Water Quality Analysis

Phase 3: Development part 1

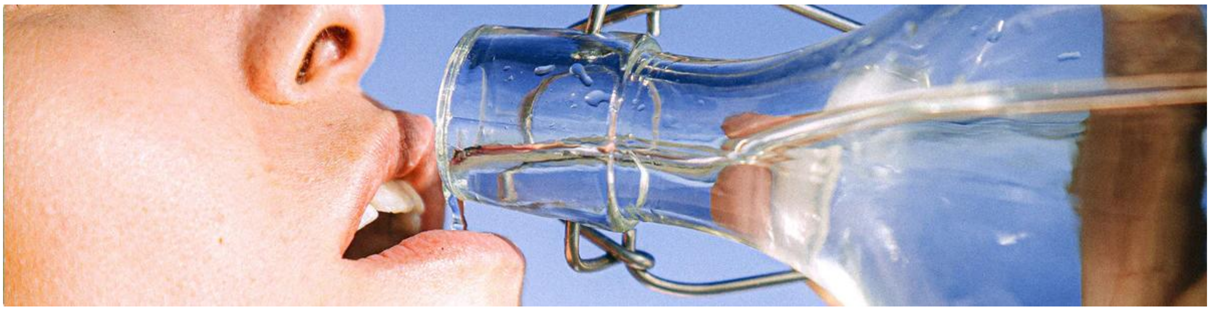
In this part you will begin building your project by loading and preprocessing the dataset.

Start building the water quality analysis preprocessing data and performing exploratory data analysis

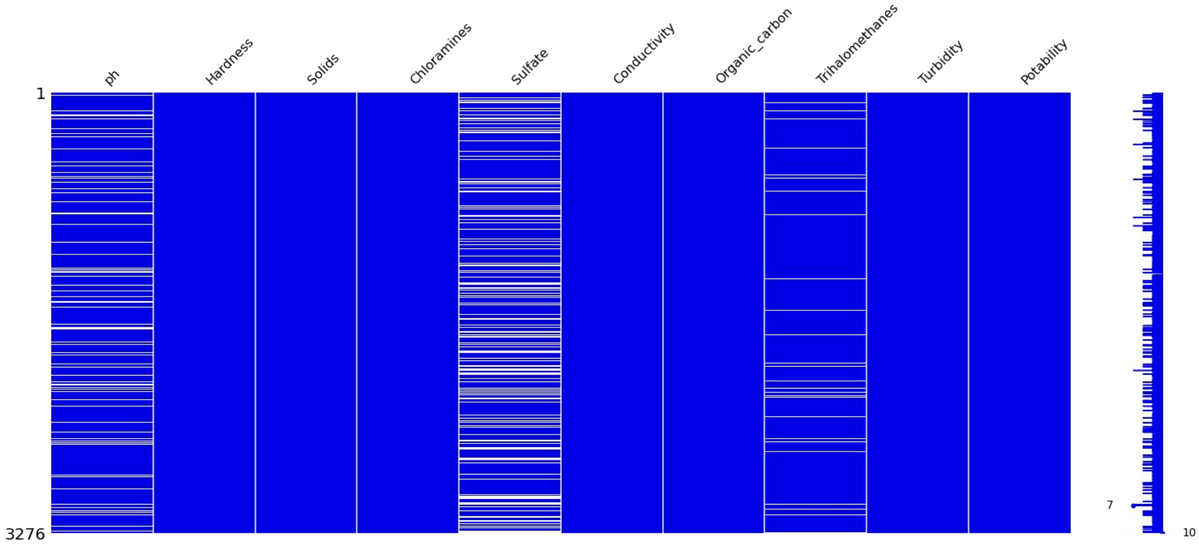
Obtain the water quality dataset and preprocess it by handling missing values

And outliers

Conduct EDA to visualize parameter distributions correlation and potential deviation from standards



Checking Missing Values



Observation:

- There are missing values for ph, sulfate, and trihalomethanes.

-
  ```
  ph 491
  ```

-
  ```
  Hardness 0 Solids 0 Chloramines 0 Sulfate 781 Conductivity 0
  Organic_carbon 0 Trihalomethanes 162 Turbidity 0 Potability 0
  is_ph_ok 0 is_Hardness_ok 0 is_Solids_ok 0 is_Chloramines_ok 0
  is_Sulfate_ok 0 is_Conductivity_ok 0 is_Organic_carbon_ok 0
  is_Trihalomethanes_ok 0 is_Turbidity_ok 0
  ```

Handling Missing Values

The method of handling missing values will differ depending on the dataset and the nature of the problem to be solved. This is a matter of determining drinkable water. Filling in missing values with certain predicted values can be a very risky decision.

For example, suppose that the missing value in the 'ph' feature is filled with a certain value. Suppose the actual value of ph is 0, but for some reason it is treated as a missing value. If the ph is 0, it is the same ph as the battery, and people should never drink this kind of wate

After Imputation of missing value

```
ph 0 Hardness 0 Solids 0 Chloramines 0 Sulfate 0 Conductivity 0
Organic_carbon 0 Trihalomethanes 0 Turbidity 0 Potability 0 is_ph_ok 0
is_Hardness_ok 0 is_Solids_ok 0 is_Chloramines_ok 0 is_Sulfate_ok 0
is_Conductivity_ok 0 is_Organic_carbon_ok 0 is_Trihalomethanes_ok 0
is_Turbidity_ok 0
```

Doing Anomaly Detection

We want to find outliers in the dataset through anomaly detection. When solving the water portability problem, it seems right to remove outliers. This is because there should not be outliers in the dataset related to life.

the size of anomaly = 101

Observation:

- There are 110 anomaly data.
- If you look at the Top 10 anomaly dates, there are many data judged to have potability.

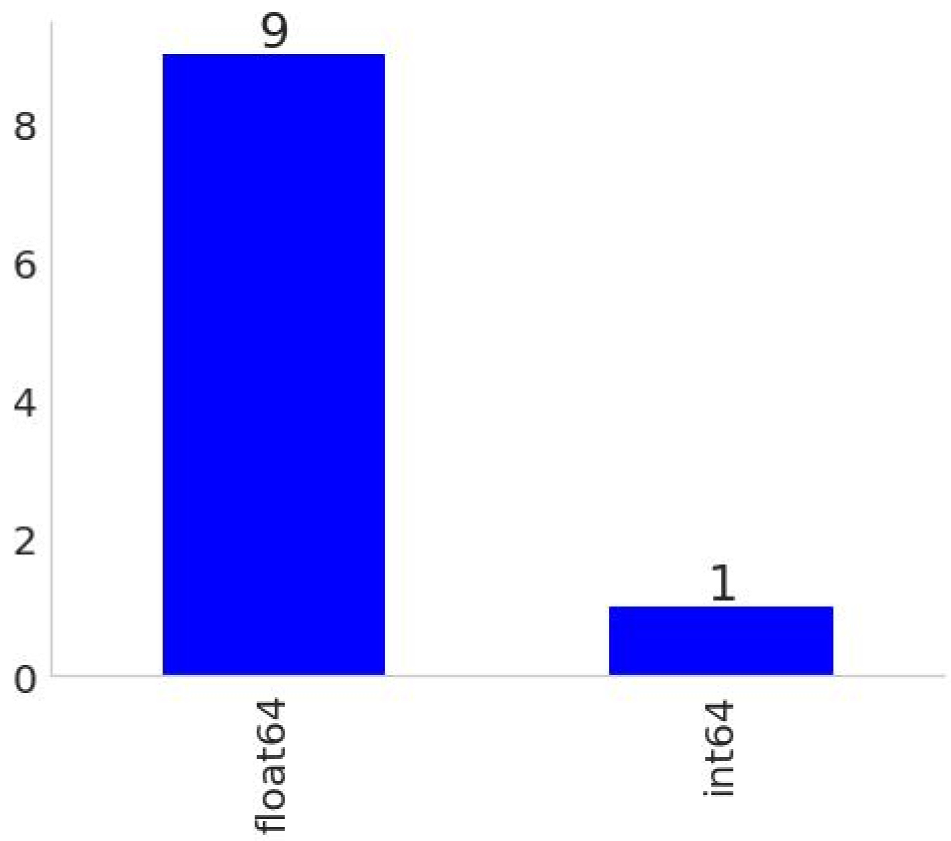
Observation:

- Looking at the target of data judged as anomaly, there are more cases judged as potability.

What can be predicted from this is that there are many cases in this dataset where undrinkable water is judged to be drinkable. That is, we can predict that recall may be low.

Exploratory Data Analysis

Checking the data type of features



Observation:

- All are numerical features.

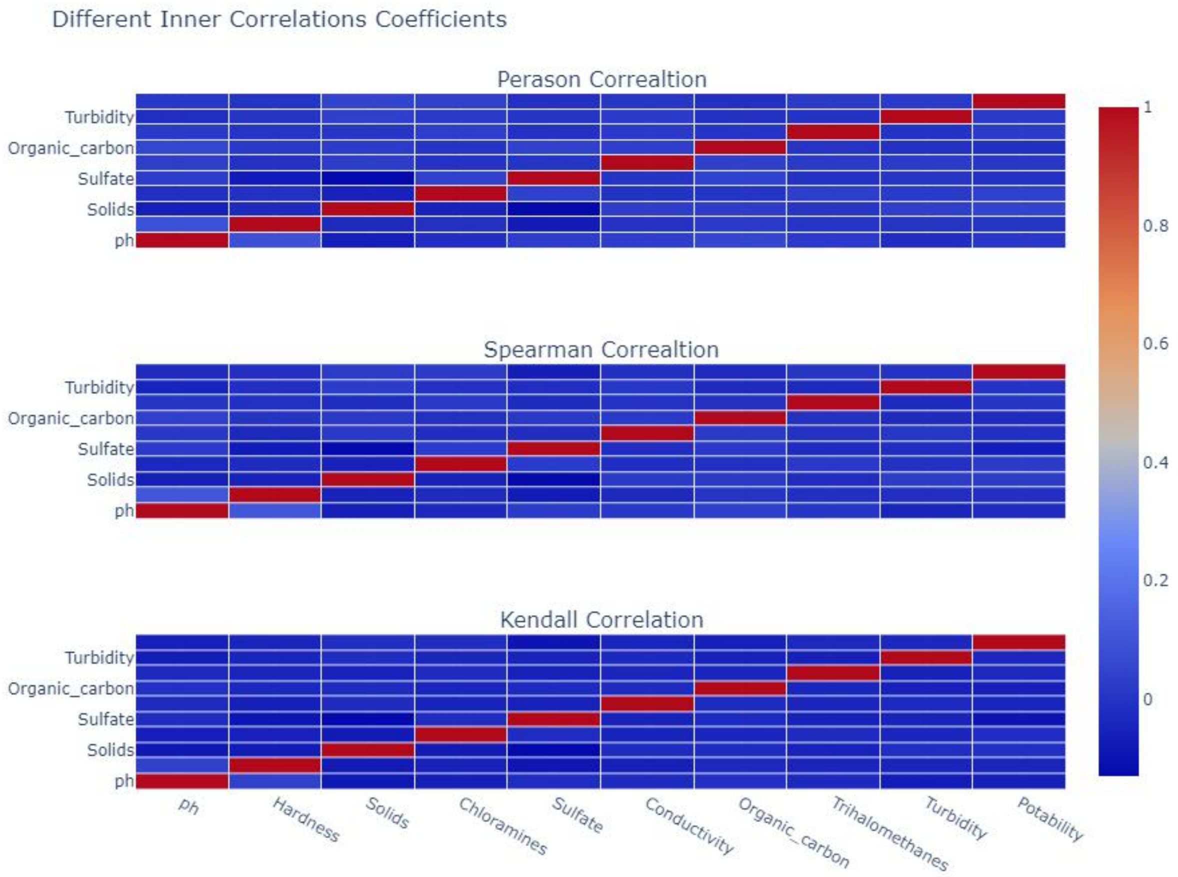
Checking Target Imbalance

Observation:

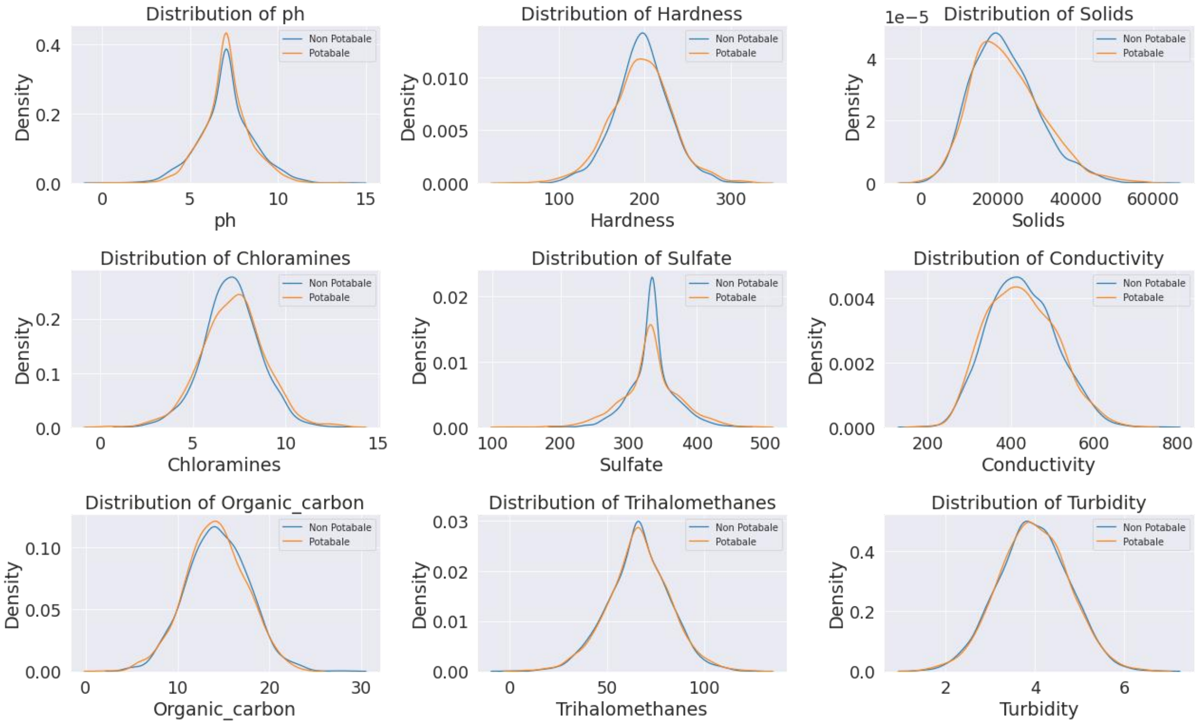
OK! Target is balanced!

General Data Analysis

Observation: We see that we have some degree of unbalancedness in our data; we will not apply any upsampling/downsampling methodology as the proportions are more close to equal than to be extremely balanced (cases like 90% / 10% where upsampling is crucial). Also, the more significant label ("Not potable") is the one with more samples; logically, we would prefer a model that will have more false negatives rather than a model that has more false positives.



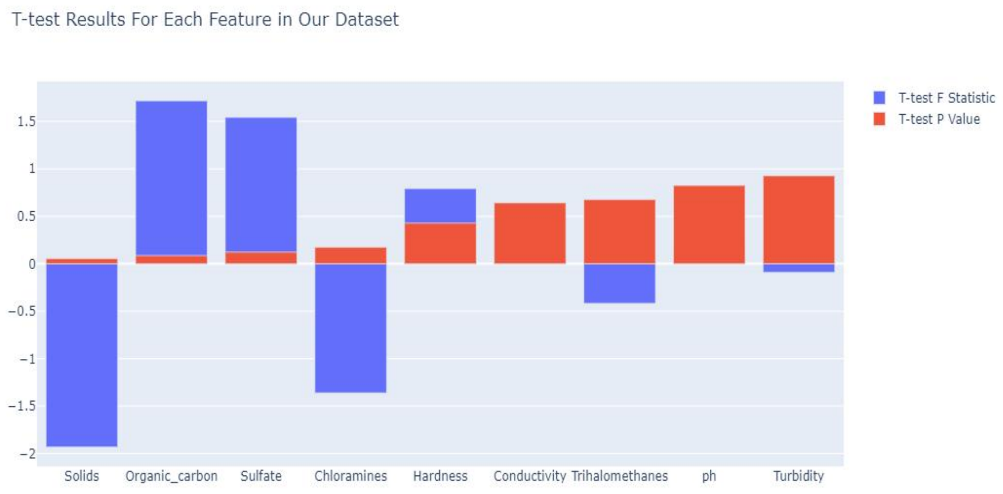
Observation: It appears that there is no linear/ranked correlation between our output label and our features, mostly due to the fact that we have a binary label and continuous features, traditional linear correlation coefficients won't tell us the true underlying story about the relationships between our features and the target variable



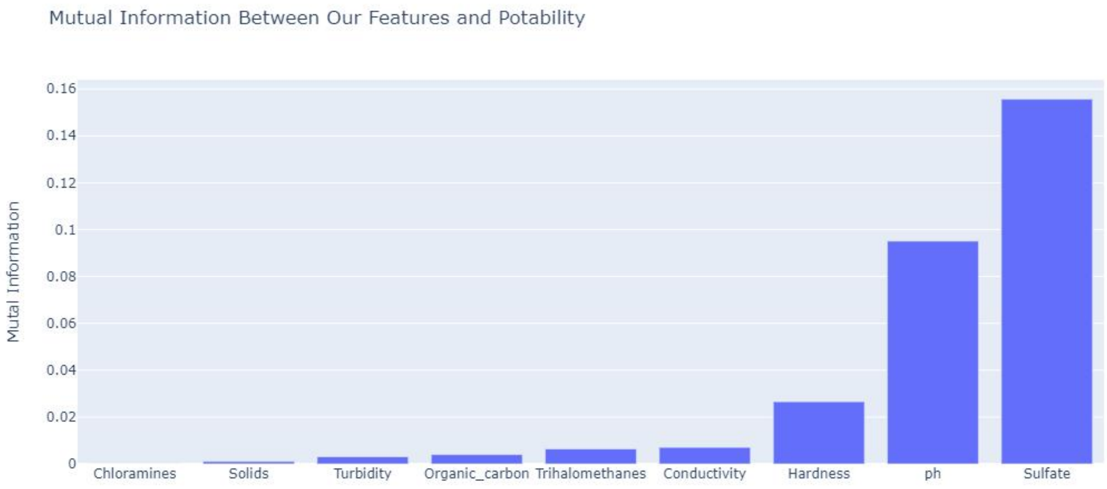
Observation: Looking at the distribution of all our features divided by our target label, we see that some of them have some difference, a key point that can help us select the features with which we will train our models. To better understand the differences between the features with respect to the target label, a more robust analysis is required to confirm any hypothesis we may have at this point just from looking at the distribution plots.

Statistical Difference Analysis

Explanation: In order to test for any significant difference between "potable" and "non-potable" water samples, we will treat both labels as two separate populations from which we sampled 'n' and 'k' samples (n = the number of "potable" samples, 'k' = the number of "non-potable" samples). We will perform a two-tailed t-test to check if there is any significant difference between the two sample means, considering the sample size differences and unequal variance. We expect to see low p-values for the features that indeed are significantly different between the labels. We will set our significance level alpha to be equal to or less than 0.1.

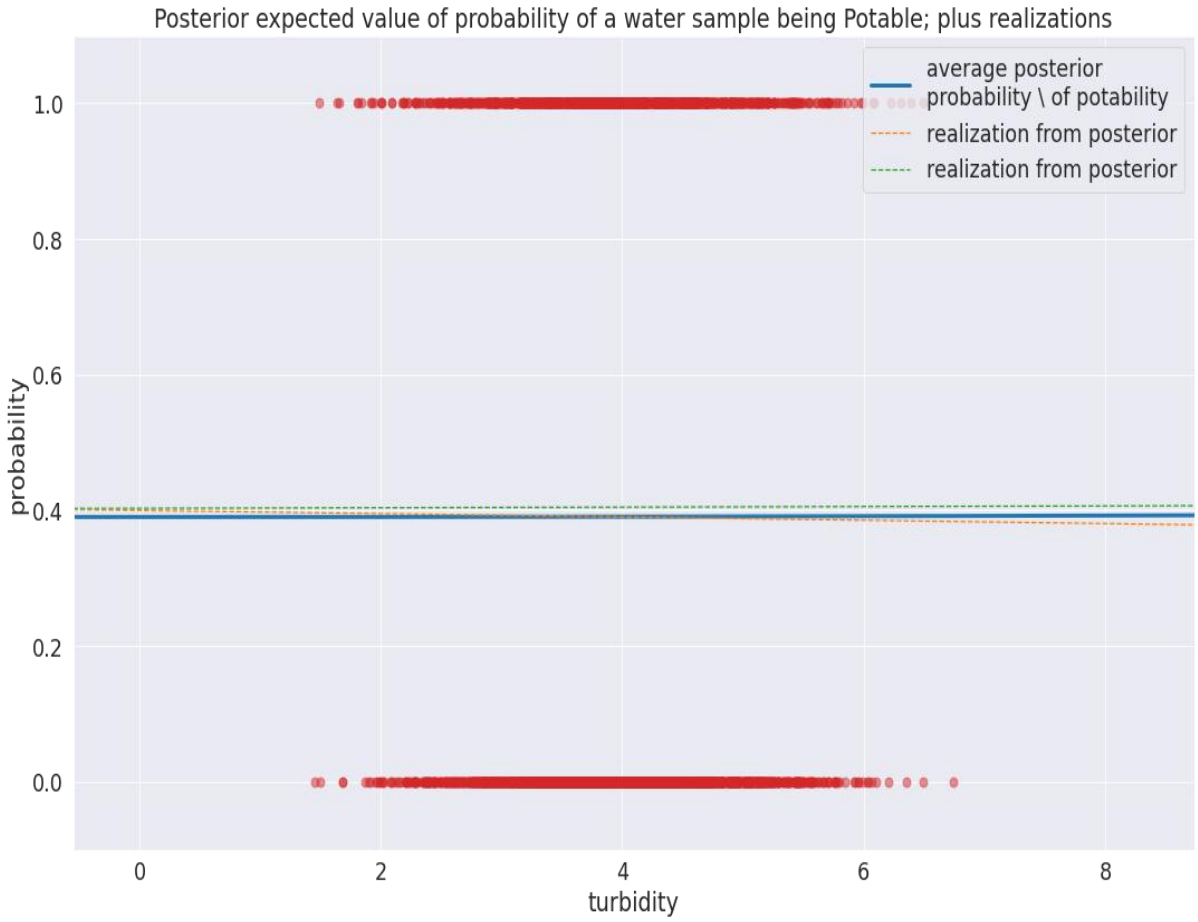


Observation: After performing the two-tailed t-test, we see that only "Solids" and "Organic carbon" have p-values below our pre-defined alpha value, even though there are two more features closer to our alpha level than the other 4. When we get to the modeling stage, the 4 features we will use will be all the features we see in the above plot with p-values below 0.18 (first 4 features in the plot)



Observation: As an additional metric for consideration, we use "Mutal Information" to test and see if there is any similarity between the probability distribution of or continuous features with the Bernoulli distribution that represent our target. We see that some of the worst scoring features in our t-test have the highest mutual information with our target label, conceptually meaning that knowing something about "Ph" decreases my uncertainty in assuming about "Potability," unfortunately, mutual information doesn't tell me exactly to what assumption does "Ph" contribute. Still, none the less it is an indicator of relationship and a strong what in the matter, so we will indeed include it as well in our modeling section.

Probabilistic Inference

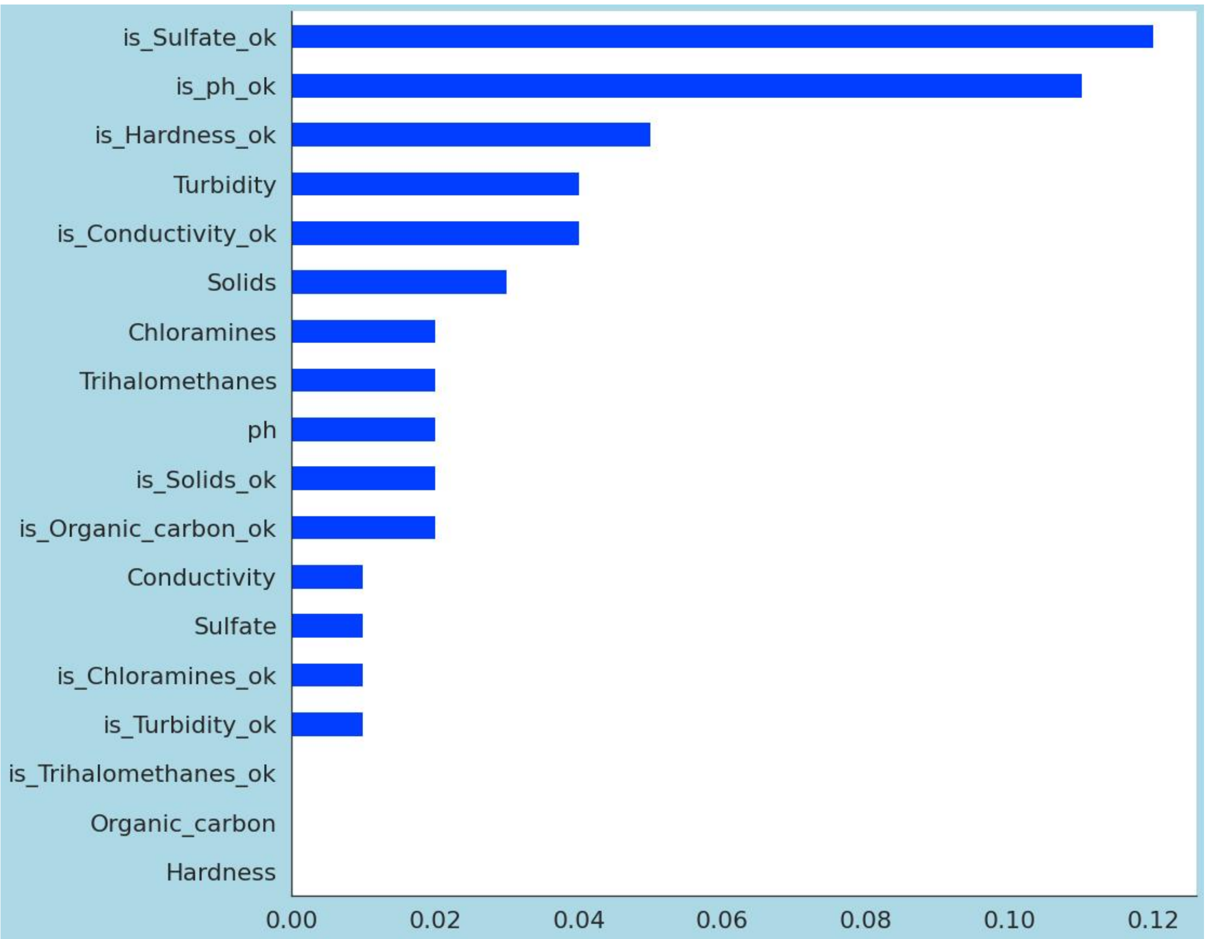


We see that after exploring and modeling the potability as a process of turbidity, the underlying posterior distribution of a logistic model that should have found a threshold of classification if it was possible to gain confidence about potability of water based on turbidity, unfortunately, this is not the case.

Checking Feature Importance

Here we check feature importance in various ways.

Knowing which features are important when judging heart disease from different features will help you make a decision. It will also be very helpful when explaining to people who have been diagnosed.

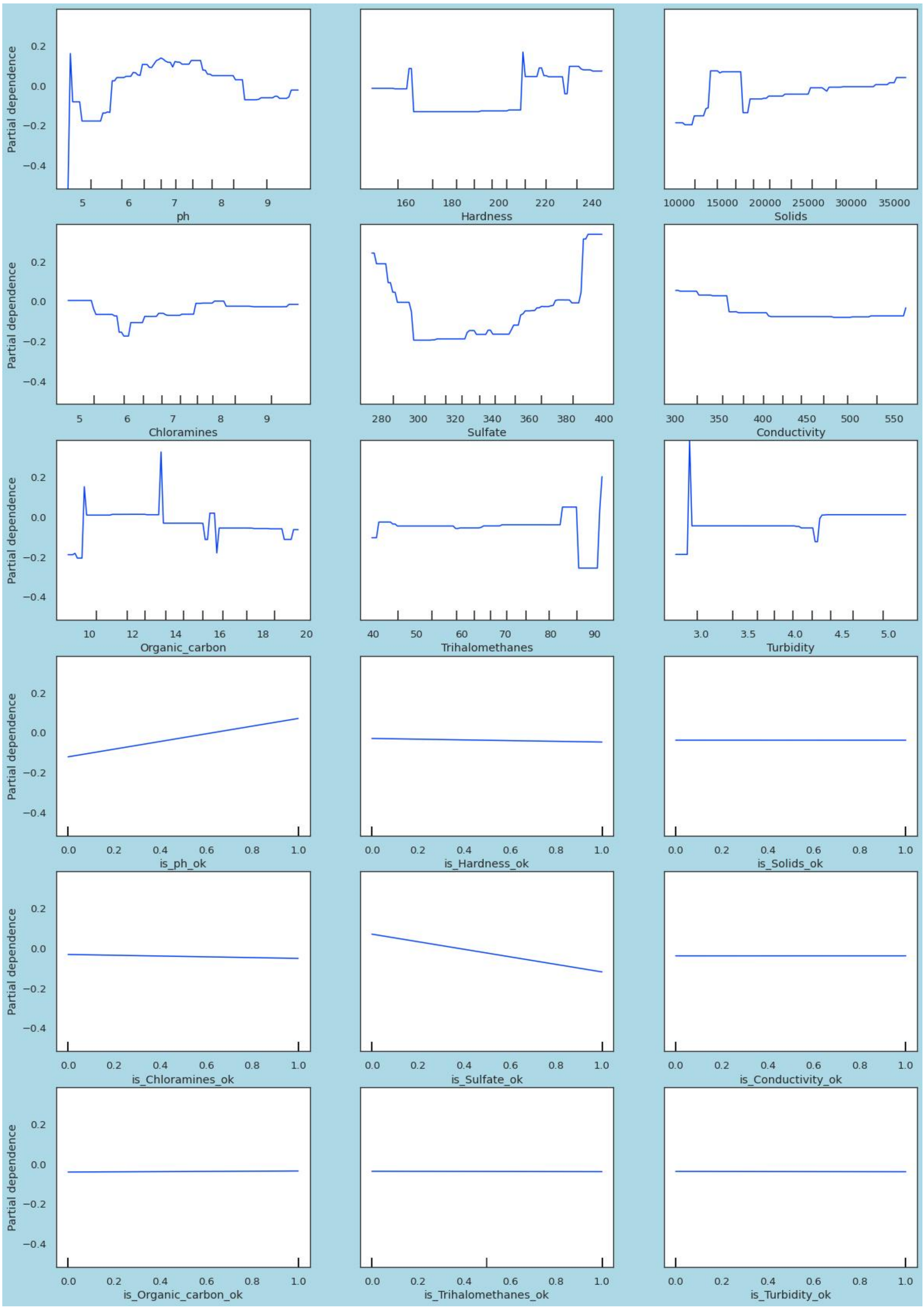


Observation:

- The correlation coefficient of all features is low.
- Correlation coefficients of newly created derived features are relatively high.

Feature importance with partial dependence

Partial dependence plots (PDP) show the dependence between the target response and a set of input features of interest, marginalizing over the values of all other input features (the‘complement’ features). Intuitively, we can interpret the partial dependence as the expected target response as a function of the input features of interest


Partial dependence
0.2
0.0
−0.2
−0.4
5
6
7
8
9
ph
160
180
200
220
240
Hardness
10000
15000
20000
25000
30000
35000
Solids
5
6
7
8
9
Chloramines
280
300
320
340
360
380
400
Sulfate
300
350
400
450
500
550
Conductivity
10
12
14
16
18
20
Organic_carbon
40
50
60
70
80
90
Trihalomethanes
3.0
3.5
4.0
4.5
5.0
Turbidity
0.0
0.2
0.4
0.6
0.8
1.0
is_ph_ok
is_Hardness_ok
is_Solids_ok
is_Chloramines_ok
is_Sulfate_ok
is_Conductivity_ok
is_Organic_carbon_ok
is_Trihalomethanes_ok
is_Turbidity_ok

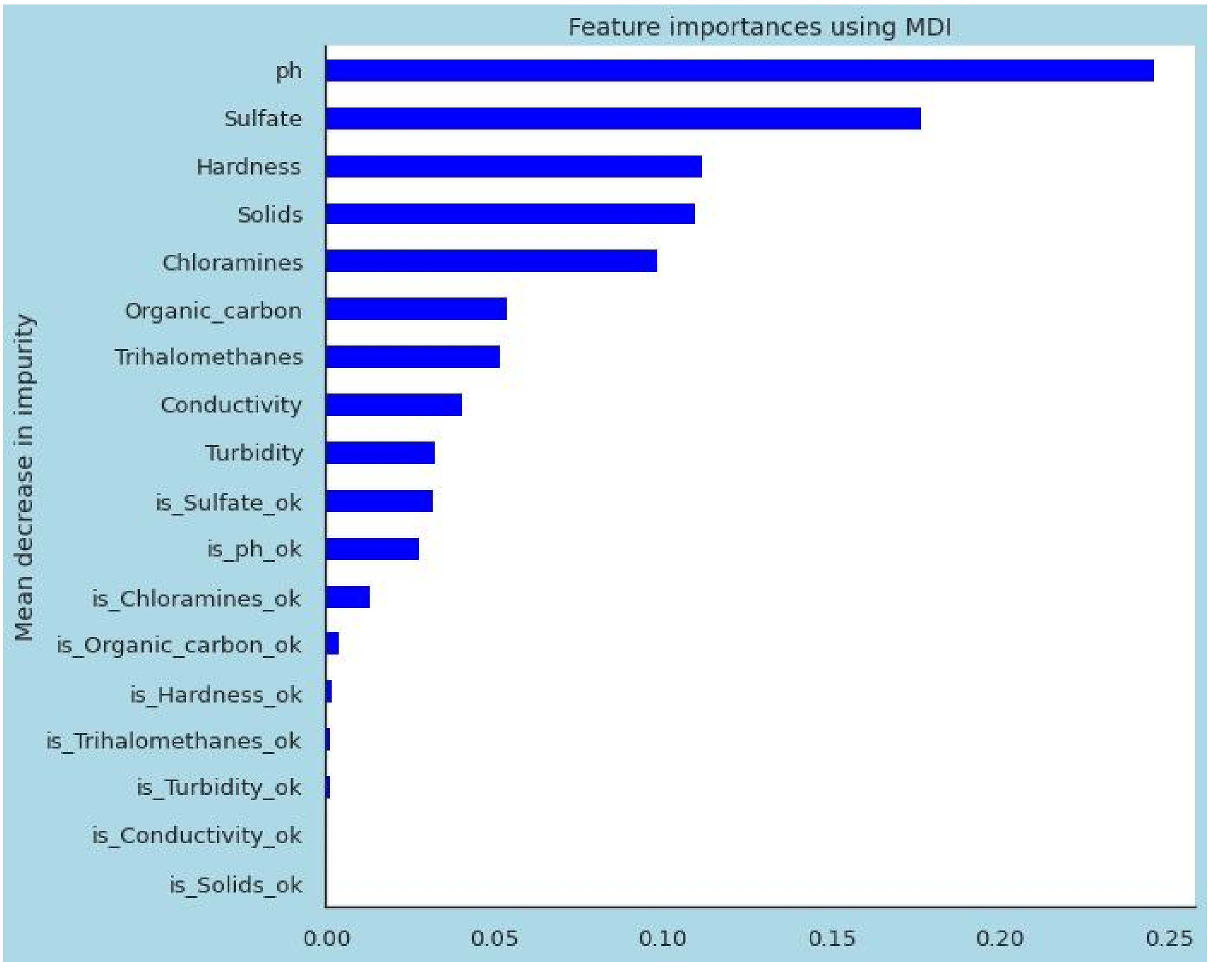
Observation:

- The ph feature has a large partial dependence at values between 6 and 8.5.
- For Hardness, the value of partial dependence rapidly increases around 210.

Feature importance based on mean decrease in impurity

After calculating the sum of the decrease of impurity at the point of splitting based on the corresponding feature in each tree, this Mean decrease Gini is the average of all tree values. This value

wi



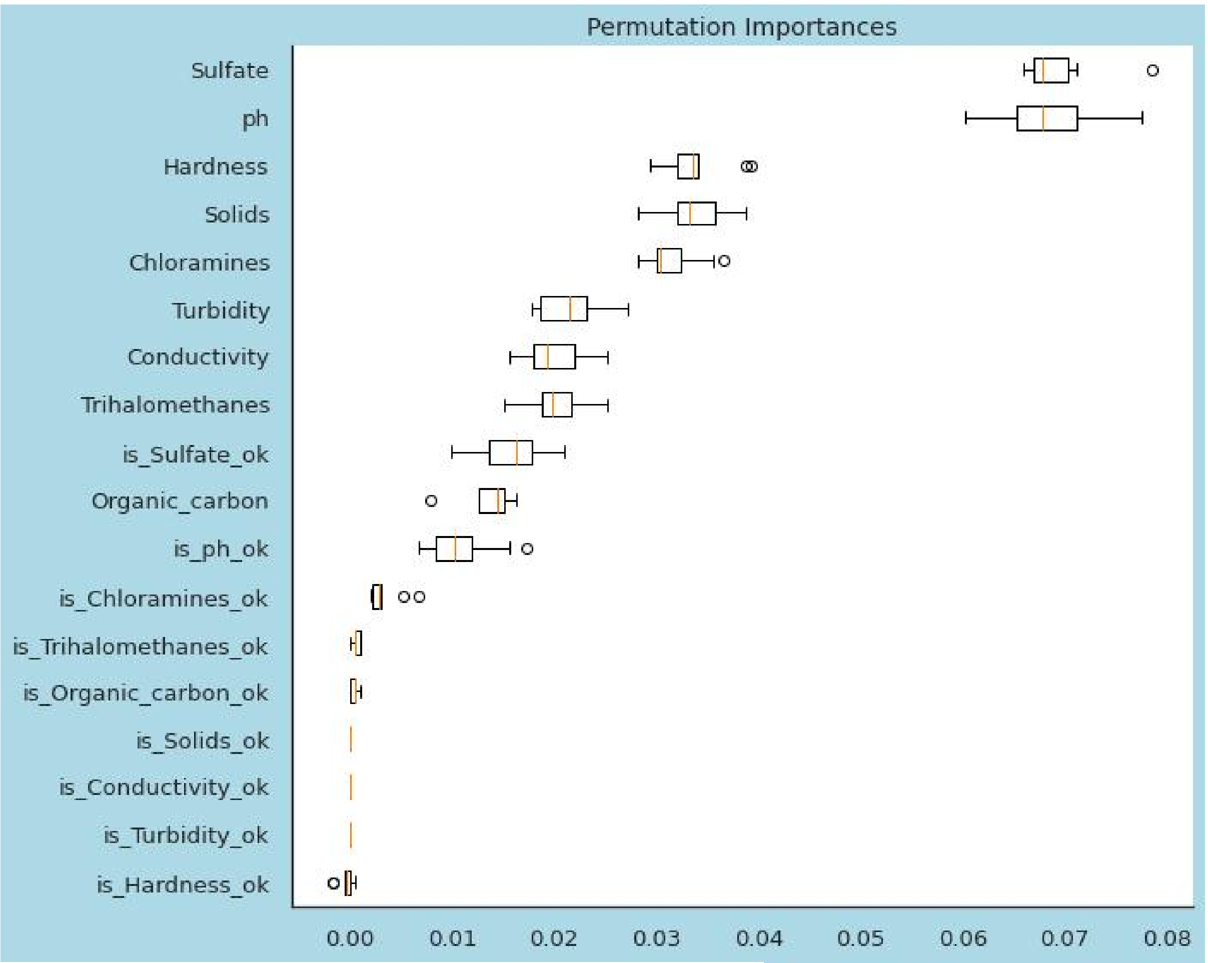
ll increase as the feature becomes important for the model to classify well.

Observation:

- ph and sulfate features were judged to be important features.

Feature importance based on feature permutation

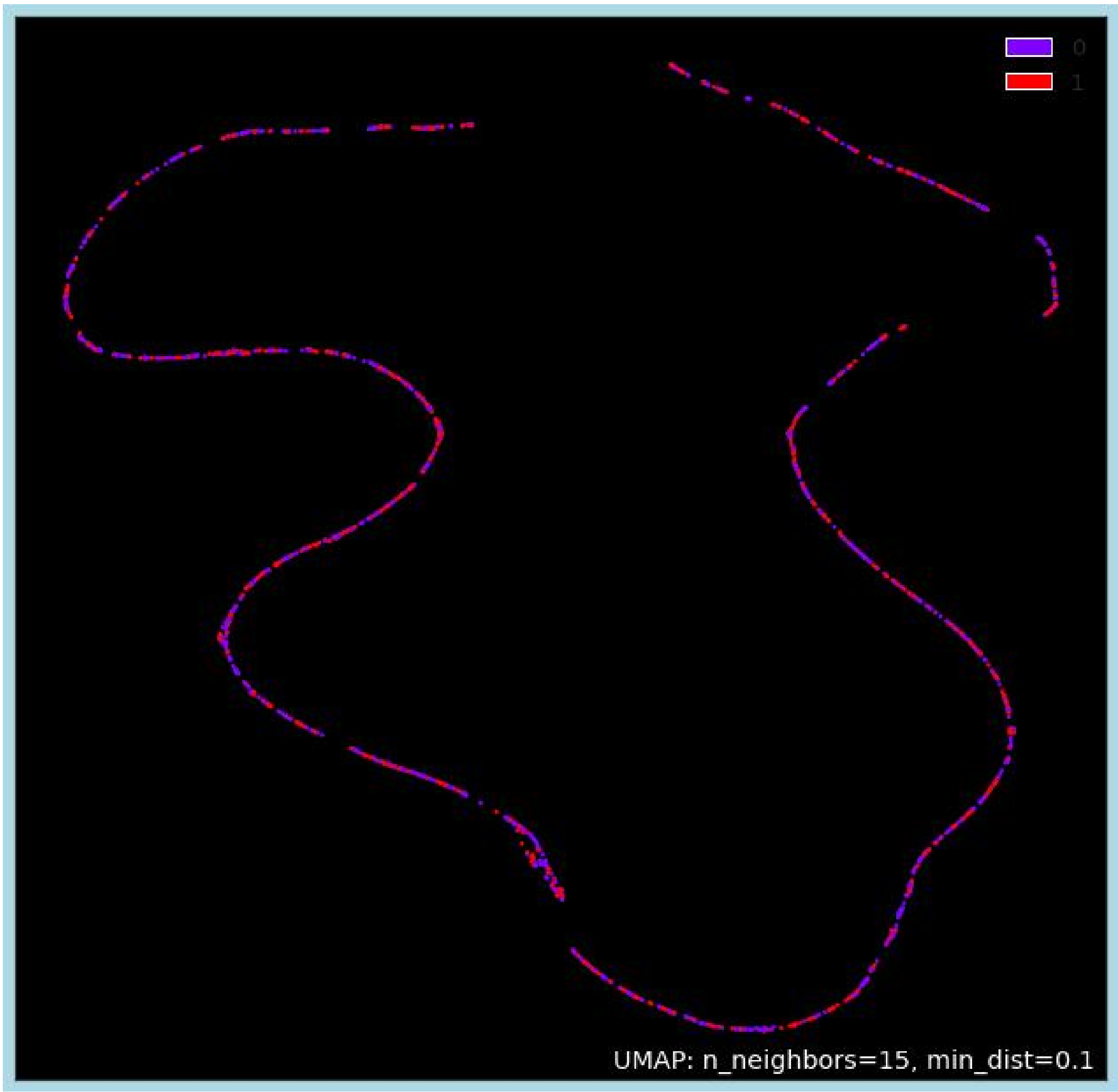
The estimator is required to be a fitted estimator. X can be the data set used to train the estimator or a hold-out set. The permutation importance of a feature is calculated as follows. First, a baseline metric, defined by scoring, is evaluated on a (potentially different) dataset defined by the X. Next, a feature column from the validation set is permuted and the metric is evaluated again. The permutation importance is defined to be the difference between the baseline metric and metric from permutating the feature column.



Observation:

- ph and sulfate features were judged to be important features.

Visualizing Training Dataset after Dimension Reduction



Observation:

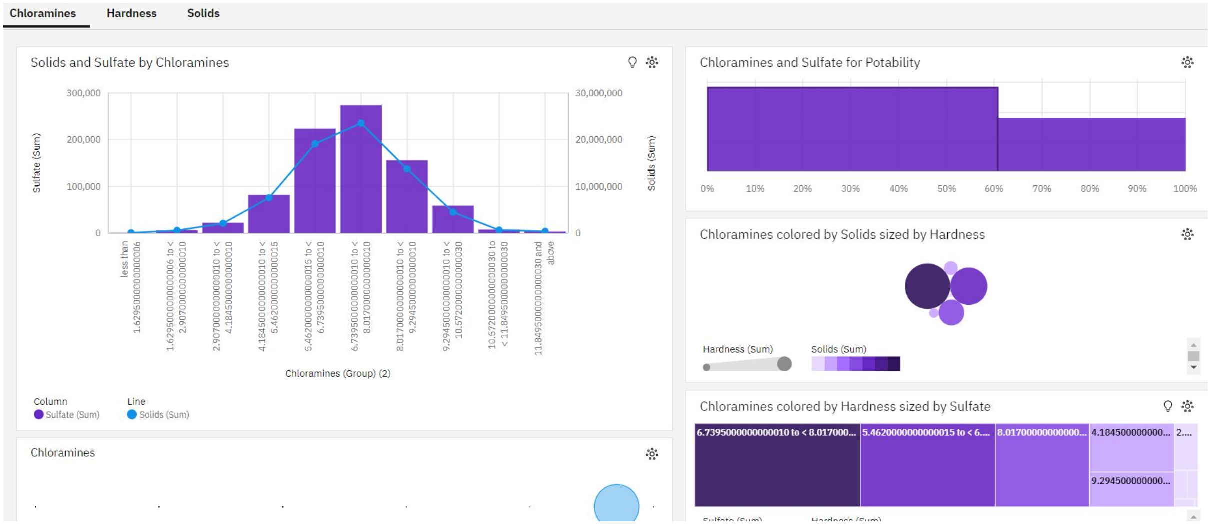
- A specific pattern or boundary is not visible.

Conclusion

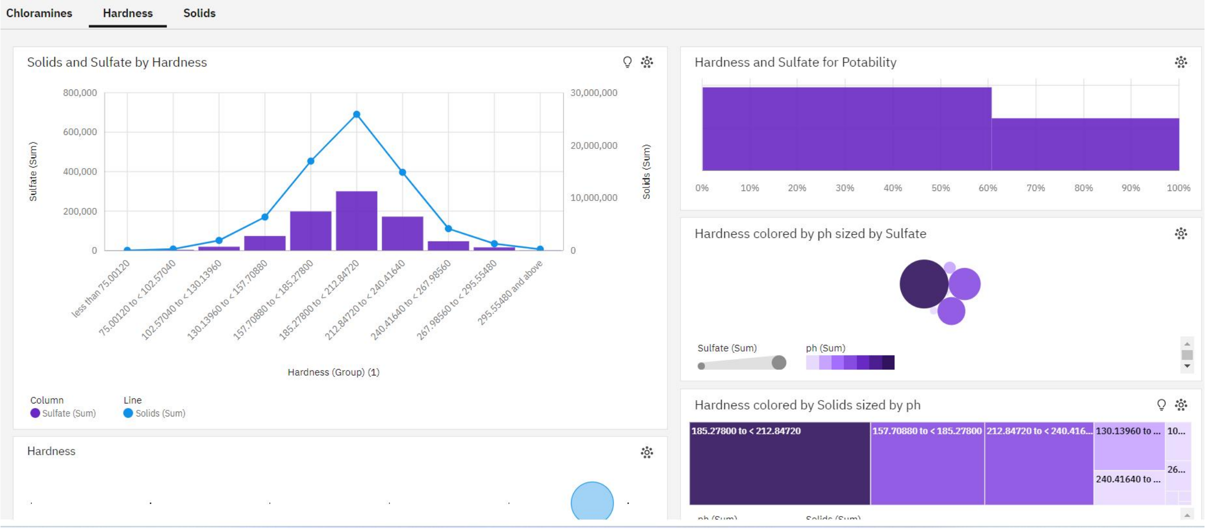
Thus implementation EDA and handing missing values performed and visualize the various parameters from dataset to made lot of insights Through charts and graphs and performed statistical methods to ensured.Standards of datasets.next will implement predictive modelling

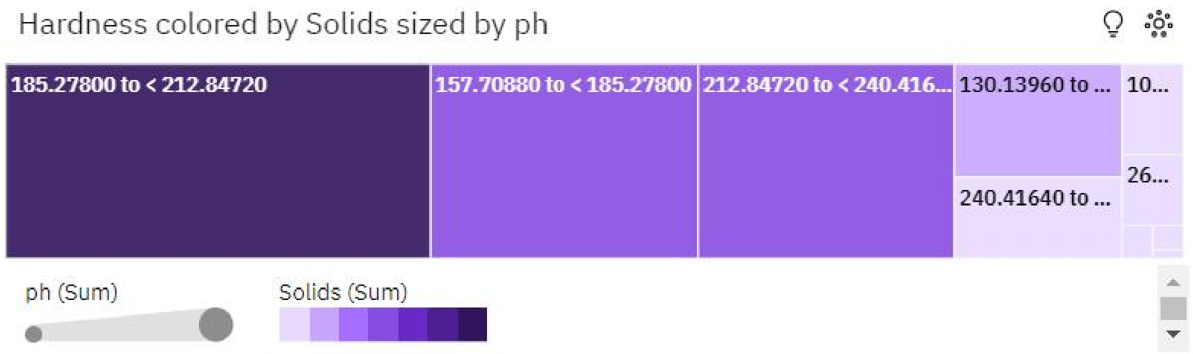
IBM Cognos Analytics

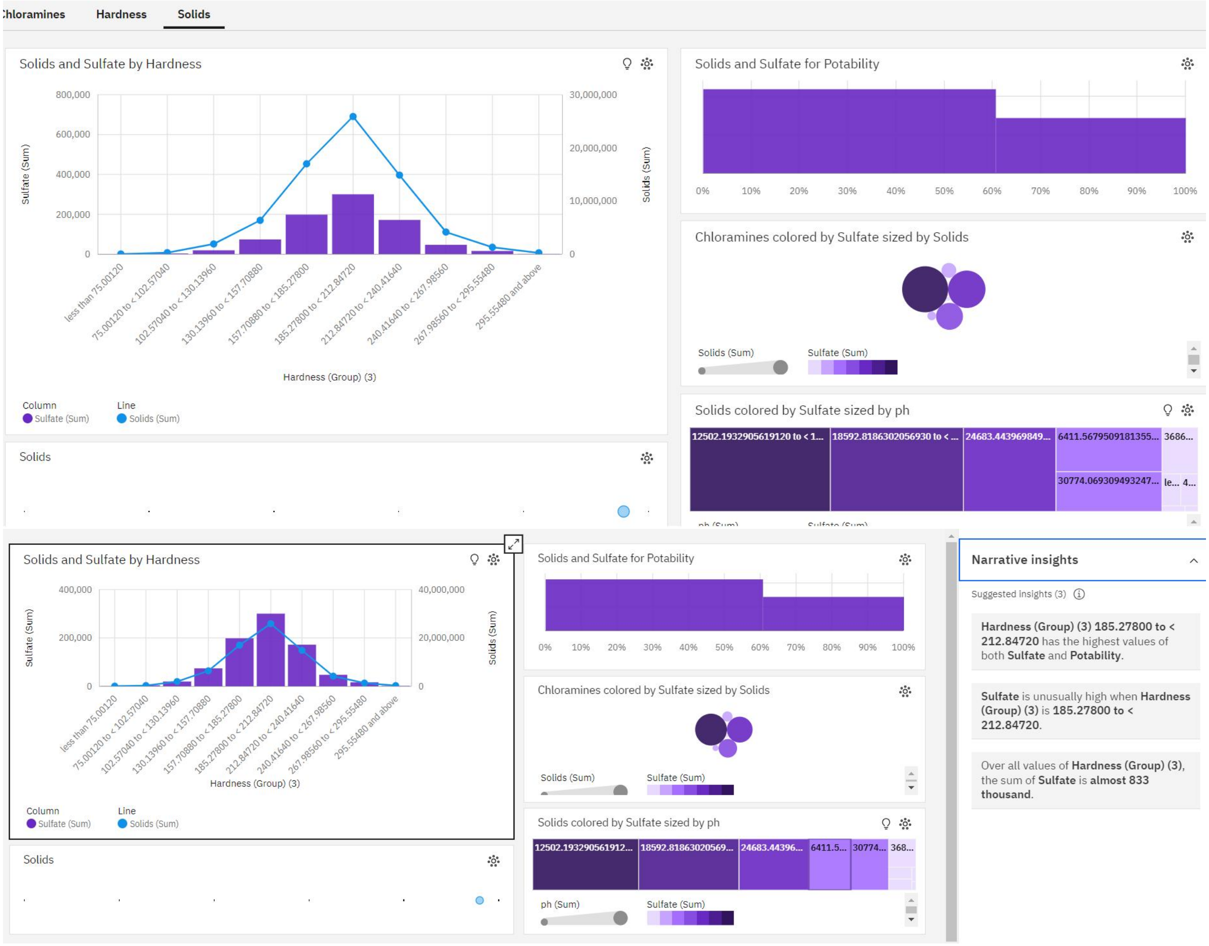
visualisation and Insights



Chloramines
Hardness
Solids
Solids and Sulfate by Chloramines
300,000
200,000
100,000
0
30,000,000
20,000,000
10,000,000
0
Sulfate (Sum)
Solids (Sum)
less than 1.629500000000006
1.629500000000006 to < 2.907000000000010
2.907000000000010 to < 4.184500000000010
4.184500000000010 to < 5.462000000000015
5.462000000000015 to < 6.739500000000010
6.739500000000010 to < 8.017000000000010
8.017000000000010 to < 9.294500000000010
9.294500000000010 to < 10.572000000000030
10.572000000000030 to < 11.849500000000030
11.849500000000030 and above
Chloramines (Group) (2)
Column
Sulfate (Sum)
Line
Solids (Sum)
Chloramines
Chloramines and Sulfate for Potability
0%
10%
20%
30%
40%
50%
60%
70%
80%
90%
100%
Chloramines colored by Solids sized by Hardness
Hardness (Sum)
Solids (Sum)
Chloramines colored by Hardness sized by Sulfate
6.7395000000000010 to < 8.017000...
5.4620000000000015 to < 6....
8.01700000000000...
4.184500000000...
2....
9.294500000000...
Sulfate (Sum)
Hardness (Sum)







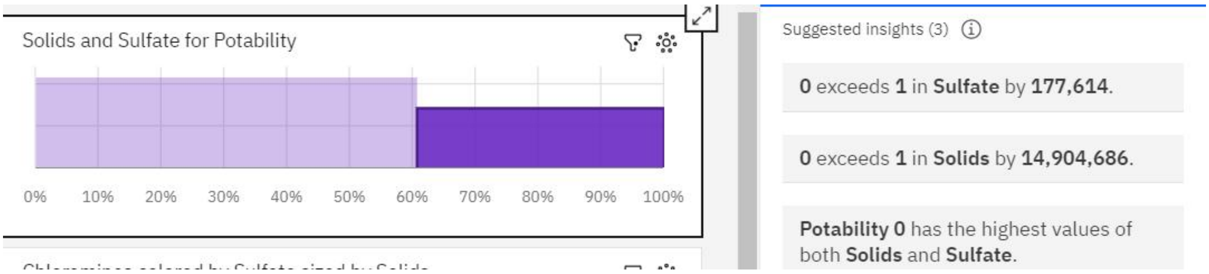
Narrative insights

Suggested insights (3)

Hardness (Group) (3) 185.27800 to < 212.84720 has the highest values of both **Sulfate** and **Potability**.

Sulfate is unusually high when **Hardness (Group) (3)** is **185.27800 to < 212.84720**.

Over all values of **Hardness (Group) (3)**, the sum of **Sulfate** is **almost 833 thousand**.

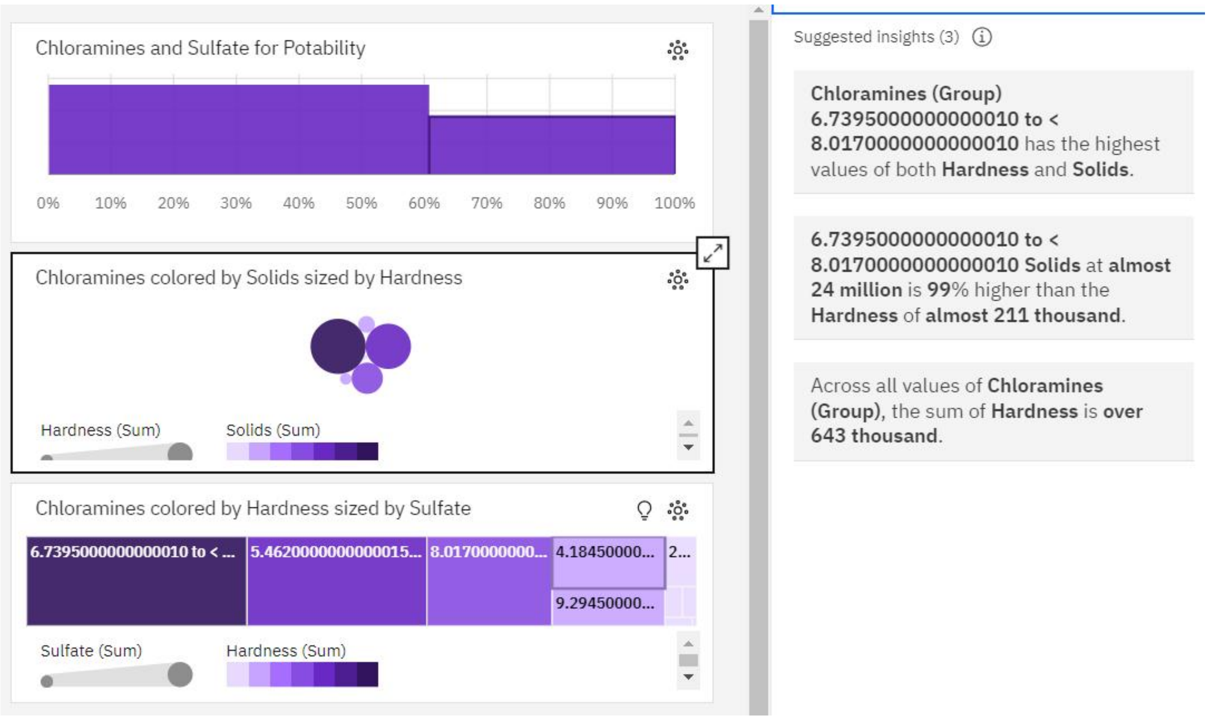

Solids and Sulfate for Potability
0%
10%
20%
30%
40%
50%
60%
70%
80%
90%
100%
Chloramines colored by Sulfate sized by Solids

Suggested insights (3)

0 exceeds **1** in **Sulfate** by **177,614**.

0 exceeds **1** in **Solids** by **14,904,686**.

Potability 0 has the highest values of both **Solids** and **Sulfate**.

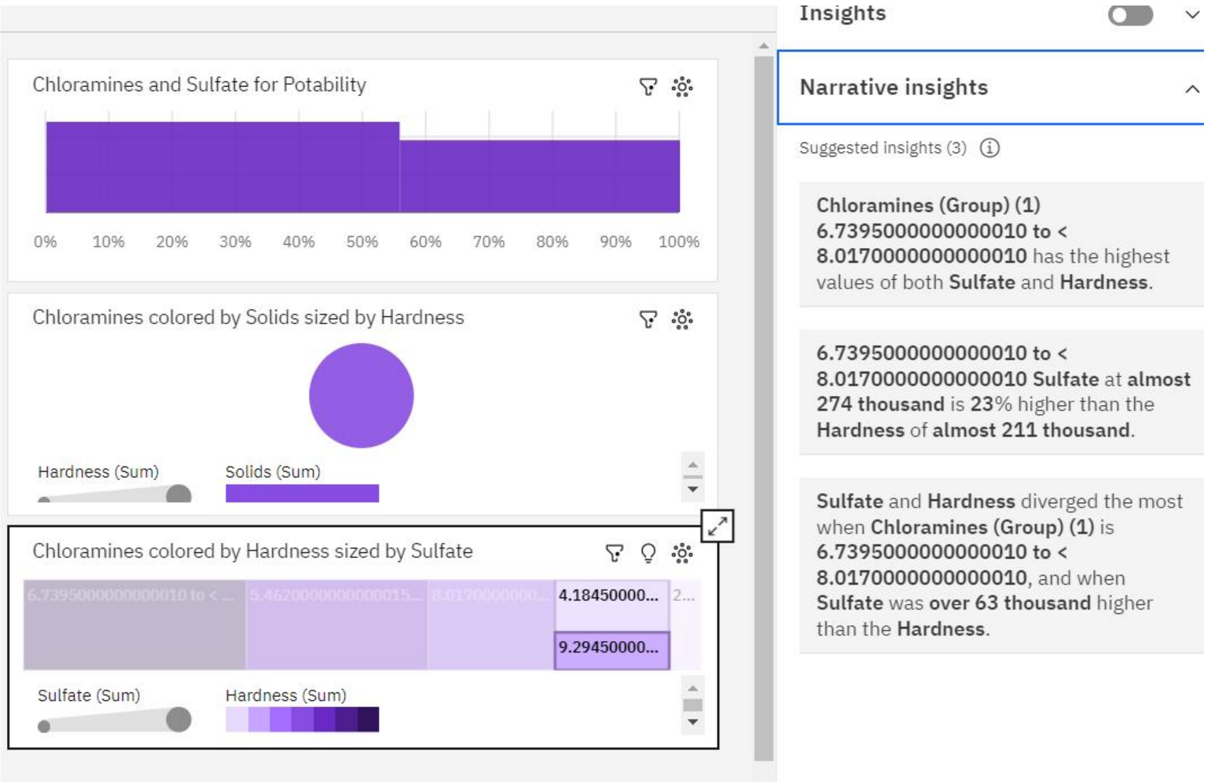


Suggested insights (3)

Chloramines (Group) 6.7395000000000010 to < 8.0170000000000010 has the highest values of both **Hardness** and **Solids**.

6.7395000000000010 to < 8.0170000000000010 Solids at **almost 24 million** is **99**% higher than the **Hardness** of **almost 211 thousand**.

Across all values of **Chloramines (Group)**, the sum of **Hardness** is **over 643 thousand**.



Insights

Narrative insights

Suggested insights (3)

Chloramines (Group) (1) 6.7395000000000010 to < 8.0170000000000010 has the highest values of both **Sulfate** and **Hardness**.

6.7395000000000010 to < 8.0170000000000010 Sulfate at **almost 274 thousand** is **23**% higher than the **Hardness** of **almost 211 thousand**.

Sulfate and **Hardness** diverged the most when **Chloramines (Group) (1)** is **6.7395000000000010 to < 8.0170000000000010**, and when **Sulfate** was **over 63 thousand** higher than the **Hardness**.