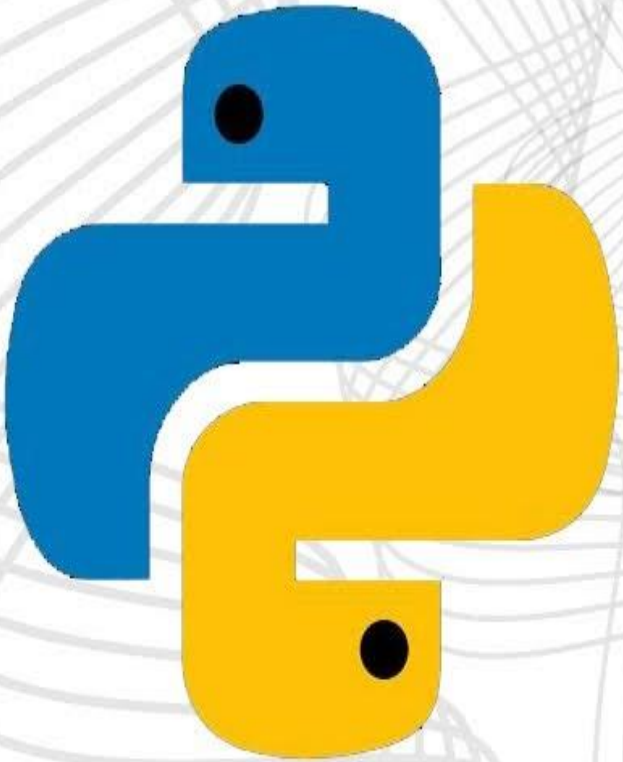
Water Quality Prediction using ML



python

Project by: Rohan Anil Gaikwad

Project Guide : Prof. Poonam Bhawke

Why classify water?

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

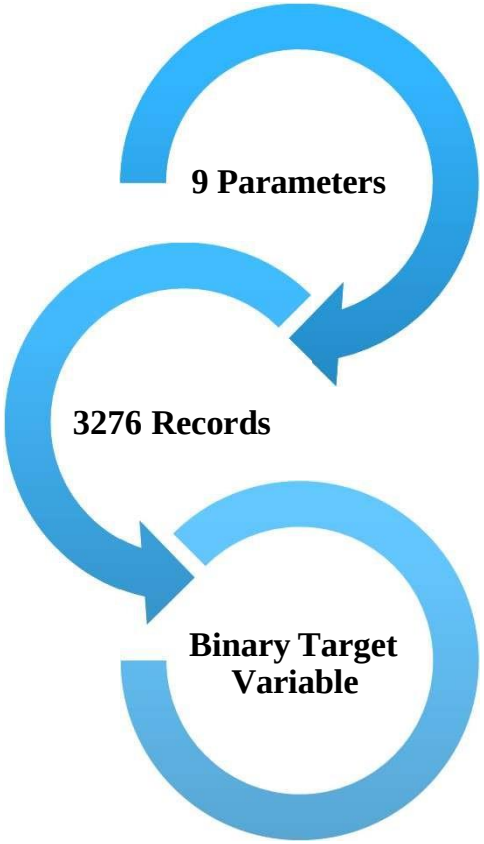
Can machine learning improve the process of water classification?

Predicting Water Potability

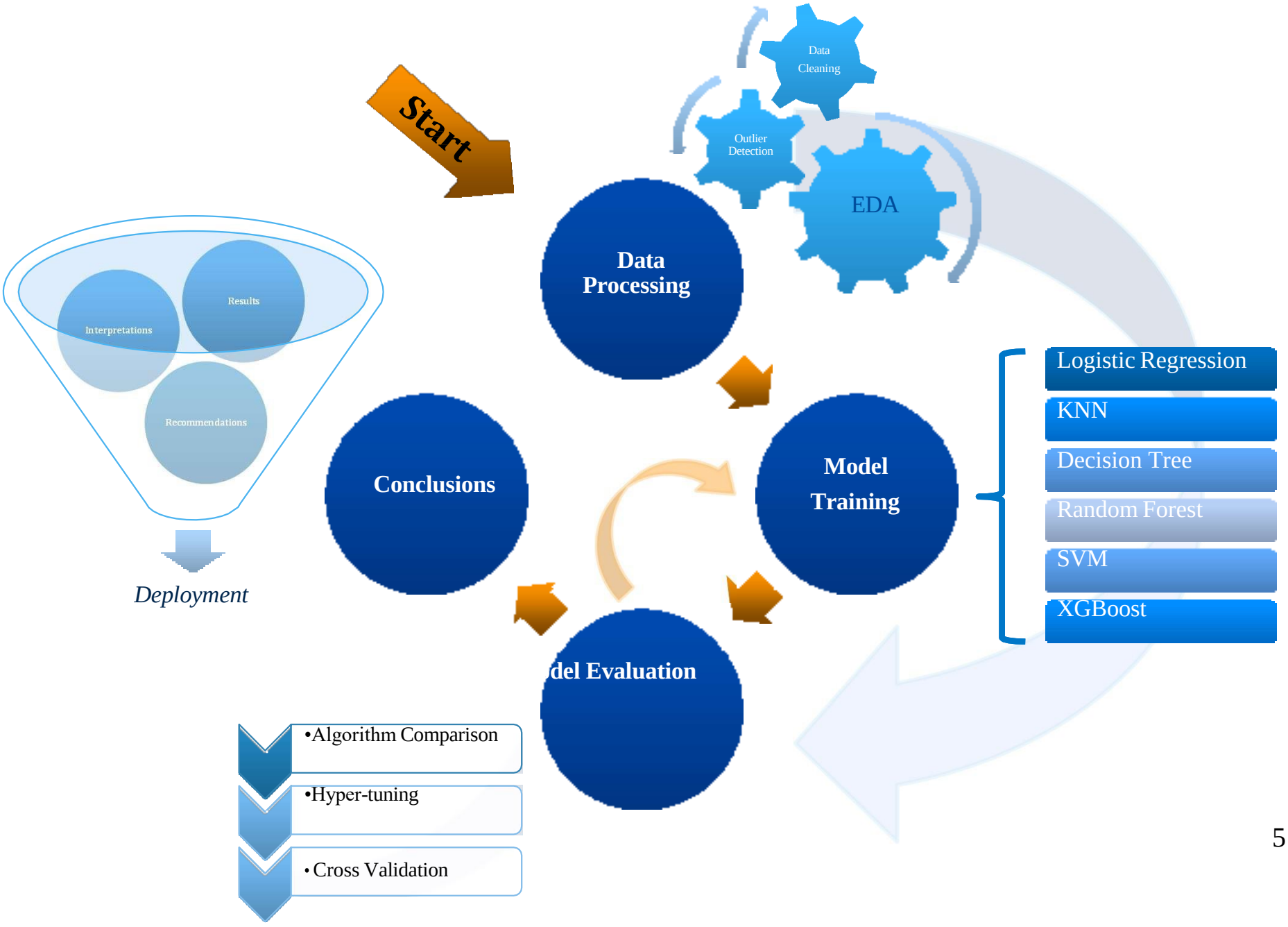
- Can we predict water potability?
- Which machine learning algorithms can yield the most efficient and accurate results?
- Can the parameters within the ML algorithms be tuned to yield the best results?
- Are the parameters within the dataset affective in water quality prediction?
- Should there be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

The Dataset

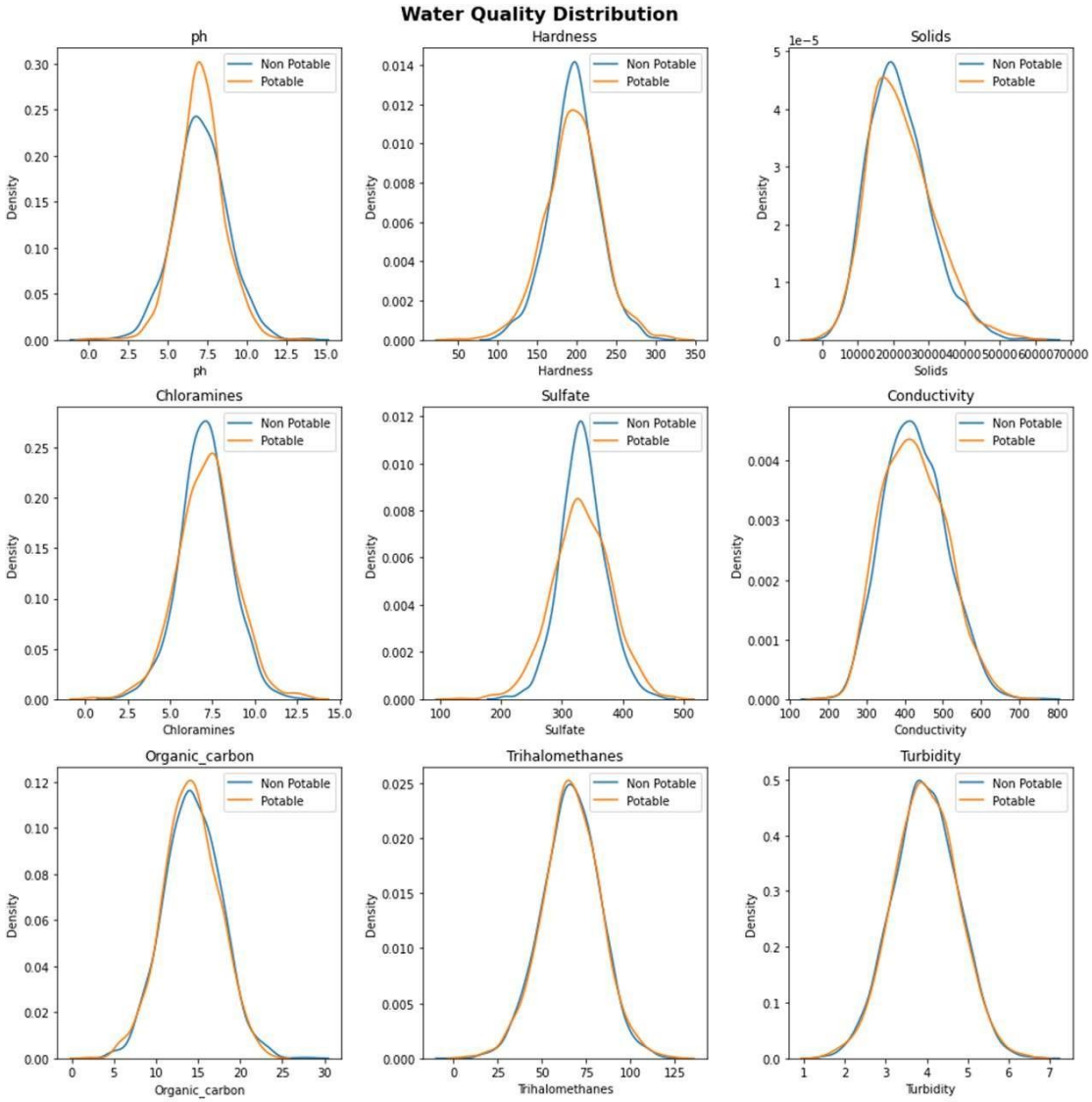
https://www.kaggle.com/datasets/adityakadiwal/water-potability/

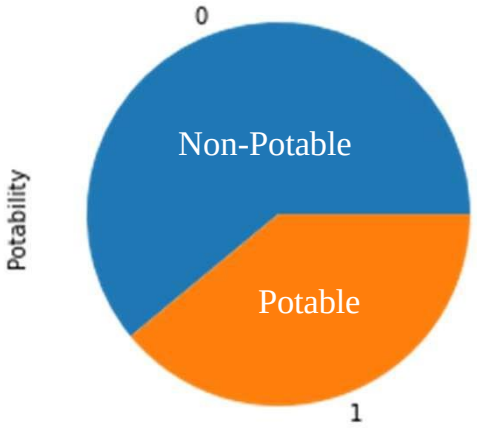


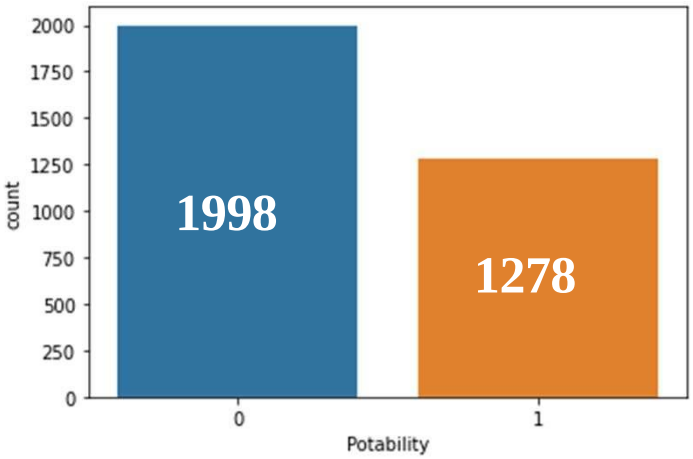
Approach Process



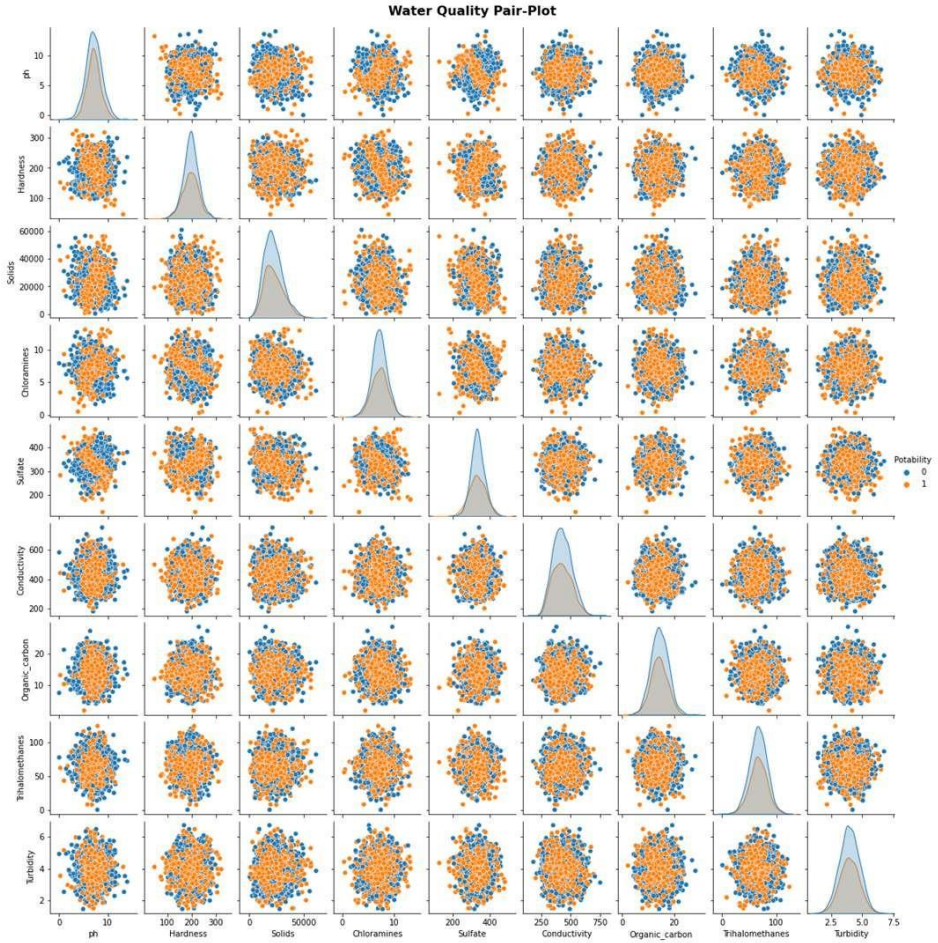
EDA: Visual Analyses

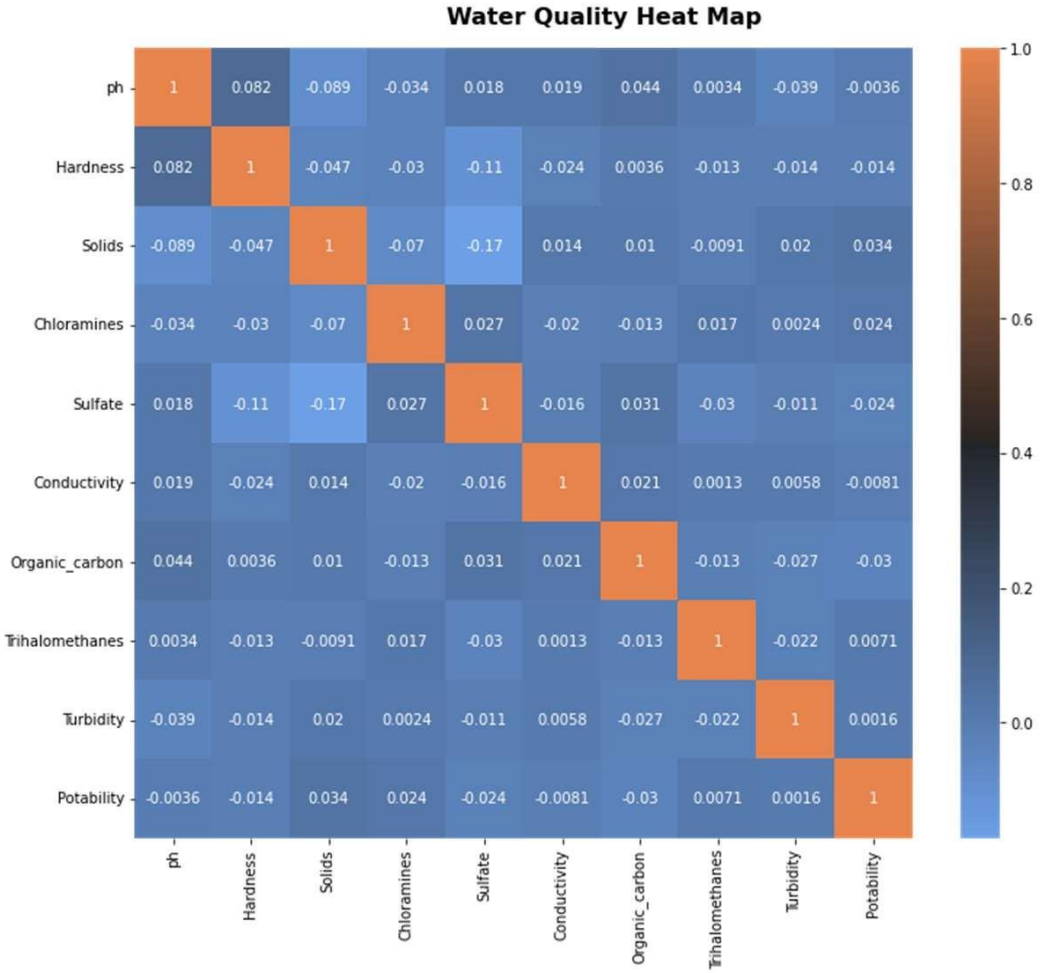




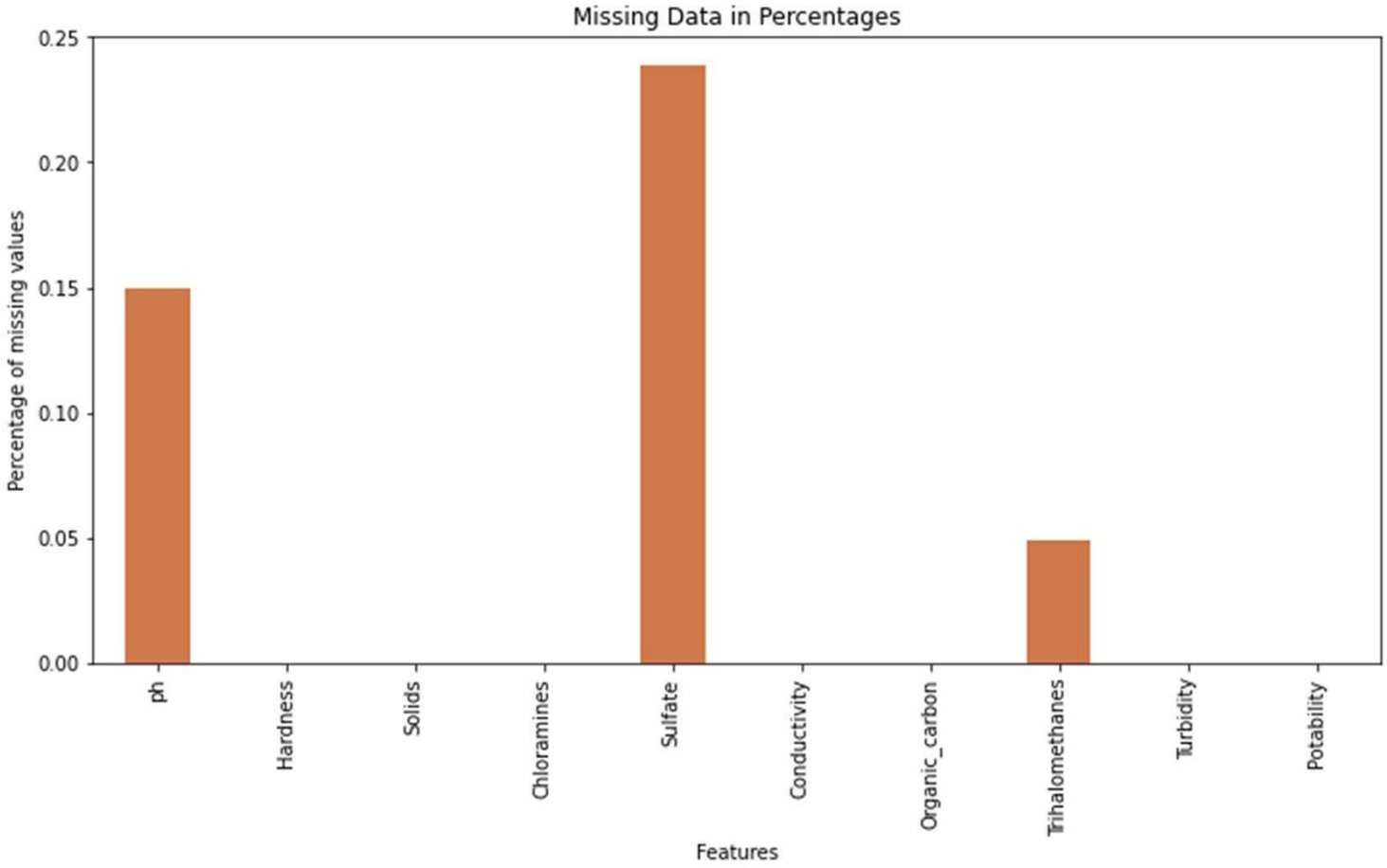


EDA: Visual Analyses



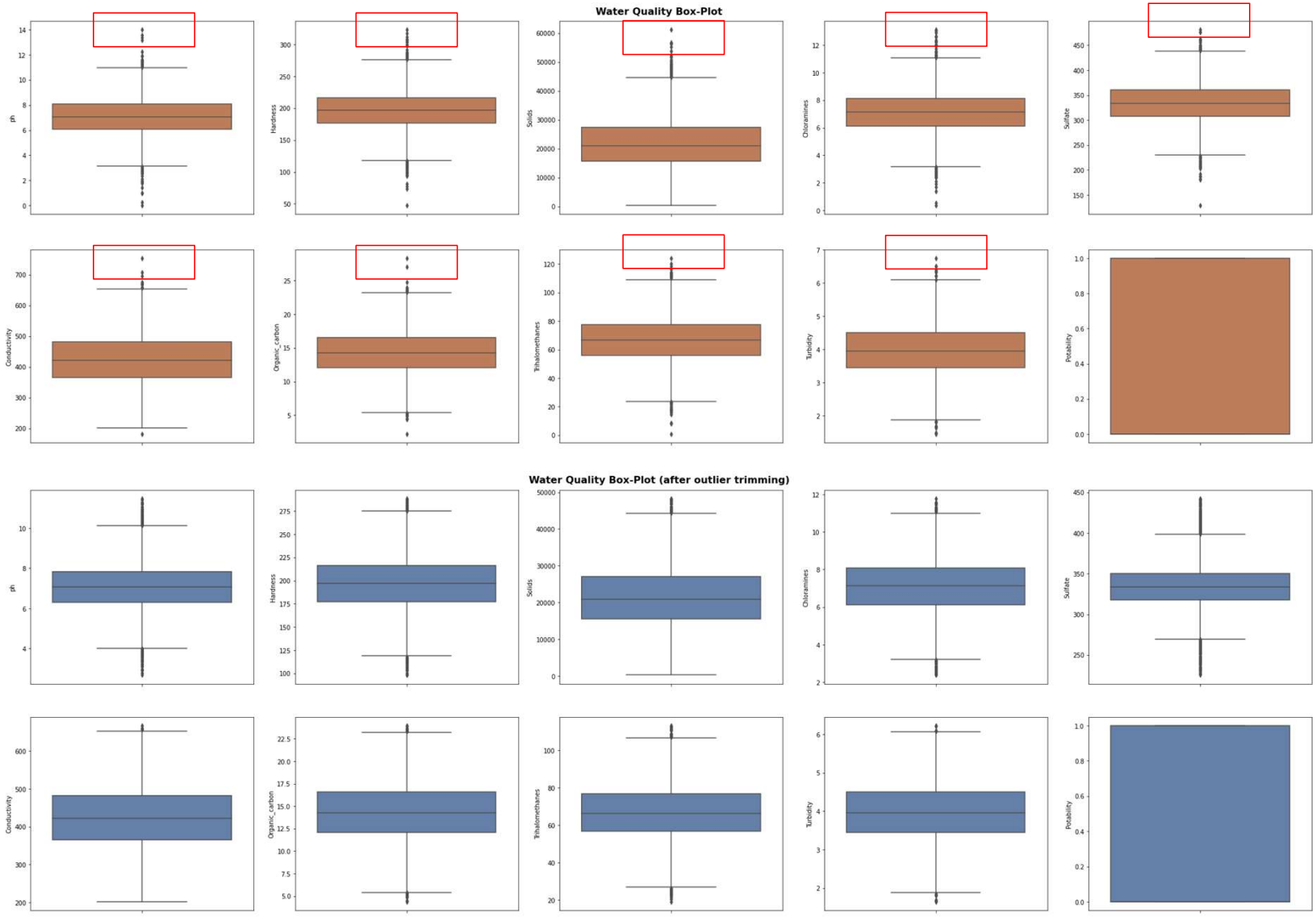


Missing Values

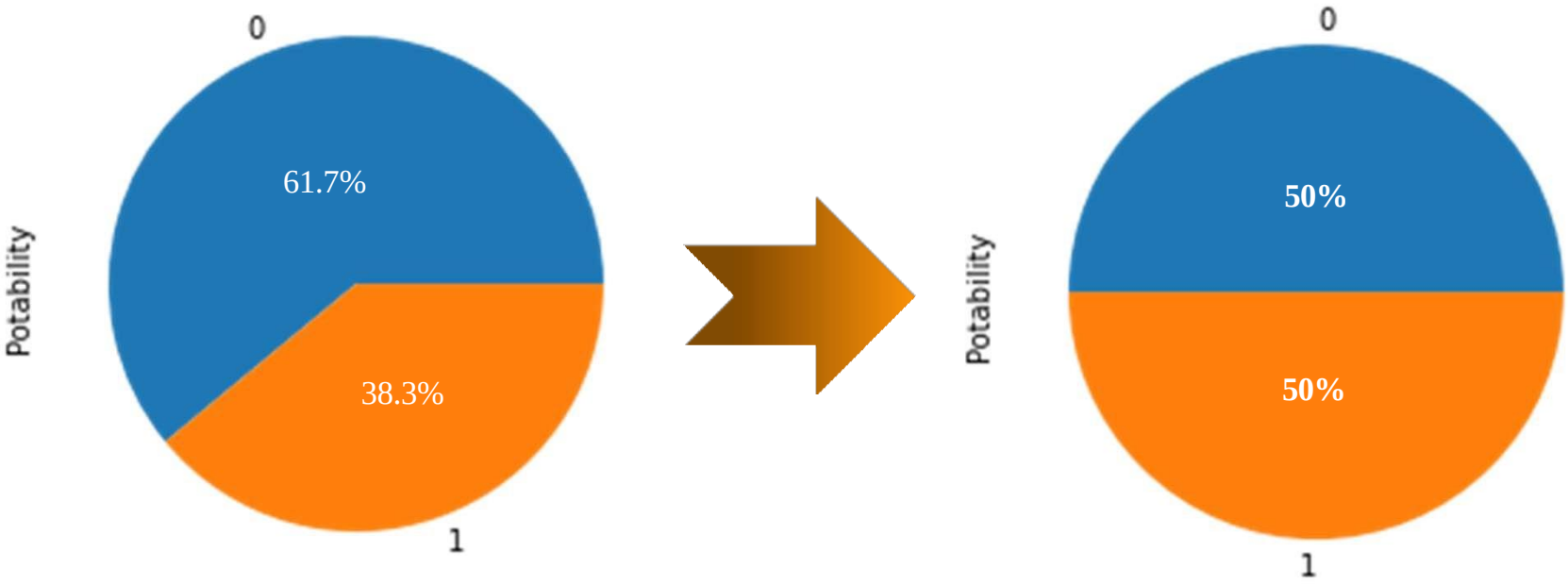


- The majority of the parameters have a Gaussian distribution therefore it was safe to replace missing values with the mean value

Outlier Detection



Class Imbalance

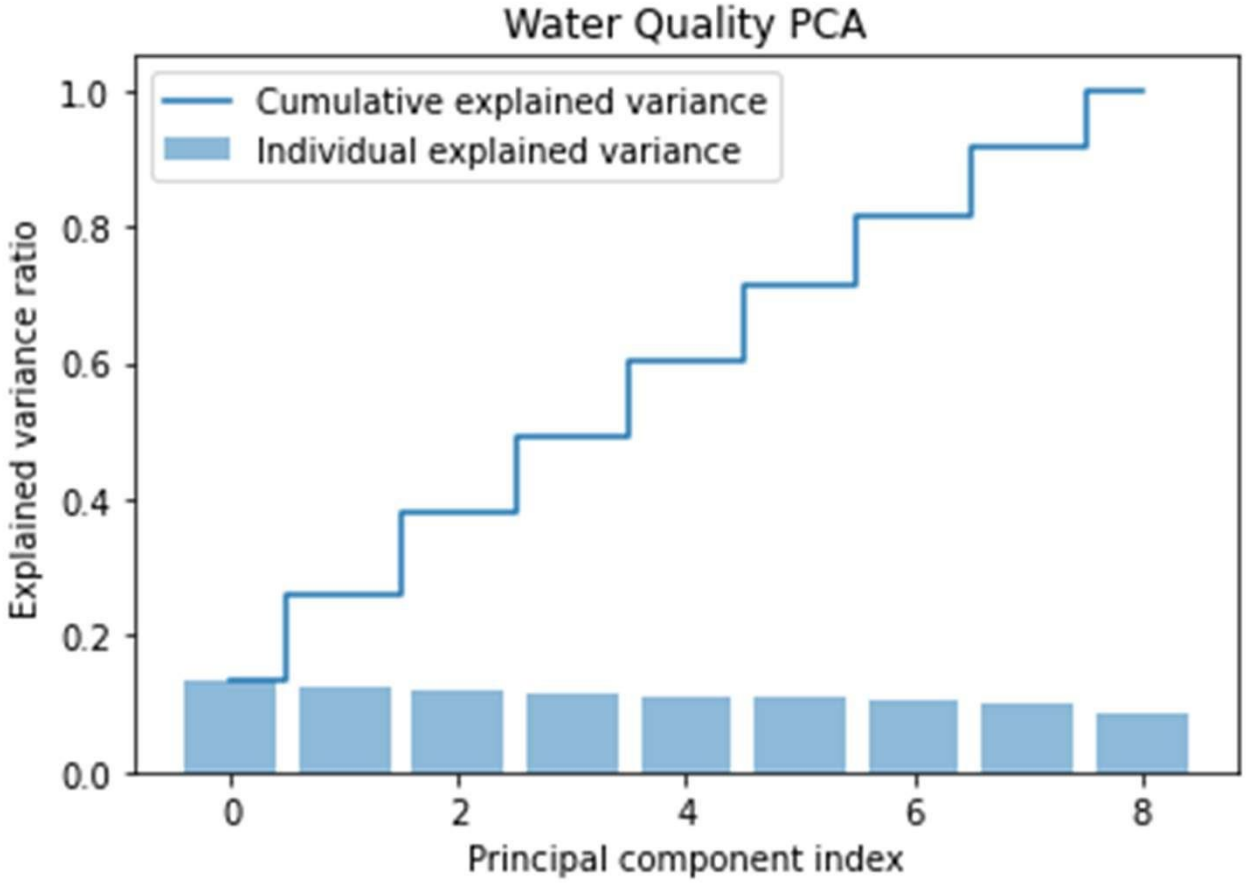


1930/1198 **1930/1930**

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

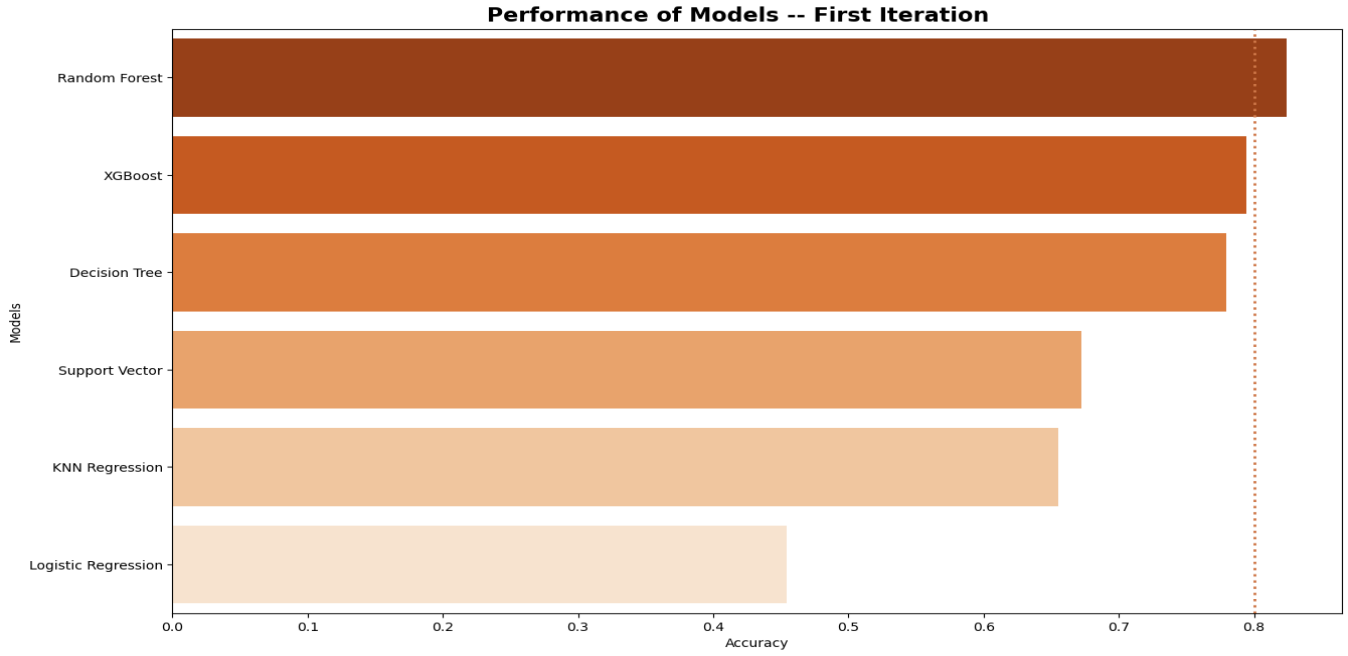
Principle Component Analysis

- 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.



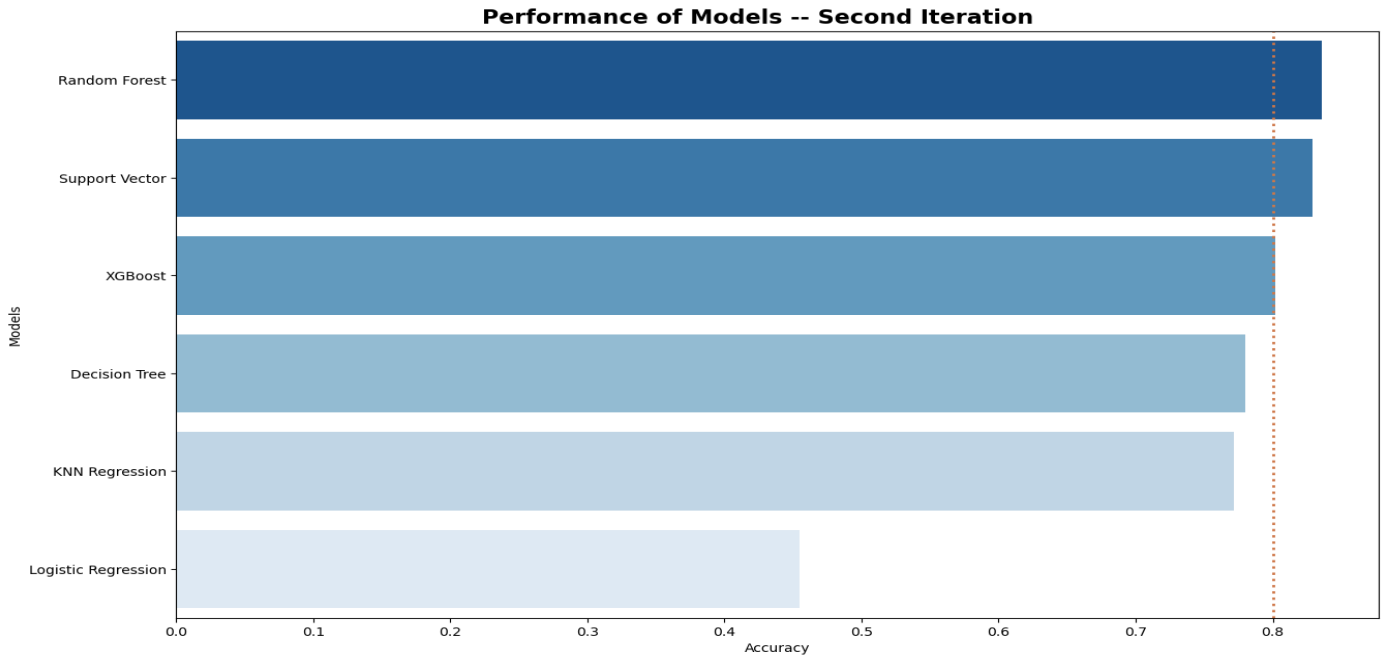
Algorithm Comparison 1st Iteration

	Model	Accuracy
3	Random Forest	0.82383
2	Decision Tree	0.77979
5	XGBoost	0.77404
4	Support Vector	0.67228
1	KNN Regression	0.65544
0	Logistic Regression	0.45460



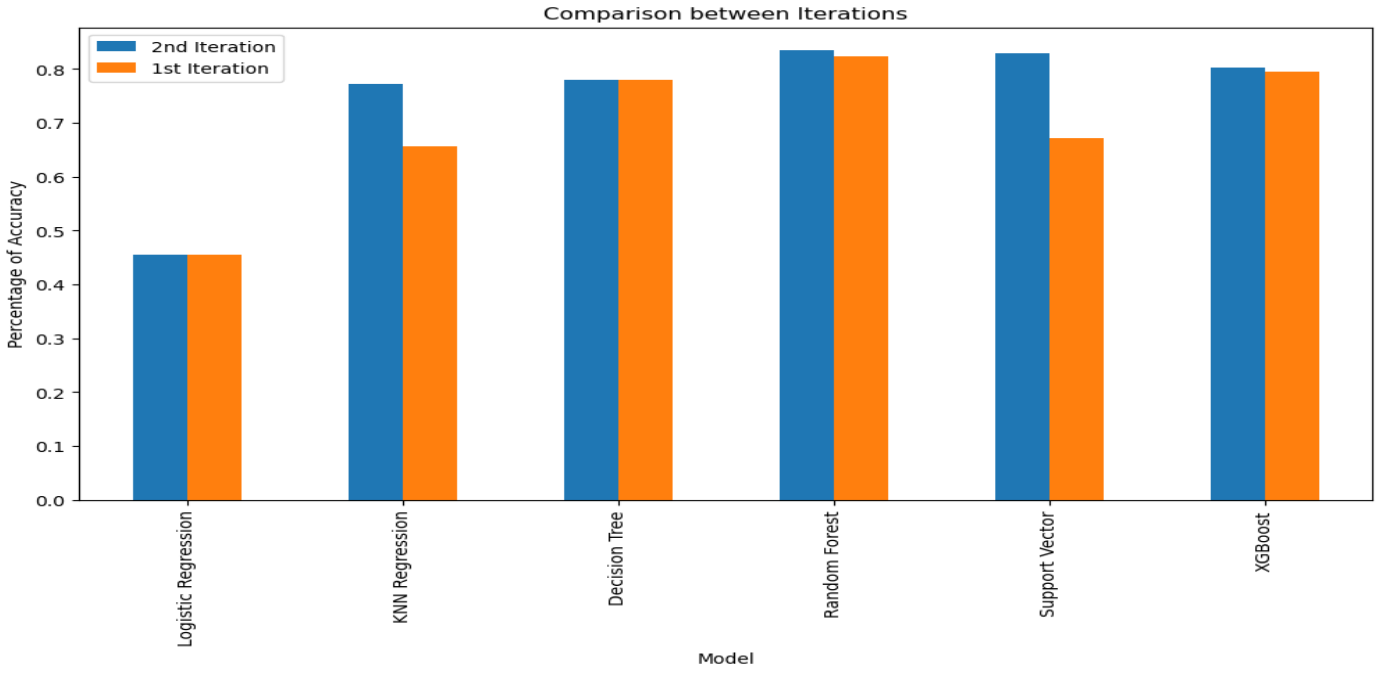
Algorithm Comparison 2nd Iteration

	Model	Accuracy
3	Random Forest	0.835482
4	Support Vector	0.829016
5	XGBoost	0.801813
2	Decision Tree	0.779793
1	KNN Regression	0.772021
0	Logistic Regression	0.454663



Model Evaluation

	Model	2nd Iteration	1st Iteration	Difference in Accuracy
0	Logistic Regression	45.46%	45.46%	0%
1	KNN Regression	77.20%	65.54%	11.66%
2	Decision Tree	77.97%	77.97%	0%
3	**Random Forest**	**83.54%**	**82.38%**	**1.16%**
4	**Support Vector**	**82.90%**	**67.22%**	**15.68%**
5	XGBoost	80.18%	77.40%	2.78%



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**


Questions?