WATER POTABILITY PREDICTION

Green AI Project - DIA3

ELVIN CHA, VINCENT COTELLA, DYLAN DRAY, STELLA HU

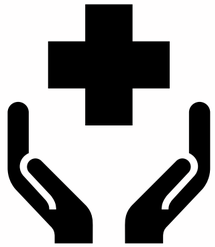
PRESENTATION OF THE DATASET

PRESENTATION OF THE DATASET

Water Quality and Potability

By Laksika Tharmalingam, found on Kaggle

Assesses water quality for human consumption



Indicating potability through specific parameters

HOW THE RAW DATA LOOKS LIKE

	ph	Hardness	Solids	Chloramines	Sulfate	Conductivity	Organic_carbon	Trihalomethanes	Turbidity	Potability
0	NaN	204.890455	20791.318981	7.300212	368.516441	564.308654	10.379783	86.990970	2.963135	0
1	3.716080	129.422921	18630.057858	6.635246	NaN	592.885359	15.180013	56.329076	4.500656	0
2	8.099124	224.236259	19909.541732	9.275884	NaN	418.606213	16.868637	66.420093	3.055934	0
3	8.316766	214.373394	22018.417441	8.059332	356.886136	363.266516	18.436524	100.341674	4.628771	0
4	9.092223	181.101509	17978.986339	6.546600	310.135738	398.410813	11.558279	31.997993	4.075075	0
...	...	...	...	...	...	...	...	...	...	...
3271	4.668102	193.681735	47580.991603	7.166639	359.948574	526.424171	13.894419	66.687695	4.435821	1
3272	7.808856	193.553212	17329.802160	8.061362	NaN	392.449580	19.903225	NaN	2.798243	1
3273	9.419510	175.762646	33155.578218	7.350233	NaN	432.044783	11.039070	69.845400	3.298875	1
3274	5.126763	230.603758	11983.869376	6.303357	NaN	402.883113	11.168946	77.488213	4.708658	1
3275	7.874671	195.102299	17404.177061	7.509306	NaN	327.459760	16.140368	78.698446	2.309149	1

10 columns and 3276 rows

DATA PREPROCESSING

Data cleaning

	0
ph	491
Hardness	0
Solids	0
Chloramines	0
Sulfate	781
Conductivity	0
Organic_carbon	0
Trihalomethanes	162
Turbidity	0
Potability	0

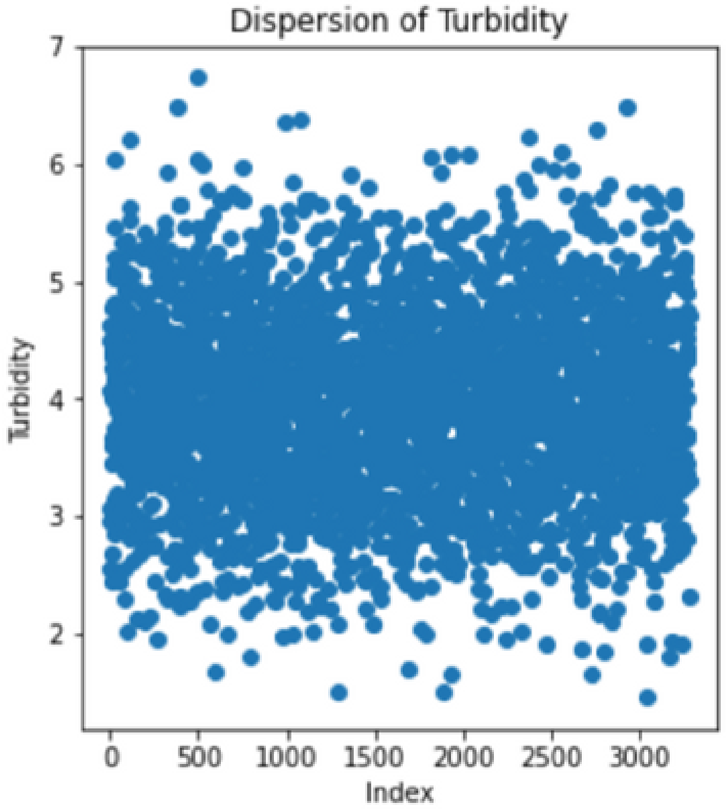
Amount of missing values

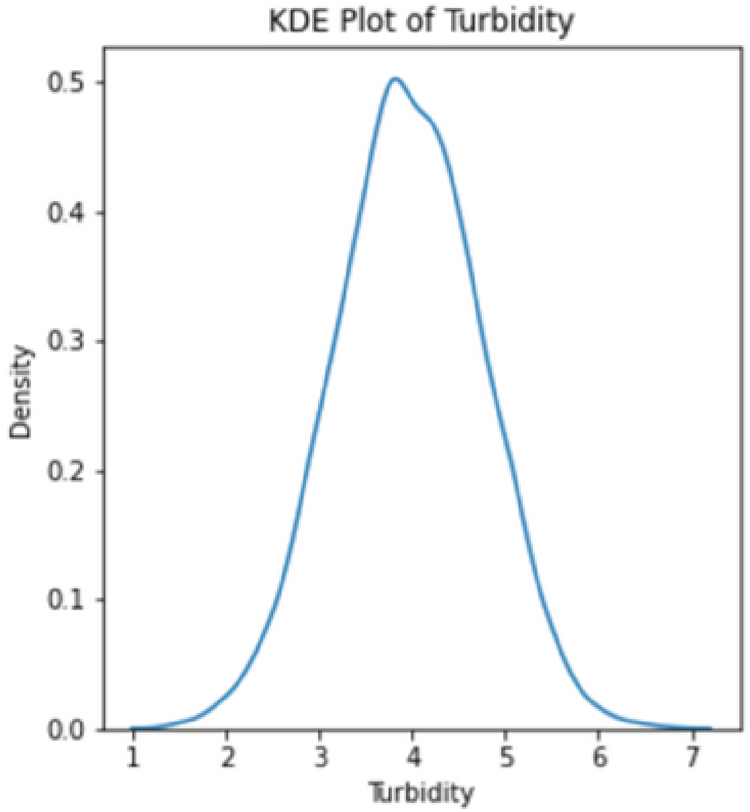
Getting rid of missing values

Replacing NaN by the mean of each column that contains missing values

Check if there are duplicates

Data cleaning





Checking the dispersion for each column

With plots, see if there are any outliers

Data cleaning

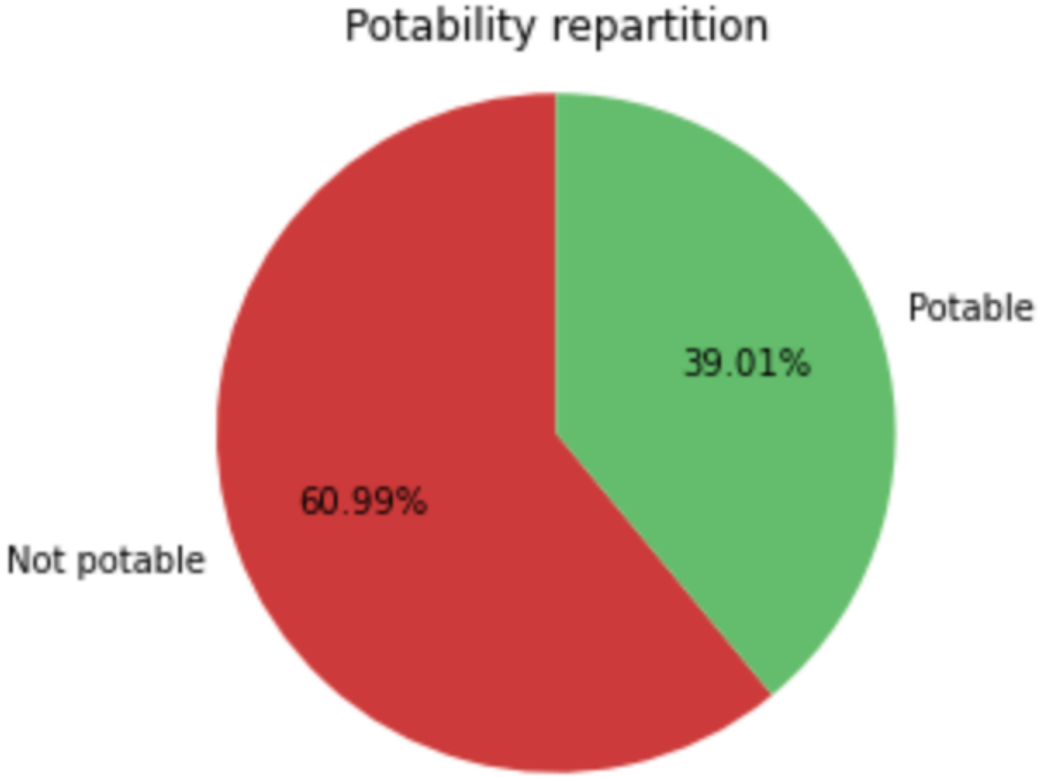
```
Data columns (total 10 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   ph               2785 non-null   float64
 1   Hardness         3276 non-null   float64
 2   Solids           3276 non-null   float64
 3   Chloramines      3276 non-null   float64
 4   Sulfate          2495 non-null   float64
 5   Conductivity     3276 non-null   float64
 6   Organic_carbon   3276 non-null   float64
 7   Trihalomethanes  3114 non-null   float64
 8   Turbidity        3276 non-null   float64
 9   Potability       3276 non-null   int64
dtypes: float64(9), int64(1)
```

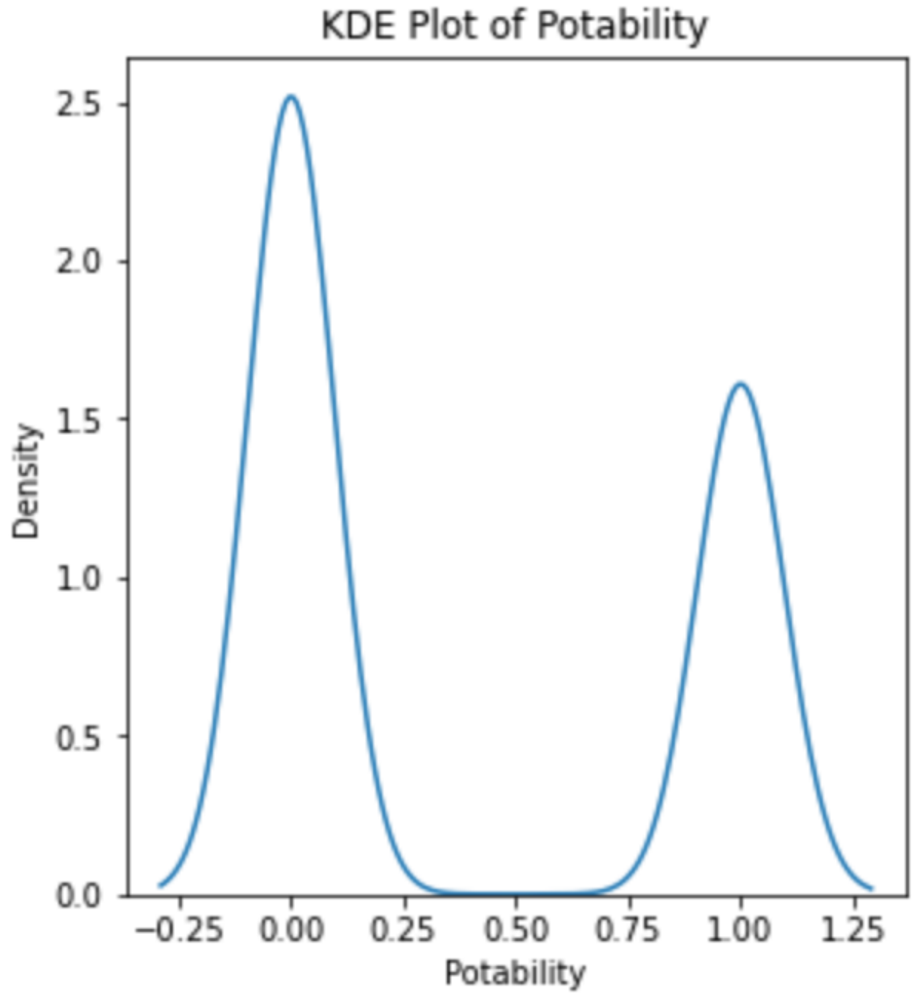
Checking types of variables

They are all numerical

DATA ANALYSIS

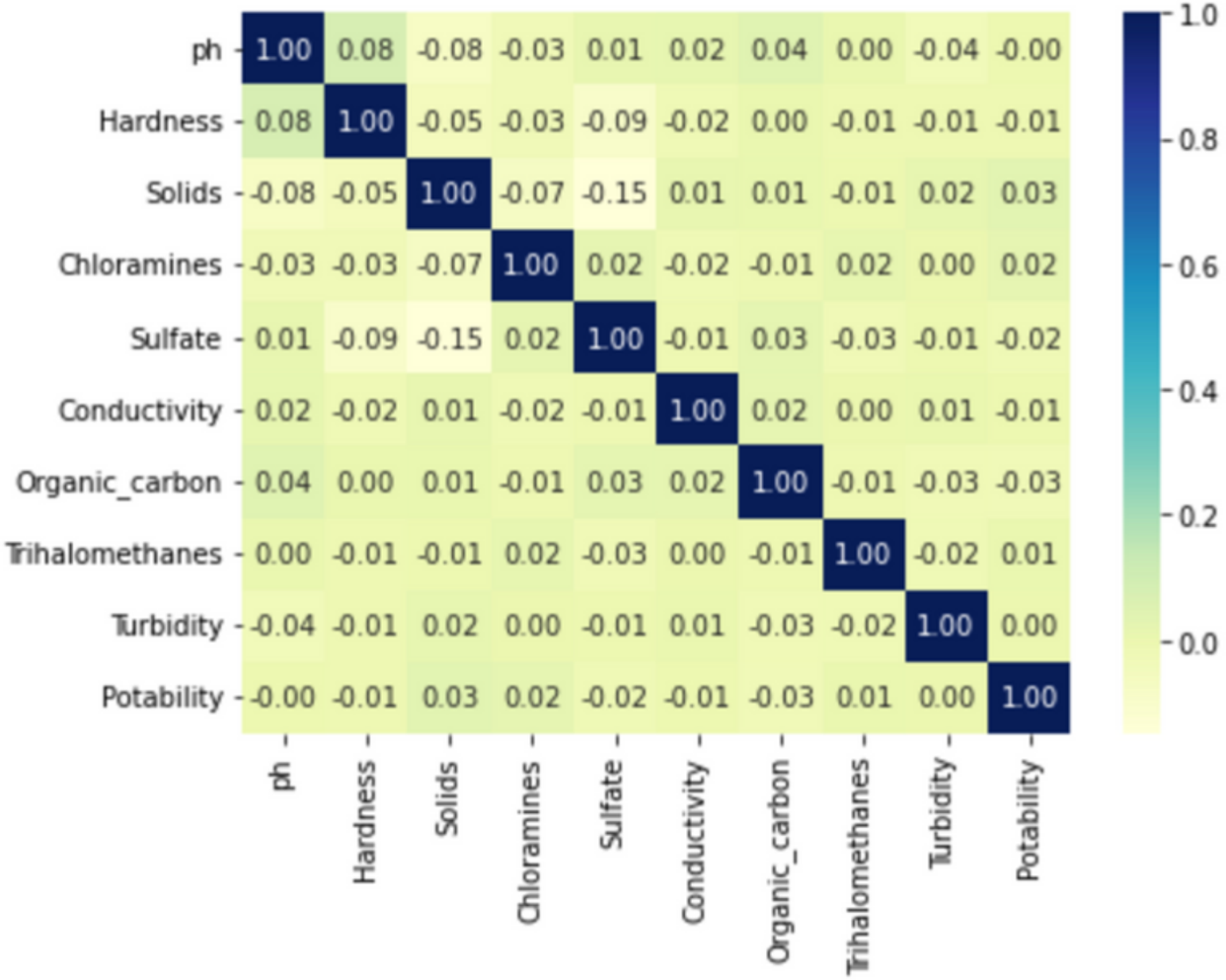
Potability repartition

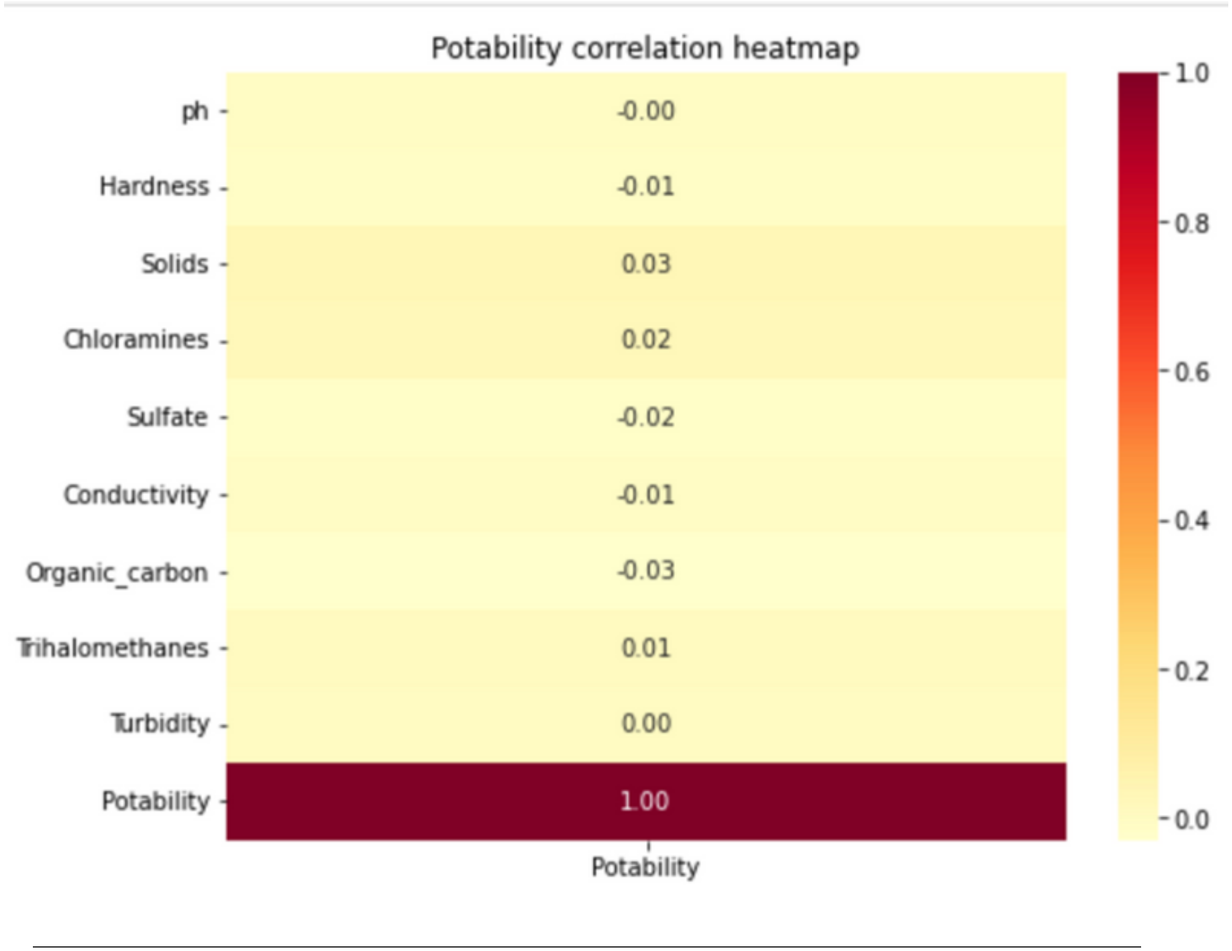




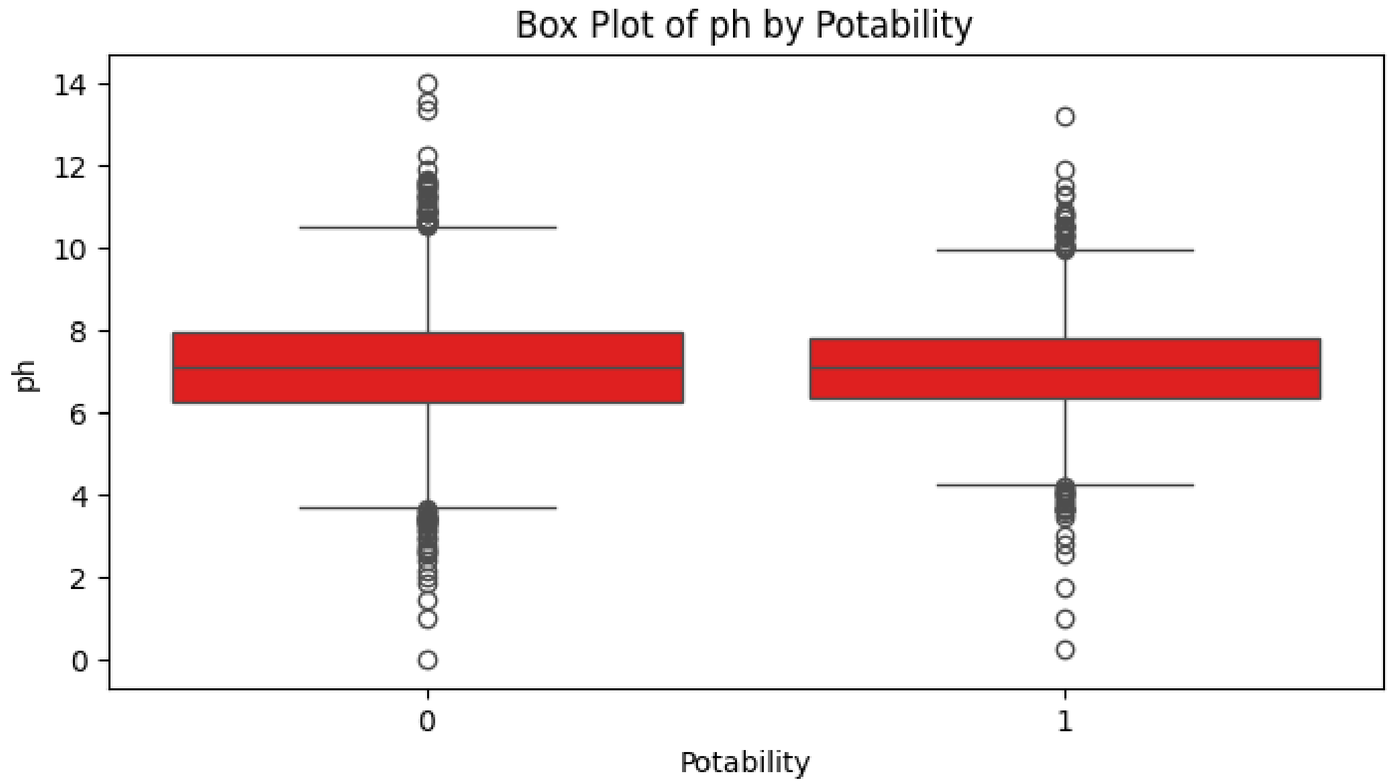
The dataset contains 2/5 of rows with Potability = 1 so the repartition isn't equal

Checking for first correlations





Box Plot Visualizations



MODELING

Modeling

1 Data preprocessing

Train-test splitting, normalization, definition of target column ‘Potability’...

2 Model testing

Decision tree, Random Forest Classifier

3 Predictions

Results, model comparison ...

Preprocessing: Train-Test Split and Normalization

1 - Features and target:

Y : 'Potability'
X : the other columns

Y :

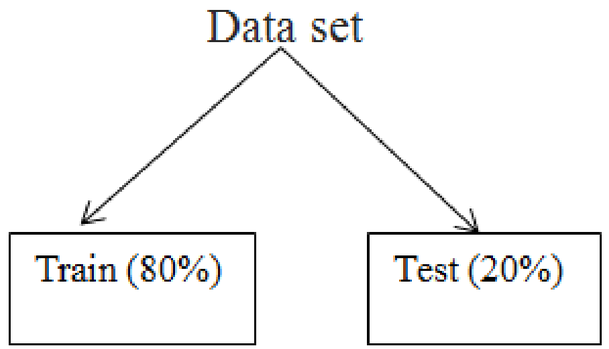
Potability
0
0
0
0
0
0
0
0
0
0

X :

	ph	Hardness	Solids	Chloramines	Sulfate	Conductivity	Organic_carbon	Trihalomethanes	Turbidity
0	null	204.8904564713	20791.318980747	7.3002118732	368.5164413498	564.3086641722	10.3797830781	86.9909704615	2.96313
1	3.7160800754	129.4229205149	18630.0578679703	6.6362458839	null	592.8853691349	15.1800131164	56.3290762845	4.50066
2	8.0991241893	224.2362593936	19909.5417322924	9.2758836027	null	418.6062130545	16.8686369296	66.4200926118	3.05693
3	8.3167668842	214.3733940856	22018.4174407753	8.0593323774	356.8861356431	363.2665161642	18.4366244955	100.3416743651	4.62877

2 - Splitting the data:

80% : training data
20% : testing data




```
Train set: (2620, 9)
Test set: (656, 9)
```

3 - Normalization:

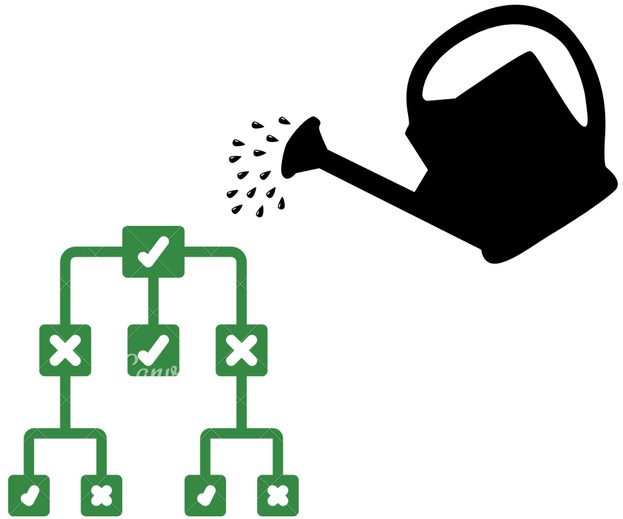
scikit StandardScaler function

X normalized :

	ph	Hardness	Solids	Chloramines	Sulfate	Conductivity	Organic_carbon	Trihalomethanes	Turbidity
0	1.2676965689	0.1069160735	-0.1900542353	0.2839021456	0.5211896559	0.8154614037	1.6119067215	-0.1466912028	-0.9901819234
1	0.5696714494	-0.764648023	-1.7393628178	-1.2791816843	1.0752377715	-1.3419755872	2.6770345761	0.4688915516	0.3937427049
2	-1.3009737197	0.0866510773	-0.5808569893	-0.2652147143	-0.193848656	-1.497450831	-0.5435006021	-0.0058069643	1.1282226478
3	0.2089498919	1.9087085658	-0.3826569483	-1.0370878265	0.1642191996	-0.8684634482	0.204602651	0.419068529	-0.2210674598
4	-0.3373475168	-0.8708613427	0.9474510861	1.5197448189	0.1755224814	0.0555447975	2.7068523859	0.4057063142	0.3999319386
5	0.9024724039	0.8963282706	-1.0322151361	-0.6021439029	-0.1777141147	-0.1678092577	0.9385545512	0.413226568	-0.1828006344

Decisision tree classifier

- Machine learning model used for classification and decision-making, splits a dataset into smaller subsets based on specific features



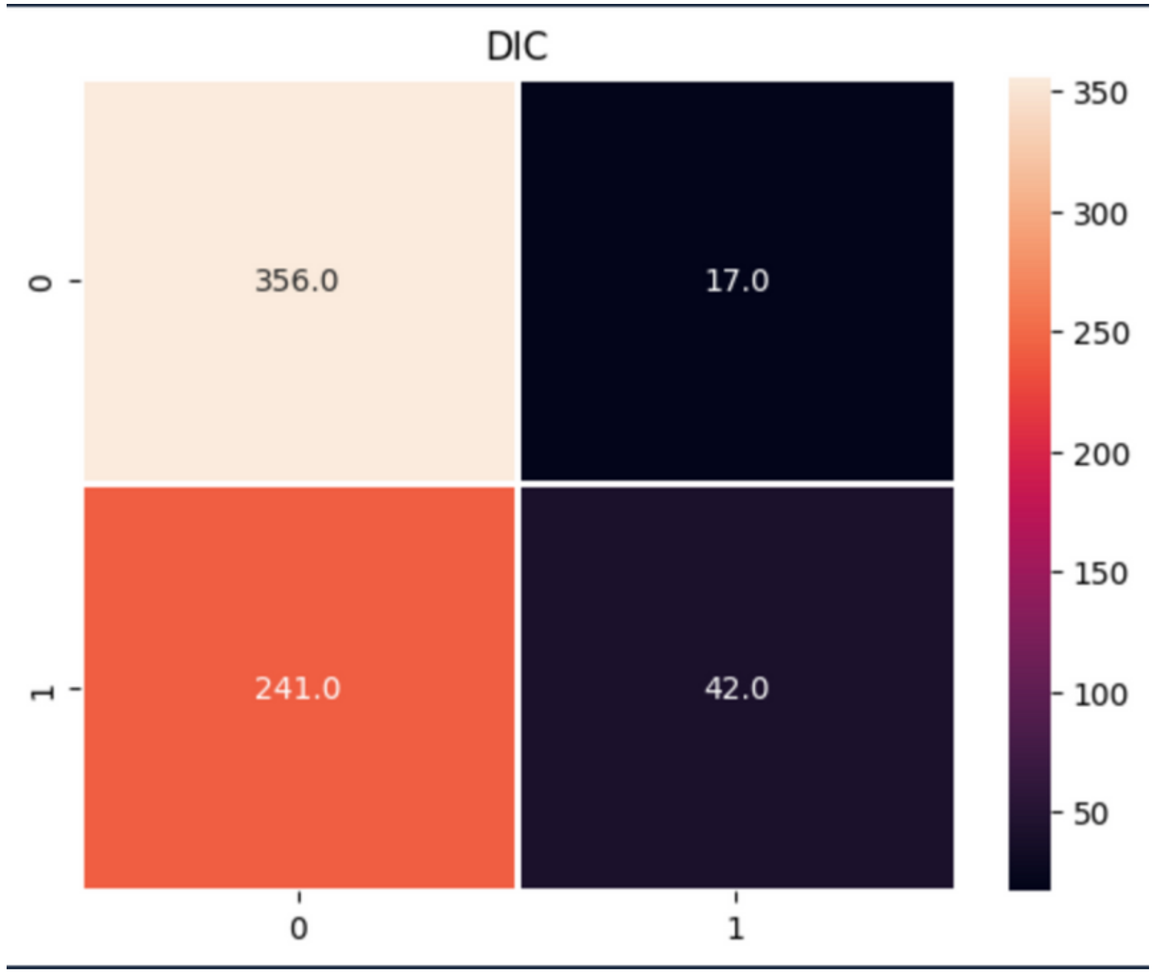
- Max depth = 3
- 80% train, 20% test
- ~~Confusion matrix and decision tree plot to see the results~~

Random Forest classifier

- Model that combines multiple decision trees to improve accuracy
- Using Grid-search for hyperparameters and cross-validation
- 80% train, 20% test
- Confusion matrix and testing methods to improve the model (changing threshold, SMOTE)

RESULTS AND PREDICTION

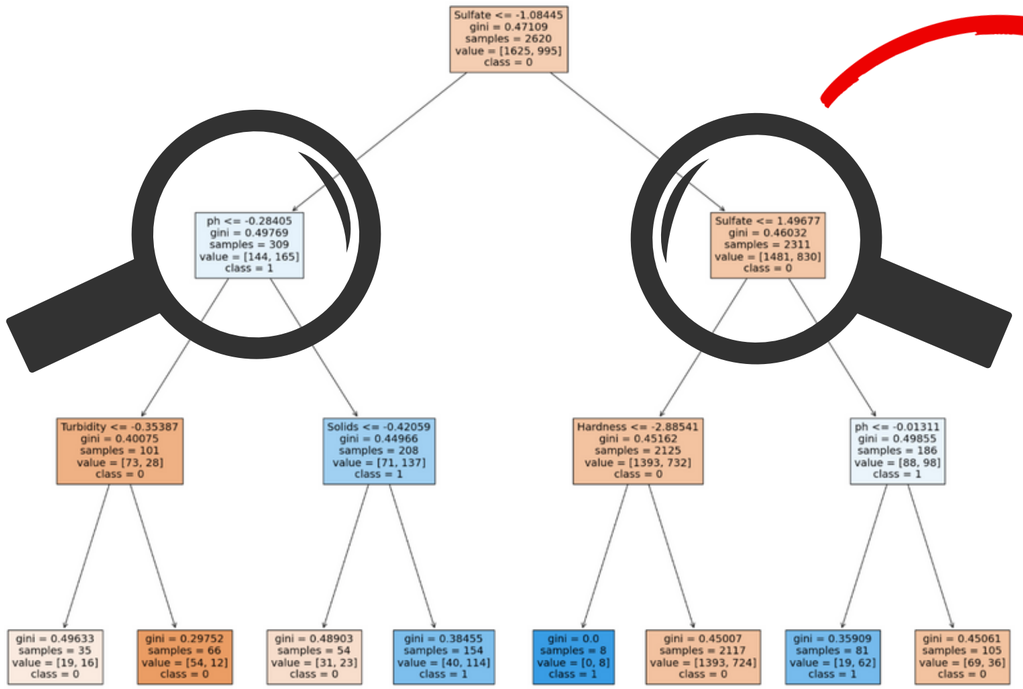
Results for Decision Tree



- Accuracy = 0.60
- Very good at predicting the non-potability of a water sample but bad at predicting true Potability = 1
- Type 2 errors (false negative)

It means that the model fails to identify positive instances, leading to misclassification

Visualization of the tree



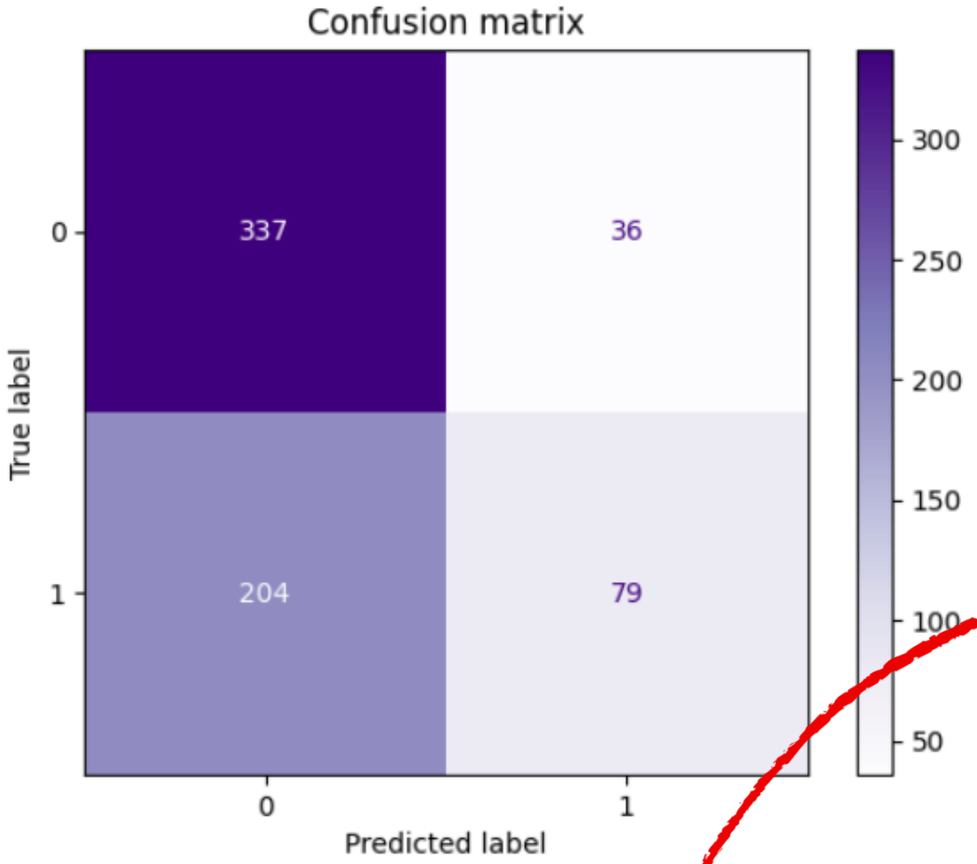
ph <= -0.28405
gini = 0.49769
samples = 309
value = [144, 165]
class = 1

Sulfate <= 1.49677
gini = 0.46032
samples = 2311
value = [1481, 830]
class = 0

- pH & Sulfate are the features which majorly relate to potability of water

- Gini values are at around 0.5: it indicates maximum impurity so it is challenging to make accurate predictions

Results for Random Forest classifier

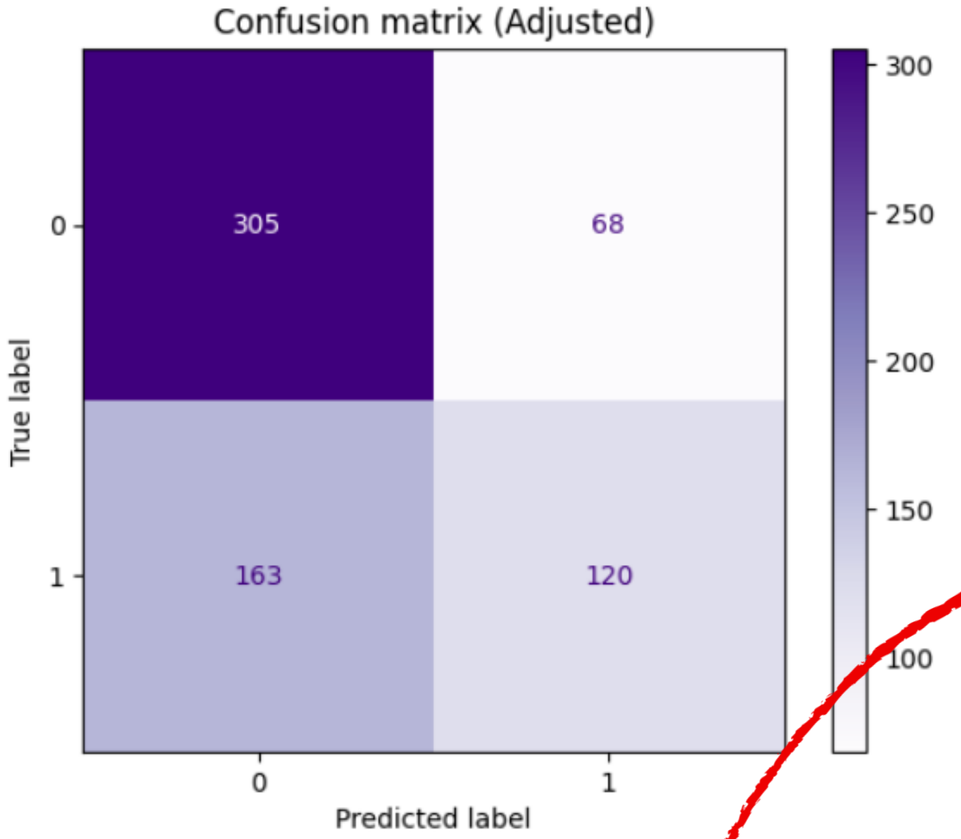



```
Classification report:
              precision    recall  f1-score   support

           0       0.68      0.89      0.77      1625
           1       0.63      0.31      0.42       995

    accuracy                           0.67      2620
```

- Accuracy = 0.67
- Good at predicting true Potability = 0 a water sample but bad at predicting true Potability = 1
- Type 2 errors (false negative)

How can we reduce type 2 errors ?

Modifying threshold



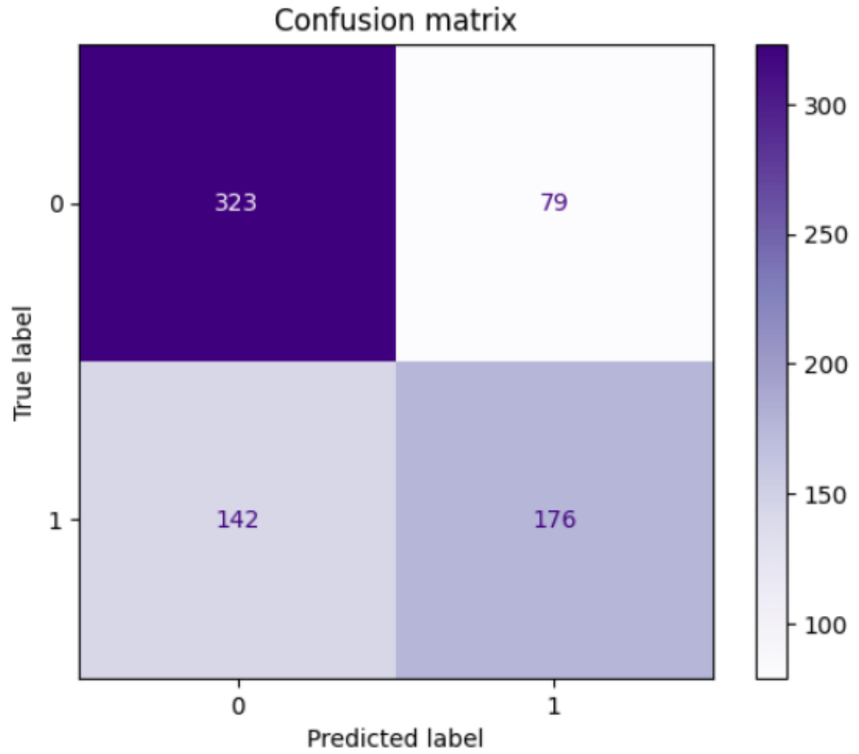
Classification report (Adjusted):

	precision	recall	f1-score	support
0	0.65	0.82	0.73	373
1	0.64	0.42	0.51	283
accuracy			0.65	656

- Accuracy = 0.65
- Threshold = 0.45
- Recall for Potability = 0 decreased but recall for Potability = 1 increased
- A bit less Type 2 errors

The model still fails to identify positive instances, but less than the default threshold, and recognizes less true negatives

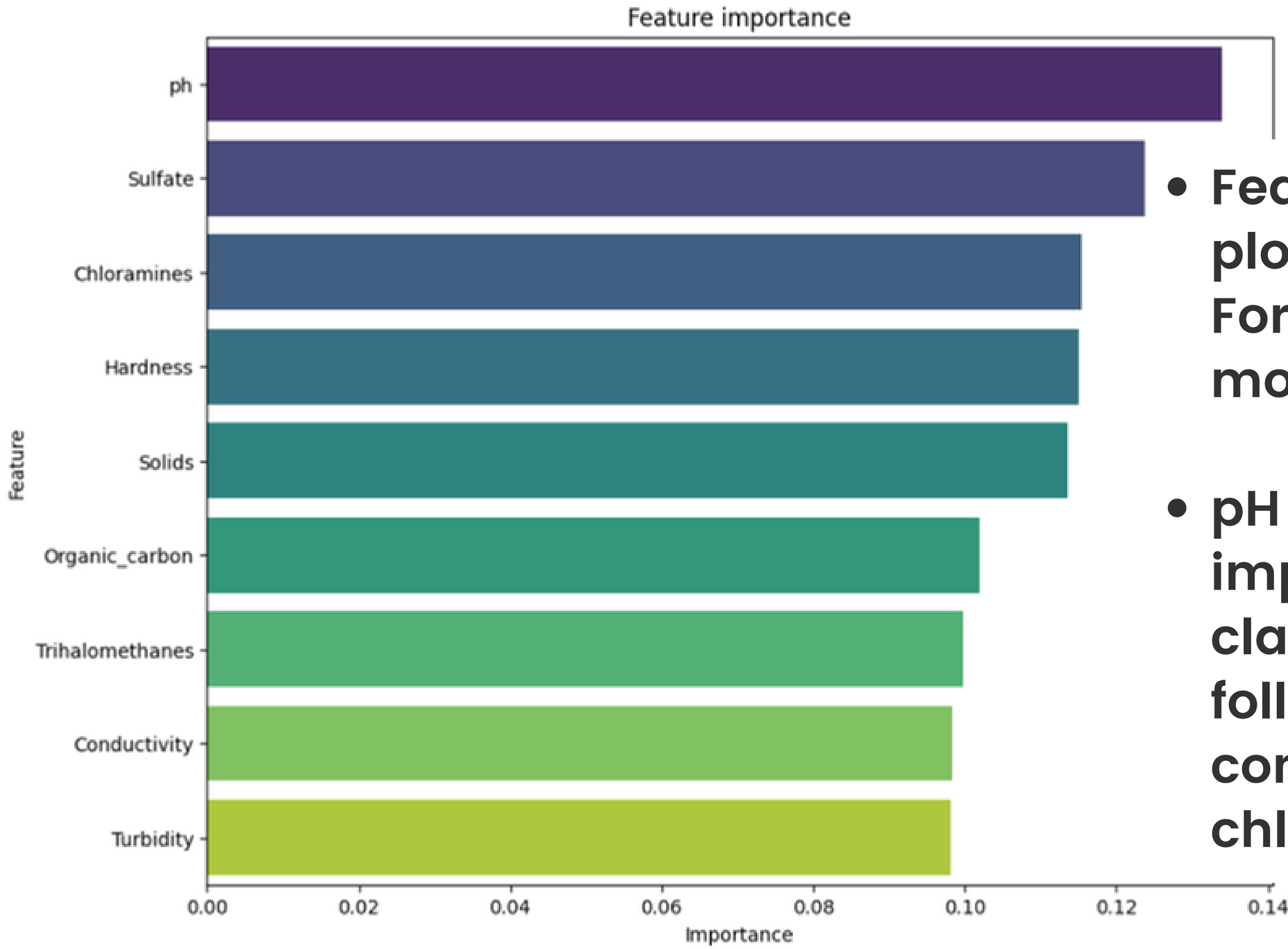
Random Forest classifier with SMOT



Classification report:				
	precision	recall	f1-score	support
0	0.69	0.80	0.75	402
1	0.69	0.55	0.61	318
accuracy			0.69	720

- Accuracy = 0.69 (best)
- Good at predicting the non-potability of a water sample and less bad at predicting true Potability = 1 than the previous ones

The accuracy improved, we reduced type 2 error but the recall for Potability = 0 isn't better, even though the recall for Potability = 1 is better



- Feature importance plot for Random Forest Classifier model

- pH is the most important feature for classification, followed by sulfate concentration and chloramines

CONCLUSION

Which model to choose ?

Decision tree classifier

- **Accuracy: 0.60**
- **Issue with recall: High misclassification of potable water as non-potable**

Random forest classifier

- **Accuracy: 0.67**
- **Similar recall problem as the Decision Tree Classifier**

Random Forest + SMOTE

- **Accuracy improvement to 0.68**
- **Addressed false negatives but a slight increase in false positives**

IT DEPENDS !

Which purpose ?

We want to focus on HUMAN CONSUMPTION

Considerations for human consumption

- Prioritize human well-being over model accuracy
- Advocate for caution and prudence in decision-making
- Highlight the need to minimize the risks associated with false positives in predicting water potability

We don't want people to drink non-potable water !

Open perspective : Impact on public health and environment

The development of a more accurate system for predicting water safety based on factors like pH, sulfate etc is crucial. Beyond its direct impact on health, such a system could contribute significantly to environmental preservation, enhance water treatment processes, and empower communities with informed decision-making abilities. This advancement has the potential to address global water challenges. By refining our ability to predict water quality, we are not only prioritizing current health concerns but also establishing foundations for a more secure and intelligent future on a global scale.

Thank you for listening !