**COS40007 Artificial Intelligence for Engineering**

**Name:** Tahmidul Haque Chowdhury  
**Student Number:** 103541308

**Studio Class:** Studio 1

# Dataset Selection

**Selected Dataset:** Water Quality

**Reason for Selection:** I chose the Water Quality dataset because it offers useful insights into the safety of drinking water by analysing different chemical and physical factors. This aligns with my interest in applying AI to support environmental monitoring and public health initiatives.

# Summary of Exploratory Data Analysis (EDA)

The Water Quality dataset has 3276 rows and 10 columns, including important features like pH, Hardness, Solids, Chloramines, Sulfate, and Conductivity. The target variable, Potability, tells us whether the water is safe to drink (1) or not (0). While exploring the data, I found that some values were missing—491 in the pH column, 781 in Sulfate, and 162 in Trihalomethanes. To avoid losing data, I filled in the missing values using the average for each column. I also noticed a few extreme values in solids and conductivity, which could affect the model’s accuracy, so I made sure to address them during feature engineering.

# Class Labelling for Target Variable

The target variable is already binary (Potable/Not Potable), so no further class labelling was needed.

**Feature Engineering and Selection**

1. **Normalization:**  
   Min-Max normalization was applied to ensure the data is within a consistent range.
2. **Composite Features:**  
   Four composite features were created:

ph\_solids = ph \* Solids

chloramines\_sulfates = Chloramines + Sulfate

hardness\_density = Hardness / Conductivity

conductivity\_organic = Conductivity \* Organic\_carbon

3. **Feature Selection:**  
 We selected a reduced feature set for stability: ph, Hardness, Solids, Chloramines, Potability.

# Feature Engineering and Selection

# I used Min-Max normalization to make sure all the features are on the same scale, so no feature dominates the others. I also created a few new features to find patterns between the variables. For example, I combined pH and Solids to make ph-solids, added Chloramines and Sulfate for chloramines-sulfates, divided Hardness by Conductivity to get hardness density, and multiplied Conductivity with Organic Carbon to form conductivity organic. After some exploration, I decided to keep things simple by focusing on the most important features: pH, Hardness, Solids, Chloramines, and Potability.

**Training and Decision Tree Model Development**

**Datasets Used**

**Original Dataset:** No normalization or composite features.

**Normalized Dataset:** Only normalization applied.

**Composite Features Dataset:** Contains new composite features.

**Selected Features Dataset (Normalized):** Selected key features with normalization.

**Selected Features Dataset (Without Normalization):** Selected features without normalization.

# Final Comparison Table

|  |  |
| --- | --- |
| **Model** | **Accuracy (%)** |
| Original Dataset | 55.21 |
| Normalized Dataset | 55.85 |
| Composite Features Dataset | 55.34 |
| Selected Features Dataset (Normalized) | 54.53 |
| Selected Features Dataset (Without Normalization) | 54.72 |

# Observations

The normalized dataset achieved the highest accuracy of 55.85%, showing that normalization helps improve model performance by ensuring all features have an equal impact. On the other hand, the selected features dataset with normalization reached an accuracy of 54.53%, indicating that refining the feature selection could further enhance the results.