**Water Quality Classification**

**CIND820: Capstone Project**

Project by: Ann Sam

ann.sam@ryerson.ca

Student #501160843

Supervisor: Dr. Ceni Baboglu

Date of submission: June 6th, 2022

Table of Contents

1. [Abstract 3](#_Toc105148290)
2. [Introduction 4](#_Toc105148291)
3. [Literature Review 5](#_Toc105148292)
4. [Data Description 8](#_Toc105148293)
5. [Exploratory Data Analysis 12](#_Toc105148294)
6. [Approach 13](#_Toc105148295)
7. [Results 14](#_Toc105148296)
8. [Conclusions 14](#_Toc105148297)
9. [References 15](#_Toc105148298)
10. [Terminology 16](#_Toc105148299)

# Abstract

There is nothing more important than safe drinking water. It is essential for the well-being of all human life but unfortunately, ensuring access to safe and clean drinking water can still be challenging, even here in Canada. Water quality has conventionally been tested through expensive and time-consuming laboratory analyses and these analyses can vary depending on the number of parameters being tested. Can the implementation of a supervised machine-learning model be used to determine if the water the safe for residents to drink?

In this capstone project, the goal is to first identify significant parameters required to predict potable water and then to use those parameters to explore a series of supervised machine learning algorithms to classify water potability as safe (potable) or unsafe (non-potable). Using the proposed Water Quality dataset that contains nine parameters that will be used as predictor variables to determine the class variable, potability (Kadiwal, 2021). This research will explore which of those parameters have the highest correlation with potability and test them against several machine learning algorithms to compare their performance in classifying safe or unsafe water. Using known techniques in Python, six classification models will be used: logistic regression, k-nearest neighbor regression, decision tree classifier, random forest classifier, Principle Component Analysis (PCA) algorithm, and XGBoost algorithm.

*\*\* Will add conclusions and results from modeling \*\**

# Introduction

Water potability is defined as water that is clean and safe for drinking. Due to many factors such as geography and the remoteness of the reserves, chronic underfunding that leads to faulty treatment facilities, and past government policies, residents may not trust the water supply fearing elevated levels of heavy metals or contaminants like E. coli. The consequences of consuming non-potable water can vary depending on the levels of dangerous contaminants or pathogens in the water and can cause long term health effects. According to the Government of Canada, there are many First Nations communities that currently do not have access to safe drinking water. The most recent update from the Government of Canada reports that 132 long-term drinking water advisories have been lifted since 2015, but there are still 33 long-term drinking water advisories in effect in 28 communities as of April 25, 2022 (Government of Canada, n.d.).

As a private citizen in the province of Ontario, the turnaround time for a water sample submitted to a Public Health Ontario Laboratory is 4-days, with samples only accepted Monday to Friday (Public Health Ontario, n.d.). Excluding the cost for one sample kit and laboratory report, simple water assessments of water quality can be time consuming, financially straining, and a reasonable amount of effort is required to collect samples and ensure they are properly shipped and processed. Water is an essential requirement for all life on Earth but more importantly, access to clean water and safe sanitation is a basic human right. The average citizen should not have to pay out of pocket to ensure clean drinking water for themselves when Canada is considered a water-rich country, with an estimated 7% of the world’s renewable freshwater supply (Government of Canada, 2015). The impact to society and its regional economy is much more detrimental than the invested cost of ensuring proper quality control, which is why water management is such an integral part of human livelihood all over the world. The approach of utilizing machine learning algorithms with reasonable accuracy to predict water potability could help ensure real-time water quality is available to all.

The data set was retrieved from:

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

The raw data and processes for this study can be accessed from: <https://github.com/annsam0115/CIND820>

# Literature Review

This investigation into methodologies using supervised machine learning will explore potential optimizations in predicting water quality, particularly for potability, water that is suitable for drinking. This study will employ six widely known machine learning methods with the intent to support the idea of employing artificial intelligence to assist in regional water quality analyses.

Previous machine learning studies on water quality have employed multiple supervised and unsupervised machine learning algorithms in an attempt to accurately predict clean water. In one study using the Pakistan Council of Research in Water Resources (PCRWR) dataset (Ahmed et al., 2019), their exploratory data analysis was able to first filter out cumbersome and irrelevant features to clean the data and subset only the most correlated features. Further data processing, such as normalization was used to calculate water quality index (WQI). Their study employed 5 different regression algorithms and 10 classification algorithms. Their methodology showed that with the use of four parameters: temperature, turbidity, pH, and total dissolved solids; they were able to achieve the best results for regression using gradient boosting and polynomial regression with a mean absolute error (MAE) of 1.9642 and 2.7273 respectively. Using the classifier, an accuracy of 85.07% was found most efficient with the use of multi-layer perception (MLP) classification.

Another study on Indian rivers that used machine learning to efficiently predict water quality (Yogalakshmi & Mahalaskhmi, 2021), utilized linear, polynomial, and logistic regression on their dataset across 13 standard water quality parameters. The most interesting issue they discovered from their studies was that water quality records could reasonably be compromised by various water poisons. They recommended to lessen the impact of tainted water, it was fundamental to establish a more sensible water quality parameter collection system, especially for the testing of pH, turbidity, temperature, and TDS parameters. In more controlled collection sites, such as tanks or treatment facilities, data collection is consistent but in rivers or lakes, the ecological state of the water source can drastically impact water quality at any given moment, but may not necessarily represent the water quality as a whole.

In Castillo et al. (2022), their study deployed both classification and regression models that found greater confidence with the use of multiple linear regression among other regression models with a residual square error (RSE) of 3.262 on 15 of the 17 parameters and the use of logistic regression model as a classification model correctly classified 93% for the 17-parameter model used. Curiously, their results lessen in performance when reducing the parameters on the classification model to from 17 to 15 parameters. In comparison of their dataset parameters and the one used in this study, they had more chemical constituents included as parameters and introduced other biological parameters such as fats, oils, and grease, and biological oxygen demand that was not seen in other similar studies that were investigated.

In Ubah et al. (2021) study, Forecasting water quality parameters using artificial neural network for irrigation purposes, they used Artificial Neural Network (ANN) which is similar to linear regression to classify water quality. Neural network algorithms are different to machine learning algorithms as they are developed to mimic the human brain by introducing the concept of bias and threshold to the modeling algorithm. Using four parameters, pH, TDS, electrical conductivity, and sodium, the performance of using ANN algorithm had R-squared values ranging from 0.951 to as high as 0.989. Artificial neural networks contain a number of layers: the input, the hidden, and the output layers. The architecture of these layers and the weight on how they are connected allows for training through feed-forward back-propagation training algorithms. Again, their study identifies that water quality parameters vast and dependent on sampling location and time of year, can vastly influence water sampling results.

Finally, researching beyond machine learning studies related to water quality predictions, a drinking water quality assessment study in Wondo genet campus in Ethiopa (Meride & Ayenw, 2016) conducted water sampling using three physico-chemical parameters and eight chemical constituent parameters to determine water drinkability but also revealed that additional testing of coliforms is necessary in conjunction with other indicators. Coliforms are bacteria from animals and are found in their wastes but can also be found in plant and soil material. While indications of coliforms in water may not cause disease, one of the major species of fecal coliform in Escherichia coli, better known as E. coli which can cause serious illness. While all the previously mentioned studies regarding water quality prediction use similar physico-chemical parameters such as: turbidity, total dissolved solids, and electrical conductivity; and chemical constituent parameters such as: pH, sulfates, and chlorides; testing for coliforms is an important factor to consider whilst establishing appropriate water testing and predicting mechanisms.

The above-mentioned studies are varied when it comes to the parameters used within their own data set but they are similar in attempting to develop a robust classifier for water quality. In all of the previously investigated studies, at minimum, three of the parameters were used from the potability data set used in this project. Further development and research into hyper-tuning algorithms would also lend to the optimization of water quality classification in future research into efficient and effective water quality predictions.

# Data Description

The Water Quality Dataset (water\_potatbility.csv) is presented as a comma-separated values format.  The dataset contains water quality metrics for 3276 different water bodies. There are nine categorical attributes to be used to predictor attributes and one class attribute, Potability. The classification attribute is represented in a binary format where 1 represents potability and 0 represents not potable. Within the dataset, 1998 records are classified as potable, while 3276 records are deemed not potable. The data was imported into Python 3.7 and a description of each attribute is provided below and a summary of the data in Table 1.

**1. pH value:**

pH is an important parameter in evaluating the acid–base balance of water. pH is measured between 0 to 14. The lower the value of pH, the more acidic and the higher the pH, the more alkaline the water condition. The WHO has recommended maximum permissible limit of pH from 6.5 to 8.5 (World Health Organization, 2022).

### 2. Hardness:

This capacity of water to precipitate soap in milligrams per litre (mg/L). 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.

### 3. Solids (Total dissolved solids - TDS):

The total dissolved solids in ppm (parts per million). 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 produce unwanted taste and diluted color in the appearance of water. Water with high TDS values indicate that water is highly mineralized. The desirable limit for TDS is 500 mg/L with a maximum limit of 1000 mg/L which prescribed for drinking purpose.

### 4. Chloramines:

The measure of chloramines in ppm. 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 mg/L or 4 ppm are considered safe in drinking water.

### 5. Sulfate:

The measure of sulfates dissolved in mg/L. Sulfates are naturally occurring substances that are found in minerals, soil, and rocks. They are present in ambient air, groundwater, plants, and food. Sulfate concentration in seawater is about 2,700 mg/L. It ranges from 3 to 30 mg/L in most freshwater supplies, although much higher concentrations (1000 mg/L) can be found in some geographic locations.

### 6. Conductivity:

Electrical conductivity of water in micro-siemens per centimetre (μ∙S/cm). Pure water is not a good conductor of electric current but it’s a good insulator. Increase in ion concentration enhances the electrical conductivity of water. Generally, the number 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 values should not exceed 400 μ∙S/cm.

### 7. Organic Carbon:

The amount of organic carbon is measured in ppm. 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/L in source water which is use for treatment.

### 8. Trihalomethanes:

The amount of trihalomethanes (THMs) in micro-grams per litre (µ∙g/L). 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 µ∙g/L is considered safe in drinking water.

### 9. Turbidity:

The measure of light emitting property of the water in nephelometric turbidity units (NTU). The turbidity of water depends on the quantity of solid matter present in the suspended state. It is also the test is used to indicate the quality of waste discharge with respect to colloidal matter. The WHO recommends value of 5.00 NTU.

### 10. Potability:

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

**Attribute Summary:**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Attributes** | **Type** | **Min** | **Max** | **Mean** | **Standard Deviation** | **Distinct Values** | **Missing Values** |
| 1 | pH | quantitative | 0.00 | 14.00 | 7.08 | 1.59 | 2785 | 491 |
| 2 | Harness | quantitative | 47.43 | 323.12 | 196.37 | 32.88 | 3276 | 0 |
| 3 | Solids | quantitative | 320.94 | 61227.20 | 22014.09 | 8768.57 | 3276 | 0 |
| 4 | Chloramines | quantitative | 0.35 | 13.13 | 7.12 | 15.8 | 3276 | 0 |
| 5 | Sulfate | quantitative | 129.00 | 481.03 | 333.78 | 41.42 | 2495 | 781 |
| 6 | Conductivity | quantitative | 181.48 | 753.34 | 426.21 | 80.82 | 3276 | 0 |
| 7 | Organic Carbon | quantitative | 2.20 | 28.30 | 14.28 | 3.31 | 3276 | 0 |
| 8 | Trihalomethanes | quantitative | 0.74 | 124.00 | 66.40 | 16.18 | 3114 | 162 |
| 9 | Turbidity | quantitative | 1.45 | 6.74 | 3.97 | 0.78 | 3276 | 0 |
| 10 | Potability | nominal | - | - | - | - | 2 | 0 |

Table 1. Attribute summary of water potability dataset

# Exploratory Data Analysis

# Approach

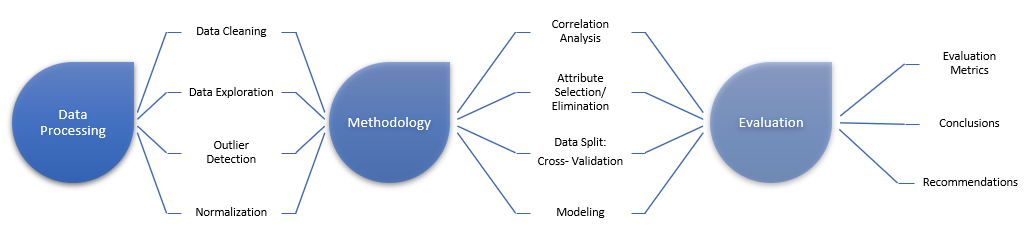


Figure 1. Approach Process

The approached used in this study involved three major stages. First stage is the data processing where the initial analytics of the raw data is performed. This process includes data cleaning, transforming and source data correction or detection of inconsistency within the records. Once the raw data is processed, it can be transformed into a data frame that can be imported and managed in Python 3.6 where the data exploration analysis (EDA) can start. The goal of EDA is to discover useful information and framing the next stage of the methodology.

Once the data has been initially processed, we can begin checking assumptions and testing those assumptions using feature correlation analyses. This process will be the foundation of how the final data set will look like after any attribute elimination. From this point, the subsetted dataset will be used for data training and run against the seven machine learning models. As the models are run, we will start to get a picture of what conclusions we can infer and support any decision making during this process to re-evaluate any previous steps.

Finally, evaluating the modeling metrics to compare performance. The model that performed the best and had the most accurate outcome in determining water potability.

# Results

# Conclusions

# References

1. Kadiwal, A. (2021). Water Quality Dataset Version 3*.* Retrieved from https://www.kaggle.com/datasets/adityakadiwal/water-potability/
2. Government of Canada. Ending long-term drinking water advisories. Retrieved from https://www.sac-isc.gc.ca/eng/1506514143353/1533317130660
3. Public Health Ontario. Drinking Water Testing – Private Citizen*.* Retrieved from https://www.publichealthontario.ca/en/laboratory-services/test-information-index/drinking-water-testing-private-citizen
4. Government of Canada. (2015) Maintaining water quality and availability. Retrieved from https://www.canada.ca/en/environment-climate-change/services/archive/sustainable-development/2015-progress-report/water-quality-availability.html
5. World Health Organization. (2022). Drinking-water quality guidelines. Retrieved from https://www.who.int/publications/i/item/9789240045064
6. Ahmed, U., Mumtaz, R., Anwar, H., Shah, A. A., Ifran, R., & Garcia-Nieto, J. (2019). *Efficient Water Quality Prediction Using Supervised Machine Learning*. www.mdpi.com/journal/water. Retrieved May 10, 2022.
7. Yogalakshmi, S. & Mahalakshmi, A. (2021). *Efficient Water Quality Prediction for Indian Reivers Using Machine Learning*. www.ajast.net. Retrieved May 31, 2022.
8. Sakizadeh, M. *Artificial intelligence for the prediction of water quality index in groundwater systems*. (2016). https://link.springer.com. Retrieved May 10, 2022.
9. Fernandez del Castillo, A., Yebra-Montes, C., Garibay, M. V., de Anda, J., Garcia-Gonzalez, A. & Gradilla-Hernandez, M. S. (2022). *Simple Prediction of an Ecosystem-Specific Water Quality Index and Water Quality Classification of a Highly Polluted River through Supervised Machine Learning*. www.mdpi.com/journal/water. Retrieved May 22, 2022.
10. Ubah, J. I., Orakwe, L. C., Ogbu, K. N., Awu, J. I., Ahaneku, I. E. & Chukwuma, E. C. (2021). *Forecasting water quality parameters using artificial neural network for irrigation purposes*. www.nature.com/scientificreports. Retrieved May 31, 2022.
11. Tan, G., Yan, J., Gao, C. & Yang, S. (2012). *Prediction of water quality time series data based on least squares support vector machine*. www.elsevier.com/locate/procedia. Retrieved May 31, 2022.
12. Meride, Y. & Ayenew, B. (2016). *Drinking water quality assessment and its effects on residents’ health in Wondo genet campus, Ethiopia*. https://environmentalsystemsresearch.springeropen.com. Retrieved May 31, 2022.

# Terminology