Water potability

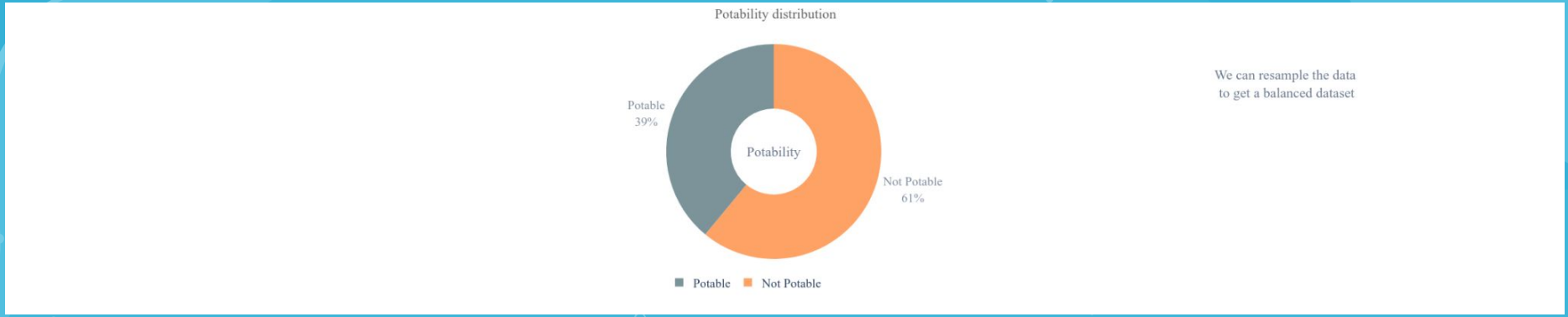
Lorenzo Arcioni 07/2023

Problem presentation

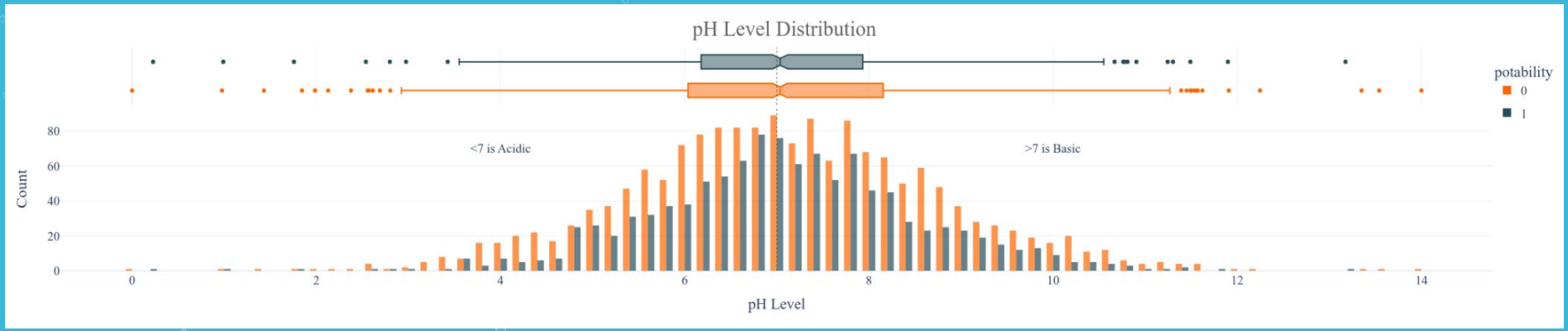
- Problem Introduction:
  - Water potability is a crucial concern for public health, as access to safe and drinkable water is essential.
  - The analysis aims to investigate and understand factors influencing water potability.
- Available Data:
  - The dataset consists of water quality measurements collected from various sources.
  - Features include pH levels, hardness, solids concentration, chloride content, and more.
  - The dataset also contains a target variable indicating the potability of water samples.
- Objectives:
  - Identify key factors impacting water potability based on the available data.
  - Develop models to predict water potability using the given features.
  - Provide insights and recommendations to improve water treatment processes.

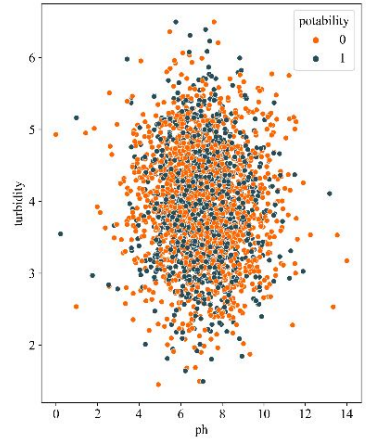
Exploratory data analysis

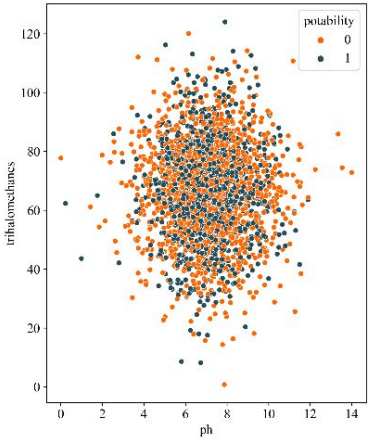
Understanding the dataset

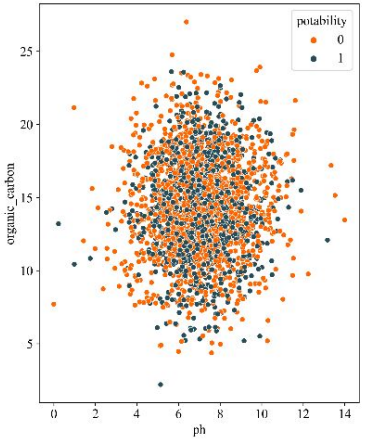


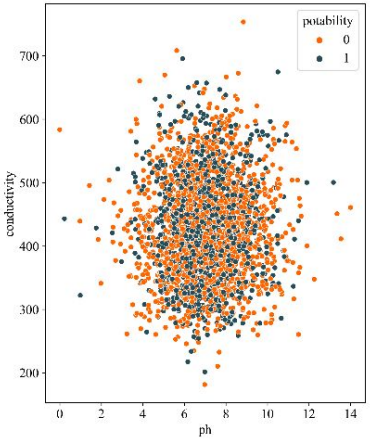
We can resample the data to get a balanced dataset

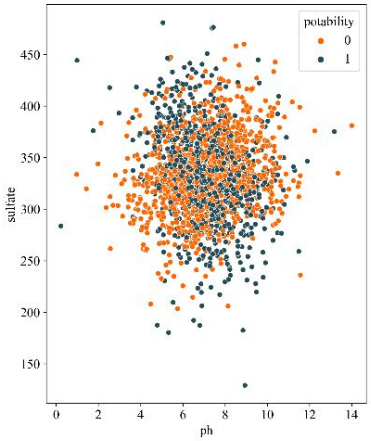


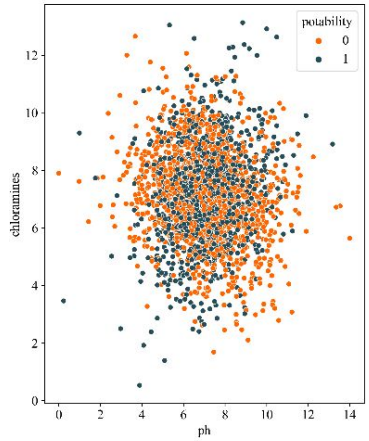

potability
0
1
turbidity
ph

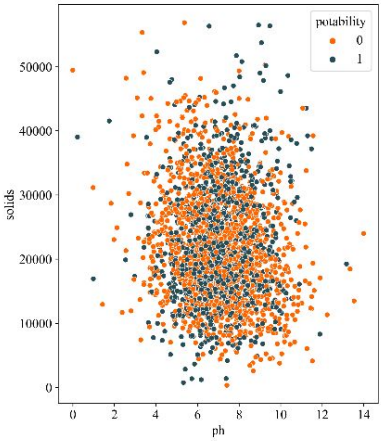

potability
0
1
trihalomethanes
ph

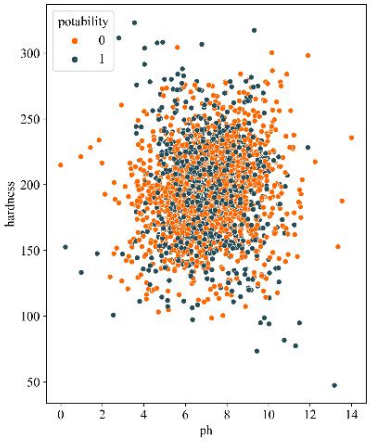

potability
0
1
organic_carbon
ph


potability
0
1
conductivity
ph


potability
0
1
sulfate
ph

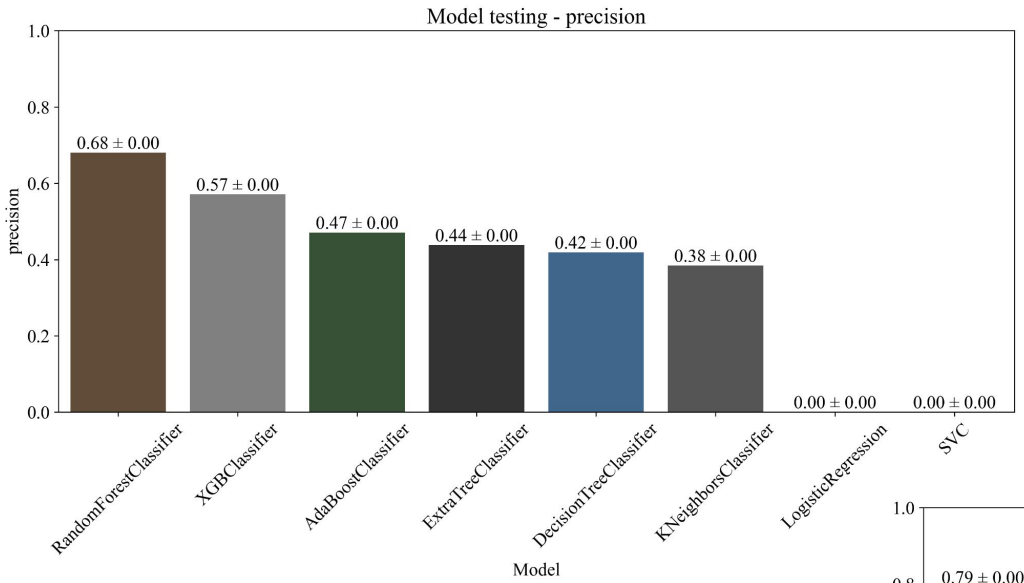

potability
0
1
chloramines
ph


potability
0
1
solids
ph


potability
0
1
hardness
ph

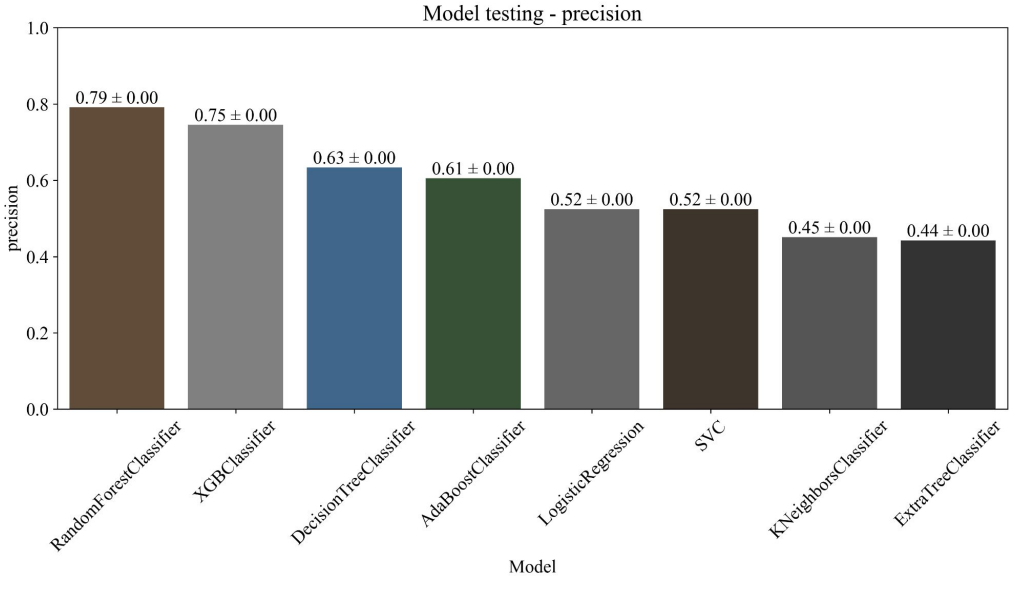
Model design and characteristics

Predictive models

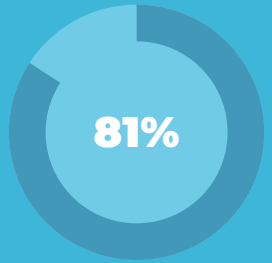


Trained on not pre-processed data

Trained on pre-processed data



Results analysis

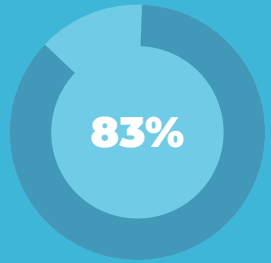


XGBoost

Slow training
Higher accuracy

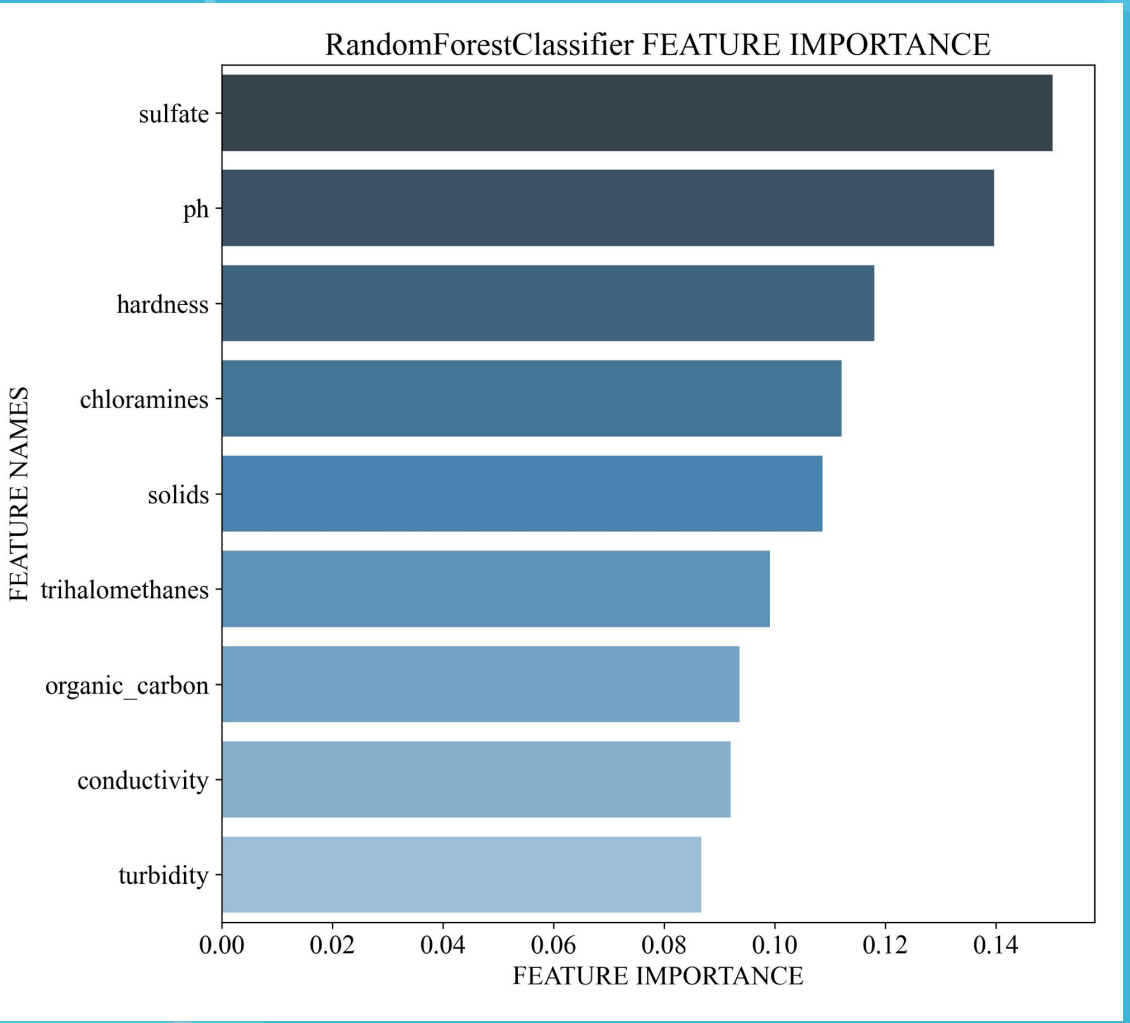
Decision tree

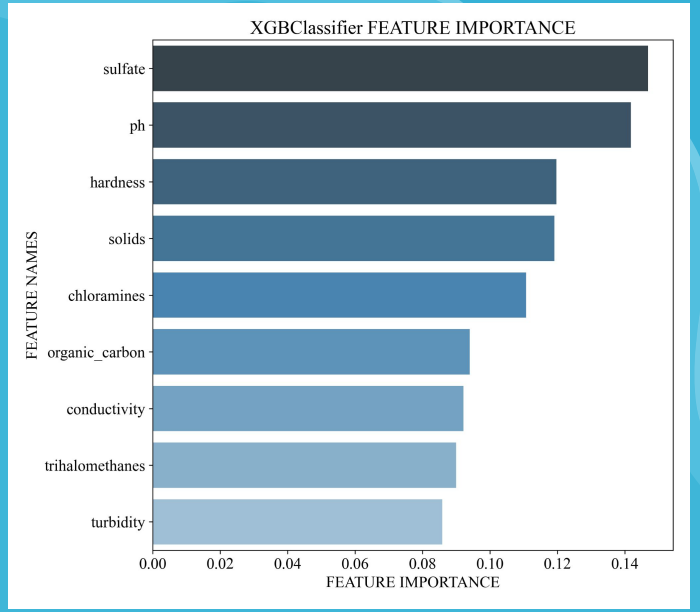
Fast training

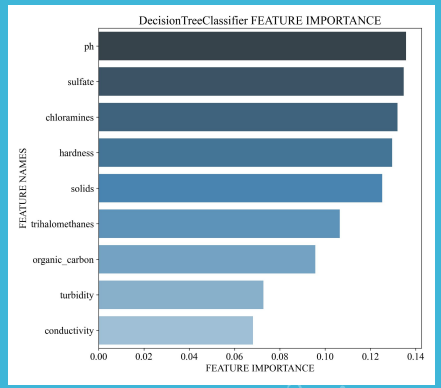


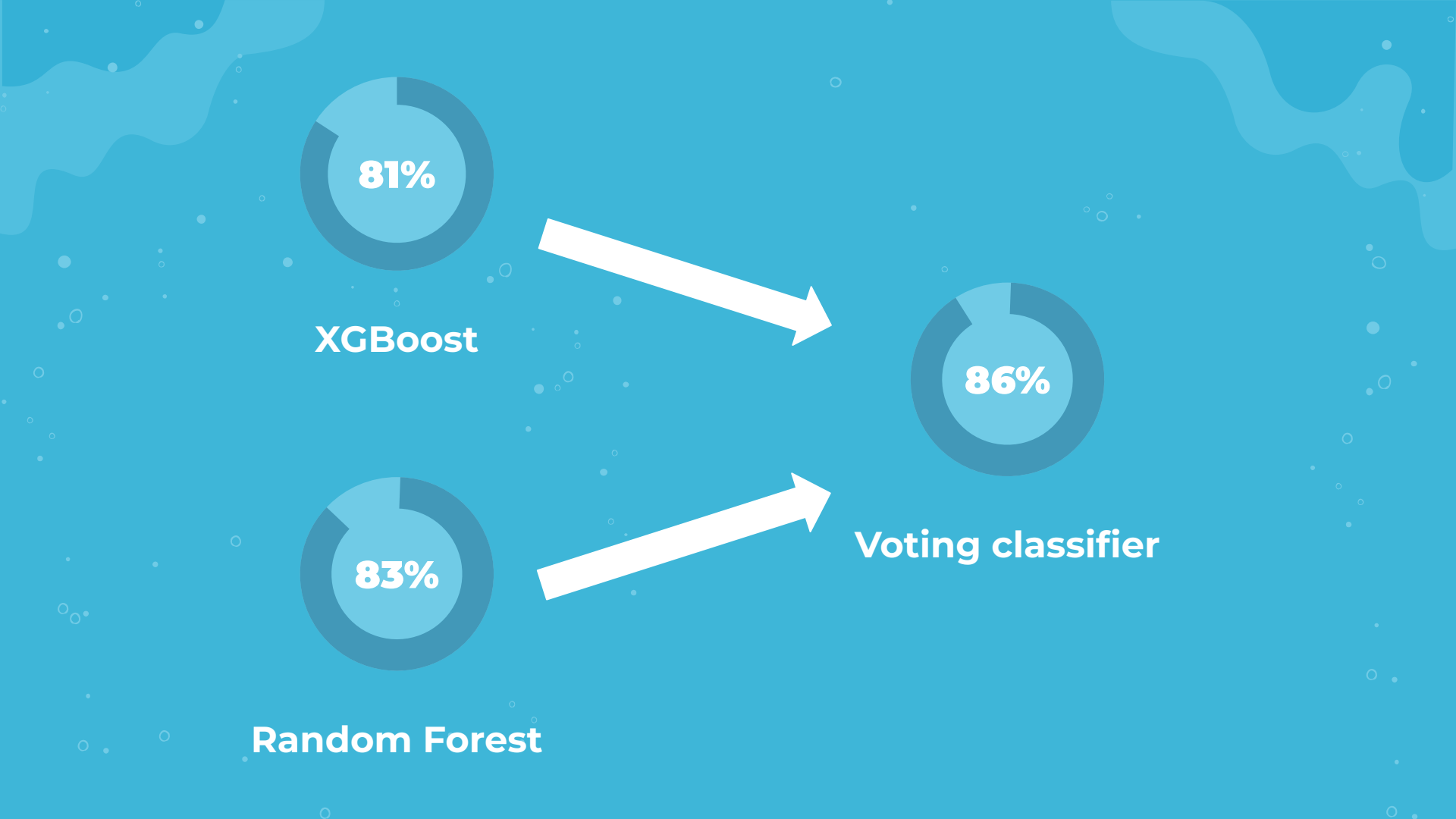
Random Forest

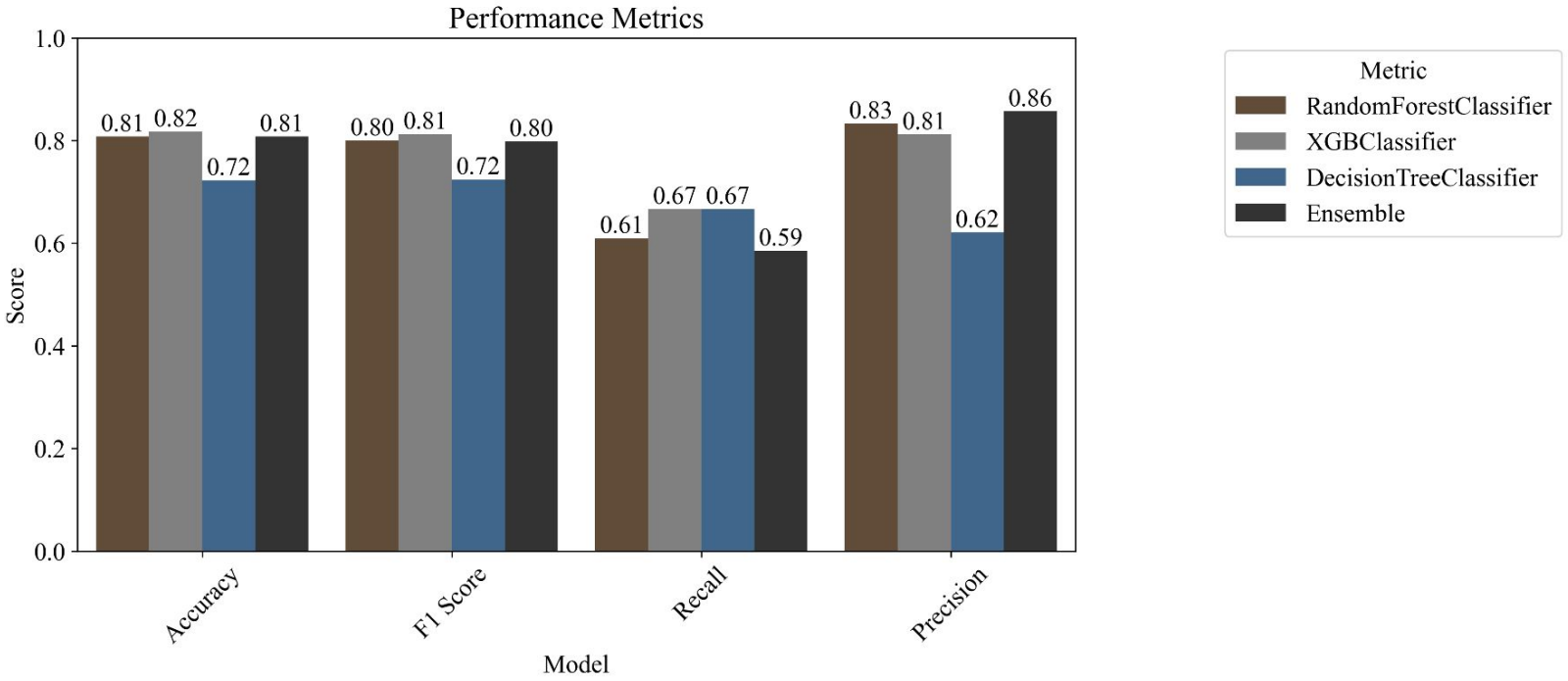
Higher precision


RandomForestClassifier FEATURE IMPORTANCE
sulfate
ph
hardness
chloramines
solids
trihalomethanes
organic_carbon
conductivity
turbidity
FEATURE NAMES
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14
FEATURE IMPORTANCE


XGBClassifier FEATURE IMPORTANCE
sulfate
ph
hardness
solids
chloramines
organic_carbon
conductivity
trihalomethanes
turbidity
FEATURE NAMES
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14
FEATURE IMPORTANCE


DecisionTreeClassifier FEATURE IMPORTANCE
ph
sulfate
chloramines
hardness
solids
trihalomethanes
organic_carbon
turbidity
conductivity
FEATURE NAMES
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14
FEATURE IMPORTANCE


81%
XGBoost
83%
Random Forest
86%
Voting classifier


Performance Metrics
Score
1.0
0.8
0.6
0.4
0.2
0.0
0.81
0.82
0.72
0.81
0.80
0.81
0.72
0.80
0.61
0.67
0.67
0.59
0.83
0.81
0.62
0.86
Accuracy
F1 Score
Recall
Precision
Model
Metric
RandomForestClassifier
XGBClassifier
DecisionTreeClassifier
Ensemble

- Conclusions:
  - Through the analysis, we have identified key factors influencing water potability.
  - The models based on decision trees have demonstrated strong predictive performance in determining water potability.
  - This analysis is of utmost importance in addressing water safety concerns and ensuring the provision of drinkable water to the population.

- Relevant Insights:
  - Our findings reveal that the levels of sulfates, pH, hardness, and chloramine are crucial features in determining water potability.
  - High sulfate levels have a negative impact on water potability, indicating the need for appropriate treatment methods to reduce their concentration.
  - pH values within a certain range contribute significantly to the overall potability, emphasizing the importance of maintaining proper pH levels in water treatment processes.
  - Water hardness and chloramine levels also play a significant role in determining water potability, requiring attention and monitoring in water treatment systems.