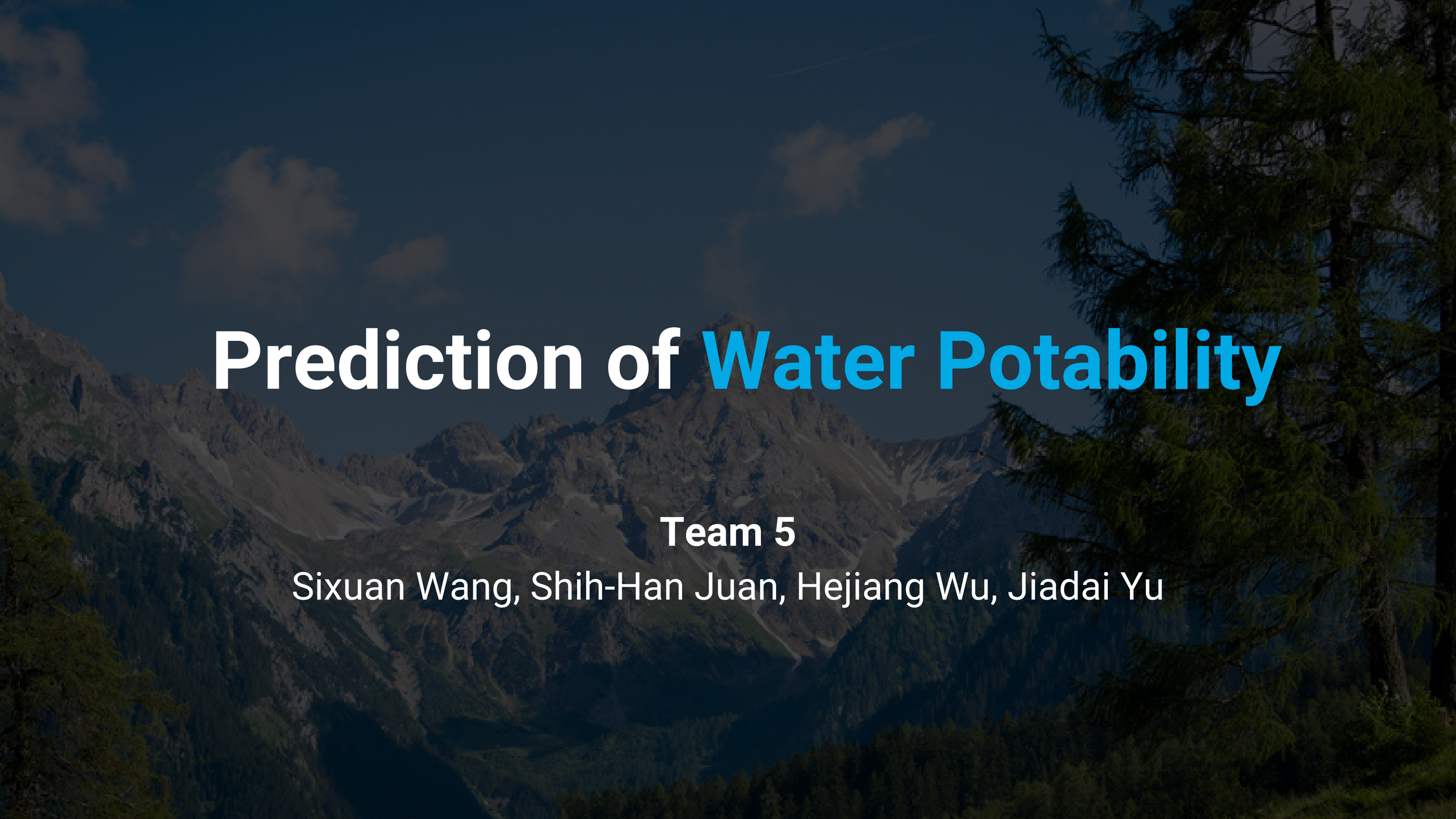


Prediction of Water Potability

Team 5

Sixuan Wang, Shih-Han Juan, Hejiang Wu, Jiadai Yu

Colab Link:

https://colab.research.google.com/drive/1ZPl5-cNtM4oPN1sOysRc9Tqj6xR17wf2?usp=sharing

AGENDA

1. Context and Motivation
2. Dataset
3. Data Cleaning and Exploration
4. Model Comparison
5. Challenges and Solutions
6. Conclusion

Prediction of Water Potability

Objective:
To develop a best-performing ML model to predict water potability

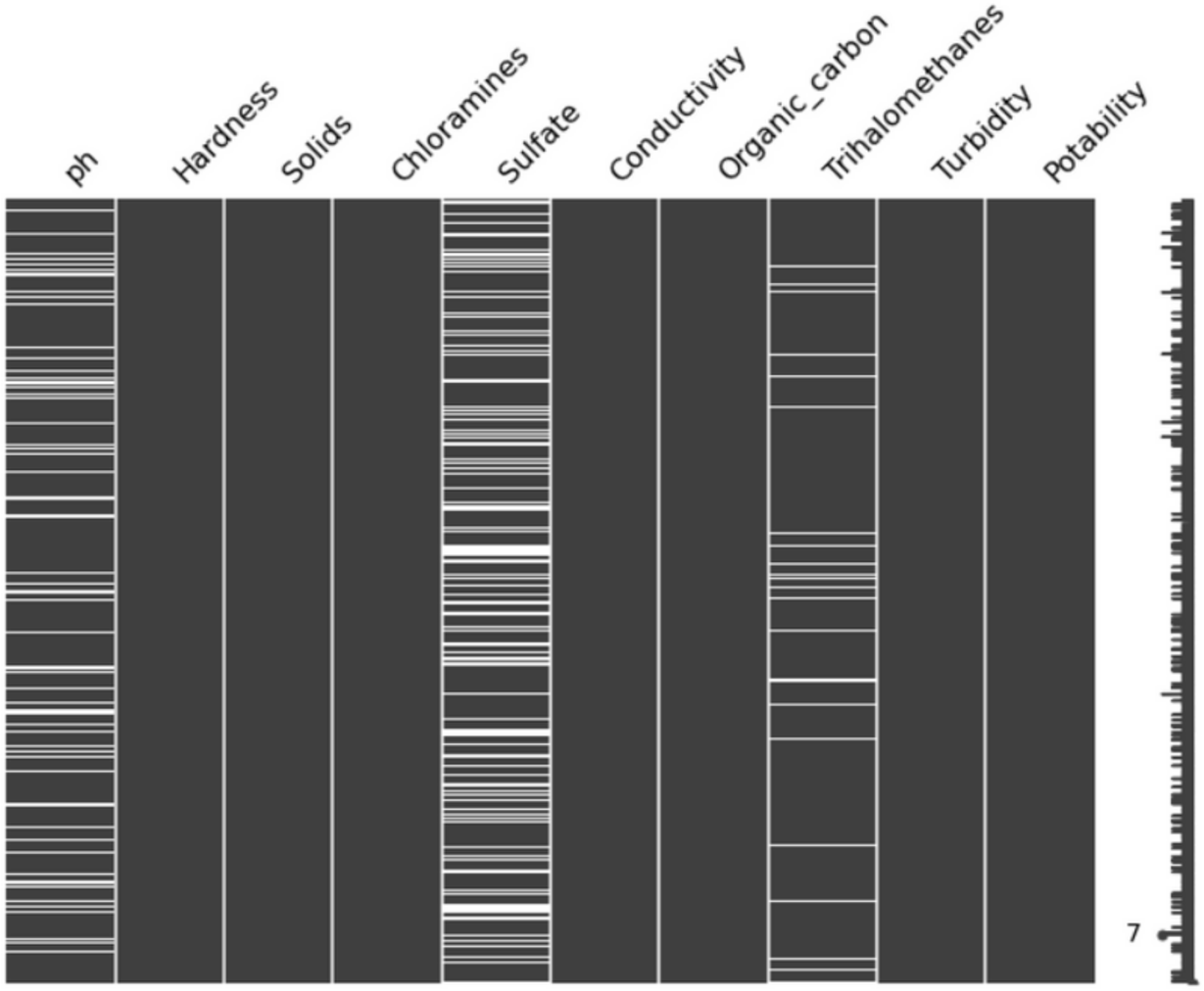
Motivation:
A model as such would serve as an economical approach to screen qualified water sources

Dataset Overview

- Data Source: Kaggle
- 3276 rows, 10 columns
- 1 target (binary), 9 predictors (numeric)

variable	description
Potability	Indicates if water is safe for human consumption where 1 means Potable and 0 means Not potable.
ph	Indicates acid–base balance of water. Maximum permissible limit of pH ranges from 6.5 to 8.5.
Hardness	Defined as the capacity of water to precipitate soap caused by Calcium and Magnesium.
Solids	Total dissolved solids. Desirable limit for TDS is 500 mg/L and maximum limit is 1000 mg/L.
Chloramines	Major disinfectants used in public water systems. Chlorine levels up to 4 milligrams per liter are considered safe in drinking water.
Sulfate	Naturally occurring substances found in minerals, soil, and rocks. It ranges from 3 to 30 mg/L in most freshwater supplies.
Conductivity	Measures the ionic process of a solution that enables it to transmit current. Should not exceeded 400 µS/cm by WHO standards.
Organic_carbon	Comes from decaying natural organic matter and synthetic sources. < 2 mg/L in treated / drinking water; < 4 mg/Lit in source water used for treatment.
Trihalomethanes	Chemicals which may be found in water treated with chlorine. Up to 80 ppm is safe in drinking water.
Turbidity	The turbidity of water depends on the quantity of solid matter present in the suspended state. WHO recommended value is 5.00 NTU.

Handling Missing Values



- 491 missing values in **ph**
- 781 in **Sulfate**
- 162 in **Trihalomethanes**
- Dropping all rows with missing values will result in losing 39% of the records

Dropping Rows ➤ **Smaller Dataset**

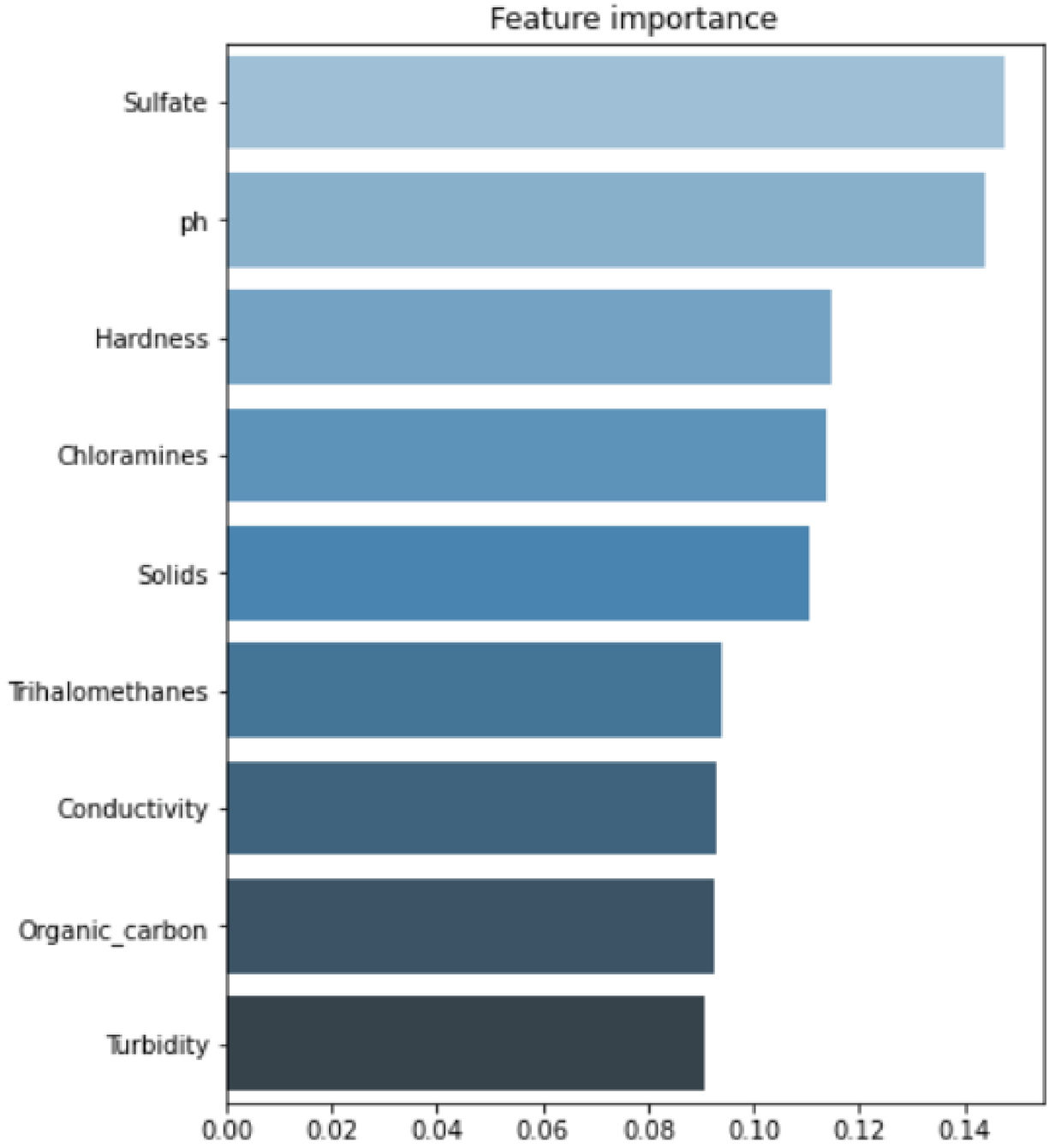
Dropping Columns ➤ **Info Loss**

Imputing ➤ **More Bias**

Handling Missing Values

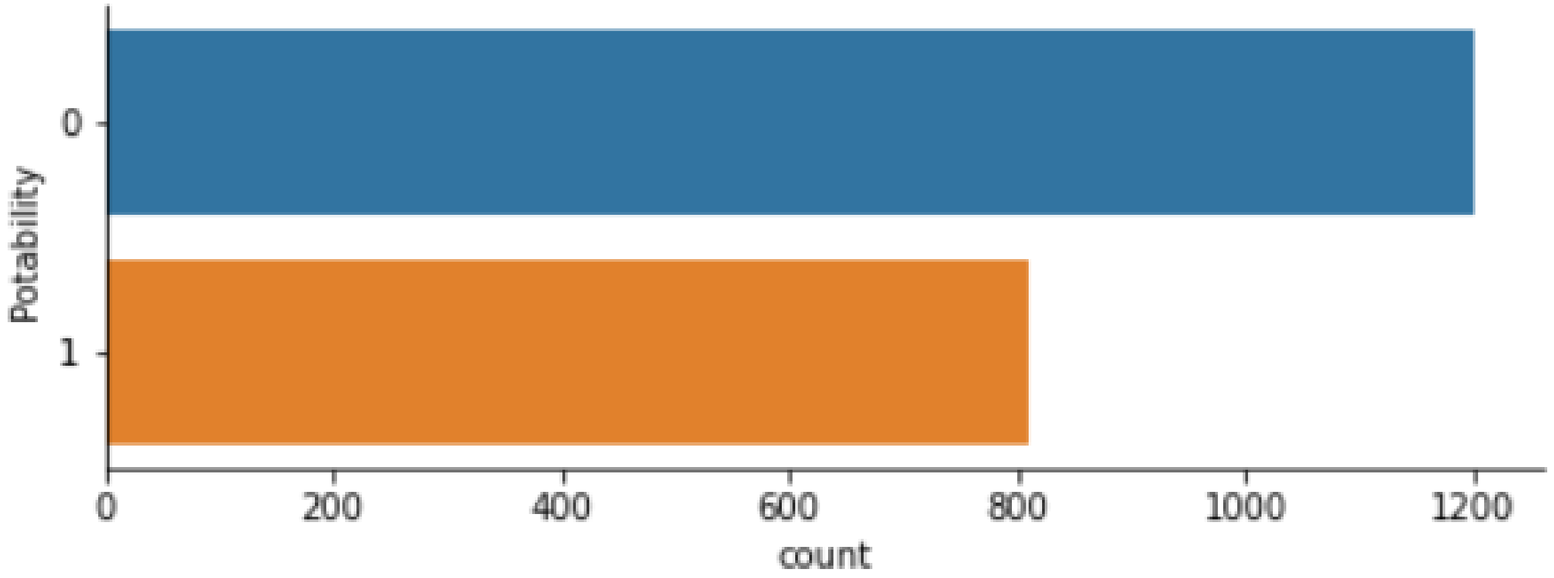
☑ **Drop Rows** (2011 rows left)

- Sulfate and ph are two main contributors to water potability
- Remain conservative about water potability (reduces false positive predictions)



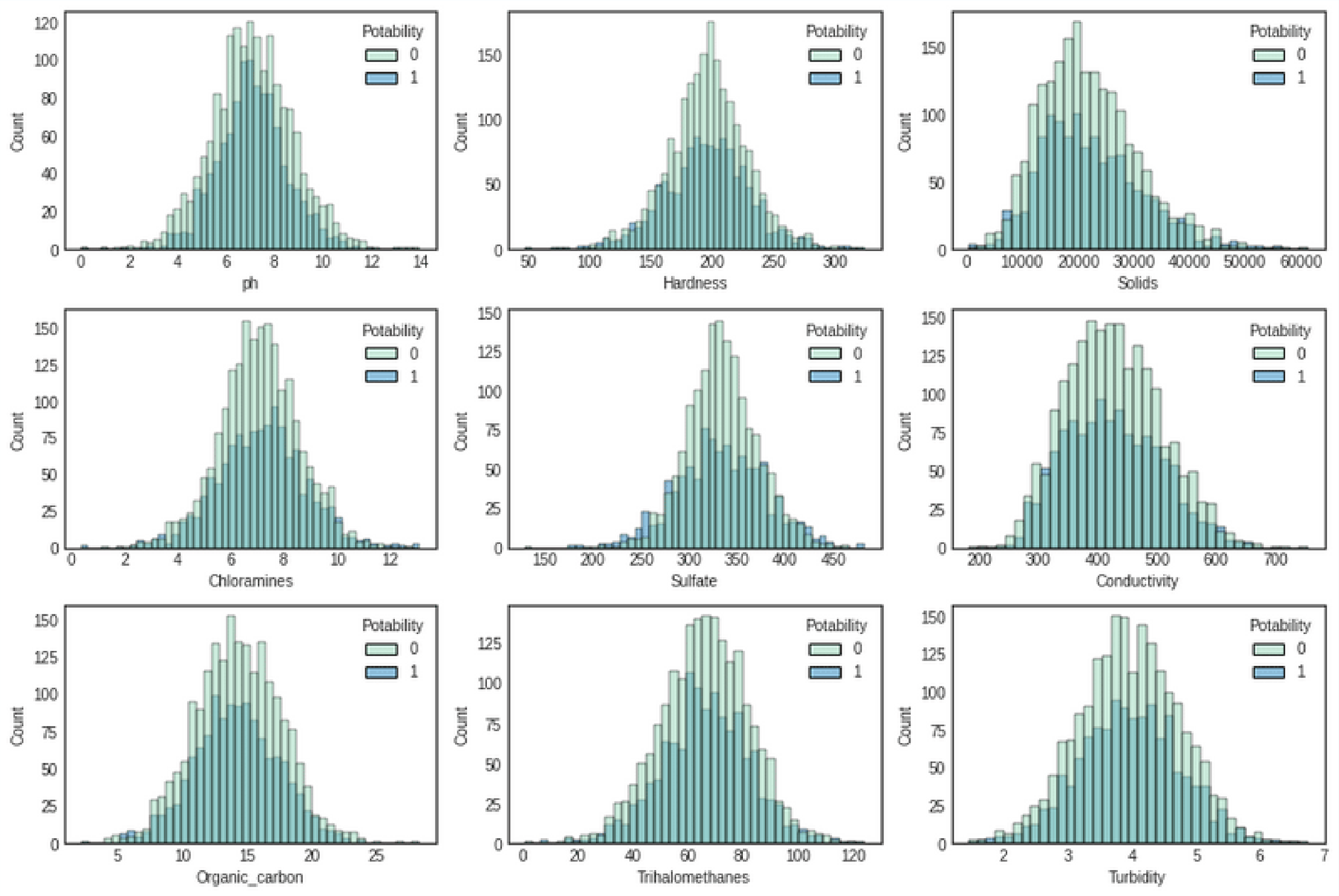
Distribution of Target Variable

- 811 records of potable water (40%)
- 1200 records of impotable water (60%)
- Relatively balanced

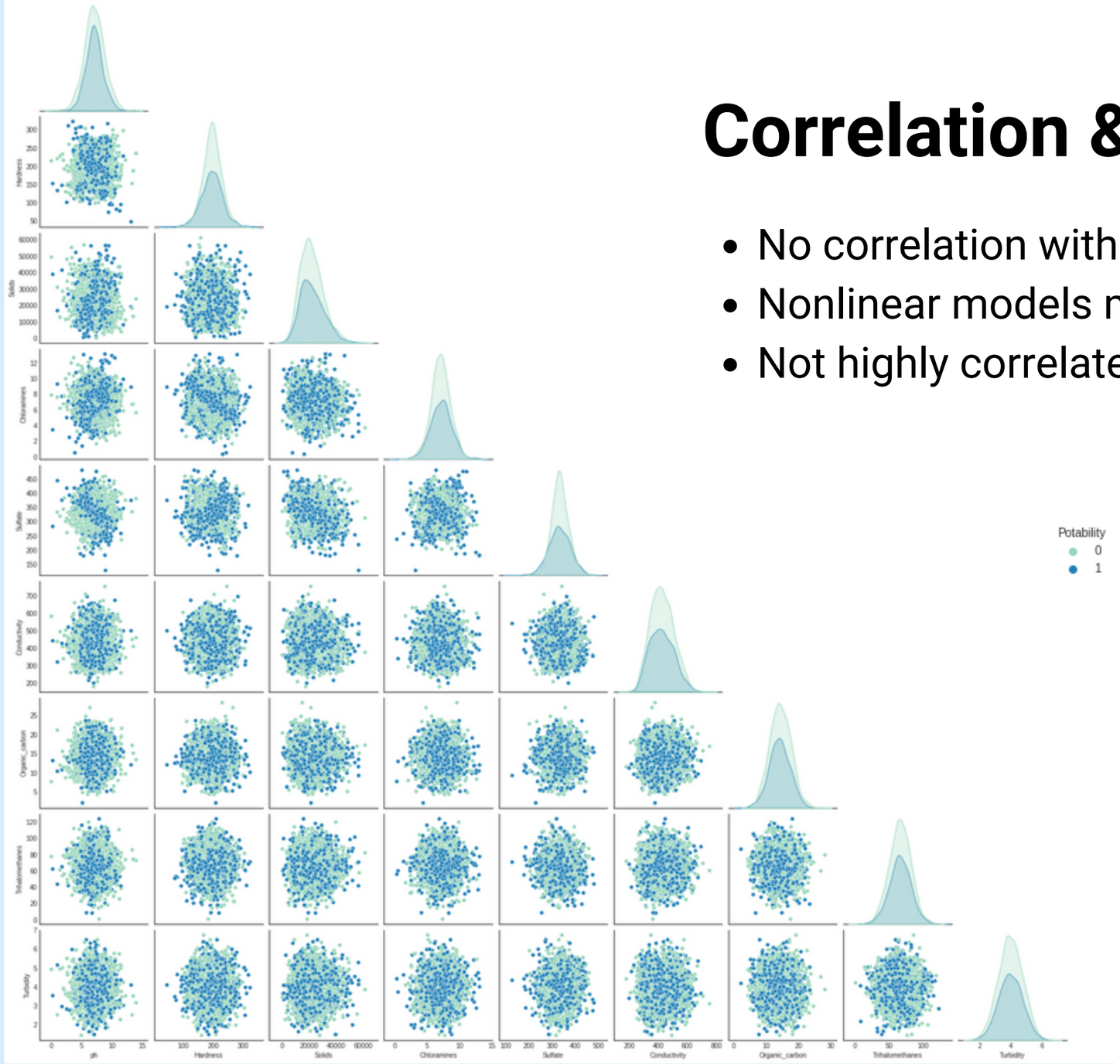


Histogram

The data is almost normally distributed.



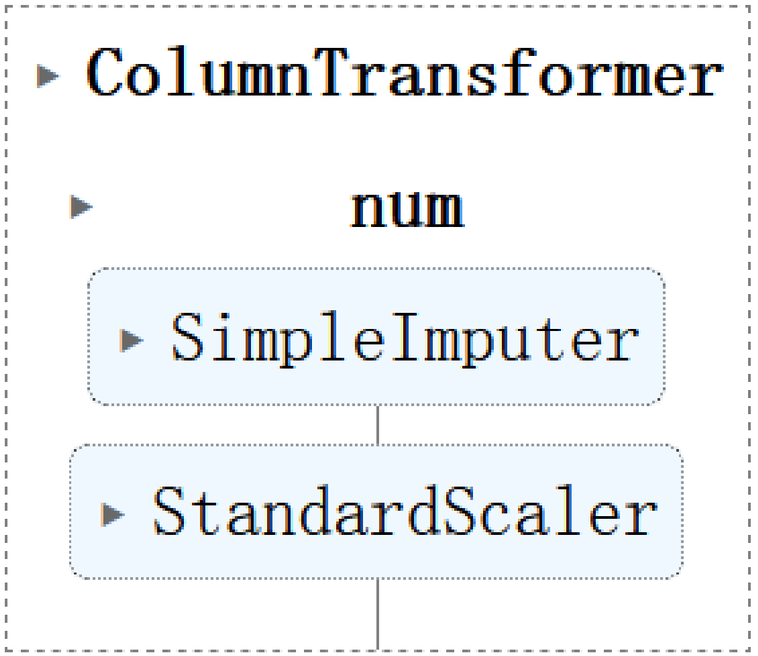
Correlation & Pairplot



- No correlation with the target
- Nonlinear models may perform better than linear models
- Not highly correlated -> No collinearity

	Potability
ph	-0.003556
Hardness	-0.013837
Solids	0.033743
Chloramines	0.023779
Sulfate	-0.023577
Conductivity	-0.008128
Organic_carbon	-0.030001
Trihalomethanes	0.007130
Turbidity	0.001581
Potability	1.000000

Pipeline



- Impute strategy: median
- StandardScaler
- Test size 20%

Which model will win?

Logistic Regression

KNN

SVM

Decision Tree

Random Forest

Ensemble-vote

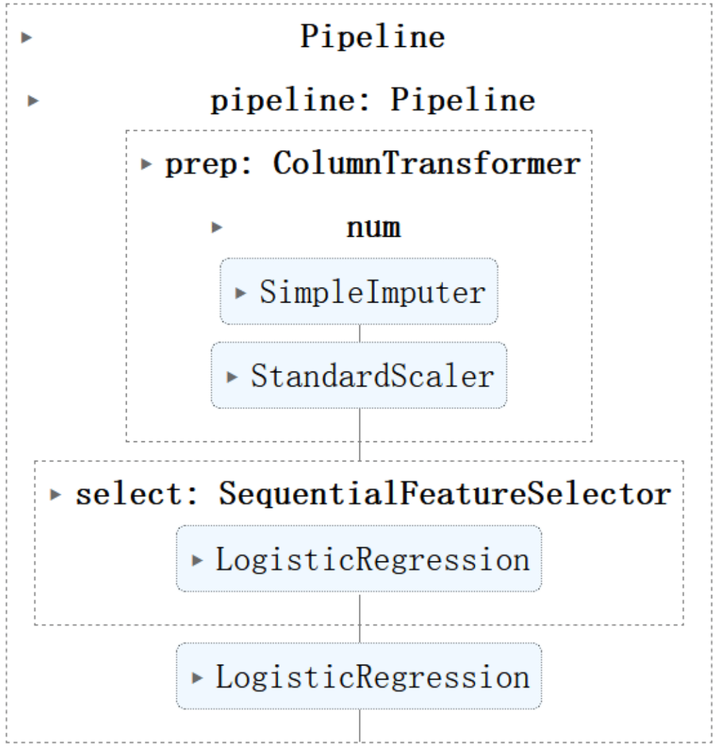
Ensemble-stack

Logistic Regression

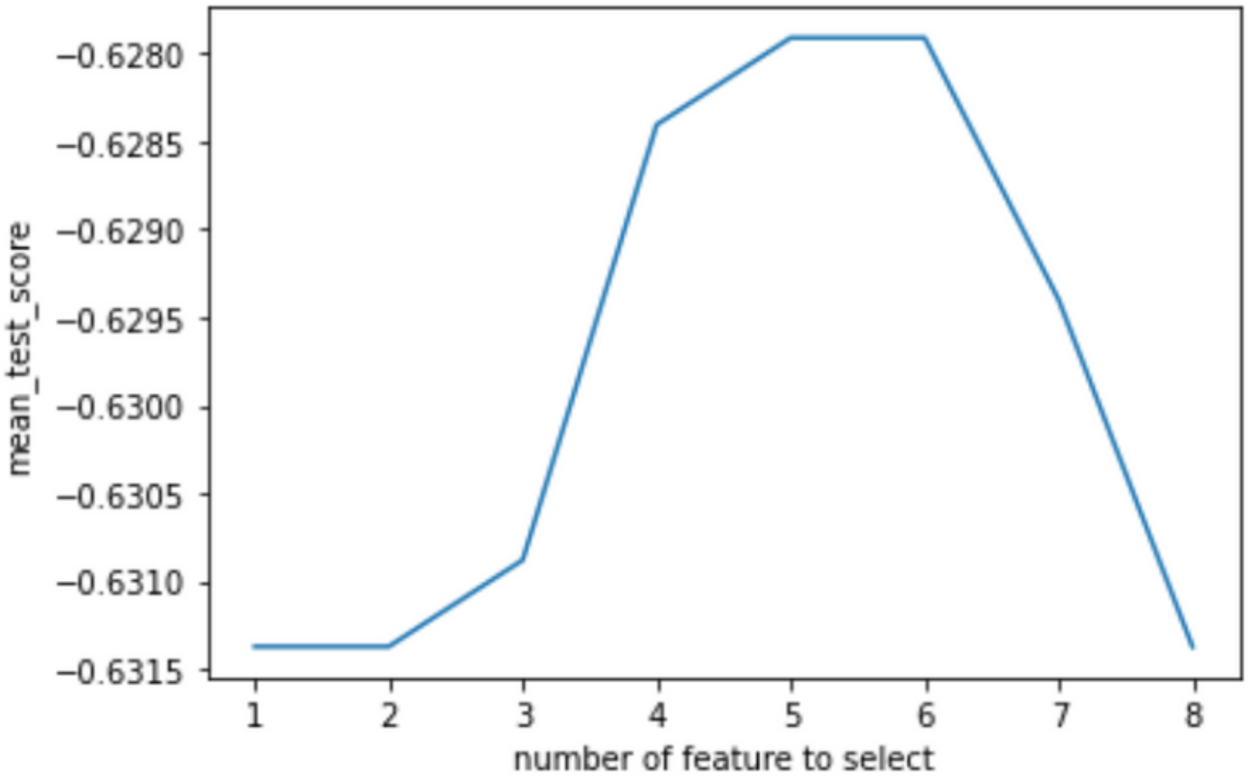
- Logistic Regression with all features:

Test accuracy: 57.07%

- Grid Search:



- mean test score is maximized to 62.79% when 5 features are selected

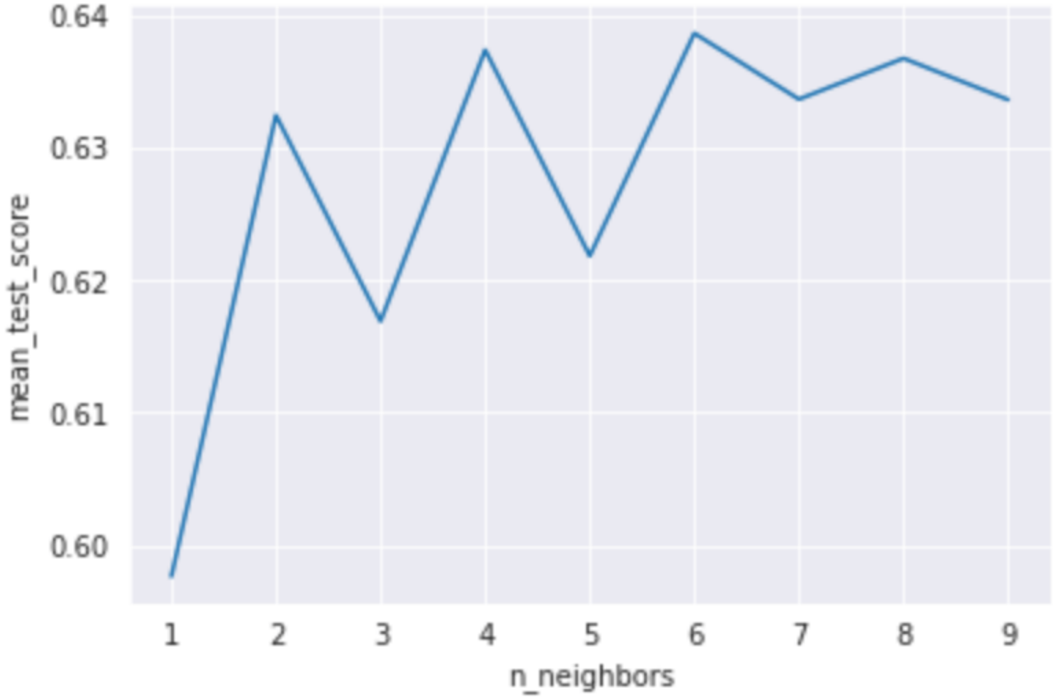


Naïve Bayes - Not Applicable

- usually performs better with a small dataset, while our dataset has more than 2000 records.
- Naïve Bayes Assumption: independence among features, but according to natural science principles, these indicators of water quality are interrelated.

KNN

- Grid Search
- n_neighbors = 6



Test accuracy: 63.77%

SVM

- Halving Random search
- Kernel
  - Linear
  - polynomial
  - Radial
- Cost

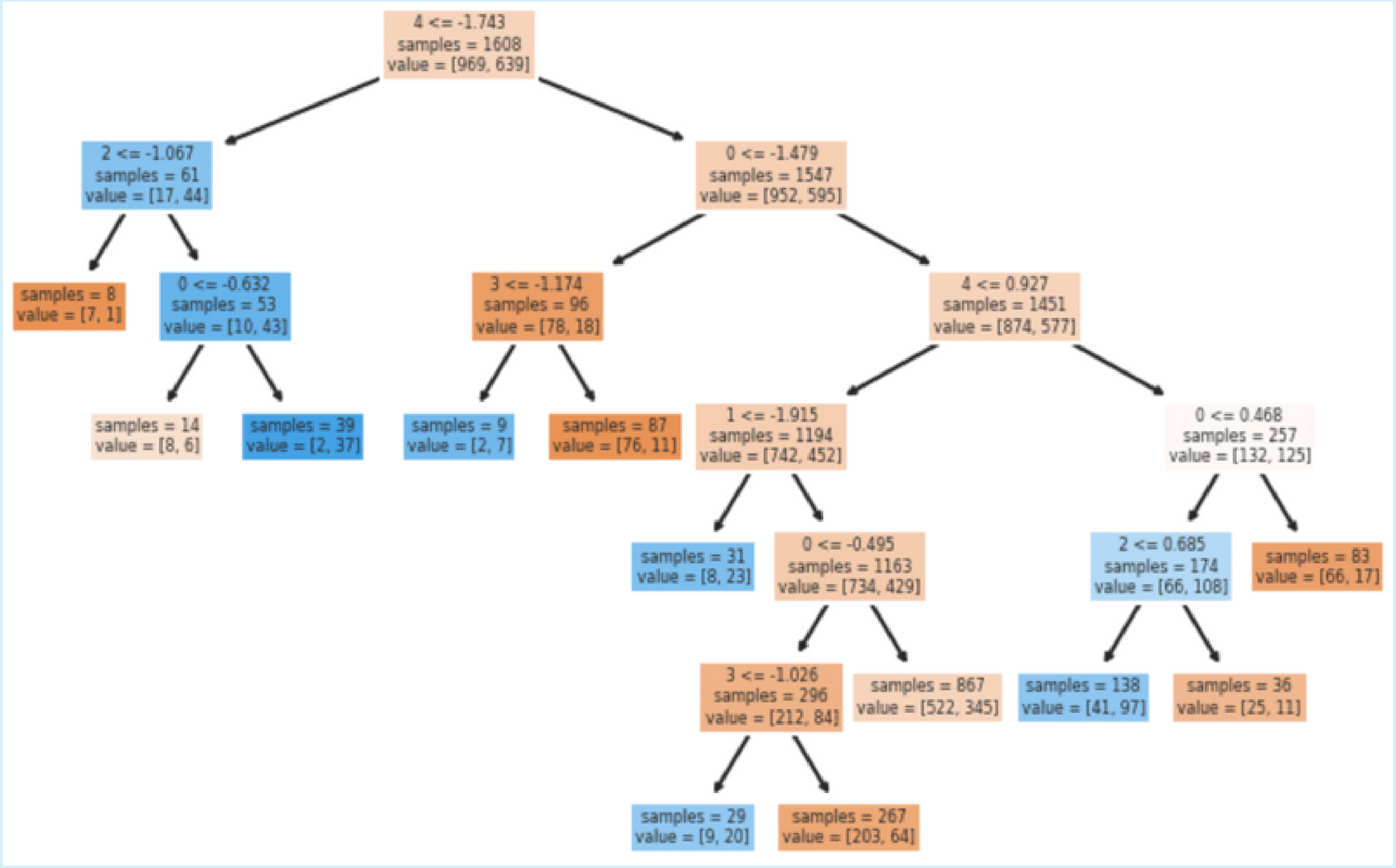
iter	n_resources	param_svm__C	param_svm__kernel	mean_test_score
2	1602	1	rbf	0.633558

Test accuracy: 68.24%

Decision Tree

- Grid Search
- ccp_alpha = 0.003017

Test accuracy: 64.26%



Random Forest

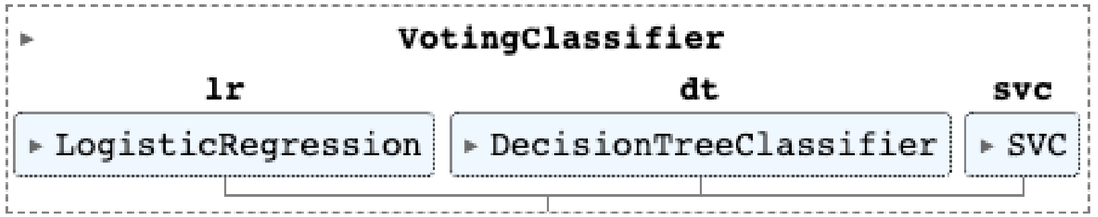
- Grid Search

- n_estimators
- min_samples_leaf

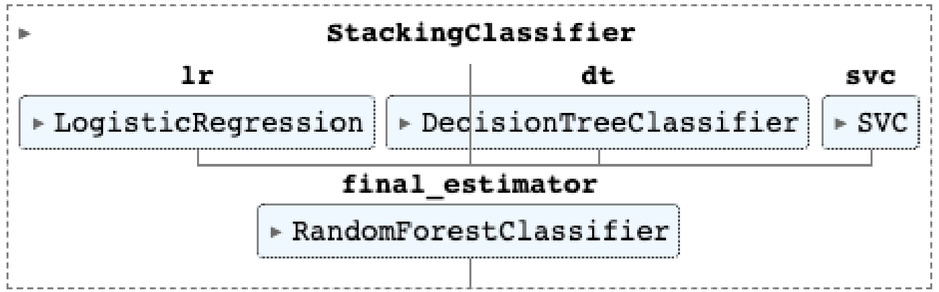
param_min_samples_split	param_n_estimators	mean_test_score
5	200	0.619415

Test accuracy: 67.49%

Ensemble

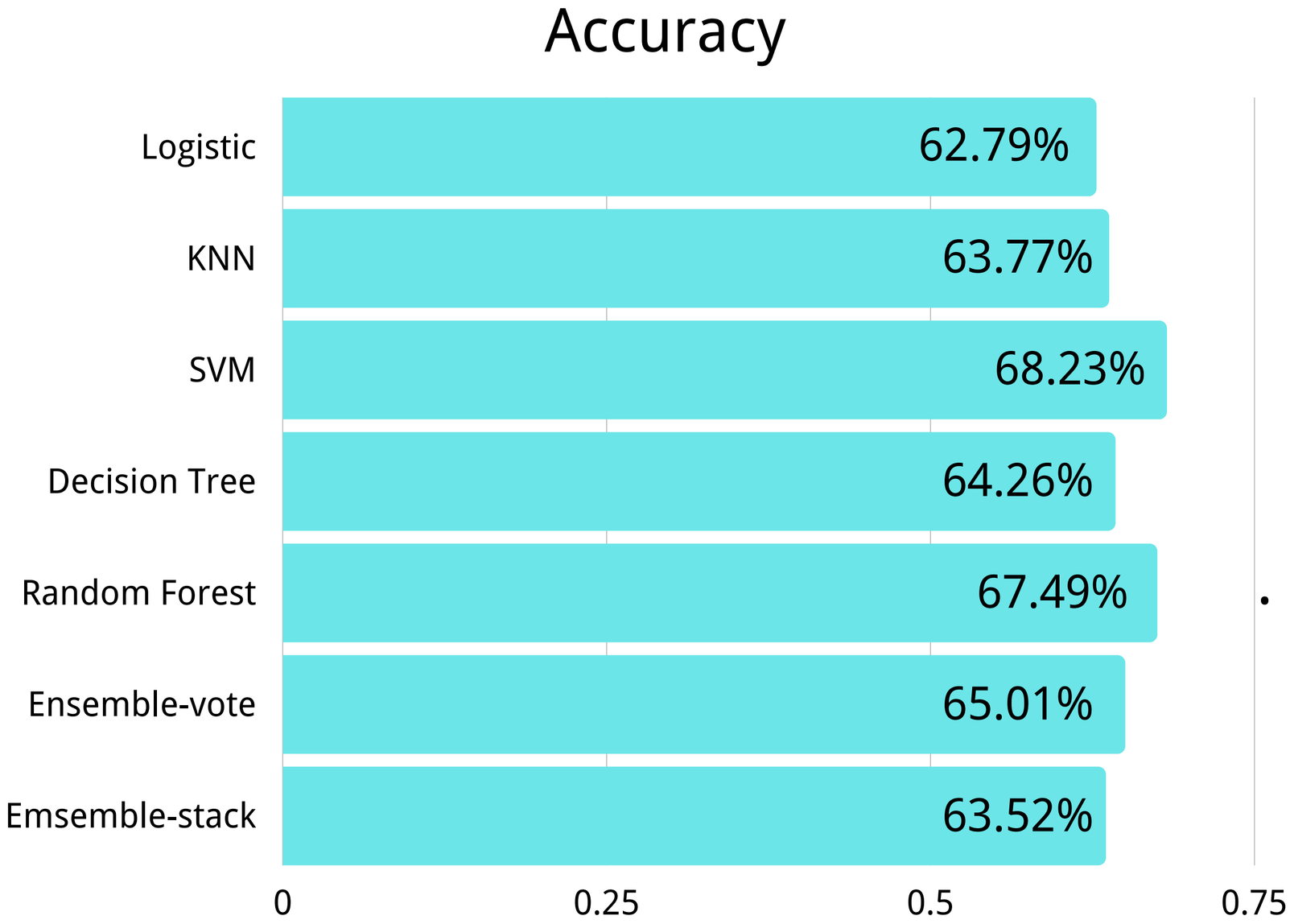


- Test accuracy: 65.01%



- Test accuracy: 63.52%

Model Selection



SVM

- Perform well on high-dimensional data with non-linear relationship
- Avoid overfitting problem

Cost Matrix

		PREDICTED CLASS	
		Class=Yes	Class=No
ACTUAL CLASS	Class=Yes	a (TP)	b (FN)
	Class=No	c (FP)	d (TN)

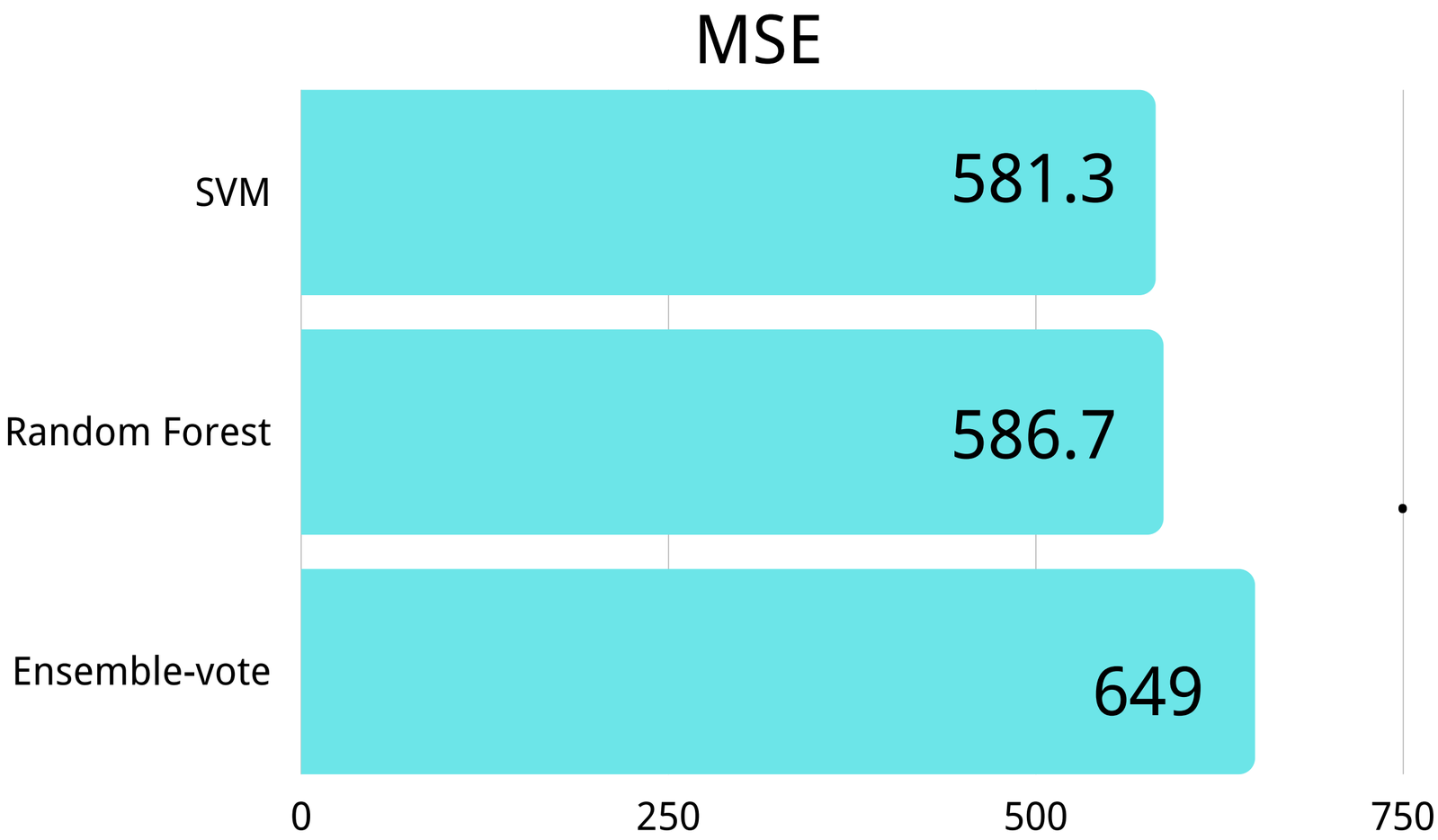
```
def default_cost(y_true, y_pred):
    cm = confusion_matrix(y_true, y_pred)
    return cm[1,0] * 4 + cm[0,1] * 1
```

Goal: Increase the criteria of potable when we select water samples.

FP cost:4

FN cost:1

Model Selection

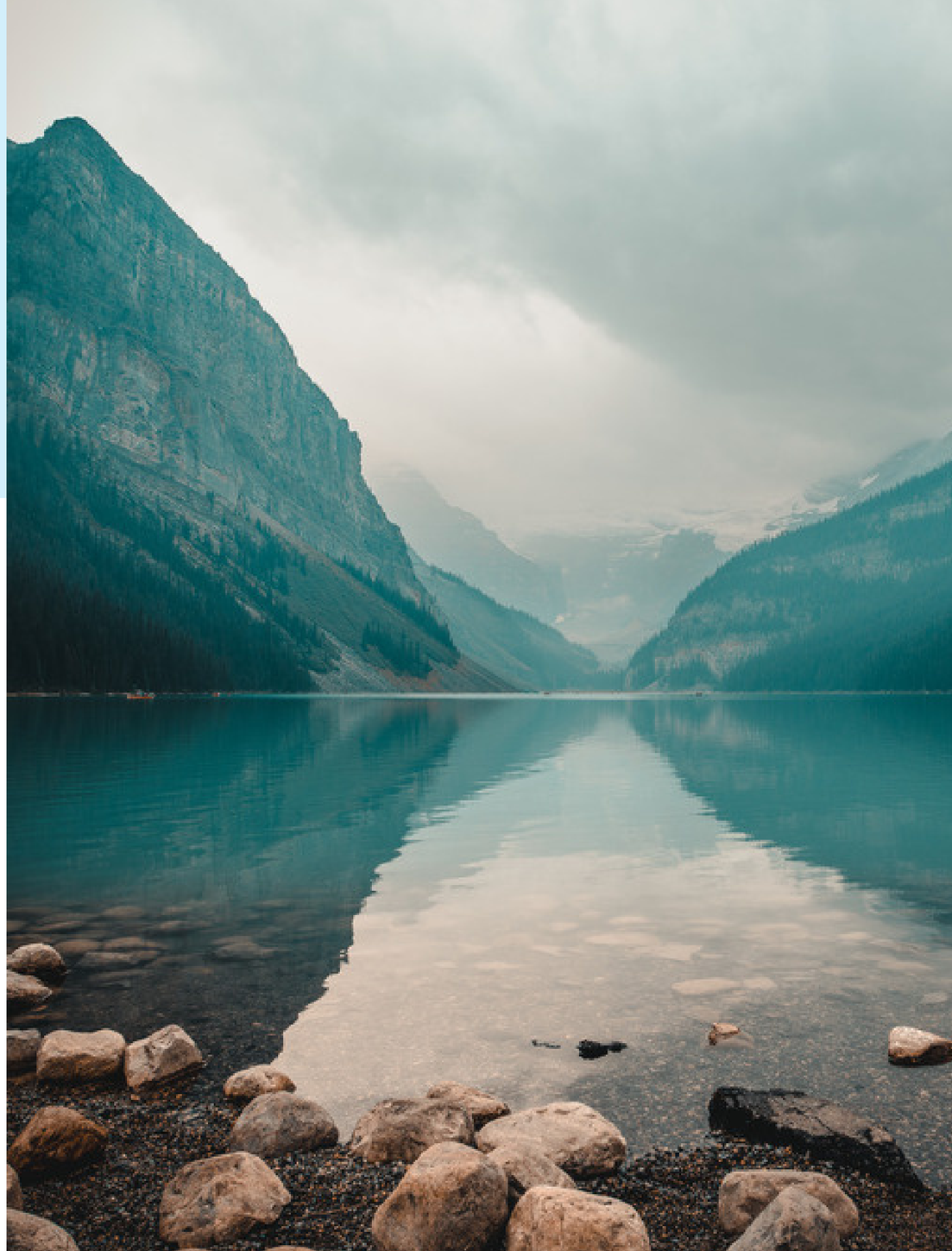


SVM

- SVM is still our best model!

Challenge and Conclusion

- [x] Best model: SVM
- [x] Challenge: Combine the model with the real-world case
- [x] Solution: Case by case study



Reference

1. http://www.mwra.com/water/html/awqr.htm
2. Drinking Water Frequently Asked Questions (FAQs) | Drinking Water | Healthy Water | CDC
3. International Decade for Action on Water for Sustainable Development, 2018-2028 (un.org)
4. https://www.kaggle.com/datasets/adityakadiwal/water-potability
5. https://www.epa.gov/ground-water-and-drinking-water/national-primary-drinking-water-regulations


Thank You