Water Quality Classification

CIND820: Capstone Project

Project by: Ann Sam
ann.sam@ryerson.ca
Student #501160843

Supervisor: Dr. Ceni Baboglu
Date of submission: July 25th, 2022

Ryerson University

Table of Contents

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 determine if we can effectively and accurate predict water quality. In doing so, we'll need to identify any 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, we will conduct a series of modeling on those predictor variables to determine the class variable, potability. 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. Six known predictive modeling techniques in Python were used: logistic regression, k-nearest neighbor regression, decision tree classifier, random forest classifier, support vector model classifier and XGBoost algorithm. Results from the modeling found Support Vector Machine Classifier had the best performance with an accuracy of 83.81% when modeled using train-test split and 87.98% using K-Fold cross-validation.

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 Ethiopia (Meride & Ayenw, 2016) conducted water sampling using three physio-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 physio-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 (Kadiwal, 2021). 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

Visual Analysis:

Plotting the distribution of the data within each of the predictor variables show both non-potable and potable records displaying a normal/Gaussian distribution pattern. With the exception of Solids where it is slightly right-skewed, this is an ideal outcome, a bell curve. This tells us that the data distribution is acceptable without having to reject any of the predictor variables and allows us to make decisions during the data cleaning stages.

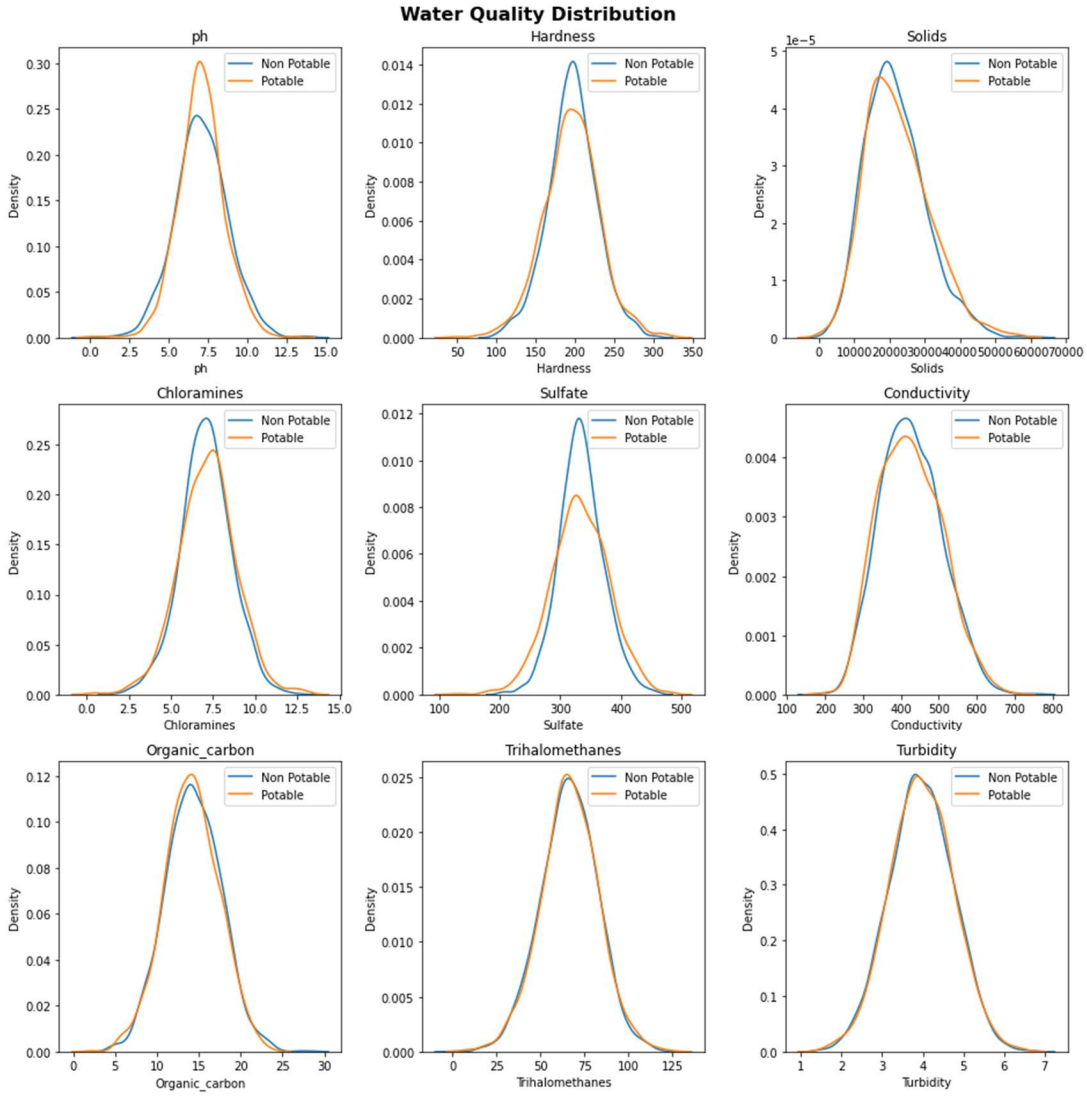


Figure 1. Distribution graph of each predictor variables differentiating between non-potable and potable records.

There are also missing data points within the source dataset in three of the predictor variables: pH, Sulfate, and Trihalomethanes. There were 491 missing values for pH, 781 missing values for Sulfate, and 162 missing values for Trihalomethanes. The percentage of missing values is significant enough that we cannot omit those records from the dataset.

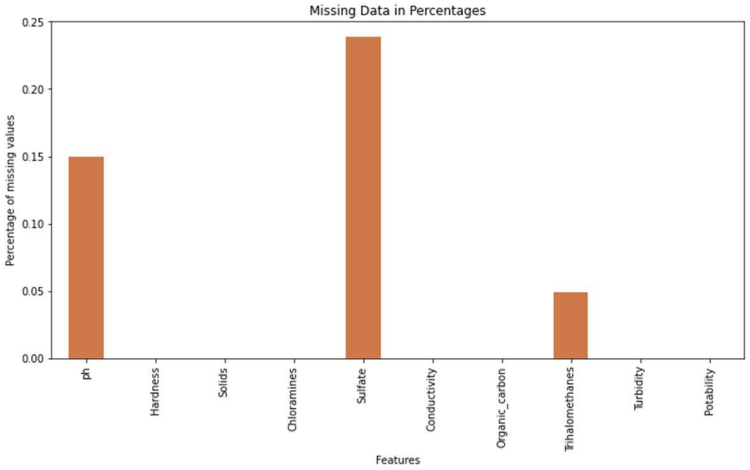


Figure 2. Percentage of missing data points in each predictor variable.

Since we have normal/Gaussian distribution, it was safe to replace all the missing values with the mean value of their corresponding variable.

Using box-plot, we can see the data distribution of the nine predictor variables and clearly identify outlier data points outside minimum and maximum whiskers within all variables. We can also confirm the skewness in Solids as seen in the distribution plots.

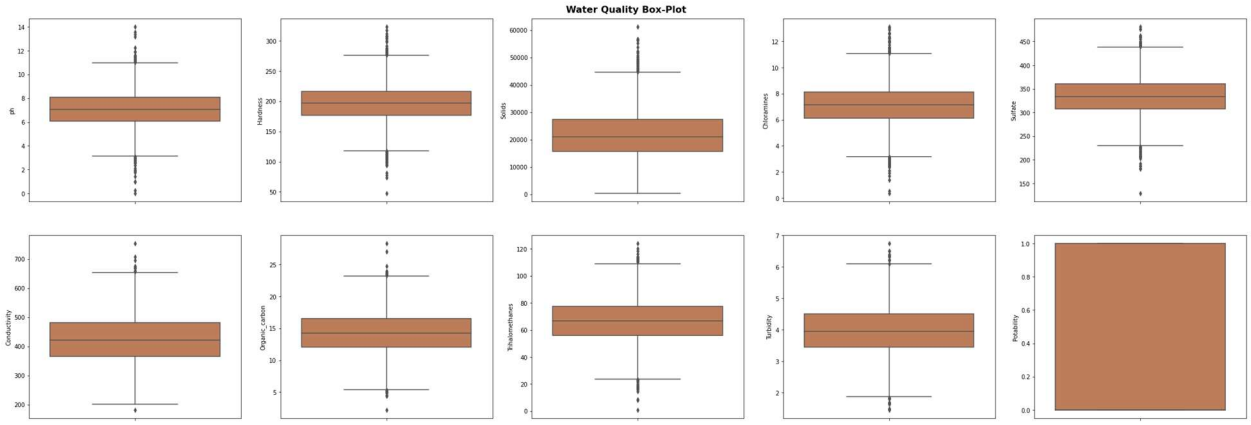


Figure 3. Box-plot of each predictor variable displaying outlier data points.

As mentioned earlier, the distribution of the dataset is mainly Gaussian and thus allows us to remove outlier data outside of three standard deviations without completely removing all the outlier data (outside the max. and min.). This allows us to keep some of the outlier data to maintain the integrity of the variability of the dataset without completely omitting them. The most visual change was again within the Solids variable where we also observed the most skewness.

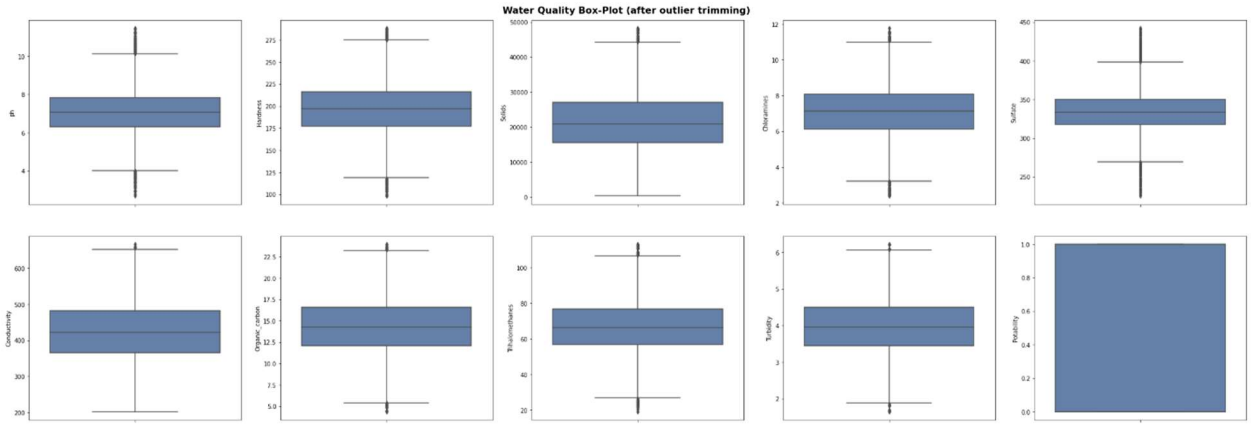


Figure 4. Box-plot of dataset after outliers greater than three standard deviations were removed.

A correlation matrix was run and visualized using a heatmap to determine any obvious correlation between any of the predictor variables and the target variable. Unfortunately, the correlation values were too low to identify any strong relationships. Specifically looking at any correlations with the target variable, the highest correlation was with Solids and Organic Carbons but at values of +0.034 and -0.030 respectively.

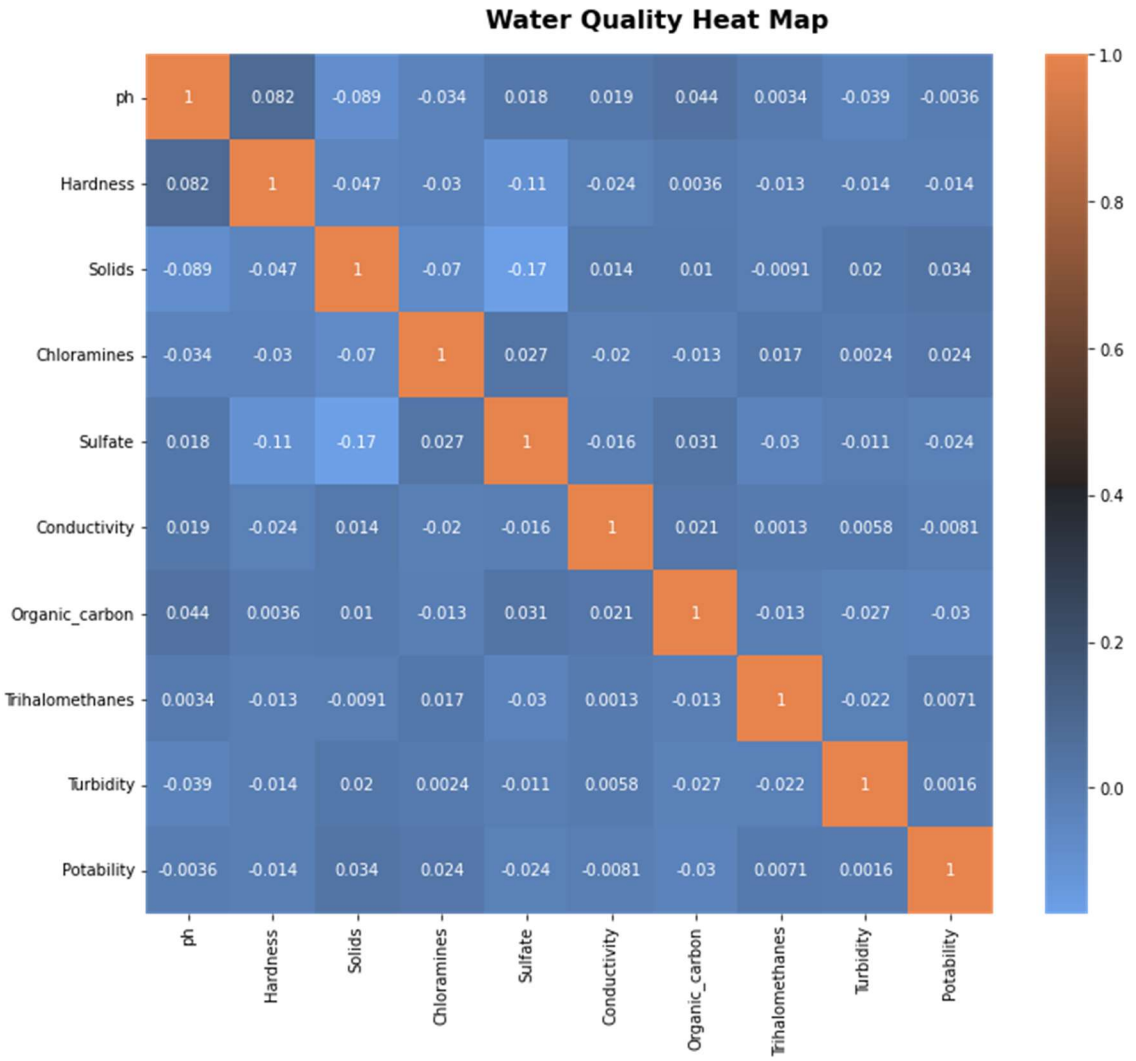


Figure 5. Heatmap displaying correlation values.

Using a pair-plot to also visualize the relationships between the predictor variables, the pairwise relationships did not yield any obvious linear relationships. As such, we can rule out using a simple linear model for classification.

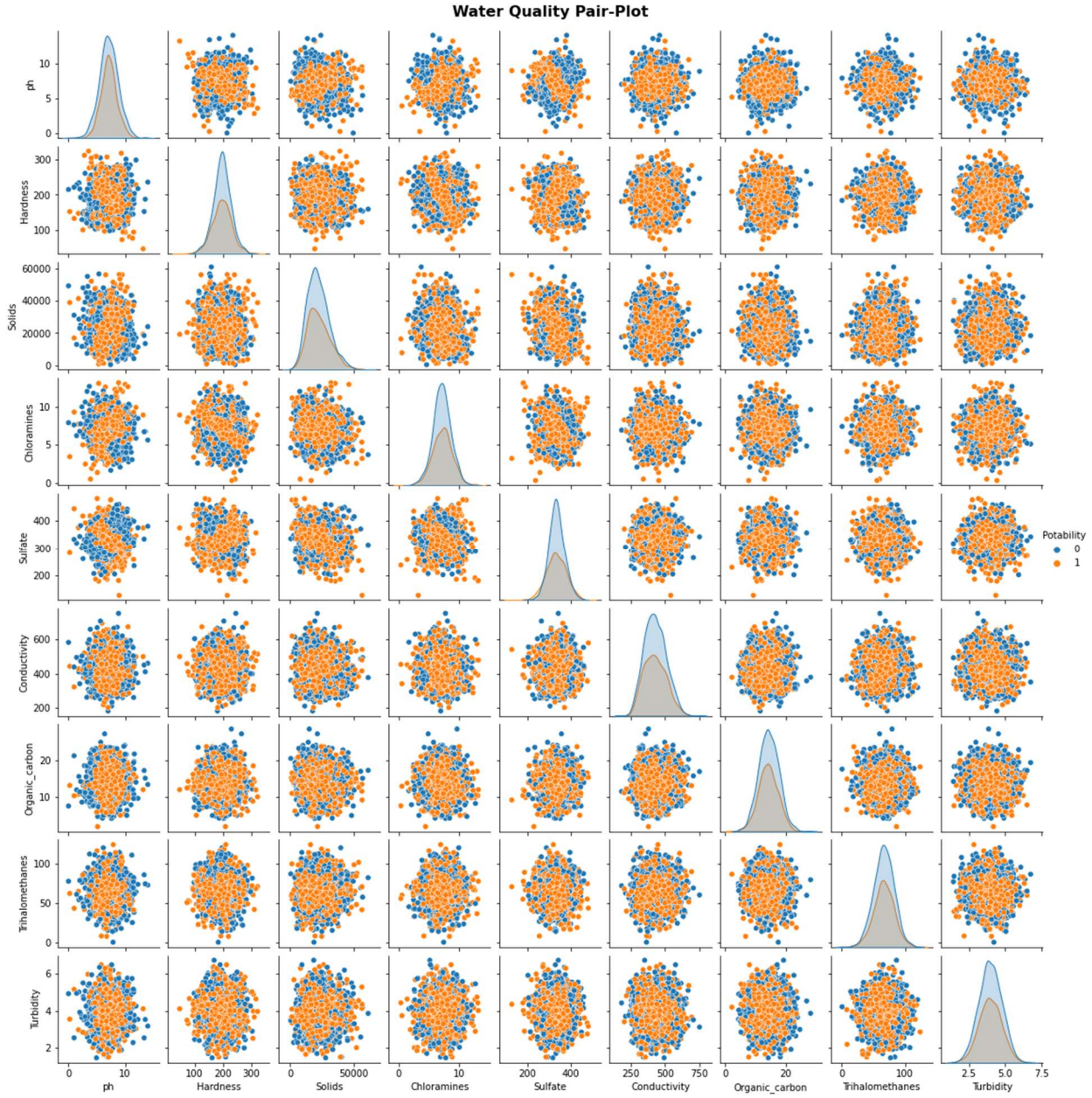
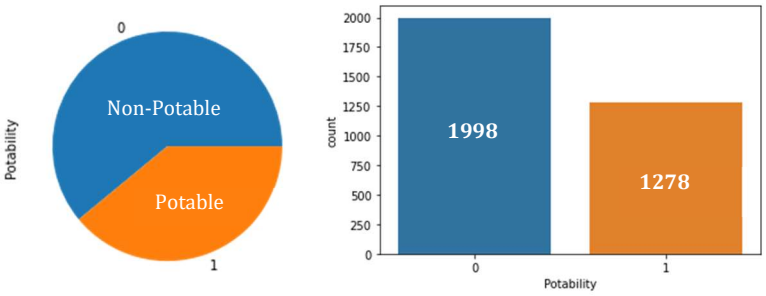


Figure 6. Pair-plot graphs of predictor variables

In the initial analysis of the source data, there is a clear imbalance of records with regards to our target variable, Potability. There are 1998 records that are non-potable, 61% of the dataset. The remaining 1278 records, 39% of the dataset are potable records. This imbalance of distribution will be problematic during model training causing the classification to be biased or skewed towards the majority class, which in this case is non-potable.



Figures 7 & 8. Pie chart of the record counts and bar chart of the original record counts within the dataset.

To handle this imbalance, Python's SKlearn resample method was used to up-sample the minority class to create a balanced dataset. It essentially will over-sample the minority class by duplicating random records to create a balanced data set.

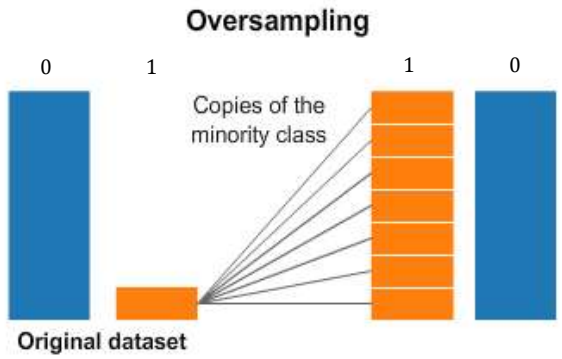


Figure 9. Visual representation of resampling method in SKlearn when dealing with imbalance in the dataset.

After the handling of missing data points and outlier detection through trimming, the total count of records were 1930 non-potable and 1198 potable. After implementing the resampling method, the balanced data set became 1930 non-potable and 1930 potable.

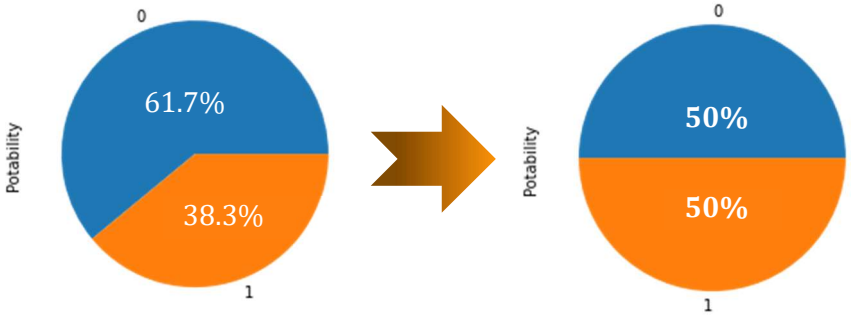


Figure 10. Pie chart of imbalanced class data to balanced class data.

Statistical Variable Analysis:

To confirm our observations from our visual analyses, a statistical analysis was conducted. Using the Analysis of Variance (ANOVA) test, we compared the variation between the sample means and the variation within each of the samples. The results of the ANOVA test produced a p-value as a metric of comparison to determine if there was in fact any statistical correlation between the nine predictor variables and the target variable. Using the standard 0.05 significance as our alpha, we were able to confirm that none of the nine predictor variables had a significant enough correlation with the target variable.

	Predictor Variable	Correlation with Potability	P-Value
0	pH	No	0.739300
1	Hardness	No	0.515201
2	Solids	No	0.243760
3	Chloramines	No	0.388958
4	Sulfate	No	0.812122
5	Conductivity	No	0.672787
6	Organic Carbon	No	0.295948
7	Trihalomethanes	No	0.680139
8	Turbidity	No	0.829477

Table 2. P-value results from ANOVA test.

Principle Component Analysis (PCA) was also used as an unsupervised statistical technique in an attempt to reduce the dimensions of the dataset (i.e., predictor variables). PCA confirmed that all nine predictor variables are independent of each other and that at least eight components (variables) are needed to explain 80% of the variance. Therefore, dimensionality reduction was not beneficial with this dataset.

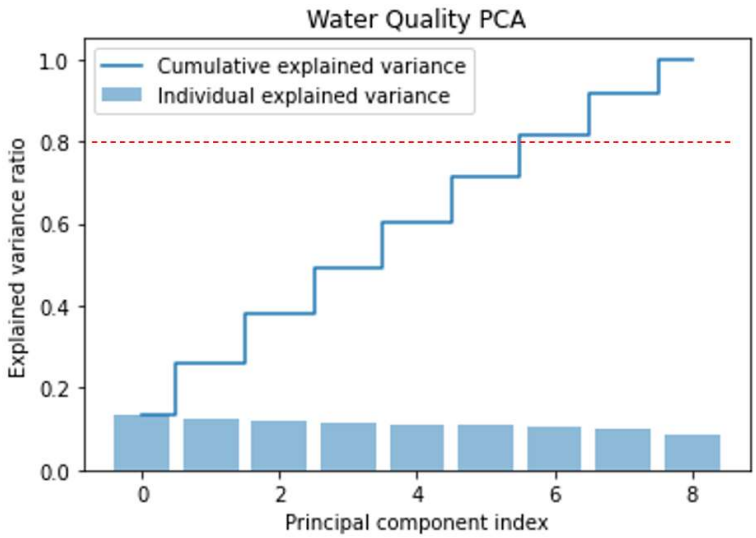


Figure 11. PCA graph depicting component variance ratio and the cumulative variance ratio.

Approach

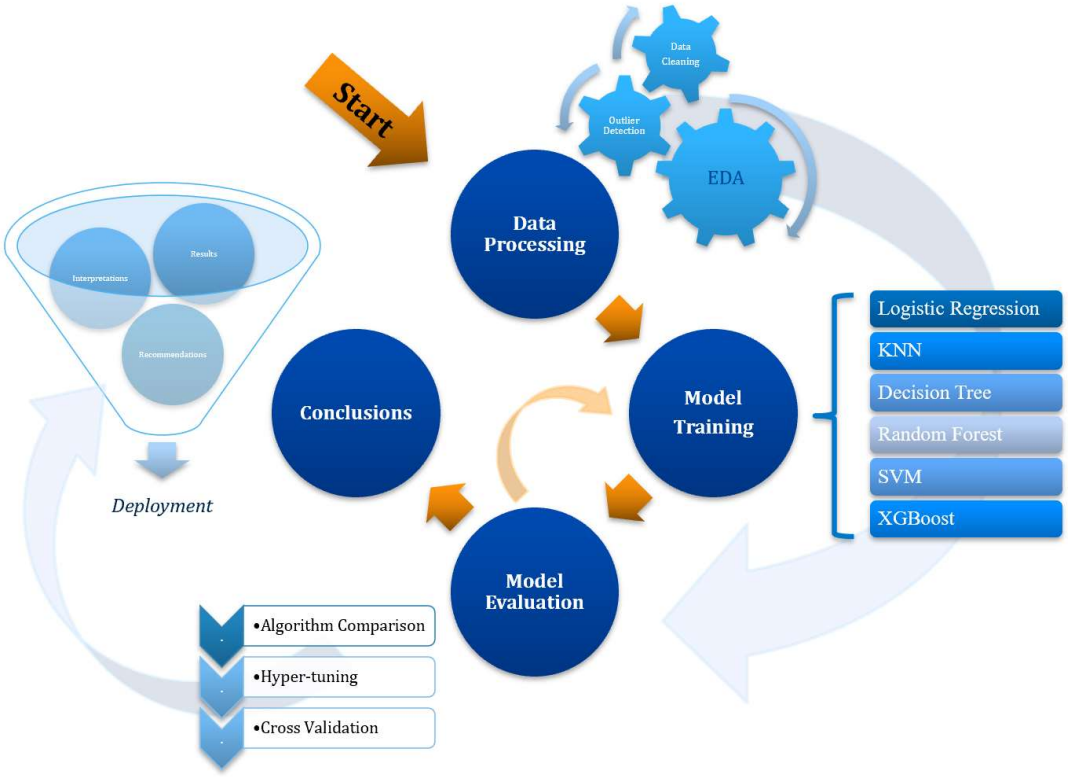


Figure 12. Approach Process

The approached used in this study involved four major stages. First stage is data processing where the initial analytics of the raw data is performed. This process includes data cleaning, transforming and source data correction. and detection of inconsistency within the records. The raw data is imported and managed in Python 3.7 where the data exploration analysis (EDA) started. The goal of EDA was to discover useful information and preparing the technically correct data for preprocessing to become the consistent data that will be used for the rest of the modeling. Once the data is considered consistent data, preprocessing and modeling was conducted using SKlearn package where the dataset is split in a train-test split of 80% to training and 20% to testing. The dataset is then scaled so that the data points fit within an appropriate scale so that higher value ranges do not dominate when data point distances are calculated. The dataset is now ready for modeling.

Six algorithms were used: logistic regression, k-nearest neighbor regression, decision tree classifier, random forest classifier, support vector modeling, and XGBoost algorithm. Two iterations of modeling were conducted, the first with default parameters and the second after hyper-tuning the parameters. After evaluation the models and their evaluation metrics, we further conducted cross validation on the top 3 performing models.

Finally, some post-prediction analyses are discussed to finalize our results and review any recommendations with our final conclusions with the intent to deploy a robust water quality classifier.

Modeling Algorithms

Logistic Regression

Logistic regression is a statistical method similar to linear regression that estimates the probability of a binary event occurring. Based on the given independent variables in a dataset the model predicts the probability a classification will occur using log odd ratios and an iterative maximum likelihood method (Hoffman, 2019).

K-Nearest Neighbour Regression

The k-nearest neighbor algorithm assumes that similar things exist in close proximity. It is a non-parametric classifier which uses similar data points that are close to each other to make classifications. By storing all available cases, it will classify new cases based on similarity measures by pattern recognition.

Decision Tree Classifier

A decision tree classifier is a flowchart-like structure where attributes or variables of the dataset are represented as nodes and the branches from each node represent a decision rule to another node. Eventually, a decision will be reach to a final leaf-node that will be either one outcome or the other, in this case, potable or non-potable.

Random Forest Classifier

Random Forest classifier consists of a large number of individual decision tress that operate as a group or a ‘forest’. Each individual tree will classify and return a prediction and the prediction with the most votes will become the final prediction. The fundamental concept of

random forest is majority wins and is beneficial in datasets where the attributes or predictor variables have low correlation. The idea of having multiple decision trees essentially protects the forest from individual error (Yiu, 2019).

Support Vector Machine Classifier

Support Vector Machine (SVM) is a linear model for classification and creates a line or a hyperplane which separates the data into classes. The hyperplane in an n-dimensional Euclidean space is flat, n-1 dimensional subset of that space that divides it into two disconnected parts (Pupale, 2018).

XGBoost Classifier

XGBoost stands for extreme gradient boosting, and is an implementation of gradient boosted decision trees designed for speed and performance. Gradient boosting is a technique where new models are created to correct the errors made by previous models. The results are combined to make the final prediction (Brownlee, 2016).

Results

First Iteration

In the first iteration of modeling, the algorithms were all run just the default parameters of their respective functions. The confusion matrices of the algorithms are as follows:

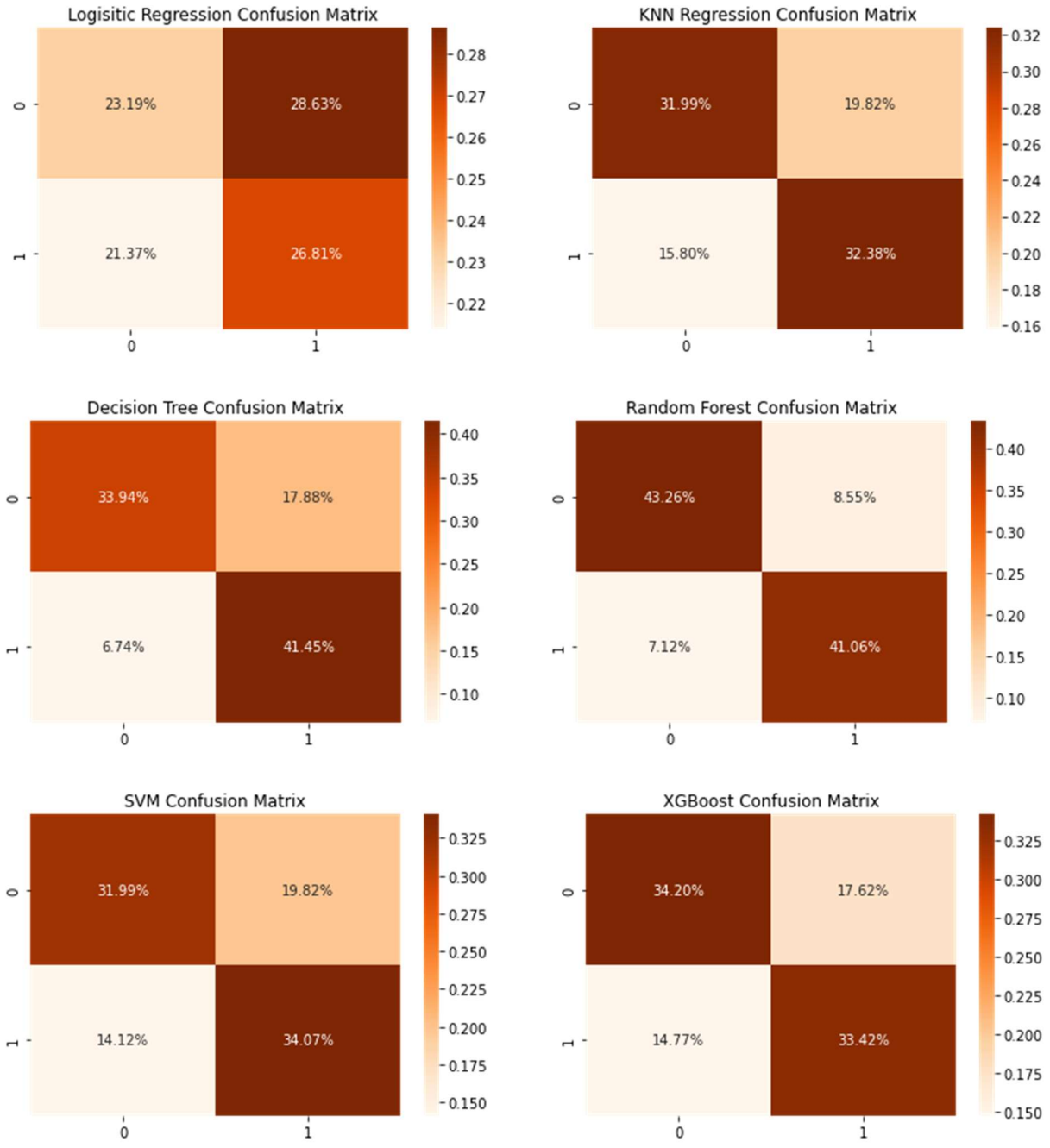


Figure 13. Confusion Matrix of each algorithm after the first iteration of modeling.

Statistically, type II errors would be considered more hazardous in particular for assessing water quality or potability. Returning a classification of a false positive would be detrimental to a community consuming unclean water. In reviewing the confusion matrices, the

highest type II error occurs in Logistic Regression at 21.37% while the lowest type II error occurs in the Decision Tree algorithm at 7.12%.

After reviewing the evaluation metrics of each algorithm for first iteration of modeling, we observed that Random Forest algorithm performed the best with an accuracy of 84.33% while Logistic Regression performed the worst at 50.00%.

	Model	Accuracy	Precision	Recall	F1 Score
3	Random Forest	0.843264	0.827676	0.852151	0.839735
2	Decision Tree	0.753886	0.698690	0.860215	0.771084
5	XGBoost	0.676166	0.654822	0.693548	0.673629
4	Support Vector	0.660622	0.632212	0.706989	0.667513
1	KNN Regression	0.643782	0.620347	0.672043	0.645161
0	Logistic Regression	0.500000	0.483645	0.556452	0.517500

Table 3. Evaluation metrics from the first iteration of modeling.

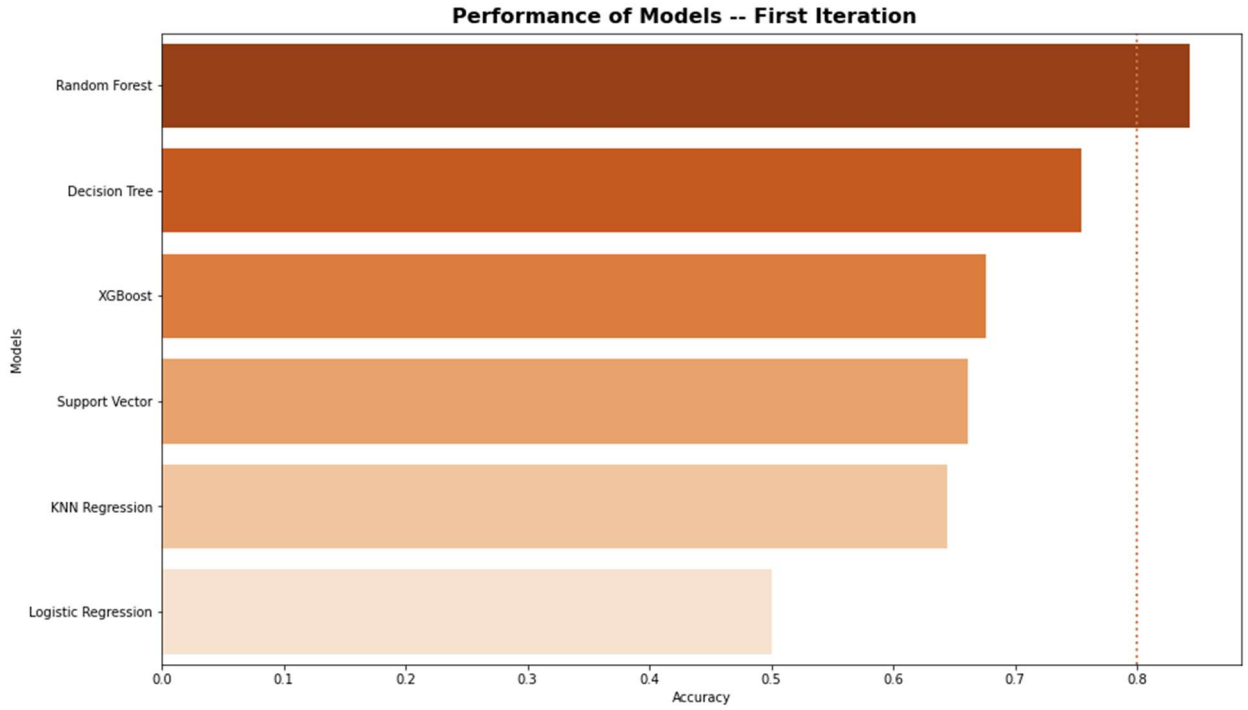


Figure 14. Performance of each algorithm after the first iteration of modeling by accuracy.

Hyper-Tunning Parameters

Using hyperparameter optimization or hyper-tunning the parameters within each algorithm, the goal is to choose a most optimal set of parameters to control the algorithm to yield the best results (i.e., performance). By deploying a Grid Search approach to five of the algorithms and Random Search to just XGBoost, these functions essentially run the algorithm across all the parameters identified and report the best parameters to use. The performance of each hyper-tunning execution is also cross validated on the training set. Hyper-tunning returned the following best parameters and were then used in the second iteration of modeling:

Algorithm	**Hyper-tuned Parameters:**
Logistic Regression	penalty: l1, solver: liblinear
KNN	algorithm: auto, n_neighbors: 1, weights: uniform
Decision Tree	criterion: entropy, max_depth: 44, min_samples_leaf: 1
Random Forest	min_samples_leaf: 2, n_estimators: 200
SVM	C: 10, gamma:1, kernel: rbf
XGBoost	n_estimators: 600, learning_rate: 0.8

Table 4. Hyper-tunning parameters for each algorithm.

Second Iteration

In the second iteration of modeling, the algorithms were all run with the hyper-tuned parameters of their respective functions (listed in Table 4). The confusion matrices of the algorithms are as follows:

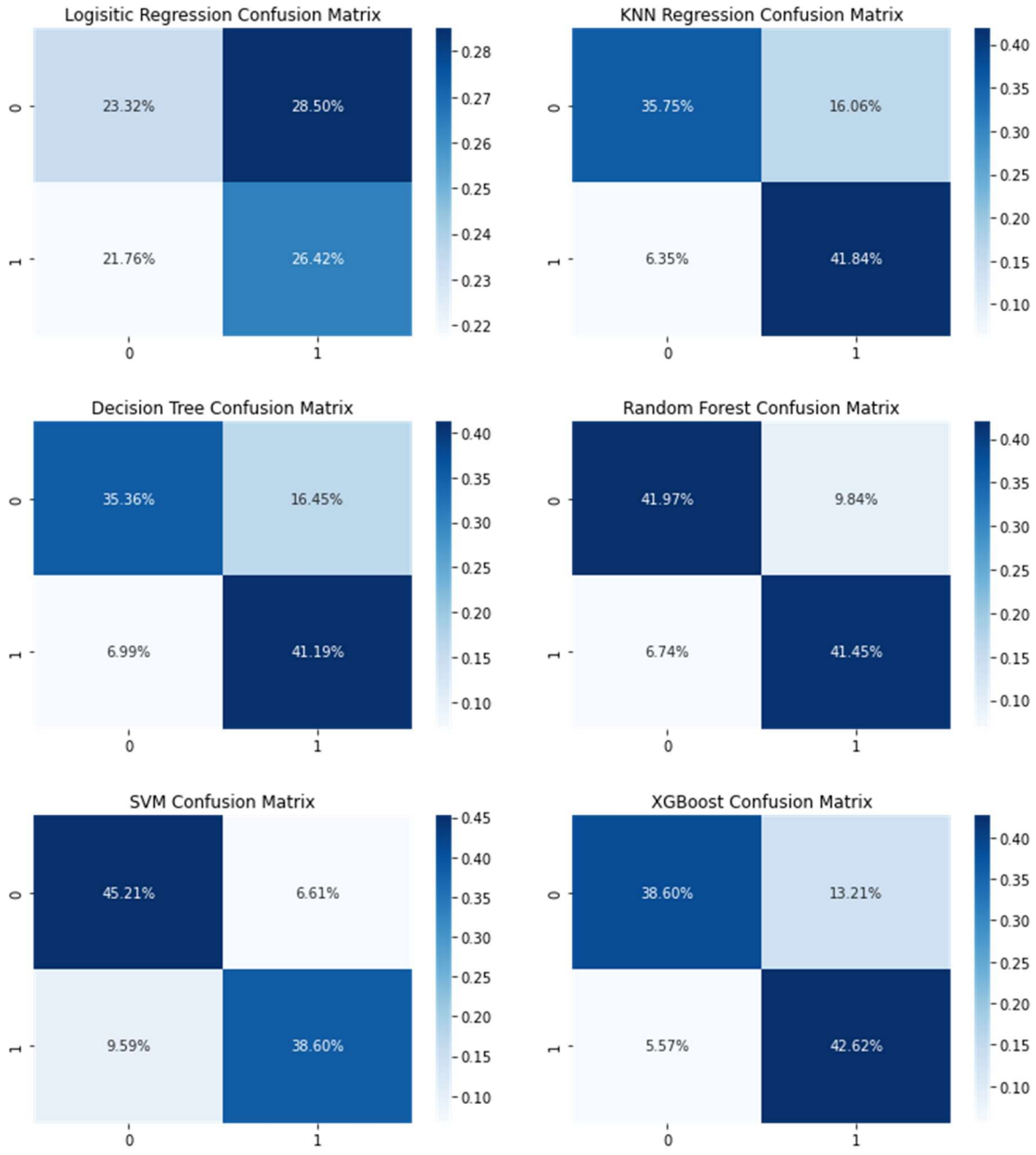


Figure 15. Confusion Matrix of each algorithm after the second iteration of modeling.

As mentioned in the first iteration, type II errors are significant with regards to classifying water potability. After hyper-tunning parameters we can see that Logistic Regression

still has the highest false positive results at 21.76%, while the rest of the algorithms have significant decreases of type II errors after hyper-tunning.

Looking at the evaluation metrics after the second iteration we can see that hyper-tunning greatly increased the performance of nearly all the algorithms. SVM and Random Forest performed the best with the highest accuracy of 83.81% and 83.42% respectively. Logistic Regression remained the lowest performing algorithm with minute changes in overall performance.

	Model	Accuracy	Precision	Recall	F1 Score
4	Support Vector	0.838083	0.853868	0.801075	0.826630
3	Random Forest	0.834197	0.808081	0.860215	0.833333
5	XGBoost	0.812176	0.763341	0.884409	0.819427
1	KNN Regression	0.775907	0.722595	0.868280	0.788767
2	Decision Tree	0.765544	0.714607	0.854839	0.778458
0	Logistic Regression	0.497409	0.481132	0.548387	0.512563

Table 5. Evaluation metrics from the second iteration of modeling.

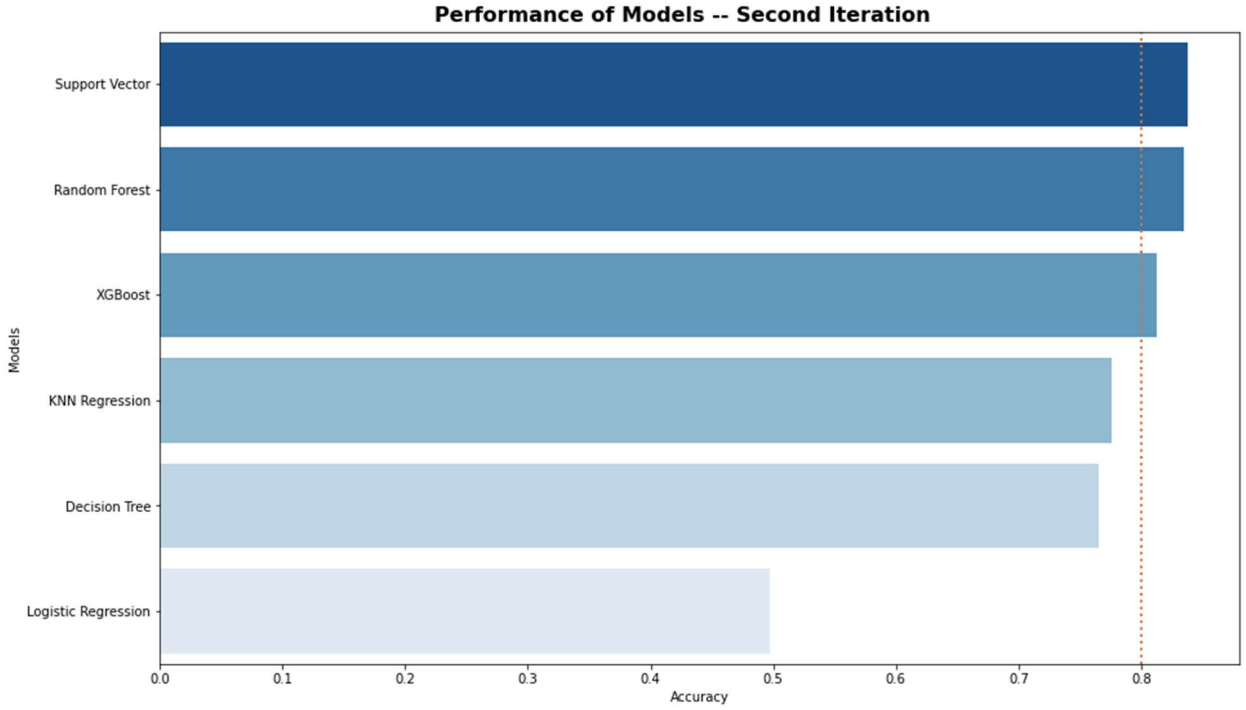


Figure 16. Performance of each algorithm after the second iteration of modeling by accuracy.

A quick comparison between the two iterations of modeling. In fact, there was a 0.26% decrease in accuracy after hyper-tunning for the Logistic Regression algorithm and 0.91% decrease in accuracy for Random Forest. SVM algorithm had the largest increase in accuracy after hyper-tunning, a difference of 17.75 %.

	Model	2nd Iteration	1st Iteration	Difference in Accuracy
0	Logistic Regression	49.74%	50.00%	-0.26%
1	KNN Regression	77.59%	64.38%	13.21%
2	Decision Tree	76.55%	75.39%	1.17%
3	Random Forest	83.42%	84.33%	-0.91%
4	Support Vector	83.81%	66.06%	17.75%
5	XGBoost	81.22%	67.62%	13.60%

Table 6. Comparison of accuracy between first and second iterations of modeling.

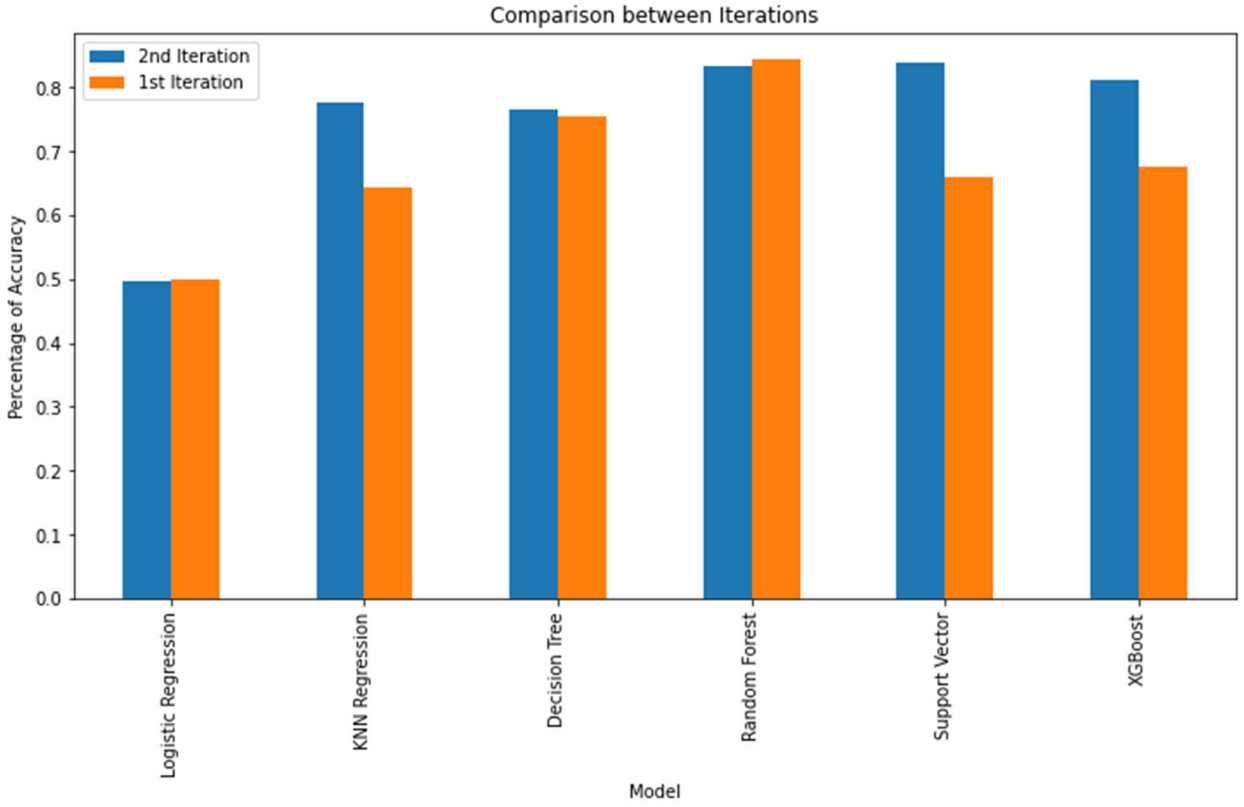


Figure 17. Graph depicting the performance of each algorithm after the first and second iteration of modeling.

Cross Validation

After the second iteration of model training, we selected the top three algorithms to apply cross validation to. Using the K-Fold Cross-Validation method, the consistent dataset (the dataset before train-test split) was used to be split into k number of subsets, where k-1 subsets are used to train the models and the last subset is kept for validation to test the models. The scores of each fold are then averaged to evaluate the overall performance of each model. Cross-validation using 10-folds, where 9 folds were used for training and 1 used for testing, returned higher accuracy results in all three algorithms: Random Forest, SVM, and XGBoost.

Algorithm	Mean Accuracy Score	Standard Deviation
Random Forest	85.28 %	1.84 %
SVM	87.98 %	1.91 %
XGBoost	80.73 %	1.77%

Table 7. Results of K-Fold Cross Validation.

SVM Performance Results

Since SVM has the highest mean accuracy score after cross validation, we returned to the second iteration of model to produce a Receiver Operating Characteristic Curve (ROC Curve) graph to visualize the SVM model's performance with respects to their classification threshold levels. The ROC Curve plots the True Positive Rate (recall) against the False Positive Rate (type II error). We can also calculate the area under the curve (ROC AUC) which will allow us to understand the classifier's performance numerically as a perfect classifier is equal to 1.0. The ROC AUC for SVM after the second iteration was 0.8368 and the cross validated ROC AUC was 0.8674 which is consistent with the rest of our evaluation metrics.

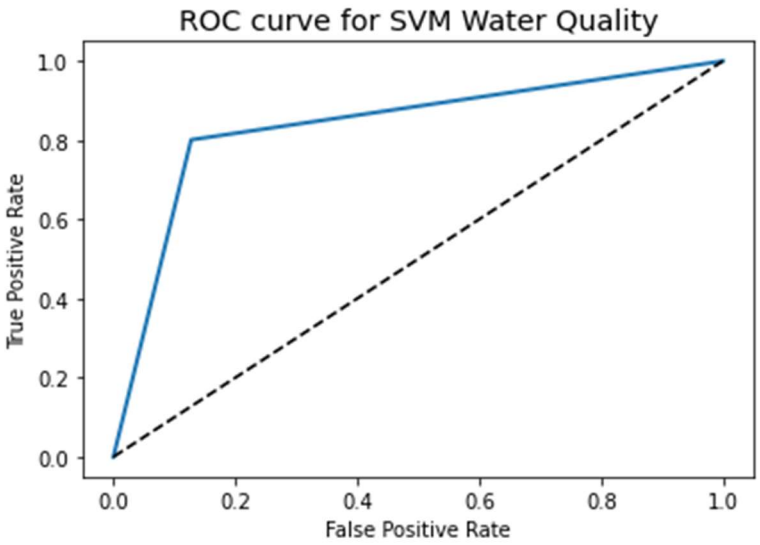


Figure 18. ROC Curve of SVM Model after second iteration.

Codes for the results section can found at: https://github.com/annsam0115/CIND820

Conclusions

Interpretation

The results of the modeling confirm that we can classify water potability using this dataset. After the first iteration, the best performing algorithm was Random Forest with a 84.33% accuracy. SVM was the best performing model after the second iteration with hyper-tunned parameters of C = 10, gamma = 1, and kernel = rbf. This yielded an accuracy of 83.81%. Using cross-validation, SVM mean accuracy rose to 87.98%, which is critical in assessing safe drinking water for communities. In terms of processing time, there were no obvious delays in when executing the models as the dataset was not very large. Thus, in terms of run-time evaluations, there was not enough a difference between the all models to determine a more efficient model in terms of efficiency.

Data preparation was essential in the modeling process. Handling missing values and outliers provided a more comprehensive dataset to model with and increased the overall accuracy. Equally important was dealing with class imbalance to ensure our modeling was not skewed or biased towards the majority class of the original dataset.

Using both the train-test split for our initial modeling and then cross validating the highest performing models gave higher confidence in final selection of the algorithm best suited for water classification. Comparing our final evaluation metrics on SVM's performance, accuracy from the first iteration was 66.06% and increased to 83.81% in the second iteration. ROC AUC values were 0.8368 and cross validated ROC AUC value was 0.8674. Final K-Fold cross validation score was 87.98%.

From literature review of previous studies conducted for water testing as well as some other general water testing research done, this particular dataset lacks, in my opinion, some critical predicting parameters like coliform or bacteria and heavy metals such as lead or copper. These particular attributes are the most noted for at-home water testing kits and for detection of water-borne diseases. These attributes do also have higher correlation with water potability which would have made initial exploratory analysis much easier.

Recommendations

Before deployment of this classifier, it would be beneficial to test against another water quality dataset with the above-mentioned additional attributes and understand what the thresholds of coliforms in our waters or heavy metals influence modeling. In the previous studies on water quality predictions, all mentioned pH, hardness, solids, sulfates, conductivity and turbidity as feature attributes and at least bacteria or coliform. As we know, fecal coliform or E.

coli can have significant consequences to consumers which should be an important predictor variable in all water classification studies. Another recommendation would to have location of water samples indicated such as treatment facility or open waters. We have learned from the literature review that water samples taken where water sources are active can affect sampling as opposed to sampling from more contained systems like treatment facilities. And finally, explore machine learning with Artificial Neural Network (ANN) which had accuracy results in the 90% range in previous studies.

Conclusions

Water classification can be improved using machine learning algorithms and can be accomplished to a high accuracy. Using Support Vector Machine Classifier yielded the best performance overall at 87.98% accuracy after hyper-tunning the algorithm parameters. While we were able to effectively classify water with the predictor variables in the dataset, a few other critical features should be included in the future before deployment such as coliform values and heavy metals as well as explore more advanced deep machine learning algorithms.

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.
13. Hoffman, J. (2019). *Basic Biostatistics for Medical and Biomedical Practitioners (2nd Edition),* https://www.sciencedirect.com/topics/medicine-and-dentistry/logistic-regression-analysis. Retrieved July 7, 2022.
14. Yiu, T. (2019). *Understanding Random Forest: How the Algorithm Works and Why it is so Effective.* https://towardsdatascience.com/understanding-random-forest-58381e0602d2. Retrieved July 7, 2022.
15. Pupale, R. (2018). *Support Vector Machines (SVM) – An Overview.* https://towardsdatascience.com/https-medium-com-pupalerushikesh-svm-f4b42800e989. Retrieved Jul 10, 2022.
16. Brownlee, J. (2016). *A Gentle Introduction to XGBoost for Applied Machine Learning.* https://machinelearningmastery.com/gentle-introduction-xgboost-applied-machine-learning/. Retrieved July 10, 2022.