Course Name: Artificial Intelligence for Engineering (COS40007)

Studio Session: Studio 1 - 7

Studio Tutor: Irfan Mirza

SWINBURNE UNIVERSITY OF TECHNOLOGY

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

Name: Ashraf Shahzad Toor

Student ID: 104586656

Submission Date: 24-03-2025

Contents

Dataset Selected 3

Dataset: 3

Reason for Dataset Choice: 3

Exploratory Data Analysis (EDA) Summary 3

Dataset shape: 3

Features include: 3

Target Variable: 3

Key Findings 3

Class Labelling for Target Variable 4

Feature Engineering and Feature Selection 5

Normalization: 5

New Features Created: 5

Feature Sets for Modeling: 5

Decision Tree Model Development 5

Model: 5

Train-Test Split: 5

Tooling: 5

Process: 5

Comparison Table 5

Summary of Observations 6

Appendix 6

Studio 1 Code Link: 6

Studio 2 Code Link: 6

Dataset Selected

Dataset: Water Potability Dataset

Reason for Dataset Choice: I chose the water potability dataset because ensuring safe drinking water is a fundamental engineering challenge, particularly in environmental and civil engineering fields. I wanted to explore patterns in water quality and build a model that can help classify potable and non-potable water based on chemical characteristics.

Exploratory Data Analysis (EDA) Summary

Dataset shape: 3276 rows × 10 columns

Features include:

- pH
- Hardness
- Solids
- Chloramines
- Sulfate
- Conductivity
- Organic_carbon
- Trihalomethanes
- Turbidity

Target Variable: Potability (binary: 0 = non-potable, 1 = potable)

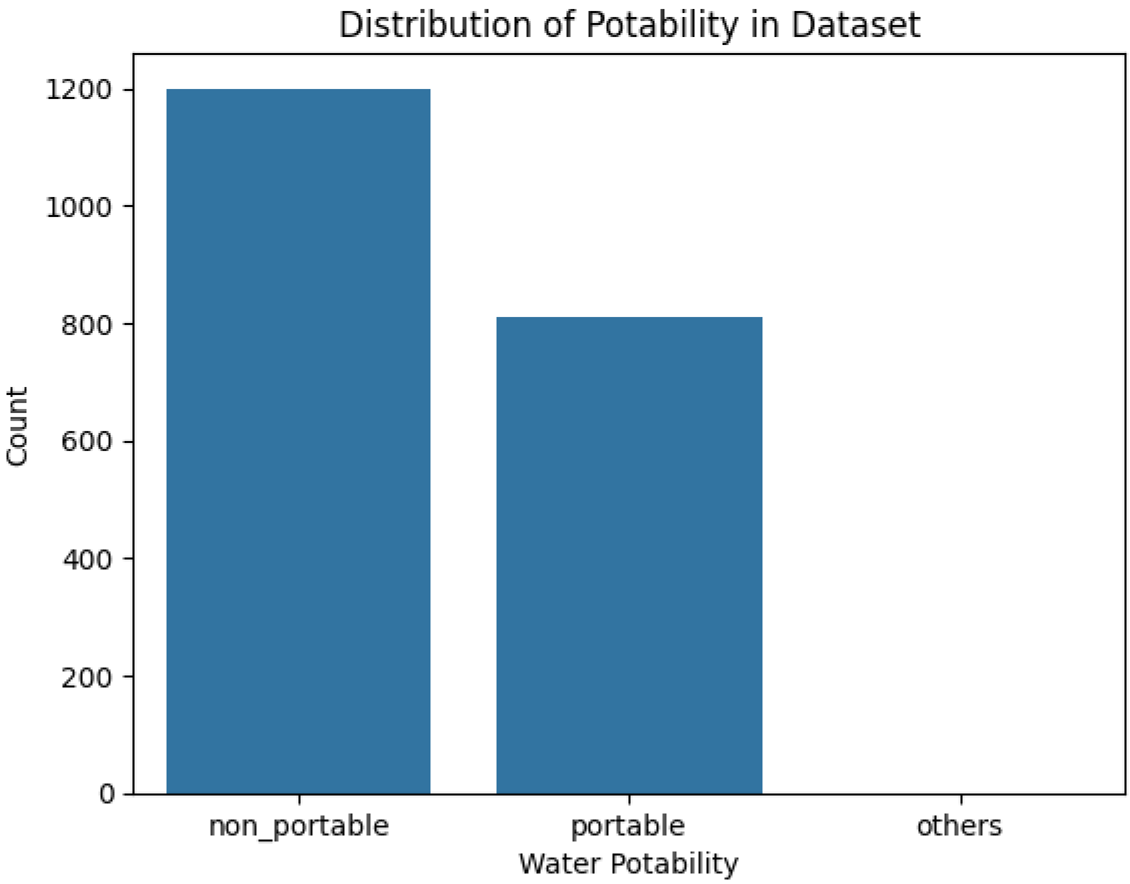
Key Findings

- The dataset had missing values in columns such as 'Sulfate' and 'Trihalomethanes'.
- The distribution of the target variable (Potability) was imbalanced:
- Non-potable: Majority class

- Potable: Minority class
- Several features showed right-skewed distributions, such as Hardness and Solids.
- Features such as Chloramines and Sulfate exhibited slightly left-skewed distributions.
- Correlation analysis revealed weak relationships between individual features and Potability. However, some weak correlations existed among independent variables:
    - Organic_carbon and pH had a weak positive correlation.
    - Trihalomethanes and Solids were weakly correlated.
    - Chloramines and Hardness shared a minor positive relationship.

Class Labelling for Target Variable

- The target variable was already categorical (binary) so no additional class labelling was needed. The class distribution looked like this:



Feature Engineering and Feature Selection

Normalization:

- To ensure all features are on the same scale, Min-Max normalization was applied to all numerical columns (excluding the target variable). This transformation scaled values to the [0,1] range, which helps improve model performance and convergence.

New Features Created:

- organic_carbon_ph: Covariance between Organic_carbon and pH.
- chloramines_hardness: Covariance between Chloramines and Hardness.
- trihalomethanes_solids: Covariance between Trihalomethanes and Solids.

Feature Sets for Modeling:

- Set 1: All features without normalisation and without composite features.
- Set 2: All features with normalisation and without composite features.
- Set 3: All features with normalisation and containing composite features.
- Set 4: Selected features with normalisation.
- Set 5: Selected feature without normalisation.

Decision Tree Model Development

Model: Decision Tree Classifier (Gini Index)

Train-Test Split: 70%-30%

Tooling: Scikit-learn

Process: Each of the 5 feature sets was used to train and test a separate decision tree.

Comparison Table

Feature Set	Accuracy (%)
Set 1	61.75
Set 2	61.92
Set 3	62.09
Set 4	58.28
Set 5	58.44

Summary of Observations

The model (Set 3) using all original features combined with composite features achieved the highest accuracy of 62.09%. This suggests that composite features (such as covariances between related variables) helped improve the model's predictive power. Moreover, models using only a subset of features (Set 4 and Set 5) underperformed compared to models that retained all original features. This indicates that reducing features may have led to loss of valuable information.

Appendix

Studio 1 Code Link:

Studio 2 Code Link: