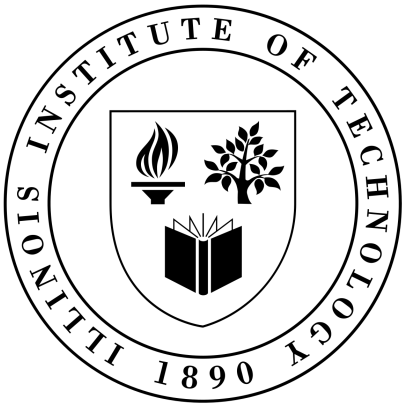


CS-584 Machine Learning

Final Project Report

Water Quality Analysis

Submitted by

Group 23

Aman Sahu (A20492367)

Devansh Goel (A20490554)

Rohit Sharma (A20475953)

GitHub Link: https://github.com/RohitSharma719/CS584_Group23_Water-Quality-Analysis

Abstract

In the past years, there has been a lot of interest in water quality and its prediction as there are many pollutants that affect water quality. The techniques provided herein will help us analyse the water quality. With the rapid increase in the volume of data on the aquatic environment, machine learning has become an important tool for data analysis, classification, and prediction. Unlike traditional models used in water-related research, data-driven models based on machine learning can efficiently solve more complex nonlinear problems. In water environment research, models and conclusions derived from machine learning have been applied to the construction, monitoring, simulation, evaluation, and optimization of various water treatment and management systems. Additionally, machine learning can provide solutions for water pollution control, water quality improvement, and watershed ecosystem security management. In this project we have gathered different features to build a machine learning model that can identify if water is safe for drinking or not. Furthermore, we propose possible future applications of machine learning approaches to water environments.

Introduction

With rapid economic development, wastewater containing various pollutants is generated, posing serious threats to natural water environments. To a large extent, water quality analysis and evaluation have substantially improved the efficiency of water pollution control. To date, many methods have been developed to monitor and assess water quality worldwide, such as the multivariate statistical method, fuzzy inference ,and the water quality index (WQI) . For evaluating water quality, although most water quality parameters can be monitored, the final water quality evaluation results may widely vary owing to the choice of parameters .Considering all water quality parameters is unrealistic because it is not only expensive and technically difficult but also fails to deal with the variability in water quality .However, in recent years, with the advances in machine learning methods, an increasing number of researchers believe that vast amounts of data can be successfully captured and analysed to meet the complex and large-scale water quality evaluation requirements.

In machine learning, a branch of artificial intelligence, algorithms are used to analyze data and attempt to mine potential patterns in the data to predict new information. As a new data analysis and processing method, machine learning has been widely used in many fields owing to its high precision, flexible customization, and convenient extensibility. Complex nonlinear relational data can be easily handled with machine learning, which facilitates the discovery of the underlying mechanisms. The excellent adaptability of machine learning has demonstrated its potential as a tool in the fields of environmental science and engineering in recent years. Therefore, more accurate evaluation results can be expected despite the complexity of using machine learning for water quality analysis and evaluation.

This project involves dataset with various features to analyze water quality using different Machine Learning algorithms. It develops a predictive model based on data about water quality parameters such as pH, hardness, solids, and others. The model is used to identify potential water contamination and alert about any hazardous water sources. The model also provides a way to monitor water sources, as well as identify areas in need of improvement. The project tries to build a model that can help ensure safe drinking water is available for everyone.

Problem Description

Water pollution is the contamination of water sources by substances which make the water unusable for drinking, cooking, cleaning, swimming, and other activities. Pollutants include chemicals, trash, bacteria, and parasites. All forms of pollution eventually make their way to water.

Water is the most significant resource of life, crucial for supporting the life of most existing creatures and human beings. Living organisms need water with enough quality to continue their lives. There are certain limits of pollutions that water species can tolerate. Exceeding these limits affects the existence of these creatures and threatens their lives. Access to safe drinking-water is essential to health, a basic human right and a component of effective policy for health protection. This is important as a health and development issue at a national, regional, and local level. In some regions, it has been shown that investments in water supply and sanitation can yield a net economic benefit, since the reductions in adverse health effects and health care costs outweigh the costs of undertaking the interventions.

Massive increases in population, the industrial revolution, and the use of fertilizers and pesticides have led to serious effects on the Water Quality environments. Thus, having models for the prediction of the Water Quality is of great help for monitoring water contamination.

In this project, we have tried build a model that can predict water is drinkable or not based on the data we have gathered. The Data contains following features:

The water_potability.csv file contains water quality metrics for 3276 different water bodies.

1. pH value:

PH is an important parameter in evaluating the acid–base balance of water. It is also the indicator of acidic or alkaline condition of water status. WHO has recommended maximum permissible limit of pH from 6.5 to 8.5. The current investigation ranges were 6.52–6.83 which are in the range of WHO standards.

2. Hardness:

Hardness is mainly caused by calcium and magnesium salts. These salts are dissolved from geologic deposits through which water travels. The length of time water is in contact with hardness producing material helps determine how much hardness there is in raw water. Hardness was originally defined as the capacity of water to precipitate soap caused by Calcium and Magnesium.

3. Solids (Total dissolved solids - TDS):

Water has the ability to dissolve a wide range of inorganic and some organic minerals or salts such as potassium, calcium, sodium, bicarbonates, chlorides, magnesium, sulfates etc. These minerals produced un-wanted taste and diluted color in appearance of water. This is the important parameter for the use of water. The water with high TDS value indicates that water is highly mineralized. Desirable limit for TDS is 500 mg/l and maximum limit is 1000 mg/l which prescribed for drinking purpose.

4. Chloramines:

Chlorine and chloramine are the major disinfectants used in public water systems. Chloramines are most commonly formed when ammonia is added to chlorine to treat drinking water. Chlorine levels up to 4 milligrams per liter (mg/L or 4 parts per million (ppm)) are considered safe in drinking water.

5. Sulfate:

Sulfates are naturally occurring substances that are found in minerals, soil, and rocks. They are present in ambient air, groundwater, plants, and food. The principal commercial use of sulfate is in the chemical industry. Sulfate concentration in seawater is about 2,700 milligrams per liter (mg/L). It ranges from 3 to 30 mg/L in most freshwater supplies, although much higher concentrations (1000 mg/L) are found in some geographic locations.

6. Conductivity:

Pure water is not a good conductor of electric current rather's a good insulator. Increase in ions concentration enhances the electrical conductivity of water. Generally, the amount of dissolved solids in water determines the electrical conductivity. Electrical conductivity (EC) actually measures the ionic process of a solution that enables it to transmit current. According to WHO standards, EC value should not exceeded 400 μS/cm.

7. Organic_carbon:

Total Organic Carbon (TOC) in source waters comes from decaying natural organic matter (NOM) as well as synthetic sources. TOC is a measure of the total amount of carbon in organic compounds in pure water. According to US EPA < 2 mg/L as TOC in treated / drinking water, and < 4 mg/Lit in source water which is use for treatment.

8. Trihalomethanes:

THMs are chemicals which may be found in water treated with chlorine. The concentration of THMs in drinking water varies according to the level of organic material in the water, the amount of chlorine required to treat the water, and the temperature of the water that is being treated. THM levels up to 80 ppm is considered safe in drinking water.

9. Turbidity:

The turbidity of water depends on the quantity of solid matter present in the suspended state. It is a measure of light emitting properties of water and the test is used to indicate the quality

of waste discharge with respect to colloidal matter. The mean turbidity value obtained for Wondo Genet Campus (0.98 NTU) is lower than the WHO recommended value of 5.00 NTU.

10. Potability:

Indicates if water is safe for human consumption where 1 means Potable and 0 means Not potable.

All null values have been handled using mean distribution and random distribution.

Methods Used:

1. Logistic Regression:
   Logistic regression classification is a statistical method used to predict the probability that an event will occur. It is a type of supervised machine learning algorithm that uses a logistic function to predict the probability of an event occurring. Logistic regression can be used for binary classification problems, where the target variable has two possible outcomes (e.g. yes/no, high/low, true/false).
   In our project, Logistic Regression algorithm has been used to classify the input as potable or not potable.

   Equation of Logistic Regression:
   $y = e^{(b0 + b1X)} / (1+ e^{(b0 + b1X)})$
   Where:
   x = input value
   y = predicted output
   b0 = bias or intercept term
   b1 = coefficient for input (x)

   This equation is similar to linear regression, where the input values are combined linearly to predict an output value using weights or coefficient values. However, unlike linear regression, the output value modelled here is a binary value (0 or 1) rather than a numeric value.

2. KNN
   K Nearest Neighbors is a classification algorithm that operates on a very simple principle. It is supervised machine learning because the data is labeled and it is used to make predictions. It is a non-parametric method which means it does not make any assumptions about the underlying data.

   The KNN algorithm uses a data-point's nearest neighbors to determine its class. The data-point's neighbors are determined by calculating the distance between the data-point and its nearest neighbors. The class of the data-point is then determined by looking at the class of its nearest neighbors. For example, if k = 5 and the five nearest

neighbors of a data-point are all labeled as “A”, then the data-point is also labeled as “A”.

KNN is often used in applications such as pattern recognition, data mining and image processing. It is also used for solving problems in areas such as finance, e-commerce, and bioinformatics.

Using Euclidean metric for distance metrics:
$d(x , x') = sqrt((x_1 – x'_1)^2 + ... + (x_n – x'_n)^2)$

where the input x assigned to the class with largest probability either potable or not.
$P(y = j \mid X = x) = 1/k \quad \Sigma_{i \in A} I(y^i = j)$

3. SVM
   Support vector machines (SVMs) are supervised learning models that are used for classification and regression tasks. SVMs are based on the concept of finding a hyperplane that separates the data points in a feature space into two categories. By finding the optimal hyperplane, SVMs can classify data points into one of two classes.

   SVMs are powerful supervised learning models that are most commonly used for binary classification tasks. SVMs are also used for regression tasks, where the goal is to predict the value of a continuous target variable. SVMs can be used for non-linear classification tasks as well, by using kernel functions to map data into a higher-dimensional space and then finding the optimal hyperplane in this space.

   SVMs are a powerful and versatile tool for supervised learning tasks and have been used in a variety of applications, such as text classification, face detection, gene expression analysis, and medical diagnosis.

4. Decision Tree
   Decision tree classification is a method of machine learning that uses a decision tree to generate rules for classification. It is a supervised learning technique that builds classification models in the form of a tree structure. The decision tree is made up of nodes which represent a test on an attribute, branches that represent the outcomes of the test, and leaves that represent the class label. The decision tree is built by splitting the training data into smaller and smaller subsets based on the attribute values. The tree is built top-down and the most significant attribute is chosen at each step. Once the tree is built, it can be used to predict the class of new instances.

   Impurity in Decision tree is calculate by following methods:
   a. Entropy : Entropy is the amount of information needed to accurately describe data. So, if data is homogenous that is all elements are similar then entropy is 0 (that is pure), else if elements are equally divided then entropy move towards 1 (that is impure).

Entropy = $- \Sigma^{n}_{i=1} p_i * \log(p_i)$

b. Gini Index: It measures impurity in the node. It has a value between 0 and 1. So the Gini index of value 0 means the sample is perfectly homogeneous and all elements are similar, whereas, the Gini index of value 1 means maximal inequality among elements.

Gini Index = $1 - \Sigma^{n}_{i=1} p_i^2$

5. Random forest

Random forest classification is a type of supervised machine learning algorithm that is used for classification tasks. It is based on the concept of decision trees, where each tree in the forest is trained using a different set of data. The output of each tree is then combined to make a final prediction. The algorithm is often used in classification problems such as facial recognition, speech recognition, and object recognition.

Working of Random Forest:

Step 1 − First, start with the selection of random samples from a given dataset.

Step 2 − Next, this algorithm will construct a decision tree for every sample. Then it will get the prediction result from every decision tree.

Step 3 − In this step, voting will be performed for every predicted result.

Step 4 − At last, select the most voted prediction result as the final prediction result.

6. XG Boost:

XGBoost is an implementation of gradient boosted decision trees designed for speed and performance that is dominative competitive machine learning. It is an optimized distributed gradient boosting library designed to be highly efficient, flexible and portable. It implements machine learning algorithms under the Gradient Boosting framework. XGBoost provides a parallel tree boosting (also known as GBDT, GBM) that solve many data science problems in a fast and accurate way.

XGBoost uses Taylor series to approximate the value of the loss function for a base learner ft(xi)

$f(x) = f(a) + f'(a)(x-a) + ½ f''(a)(x-a)^2 + ...f^n(a)(x-a)^n$

$L^{(t)} = \Sigma^{n}_{i=1} [l(y_i , y'^{(t-1)}) + g_i f_t(x_i) + ½ h_i f_t^2(x_i)] + \Omega(f_t)$

Where:

$g_i = \partial_{y'^{(t-1)}} l(y_i , y'^{(t-1)})$ and $h_i = \partial^2_{y'^{(t-1)}} l(y_i , y'^{(t-1)})$

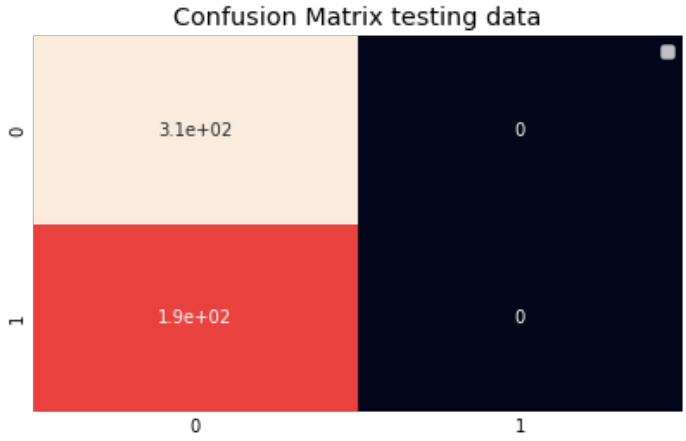
Result

To classify and predict the potability of water, we have used algorithms described earlier.

The results of the machine learning project about water quality analysis show that the Random Forest algorithm provides the best performance. The algorithm achieved an accuracy score of 0.6856, which is significantly higher than the other tested algorithms. Additionally, the Random Forest algorithm is able to accurately classify the different water quality levels. These results demonstrate that the Random Forest algorithm is an effective method for accurately predicting water quality levels.

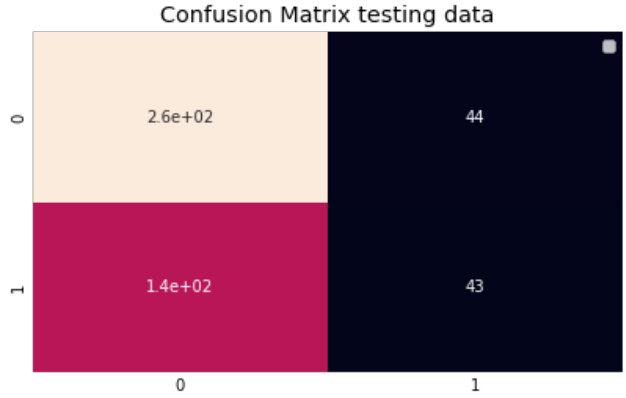
Logistic Regression :

Accuracy: 62.76%



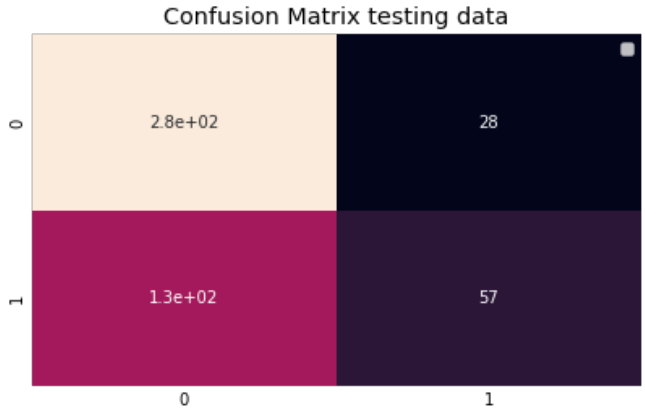
KNN

Accuracy: 61.99%



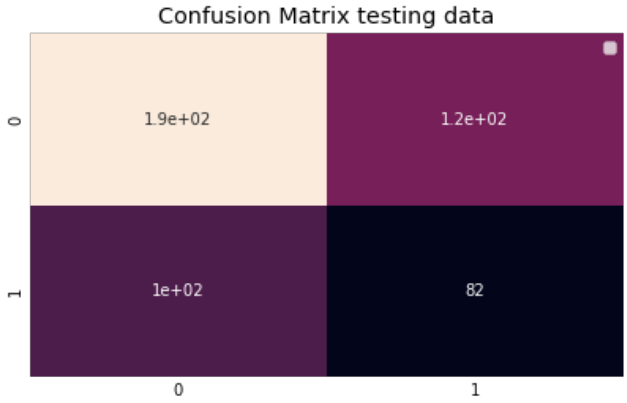
SVM:

Accuracy: 68.08%



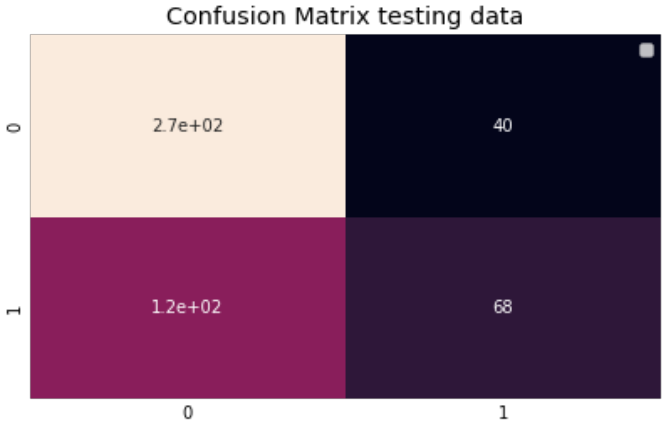
Decision Tree:

Accuracy: 55.48%



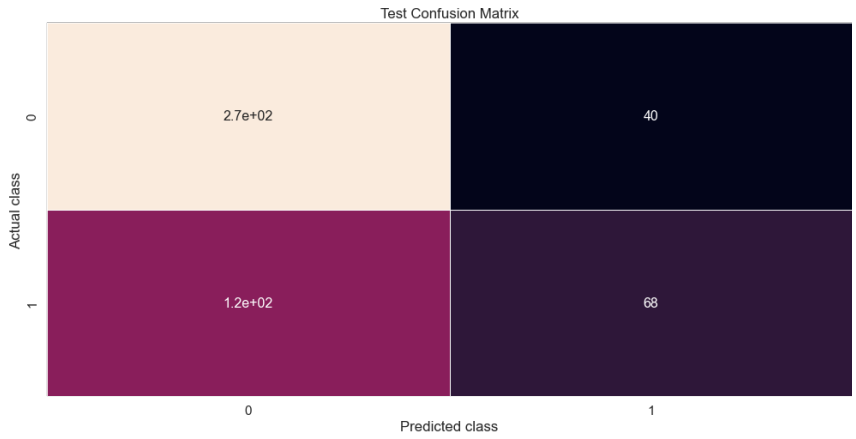
Random Forest:

Accuracy: 67.88%



XG Boost:

Accuracy: 65%



Conclusion and Future Work

The Water Quality analysis Machine Learning project has been a successful effort to use machine learning algorithms to analyze and classify water quality data. The project has demonstrated that machine learning can be used to accurately predict water quality from a variety of sources. The main contributions of this project include:

• Developing an efficient water quality analysis framework based on machine learning algorithms.

• Evaluating the accuracy of machine learning models for water quality analysis.

• Demonstrating the potential of machine learning for water quality analysis in different contexts.

The project has also highlighted the importance of incorporating domain-specific knowledge into machine learning models. In addition, it has demonstrated the need for careful data pre-processing and feature engineering to ensure high accuracy.

This project has provided valuable insight into the capabilities of machine learning for water quality analysis and has laid the foundation for future work in this area. Potential future work could include:

• Exploring different machine learning algorithms and frameworks for water quality analysis.

• Incorporating domain-specific knowledge into machine learning models.

• Investigating approaches for dealing with imbalanced datasets.

• Applying machine learning models to other related water quality datasets.