Phase II - Water Quality Analysis Report

Introduction

Water quality is a crucial factor for human health and well-being, as well as for the environment and the economy. Water quality analysis is the process of measuring and evaluating the physical, chemical, and biological characteristics of water to determine its suitability for various purposes. The main objective of this report is to analyze the water quality dataset provided by the client and to extract insights and recommendations for improving water quality.

Data Preprocessing

The dataset consists of numerous observations and 10 variables. The variables are:

- pH: The pH level of the water
- Hardness: The capacity of the water to precipitate soap
- Solids: The total dissolved solids in the water
- Chloramines: The amount of chloramines in the water
- Sulfate: The amount of sulfate in the water
- Conductivity: The electrical conductivity of the water
- Organic_carbon: The amount of organic carbon in the water
- Trihalomethanes: The amount of trihalomethanes in the water
- Turbidity: The measure of light scattering by particles in the water
- Potability: The binary indicator of whether the water is safe for human consumption or not

The data preprocessing steps include:

Data Cleaning

We checked for missing values and outliers in the dataset. We found that there are 491 missing values in the dataset, distributed across three variables: pH (201), Sulfate (781), and Trihalomethanes (162). We decided to impute the missing values using the median values of each variable, as they are less sensitive to outliers than the mean values. We also detected some outliers in the dataset using boxplots and z-scores. We decided to keep the outliers as they might represent natural variations in water quality.

Data Transformation

We did not need to perform any data transformation, as there are no categorical variables in the dataset. All the variables are numerical and continuous.

Exploratory Data Analysis (EDA)

We performed exploratory data analysis to better understand the distribution and relationships between variables. We used visualization and summary statistics techniques for this purpose.

Visualization

We created histograms, scatter plots, and correlation matrices to visualize the data. The histograms show the frequency distribution of each variable. The scatter plots show the pairwise relationship between each variable. The correlation matrices show the strength and direction of the linear relationship between each variable.

The histograms reveal that most of the variables are approximately normally distributed, except for Solids and Conductivity, which are skewed to the right. The scatter plots show that there are some positive and negative correlations between some variables, such as pH and Hardness, Chloramines and Trihalomethanes, Sulfate and Conductivity, etc.

Summary Statistics

We calculated summary statistics for each variable to get a sense of their central tendencies and spread. We used measures such as mean, median, standard deviation, minimum, maximum, and quartiles for this purpose.

The summary statistics show that there are significant variations in water quality parameters across different observations. For example, the mean pH level is 7.08, but it ranges from 0 to 14. Similarly, the mean Hardness level is 196.37 mg/L, but it ranges from 47.43 to 323.12 mg/L. The summary statistics also show that there is an imbalance in the Potability variable, as only 40% of the observations have potable water.

Feature Engineering

We performed feature engineering to derive additional features from the existing dataset that could be useful for our water quality analysis. We calculated the Water Quality Index (WQI) based on the variables in the dataset.

The WQI is a composite indicator that reflects the overall quality of water based on multiple parameters. It is calculated by aggregating sub-indices for each parameter using weights and functions that reflect their relative importance and impact on water quality. We used the following formula to calculate the WQI:

WQI = w1 * f1(pH) + w2 * f2(Hardness) + w3 * f3(Solids) + w4 * f4(Chloramines) + w5 * f5(Sulfate) + w6 * f6(Conductivity) + w7 * f7(Organic_carbon) + w8 * f8(Trihalomethanes) + w9 * f9(Turbidity)

where wi are weights that sum up to 1, and fi are functions that map each parameter to a sub-index between 0 and 100. We used the following weights and functions based on literature review and expert judgment:

Parameter	**Weight**	**Function**
pH	0.11	f1(pH) = 100 - 5 *
Hardness	0.10	f2(Hardness) = 100 - 0.5 * Hardness
Solids	0.08	f3(Solids) = 100 - 0.1 * Solids
Chloramines	0.10	f4(Chloramines) = 100 - 2 * Chloramines
Sulfate	0.10	f5(Sulfate) = 100 - 0.5 * Sulfate
Conductivity	0.09	f6(Conductivity) = 100 - 0.02 * Conductivity
Organic_carbon	0.12	f7(Organic_carbon) = 100 - Organic_carbon
Trihalomethanes	0.07	f8(Trihalomethanes) = 100 - Trihalomethanes
Turbidity	0.08	f9(Turbidity) = 100 - Turbidity

We added the WQI as a new variable to the dataset and visualized its distribution and relationship with Potability. The histogram shows that the WQI is skewed to the left, with most of the observations having a low WQI.

Water Quality Analysis

We applied appropriate statistical and machine learning techniques to analyze the data and extract insights. We performed the following analyses:

Correlation Analysis

We examined the relationships between water quality parameters using correlation coefficients and p-values. We used Pearson's correlation coefficient for continuous variables and Spearman's rank correlation coefficient for ordinal variables

The correlation analysis results are shown in the table below:

Parameter1	Parameter2	Correlation Coefficient	P-value
pH	Hardness	0.08	<0.001
pH	Solids	-0.09	<0.001
pH	Chloramines	-0.03	0.18
pH	Sulfate	-0.01	0.64
pH	Conductivity	-0.02	0.38
pH	Organic_carbon	-0.04	0.06
pH	Trihalomethanes	-0.01	0.74
pH	Turbidity	-0.04	0.07
pH	Potability (Spearman)	-0.01	0.83

Team Members

This report was submitted by,

- Yuvaraj V **(2021115125)**
- Shadhani TT **(2021115302)**
- Adnan Khan S **(2021115304)**
- Vishal Raj Vellaisamy **(2021115120)**
- Vikhram SS **(2021115119)**
- Sabarisrinath R **(2021115335)**