Joël Davila Estevez

Estimating the Hazardousness of Urban Waters from the Pollutant Concentrations:

Abstract

This work aims to define the hazardousness of urban waters using concentrations of pollutant elements (such as arsenic, copper) or bacteria and viruses based on machine-learning methods (Binary classification, logistic regression and Random Forest).

1. Introduction

The machine learning is an interesting method for the analysis of aquatic environments, because it allows to reduce the number of surveys and thus to save more time. In addition, the addition of complex mathematical functions, allows to see non-linear relationships that are difficult to observe in a conventional way (Najah Ahmed et al., 2019).Moreover, the models used in machine learning have better accuracy than previously used models not based on ML (Haghiabi, Nasrolahi, Parsaie, 2018).

2. Data

As for my master work I don't have any data yet, I tried to take data corresponding to figures that I could have to do in my future job. So, I downloaded some data from Kaggle. These are data that represent the water quality in the urban environment. It gives us the information about the concentration in water [mg/L] of 20 elements and a column that gives us the information if it is safe or not with a Boolean [0 = it's not safe / 1 = it's safe].

aluminium	ammonia	arsenic	barium	cadmium	chloramine	chromium	copper	flouride	bacteria	
1.65	9.08	0.04	2.85	0.007	0.35	0.83	0.17	0.05	0.2	
viruses	lead	nitrates	nitrites	mercury	perchlorate	radium	selenium	silver	uranium	is_safe
0	0.054	16.08	1.13	0.007	37.75	6.78	0.08	0.34	0.02	1

Figure 1. Data used with the first row as example, the concentration are in [mg/L]

2.1. Visualization

As this data is obtained from the internet, it is important to analyse it before launching into machine learning, to check that there is no problem with one of the variables or variables that are too correlated with each other. So a visualization has been made of the data with histograms, to get a picture of the kind of concentration we can have. Here is an example of the histograms I was able to obtain.

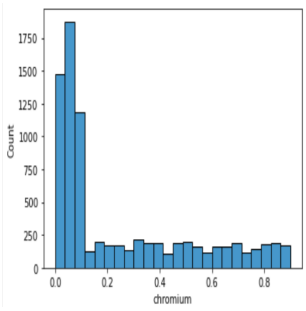


Figure 2. Chromium concentration [mg/L] histogram on the 7996 data

Not all data has the same fineness or order of magnitude, these are issues that need to be considered if there is a problem with my future prediction, as this can create bias. Therefore, to get a better idea of the different variables, it is important to know the mean, variance, maximum and minimum of each of them.

	Average [mg/L]	Variance[mg/L]	Minimum [mg/L]	Maximum [mg/L]
aluminium	0.666396	1.600842	0.0	5.05
arsenic	0.161477	0.063815	0.0	4.94
barium	1.567928	1.479024	0.0	0.13
cadmium	0.042803	0.001299	0.0	8.68
chloramine	2.177589	6.589741	0.0	0.9
chromium	0.247300	0.073250	0.0	2.0
copper	0.805940	0.427133	0.0	1.5
flouride	0.771646	0.189570	0.0	1.0
bacteria	0.319714	0.108554	0.0	1.0
viruses	0.328706	0.142952	0.0	1.0
lead	l0.099431	0.003383	0.0	0.2
nitrates	9.819250	30.709663	0.0	19.83
nitrites	1.329846	0.328598	0.0	2.93
mercury	0.005193	0.000009	0.0	0.01
perchlorate	16.465266	312.855455	0.0	60.01
radium	2.920106	5.394749	0.0	7.99
selenium	0.049684	0.000828	0.0	0.1
silver	0.147811	0.020609	0.0	0.5
uranium	0.044672	0.000724	0.0	0.09

Figure 3. Data analysis of the different concentration of the elements

As it is elements that can act in synergy, it is interesting to see how they will interact with each other. To visualise, these synergies a correlation matrix has been produced to determine whether elements can be removed.

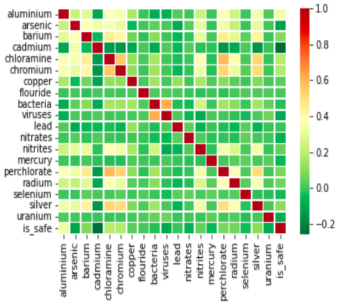


Figure 4. Correlation matrix

Looking at this figure 4, it seems that viruses and bacteria have a strong correlation with perchlorate and chloramine, it could be interesting to see in the literature if there is a synergy in these elements in the water. However, in view of the correlation with the "is safe" variable, it does not seem obvious that any variable should be deleted.

3. Methodologies

Firstly, the data has been copied in two parts, one part is the different elements that will be useful to predict the variable of interest which is the "is safe" variable, the Boolean that gives the information if the water is safe or not. Subsequently, the data was divided into three, a training, a validation and a test set. For the length of each set, It was taken 60% for the training and 20% for the validation and 20% for the test.

3.1. Logistic regression

The first model used is the logistic regression, it is known to be very efficient for binary classification problems. Moreover, it does not require a linear relationship between input variables and outputs, which is certainly the case in the pollutant domain.

3.2. Random Forest

Random forest (RF) is a well-known machine learning algorithm that randomly creates different decissional trees from the data and averages them to obtain interesting prediction results, and seems to have good performance when predicting binary classification.

3.3. Hyperparameter

To know the performance of my hyperparameter two performance metrics have been used the "mean test score" and the "std test score". A grid search cross validation was therefore performed, using all possible combinations of hyperparameters. This search was performed on the validation set.

3.3.1. Logistic regression

The solver is an important choice as it will dictate the optimisation of our algorithm. The penalty will define how the model reacts to an error. The C value, on the other hand, gives information on the regularisation of the model, the higher the value the less the regularisation.

Hyperparameter	Values	Best
Solvers	newton-cg, lbfgs, liblinear	liblinear
Penalty	l2,l1',elasticnet	l1
C value	100, 10, 1.0, 0.1, 0.01	100

The metrics obtained with the best hyperparameters are 0.910 for the mean test score and 0.0158 for the std test score.

3.3.2. Random Forest

Numbers of estimator indicates the number of trees in the forest while the maximum of features gives information on the splitting of the trees

Hyperparameter	Values	Best
Number of estimators	10, 100, 1000	1000
Maximum of features	'sqrt', 'log2'	Log2

The metrics obtained with the best hyperparameters are 0.938 for the mean test score and 0.0131 for the std test score.

4. Results and Discussions

4.1. Logistic regression

Looking at the performance of my data with the logistic regression, it seems satisfactory.

Accuracy of logistic regression classifier on test set:	0.90
Accuracy of logistic regression classifier on validate set:	0.91
Accuracy of logistic regression classifier on train set:	0.90

But focusing on the two classes, we notice that we have a problem on the prediction. For this purpose, a normalized matrix confusion was performed. It gives information on the number of true positive, false positive, false negative and true negative.

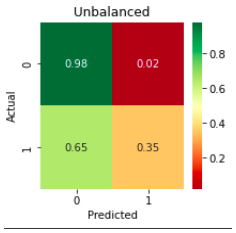


Figure 5. normalized matrix confusion of unbalanced data

It can be seen that this is good information on the true positives. However, there is a big mistake on the false negatives. As can be seen from my various results, it seems that the project is good at predicting whether water is unsafe. But it is bad at describing water that is safe. Of course, it is better to have accurate data to know if the water is dangerous. However, it is necessary to improve it to be able to better predict whether the water is safe.

To do this, it is necessary to rely on data. The data is unbalanced, there are more samples of dangerous water than safe water.

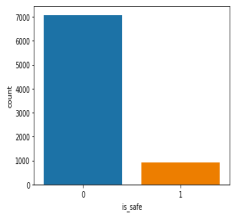


Figure 6. Count of "is safe"

To overcome this problem two methods were made in parallel. The first one was an oversampling and the second one an undersampling. The data were oversampled and undersampled randomly with the aim of having the same number of data with safe and unsafe waters

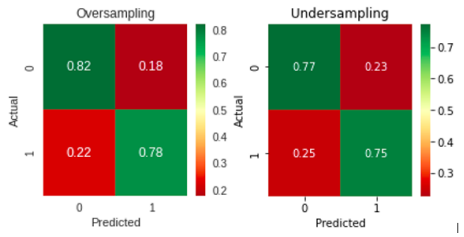


Figure 7. normalized matrix confusion of oversampled and undersampled data

Looking at these different results we can see that oversampling is better than undersampling in every aspect. Moreover, it is better than the unbalanced data to describe the waters when they are safe. However, there is a loss of performance when it comes to knowing if our water is dangerous. So, it is difficult to say which method is best to use, if we really want to be sure if our water is dangerous it is better to use our basic data, however if we want better overall performance it is better to use data with oversampling. However, to know if our model is performing well, it is useful to use another model. To do this, the random forest was used.

4.2. Random Forest

As it is hard to know which data are going to be the best, the random forest was performed on the unbalanced and oversampled data.

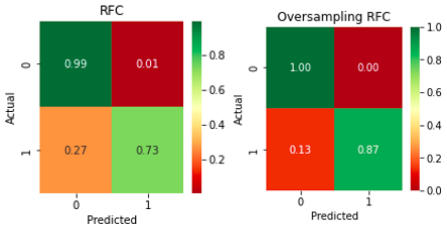
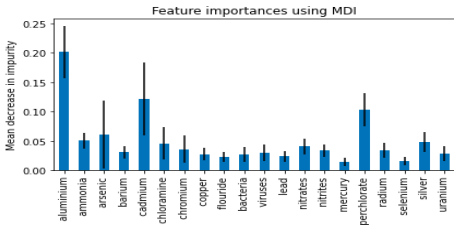


Figure 8. normalized matrix confusion of RFC with unbalanced data and oversampled data

It seems that the random forest has better results than the logistic regression. Moreover, oversampling seems to be positive in this case in both cases, for defining when the water is dangerous and for defining when the water is not dangerous. However, in view of the results of when our water is polluted, there is doubt of overfit.

To improve the different models it is possible that not all the different variables that have been put in are necessary and that dangerousness is explained by fewer variables. To do this, a feature importance was performed using the Mean Importances in Impurity (MID)



5. Conclusion

We notice that machine learning can be a very efficient tool to determine water quality, however, it is possible that it forgets some parameters that appear in a rarer way in our concentrations, so it would be necessary to study more extensively each pollutant separately.

6. References

Haghiabi, A. H., Nasrolahi, A. H., Parsaie, A. (2018). Water quality prediction using machine learning methods. Water Quality Research Journal, 53(1), 3-13. https://doi.org/10.2166/wqrj.2018.025

Najah Ahmed, A., Binti Othman, F., Abdulmohsin Afan, H., Khaleel Ibrahim, R., Ming Fai, C., Shabbir Hossain, M., ... Elshafie, A. (2019). Machine learning methods for better water quality prediction. Journal of Hydrology, 578, 124084. https://doi.org/10.1016/j.jhydrol.2019.124084

7. Annexes

Link of the Code :

shorturl.at/oqsxM

Link of the data :

https://www.kaggle.com/datasets/mssmartypants/water-quality