**Portfolio Assessment 1 Report**

**A. Personal Information**

* **Name**: Dang Khoa Le
* **Student ID**: 103844421
* **Studio Class**: Studio 1-7 (Tuesday 6.30PM – 8.30PM)
* **Dataset Chosen**: Water Quality in Water Engineering

**Explanation of Dataset Choice**

As a Software Engineer, I am interested in datasets that involve real-world applications where technology and data science intersect with practical problem-solving. The Water Quality in Water Engineering dataset offers an opportunity to explore how data science can be applied to environmental science, specifically in assessing the potability of water. This dataset aligns with my interest in using data-driven approaches to address important issues like water quality, which has significant implications for public health and safety.

**B. Dataset Context**

The dataset contains measurements related to the quality of water and aims to assess whether the water is suitable for human consumption. Each entry in the dataset represents a water sample with various attributes, and the "Potability" column indicates if the water is potable (1) or not (0).

**Columns:**

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

**C. Exploratory Data Analysis (EDA)**

**1. Variable Identification**

* **Target Variable**: Potability
* **Predictors (Input Variables)**: pH, Hardness, Solids, Chloramines, Sulfate, Conductivity, Organic\_carbon, Trihalomethanes, Turbidity

**2. Univariate Analysis**

To understand the distribution and characteristics of each variable, we performed univariate analysis on the numerical columns. Below are the observations based on the analysis:

**Numerical Columns Analysis**

**A computer screen with text on it

Description automatically generated**We created visualizations for the distribution of the following columns: pH, Hardness, Solids, Chloramines, Sulfate, Conductivity, Organic\_carbon, Trihalomethanes, Turbidity.

Fig. Univariate Analysis in Python

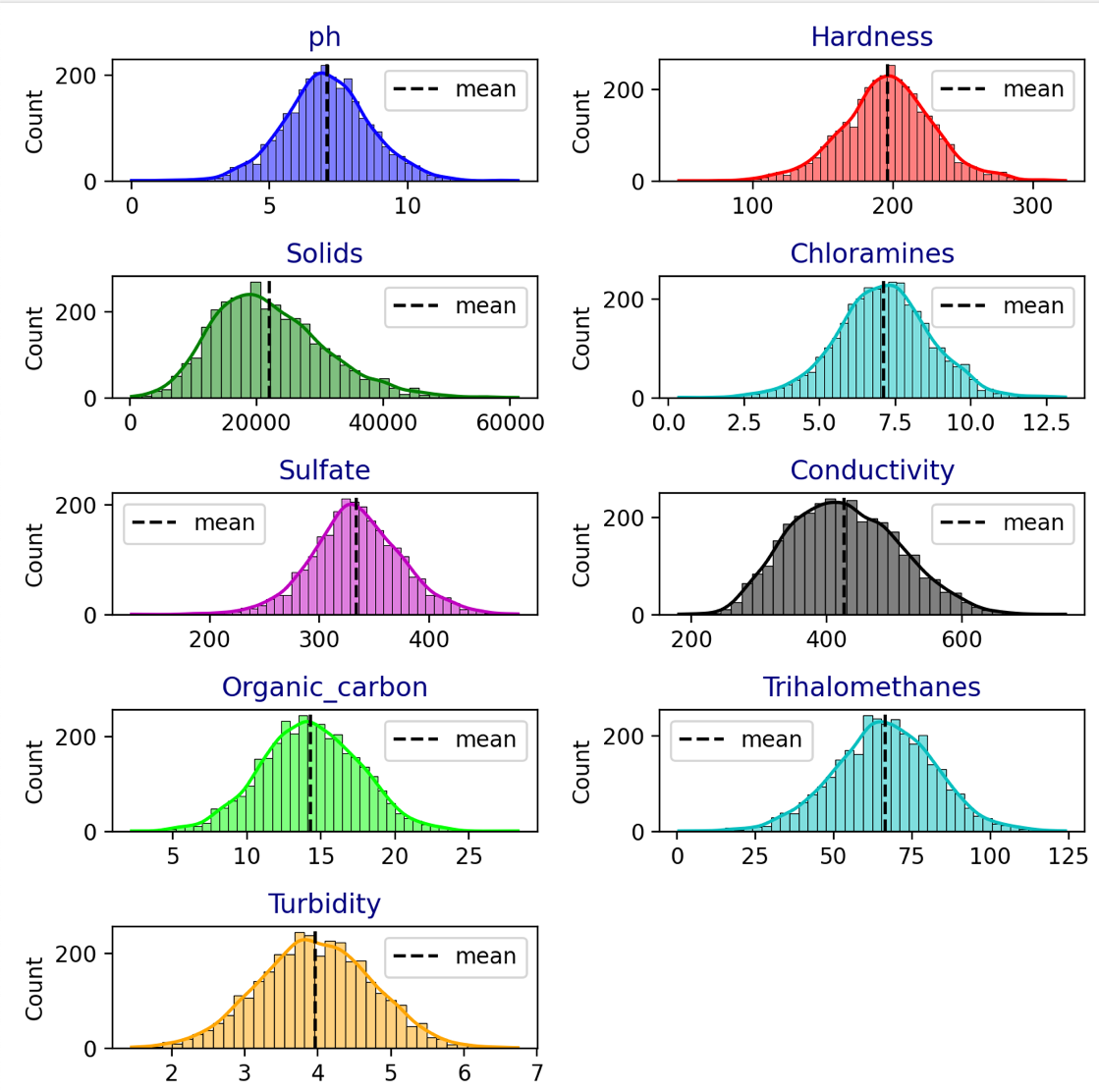
**Observations:**

Fig. Univariate Analysis illustrations

* **pH**: The pH level has a distribution with a peak around 7, indicating it is relatively balanced but with a right skew, indicating some higher values.
* **Hardness**: The distribution shows a peak around 196 with a right skew, suggesting higher hardness values are less common but present.
* **Solids**: This column exhibits a right skew with a broad distribution, indicating a large range of total dissolved solids.
* **Chloramines**: The distribution is slightly right-skewed with a peak around 7, showing most values cluster around this point.
* **Sulfate**: This column has a right-skewed distribution with a peak around 334, indicating a broad range of sulfate concentrations.
* **Conductivity**: The distribution shows a right skew with most values clustered around 426, indicating a broad spread.
* **Organic\_carbon**: The distribution is slightly right-skewed with most values clustering around 14.
* **Trihalomethanes**: This variable has a right-skewed distribution, with a peak around 66.
* **Turbidity**: The distribution is moderately right-skewed with most values around 4, suggesting a higher concentration of lower turbidity values.

**Target Variable Analysis**

The target variable, **Potability**, shows a distribution with a mean value of 0.39, indicating an imbalance with more samples classified as not potable (0) compared to potable (1).

**3. Summary Statistics**

The summary statistics provide a comprehensive overview of the distributions and central tendencies of each variable in the dataset.

Here's the breakdown:

a) The pH variable has a mean close to the neutral value (7). However, the wide range from 0 to 14 suggests significant variability. The standard deviation is relatively high, indicating some degree of dispersion from the mean.

b) Hardness has a relatively high mean and a considerable range, suggesting variability in water hardness. The distribution appears moderately skewed, with the majority of values clustering around the mean.

c) Solids have a very high mean and a broad range. The large standard deviation indicates a wide spread of values around the mean, suggesting significant variability in the amount of solids.

d) Chloramines have a mean value that is somewhat higher than the median, indicating a slight right skew. The range and standard deviation are moderate.

e) Sulfate values are spread out, with a relatively high mean and substantial range. The standard deviation suggests moderate variability.

f) Conductivity has a high mean and a broad range. The high standard deviation indicates significant variability, with values spread over a wide range.

g) Organic Carbon has a mean with a moderate range. The standard deviation is moderate, indicating a spread of values around the mean.

h) Trihalomethanes have a wide range with considerable variability. The mean is slightly higher than the median, suggesting a possible right skew.

i) Turbidity has a low mean and a relatively narrow range, with a standard deviation indicating low variability. The distribution is likely close to normal.

j) Potability is a binary variable with a mean suggesting that about 39% of the samples are potable.

**Summary:**

* **pH** and **Chloramines** show moderate variability around their means, with potential skew.
* **Hardness** and **Sulfate** have higher means and broader ranges, indicating significant variability.
* **Solids** exhibit very high variability, suggesting diverse solid content.
* **Conductivity** and **Trihalomethanes** show substantial spread in values, with possible right skew.
* **Organic Carbon** and **Turbidity** are less variable with relatively narrow ranges.
* **Potability** is a binary variable with a clear bimodal distribution.

**4. Multivariate Analysis**

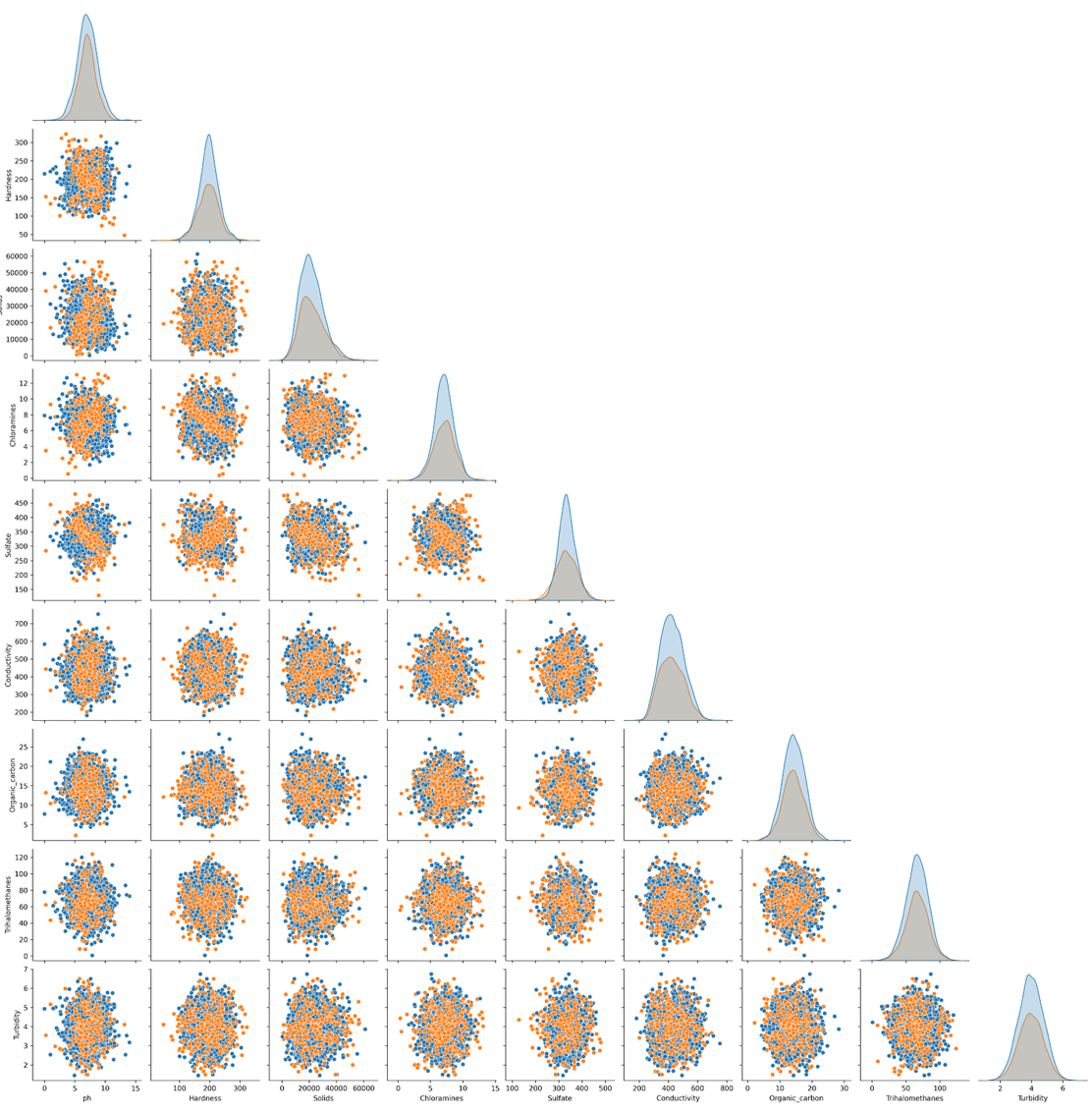
**Pairplot Analysis**: The pairplot shows relationships between different features.

Fig. Pairplot Multivariate Analysis illustrations

**Diagonal Analysis**:

The KDE diagonal plots show mostly unimodal distributions for most variables, with a few exceptions.

**pH** and **Turbidity** show slightly right-skewed distributions.

**Conductivity** shows a more pronounced right-skewed distribution.

**Organic\_carbon** appears to have a bimodal distribution, suggesting two distinct groups in the data for this parameter.

**Off-Diagonal Analysis**: Relationship between indpendent attributes

* pH vs other attributes: pH shows weak to moderate negative correlations with most other attributes, particularly noticeable with Conductivity and Sulfate.
* Hardness vs other attributes: Hardness shows strong positive correlations with Solids, Chloramines, Sulfate, and Conductivity. This suggests these parameters are closely related in water quality.
* Solids vs other attributes: Similar to Hardness, Solids show strong positive correlations with Chloramines, Sulfate, and Conductivity.
* Chloramines vs other attributes: Besides its correlation with Hardness and Solids, Chloramines also show a moderate positive correlation with Sulfate and Conductivity.
* Sulfate vs other attributes: Sulfate shows strong positive correlations with Hardness, Solids, Chloramines, and Conductivity.
* Conductivity vs other attributes: Conductivity is strongly positively correlated with Hardness, Solids, Chloramines, and Sulfate, indicating it's a good overall indicator of dissolved substances in the water.
* Organic\_carbon vs other attributes: Organic\_carbon shows weak correlations with most other attributes, suggesting it might be an independent factor in water quality.
* Trihalomethanes vs other attributes: Trihalomethanes show moderate positive correlations with several attributes, particularly Chloramines and Conductivity.

**Potability attribute: Relationship between dependent and independent attributes**

* Potability vs pH: There's no clear relationship, with potable and non-potable samples spread across the pH range.
* Potability vs Hardness: Slight tendency for potable water to have lower hardness, but the relationship is weak.
* Potability vs Solids: No clear relationship is visible.
* Potability vs Chloramines: Potable water samples tend to have slightly lower Chloramines levels, but there's significant overlap.
* Potability vs Sulfate: No strong relationship is apparent.
* Potability vs Conductivity: Potable water samples tend to have slightly lower Conductivity, but there's significant overlap.
* Potability vs Organic\_carbon: No clear relationship is visible.
* Potability vs Trihalomethanes: Potable water samples tend to have slightly lower Trihalomethanes levels, but the relationship is weak.
* Potability vs Turbidity: No strong relationship is apparent.

**5. Study Correlation**

A diagram of a diagram

Description automatically generated with medium confidenceA screen shot of a computer program

Description automatically generated**:**

Fig. Correlation Analysis table

Fig. Pairplot Correlation Analysis illustrations

Fig. Correlation analysis data execution program

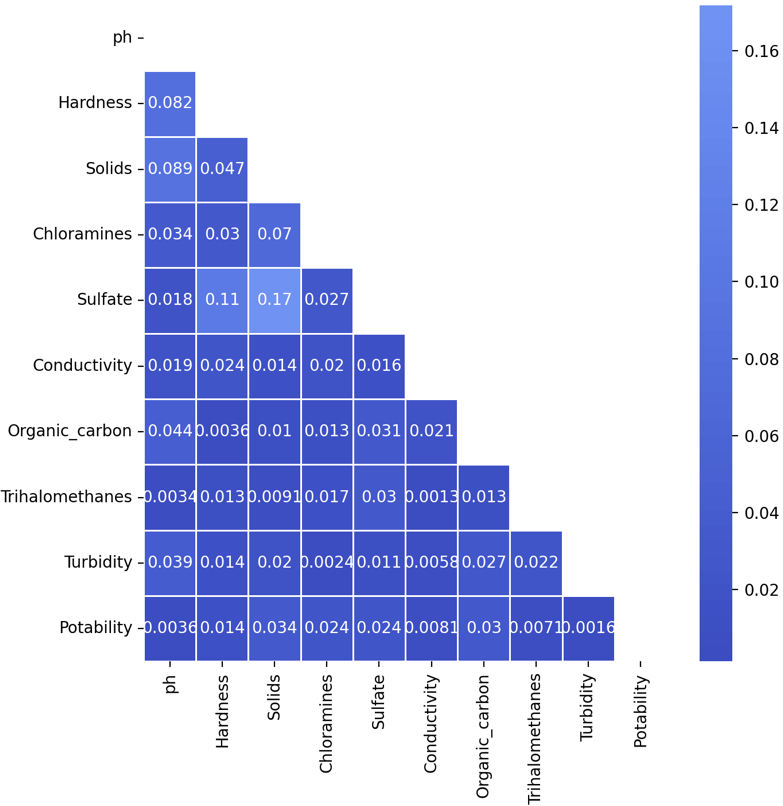
**Observations:**

Fig. Correlation Analysis Heatmap

* **pH**:
  + Shows negligible correlation with most features.
  + Very low positive correlation with **Hardness**.
  + Very low negative correlation with **Solids** and **Turbidity**.
* **Hardness**:
  + Very low negative correlation with **Solids**.
  + Very low negative correlation with **Sulfate**.
* **Solids**:
  + Very low negative correlation with **Sulfate** and **Chloramines**.
* **Chloramines**:
  + Shows negligible correlation with most features.
  + Very low positive correlation with **Sulfate**.
* **Sulfate**:
  + Very low negative correlation with **Solids**.
  + Very low positive correlation with **Organic Carbon**.
* **Conductivity**:
  + Shows negligible correlation with most features.
  + Very low positive correlation with **Organic Carbon**.
* **Organic Carbon**:
  + Shows negligible correlation with most features.
  + Very low positive correlation with **Sulfate**.
* **Trihalomethanes**:
  + Shows negligible correlation with most features.
* **Turbidity**:
  + Shows negligible correlation with most features.
  + Very low negative correlation with **Organic Carbon**.
* **Potability**:
  + Shows negligible correlation with most features.
  + Very low positive correlation with **Solids**.

By analysing data From these correlation table and figures, we observe that there are no strong correlations between the features. Most features exhibit negligible to very low correlations with each other. This implies that the features in this dataset are largely independent of one another, and there are no redundant features based on their pairwise correlations.

### **6. EDA (Exploratory Data Analysis) Summary**

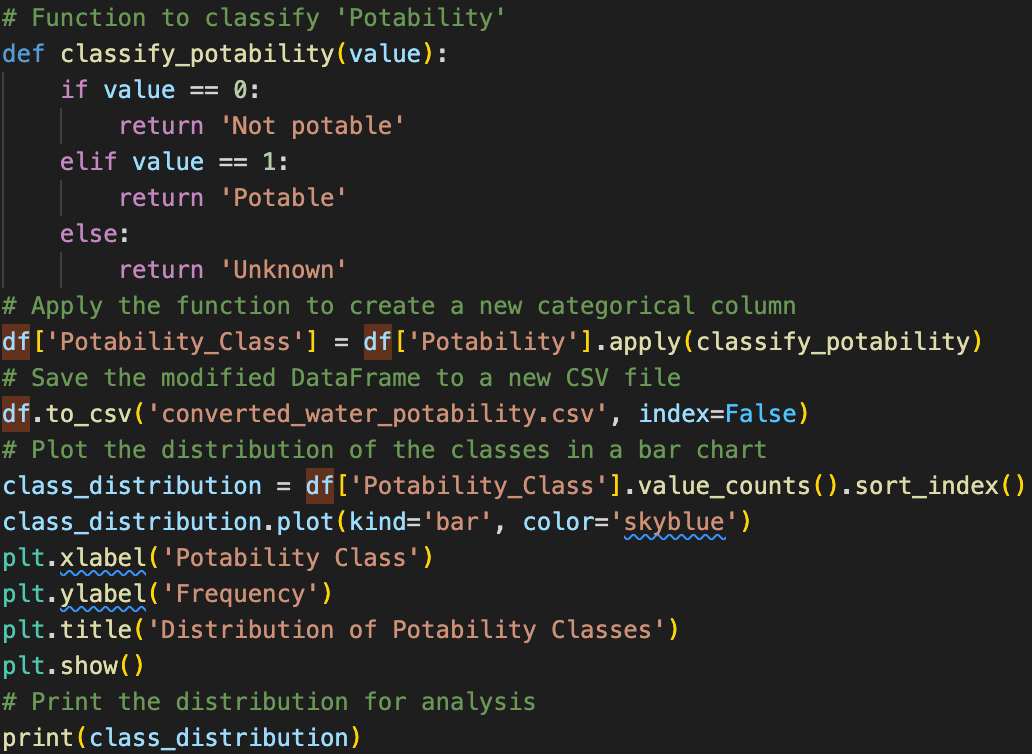
1. Except for 'Hardness', 'Sulfate', 'Conductivity', and 'Organic\_carbon' features, all other features have very weak relationships with the 'Potability' feature and do not account for making statistical decisions (of correlation).
2. 'Hardness' has a very low positive correlation with 'Solids' and 'Chloramines' features, perhaps we can create additional features like (hardness + solids) and (hardness + chloramines) to predict the potability.
3. 'Sulfate' has a very low positive correlation with 'Organic\_carbon' feature, perhaps we can create additional features like (sulfate + organic\_carbon) to predict the potability.
4. 'Conductivity' has very low positive correlations with 'Organic\_carbon' and 'Trihalomethanes' features, perhaps we can create additional features like (conductivity + organic\_carbon) and (conductivity + trihalomethanes) to predict the potability.
5. Range of clusters in this dataset is 2 to 6.

**D. Class labelling / Creating ground truth data**

#### **1. Task Description**

Using the water potability dataset to create categorical labels based on the ‘Potability’ value of the water. The goal is to classify the water into 2 potability categories based on potability levels as follows:

1. Not potable: Potability = 0
2. Potable: Potability = 1

**2. Python program**

**3. Analysis**

**Distribution:**

* + 1. Class ‘Not potable’: 1998 samples
    2. Class ‘Potable’: 1278 samples

**A graph of a number of potability classes

Description automatically generatedObservation:**

Fig. Distribution of Potability classes

There is a clear imbalance in the distribution of pH classes in this dataset. The distribution is not equal across the five classes, and there are significant differences (more than 200) between some class samples. Here's a detailed breakdown:

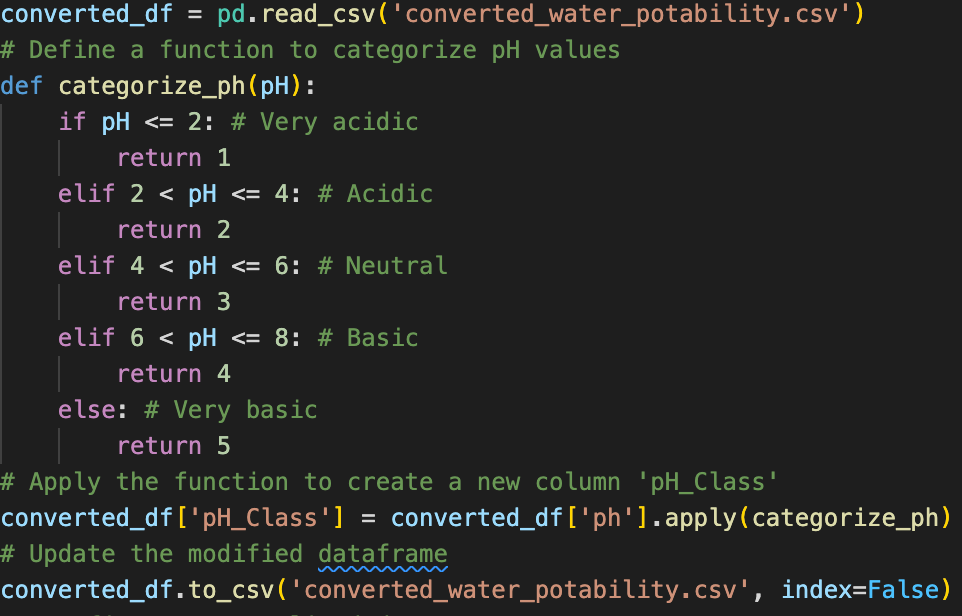
* Class 1 (Not potable) is the most frequent, with 1998 samples. This represents a majority of the data, indicating that non-potable samples are more common in this dataset.
* Class 2 (Potable) has 1278 samples, which is notably fewer compared to Class 1.

**Imbalance assessment:**

* Imbalance: There is a significant imbalance between the two classes, with the difference being 720 samples. This indicates a substantial imbalance in the dataset, with the 'Not potable' class being significantly more prevalent than the 'Potable' class.

In summary, the dataset exhibits a noticeable imbalance between the two classes, with the 'Not potable' class having a higher representation compared to the 'Potable' class. This imbalance could affect the performance of classification models and should be considered during model training and evaluation.

**E. Feature Engineering and Feature Selection**

**1. Feature Engineering**

**a. Simplify the pH feature**:

First, we will convert the pH feature into categorical classes based on the provided ranges:

* **Class 1 (pH ≤ 2)**
* **Class 2 (2 < pH ≤ 4)**
* **Class 3 (4 < pH ≤ 6)**
* **Class 4 (6 < pH ≤ 8)**
* **Class 5 (pH ≥ 8)**

Fig. Python program classify pH classes

**Then, update these pH classes to the ‘converted\_water\_potability.csv’ file made previously.**

**b. Normalize selected features**:

Normalize 'Hardness', 'Sulfate', 'Conductivity', 'Organic\_carbon', 'Chloramines', 'Solids', and 'Trihalomethanes' using MinMaxScaler. Then save these normalised features to ‘normalised\_water\_potability.csv’.

**c. Create new composite features**:

Create composite features based on the given relationships from the EDA summary:

* + - (Hardness + Solids)
    - (Hardness + Chloramines)
    - (Sulfate + Organic\_carbon)
    - (Conductivity + Organic\_carbon)
    - (Conductivity + Trihalomethanes)

**d. Save the new features and class labels**:

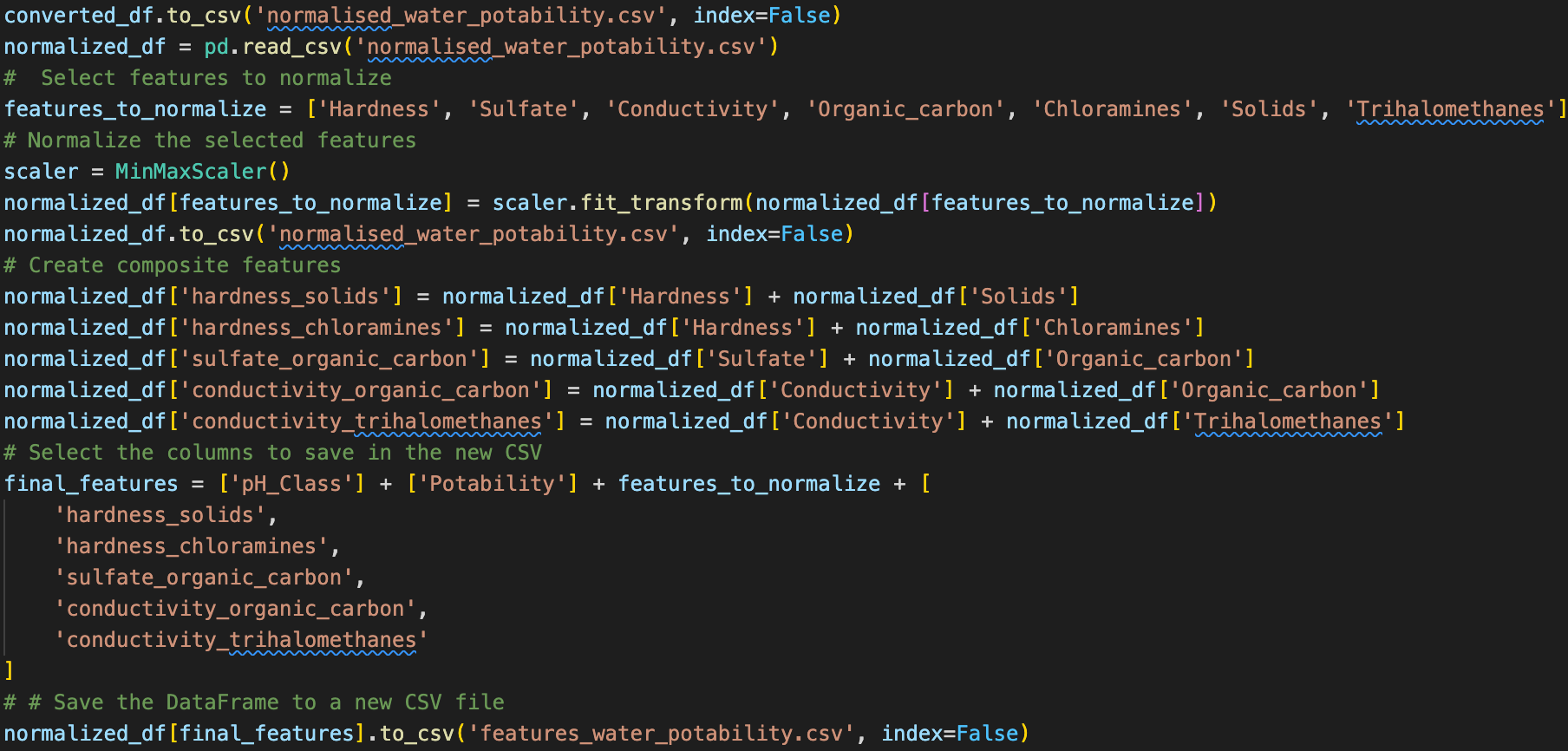
Save these 12 features (1 pH class + ‘Potability’ + 7 normalized features + 4 composite features) along with the class label to ‘features\_water\_potability.csv’.

Fig. Python program for b, c and d sections

**2. Feature Selection:**

Based on the EDA observations, we will select specific features for further analysis.

* Keep the important features and composite features identified earlier.
* Drop other features that have a weak relationship with the target variable.
* A screen shot of a computer

  Description automatically generatedSave the new dataset with the selected features.

**a. Identify important features**:

Keep the 'pH\_Class', 'Hardness', 'Sulfate', 'Conductivity', and 'Organic\_carbon' features along with the composite features created in the previous task.

**b. Drop other features**:

Remove features that are not strongly correlated with the target variable.

**c. Save the selected features**:

Save the new dataset with the selected features to a new CSV file named ‘selected\_features\_water\_potability.csv’.

Fig. Excel table converted from selected\_features\_water\_potability.csv

**F. Training and decision tree Model development**

A screen shot of a computer program

Description automatically generated**1. Create Selected Converted file**

Fig. Python program to create selected converted file

We firstly create the ‘selected\_converted\_water\_potability.csv’ file, which is defined to be the selected feature without normalisation, here select only 'Hardness', 'Sulfate', 'Conductivity', 'Organic\_carbon', 'Chloramines', 'Solids', 'Trihalomethanes', 'pH\_Class' and ‘Potability’ as features.

**2. Develop decision tree with training model**

* **Loading Data**: The function load\_and\_train\_model loads the dataset and splits it into features and target variable.
* **Splitting Data**: It uses a 70-30% split for training and testing.
* **Training the Model**: It trains a DecisionTreeClassifier on the training set.
* **Evaluating Accuracy**: It computes the accuracy on the test set and prints it.

A screen shot of a computer screen

Description automatically generated**G. Model Comparison table**

Fig. Python program to create selected converted file

The model comparison was performed across five datasets derived from the water potability dataset. Each dataset underwent different preprocessing techniques to observe their effects on the model's accuracy. Below are the results and insights based on the comparison:

**Dataset to be compared:**

* converted\_water\_potability.csv
* normalised\_water\_potability.csv
* features\_water\_potability.csv
* selected\_features\_water\_potability.csv
* selected\_converted\_water\_potability.csv

**Demonstrated on Excel table:**

**Converted vs. Normalized Data:**

* The model trained on the normalized dataset (normalised\_water\_potability.csv) achieved slightly better accuracy (0.5748) compared to the unnormalized converted data (converted\_water\_potability.csv) with an accuracy of 0.5738. This suggests that normalization may have provided a minor improvement in model performance, potentially due to scaling effects.

**Feature Engineering Impact:**

* Introducing new composite features and normalizing certain variables in the features\_water\_potability.csv dataset did not significantly improve model accuracy, with a slightly lower performance (0.5707) compared to the normalized dataset. This indicates that the added complexity might not have provided substantial benefits in predicting potability.

**Feature Selection Importance:**

* The selected\_features\_water\_potability.csv dataset, which involved selecting important features based on earlier analyses, resulted in the highest accuracy (0.5819). This demonstrates the importance of carefully selecting features that are most relevant to the target variable, leading to improved model performance.

**Selected Converted Data:**

* The model trained on the selected\_converted\_water\_potability.csv dataset, which retained selected features without normalization, performed the poorest with an accuracy of 0.5585. This suggests that for this dataset, normalization or retaining the natural feature distribution might have been more beneficial.

Overall, the results indicate that feature selection has a significant impact on model performance, with the selected\_features\_water\_potability.csv dataset yielding the highest accuracy. Normalization provided minor improvements, while feature engineering through composite features did not lead to substantial accuracy gains. The findings emphasize the importance of both feature selection and data preprocessing in developing effective predictive and data training models.

**H. Summary**

This report provided a comprehensive analysis of the water potability dataset, focusing on exploratory data analysis (EDA), feature engineering, class labelling, and model development using decision trees. The primary objective was to assess the water's suitability for human consumption, with the potability of water as the target variable.

1. **Dataset Exploration:**
   * The dataset contains various water quality measurements, and initial exploratory data analysis revealed important insights about the distribution and relationships among features. Most features exhibited weak correlations, suggesting limited redundancy.
2. **Class Imbalance:**
   * The dataset exhibited a noticeable class imbalance, with more non-potable water samples than potable ones. This imbalance could potentially impact model performance, highlighting the need for techniques such as resampling or class-weight adjustments.
3. **Feature Engineering:**
   * Several new features were created to capture potential interactions between variables. These included composite features such as Hardness + Solids and Sulfate + Organic\_carbon. Additionally, features were normalized to ensure they were on the same scale.
4. **Model Development:**
   * A decision tree classifier was trained on various versions of the dataset to evaluate the impact of different preprocessing steps. The selected features dataset, which involved careful feature selection, yielded the best accuracy, underscoring the importance of selecting relevant features for model training.
5. **Model Comparison:**
   * The model comparison demonstrated that while normalization and feature selection positively impacted model accuracy, the addition of composite features did not lead to significant improvements. The highest accuracy was achieved with the selected features dataset.

**Conclusion:** This analysis underscores the importance of thorough data preprocessing, feature selection, and addressing class imbalance when developing predictive models. The findings from this report provide valuable insights into optimizing the decision tree model for water potability classification, with potential applications in ensuring safe water quality.

**I. Appendix**

All dataset (CSV files) and Python program as “a1.py” is provided via Google Drive with the access link below:

<https://drive.google.com/drive/folders/1gckgh-dtlzQI2f3rvXXGWB-JRKSOQHj4?usp=sharing>

Google Drive link is accessible publicly.

The Python program is categorised in section (A to F sections) corresponding to the report. Please make sure while running the Python program, you must comment-in sections which you do not plan to run, and only comment-out (leaves uncomment) sections you would wish to run.