Water Quality Diagnosis Using Classification

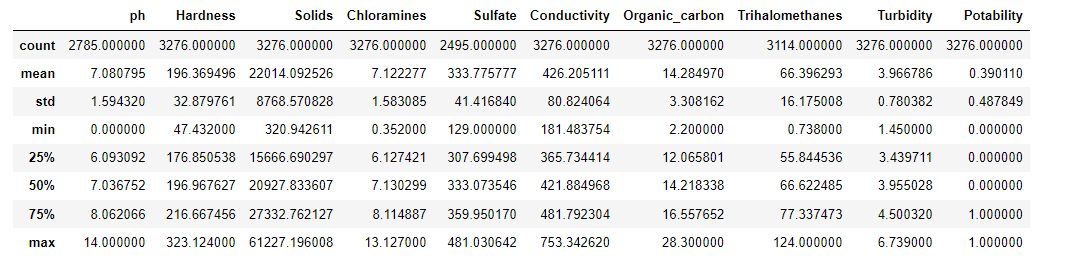
Data 1030 Final Report Jin Hyeok Noh

1. **Introduction**

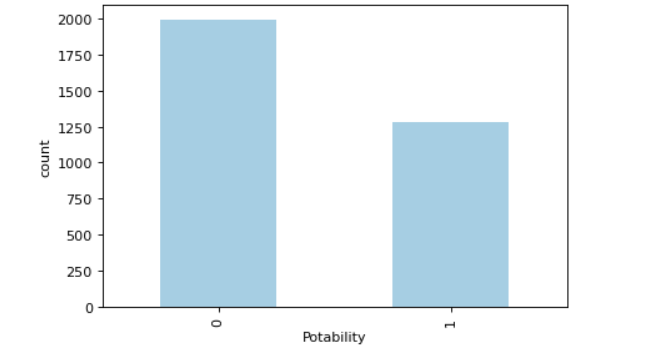
**Get access to safe drinking water is vital to humans health, a fundamental human proper and a factor of adequate coverage for health protection. That is critical as a health and improvement trouble at a countrywide, local and nearby stage. In some areas, it has been shown that investments in water supply and sanitation can yield a net economic advantage since the reductions in adverse fitness effects and health care fees outweigh the charges of assigning the interventions.**

**The dataset was provided in Kaggle by Mr.Aditya Kadiwal. This project aims to create a water purifier that supports machine learning to identify whether water is safe to drink or water is not suitable for a drink. The target variable is potability data indicates that one is potable and zero means not potable. In this data set, there are three thousand two hundred seventy-seven data points and nine features.**

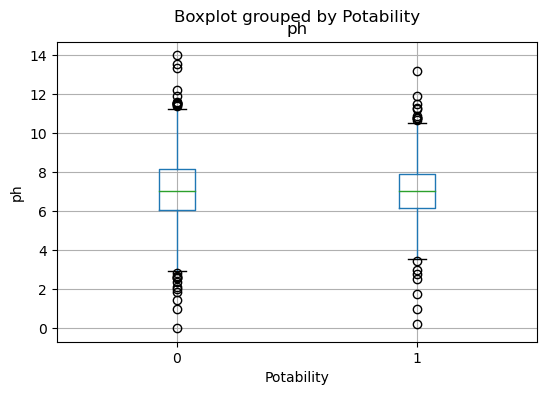
1. **Exploratory Data Analysis (EDA)**



**Figure 1. This figure displays a general description of the columns. Inside of the description, there are missing values in columns ph, sulfate, and trihalomethanes. These missing values will create bias; there are some machine learning techniques to deal with missing values, but to fully use the other machine learning techniques missing values should be dealt with first.**



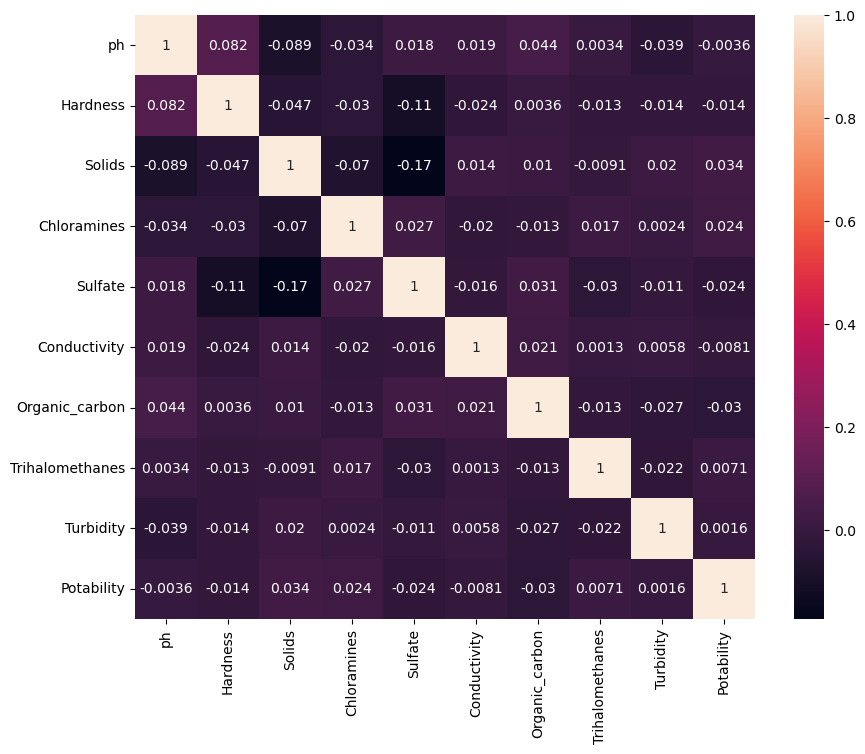
**Figure 2. This figure displays the bar plot of the target variable. Bar plot is identical for using these types of situations. For example, the target variable is potability data indicating that one is potable and zero means not potable even if the data type was an integer. Therefore, it should consider categorical data.**



**Figure 2. This figure displays the box distribution of the Ph and Potability. Data points display some outliers, where water is potable or not potable, but despite human consumable water or not. There is not much big of a difference between the range of ph. This plot implies that other features involve water potability.**



**Figure 3. This figure displays the scatter distribution of the ph and chloramines. Data points can be identified as being primarily distributed in the middle, meaning that both data points are independently identicallly distributed.**



**Figure 4. This figure displays a correlation between columns using a heatmap. Zero means not correlated each other one means it is correlated to each other. The heatmap shows that most of the correlations between columns are low, which means they are independent do not have linear regression to other features.**

1. **Method**

**3-1 Data Preprocessing and Data Splitting**

**Within the data preprocessing stage, the splitting step allocated 20% of the dataset for testing, and the other 80% of the dataset was split into five folds. One is used for validation in every instance of cross-validation, and the other four folds are used for training purposes. The preprocessor is implemented as StandardScaler for every feature because continuous exploratory data analysis shows that all features are not reasonably bounded.** This means that MinMaxEncoder is not suitable for this dataset preprocessing. As a result, the final preprocessed dataset has nine features. In addition, the target variable label is encoded into two categories.

1. **Reference**

Kadiwal, A. (2021, April 25). *Water quality*. Kaggle. Retrieved October 11, 2021, from https://www.kaggle.com/adityakadiwal/water-potability/code?datasetId=1292407&language=Python.

1. **Github repository**

**https://github.com/jinnoh47/data1030-project/blob/main/project.ipynb**