Explaining water potability results

(MTQ315 Project)

Yash Gohel(2018MT10773)

Under the guidance of
Prof. Niladri Chatterjee

Department of Mathematics
IIT Delhi

30th March 2022

Contents

1 Introduction **2**

2 Data description **2**

2.1 Parameters 2

2.2 Permissible range 3

2.3 Data visualization 3

3 Prediction model **4**

3.1 Data pre-processing 4

3.2 Model selection 4

4 Explainability **5**

4.1 SHAP and LIME 5

4.1.1 Explaining one point 5

4.1.2 Explaining the dataset 7

4.1.3 Some wrong interpretations 7

4.2 Counterfactual 8

4.2.1 Explaining one point 8

4.2.2 Some wrong interpretations 9

4.3 Conclusion 9

1 Introduction

Water is vital component of everyday life. We use water for many purposes such as drinking, generating electricity in power plants, transportation etc. Water quality is one of the most important factor for person's health. Clean water is very important for growth of any living organism. Drinking water with poor quality may result in poor health. Contaminated water results in transmission of diseases such as Cholera and Typhoid. According to WHO, 3 out of 4 people have ready access to safe drinking water. This number is pretty bad and other 25% person's health may be in danger. So, quality of drinking water is important factor in one's health. In this project we are checking potability of water on certain factors given to us.

2 Data description

In the data given to us we are given 9 parameters from which we are determining whether the water is potable or not.

2.1 Parameters

- **pH** - It is the pH value of the water. pH values range from 0-14.
- **Hardness** - Hardness of water is calculated as the concentration of Ca^{+2} and Mg^{+2} ions present in the water. It is measured in ppm(mg/L).
- **Solids** - Concentration of total dissolved solids in water. It is generally caused due to industrial sewage and waste being disposed in water bodies. It is measured in ppm.
- **Chloramines** - Water is disinfected using chlorine and chloramines. It is concentration of the chloramines(including chlorine) in water. It is measured in ppm.
- **Sulfate** - It is concentration of sulfate in the water. It is often caused from runoff from fertilized agricultural lands. It is also measured in ppm.
- **Conductivity** - Water is conductive due to presence of ions present in the water. It is measured in $\mu S/cm$ where S is siemens SI unit of conductivity.
- **Organic Carbon** - Total amount of carbon present in organic compounds in water. It is measured in ppm.
- **Trihalomethanes** - This chemicals forms when water is treated with chloramines for disinfection. It is measured in ppb(parts per billion).
- **Turbidity** - It is the quantity of solid matter present in water when water is in suspended state. It is measured in NTU(Nephelometric Turbidity Unit).

2.2 Permissible range

The following are permissible range given by WHO for the given parameters.

Sr. No.	feature	permissible range
1.	pH	6.5-8.5
2.	Hardness	below 200 ppm
3.	Solids	below 500 ppm
4.	Chloramines	below 4 ppm
5.	Sulfate	below 250 ppm
6.	Conductivity	200-800$\mu S/cm$
7.	Organic Carbon	below 25 ppm
8.	Trihalomethanes	below 80 ppb
9.	turbidity	below 4 NTU

2.3 Data visualization

Below are the graphs for the given parameters.

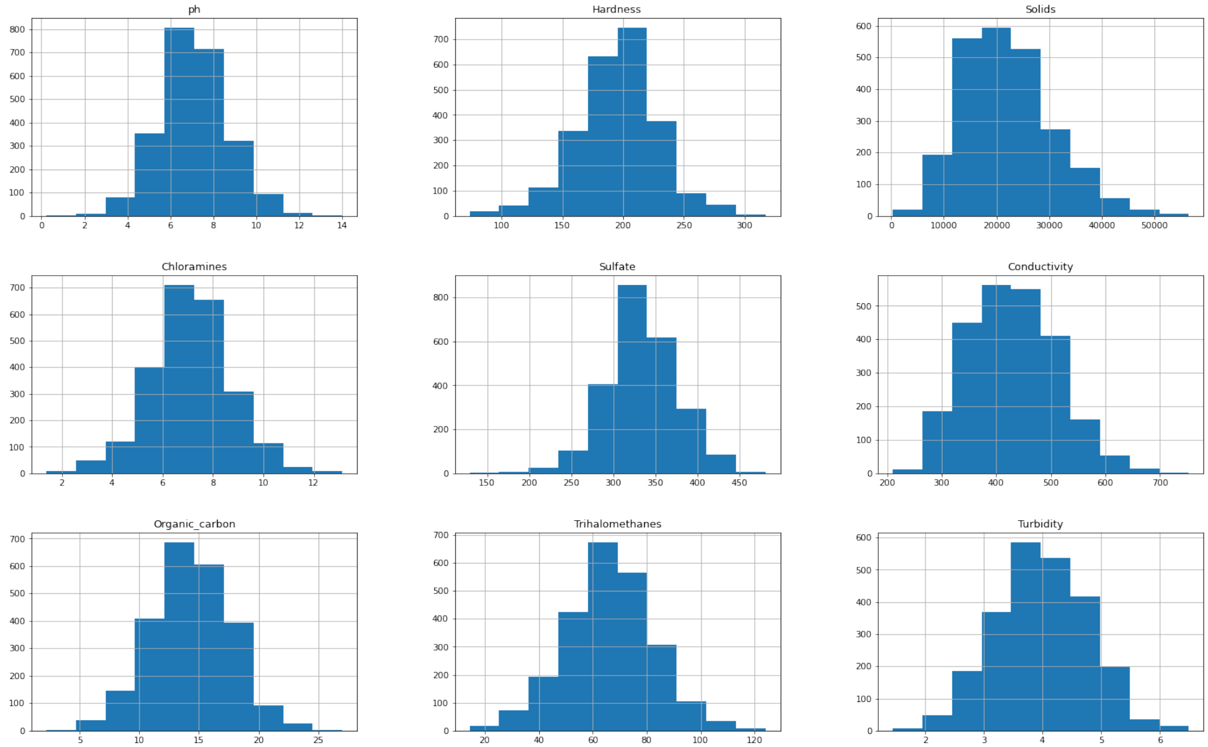


Figure 1: Distribution of the given parameters

In this distribution we can see that for some features such as conductivity and organic carbon we have almost all the values in the given permissible range by WHO. Whereas there are also features such as Solids and Chloramines where only few values are there in the permissible range given by WHO.

3 Prediction model

3.1 Data pre-processing

We were given with 3276 data points in total for the given problem. Out of this only 2011 data points were containing all the values whereas other points were having some values as NaN. We dropped all the points which had one of the values as NaN. The reason behind this was if we had generated some value in place of that NaN values with the help of some imputer then there was possibility that the result of potability may have changed. As it is very imporatant decision we dropped the points containing NaN values.

In the remaining data, there were 1200 values for which potability was 0 whereas only 811 points for which the potability was 1. It is great imbalance and model which we are training for prediction may interpret wrong if we do not resolve this imbalance. This imbalance was resolved using the resampler provided by sklearn library.

We used two strategies while imbalancing the data. First was to upscale the data, i.e. increasing the data points. In this case we increase the data by selecting the same data from the dataset. Opposite to this case we decrease the data points for which the value of potability is given 0. We downscale it and remove some data points randomly using resampler.

3.2 Model selection

We selected two models for task of prediction namely Random forest and Support Vector Machine(SVM). We splitted the data generated after pre-processing in training and testing such that 20% of our data will be used for testing the model. We trained both models using the training data on both upsampled and downsampled data. The results(accuracy values) can be summarized as below,

	Upsampled Data	Downsampled Data
SVM	57.25	50.625
Random Forest	82.23	81.732

From the above values we can say that there is minor difference between accuracy when data is upsampled or downsampled. Hence, we use model with upsampled data for prediction task. We can see that clearly random forest is working way better than SVM model. Hence, we use Random forest with upsampled data as our model for prediction task. After resampling the data overfitting was happening as the accuracy on training data was around 97%.

Confusion matrix for random forest on upsampled data with best accuracy(above accuracy is average) of 85.21% is given below

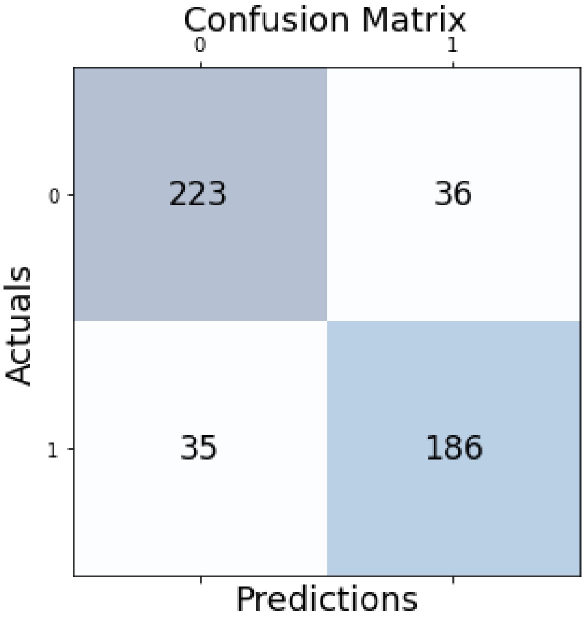


Figure 2: Confusion Matrix

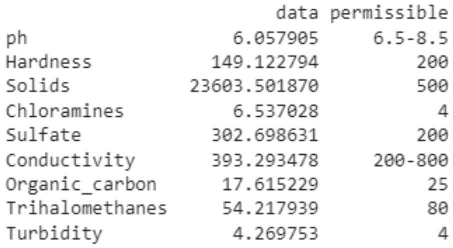
4 Explainability

4.1 SHAP and LIME

We used SHAP values and LIME values for global and local explanations. We used the *shap* library in python and *interpret* to use lime.

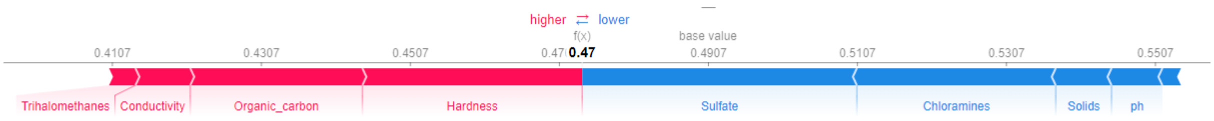
4.1.1 Explaining one point

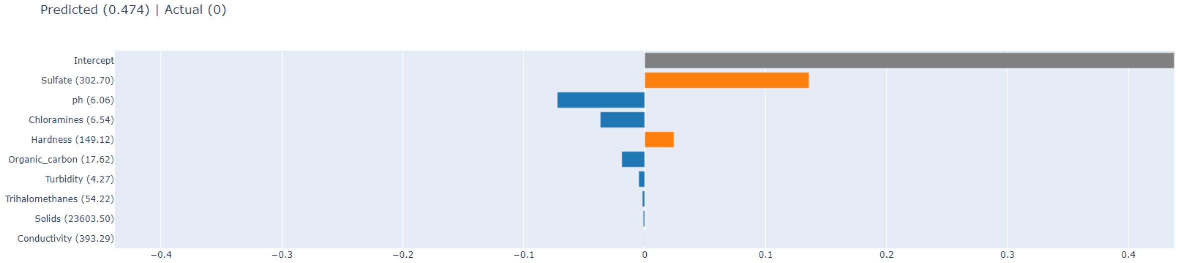
The point given to us is as given below,



```
                         data permissible
ph                   6.057905     6.5-8.5
Hardness           149.122794         200
Solids           23603.501870         500
Chloramines          6.537028           4
Sulfate            302.698631         200
Conductivity       393.293478     200-800
Organic_carbon      17.615229          25
Trihalomethanes     54.217939          80
Turbidity            4.269753           4
```

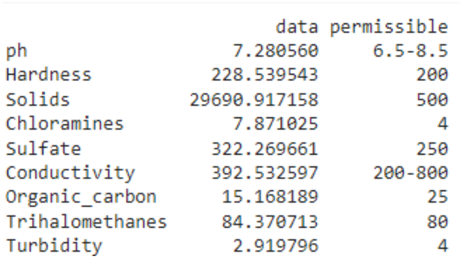
SHAP and LIME values for this point is as follows,





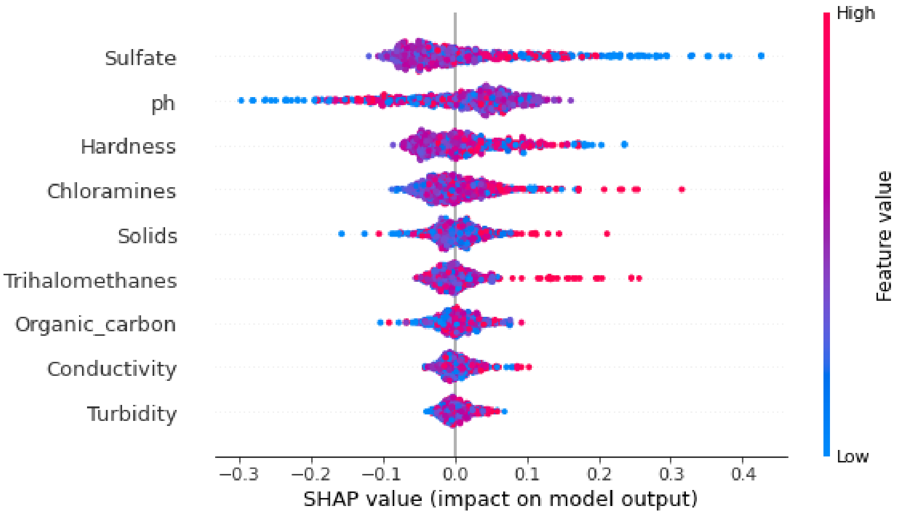
In the above figure we can see that SHAP is able to explain pretty well on the WHO permissible ranges while LIME is not able to explain features such as Sulfate, Trihalomethanes and Conductivity. Same trend was followed when we tried to explain some data points with the help of both SHAP and LIME values. Hence, for this model SHAP values were able to explain the results in much better way than LIME values.

One of the facts that one may wonder is that difference between actual value and permissible range is much greater in case of Solids when compared with Sulfate yet Sulfate is contributing more than Solids. The reason with which one may explain is that only one value is in permissible range according to WHO data when we see the value of Solids. Hence we can say that model is not able to learn well on Solids. We tried to plot the graph between value of solids and potability and check whether we can find any trend that between which values of Solids potability value 1 was coming more than 0. But we were not able to find any trends hence we were not able to explain the data when it comes to Solids. Similar trend follows in case of Chloramines.



```
                       data  permissible
ph                 7.280560      6.5-8.5
Hardness         228.539543          200
Solids         29690.917158          500
Chloramines        7.871025            4
Sulfate          322.269661          250
Conductivity     392.532597      200-800
Organic_carbon    15.168189           25
Trihalomethanes   84.370713           80
Turbidity          2.919796            4
```

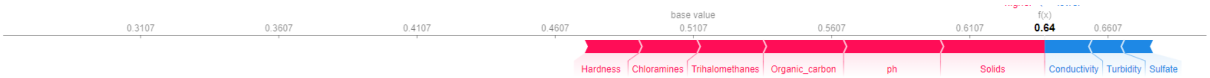
4.1.2 Explaining the dataset



From the above figure we can see that because some values are not falling in the range of permissible values hence we are not able to explain them. We are able to explain ph as for very high or very low ph value the impact is coming negative which is true in the case of ph.

4.1.3 Some wrong interpretations

The point we are looking for is as the following,

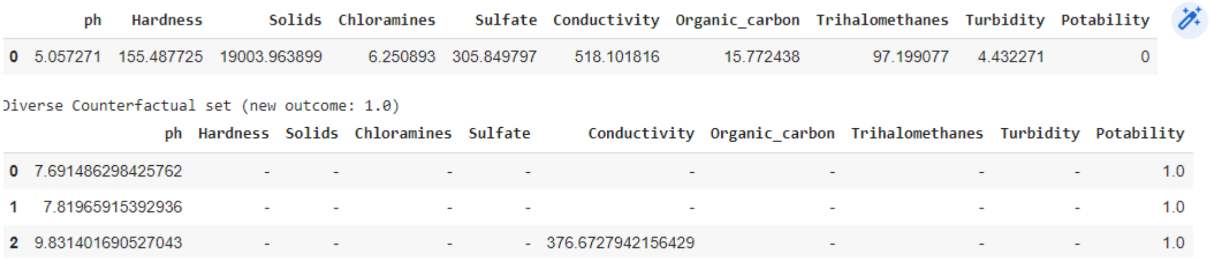


From the values we can see that Hardness are not coming in permissible range yet it is pushing the value to be 1. Similarly there are many values which SHAP is not able to explain. Hence, in SHAP also there are some wrong interpretations.

4.2 Counterfactual

We have used DiCE (Diverse Counterfactuals Explanations) for generating counterfactuals. Counterfactuals are explanations that tells us that after changing which feature values we will get change at the output.

4.2.1 Explaining one point



	ph	Hardness	Solids	Chloramines	Sulfate	Conductivity	Organic_carbon	Trihalomethanes	Turbidity	Potability
0	5.057271	155.487725	19003.963899	6.250893	305.849797	518.101816	15.772438	97.199077	4.432271	0

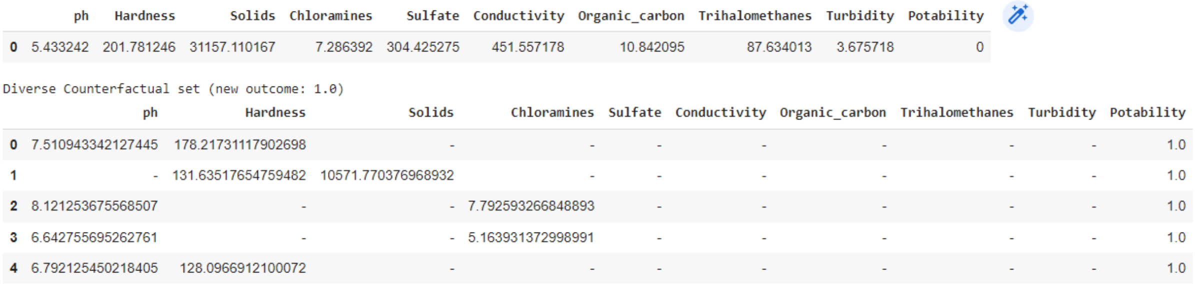
Diverse Counterfactual set (new outcome: 1.0)

	ph	Hardness	Solids	Chloramines	Sulfate	Conductivity	Organic_carbon	Trihalomethanes	Turbidity	Potability
0	7.691486298425762	-	-	-	-	-	-	-	-	1.0
1	7.81965915392936	-	-	-	-	-	-	-	-	1.0
2	9.831401690527043	-	-	-	-	376.6727942156429	-	-	-	1.0

Figure 3: All values varied

From the above figure we can see that when we are varying all the values then we are getting these counterfactuals. In these we can see that first 2 are still good counterfactuals as they are able to explain the ph range in permissible values but last counterfactual is somewhat wrong. This may be due to data given to us is not according to the WHO guidlines.

We know that some features are interdependent on each other. For example the value of Trihalomethanes concentration depends on the amount of Chloramines added, i.e. value of Chloramines. We know that conductivity of water is due to ions present in it. So, if we are reducing the hardness of the water it will have effect on conductivity of the water as the ions present in the water are decreasing. So, in the next step we varied features that are not interdependent.



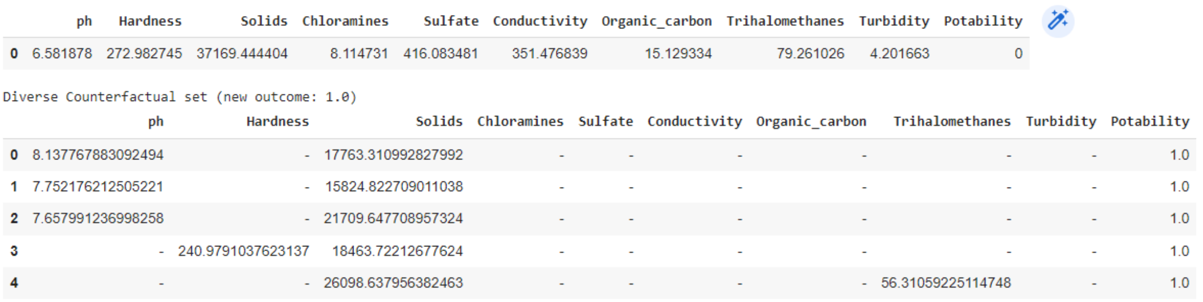
	ph	Hardness	Solids	Chloramines	Sulfate	Conductivity	Organic_carbon	Trihalomethanes	Turbidity	Potability
0	5.433242	201.781246	31157.110167	7.286392	304.425275	451.557178	10.842095	87.634013	3.675718	0

Diverse Counterfactual set (new outcome: 1.0)

	ph	Hardness	Solids	Chloramines	Sulfate	Conductivity	Organic_carbon	Trihalomethanes	Turbidity	Potability
0	7.510943342127445	178.21731117902698	-	-	-	-	-	-	-	1.0
1	-	131.63517654759482	10571.770376968932	-	-	-	-	-	-	1.0
2	8.121253675568507	-	-	7.792593266848893	-	-	-	-	-	1.0
3	6.642755695262761	-	-	5.163931372998991	-	-	-	-	-	1.0
4	6.792125450218405	128.0966912100072	-	-	-	-	-	-	-	1.0

Figure 4: pH, Solids, Hardness and Chloramines varied

In the above figure almost every counterfactual is fine but counterfactual 2 is bit confusing. It is because if we increase the value of Chloramines in water and pH of the water it may increase the concentration of trihalomethanes. Because of that water may become not potable.



	ph	Hardness	Solids	Chloramines	Sulfate	Conductivity	Organic_carbon	Trihalomethanes	Turbidity	Potability
0	6.581878	272.982745	37169.444404	8.114731	416.083481	351.476839	15.129334	79.261026	4.201663	0

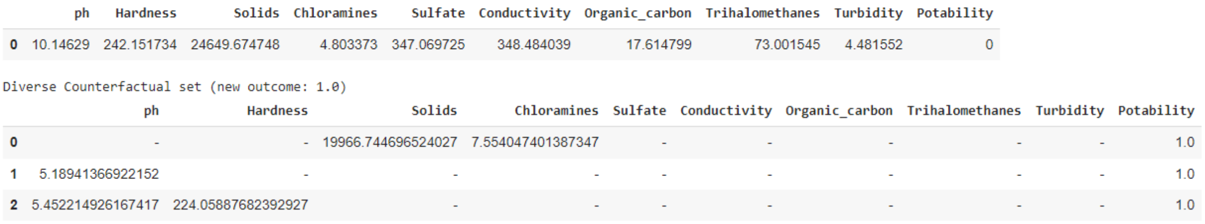
```
Diverse Counterfactual set (new outcome: 1.0)
```

	ph	Hardness	Solids	Chloramines	Sulfate	Conductivity	Organic_carbon	Trihalomethanes	Turbidity	Potability
0	8.137767883092494	-	17763.310992827992	-	-	-	-	-	-	1.0
1	7.752176212505221	-	15824.822709011038	-	-	-	-	-	-	1.0
2	7.657991236998258	-	21709.647708957324	-	-	-	-	-	-	1.0
3	-	240.9791037623137	18463.72212677624	-	-	-	-	-	-	1.0
4	-	-	26098.637956382463	-	-	-	-	56.31059225114748	-	1.0

Figure 5: pH, Solids, Hardness and Trihalomethanes varied

From above two counterfactuals we can see that it is able to give some valid explanations for given data according to guidelines.

4.2.2 Some wrong interpretations



	ph	Hardness	Solids	Chloramines	Sulfate	Conductivity	Organic_carbon	Trihalomethanes	Turbidity	Potability
0	10.14629	242.151734	24649.674748	4.803373	347.069725	348.484039	17.614799	73.001545	4.481552	0

```
Diverse Counterfactual set (new outcome: 1.0)
```

	ph	Hardness	Solids	Chloramines	Sulfate	Conductivity	Organic_carbon	Trihalomethanes	Turbidity	Potability
0	-	-	19966.744696524027	7.554047401387347	-	-	-	-	-	1.0
1	5.18941366922152	-	-	-	-	-	-	-	-	1.0
2	5.452214926167417	224.05887682392927	-	-	-	-	-	-	-	1.0

The above counterfactual is wrong as we can see that it is not able to give valid reasons. If we keep pH very low or very high, the water will not be potable but it is giving that water is potable.

4.3 Conclusion

We can say that SHAP values and counterfactuals were good methods to get explanations on the given data. The data given to us was small and imbalanced hence we were not able to explain the results well.