**Water Quality Classification**

**CIND820: Capstone Project**

Project by: Ann Sam

ann.sam@ryerson.ca

Student #501160843

Supervisor:

Date of submission: May 16th, 2022

**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. 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[[1]](#footnote-1). 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. 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 quality as safe (potable) or unsafe (non-potable). Using the proposed Water Quality dataset[[2]](#footnote-2) that contains nine parameters that will be used as predictor variables to determine the class variable, potability. The dataset also contains water quality metrics from 3276 different water bodies which 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 algorithm, and XGBoost algorithm.

This project will aim determine weather specific water quality parameters can yield more correlated potability results, if using a machine learning approach can reduce tedious and time-consuming water sampling analyses, and of those algorithms, which can execute the most reliable outcome. In assessing the results from the models and comparing their respective evaluation metrics, the intent will be to understand the significance of the water quality parameters of the dataset and further develop a quicker and more robust classifier for those who maybe at risk for being under drinking water advisories.

1. *Ending long-term drinking water advisories.* Government of Canada. https://www.sac-isc.gc.ca/eng/1506514143353/1533317130660 [↑](#footnote-ref-1)
2. Kadiwal, A. *Water Quality Dataset Version 3.* Kaggle. https://www.kaggle.com/datasets/adityakadiwal/water-potability [↑](#footnote-ref-2)