Portfolio – Week 2

Studio 1, 2

**Dataset selected in Studio 1**: Water Quality in Water Engineering

All the datasets in Studio 1 are from the Engineering field, which is not my major (I'm studying Computer Science, majoring in Data Science). Therefore, I chose a dataset based on my personal interest: the water potability dataset. Specifically, I’ve always been concerned about the global issue of saltwater intrusion, including its impact in my home country. By working with this dataset, I aim to enhance my skills in Data and ML engineering while also gaining a deeper understanding of the factors that influence water potability.

# EDA (Exploratory Data Analysis) Summary

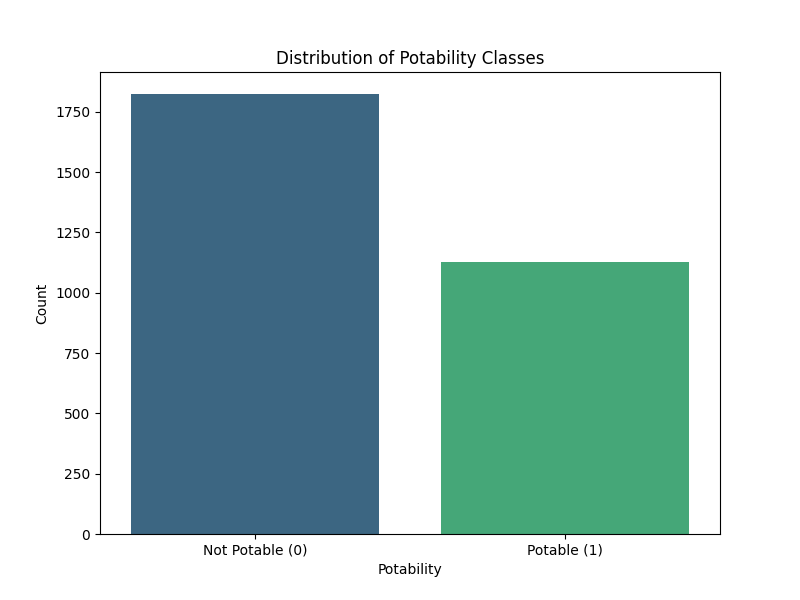
1. Except for the 'Conductivity', 'Sulfate', 'Organic\_carbon', and 'Trihalomethanes' features, all other features show very weak relationships with the 'Potability' target variable, making them less significant for correlation-based decision-making.
2. The 'Conductivity' feature has a low positive correlation with the 'Sulfate' feature, suggesting that we can create additional features like (Conductivity + Sulfate) to improve the prediction of water potability.
3. The 'Turbidity' feature shows a low positive correlation with the 'Organic\_carbon' feature, indicating the possibility of creating new features like (Turbidity + Organic\_carbon) to enhance predictions.
4. The 'Trihalomethanes' feature exhibits a low positive correlation with the 'Chloramines' feature, suggesting that combining these two features (Trihalomethanes + Chloramines) could improve model performance.

# Class labelling for target variable / developing ground truth data

For the "Water Portability" dataset, I believe the target variable, ‘Portability’, doesn't need to be relabeled for class classification, given its binary nature (0 and 1). The binary format already provides clear and distinct classes, which are sufficient for most classification models.

Additionally, I have implemented a bar chart to visualize the distribution of classes for the Potability variable, showing the counts for Non-portable (0) and Portable (1).





The bar chart clearly shows an imbalance in the distribution of the two potability classes:

1. **Not Potable:** This class has a larger number of samples, indicating that most of the samples are non-potable.
2. **Portable:** This class represents a smaller proportion of the samples, as it has fewer instances, indicating that a minority of the water samples are drinkable.

# Feature engineering

* 1. Convert the target values to the new categorial values (0,1, 2,…)

This conversion is unnecessary because, as mentioned, the existing binary labels—0 for non-potable and 1 for potable—are already appropriate for classification tasks.

* 1. Normalization

Normalizing input features, except for the Potability feature, helps ensure that all numerical features are on a similar scale, which can improve the performance and convergence of machine learning models. Then, I saved it as “normalised\_water\_potability.csv” file.



And its output:

A screenshot of a computer

Description automatically generated

1. Composite features using covariance

Based on the insights from the EDA, I have created four new input features by combining four pairs of existing features and then saved as “normalised\_water\_potability\_with\_composites.csv” file



And its output:

A black screen with white text

Description automatically generated

# Feature selection

In the EDA section, I noted that the features with the weakest relationships are ‘ph’, ‘Hardness’, ‘Sulfate’, and ‘Turbidity’. Therefore, I have created datasets with only the selected features, for both normalized and non-normalized versions.

Non-normalized dataset:



Normalized dataset:



And its output:

A screenshot of a computer screen

Description automatically generated

# Model development

So far, I have created four new datasets based on the original dataset (which includes all features without normalization or composite features – “water\_potability\_no\_outliers.csv”), as follows:

* All features with normalization and without composite features: “normalised\_water\_potability.csv”
* All features with normalization and including composite features: “normalised\_water\_potability\_with\_composites.csv”
* Selected features with normalization: “selected\_normalised\_features\_water\_potability.csv”
* Selected features without normalization: “selected\_features\_water\_potability.csv”

Then, I will develop a decision tree classifier using these five datasets:









And its output to compare 5 different accuracy value:

A screen shot of a computer code

Description automatically generated

A graph of blue squares

Description automatically generated

# Summarisation

I have made a table to compare 5 different models:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model 1:  water\_potability\_no\_outliers.csv | Model 2:  normalised\_water\_potability.csv | model 3:  normalised\_water\_potability\_with\_composites.csv | model 4:  selected\_normalised\_features\_water\_potability.csv | Model 5:  selected\_features\_water\_potability.csv |
| 57.79% | 57.79% | 55.42% | 50.56% | 50.56% |

**Observation:**

1. **Best Accuracy**: The datasets "normalised\_water\_potability.csv" and "water\_potability\_no\_outliers.csv" achieved the highest accuracy, both around 0.5779.
2. **Impact of Composite Features**: The dataset "normalised\_water\_potability\_with\_composites.csv," which includes composite features, yielded slightly lower accuracy at approximately 0.5542, suggesting that the addition of composite features did not significantly improve model performance.
3. **Feature Selection**: The feature selection process, which involved removing weakly correlated features, resulted in lower accuracy for both normalized and non-normalized selected feature datasets, with an accuracy of about 0.5056.

# Appendix

Link to the source code Portfolio – Week 2:

1. Studio1 <https://github.com/thinhpham1807/COS40007_Artificial_Intelligence_for_Engineering/tree/main/Studios/Studio01>
2. Studio 2 <https://github.com/thinhpham1807/COS40007_Artificial_Intelligence_for_Engineering/tree/main/Studios/Studio02>