**American International University-Bangladesh**



**Course:** INTRODUCTION TO DATA SCIENCE

**Assignment Title:** Final-term assignment

**Submitted by:**

Name: Md. Shamim Siddiky  
Id: 20-42649-1  
Section: D

**Submitted to:**

TOHEDUL ISLAM

Assistant Professor, Computer Science  
American International University-Bangladesh

**Submission Date:** December 11, 2022

**Dataset Description**

**Name**: Water Potability

**Source**: <https://www.kaggle.com/datasets/adityakadiwal/water-potability>

**Attributes**: There are total 10 attributes in the dataset.

1. Ph: PH is an important parameter in evaluating the acid–base balance of water. WHO has recommended maximum permissible limit of pH from 6.5 to 8.5. The current investigation ranges were 6.52–6.83 which are in the range of WHO standards.
2. Hardness: Hardness is mainly caused by calcium and magnesium salts. These salts are dissolved from geologic deposits through which water travels.
3. Solids: Water has the ability to dissolve a wide range of inorganic and some organic minerals or salts. These minerals produced un-wanted taste and diluted color in appearance of water.
4. Chloramines: Chlorine and chloramine are the major disinfectants used in public water systems.
5. Sulfate: Sulfates are naturally occurring substances that are found in minerals, soil, and rocks.
6. Conductivity: Pure water is not a good conductor of electric current rather’s a good insulator. Increase in ions concentration enhances the electrical conductivity of water.
7. Organic carbon: Total Organic Carbon (TOC) in source waters comes from decaying natural organic matter (NOM) as well as synthetic sources.
8. Trihalomethanes: THMs are chemicals which may be found in water treated with chlorine. The concentration of THMs in drinking water varies according to the level of organic material in the water.
9. Turbidity: The turbidity of water depends on the quantity of solid matter present in the suspended state.
10. Potability: The class attribute which denotes if the water is safe or not.

**Instances**: There are total 3276 instances in the dataset.

**Classes**: The class attribute of the dataset is “Potability”. The dataset is classified with two classes;

1. Potable
2. Not Potable

**Performed Tasks**

1. Import and view dataset

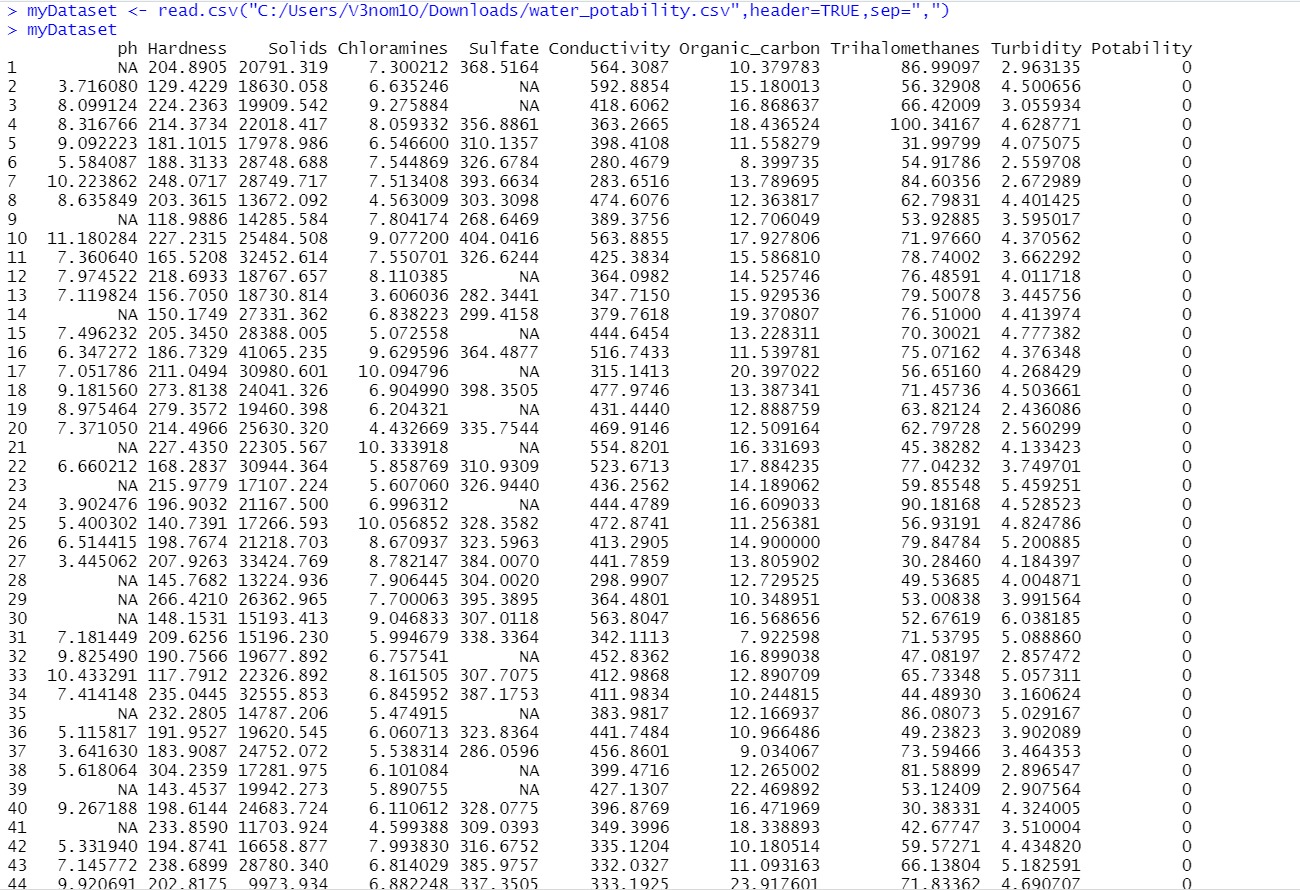


Fig 1: Dataset

1. Normalize Attribute value

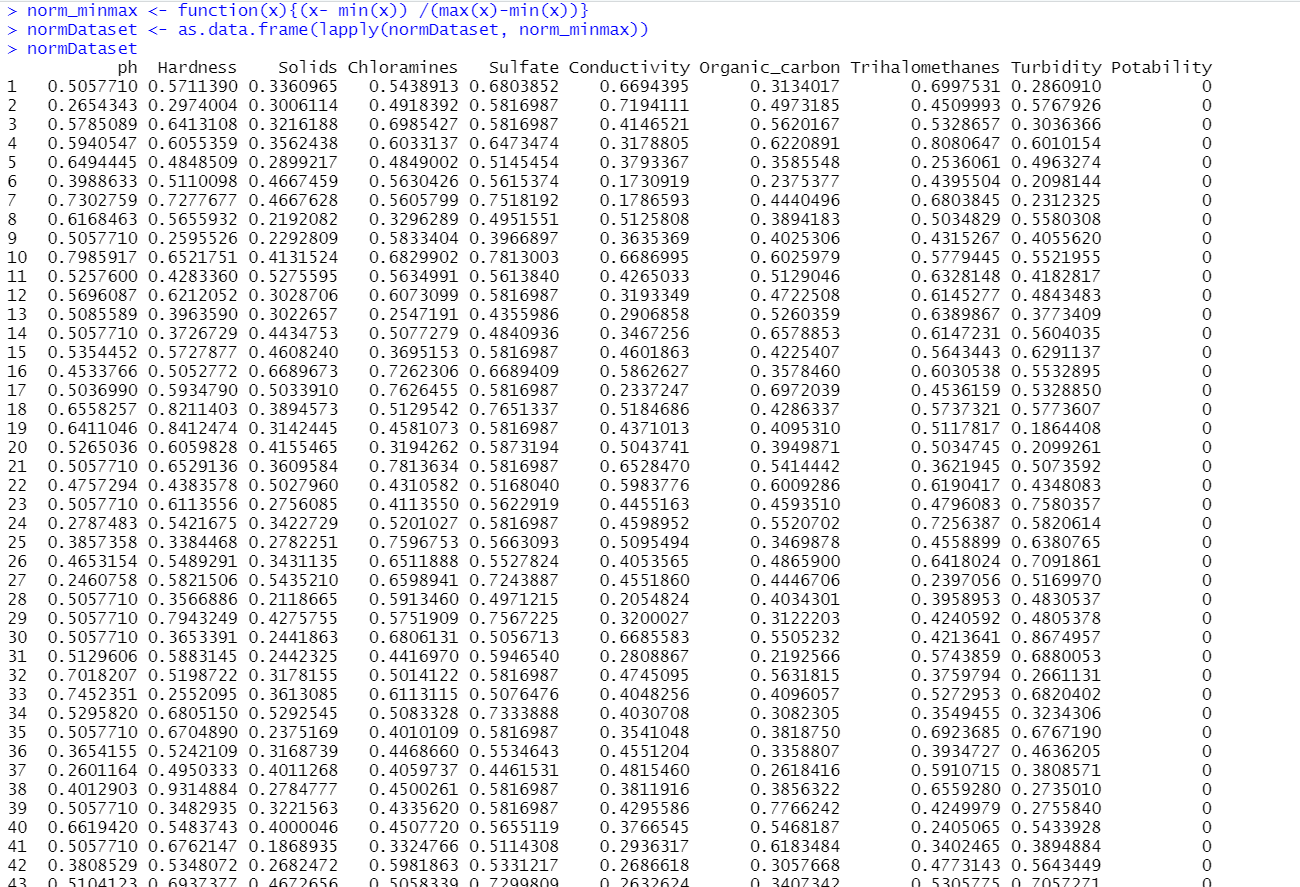


Fig 2: Normalized dataset

Before normalizing the dataset, the dataset was prepared. As there were many missing values for multiple instances in each attribute, so the missing values were replaced with the mean value for each attribute.

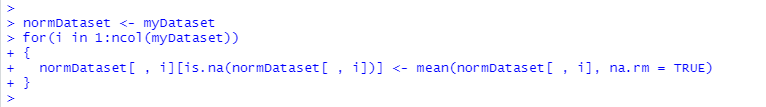


Fig 3: Replace missing values with mean

1. Split dataset into Training and Test dataset

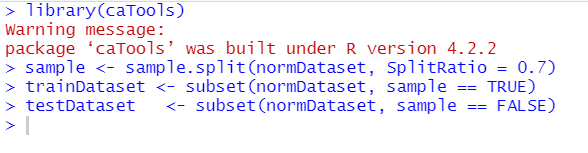


Fig 4: Splitting dataset in Training and Test dataset

1. Machine Learning model using KNN



Fig 5: Labeling class attribute

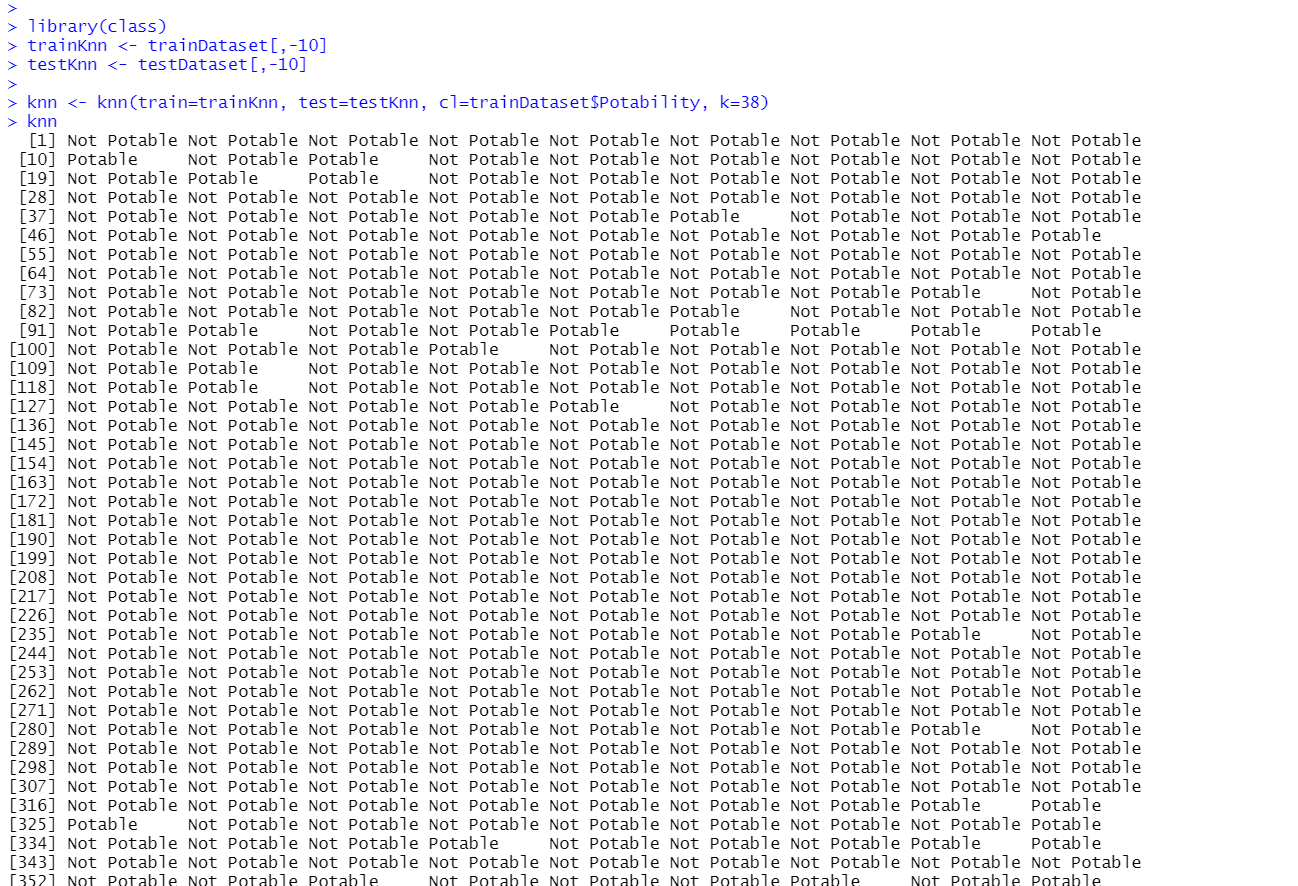


Fig 6: Perform KNN

1. Evaluate the Model, Create Confusion matrix and calculate accuracy

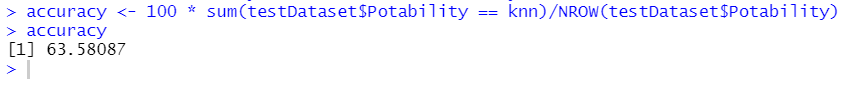


Fig 7: Calculate accuracy

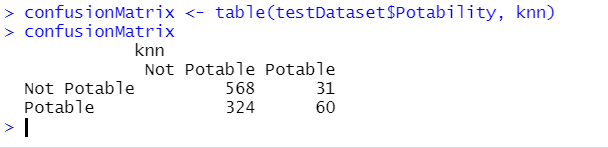


Fig 8: Generate confusion matrix

**Confusion Matrix**

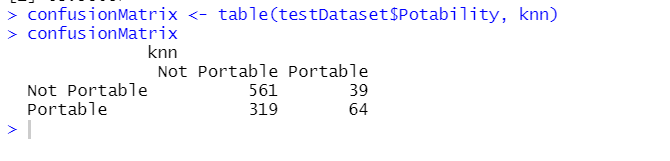


Fig 9: Confusion matrix

A confusion matrix is a table that is used to define the performance of a classification algorithm. Here the class Potability determines if water is adequate for human consumption, with 1 denoting potable and 0 representing unfit for potable. For better understanding, the values 0 and 1 were replaced with Not Potable and Potable before showing the confusion matrix.

In the above fig 9, the first row, first column (1,1) position indicates the number of not Potable instances that were correctly guessed by the model accurately. So, it is the **True Negative** value of the confusion matrix. The model guessed 561 not Potable instances correctly from the training dataset. The (2,1) position of the confusion matrix indicates the number of Potable instances that were classified as not Potable by the model. So, it is the **False Negative** portion of the confusion matrix. 319 Potable instances were guessed as Not Potable by the model.

Then, the first row-second column (1,2) position indicates the not Potable instances which were classified by the model as Potable. Therefore, it denotes the **False Positive** portion of the confusion matrix as the number of actual negative examples are classified as positive. Out of 983 instances, 39 instances were false positive. The (2,2) position indicates the Potable instances which were correctly classified by our model as Potable instances. A total of 64 Potable instances were classified correctly. It is the **True Positive** portion of the matrix.