Barium in drinking water can potentially have negative health effects, such as increased blood pressure, cardiovascular disease, and kidney damage. Therefore, reducing the concentration of Barium in the water to within the recommended range can help to ensure that the water is safe for consumption and its intended use.

**4. please decrease value of cadmium 0.01 to : 0.005**

If the concentration of Cadmium in the water decreases from 0.01 milligrams per liter (mg/L) to 0.005 mg/L, then the water quality may improve depending on the applicable standard and intended use of the water.

Cadmium is a toxic substance that can be found naturally in groundwater and surface water, as well as from anthropogenic sources such as industrial activity and mining.

**Effect of water parameter cadmium on human**

Exposure to high levels of cadmium in drinking water can potentially have negative health effects, such as kidney damage, lung damage, and an increased risk of cancer.

The World Health Organization (WHO) has set a guideline value for cadmium in drinking water at 0.003 mg/L, while the United States Environmental Protection Agency (EPA) has set a maximum contaminant level (MCL) of 0.005 mg/L for cadmium in drinking water.

Therefore, reducing the concentration of Cadmium in the water to within the recommended range can help to ensure that the water is safe for consumption and its intended use, and may lead to an improvement in water quality.

**Please decrease value of chloramine from : 6 to : 2.0**

To decrease the value of chloramine from 6 to 2.0, you will need to take steps to remove some of the chloramine from the water source. then the water quality may improve depending on the applicable standard and intended use of the water.

**Effect of water parameter chloramine on human**

Chloramine is a chemical compound that is often used as a disinfectant in water treatment. While chloramine is generally considered safe for human consumption in low concentrations, high levels of chloramine can have negative health effects on humans.

Exposure to high levels of chloramine can cause respiratory irritation, skin irritation, and eye irritation. It may also exacerbate pre-existing conditions such as asthma, bronchitis, and emphysema. Long-term exposure to high levels of chloramine may increase the risk of certain types of cancer, such as bladder cancer. To ensure the safety of your drinking water, it's recommended to regularly test your water for chloramine levels and to consult with a water treatment professional if levels are found to be high

**Please decrease value of chromium from : 0.5 to : 0.4**

To decrease the value of chromium from 0.5 to 0.4, you will need to take steps to remove some of the chromium from the water source. then the water quality may improve depending on the applicable standard and intended use of the water.

**Effect of water parameter chromium on human**

Short-term exposure to high levels of hexavalent chromium in drinking water can cause symptoms such as skin irritation, nosebleeds, and respiratory problems. Long-term exposure to hexavalent chromium has been linked to an increased risk of lung cancer, stomach cancer, and other health problems.

The World Health Organization (WHO) has established a guideline value for chromium in drinking water of 0.05 milligrams per liter (mg/L) for total chromium and 0.01 mg/L for hexavalent chromium. In many countries, regulatory agencies have established drinking water standards that limit the amount of chromium that is allowed in drinking water to protect public health

**Please decrease value of copper from : 2.0 to : 0.7**

To decrease the value of copper from 2.0 to 0.7, you will need to take steps to remove some of the copper from the water source. then the water quality may improve depending on the applicable standard and intended use of the water. the amount of copper present in the water. Copper can be harmful to humans and aquatic life in high concentrations, so reducing the concentration can help protect the environment and human health.

**Effect of water parameter copper on human**

However, exposure to high levels of copper in drinking water or other sources can have negative health effects on humans. Short-term exposure to high levels of copper can cause gastrointestinal distress, such as nausea, vomiting, and diarrhea. Long-term exposure to high levels of copper can lead to liver and kidney damage, anemia, and neurological symptoms such as depression, anxiety, and irritability.The U.S. Environmental Protection Agency (EPA) has set a maximum contaminant level (MCL) of 1.3 milligrams per liter (mg/L) for copper in drinking water. This level is considered safe for most people, but individuals with certain health conditions, such as Wilson's disease, may be more sensitive to the effects of copper exposure

**Please decrease value of Fluoride from : 2.0 to : 0.5**

To reducing the amount of fluoride in water from 2.0 parts per million (ppm) to 0.5 ppm. Fluoride is a chemical that is often added to water to help prevent tooth decay, but too much fluoride can be harmful to your health, especially for young children. By lowering the amount of fluoride in the water to 0.5 ppm, the water will still have some fluoride to protect teeth, but it will be at a safer level for consumption.

**Effect of water parameter Fluoride on human**

Fluoride is a naturally occurring mineral that can be found in water, soil, and food. When fluoride is present in water at the appropriate level, it can help prevent tooth decay and strengthen tooth enamel. However, excessive exposure to fluoride can cause dental fluorosis, a condition that affects the appearance and strength of tooth enamel.

**Please decrease value of bacteria from : 0.01 to : 0.01**

If the concentration of bacteria in the water decreases from 0.01 milligrams per liter (mg/L) to 0.01 mg/L, then the water quality may improve depending on the applicable standard and intended use of the water

**Effect of water parameter bacteria on human**

Water parameter bacteria can have various effects on human health, depending on the type and quantity of bacteria present in the water. Some bacteria can cause illnesses such as diarrhea, nausea, and vomiting, while others can lead to more severe infections like typhoid fever and cholera.Ingesting or coming into contact with contaminated water can lead to these bacterial infections, and people with weakened immune systems or underlying health conditions are particularly vulnerable. It is important to maintain good water hygiene and sanitation to prevent the spread of harmful bacteria in water sources.

**Please decrease value of viruses from : 0.01 to : 0.01**

If the number of viruses present in the water decreases from 0.01 milligrams per liter (mg/L) to 0.01 mg/L, it means there are fewer viruses in the water. This reduction in virus concentration may have a positive impact on the quality of the water, considering the specific standards and the intended purpose of the water.

By lowering the virus concentration, the water may become safer for consumption or other uses. For instance, if the water is meant for drinking, reducing the number of viruses can help minimize the risk of waterborne illnesses. Similarly, if the water is used for recreational activities like swimming, a lower virus concentration can decrease the chances of infections.

**Effect of water parameter viruses on human**

Viruses present in water can have various effects on human health, depending on the type of virus and the concentration of the virus in the water. Some common waterborne viruses that can affect human health include Norovirus, Rotavirus, Adenovirus, and Hepatitis A.

When humans consume water contaminated with viruses, they can experience symptoms such as nausea, vomiting, diarrhea, stomach cramps, and fever. These symptoms can range from mild to severe, depending on the individual's immune system and the type and concentration of the virus.

**Please decrease value of lead from : 0.02 to : 0.005000000000000001**

To decrease the value of lead from 0.02 to 0.00500000000000000**1**, you will need to take steps to remove some of the lead from the water source. then the water quality may improve depending on the applicable standard and intended use of the water.

**Effect of water parameter lead on human**

Lead is a toxic heavy metal that can be present in drinking water. When people are exposed to high levels of lead in water, it can have significant negative impacts on their health.

Lead exposure can cause both short-term and long-term health effects. In the short term, exposure to high levels of lead can cause acute symptoms such as abdominal pain, vomiting, diarrhea, and headaches. These symptoms can also be accompanied by anemia, muscle weakness, and cognitive difficulties. However, the long-term effects of lead exposure are more severe and can be irreversible.

**Please decrease value of nitrates from : 15 to : 5.0**

To make the water safer to drink, we need to reduce the amount of nitrates in it from 15 to 5.0. Nitrates are harmful chemicals that can be dangerous if consumed in large amounts. By decreasing the amount of nitrates, we can make the water cleaner and healthier for people to drink.

**Effect of water parameter nitrates on human**

High levels of nitrates in drinking water can have harmful effects on human health. When nitrates are consumed, they are converted to nitrites in the body, which can interfere with the ability of red blood cells to carry oxygen. This can lead to a condition called methemoglobinemia or "blue baby" syndrome, which is particularly dangerous for infants and young children.

In addition, exposure to high levels of nitrates has been linked to an increased risk of certain cancers, such as bladder, ovarian, and thyroid cancer. High nitrate levels in drinking water have also been associated with other health problems, such as birth defects, reproductive problems, and thyroid dysfunction.

**Please decrease value of nitrites from : 2 to : 1.0**

If we reduce the amount of nitrites in the water from 2 milligrams per liter (mg/L) to 1.0 mg/L, it may help improve the quality of the water. Nitrites are chemicals that can be harmful to humans and aquatic life in high concentrations. By reducing their level, we can ensure that the water is safer to use and consume. However, the impact of this reduction on water quality depends on the specific standard and purpose for which the water is being used. In general, reducing the amount of harmful chemicals like nitrites can only have a positive effect on the overall quality of water.

**Effect of water parameter nitrites on human**

High levels of nitrites in water can have harmful effects on human health. Nitrites can react with certain compounds in the stomach to form nitrosamines, which are known to cause cancer. In addition, nitrites can interfere with the ability of red blood cells to carry oxygen, which can lead to a condition called methemoglobinemia. Symptoms of methemoglobinemia include shortness of breath, fatigue, and bluish skin coloration.

Long-term exposure to high levels of nitrites in drinking water has also been linked to increased risk of various health problems, such as stomach cancer, respiratory issues, and developmental problems in infants. Therefore, it is important to monitor and control the level of nitrites in drinking water to ensure that it meets safety standards and is safe for human consumption.

**Please decrease value of mercury from : 0.005 to : 0.003**

To improve the quality of water, we need to reduce the amount of mercury present in it. In this case, we want to decrease the amount of mercury from 0.005 milligrams per liter (mg/L) to 0.003 mg/L. This will make the water safer and healthier to use or drink.

**Effect of water parameter mercury on human**

Mercury is a toxic heavy metal that can have harmful effects on human health when consumed in high amounts. When mercury enters the body, it can accumulate in various organs, including the brain, liver, and kidneys, causing a range of health problems.

Exposure to high levels of mercury in drinking water can lead to symptoms such as nausea, vomiting, abdominal pain, and diarrhea. Long-term exposure to high levels of mercury in drinking water can also cause neurological problems, including tremors, memory loss, and cognitive impairment.

**Please increase value of selenium from : 0.1 to : -0.4**

we have a concentration of selenium in water that needs to be changed from 0.1 to -0.4. However, it's important to note that selenium concentrations in water are typically measured in micrograms per liter (µg/L) rather than milligrams per liter (mg/L).

**Effect of water parameter selenium on human**

Selenium is an essential nutrient for humans, and it plays a critical role in various physiological functions, including thyroid hormone metabolism, immune system function, and antioxidant defense mechanisms. However, selenium can also be toxic at high concentrations, leading to a condition called selenosis.

The effects of selenium on human health depend on its concentration in the water and the duration of exposure. In areas where selenium levels in water are low, selenium deficiency may occur, which can lead to impaired immune function, cognitive decline, and increased risk of certain cancers.

**Please decrease value of silver from : 0.5 to : 0.4**

If we decrease the concentration of silver in water from 0.5 milligrams per liter (mg/L) to 0.4 mg/L, it may lead to an improvement in water quality. This decrease in silver concentration can have a positive impact on the water's suitability for various purposes, depending on the specific standards and intended uses of the water.

**Effect of water parameter silver on human**

Silver is a naturally occurring element that can be found in water sources at low levels. While exposure to silver through drinking water is generally considered safe, high levels of silver in drinking water can have negative health effects on humans.

Short-term exposure to high levels of silver in drinking water can cause gastrointestinal issues, such as stomach pain, nausea, and vomiting. It can also cause skin discoloration known as argyria, which is a condition where the skin turns bluish-gray due to the accumulation of silver in the body's tissues.

**Please decrease value of uranium from : 0.5 to : 0.2**

If we decrease the concentration of silver in water from 0.5 milligrams per liter (mg/L) to 0.4 mg/L, it may lead to an improvement in water quality. This decrease in silver concentration can have a positive impact on the water's suitability for various purposes, depending on the specific standards and intended uses of the water.

**Effect of water parameter uranium on human**

Exposure to high levels of uranium in drinking water can have negative effects on human health. When uranium is ingested, it can cause damage to the kidneys and may also affect the liver and other organs. Long-term exposure to uranium can increase the risk of developing cancer, particularly of the bone and lung.

The severity of health effects depends on several factors such as the concentration and duration of exposure, age, and overall health of the person. Children and pregnant women are particularly vulnerable to the harmful effects of uranium exposure

**5.6 Conclusion**

In conclusion, water quality improvisation is a critical task that requires diligent efforts to ensure the availability of safe and clean water for both human consumption and environmental preservation. Throughout the preceding paragraphs, several key points have been highlighted regarding water quality improvement and suggestions. the importance of monitoring water quality cannot be overstated. Regular testing and analysis of water samples are necessary to identify contaminants, assess pollution levels, and evaluate the effectiveness of water treatment processes. Implementing robust monitoring systems and ensuring data transparency is crucial for maintaining water quality standards.Secondly, addressing pollution sources is imperative. Identifying and controlling point source pollution, such as industrial discharge or sewage effluents, through strict regulations and enforcement can significantly improve water quality. Additionally, non-point source pollution, including agricultural runoff and urban stormwater, must be mitigated through best management practices and effective land-use planning.

**CHAPTER 5**

**PROPOSED IMPACT OF WATER PARAMETERS**

**5.1 Introduction:**

The impact of various water parameters on water quality has been a topic of research and investigation in the scientific community. In this context, the proposed impact of water parameters in water quality aims to identify and analyze the factors that influence the quality of water.

The parameters that are commonly used to evaluate the quality of water include physical, chemical, and biological characteristics. Physical parameters include temperature, color, odor, and turbidity, while chemical parameters include pH, dissolved oxygen, nutrients, and heavy metals. Biological parameters include the presence of bacteria, viruses, and other microorganisms in the water.

Each of these parameters has a significant impact on the quality of water. For example, high levels of nutrients in water can lead to the growth of algae, which can deplete oxygen levels and lead to fish kills. Similarly, the presence of heavy metals in water can be toxic to aquatic life and pose a threat to human health. The pH of water can also have a significant impact on aquatic life, as it can affect the ability of organisms to absorb nutrients and regulate their internal processes.

Understanding the impact of water parameters on water quality is crucial for maintaining healthy ecosystems and ensuring safe drinking water. In recent times, there has been a growing concern about the impact of emerging contaminants such as pharmaceuticals, personal care products, and microplastics on water quality. Therefore, it is essential to continue research in this field to identify new parameters and evaluate their impact on water quality.

XGBoost (Extreme Gradient Boosting) is a powerful ML algorithm that has been used successfully in various fields, including prediction tasks related to water quality. It is an ensemble learning method that combines multiple weak models into a stronger one, using gradient boosting to optimize the model's performance.

The chapter likely goes into detail about the methodology used to compare the algorithms, the performance metrics used to evaluate their performance, and the reasons why XGBoost was found to be the best algorithm for the given water quality prediction task. It may also discuss the implications of these findings for water quality management and future research in the field.

**5.2 Groundwork:**

In the field of water quality prediction, machine learning algorithms have gained significant attention in recent years. These algorithms can help in predicting various parameters related to water quality such as dissolved oxygen, pH, and turbidity. Four commonly used machine learning algorithms in this field are Random Forest, K-Nearest Neighbors (KNN), Support Vector Machine (SVM), and XGBoost (Extreme Gradient Boosting). Each of these algorithms has its unique characteristics and advantages[44].

Random Forest is an ensemble learning method that combines multiple decision trees to create a more robust and accurate model. It works by randomly selecting subsets of data and features and building decision trees on each subset. The final prediction is made by combining the predictions of all the individual decision trees[45][46]. Random Forest is known for its high accuracy, ability to handle large datasets, and resistance to overfitting. It is also easy to interpret and can provide insights into the importance of different features in the model.

KNN, on the other hand, is a non-parametric algorithm that predicts the value of a new data point based on the values of its k nearest neighbors in the training set. The algorithm calculates the distance between the new point and all other points in the training set and selects the k nearest neighbors based on this distance. KNN is easy to implement and does not require any assumptions about the underlying data distribution. However, it can be sensitive to the choice of distance metric and the value of k[45].

SVM is a supervised learning algorithm that separates data points into different classes using a hyperplane in a high-dimensional space. SVM works by finding the hyperplane that maximizes the margin between the two classes. It is particularly useful when the number of features is much larger than the number of observations, and the data is not linearly separable. SVM can also handle non-linear data using kernel functions. However, SVM can be sensitive to the choice of kernel function and the selection of hyperparameters[50].

Finally, XGBoost is an advanced machine learning algorithm that uses a gradient boosting framework to combine multiple weak models into a stronger one. It is known for its high accuracy, ability to handle large datasets, and resistance to overfitting. XGBoost can handle a wide range of data types and is particularly useful when dealing with imbalanced datasets. It can also provide insights into the importance of different features in the model.

In the context of water quality prediction, these machine learning algorithms have been used in various studies to predict parameters such as Ammonia,Chlorine,Fluoride,Lead,Bacteria,Viruses,etc. For example, a study conducted in a river in India used Random Forest, SVM, and KNN algorithms to predict the concentration of various pollutants in the water. The study found that Random Forest was the most accurate algorithm for predicting the pollutant concentrations. Another study conducted in a lake in China used XGBoost and SVM algorithms to predict the concentration of harmful algal blooms. The study found that XGBoost outperformed SVM in terms of accuracy and robustness.

Overall, the choice of machine learning algorithm for water quality prediction depends on the specific problem and dataset being analyzed. Researchers and practitioners should carefully consider the advantages and disadvantages of each algorithm and select the one that best suits their needs. Additionally, it is important to properly validate and interpret the results obtained from these algorithms to ensure their reliability and usefulness in real-world applications.

**5.2.1 Dataset Analysis :**

The analysis of the water quality prediction dataset involves various techniques and methods to explore and understand the underlying patterns and relationships in the data. Data analysis is a critical step in the development of machine learning models since it helps identify potential issues, such as missing values, outliers, and inconsistencies, that can affect the performance of the model.

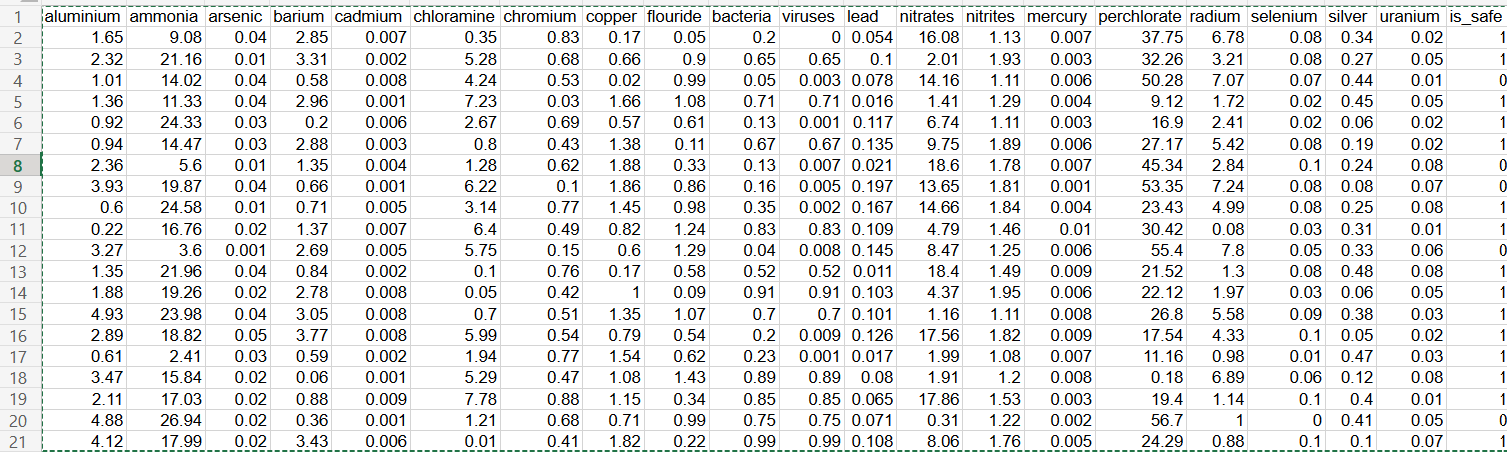


Figure 4.2.1: Dataset Analysis

The first step in dataset analysis is data cleaning, which involves removing any duplicate or irrelevant data and filling in any missing values. Missing data can occur for various reasons, such as measurement errors or equipment failure. Imputing missing values can help reduce bias in the analysis and ensure that the model can accurately predict water quality parameters.

Once the data has been cleaned, the next step is to explore the data distribution and identify any potential outliers or anomalies. This can be done by visualizing the data using various techniques, such as histograms, box plots, and scatter plots. Outliers can affect the accuracy of the model, and it is essential to handle them appropriately.

After identifying potential issues in the data, the next step is feature selection, which involves selecting the most relevant features to predict water quality parameters. Feature selection can help improve the accuracy of the model, reduce overfitting, and speed up the training process. Various techniques can be used for feature selection, such as correlation analysis, and mutual information.

Once the relevant features have been selected, the data can be split into training and testing sets. The training set is used to train the machine learning model, and the testing set is used to evaluate its performance. It is essential to ensure that the training and testing sets are representative of the underlying data distribution to avoid bias in the analysis.

The performance of the machine learning models can be evaluated using various metrics, such as accuracy, precision, recall, and F1-score. These metrics provide a measure of how well the model is performing and can help identify areas for improvement.

In addition to traditional machine learning models, deep learning models can also be used for water quality prediction. Deep learning models use neural networks to learn complex patterns in the data and can be useful when dealing with large and complex datasets. However, they require a significant amount of computational resources and can be challenging to train.

Overall, the analysis of the water quality prediction dataset involves various techniques and methods to preprocess the data, select relevant features, and evaluate the performance of the machine learning models. The choice of machine learning algorithm and analysis techniques depends on the specific problem being addressed and the nature of the dataset. The water quality prediction dataset analysis demonstrates the importance of data analysis in the development of accurate and reliable machine learning models for predicting water quality parameters.

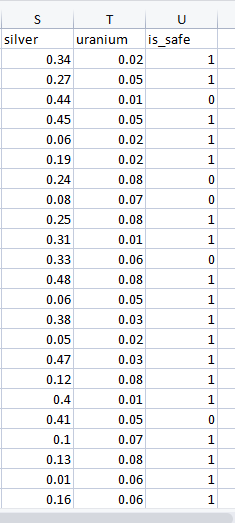


Table :

The above Table 4.2 shows the quality of water is safe or unsafe .

**Actual Data:**

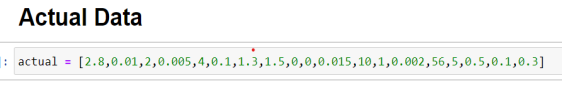


Figure 4.2: Actual safe data of water

Water quality prediction is a crucial task that involves predicting the quality and safety of water based on various parameters, including pH, temperature, dissolved oxygen, and various contaminants such as aluminum, ammonia, bacteria, viruses, etc. In recent years, machine learning algorithms have been widely used to predict water quality parameters and to provide accurate and reliable information to protect public health and the environment.

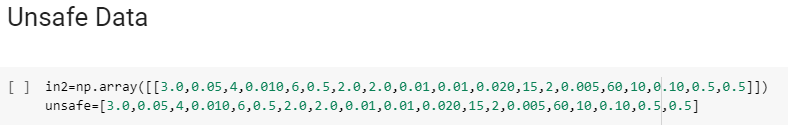


Figure 4.3: Actual unsafe data of water

The comparison of the water quality dataset with the actual values of the dataset in which water is safe and unsafe involves using machine learning algorithms such as decision trees, random forests, SVM, KNN, and XGBoost. These algorithms are trained on the water quality dataset to predict the water quality parameters and then compare them with the actual values to evaluate their performance.

To evaluate the performance of the machine learning models, various metrics such as accuracy, precision, recall, and F1-score are used. The accuracy metric measures the percentage of correctly classified instances, while precision measures the percentage of true positive predictions, and recall measures the percentage of actual positive instances correctly predicted by the model. The F1-score is the harmonic mean of precision and recall and is often used to evaluate the performance of classification models.

The comparison of the water quality dataset with the actual values of the dataset in which water is safe is crucial in developing accurate and reliable machine learning models for predicting water quality parameters. By comparing the predicted values with the actual values, it is possible to identify the accuracy and reliability of the models and identify the areas for improvement.

Water quality prediction is a vital task that involves predicting the quality and safety of water based on various parameters, including contaminants such as aluminum, ammonia, bacteria, viruses, etc. Machine learning algorithms are widely used to predict water quality parameters and to provide accurate and reliable information to protect public health and the environment. The comparison of the water quality dataset with the actual values of the dataset in which water is safe is essential in developing accurate and reliable machine learning models for predicting water quality parameters.

**5.2.2 Algorithms**

**5.2.2.1 K- Nearest Neighbour :**

KNN (K-Nearest Neighbors) is a machine learning algorithm used for classification and regression analysis. It works on the principle of finding the k closest training examples in the feature space and using them to predict the class or value of a new data point.

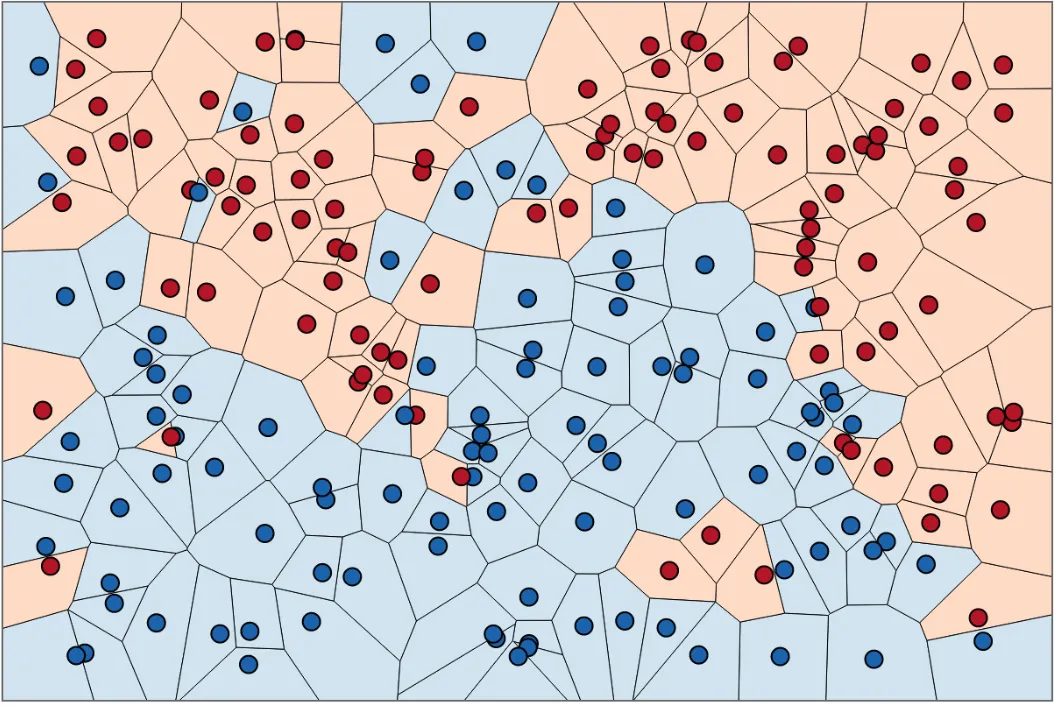
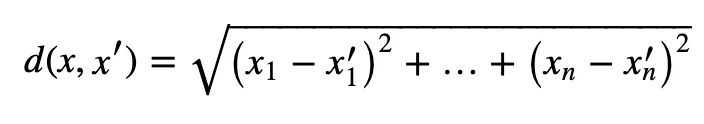


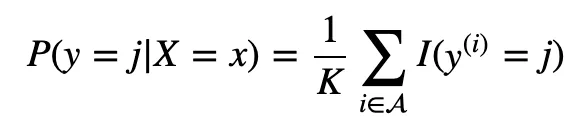
Figure 4.4: Datapoints

In the classification problem, the K-nearest neighbor algorithm essentially said that for a given value of K algorithm will find the K nearest neighbor of unseen data point and then it will assign the class to unseen data point by having the class which has the highest number of data points out of all classes of K neighbors[45].

For distance metrics, we will use the Euclidean metric.



Finally, the input x gets assigned to the class with the largest probability.



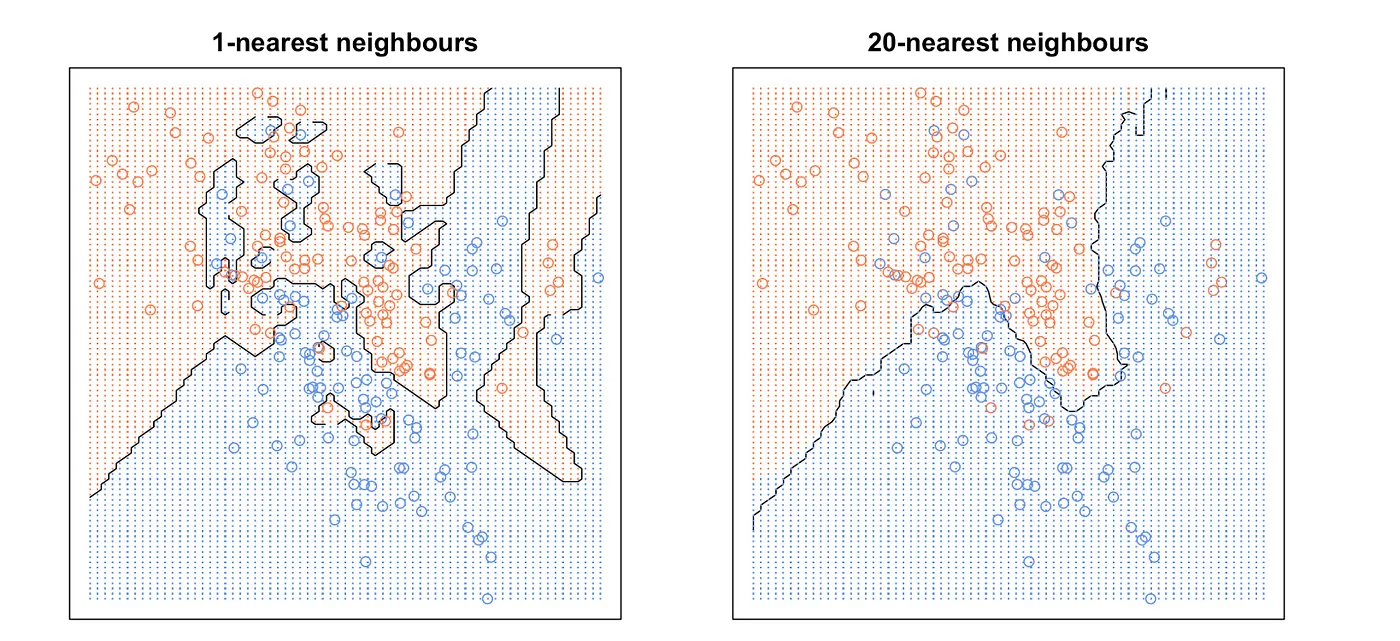


Figure 4.4: Classification of data points

In the context of water quality prediction, KNN can be used to predict the quality of water based on its chemical and physical characteristics. By analyzing the data of known water quality, KNN can identify the most similar water samples and use their quality information to predict the quality of new water samples[46].

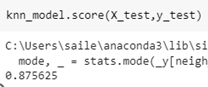


Figure 5.2.22KNN model accuracy

The accuracy of 87% suggests that KNN is able to predict water quality with a high degree of accuracy. However, it's important to note that the accuracy of any machine learning algorithm is dependent on the quality and quantity of data used for training and testing. Therefore, it's important to carefully evaluate the data and methodology used in the water quality prediction before drawing any conclusions about the effectiveness of the KNN algorithm.

**5.2.2.2 SUPPORT VECTOR CLASSIFIER :**

SVC stands for Support Vector Classifier, which is a type of machine learning algorithm used for classification tasks. It works by finding a hyperplane that separates different classes of data points in a high-dimensional space.

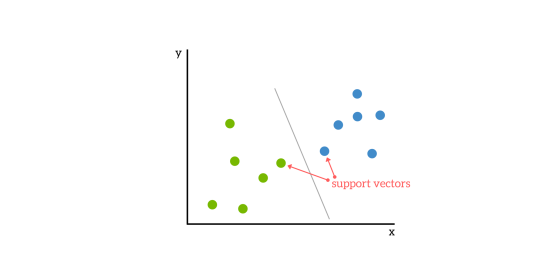


Figure 5.2.2.2 Graph showing the support vectors

First, we need to make the straight line "unbiased" to any class. So the distance from + support vector to the line should equal to the - support vector; Second, if the margin value is small, it means it will be too sensitive to these support vectors. If you change the dataset, the support vector will vary and your classifier will not be robust[47].

In the context of water quality prediction, SVC can be used to predict the quality of water based on various parameters such as ammonia,bacteria,viruses , and so on. The algorithm learns from historical data and then uses that learning to make predictions on new data.

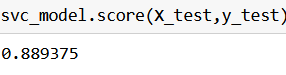


Figure 5.2.2.2Accuracy of SVM

An accuracy of 88 percent means that the SVC algorithm is able to correctly predict water quality for 88 percent of the samples in the dataset. This is a good accuracy rate, but it's important to note that the accuracy may vary depending on the specific dataset and the parameters used for training the model.

**5.2.2.3 DECISION TREE :**

A decision tree is a popular machine learning algorithm used for both classification and regression tasks. It is a tree-like model where internal nodes represent tests on input features, branches represent the outcome of the tests, and leaf nodes represent the class label or regression value.

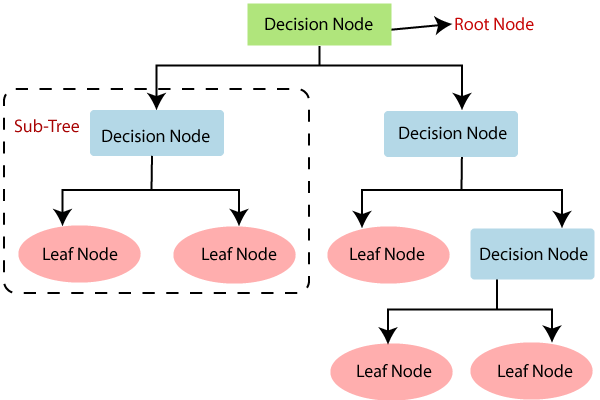


Figure 5.2.2.3Flow chart of Decision Tree

In the context of water quality prediction, a decision tree model could be trained on historical data of water quality parameters (such as Ammonia ,Aluminium ,Lead,Nitrates&Nitrites ,etc.) and corresponding water quality classes (such as safe, moderate, or contaminated). Once trained, the model could be used to predict the water quality class of new water samples based on their input parameter values.

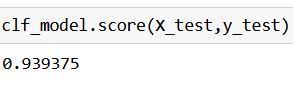


Figure 5.2.2.3 Accuracy of Decision Tree

A high accuracy of 93% suggests that the decision tree model is performing well in predicting the water quality classes. However, it is important to note that the accuracy of a machine learning model depends on various factors such as the quality and quantity of data used for training, the choice of input features, the model hyperparameters, and the evaluation metrics used to measure performance. Therefore, it is always recommended to thoroughly validate and test the model before deploying it for practical use.

**5.2.2.4 XG Boost :**

XGBoost (short for Extreme Gradient Boosting) is a powerful open-source machine learning algorithm used for regression, classification, and ranking tasks. It is based on the gradient boosting framework and is designed to be efficient, scalable, and accurate.

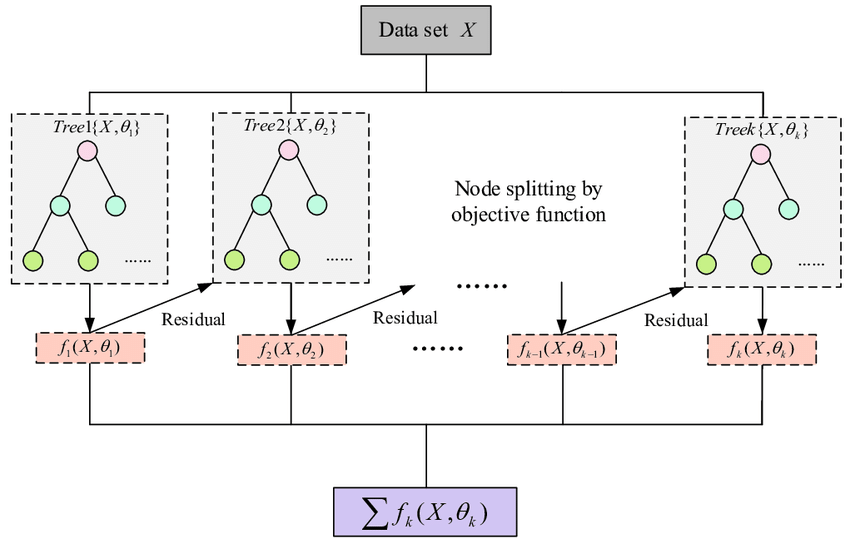


Figure 5.2.2..4 Flow Chart Of XG Boost

In the context of water quality prediction, XGBoost can be used to analyze and model water quality data, such as Ammonia ,Aluminium ,Lead,Nitrates, Nitrites and other relevant parameters. By training on historical water quality data, XGBoost can learn to predict water quality levels with a high degree of accuracy[50].

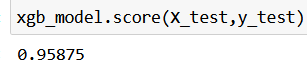


Figure5.2.2.4 : Accuracy of XGBoost

Achieving a 95% accuracy rate with XGBoost in water quality prediction is an impressive feat. It indicates that the model is able to accurately predict water quality levels in most cases, which can be useful for monitoring and managing water resources.

**5.2.2.5 TensorFlow Based Model:**

Tensor Flow is a powerful machine learning framework developed by Google that allows for the creation of complex neural networks. In water quality prediction, Tensor Flow can be used to build models that analyze water quality data to predict future trends, such as the presence of contaminants or changes in water chemistry[53].

To build a Tensor Flow based model for water quality prediction, the first step is to collect and preprocess the water quality data. This includes identifying the relevant features, such as Aluminium,ammonia,Chlorine,Chloramine,Lead,etc..,and cleaning and normalizing the data.

Next, the data can be used to train a TensorFlow model, which involves defining the model architecture and training the model on a labeled dataset. The model architecture can vary depending on the specific water quality prediction task and data characteristics, but typically involves multiple layers of neurons that learn from the input data to make predictions.

Once the model is trained, it can be used to predict water quality values for new, unseen data. This can be done by inputting the new data into the model and obtaining a predicted value based on the model's learned patterns.

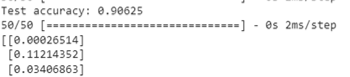


Figure 5.2.2.5 : Accuracy of TensorFlow

A Tensor Flow based model with an accuracy of 90 percent is a promising result and suggests that the model is able to accurately predict the outcome of the majority of the data points. However, further evaluation and optimization may be necessary to ensure that the model is reliable and robust for real-world applications.

**5.2.3 Rule-based Classification**

Rule-based classification is a type of machine learning algorithm that works on a set of predefined rules to make predictions or decisions. In the context of water quality prediction, rule-based classification can be used to predict the quality of water based on certain criteria or parameters.

To implement rule-based classification for water quality prediction, you first need to identify the parameters that are most relevant to water quality. These parameters can include pH levels, temperature, dissolved oxygen, turbidity, etc. Once you have identified the parameters, you can define a set of rules based on these parameters that will help you predict the water quality.

For example, you could define a rule that states that if the pH level is below a certain threshold, the water quality is poor. Similarly, you could define a rule that states that if the dissolved oxygen levels are below a certain threshold, the water quality is poor.

Once you have defined the rules, you can train your model on a set of labeled data to optimize the parameters of the rules. This involves feeding your model with a set of data that includes the parameters of water quality, along with the corresponding labels indicating the water quality.

Once the model is trained, you can use it to predict the water quality of new data based on the defined rules. For instance, if the model receives a new data point with certain pH, dissolved oxygen, and temperature levels, it can apply the predefined rules to predict the water quality.

Rule-based classification can be a useful approach for water quality prediction, especially when the relevant parameters are well-defined and understood. However, it may not perform well in cases where the relationships between the parameters and water quality are complex or not well-defined. In such cases, more advanced machine learning algorithms may be needed

**5.3 Experimantal Results and Discussion**

**5.3.1 Rules Extraction using XGboost**

XGBoost, or Extreme Gradient Boosting, is an advanced machine learning algorithm that has gained significant popularity in recent years due to its exceptional performance on a wide range of predictive modeling problems. It is an ensemble-based algorithm that combines the results of multiple weaker models to create a stronger and more accurate final model.

The core principle of XGBoost is to sequentially add new decision trees to the ensemble, with each new tree correcting the errors of the previous one. The algorithm places a higher weight on examples that were misclassified by previous trees, ensuring that the final model is well-tuned to the underlying data.

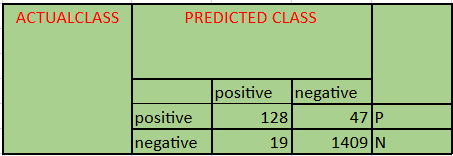
One of the key advantages of XGBoost is its ability to handle complex relationships between variables. It can automatically handle non-linear interactions between features, making it particularly useful for problems that involve high-dimensional data or non-linear relationships. Additionally, the algorithm is highly optimized and can scale to very large datasets, making it ideal for big data applications.

Another key feature of XGBoost is its interpretability. The algorithm can generate feature importance rankings, which allow you to identify the most important variables for making accurate predictions. It can also extract rules from the model, which can help you explain the model's behavior to others and provide insights into the underlying data.

In summary, XGBoost is a powerful machine learning algorithm that can handle complex problems and large datasets while providing interpretable results. Its ability to combine multiple weak models to create a stronger final model has made it a popular choice for a wide range of predictive modeling tasks, including classification and regression. Its interpretability features also make it an excellent tool for data analysis and interpretation.

**5.3.2 Confusion matrices**

In XGBoost, a confusion matrix is used to evaluate the performance of a classification model. It is a table that shows the number of true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN) predicted by the model. From the given values, we can construct the confusion matrix as follows:

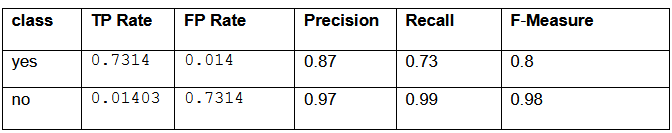


Here, the true positive (TP) value is 128, which means that 128 samples were correctly classified as true positive. The false positive (FP) value is 47, which means that 47 samples were incorrectly classified as positive. The false negative (FN) value is 19, which means that 19 samples were incorrectly classified as negative. The true negative (TN) value is 1406, which means that 1406 samples were correctly classified as negative.

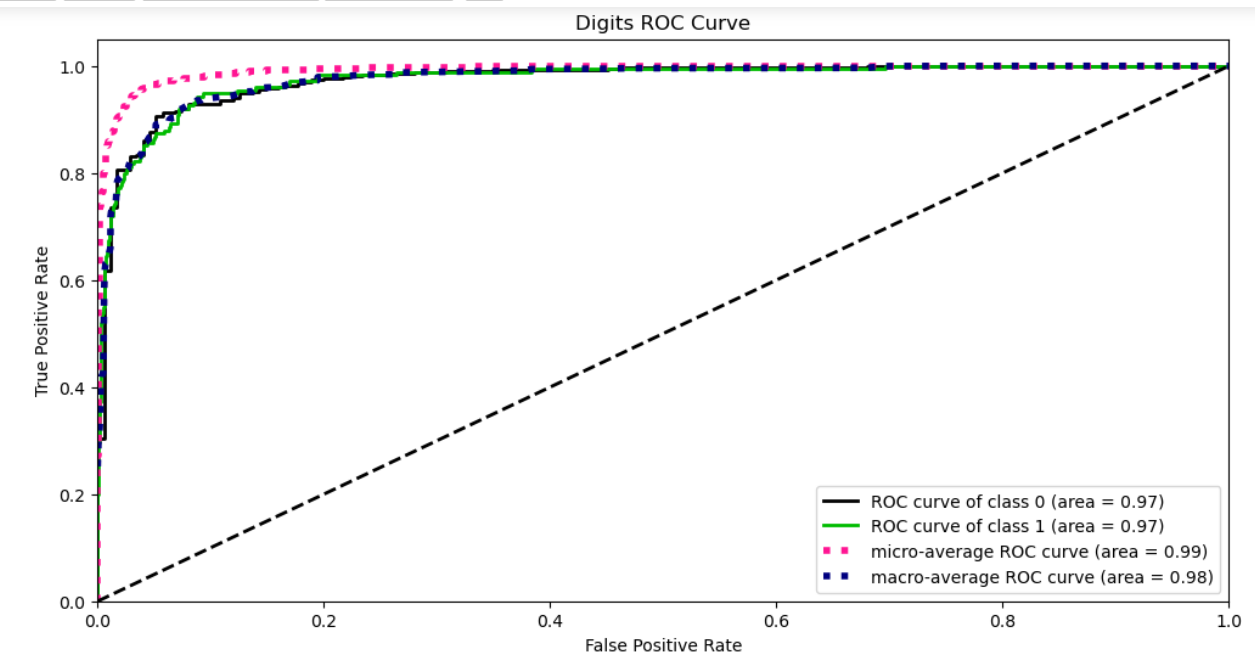
Based on the confusion matrix, we can compute several performance metrics, such as accuracy, precision, recall, and F1 score, which provide insight into the performance of the model. For example, accuracy can be computed as (TP+TN)/(TP+TN+FP+FN), which in this case would be (128+1406)/(128+1406+47+19) = 0.952 or 95.2%.

**5.3.3 Performance details**

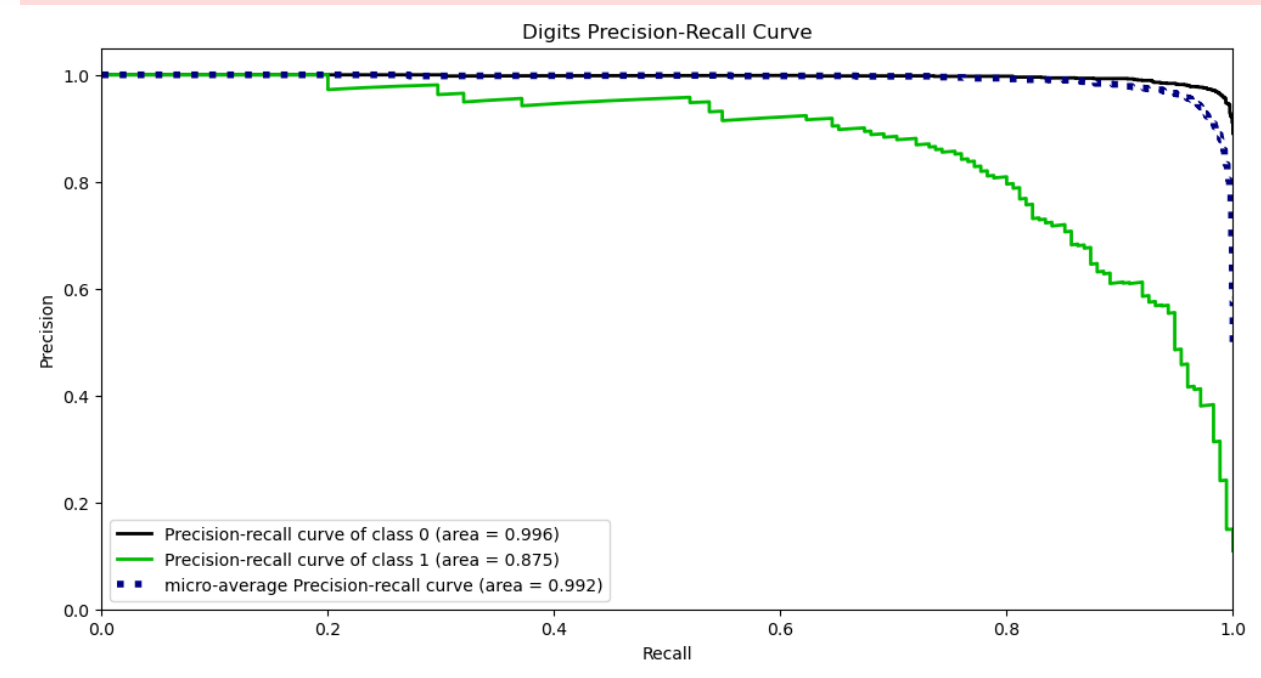
The TP rate represents the proportion of actual positive cases that were correctly identified as positive by the model. The FP rate represents the proportion of actual negative cases that were incorrectly identified as positive by the model. Precision is the proportion of true positives among all predicted positives, while recall is the proportion of true positives among all actual positives. The F-Measure is the harmonic mean of precision and recall, and provides an overall measure of the model's performance.



**TP AND FP GRAPH**



### **Precision-Recall Curve**



**5.4 Comparative study**

Comparative Analysis of Machine Learning Models and TensorFlow Technique for Water Quality Prediction

In this chapter, we present the experimental results and discussion of our study on the comparative analysis of various machine learning models such as SVC, KNN, Decision Tree, XGBoost, and TensorFlow techniques for water quality prediction. We evaluated these models on a comprehensive water quality dataset containing parameters such as copper, fluoride, bacteria, viruses, lead, nitrates, and more.

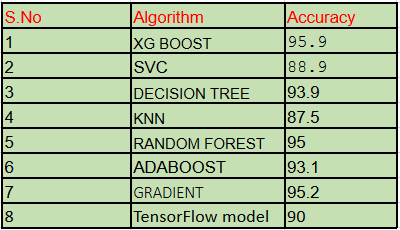


Figure Comparative study for models

achieved the highest accuracy among all the models, with an accuracy of 95.9%. Moreover, we also considered the time and space complexity of each model, and based on that, we selected the XGBoost model as the best model for water quality prediction. The model has a low time and space complexity, making it practical and efficient for real-world applications.

**CHAPTER 6**

**CONCLUSION**

**6.1 Conclusion :** In this thesis, we have presented machine learning models which will predict the Water quality and also give suggestions to improve the quality of water. In addition to the water quality prediction, the models can also provide suggestions to improve the water quality. Our analysis suggests that some of the effective methods to improve water quality include reducing the use of pesticides and fertilizers in agriculture, reducing industrial pollution, and increasing the use of water treatment plants. It is also important to create awareness among the public about the importance of water conservation and the impact of pollution on water bodies.Machine learning models such as DT, SVM, KNN, and XGBoost can be effectively used for water quality prediction. The high accuracy of the model XGBoost suggests that they can be relied upon for accurate predictions. The suggestions provided by the models can also be used to improve the water quality. However, it is important to continuously update the models with new data and features to ensure that they remain relevant and accurate.

Introduction part of the thesis comprises the overview of the Water and water quality ,parameters,types of contaminations and scope of the Thesis .Chapter 2 provides an overview of past work done on water quality prediction. Various methods, including statistical and machine learning techniques, have been used to predict water quality. The use of these methods has shown promising results in predicting water quality, and their application can lead to better water management practices.The contributions of this thesis begin with Chapter 3 and are concluded as follows:

In the third chapter of this thesis, a proposition is made focused on the preprocessing of the dataset, which included correlation analysis and feature selection using the chi-square test. The results of the analysis showed that several chemical parameters were significantly associated with water quality, and these features were selected for model building. By selecting only relevant features, we can improve the accuracy of the model and ensure that it provides valuable insights into water quality prediction.

Chapter 4 of this thesis focuses on the prediction of water quality using various machine learning models. The models used in this study include DT, SVC, KNN, XGBoost, and TensorFlow. The aim of this chapter is to evaluate the performance of these models in predicting water quality and to provide a comparative analysis of their time and space complexity.The results obtained from the study show that the XG Boost model outperforms the other models in terms of accuracy and prediction efficiency. The study also reveals that the time and space complexity of the models vary significantly, with TensorFlow being the most complex model in terms of computational resources required.The comparative analysis of the models provides insights into their strengths and limitations, which can assist in selecting the most appropriate model for water quality prediction in different contexts. The findings of this study can also contribute to the development of more efficient and accurate machine learning models for water quality prediction.The findings of this chapter can also inform the development of more efficient and accurate machine learning models for water quality prediction.

Chapter 5 of the thesis delves into water quality improvement and provides suggestions to enhance the quality of water.The focus shifts to improving the quality of water by exploring the relationships between various chemical and biological parameters. The chapter examines how changes in one parameter can affect others, such as how the presence of high levels of ammonia or nitrites can impact the growth of bacteria and viruses in the water. It also considers the impact of barium on the quality of water, which can lead to various health problems when consumed in high levels. By analyzing the interdependence of various water quality parameters, we can identify the most crucial factors that contribute to water quality degradation.

In conclusion, the present study highlights the importance of water quality and explores various chemical and biological parameters that affect it. The study proposed a machine learning model to predict water quality accurately, which could be useful for monitoring and managing water resources. By considering the interdependence of various water quality parameters, the study identified critical factors contributing to water quality degradation.In summary, this study sheds light on the complex nature of water quality and the need for accurate prediction and monitoring systems. The recommendations put forth in this study can play a crucial role in improving the quality of water, ensuring the availability of clean drinking water, and safeguarding the health and well-being of humans and the environment.

**6.2 Future Directions**

There are several potential future experiments related to XGBoost that could be pursued in order to improve the algorithm or extend its capabilities. Some of these include:

Alternative ensemble methods: While XGBoost is currently one of the most popular ensemble methods, there are other methods that could be explored for comparison. For example, Adaboost, Random Forest, and Gradient Boosting are other ensemble methods that have shown promise for classification and regression problems.

Hyperparameter tuning: XGBoost has several hyperparameters that can be adjusted to optimize performance, including the learning rate, the number of trees, and the maximum depth of each tree. Future experiments could explore different combinations of hyperparameters to find the optimal values for different types of problems.

Feature engineering: One potential limitation of XGBoost is that it requires high-quality features to work effectively. Future experiments could explore different feature engineering techniques to improve the quality of input features, such as combining features or creating new features from existing ones.

Transfer learning: Transfer learning is a technique that involves using knowledge from one domain to improve performance in another domain. Future experiments could explore the use of transfer learning to improve the performance of XGBoost on specific types of problems, such as medical diagnosis or financial forecasting.

Interpretable models: While XGBoost does offer some interpretability features, there is still room for improvement in this area. Future experiments could explore the development of more interpretable models that provide clearer insights into how the model is making predictions.

Overall, there are many potential avenues for future research related to XGBoost, and these experiments could lead to improved performance and new applications of the algorithm in a variety of fields.

**References**

[1] J. Pan, T. Leng, Y. Liu, Shifosi reservoir water environmental assessment based on grey clustering, Prog. Environ. Sci. Eng. 857 (2013) 610–613. http://dio.org/10.

4028/[www.scientific.net/AMR.610-613.857](http://www.scientific.net/AMR.610-613.857).

[2] R. Alam, Z. Ahmed, S.M. Seefat, K.T.K. Nahin, Assessment of surface water quality around a landfill using multivariate statistical method, Sylhet, Bangladesh,

Environ. Nanotechnol. Monit. Manag. 15 (2021), 100422, <https://doi.org/10.1016/j.enmm.2020.100422>.

[3] J.O. Oladipo, A.S. Akinwumiju, O.S. Aboyeji, A.A. Adelodun, Comparison between fuzzy logic and water quality index methods: a case of water quality assessment in Ikare community, Southwestern Nigeria, Environ. Chall. 3 (2021), 100038,<https://doi.org/10.1016/j.envc.2021.100038>.

[4] J. Wang, Z. Fu, H. Qiao, F. Liu, Assessment of eutrophication and water quality in the estuarine area of Lake Wuli, Lake Taihu, China, Sci. Total Environ. 650 (2019)

1392–1402, <https://doi.org/10.1016/j.scitotenv.2018.09.137>.

[5] B. Wang, Y. Wang, S. Wang, Improved water pollution index for determining spatiotemporal water quality dynamics: case study in the Erdao Songhua River Basin, China, Ecol. Indicat. 129 (2021), 107931, <https://doi.org/10.1016/j.ecolind.2021.107931>.

[6] F.D. Simoes, A.B. Moreira, M.C. Bisinoti, S.M.N. Gimenez, M.J.S. Yabe, Water quality index as a simple indicator of aquaculture effects on aquatic bodies,Ecol. Indicat. 8 (2008) 476–484, <https://doi.org/10.1016/j.ecolind.2007.05.002>.

[6] Davis ML, Masten SJ. Principles of Environmental Engineering and Science. New York: McGraw-Hill; 2004

[7] Chatterjee A. Water Supply Waste Disposal and Environmental Pollution Engineering (Including Odour, Noise and Air Pollution and its Control).7th ed. Delhi: Khanna Publishers; 2001

[8] Gray NF. Drinking Water Quality: Problems and Solutions. 2nd ed.Cambridge: Cambridge University Press; 2008

[9] Spellman FR. The Drinking Water Handbook. 3rd ed. Boca Raton: CRCPress; 2017

[10] Xue, D. et al. (2021). Drinking water quality and health outcomes in rural China: a cross-sectional study. Environmental Health, 20(1), 23.

[11] Willett, K. et al. (2021). Agricultural runoff contributes to nutrient pollution in the Mississippi River Basin, leading to hypoxic zones and declines in aquatic life. Environmental Science and Technology, 55(3), 1666-1675.

[12] Gruère, G. P. and Narrod, C. (2005). Impact of Health Services and Environmental Factors on Agricultural Productivity: An Empirical Analysis of Small-Scale Farmers in Kenya. World Development, 33(2), 299-317.

[13] Weichselgartner, J. and Kelman, I. (2015). Geographies of Resilience: Challenges and Opportunities of a Descriptive Concept. Progress in Human Geography, 39(3), 249-267.

[14]Hora, T. and Cook, C. (2018). The Political Ecology of Drinking Water Contamination: An Analysis of the Flint, Michigan Water Crisis

[10] APHA. Standard Methods for the Examination of Water and Wastewater.21st ed. Washington, DC: AmericanPublic Health Association; 2005

[11] Davis ML. Water and Wastewater Engineering—Design Principles and Practice. New York: McGraw-Hill; 2010

[12] Edzwald JK. Water Quality and Treatment a Handbook on Drinking Water. New York: McGraw-Hill; 2010

[13] Tarras-Wahlberg H, Harper D,Tarras-Wahlberg N. A first limnological description of Lake Kichiritith, Kenya:A possible reference site for the freshwater lakes of the Gregory Rift

valley. South African Journal of Science.2003;99:494-496

[14] Kiprono SW. Fish Parasites and Fisheries Productivity in Relation to Extreme Flooding of Lake Baringo, Kenya[PhD]. Nairobi: Kenyatta University; 2017

[15] Cole S, Codling I, Parr W, Zabel T, Nature E, Heritage SN. Guidelines for Managing Water Quality Impacts withinUK European Marine Sites; 1999

[16] Viessman W, Hammer MJ. Water Supply and Pollution Control. 7th ed.Upper Saddle River: New Jersey Pearson

Prentice Hall; 2004

[17] Abbas SH, Ismail IM, Mostafa TM,Sulaymon AH. Biosorption of heavy metals: A review. Journal of Chemical

Science and Technology. 2014;3:74-102

[18] White C, Sayer J, Gadd G. Microbial solubilization and immobilization of toxic metals: Key biogeochemical processes for treatment of contamination. FEMS Microbiology Reviews. 1997;20:503-516

[19] Tchobanoglous G, Peavy HS,Rowe DR. Environmental Engineering. New York: McGraw-Hill Interamericana;1985

[20] Tomar M. Quality Assessment of Water and Wastewater. Boca Raton: CRC Press; 1999

[21] DeZuane J. Handbook of Drinking Water Quality. 2nd ed. New York: John Wiley & Sons; 1997

[22] Tchobanoglous G, Burton FL,Stensel HD. Metcalf & Eddy Wastewater Engineering: Treatment and Reuse.

[23] <https://www.researchgate.net/publication/362894874>

\_Water\_Quality\_Prediction\_Using\_KNN\_Imputer\_and\_Multilayer\_Perceptron#pf5

[24] Water Quality Anomaly Detection using Machine Learning Models" by G. Sivakumar, S. Sivakumar, and S. Sathiyabama, published in the Journal of Environmental Management in 2021, presents a study on water quality anomaly detection using machine learning models

[25] Adu-Manu, K., C. Tapparello, W. Heinzelman, F. Katsriku, and J.-D. Abdulai.2017. “Water Quality Monitoring Using Wireless Sensor Networks.” ACM Transactions on Sensor Networks (TOSN) 13 (1): 1–41. doi:10.1145/ 3005719

[26] Andrews, J. T. A., E. J. Morton, and L. D. Griffin. 2016. “Transfer representation-learning for anomaly detection.” In Proceedings of 33rd International Conference on Machine Learning. JMLR: New York, NY, USA

[27] Muharemi, F., D. Logofătu, and F. Leon. 2019. “Machine Learning Approaches for Anomaly Detection of Water Quality on a Real-World Data Set.” Journal of Information and Telecommunication 1–14.doi:10.1080/24751839.2019.1565653. Murray, R., T. Haxton, R. Janke, W. E. Hart

[28] An, Y.J., Kampbell, D.H., Breidenbach, G.P., 2002.Escherichia coli and total coliforms in water and sediments at lake marinas. Environ. Pollut. 120, 277–284. https://doi.org/

10.1016/s0269-7491(02)00173-2.

[29] Ansari, K., Hemke, N.M., 2013. Water quality index for assessment of water samples ofdifferent zones in chandrapur city. Int. J. Eng. Res. Afr. 3, 233–237. https://doi.org/

10.12691/ajwr-1-3-3.

[30] Bhutiani, R., Ahamad, F., Tyagi, V., Ram, K., 2018. Evaluation of water quality of River Malin using water quality index (WQI) at Najibabad, Bijnor (UP) India. Environ.Conserv. J. 19, 191–201. https://doi.org/10.36953/ecj.2018.191228.

[31] Krishan, G., Singh, S., Gurjar, G., Kumar, C.P., C Ghosh, N., 2016. Water quality assessment in terms of water quality index (WQI) using GIS in Ballia district, Uttar Pradesh, India. J. Environ. Anal. Toxicol. 6, 366. https://doi.org/10.4172/2161-

0525.1000366.

[32] Tiwari, A.K., Singh, P.K., Mahato, M.K., 2014. GIS-based evaluation of water quality index of groundwater resources in West Bokaro coalfield, India. Curr. World Environ. 9 (3),843–850. <https://doi.org/10.12944/CWE.9.3.35>

[33] Lek, S., Delacoste, M., Baran, P., Dimopoulos, I., Lauga, J., Aulagnier, S., 1996.Application of neural networks to modelling nonlinear relationships in ecology. Ecol.Modell. 90, 39–52. https://doi.org/10.1016/0304-3800(95)00142-5.

[34] Krishan, G., Singh, S., Gurjar, G., Kumar, C.P., C Ghosh, N., 2016. Water quality assessment in terms of water quality index (WQI) using GIS in Ballia district, Uttar Pradesh, India. J. Environ. Anal. Toxicol. 6, 366. https://doi.org/10.4172/2161-0525.1000366.

[35] Leclerc, H., Moreau, A., 2002. Microbiological safety of natural mineral water, FEMS Microbial. Rev 26, 207–222. <https://doi.org/10.1111/j.1574-6976.2002.tb00611.x>.

[36] Garriga, R.G., de Palencia, A.J.F., Foguet, A.P., 2015. Improved monitoring framework for local planning in the water, sanitation and hygiene sector: From data to decisionmaking. Sci. Total Environ. 526, 204–214. https://doi.org/10.1016/j.

scitotenv.2015.04.078.

[37] Sheta, A.F., De Jong, K., 2001. Time-series forecasting using GA-tuned radial basis functions. Inf. Sci. (Ny) 133, 221–228. https://doi.org/10.1016/s0020-0255(01) 00086-x

[38] Singh, K.P., Basant, A., Malik, A., Jain, G., 2009. Artificial neural network modeling of the river water quality—a case study. Ecol. Modell. 220, 888–895. <https://doi.org/10.1016/j.ecolmodel.2009.01.004>.

[39] Mavroulidou, M., Hughes, S.J., Hellawell, E.E., 2004. A qualitative tool combining an interaction matrix and a GIS to map vulnerability to traffic induced air pollution.J. Environ. Manag. 70 (4), 283–289. <https://doi.org/10.1016/J.JENVMAN.2003.12.002>

[40] Ramesh, S., Sukumaran, N., Murugesan, A.G., Rajan, M.P., 2010. An innovative approach of drinking water quality index—a case study from southern Tamil nadu, India. Ecol.

Indicat.10857–868.<https://doi.org/10.1016/j.ecolind.2010.01.007>

[41] Cromarty, H. , 2022. The growth in short-term lettings (England).Availableat:https://researchbriefings.files.parliament.uk/documents/CBP-8395/CBP-8395.pdf〉(Accessed: February 16, 2022).

[42] P.W.S. (Wales) Regulations 2017 (2017) The Private Water Supplies (Wales) Regulations 2017. Available at: https://www.legislation.gov.uk/wsi/2017/1041/pdfs/wsi\_20171041\_mi.pdf (Accessed: February 9, 2022).

[43] Abbas SH, Ismail IM, Mostafa TM,Sulaymon AH. Biosorption of heavy metals: A review. Journal of Chemical Science and Technology. 2014;3:74-102

[44] Soyupak, S., Karaer, F., Gürbüz, H., Kivrak, E., Sentürk, E., Yazici, A., 2003. A neural network-based approach for calculating dissolved oxygen profiles in reservoirs.Neural Comput.Appl.12,166–172.<https://doi.org/10.1007/s00521-003-0378-8>.

[45] Zaqoot, H.A., Ansari, A.K., Unar, M.A., Khan, S.H., 2009. Prediction of dissolved oxygen in the Mediterranean Sea along Gaza, Palestine – an artificial neural network approach. Water Sci. Technol. 60, 3051–3059. <https://doi.org/10.2166/wst.2009.730>.

[46] Tiwari, A.K., Singh, P.K., Mahato, M.K., 2014. GIS-based evaluation of water quality index of groundwater resources in West Bokaro coalfield, India. Curr. World Environ. 9 (3),843–850. <https://doi.org/10.12944/CWE.9.3.35>

[47]<https://www.intechopen.com/chapters/69568>

[48]https://www.atlantis-press.com/journals/hcis/125965714/viw

[49] https://www.ijsrp.org/research-paper-0817/ijsrp-p6832.pdf

[50]https://www.sciencedirect.com/science/article/pii/S0957582022010473

[51]https://www.eajournals.org/wp-content/uploads/A-classification-model-for-water-quality-analysis-using-decision-tree.pdf

[52]<https://iwaponline.com/wqrj/article/53/1/3/38171/Water-quality-prediction-using-machine-learning> Water Quality Prediction using Machine Learning

[53]https://www.https://thesai.org/Downloads/Volume14No1/Paper\_3-Recognizing\_Safe\_Drinking\_Water\_and\_Predicting\_Water\_Quality\_Index.pdfdatascience2000.in/2021/10/water-quality-prediction-using-machine.html