Water Quality

Phuc Hoang Pham
Data Science
Wentworth Institute of Technology
Boston,MA
phamp8@wit.edu

ABSTRACT

Water quality is a critical issue that affects people around the world. We rely on water every day — for drinking, cooking, bathing, and cleaning — yet many of us may not know whether the water we use is truly safe. Understanding and assessing water safety is essential to protecting our health and well-being.

KEYWORDS

- Water Quality
- Machine Learning
- Safety

1 Introduction

Water quality plays a vital role in human health, as clean and safe water is essential for daily activities such as drinking, cooking, bathing, and cleaning. However, in many regions around the world, water contamination remains a serious concern, exposing millions to harmful pollutants and waterborne diseases. To better understand and predict water safety, data-driven approaches have become increasingly important.

2 Data

The Water Potability Dataset, sourced from Kaggle, provides physicochemical measurements of water samples — including pH, hardness, total dissolved solids, chloramines, sulfate, conductivity, organic carbon, trihalomethanes, and turbidity — along with a label indicating whether each sample is potable.

2.1 Source of dataset

Source: https://www.kaggle.com/datasets/adityakadiwal/water-potability

Source From Kaggle, It was generated by **Aditya Kadiwal** 5 years ago

2.2 Characters of the datasets

The file contains water quality metrics for 3267 different water bodies.

Columns:

1. PH Value: PH is an important parameter in evaluating the acid–base balance of water. WHO has recommended maximum permissible limit of pH from 6.5 to 8.5.

2. Hardness: Hardness is mainly caused by calcium and magnesium salts

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. 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. 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.

3 Methodology

The methodology will involve data cleaning and analysis using visualization to check which feature is most important. Then, several Predictive Models will be trained and compared for their effectiveness in classifying potable versus non-potable water, using specialized techniques to handle data

3.1 Linear Regression

In this step, we use the features that showed the strongest absolute correlation in the initial analysis to train the Linear Regression model for the binary outcome (Potable/Non-Potable). Then Using result to compare with other model

3.2 Logistic Regression

In this step, we use the features that showed the strongest absolute correlation in the initial analysis to train the Logistic Regression model for the binary outcome (Potable/Non-Potable). Then Using results to compare with other model. This model could be the same as Linear regression, but will it be?

3.3 Random Forest Classification

This section outlines the process for using the **Random Forest Classifier** to predict water potability. Unlike linear models, the Random Forest excels at capturing the complex, non-linear relationships revealed by your analysis and are implemented with a specific technique to handle data imbalance. Analysis to train the Logistic Regression model for the binary outcome (Potable/Non-Potable). Then Using results to compare with another model

4 Results and Discussion

Both Linear Regression and Logistic Regression achieved an identical overall accuracy of 61% This figure is misleading because both models failed to predict any potable water 0% Recall). This happened because of the data imbalance.

The Random Forest Classifier was the best model, achieving 66 %accuracy. More importantly, it successfully identified potable water, raising its Recall for the Potable class to 28% and Precision to 65%

6 Conclusion

For this water quality data, the random forest is best model overall.

REFERENCES

[1] Dataset Source Aditya Kadiwal (n.d.). *Water Potability* [Data set]. Kaggle. Retrieved from https://www.kaggle.com/datasets/adityakadiwal/water-potability/data

[2] Random Forest Algorithm L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[3] Logistic Regression and Linear Models J. Faraway, *Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models*. Chapman and Hall/CRC, 2016.

[4] Machine Learning Framework F. Pedregosa *et al.*, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.