Water quality prediction

DATA 606

Capstone in Data Science

Project by: Rashmitha Challa(IZ48096)

rchalla1@umbc.edu

Supervised by: Chaojie (Jay) Wang, D.Sc.

Why classify water?

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

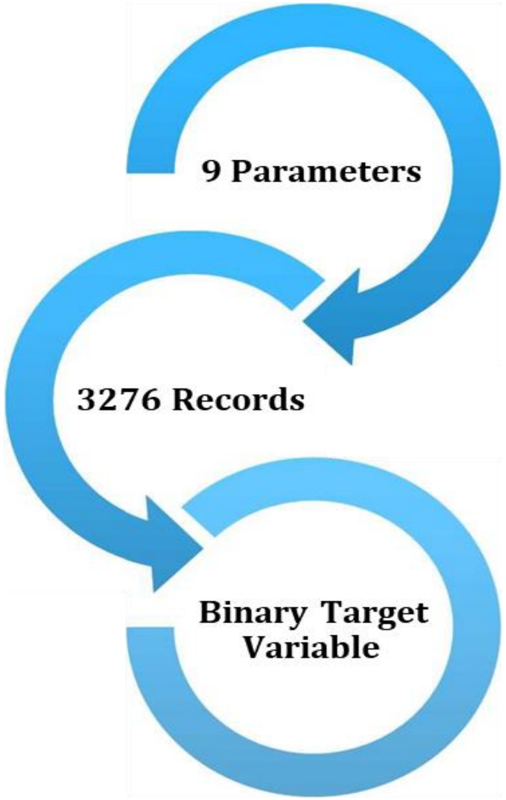
Can machine learning improve the process of water classification?

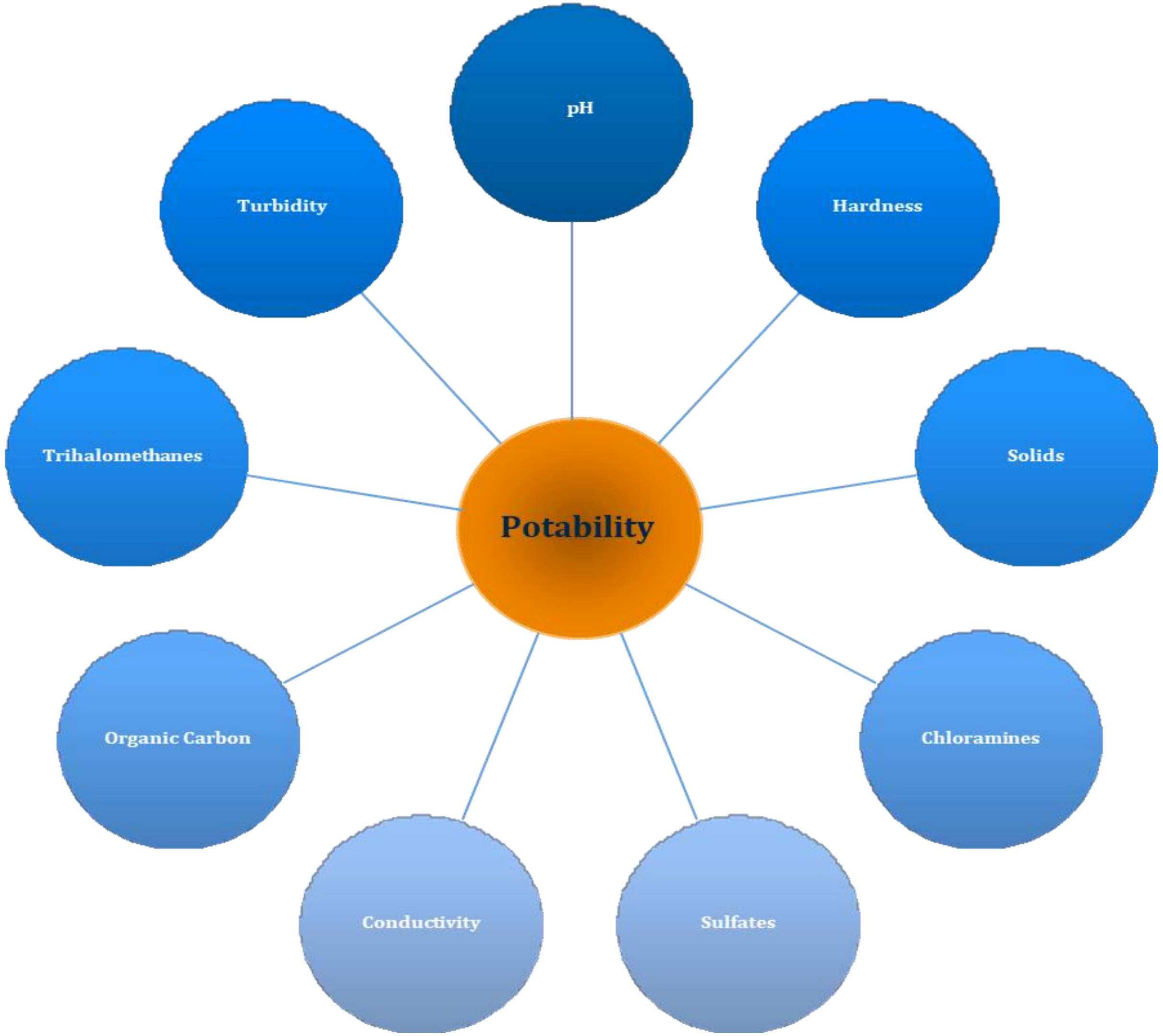
Predicting Water Potability

- 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?

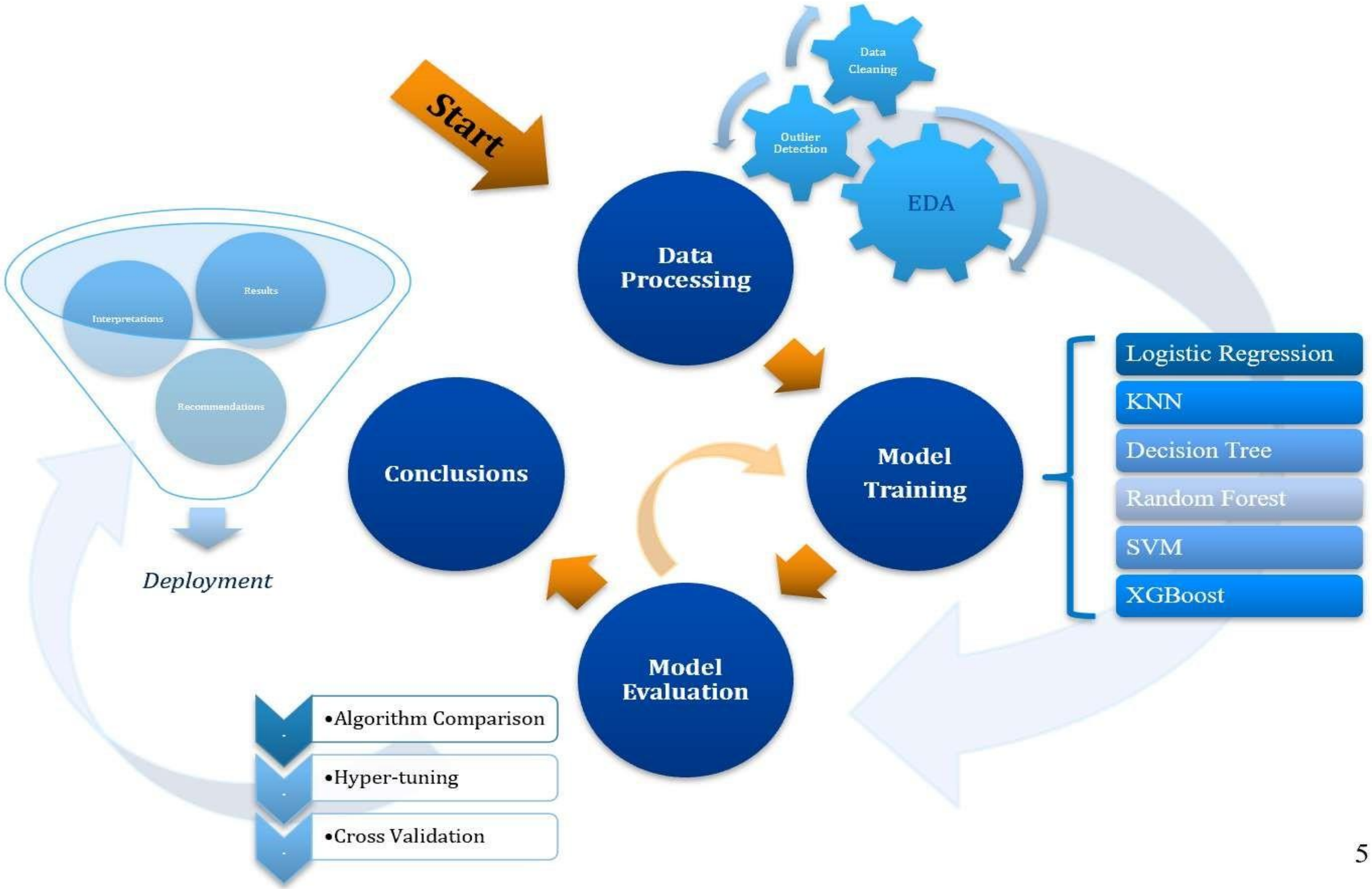
The Dataset

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



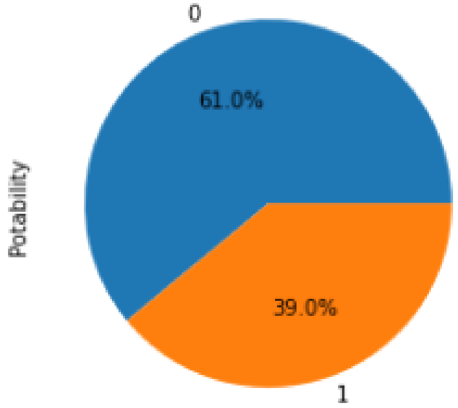


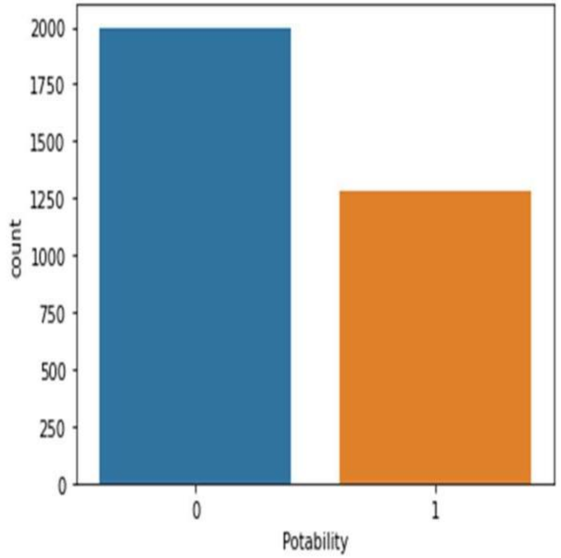
Approach

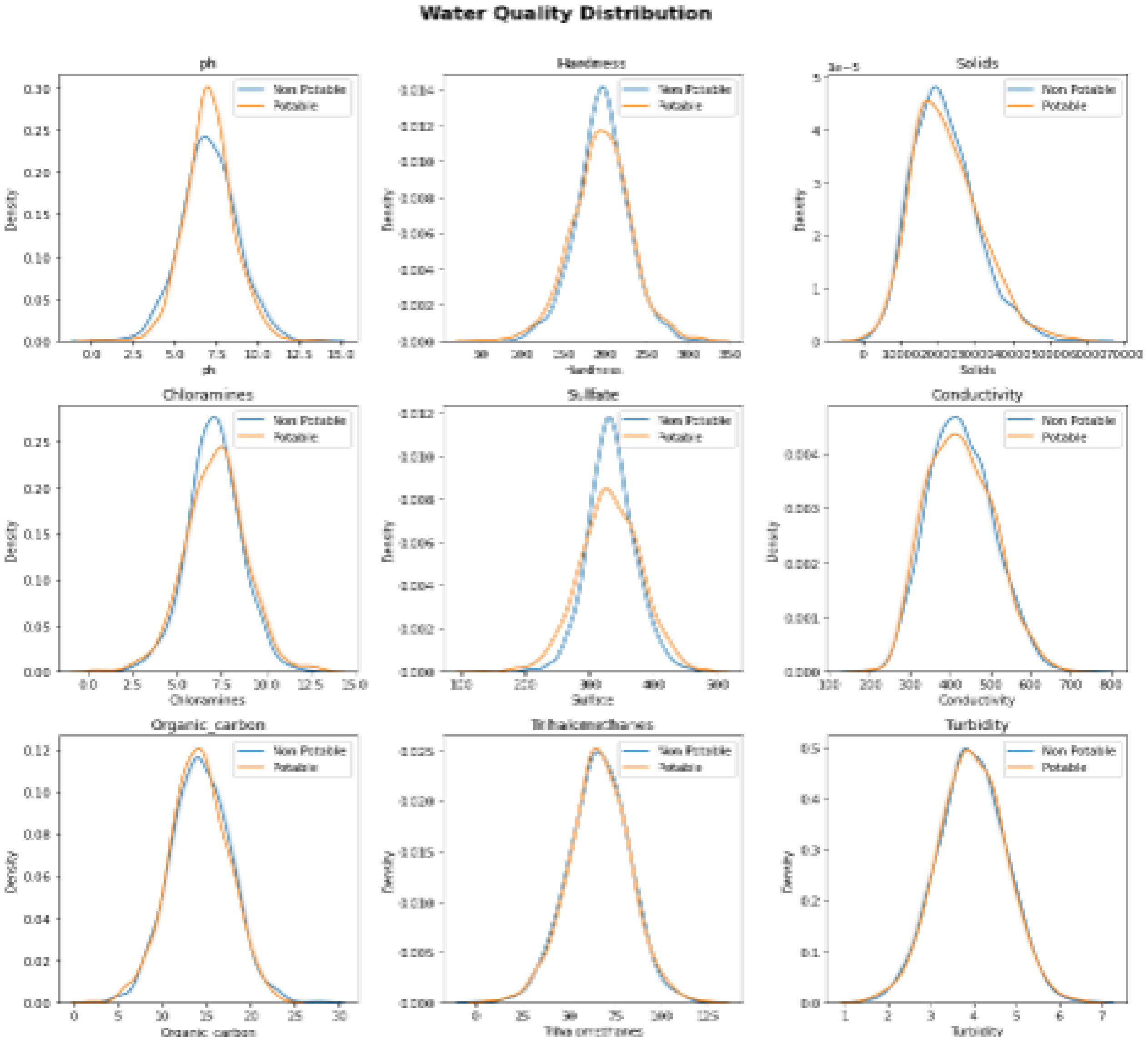

Start
Data Cleaning
Outlier Detection
EDA
Data Processing
Logistic Regression
KNN
Decision Tree
Random Forest
SVM
XGBoost
Model Training
Model Evaluation
Conclusions
Results
Interpretations
Recommendations
Deployment
•Algorithm Comparison
•Hyper-tuning
•Cross Validation

EDA: Visual Analyses

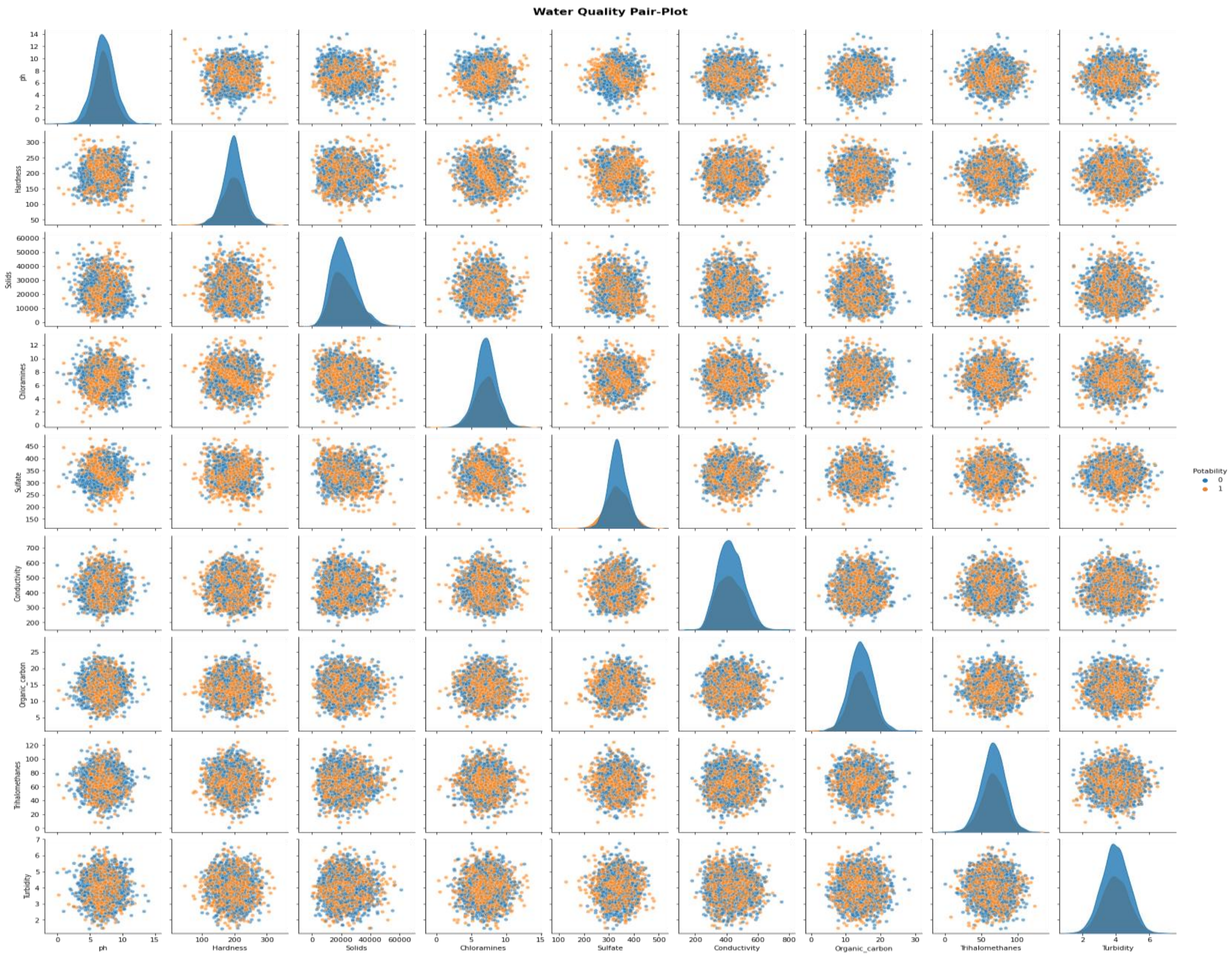
Portability Value Counts

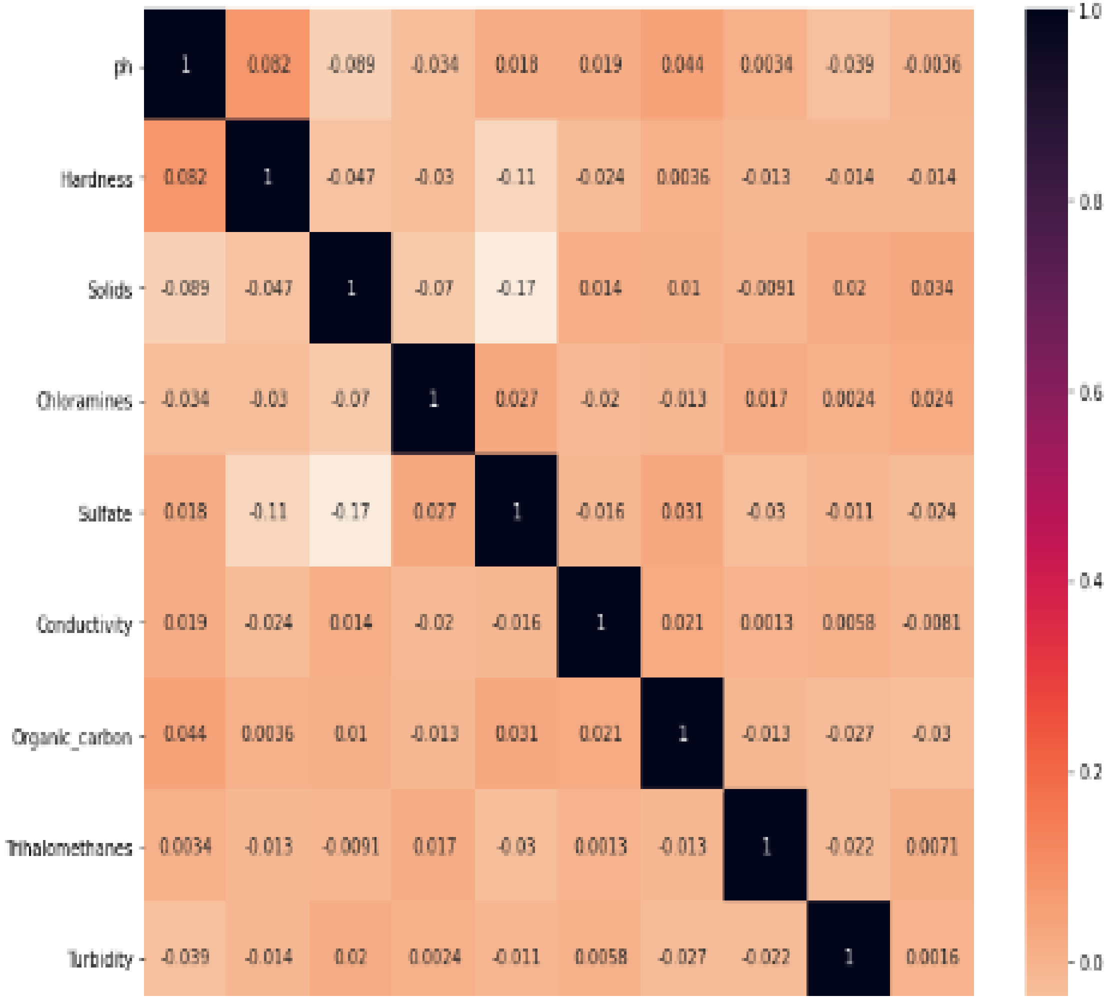




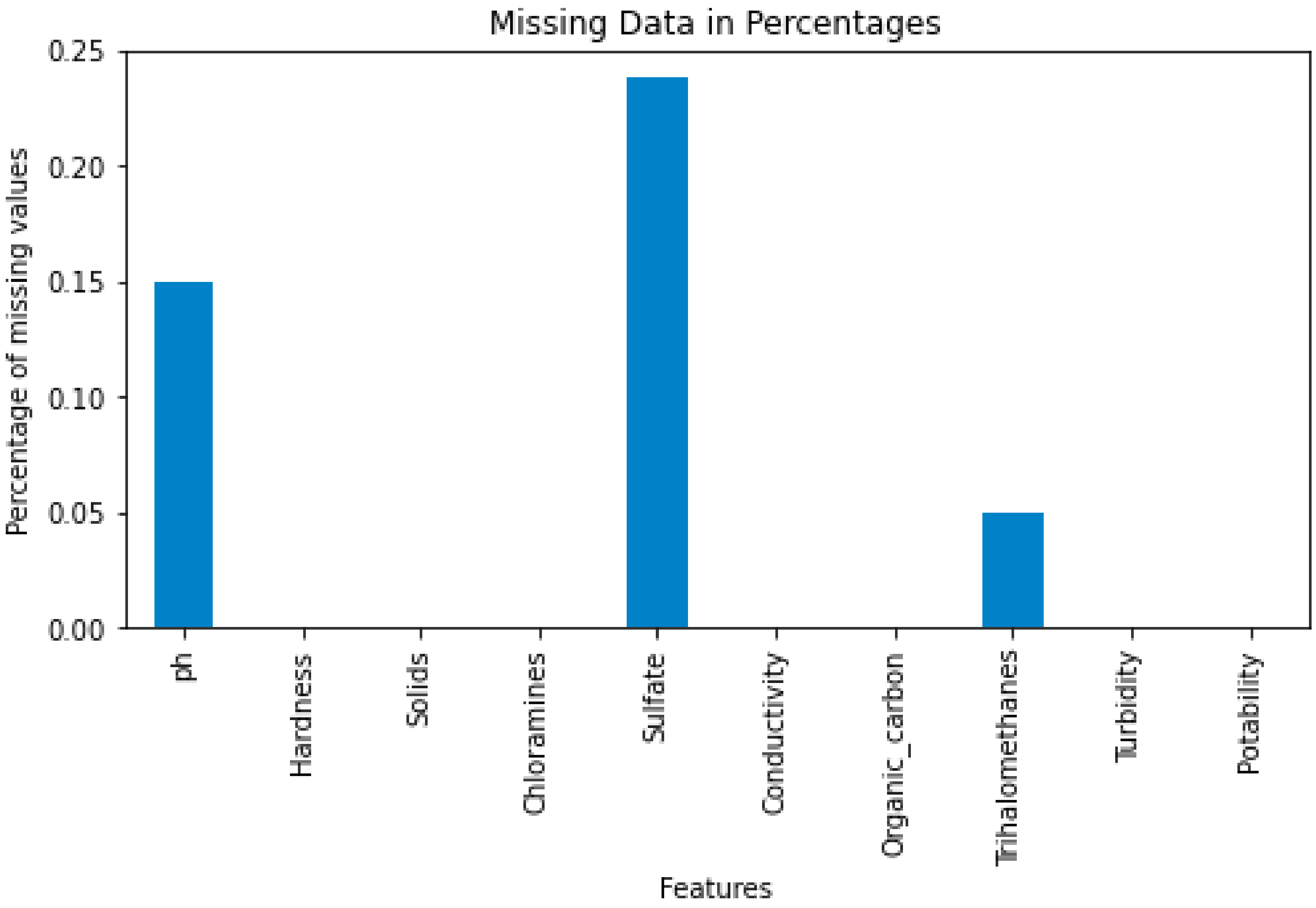


EDA: Visual Analyses

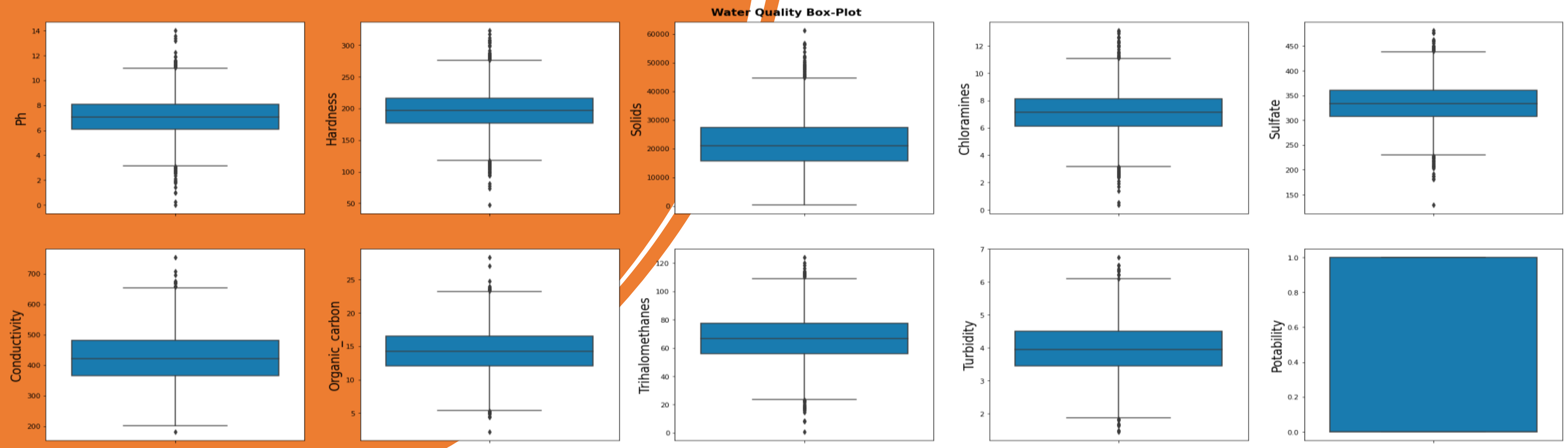


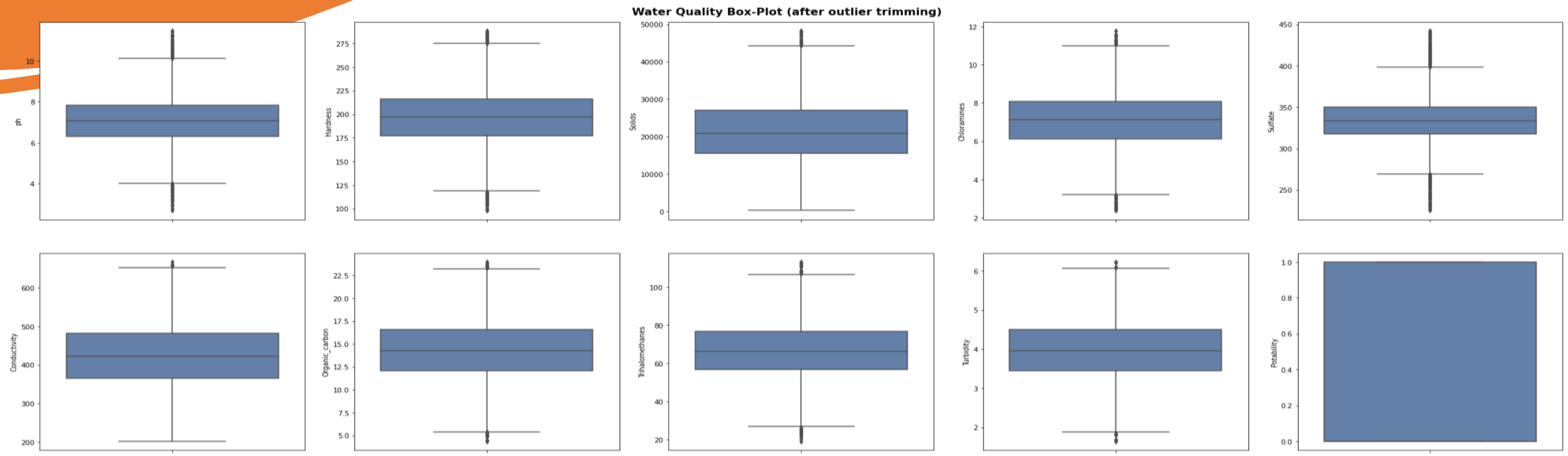


Missing Values



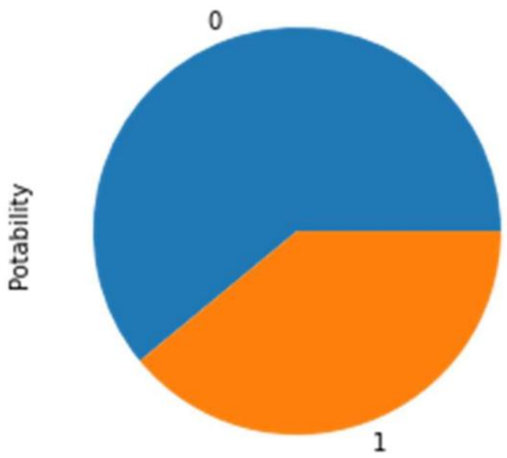
Outlier Detection



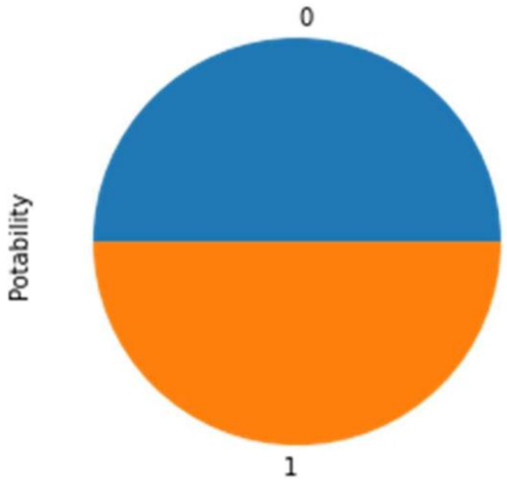


Class Imbalance

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



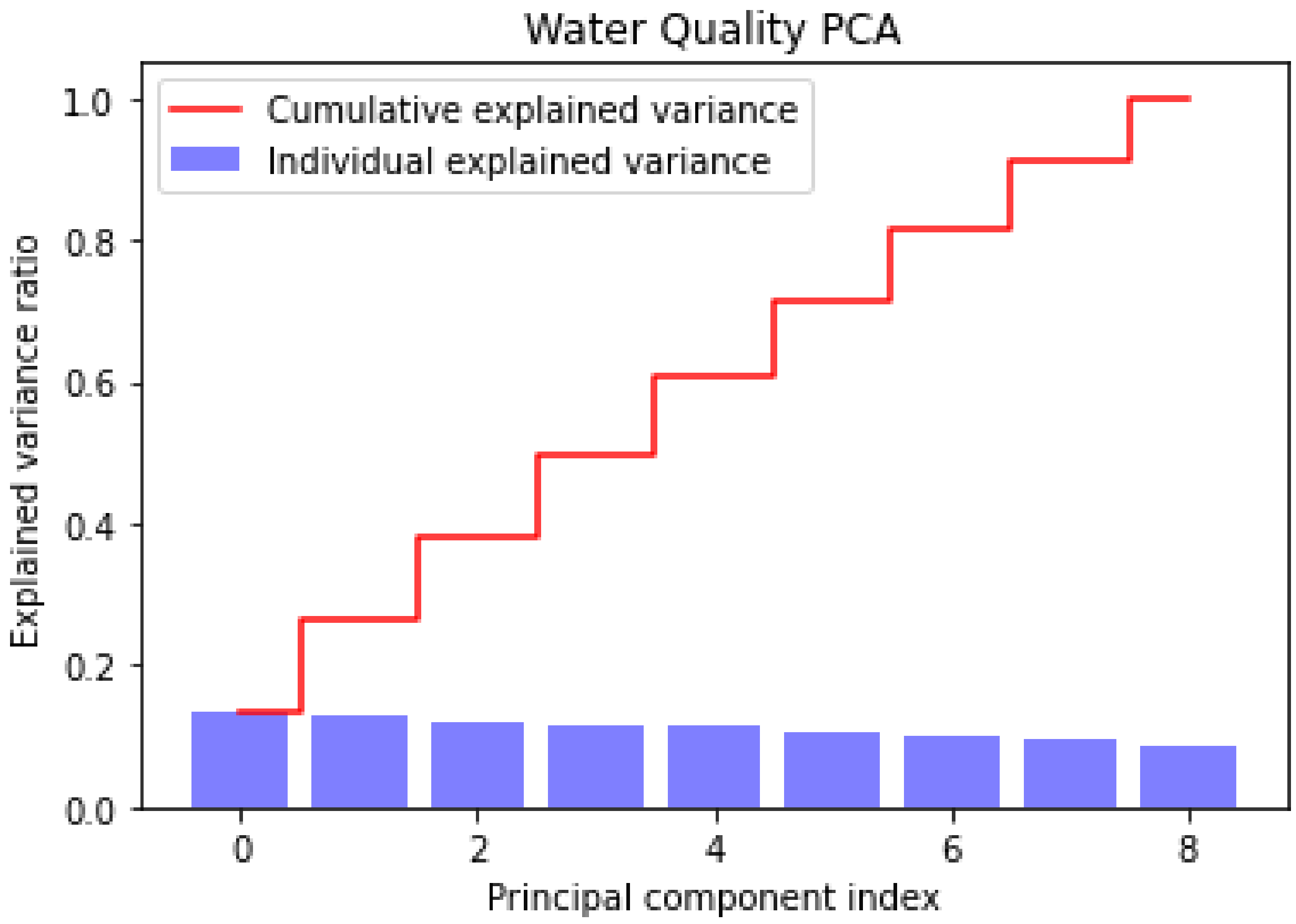
1930/1198



1930/1930

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

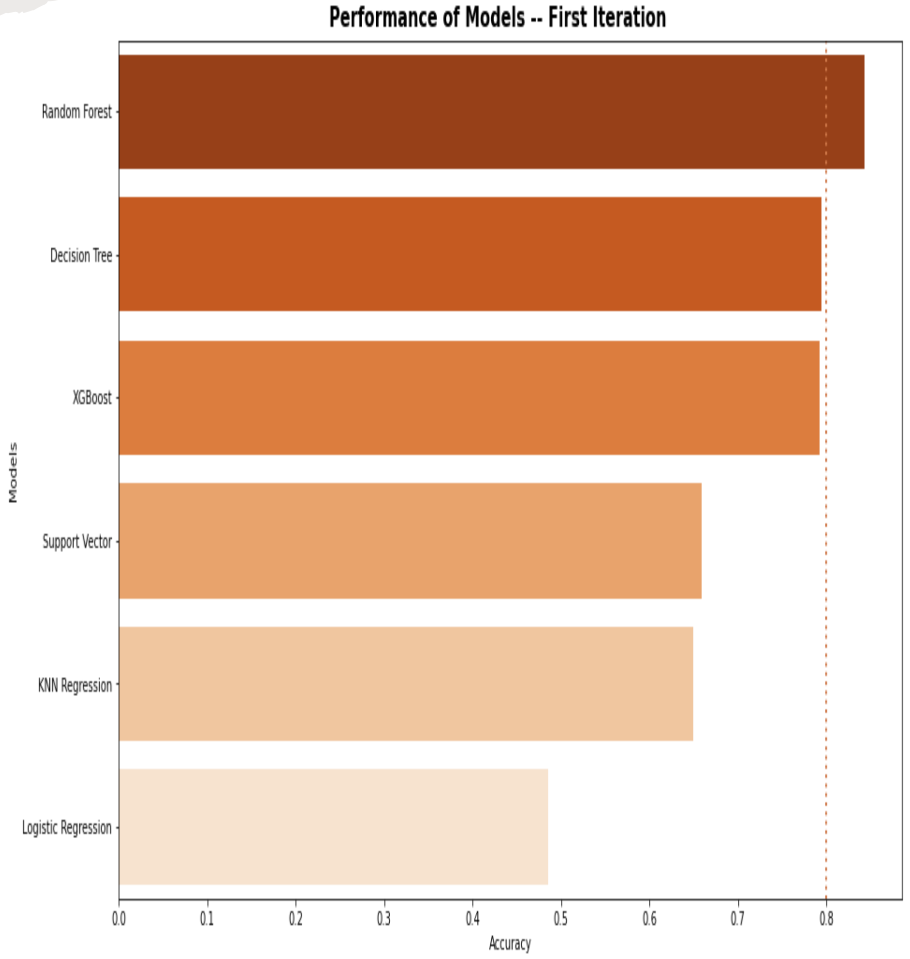


Modelling algorithms

- Logistic Regression
- K-Nearest Neighbour Regression
- Decision Tree Classifier
- Random Forest Classifier
- XGBoost Classifier

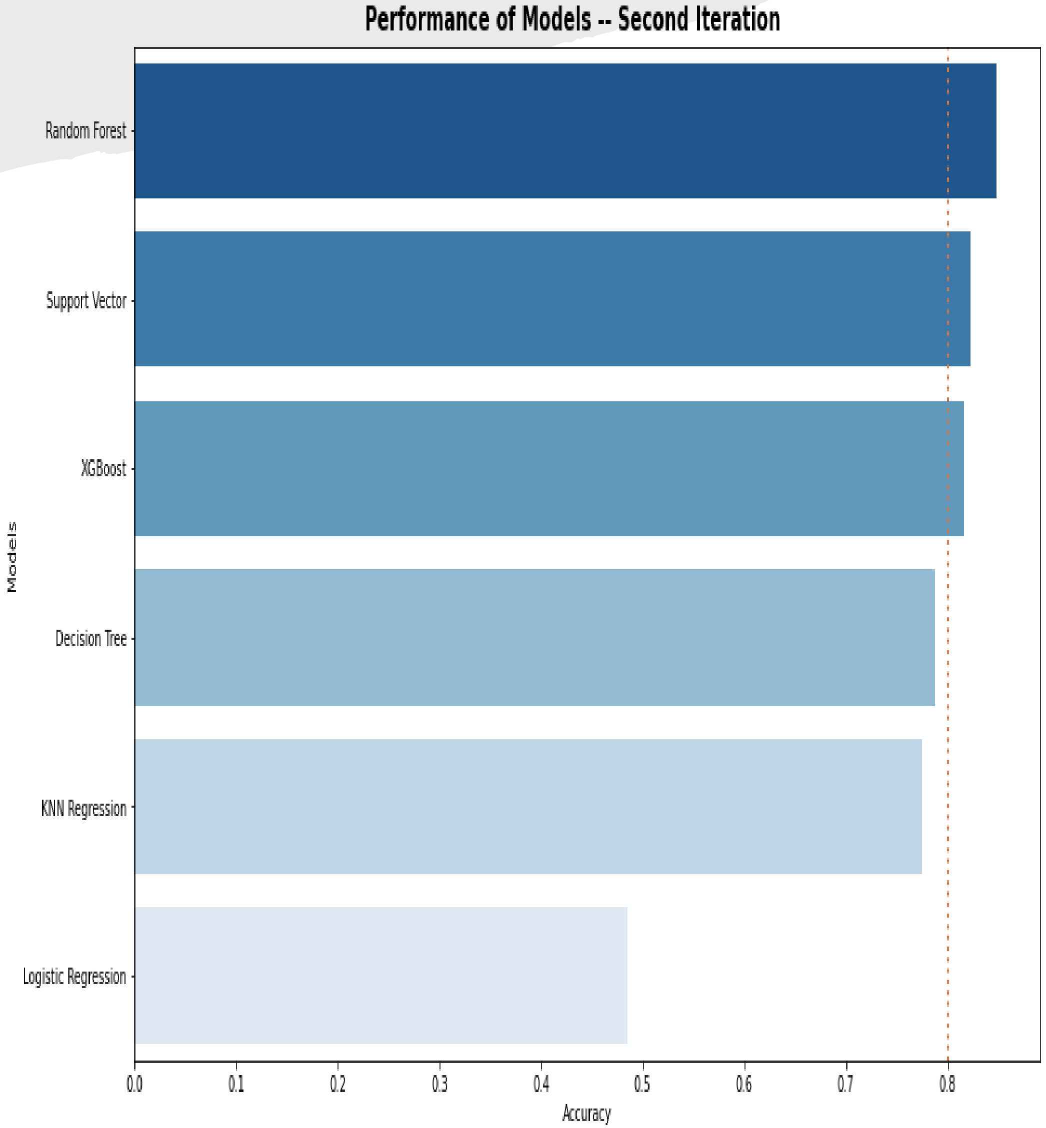
Algorithm Comparison 1st Iteration

Model	Accuracy	Precision	Recall	F1 Score	
3	Random Forest	0.843264	0.838275	0.836022	0.837147
2	Decision Tree	0.794041	0.738255	0.887097	0.805861
5	XGBoost	0.792746	0.754808	0.844086	0.796954
4	Support Vector	0.659326	0.640103	0.669355	0.654402
1	KNN Regression	0.648964	0.622871	0.688172	0.653895
0	Logistic Regression	0.485751	0.470998	0.545699	0.505604



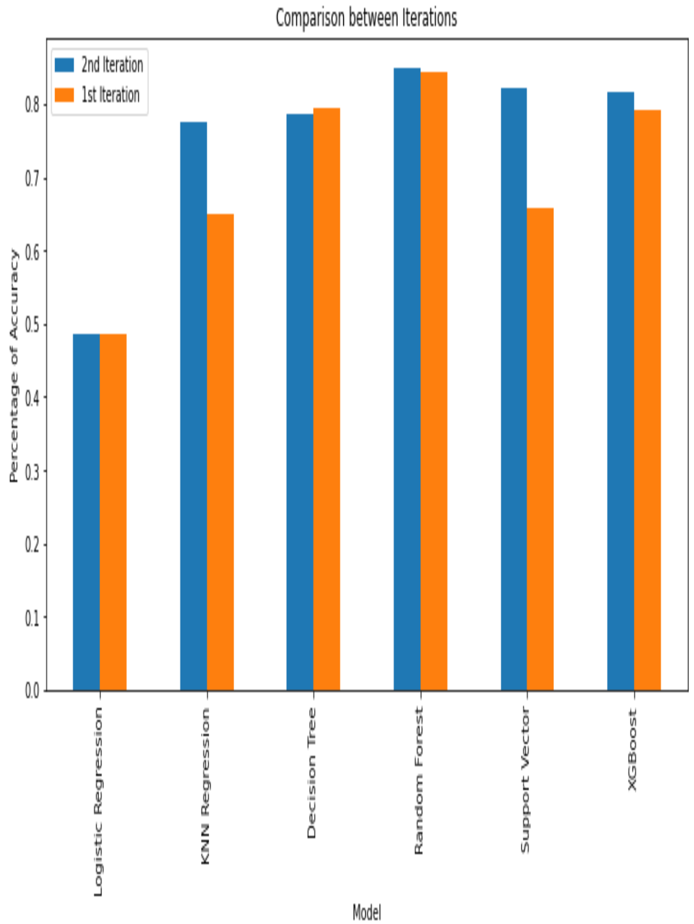
Algorithm Comparison 2nd Iteration

Model	Accuracy	Precision	Recall	F1 Score	
3	Random Forest	0.848446	0.840000	0.846774	0.843373
4	Support Vector	0.822539	0.844575	0.774194	0.807854
5	XGBoost	0.816062	0.780488	0.860215	0.818414
2	Decision Tree	0.787565	0.734234	0.876344	0.799020
1	KNN Regression	0.774611	0.722973	0.862903	0.786765
0	Logistic Regression	0.485751	0.470998	0.545699	0.505604

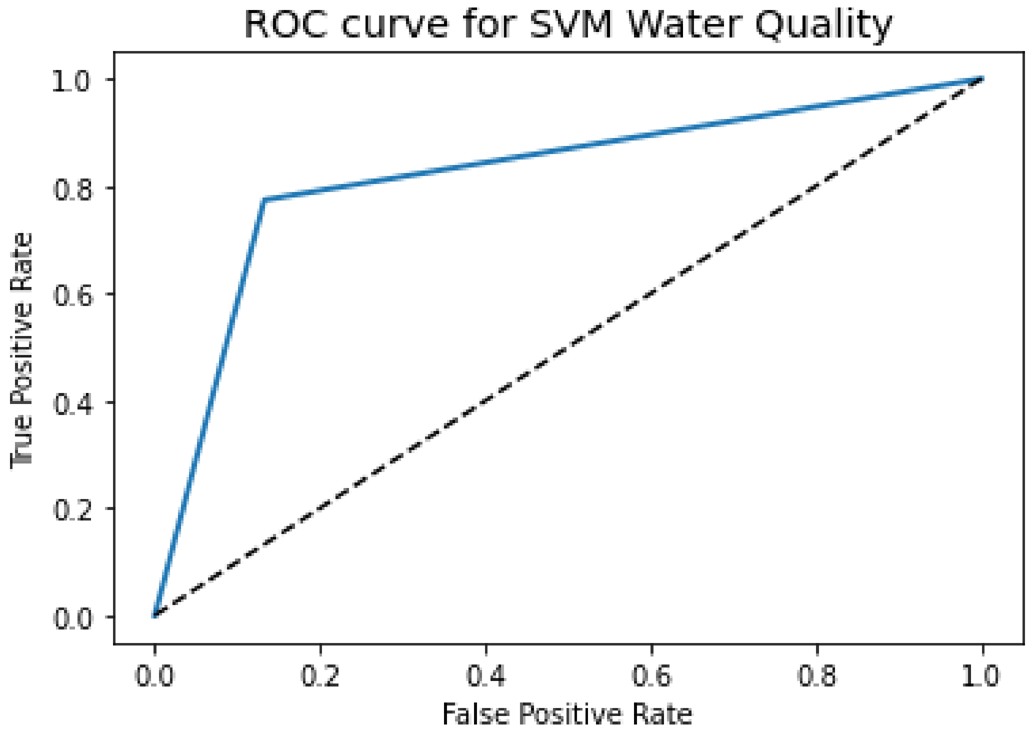


Model Evaluation

Model	2nd Iteration	1st Iteration	Difference in Accuracy
Logistic Regression	48.58%	48.58%	0.00%
KNN Regression	77.46%	64.90%	12.56%
Decision Tree	78.76%	79.40%	-0.65%
Random Forest	84.84%	84.33%	0.52%
Support Vector	82.25%	65.93%	16.32%
XGBoost	81.61%	79.27%	2.33%



Cross Validation



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

Interpretation & Recommendations

- After 2nd iteration and hyper-tunning parameters, SVM performed with the greatest accuracy 82.25%
- 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

Conclusions

Can we predict water potability?	✓ Using ML it is possible to predict water potability
Which machine learning algorithms can yield the most efficient and accurate results?	✓ Support Vector Machine Classifier best performance 87.98% accuracy
Can the parameters within the ML algorithms be tuned to yield the best results?	✓ Hyper-tunning did increased accuracy in modeling for most of the algorithms
Are the parameters within the dataset affective in water quality prediction?	✓ The parameters within the dataset were affective in prediction although had low correlation
Should there be other parameters to consider?	✓ Through research from other studies, additional attributes such as coliform and heavy metals should be included
How confident are we in our findings?	✓ Confident in our findings but room for improvement


ANY QUESTIONS?