Water Quality Analysis and Modeling

Katie Chan

Department of Atmospheric and Oceanic Science, University of California Los Angeles

AOS C111: Introduction to Machine Learning for Physical Sciences

Dr. Alexander Lozinski

1. Introduction

Water is essential to health, poverty reduction, food security, ecosystems and education. Accessibility to water resources and potable water is a necessity as well as a global issue for many people worldwide. Water quality analysis is crucial for public health, environmental conservation and sustainable resource management. It is fundamental in ensuring the safety of drinking water as it can help identify the presence of harmful substances such as pollutants, bacteria and contaminants. Compliance with established water quality standards is central to providing safe and potable drinking water to communities. Ensuring that water is free from contaminants is vital for public health, especially because waterborne diseases such as cholera and dysentery can spread through contaminated water sources. Many countries around the world, most notably low-income countries, rely on nearby water sources such as streams or lakes and drink directly from these sources. According to the United Nations, in 2022, it was reported that 2.2 billion people lacked safe portable water and sanitation stations. With climate change and water pollution, more countries are facing challenges in safely managing their water resources with extreme weather leading to water scarcity and the degradation of water ecosystems.

With changing climate dynamics and feedback systems within the water cycle, it has become increasingly challenging to assess and manage water resources. Creating a water quality model to predict safe drinking water and distinguish portable water could inform the public about what water source is safest to drink from. Determining the most important feature in measuring water quality using machine learning could also reduce the process in keeping track of each variable and help the decision-making for communities to switch their water sources.

2. Data

For this project, I will base my model on a public domain dataset found on Kaggle that contains water quality metrics for 3,276 different water bodies throughout the world:
https://www.kaggle.com/datasets/adityakadiwal/water-potability

Within this dataset, there are nine variables to determine if water would be portable and safe drinking. The following variables are:

1. pH value: (0 - 14)
    - pH is an indicator of acidic or alkaline water conditions. The World Health Organization has recommended the pH of water to be within a range of 6.5 to 8.5.
2. Hardness: (mg/L)
    - Hardness of water is caused by calcium and magnesium salts, which are dissolved from deposits when water travels. Thus, it is defined as the capacity of water to precipitate soap.
3. Total Dissolved Solids: (mg/L)
    - Since water has the ability to dissolve various inorganic and organic minerals, these minerals have the ability to produce harsh tasting water and dilute the clarity of water. This parameter is important to figuring out the use of water such as drinking, personal hygiene and agricultural use. For drinking purposes, a range of 500 mg/L to 1,000 mg/L.
4. Chloramines: (mg/L)
    - Chlorine and chloramine are major disinfectants in the treatment of water. Acceptable chlorine levels are 4 mg/L for drinking water.

5. Sulfate: (mg/L)
    - Sulfates can be found in soil and minerals, which makes its way into water sources. Chemical industries are also known to release sulfates into nearby water resources. Thus, a sulfate concentration of 3 mg/L to 30 mg/L is relatively common in freshwater systems considered safe to drink.
6. Conductivity: (µS/cm)
    - Drinkable water should not be a good conductor of electricity and should act as a good insulator. The World Health Organization has recommended values near 400 µS/cm for drinking purposes.
7. Organic Carbon: (ppm)
    - Decaying natural organic matter created organic carbon within water resources. This decaying matter is a good indicator of possible bacteria and diseases within the water. According to the Environment Protection Agency, the typical value in treated water from wastewater treatment plants and drinking water is less than 2 ppm.
8. Trihalomethanes: (µg/L)
    - Trihalomethanes is usually found in drinking water that was treated with chlorine. The concentrations of THMs vary depending on the concentration of organic matter. A maximum THM concentration of 80 µg/L is considered safe in drinking water.
9. Turbidity: (NTU)
    - Turbidity of water greatly depends on the quantity of dissolved solids, minerals and suspended material within the water. The recommended value is about 5 NTU according to the World Health Organization.

This dataset includes a category (Potability) that determines if the water was considered portable or not. A value of 1 indicates that the water is portable and a value of 0 indicates that that water is not potable.

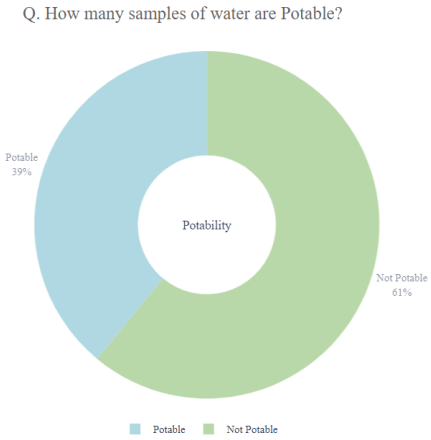
To prepare the dataset for modeling, the dataset was first uploaded to Google Drive and loaded onto a Google Colab Notebook. I used the pandas library to load the comma-separated values file as a dataframe object. After this, I found any missing data or invalid values within the dataset. There were 491 invalid pH values, 781 invalid sulfate values and 162 invalid trihalomethanes values. I calculated the mean and median values for the numeric pH, sulfate and trihalomethanes values found within the dataset. Comparing the mean and median values for potable and non-potable water, it could be seen that there were small differences between the mean and median values (Figure 1). Thus, it was decided to use the median values to replace the invalid values since it is not sensitive to outliers, if any are present within the dataset. After checking if all NaN values were replaced, the data frame was updated and successfully preprocessed. A visual representing the amount of samples that are considered to be potable and non-potable within the dataset is shown in Figure 2.

Figure 1. Mean and Median values of Potable and Non-Potable Water

Potable		
Variable	Mean	Median
pH	7.073783	7.036752
Sulfate	332.56699	331.838167
Trihalomethanes	66.539684	66.678214

Non-Potable		
Variable	Mean	Median
pH	7.085378	7.035456
Sulfate	334.56429	333.389426
Trihalomethanes	66.303555	66.542198

Figure 2. How many samples of water are Potable?



3. Modeling

For this project, the main purpose of the investigation is to test some modeling algorithms that would best fit the dataset. Thus, I will be testing logistic regression, decision tree regression and random forest regression. Logistic regression, decision tree regression, and random forest are all machine learning algorithms commonly used for classification tasks, such as determining if water is potable or non-potable. Each of these tactics has its own strength and weakness so I will test the accuracy of each algorithm.

Logistic regression is well-suited for binary classification, where the outcome is either yes or no. In this case, the binary classification would determine if the water is potable or non-potable. However, it assumes a linear relationship between the input features (eg. pH, hardness, total dissolved solids, chloramines, etc.) and target variable (eg. potability). To prepare the data to get trained and tested, I isolated the target variable as potability and dropped the target variable from the input features. Normalizing the data is recommended for logistic regression so that each feature is taken into account in the model on a similar scale. After normalizing the dataset's input features, the training and testing data was split through using sklearn's *train_test_split* function. With a test size of 20% and random state of 42, the model is set to use 80% of the data for training the model and the remaining 20% will be used for evaluating the model's performance on data it has not seen before. After training and testing, the logistic regression model resulted in a low accuracy of 62.80%. Figure 3 shows the ROC curve for the logistic regression model.

Decision trees are capable of capturing nonlinear relationships within the dataset. If the relationship between input features and water potability is complex and nonlinear, the accuracy of modeling the dataset with a decision tree could increase. Since normalizing data is not necessary or recommended for decision trees, the data used for training and testing was not normalized. Initially, the decision tree was allowed to grow and split without any limits. This led to an accuracy of 57.62% for the decision tree without constraints. Considering the low accuracy of the decision tree, I decided to conduct another decision tree with a maximum depth of three. This sets a constant on the number of levels within the tree during the training process. This improved the accuracy of the decision tree and resulted in an accuracy of 63.87%. Figure 4 and 5 showcases the decision trees without constraints and with a maximum depth of three.

Random Forest is an ensemble learning method that builds multiple decision trees and combines their predictions. This helps improve predictive accuracy and generalization by reducing overfitting. Similarly to the decision tree, the dataset that was not normalized was used. This led to an accuracy of 67.38% for the random forest model, which was the best accuracy obtained. The ROC curve of the resulting model is shown in Figure 6. Since this was the best performing model, the importance of each feature was calculated and is shown in Figure 7.

4. Results

Figure 3. ROC Curve for Linear Regression

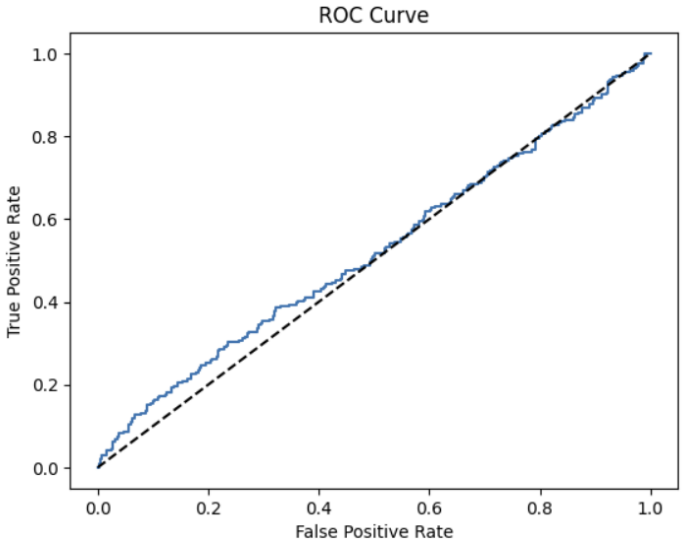


Figure 4. Decision Tree without constraints

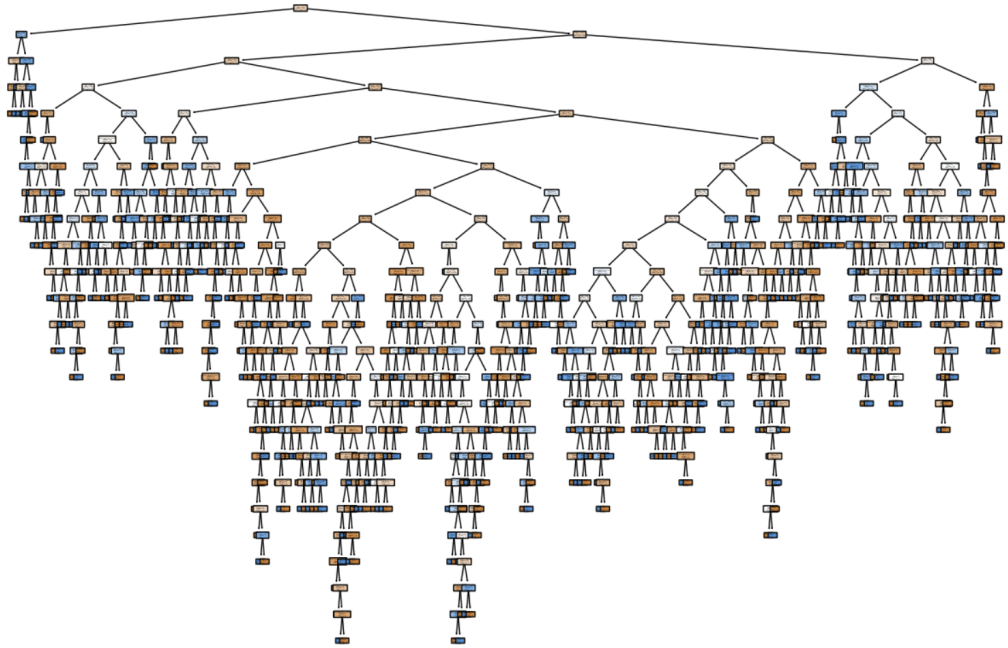


Figure 5. Decision Tree with a maximum depth of 3

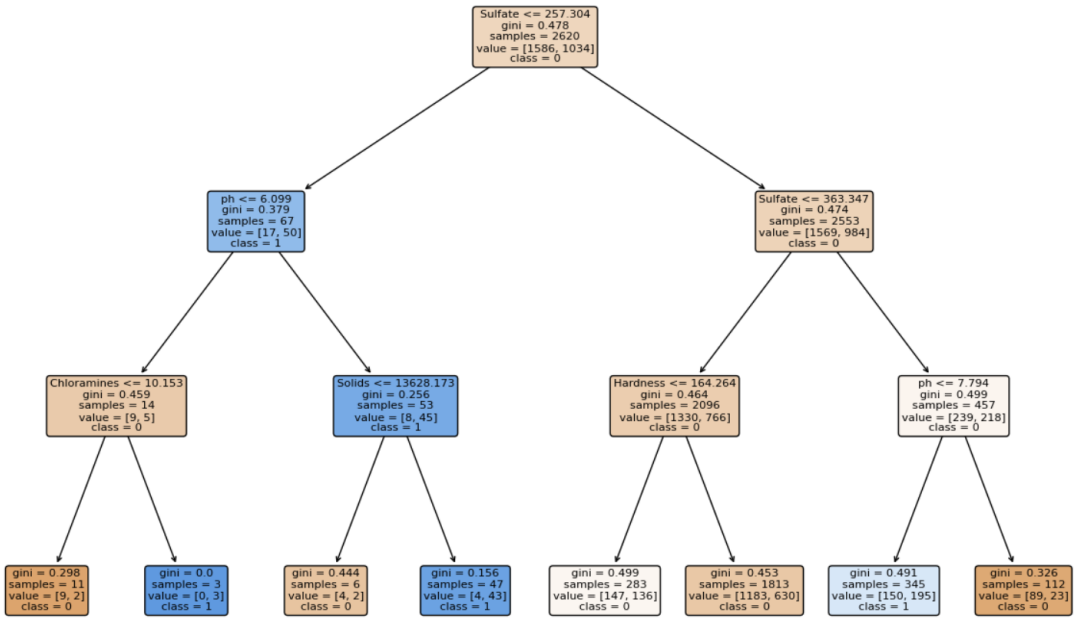


Figure 6. ROC Curve for Random Forest

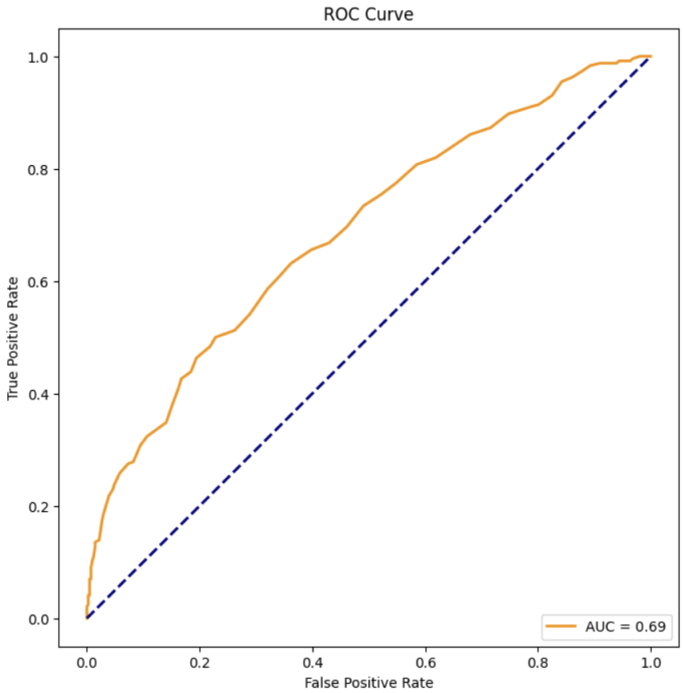


Figure 7. Feature Importances for Random Forest

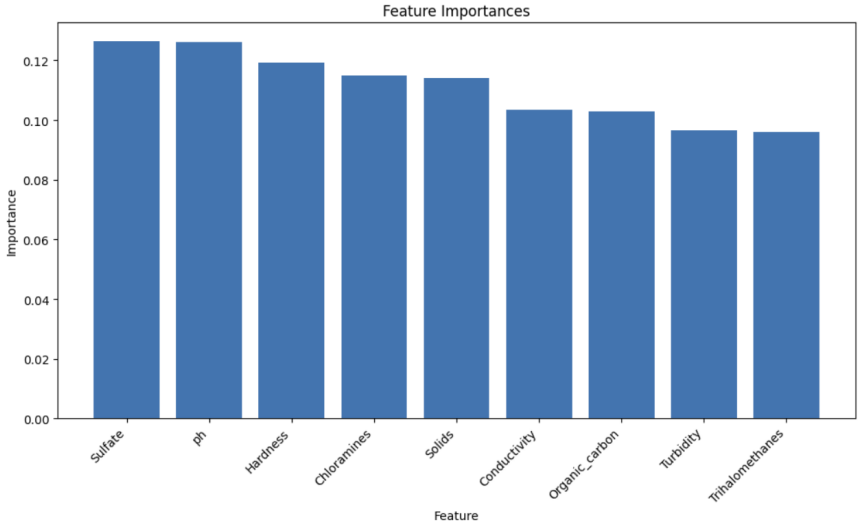


Figure 8. Overall Accuracy of Models Trained and Tested

Model	Accuracy
Random Forest	67.38%
Decision Tree *(with max depth of 3)*	63.87%
Linear Regression	62.80%
Decision Tree *(without constraints)*	57.62%

5. Discussion

Based on the low accuracy results of the logistic regression model, it could be claimed that the relationship between the input features and target variable is nonlinear. Thus, the logistic regression fails to capture and model the complex patterns found within the dataset. This can be clearly seen within the ROC curve shown in Figure 3. Figure 3 shows that the logistic regression model has an almost linear ROC curve, which suggests that the model's performance is not much better than random chance. Thus, it can be concluded that the model is not creating meaningful distinctions between the true positive rate and true negative rate. Therefore, the logistic regression model is essentially random in its predictions on whether a body of water could be classified as potable or non-potable.

During modeling the dataset with decision trees, the initial decision tree was allowed to grow without any constraints leading to an overly complex tree that captured noise. This resulted in a low accuracy of 57.62% due to the overly complicated decision tree that doesn't generalize the dataset correctly. When maximizing the depth of the decision tree to three, it increased the accuracy by 6.25% to an overall accuracy of 63.87%. Thus, it can be claimed the maximum depth led to the decision tree model to capture

the complex relationships of the dataset while reducing bias and correctly modeling the nonlinearity of the dataset.

Since a random forest is less prone to overfitting, it can deal with a noisy or complex dataset more efficiently. During the process of modeling, it was attempted to constraint the model with a maximum depth like the decision trees. However, it led to a decrease in accuracy than a random forest created with no constraints. This could be because the depth of the random forest is limited, the random forest regressor is forced to create weaker and less complex trees. Thus, the accuracy of random forest with maximum depth constraints decreased when compared to random forest that is allowed to grow without constraints. Based on the ROC curve, the random forest model performs the best out of all the models tested within this project. Since the ROC curve is concave slightly upward above the random classifier line, it indicates that the model is performing better than random chance but the model does not achieve high levels of discrimination. The shape suggests that the true positive rate (sensitivity) is increasing at a slightly faster rate than the false positive rate (1 - specificity), leading to the upward curve. Yet, the random forest model worked the best for this dataset. Calculating the feature importances of water quality prediction shows that the top three features of importance are sulfate concentration, pH level and hardness of the water. Although the importance of features is roughly similar based on Figure 7, it is commonly accepted that pH is the most important feature in water quality.

It is important to note that this dataset contains information on water bodies from different locations. This could've affected the accuracy of the models since it is possible that the importance of certain input features are weighted differently in different areas. For example, in one area the most factor in accessing water quality could be the pH while in another area it could be the turbidity of the water. Majority of water quality sampling comes from scientists visiting water sources and taking a sample of the vertical water column to measure the variables included in this model. Forecasting water quality with instruments to detect the important features in water quality could quicken the analysis that scientists need to focus on. This project was aimed to encourage scientists and researchers to invest their time in creating a water quality model that could quickly gather information on water quality. This information could be used by the public in accessing if their water sources are safe to use and drink from, which lessens the risk of public health within low-income areas.

6. Conclusion

From this study, the following conclusions can be made:

- Random Forest worked the best to train the model.
- The three most important features are sulfates, pH and hardness within the Random Forest model.
- The ROC curve of Random Forest indicates that the model may not be achieving high levels of discrimination between positive and negative classes.
- The Decision Tree model worked better with a maximum depth.
- Logistic Regression model failed since the relationship between input features and target variable is nonlinear.