Predicting Water Potability Using Statistical Algorithms

Abhishek Anand, Pratapaditya Ghosh, Devashri Karve, Tanay Kulkarni
36-600: Overview of Statistical Learning and Modeling, Fall 2021

INTRODUCTION AND MOTIVATION

- Access to safe drinking water is essential for human health
- Potability (drinkability) of water depends on several factors

Can we construct a classification model to determine the potability of water by using different water quality parameters?

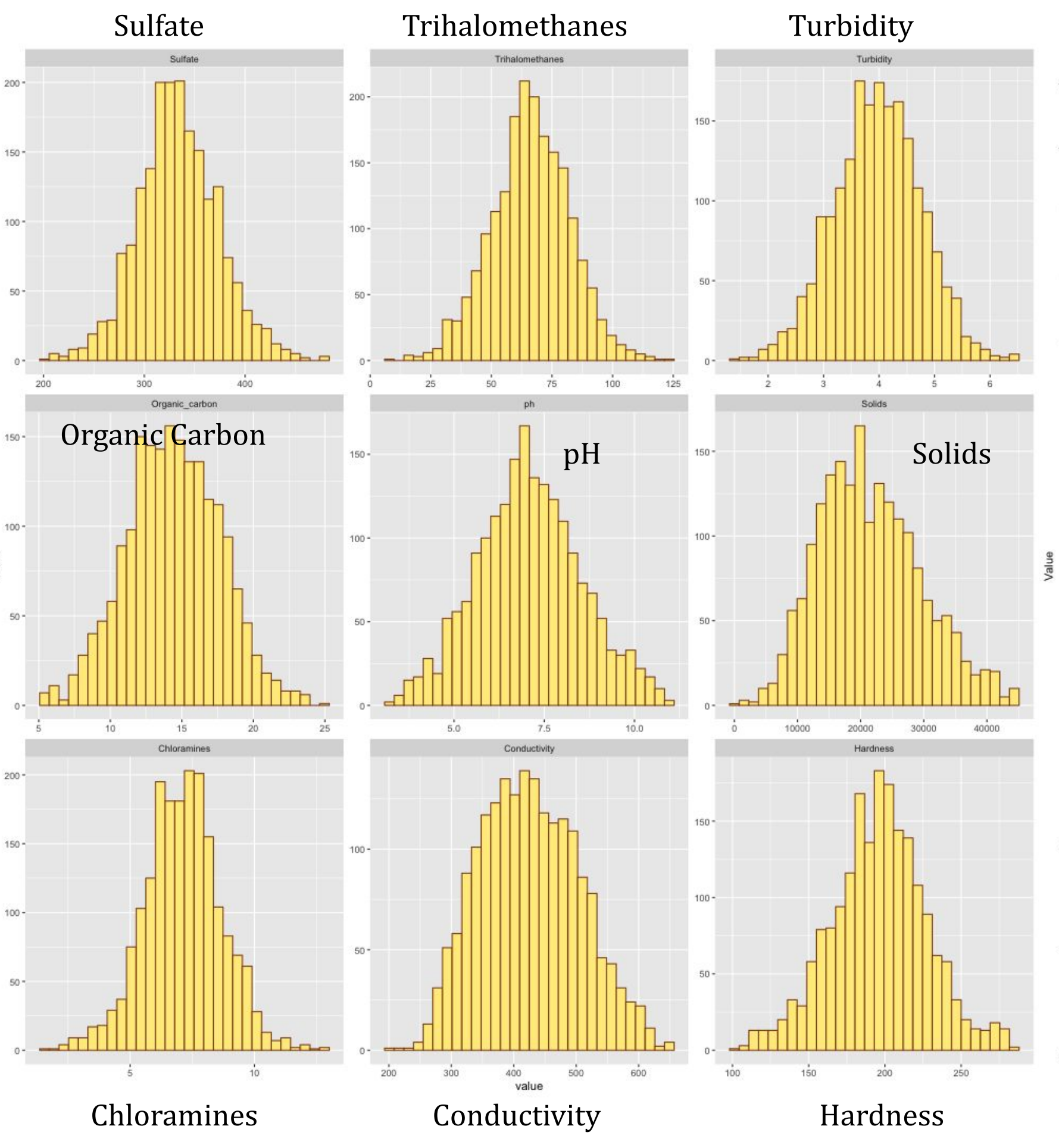
METHODS

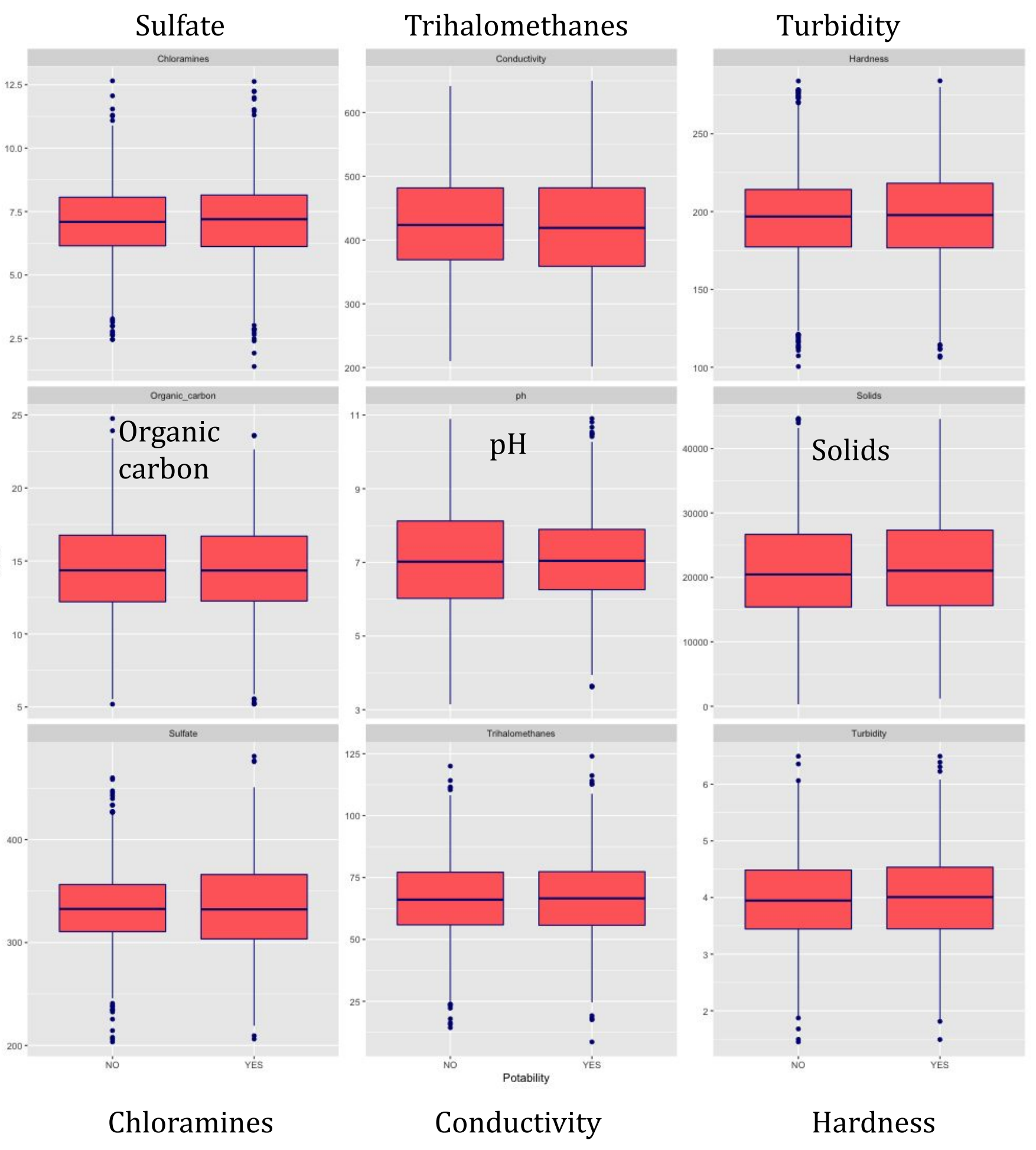
- We use various models for binary classification of water samples into potable and non-potable using nine independent variables.
- The data was split into 'train' (70%) and 'test' (30%), where train was used for developing the models and test for validating its performance.
- Models used for the analysis are decision tree (DT), best linear model (BLM), logistic regression (LR), random forest (RF), support vector machine (SVM), XGBoost (XGB), and k-nearest neighbors (KNN).
- ROC curves are receiver operating characteristics curves that illustrate the tradeoff between classifying members of each class well. The area under a ROC curve is dubbed AUC; model with the highest AUC value is selected as the best model.
- We maximize Youden's J statistic to generate class predictions.
- Finally, the prediction values are compared with the response to derive confusion matrix.
- Elements of the confusion matrix can be used to estimate the probability of model making prediction errors.

EXPLORATORY DATA ANALYSIS (EDA)

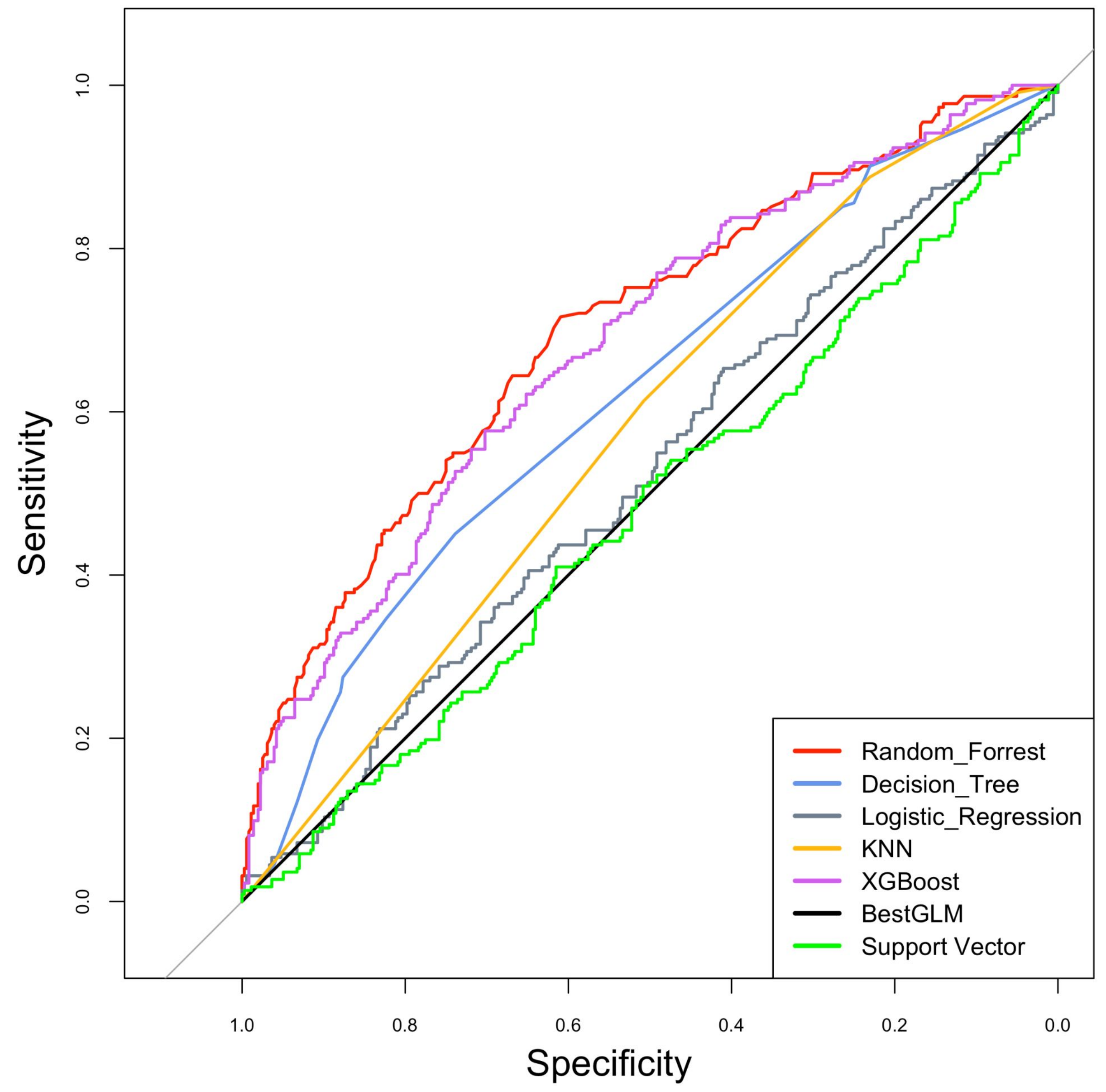
Our dataset contains nine dependent variables (water quality parameters) such as pH, turbidity, solids, etc (shown in the table below). Potability of water depends on these parameters. We prepare histograms of these parameters and find that all of them are normally distributed. Then we remove some outliers and prepare boxplots showing range of values of different parameters for which the water is potable. However, visually we do not find any clear difference between parameter ranges for potable and non-potable water.

Sulfate	Trihalomethanes	Turbidity	Organic_Carbon	pH
Independent	Independent	Independent	Independent	Independent
Solids	Chloramines	Conductivity	Hardness	Potability
Independent	Independent	Independent	Independent	Dependent





RESULTS



- ROC curves for different models is shown here
- Random Forest Algorithm has highest AUC (area under curve)

Models	DT	GLM	LR	RF	SVM	XGB	KNN
AUC	0.612	0.5	0.523	0.701	0.481	0.682	0.577

CONFUSION MATRIX FOR RANDOM FOREST

	Response = YES	Response = NO
Prediction = YES	159	139
Prediction = NO	63	217

- The misclassification rate is 0.349, i.e., the RF model makes flawed prediction 34.9% of times.
- There are 139 cases out of 578 (24%) when the model labels non-potable water as safe for drinking.
- 63 times out of 578 potable water is labelled as unsafe.

CONCLUSIONS

In this project, we trained seven classification models in an attempt to determine water potability given water properties. We find that our random forest model performs best among all models, with an area under curve of 0.701 and a misclassification rate of 34.9%. These numbers indicate that potability prediction is a difficult problem and that perhaps other water property measurements are needed to assess potability.