

Image: Unslash/Anderson Rian

**Water Quality Classification**

**Table of Contents**

[**Water Quality Classification** 1](#_Toc117155249)

[**Water is Life** 2](#_Toc117155250)

[**About the Dataset** 2](#_Toc117155251)

[**Cleaning the Data** 4](#_Toc117155252)

[**Pre-processing the Data** 5](#_Toc117155253)

[**Fitting the Model** 6](#_Toc117155254)

[**Predicting with the Model** 7](#_Toc117155255)

[**Evaluating the Model** 8](#_Toc117155256)

[**Accuracy of the Model** 10](#_Toc117155257)

[**Python Code** 11](#_Toc117155258)

**Water is Life**

Water is essential to life. It is part of human existence and daily activities. To be safe for drinking, the water must satisfied some criteria like being tasteless, odourless and colourless. But with recent events and human activities, the criteria above can no longer satisfy how safe water can be for drinking or use as it comes polluted. Impurities like ammonia, bacteria, copper, viruses, mercury, lead and a host of other are not visible to the bare eyes.

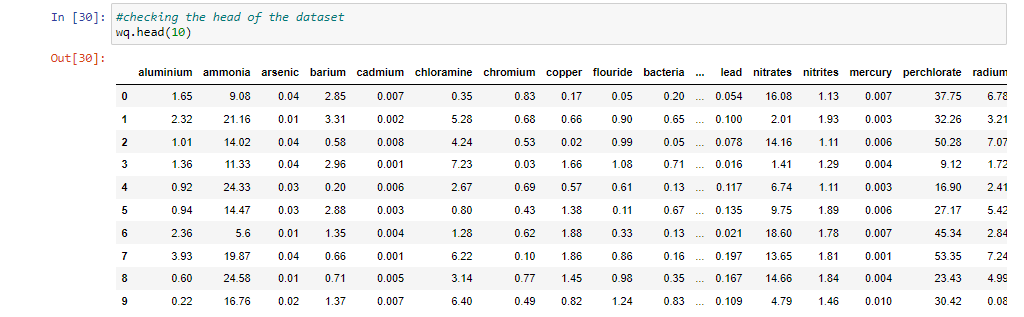
But with laboratory testing, they can be detected if present in the water sample and based on acceptable values, the water can be determine if its fit for use or not and classification machine learning model can assist to classify into weather is safe or not safe

The dataset is a not real but it’s for the purpose of training the model and it can be pulled from [Kaggle](https://www.kaggle.com/datasets/mssmartypants/water-quality).

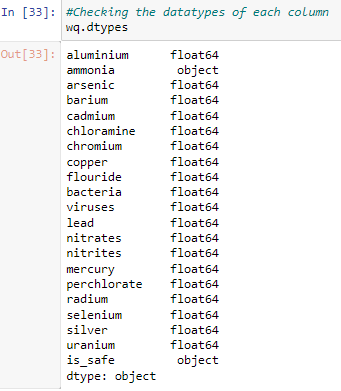
**About the Dataset**



The dataset has 7, 999 observations and 21 features/variables.



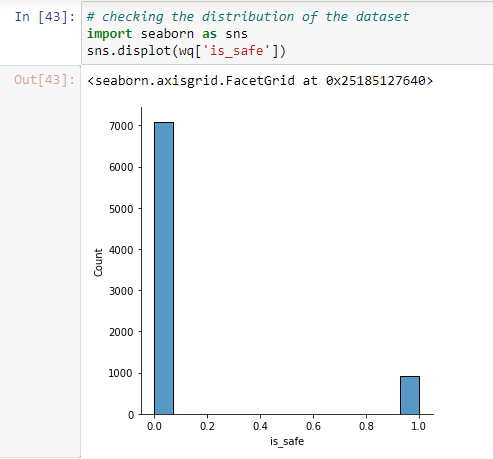
Out of the 21 features, 2 had incorrect data types because some observations were recorded as **#NUM!**



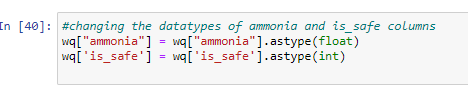
**Cleaning the Data**

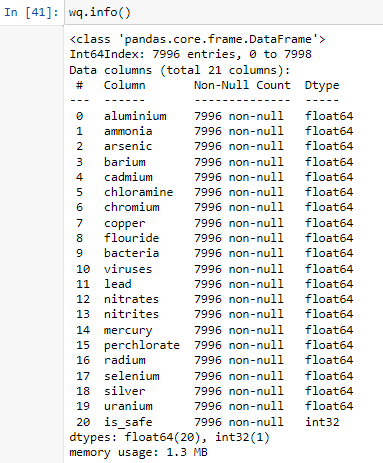
The rows with **#NUM!** values were deleted from the record. Deleting incomplete record was opted for to prune the dataset which is good for classification model, small dataset. After the deletion, the distribution of the dataset has 7,996 observations (3 rows were deleted) and 21 features/variables with 7,084 rows classified as unsafe as shown below

|  |  |
| --- | --- |
| 0 | 7,084 |
| 1 | 912 |
| **Total** | **7,996** |



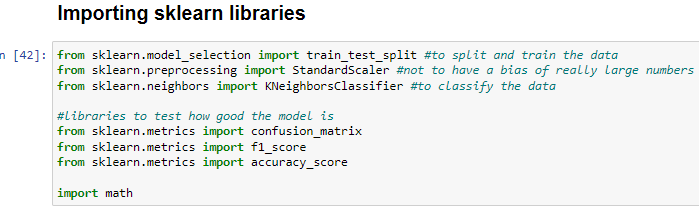
The 2 columns (ammonia and is\_safe) were changed to the appropriate data types



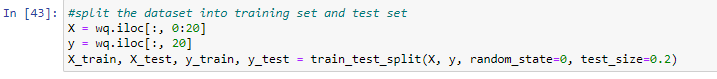


**Pre-processing the Data**

**Libraries:** the necessary libraries to pre-process the data and model it were imported

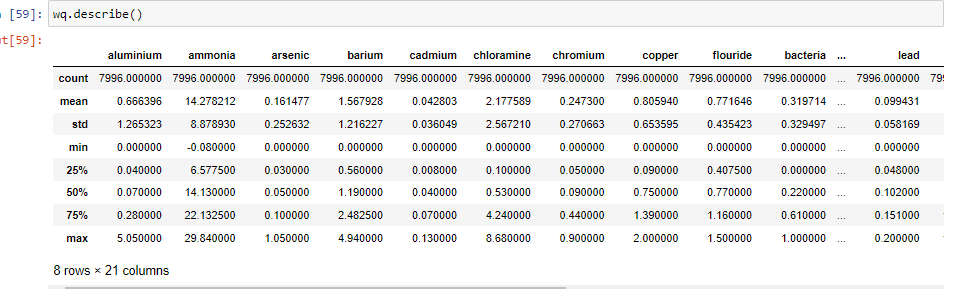


The dataset was splitted into 80% training set and 20% testing set with all the rows and first 20 features as independent variable and all the rows and the last column as the dependent variable.

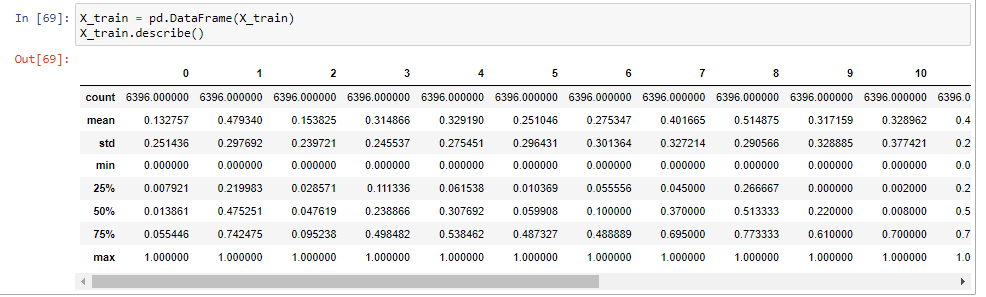


**Feature Scaling:** We will be performing normalisation on the dataset because the independent variables are not too wide apart (values ranges from -0 to 29). The feature scaling will make the model perform optimally or converge faster when the different observations (min and max) are on a smaller scale, say between 0 and 1, thereby leading to an accurate model.

The image below is the description of the dataset before and after normalisation was performed on it



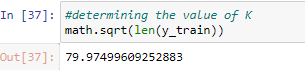
Before normalisation



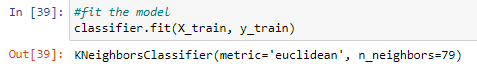
After normalisation

**Fitting the Model**

Since K in the KNN model is a hyper parameter, its value was calculated, the result came up to 79; this will be perfect to use because it’s not an even number. The model was initialised with k = 79, p = 2 (since the target variable is either safe or is\_not safe) using the Euclidean distance.

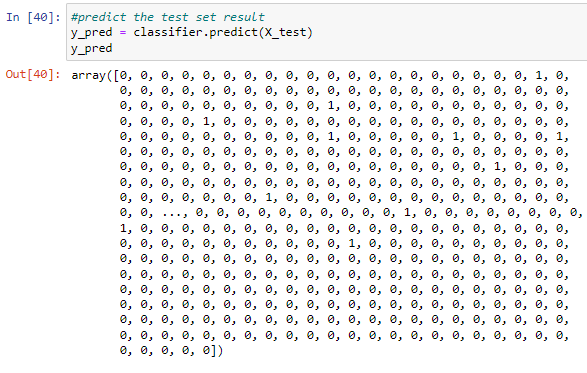






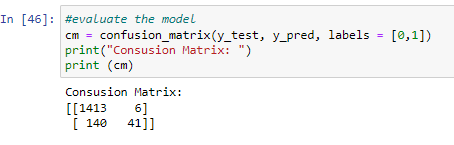
**Predicting with the Model**

Below is the result of the prediction



**Evaluating the Model**

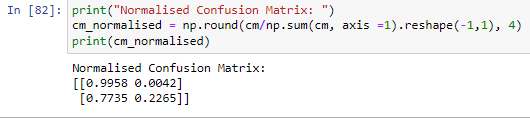
We implored the use of confusion matrix to evaluate the model and the output is displayed below with its heatmap

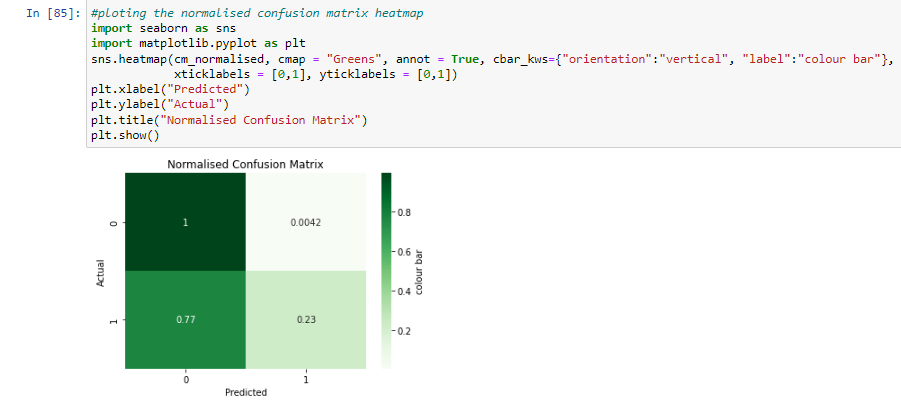




From the output, there were 41 cases whose actual label was one and were predicted as one; 1,413 cases whose actual label was zero and was predicted as zero. Six cases whose actual label was zero but predicted as one; 140 cases where the actual label were one but predicted as zero.

To clearly understanding the ratio in the confusion matrix, we normalised the matrix as shown below;





From the figure above, it shows that the true negative (TP) were 100% predicted correctly and the true positive (TP) were 23% predicted correctly.

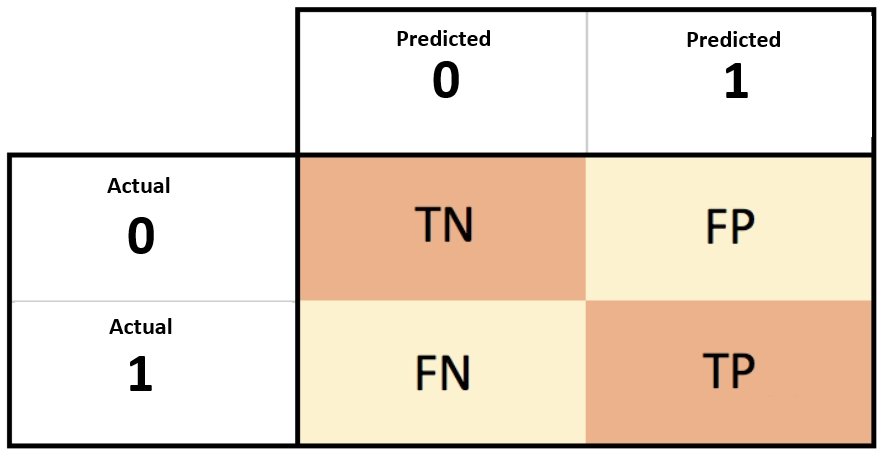
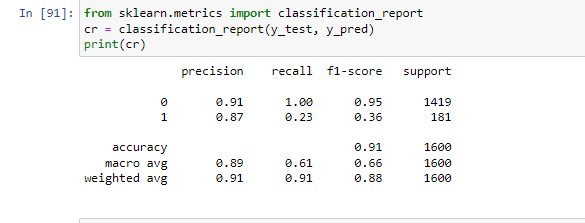


Image Source: Google

**Accuracy of the Model**

For a concise accuracy report, classification\_import module was used.

Below is the accuracy of the model when tested. The result is discussed as follows.



**Precision:** Only 87% of the total water sample that the model predicted as ‘is\_safe’ was actually safe.

**Recall:** Out of all the water sample that actually is\_safe, the model only predicted this outcome correctly for 23% of those samples.

**F1 Score:** Since the value isn’t very close to 1, it tells that the model does a poor job of predicting whether or not the water is safe

**Support:** The values under the support column relate how many water samples belonged to each class in the test dataset, among the water samples in the test dataset, 1,419 was not safe and 181 were safe.