Water Quality Classification

CIND820:
Big Data Analytics Project

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

Supervisor: Dr. Ceni Baboglu

Spring-Summer 2022

Ryerson University

Why classify water?

- Basic necessity for all human life
- Process of water testing is time consuming: water collection and laboratory testing
- Costly

Can machine learning improve the process of water classification?

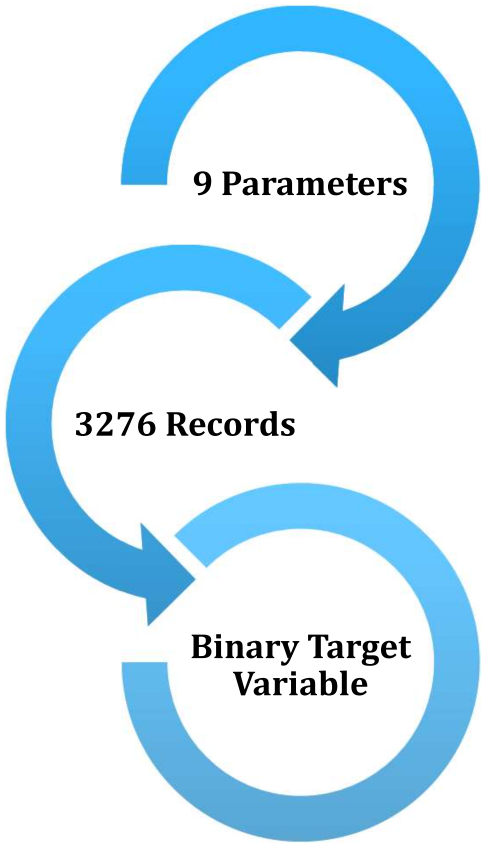
Predicting Water Potability

- Can we predict water potability?
- Which machine learning algorithms can yield the most efficient and accurate results?
- Can the parameters within the ML algorithms be tuned to yield the best results?
- Are the parameters within the dataset affective in water quality prediction?
- Should there be other parameters to consider?
- How confident are we in our findings?

Ryerson University

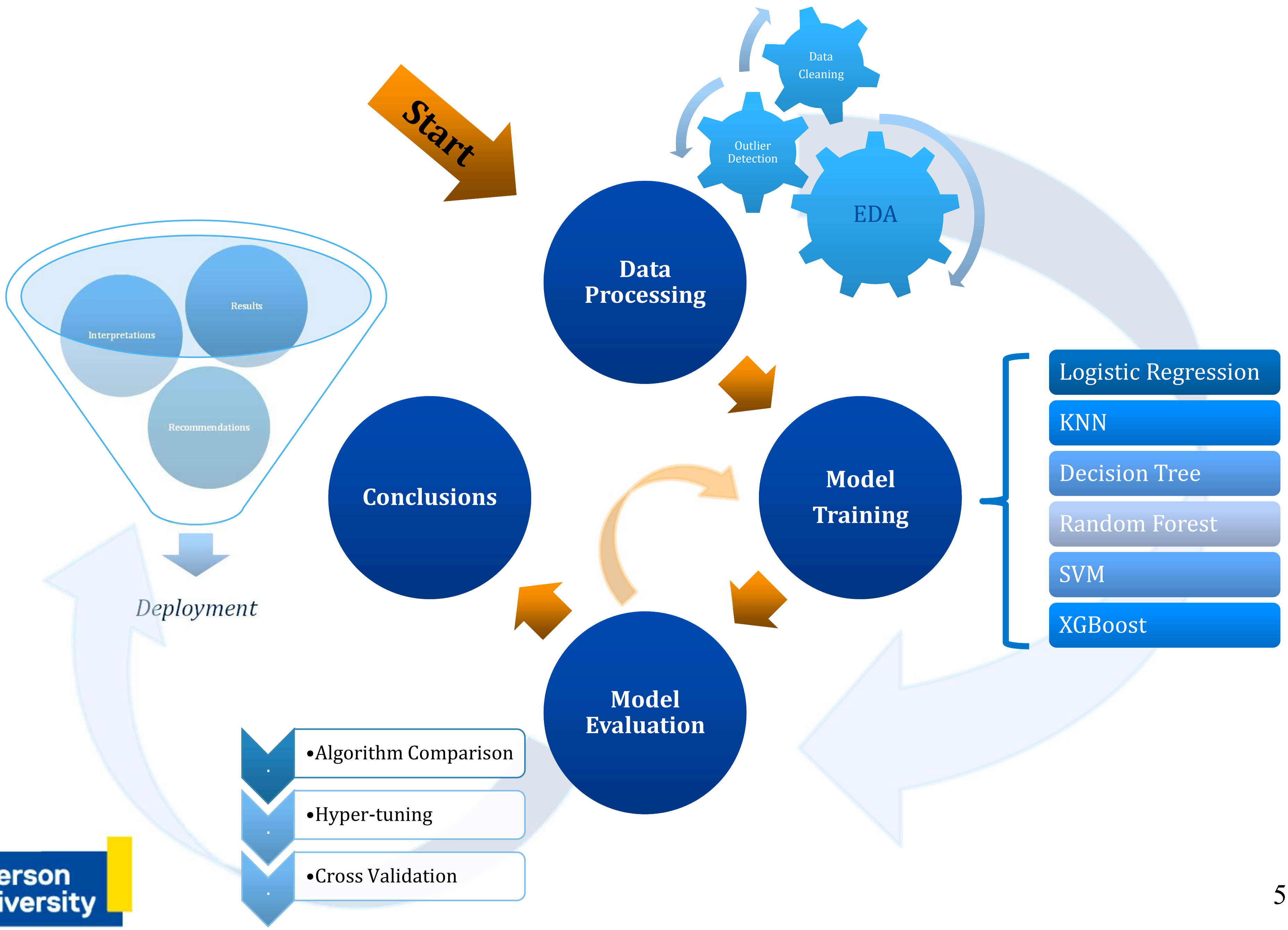
The Dataset

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



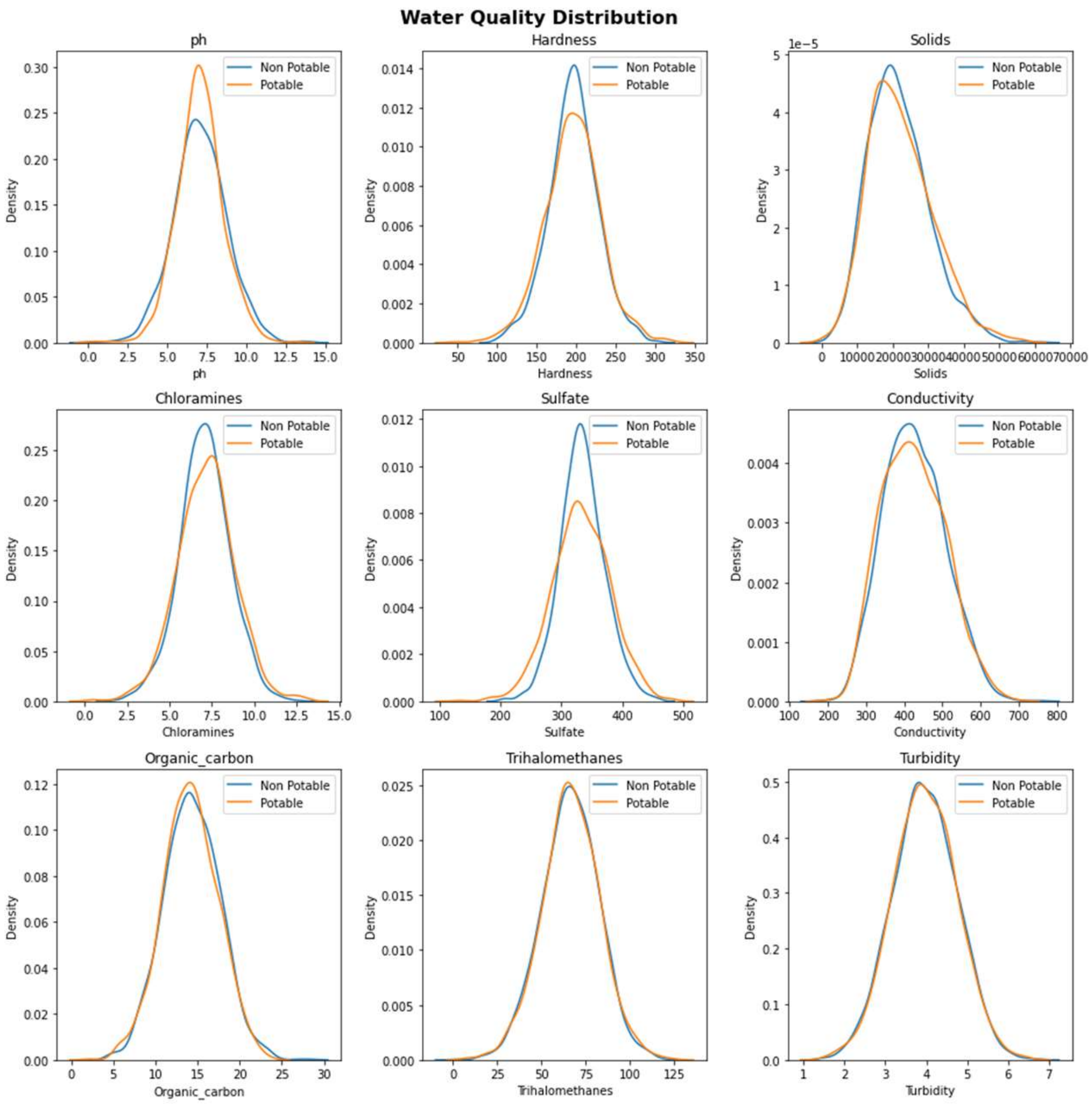
Ryerson University

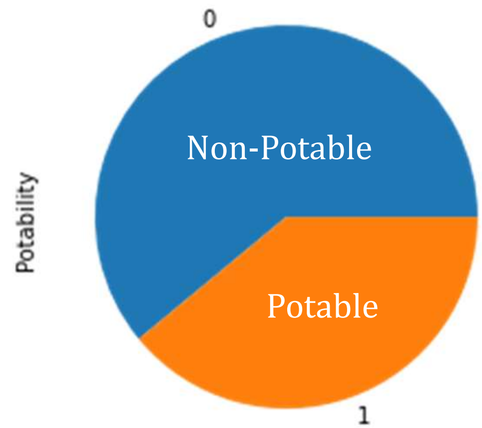
Approach Process

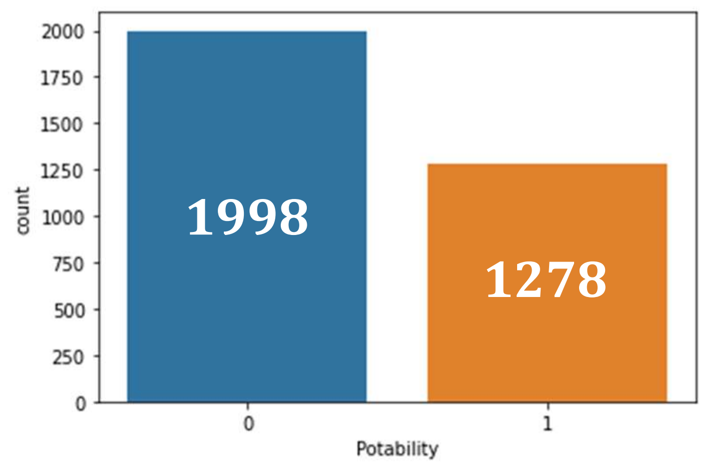


Ryerson University

EDA: Visual Analyses

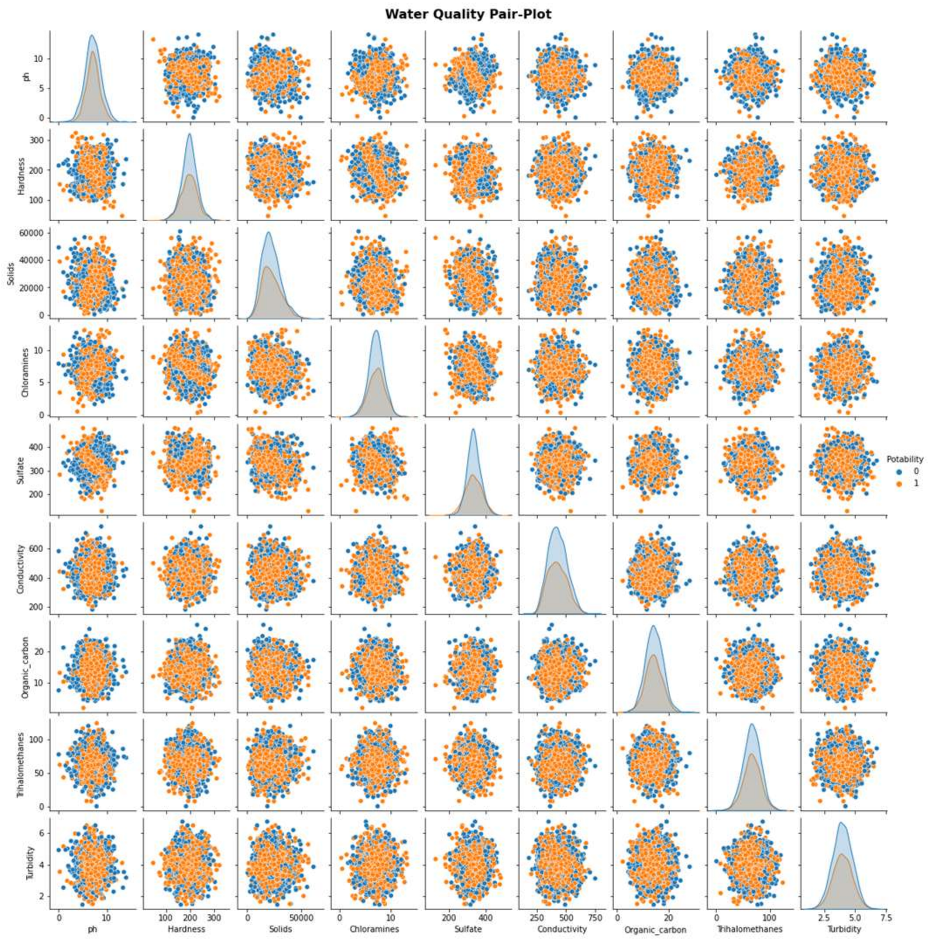


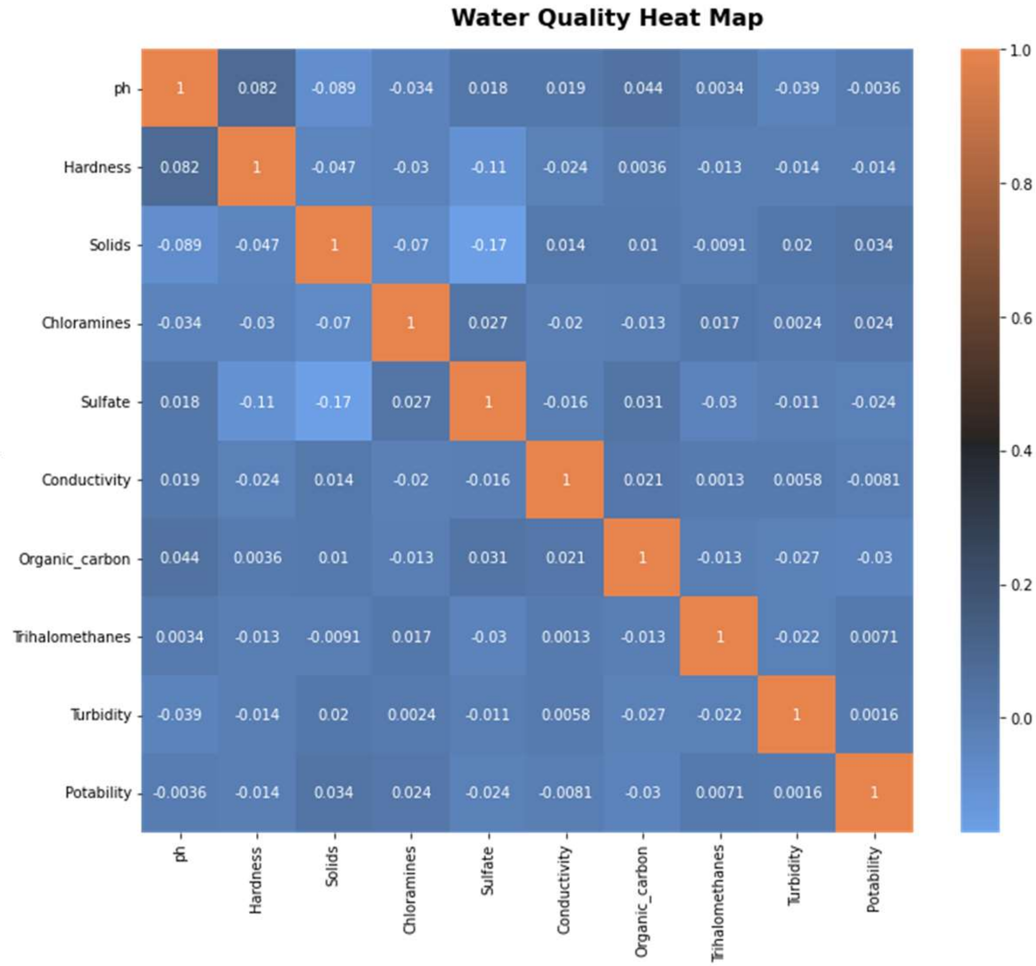




Ryerson University

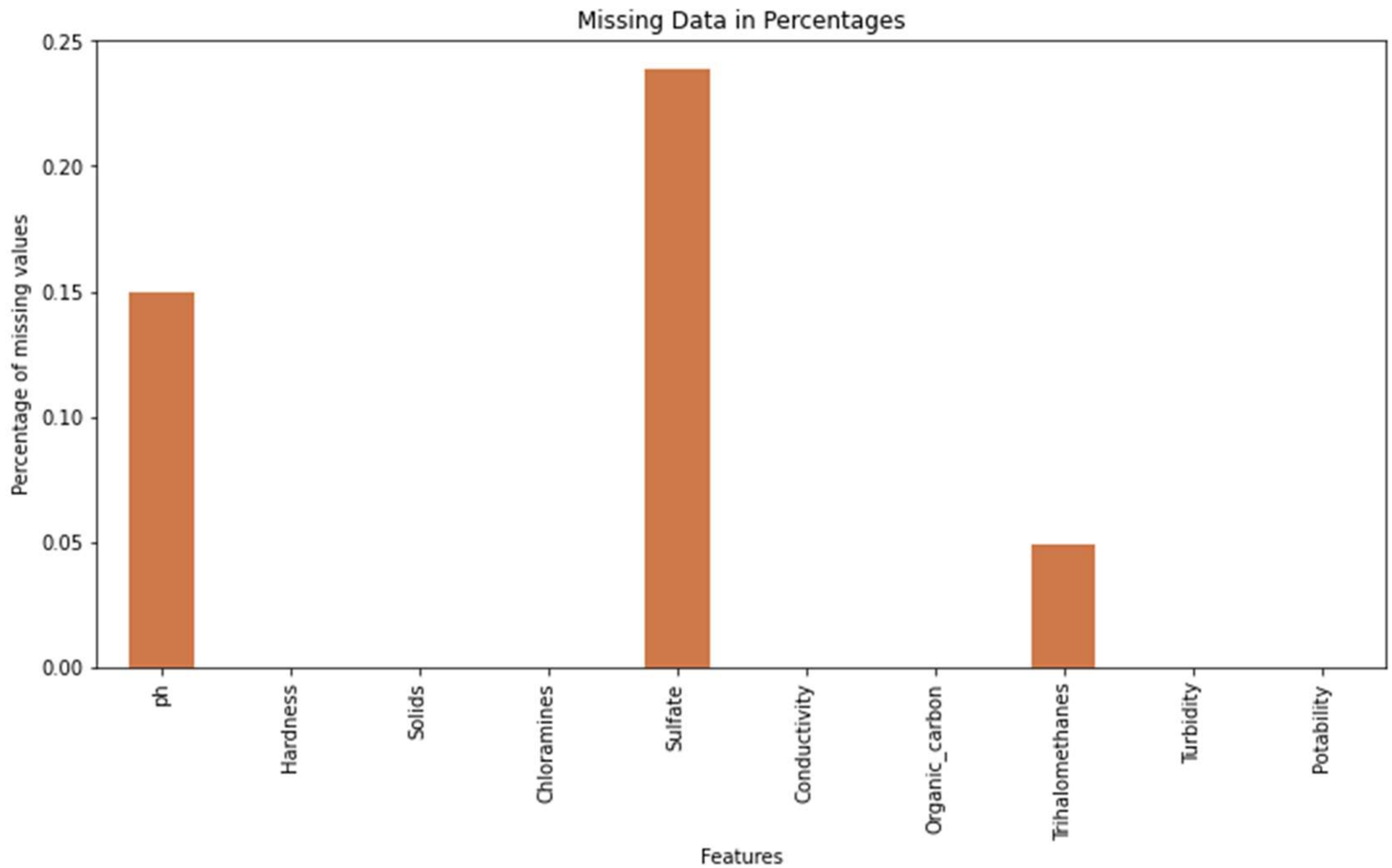
EDA: Visual Analyses





Ryerson University

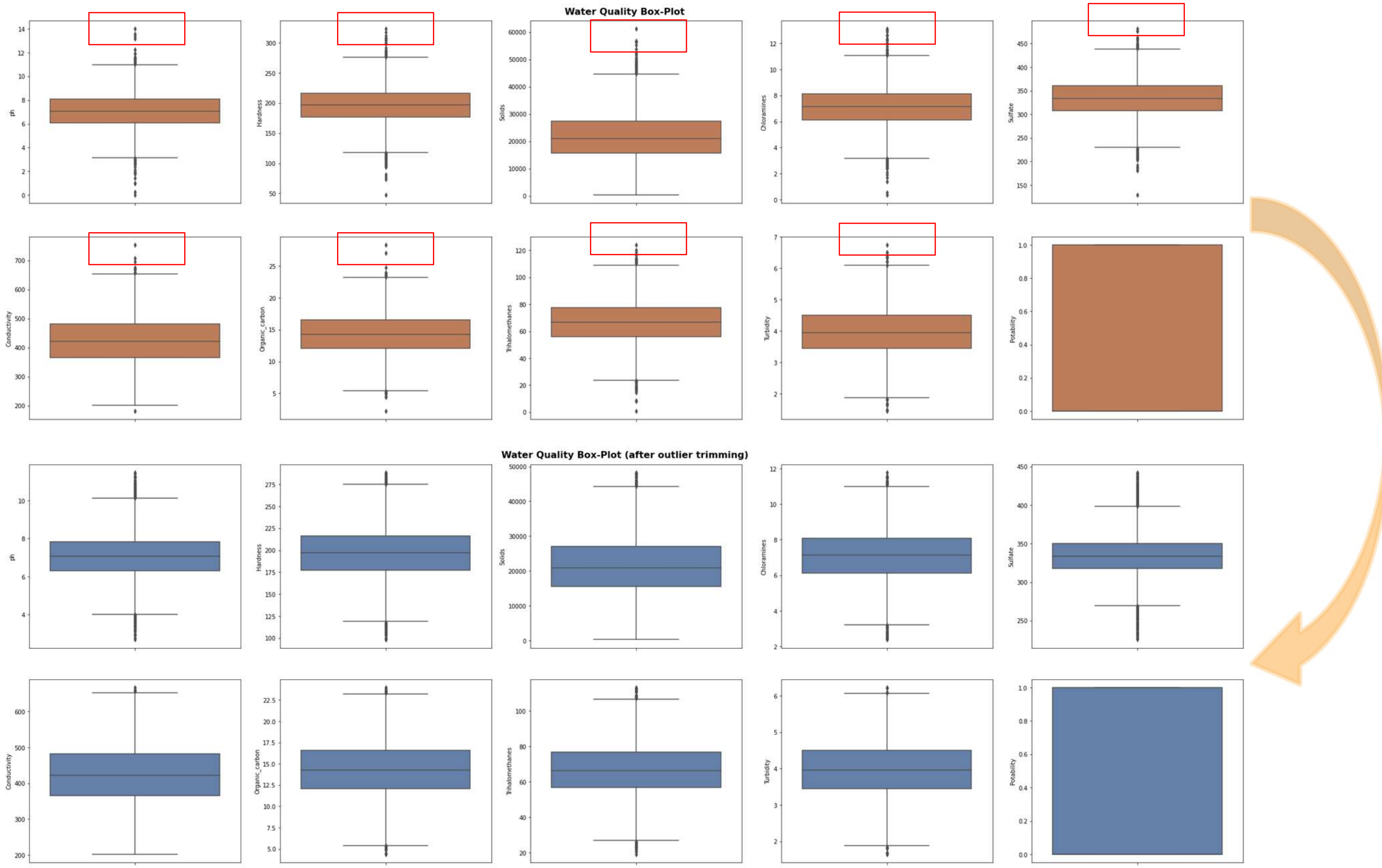
Missing Values



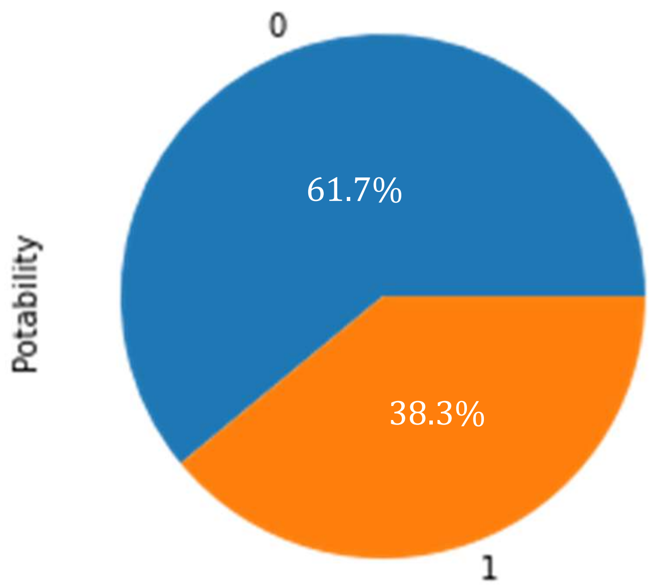
- The majority of the parameters have a Gaussian distribution therefore it was safe to replace missing values with the mean value

Ryerson University

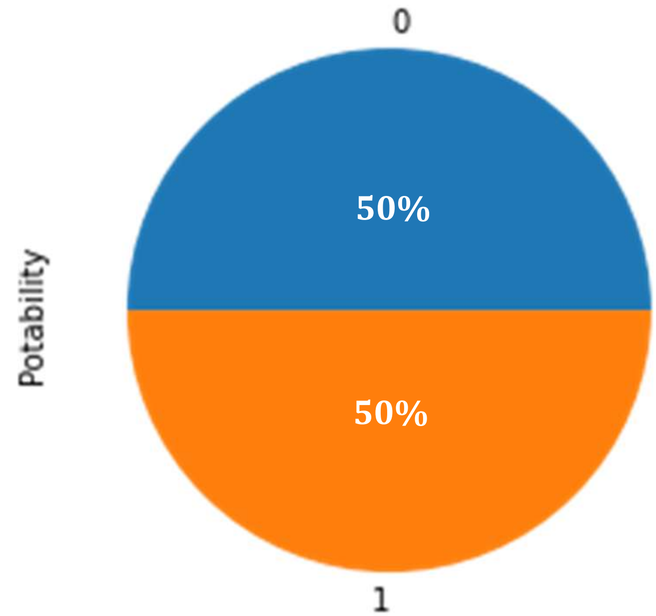
Outlier Detection



Class Imbalance



1930/1198

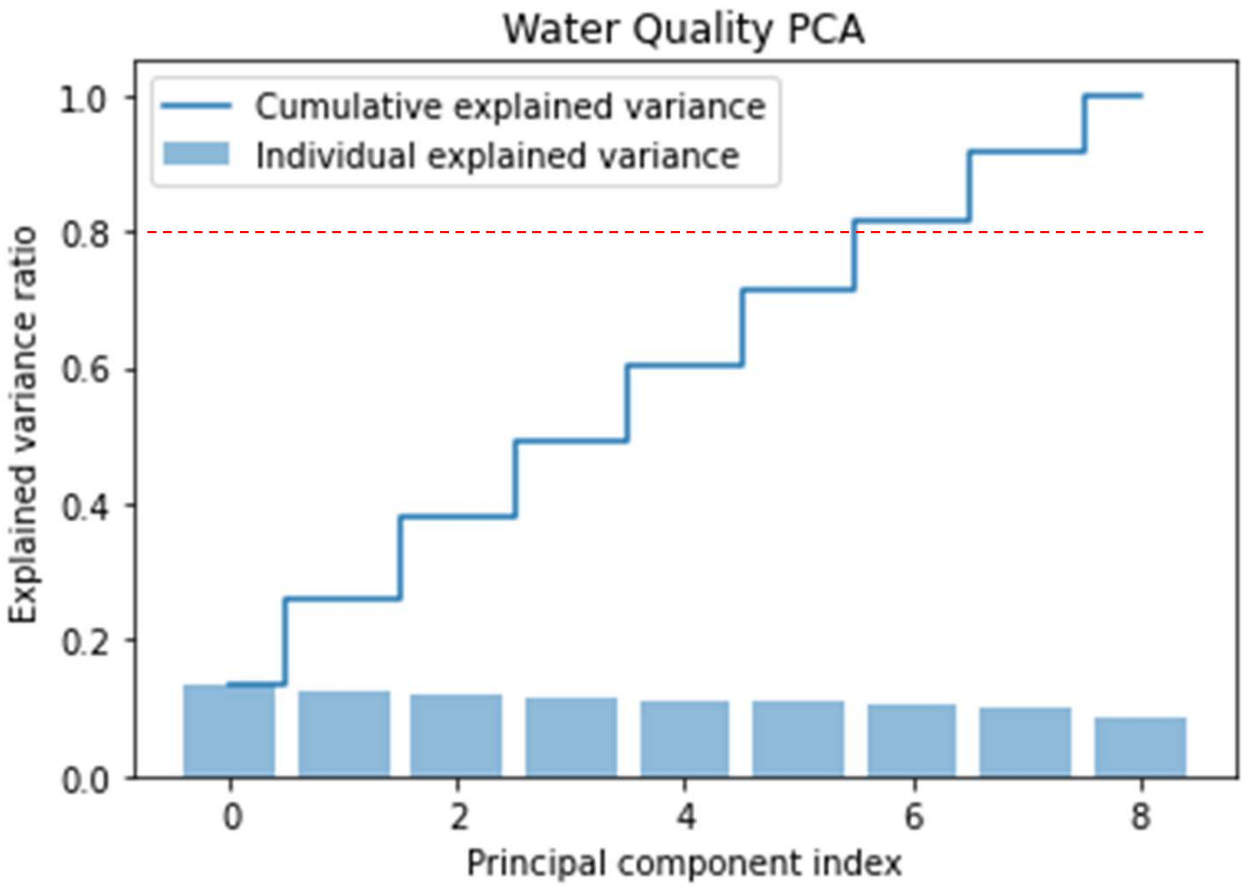


1930/1930

- Up-sampling the minority class to balance the data for training to prevent bias to the majority class

Principle Component Analysis

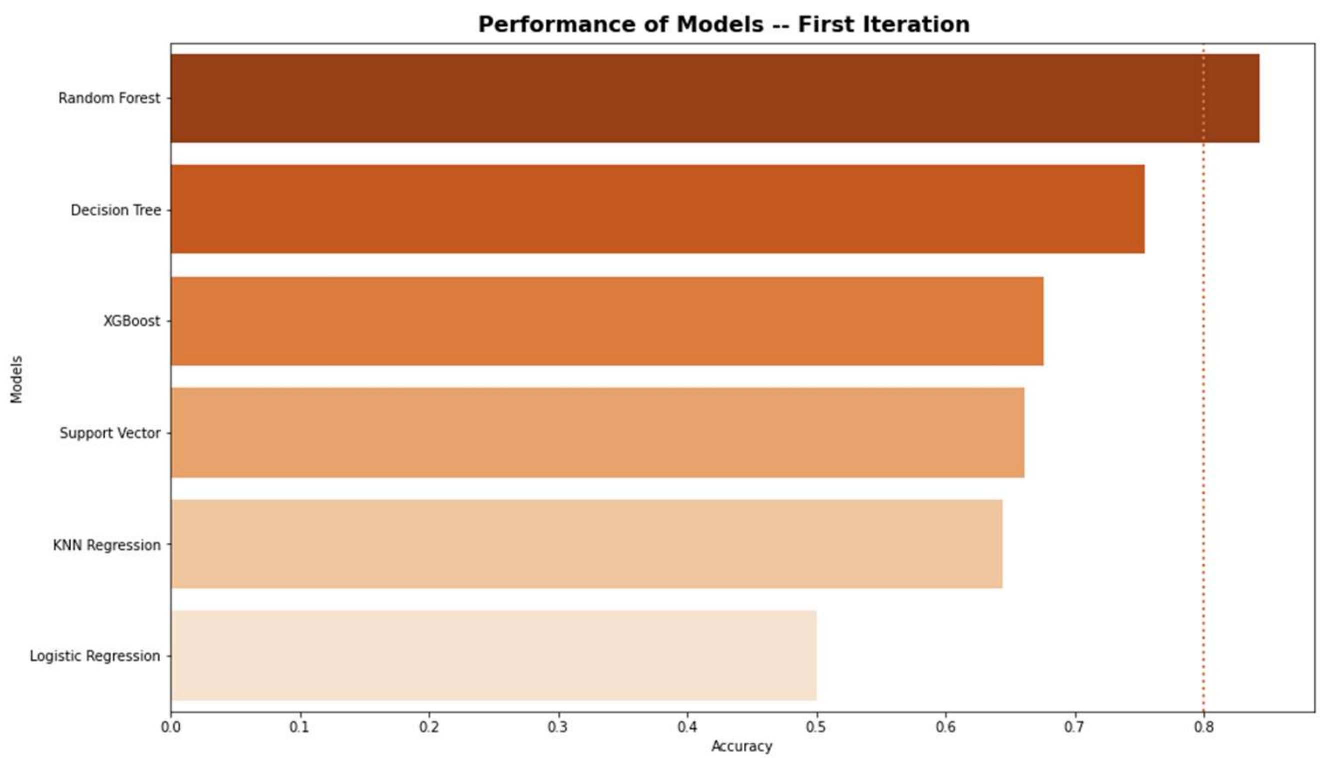
- Exploring dimensionality reduction using **PCA** tells us that all the variables are independent from each other and further confirms our previous observations from the heatmap.



Ryerson University

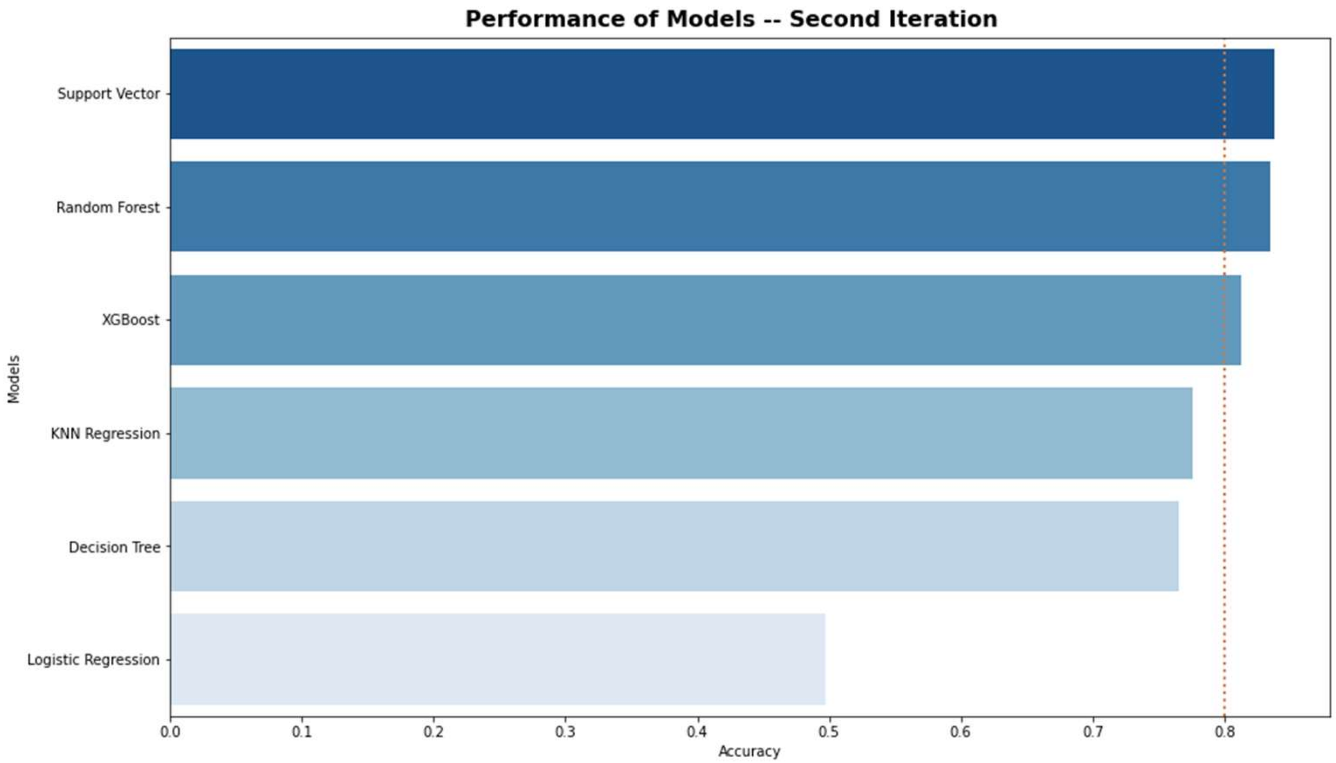
Algorithm Comparison 1st Iteration

	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



Algorithm Comparison 2nd Iteration

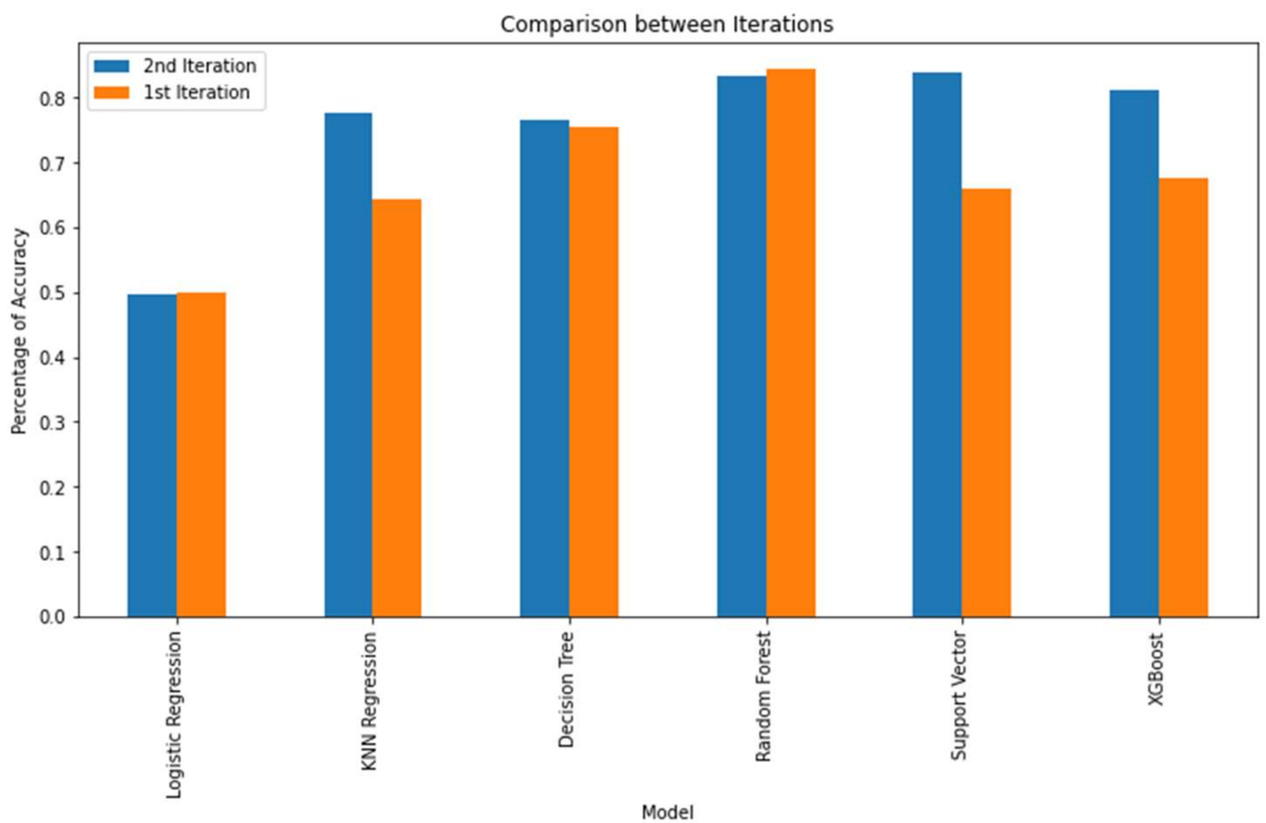
	Model	Accuracy	Precision	Recall	F1 Score
4	Support Vector	0.838083	0.853868	0.801075	0.826630
3	Random Forest	0.834197	0.808081	0.860215	0.833333
5	XGBoost	0.812176	0.763341	0.884409	0.819427
1	KNN Regression	0.775907	0.722595	0.868280	0.788767
2	Decision Tree	0.765544	0.714607	0.854839	0.778458
0	Logistic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



Ryerson University

Model Evaluation

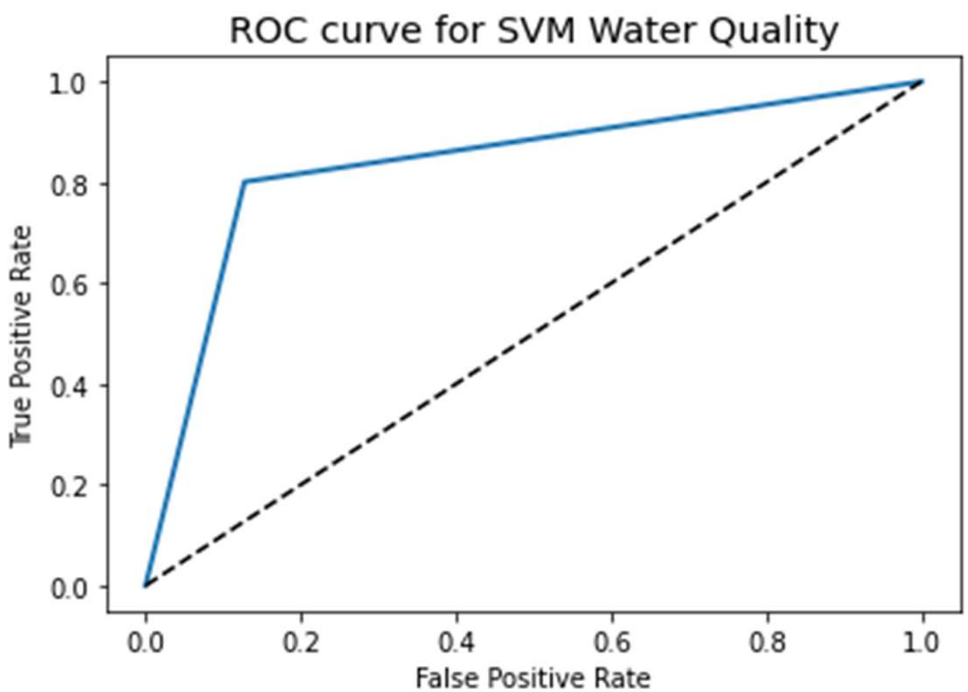
	Model	2nd Iteration	1st Iteration	Difference in Accuracy
0	Logistic Regression	49.74%	50.00%	-0.26%
1	KNN Regression	77.59%	64.38%	13.21%
2	Decision Tree	76.55%	75.39%	1.17%
3	**Random Forest**	**83.42%**	**84.33%**	**-0.91%**
4	**Support Vector**	**83.81%**	**66.06%**	**17.75%**
5	XGBoost	81.22%	67.62%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



Cross Validation

K-Fold CV

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



ROC AUC: **0.8368**
CV ROC AUC: **0.8674**

Interpretation & Recommendations

- After 2nd iteration and hyper-tunning parameters, SVM performed with the greatest accuracy 84.33%
- After k-Fold cross validation, SVM's accuracy increased to 87.98%

- Increasing parameters: coliforms and heavy metals
- Explore deeper machine learning such as ANN (artificial neural network)

Ryerson University

Conclusions

- Can we predict water potability?
- Which machine learning algorithms can yield the most efficient and accurate results?
- Can the parameters within the ML algorithms be tuned to yield the best results?
- Are the parameters within the dataset affective in water quality prediction?
- Should there be other parameters to consider?
- How confident are we in our findings?

✓ **Using ML it is possible to predict water potability**

✓ **Support Vector Machine Classifier best performance 87.98% accuracy**

✓ **Hyper-tunning did increased accuracy in modeling for most of the algorithms**

✓ **The parameters within the dataset were affective in prediction although had low correlation**

✓ **Through research from other studies, additional attributes such as coliform and heavy metals should be included**

✓ **Confident in our findings but room for improvement**

Ryerson University