

**SWINBURNE UNIVERSITY OF TECHNOLOGY**

COS40007 – Artificial Intelligence for Engineering

Portfolio Assessment-1: “Hello Machine Learning for Engineering”

Student Name: Nguyen Cuong Nhat

Student ID: 104178590

Instructor: Dr. Pham Thi Kim Dung

Submission Date: 22/9/2024

# I. Dataset

This dataset contains water quality measurements aimed at assessing the potability of water, indicating whether it is suitable for human consumption. It includes features such as pH level, hardness, dissolved solids, chloramine concentration, sulfate levels, and other key water quality parameters. The target variable, "Potability," is binary, with 1 indicating potable water and 0 indicating non-potable water. The primary goal is to use these attributes to predict water safety, making the dataset suitable for a supervised binary classification task. This dataset offers valuable insights for water quality assessment and decision-making in public health and environmental management.

As an AI major student, I chose this dataset because it presents a real-world application of machine learning in a critical area: water safety. Predicting water potability is a key challenge for public health, and applying AI models to assess water quality attributes can significantly aid in making informed decisions for water treatment and ensuring safe consumption. This dataset allows me to explore and develop supervised classification models, enhancing my understanding of AI techniques while contributing to an impactful environmental and societal issue.

# II. Exploratory data analysis (EDA) summary

## Data understanding

The dataset consists of 3,276 rows and 10 columns, each representing a water sample with various attributes related to water quality. The columns include:

* pH: The pH level of the water.
* Hardness: Water hardness, a measure of mineral content.
* Solids: Total dissolved solids in the water.
* Chloramines: Chloramines concentration in the water.
* Sulfate: Sulfate concentration in the water.
* Conductivity: Electrical conductivity of the water.
* Organic\_carbon: Organic carbon content in the water.
* Trihalomethanes: Trihalomethanes concentration in the water.
* Turbidity: Turbidity level, a measure of water clarity.
* Potability: Target variable; indicates water potability with values 1 (potable) and 0 (not potable).

These attributes are all continuous variables represented as `float64` data types. The target variable, "Potability," is a binary `int64` column that classifies water samples as either potable (1) or non-potable (0).

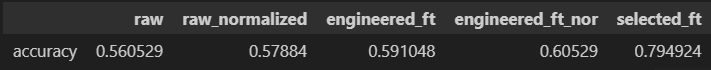
There are 9 features in the dataset with `float64` types, and 1 target variable with `int64`. This structure provides a solid foundation for a binary classification task, where machine learning models can be applied to predict whether a water sample is suitable for consumption based on the given features.

* Because our target variable “Portability” is already a categorical variable, we can skip the step for class labelling for target variable and developing ground truth data.
* **Correlations**: Most features show weak correlations with the target variable, "Potability." None of the features exhibit strong positive or negative relationships with "Potability." The heatmap confirms this, as the correlation values are close to zero. This indicates that each feature has limited linear impact on water potability classification.
* **Feature Distributions**:
* Most features, such as **ph**, **Hardness**, **Solids**, **Chloramines**, **Sulfate**, **Conductivity**, **Organic\_carbon**, **Trihalomethanes**, and **Turbidity**, have approximately normal distributions. This suggests that the data is relatively balanced without strong skewness or extreme outliers for these features.
* However, the **Potability** feature is imbalanced, with the majority of the water samples being classified as **non-potable** (label 0).
* **Outliers**: There are notable outliers in features such as **ph**, **Hardness**, **Sulfate**, **Trihalomethanes**, and **Solids** based on the boxplots. These outliers might require special treatment during modeling, such as capping or transformation.
* **Data Transformations**: Given the weak correlations and presence of outliers, it might be beneficial to explore non-linear transformations or combinations of features (such as feature engineering) to improve the prediction accuracy. Feature scaling may also be useful, given the differences in scales across the features like **Solids** and **Conductivity**.
* **Next Steps**:
* Handle outliers by either capping extreme values or using robust scaling techniques.
* Consider using feature engineering to create interaction terms that may provide better predictive power, especially between features with weak correlations.

# III. Feature engineering and Feature selection

# IV. Decision tree model development

# V. Experimental results



# VI. Conclusion

# VII. Appendix