**COS40007 Artificial Intelligence for Engineering**

**Portfolio Assessment 1: "Hello Machine Learning for Engineering"**

**Student Name:** Nguyen Quy Hung  
**Student Number:** 104850199 **Studio Class:** Studio 1&2  
**Submission Date:** September 14, 2025

**1. Dataset Selection**

**Selected Dataset:** Water Potability Dataset (water\_potability.csv)

I selected a dataset comprising 3,276 water samples with 9 physicochemical parameters and 1 target variable indicating water potability. The original features include: pH, Hardness, Solids, Chloramines, Sulfate, Conductivity, Organic\_carbon, Trihalomethanes, and Turbidity.

**2. Rationale for Dataset Selection**

I chose the Water Potability Dataset because it directly relates to environmental engineering applications, which aligns with my interests in data-driven solutions for real-world problems. As a third-year engineering student, I find water quality assessment particularly relevant since it represents a critical domain where machine learning can provide substantial value in automated monitoring systems. The binary classification problem of determining water safety based on quantitative chemical measurements reflects scenarios I might encounter in my future career, where data-driven approaches can complement traditional analytical methods in environmental monitoring and regulatory compliance.

**3. Exploratory Data Analysis Summary**

**3.1 Dataset Overview**

I began my analysis by examining the dataset structure and quality**:**

A screen shot of a computer code

AI-generated content may be incorrect.

**My Results:**

A screenshot of a computer

AI-generated content may be incorrect.

I discovered that three features had missing values that needed to be addressed during preprocessing.

**3.2 Target Variable Analysis**

I analyzed the target variable distribution to understand the classification problem:

A screen shot of a computer code

AI-generated content may be incorrect.

**My Results:**

A screen shot of a computer

AI-generated content may be incorrect.

I found the dataset has a moderate class imbalance with approximately 61% non-potable and 39% potable water samples.

**3.3 Feature Correlation Analysis**

I investigated the relationships between predictor variables and the target:

A screen shot of text

AI-generated content may be incorrect.

**My Results:**

A screenshot of a computer program

AI-generated content may be incorrect.

I observed that all features show weak linear correlations with the target variable, suggesting that water potability determination involves complex, potentially non-linear interactions between multiple physicochemical parameters rather than simple dependence on individual variables.

**4. Class Labelling for Target Variable**

**4.1 My Target Variable Classification Strategy**

I assessed the target variable to determine the optimal classification approach:

A computer code with colorful text

AI-generated content may be incorrect.

**My Results:**

A black screen with white text

AI-generated content may be incorrect.

My Decision: I found that the target variable is already categorical with a binary classification structure. Since the minority class represents 39% of observations (above the 0.30 threshold for acceptable balance), I decided to preserve the original binary classification scheme to maintain engineering relevance and regulatory compliance interpretation.

My Alternative Demonstration: To show my understanding of multi-class labelling, I created a 4-class pH-based categorization:

A screen shot of a computer program

AI-generated content may be incorrect.

**My Results:**



**5. Feature Engineering and Feature Selection**

**5.1 My Data Preprocessing and Normalization**

**5.1.1 Missing Value Treatment**

I handled missing values using median imputation:

A screen shot of a computer code

AI-generated content may be incorrect.

**My Results:**

A screenshot of a computer

AI-generated content may be incorrect.

**5.1.2 Feature Normalization**

I applied StandardScaler normalization to ensure fair feature contribution:

A screen shot of a computer code

AI-generated content may be incorrect.

**My Results:**



**5.2 My Integer Categorization of Features**

I systematically converted continuous variables to categorical representations:

A screen shot of a computer code

AI-generated content may be incorrect.

**My Results:**

A computer screen shot of a black screen

AI-generated content may be incorrect.

**5.3 My Composite Feature Development**

**5.3.1 Ratio-Based Features**

I created ratio-based features inspired by my understanding of water chemistry:

A screen shot of a computer code

AI-generated content may be incorrect.

**5.3.2 Interaction Features**

I developed interaction terms to capture synergistic effects:



**My Final Engineered Features:**

1. ph\_hardness\_ratio
2. solids\_conductivity\_ratio
3. chloramines\_sulfate\_ratio
4. organic\_carbon\_trihalomethanes\_product
5. ph\_conductivity\_interaction
6. turbidity\_cat\_num
7. hardness\_cat\_num
8. ph\_cat\_num
9. chloramines\_cat\_num

**6. Decision Tree Model Development and Training**

**6.1 Model Architecture and Hyperparameter Configuration**

A consistent decision tree architecture was implemented across all feature set evaluations to ensure fair comparative analysis:

A computer screen shot of a program code

AI-generated content may be incorrect.

**Hyperparameter Justification:**

* **max\_depth=10:** Balances model complexity with interpretability
* **min\_samples\_split=10:** Prevents excessive partitioning of small subsets
* **min\_samples\_leaf=5:** Ensures statistical significance of leaf nodes

**6.2 Feature Set Design and Experimental Framework**

Six distinct feature sets were systematically designed to evaluate different aspects of feature engineering effectiveness:

A black background with orange text

AI-generated content may be incorrect.

**Experimental Design:**

1. **Set1\_Original\_9Features:** Baseline performance using raw physicochemical parameters
2. **Set2\_Top5Correlated:** Feature selection based on correlation magnitude
3. **Set3\_Original\_Plus\_Ratios:** Integration of original and ratio-based features
4. **Set4\_Engineered\_Features:** Exclusive use of composite features
5. **Set5\_Best\_Mixed:** Curated combination of high-performing features
6. **Set6\_All\_Categorical:** Discretised categorical representations

**6.3 Model Training and Validation Protocol**

A black background with white text

AI-generated content may be incorrect.

A screen shot of a computer code

AI-generated content may be incorrect.

The stratified sampling approach ensures representative class distribution in both training and testing subsets, critical for unbiased performance estimation.

**7. Comparative Performance Analysis**

**7.1 Quantitative Results Summary**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Feature Set** | **Features (n)** | **Accuracy** | **Tree Depth** | **Leaves** | **Dominant Feature** | **Importance** |
| Set3\_Original\_Plus\_Ratios | 12 | 0.6419 | 10 | 114 | Sulfate | 0.1623 |
| Set1\_Original\_9Features | 9 | 0.6338 | 10 | 125 | Sulfate | 0.1987 |
| Set6\_All\_Categorical | 4 | 0.6073 | 9 | 57 | ph\_cat\_num | 0.4643 |
| Set5\_Best\_Mixed | 6 | 0.5972 | 10 | 94 | ph\_hardness\_ratio | 0.2986 |
| Set4\_Engineered\_Features | 5 | 0.5921 | 10 | 95 | ph\_hardness\_ratio | 0.3293 |
| Set2\_Top5Correlated | 5 | 0.5788 | 10 | 108 | ph\_hardness\_ratio | 0.3317 |

**8. Summary of My Observations**

Through my experimental analysis, I discovered several critical insights about feature engineering effectiveness in water potability classification:

My Key Findings:

1. Feature Integration Works Best: My Set3\_Original\_Plus\_Ratios achieved the highest performance (64.19% accuracy), which taught me that combining domain-specific engineered features with original measurements enhances predictive capability. I achieved a 0.81 percentage point improvement over baseline features.
2. Categorical Features Are Surprisingly Effective: My Set6\_All\_Categorical achieved competitive performance (60.73% accuracy) using only 4 discretized features. This showed me that decision trees can effectively exploit categorical boundaries while reducing computational complexity.
3. Correlation-Based Selection Has Limitations: My Set2\_Top5Correlated produced the lowest performance (57.88%), which taught me that simple correlation-based feature selection doesn't adequately capture the complex multivariate relationships essential for water potability determination.
4. Chemical Knowledge Matters: I consistently found Sulfate as the dominant decision factor across multiple feature sets, which aligns with what I learned about water quality standards in my coursework, validating that the model captures chemically relevant patterns.
5. Balance Between Complexity and Performance: My optimal model balanced predictive performance with reasonable interpretability (10 depth, 114 leaves), which I believe is crucial for engineering applications where I need to explain my decisions.

**9. Appendix: Source Code Repository**

My Source Code Access: I have made my complete Jupyter notebook implementation and associated data files available through the following shared repository:

<https://github.com/S1zzX/COS40007-Artificial-Intelligence-for-Engineering-/tree/main/Portfolio_Assessment/Assessment1>