Water Quality

Team members

- Tassneem Hamdy Kourany
- Manar Maged Mamdouh El Bagoury
- Habiba Mohamed Galal
- Karma Yasser Ismail Mahmoud
- Mariam Mohamed Attya Hamed

- Understand the dataset attributes

Data Contains 9 Attributes:

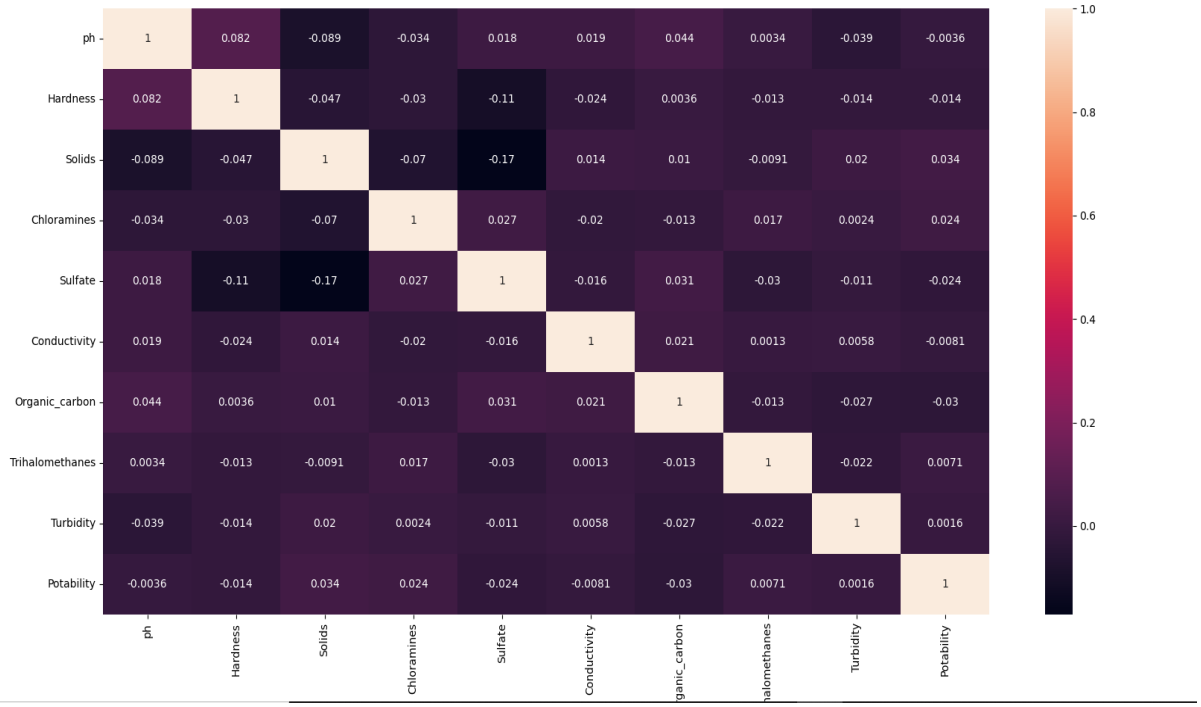
➢ **Type:**

- Numeric. [Float64 , Except Label is Int64]
- Non ordinal attributes except pH and Hardness.

```
RangeIndex: 3276 entries, 0 to 3275
Data columns (total 10 columns):
 #    Column           Non-Null Count  Dtype
---   ------           --------------  -----
 0    ph               2785 non-null   float64
 1    Hardness         3276 non-null   float64
 2    Solids           3276 non-null   float64
 3    Chloramines      3276 non-null   float64
 4    Sulfate          2495 non-null   float64
 5    Conductivity     3276 non-null   float64
 6    Organic_carbon   3276 non-null   float64
 7    Trihalomethanes  3114 non-null   float64
 8    Turbidity        3276 non-null   float64
 9    Potability       3276 non-null   int64
dtypes: float64(9), int64(1)
```

➢ **Linear Relation**

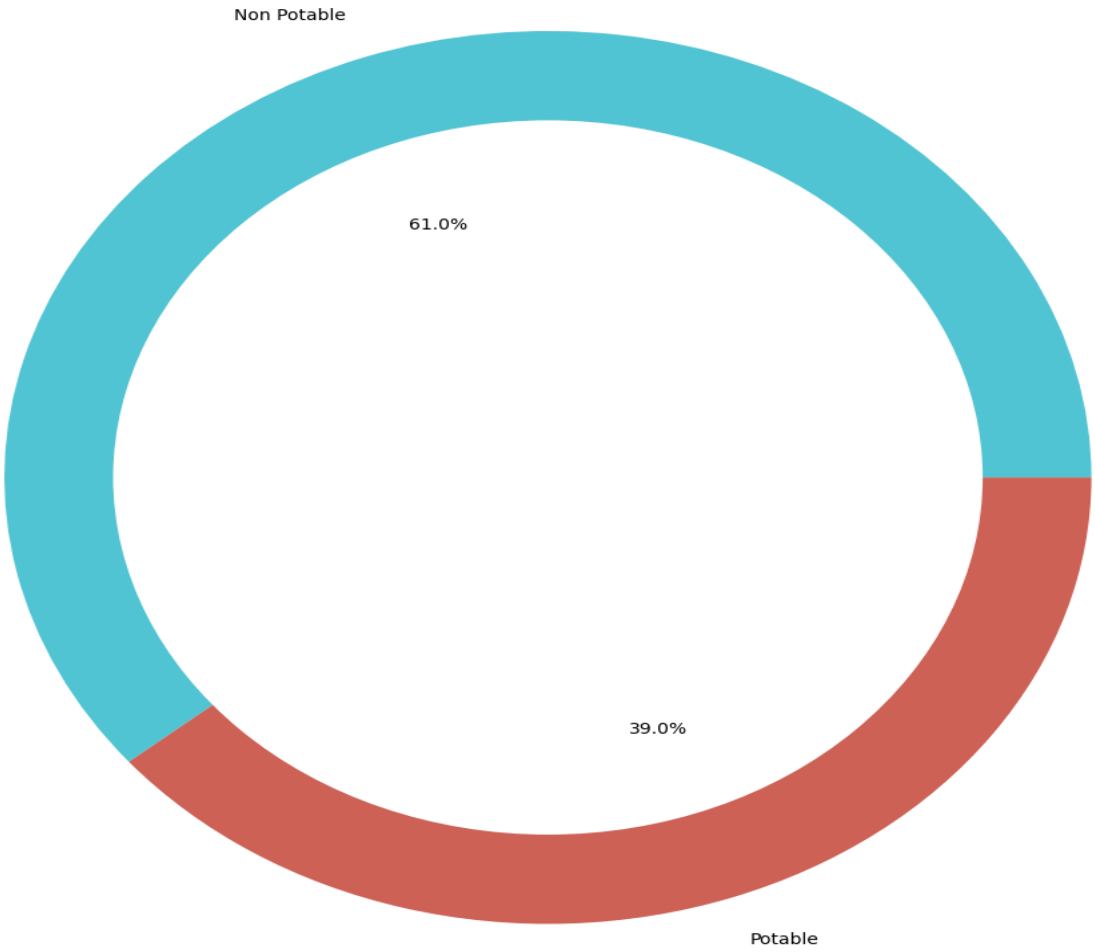
- No linear relation between attributes: As correlation factor close to 0
- The 2 features with the highest correlation are sulfate and solids.



- **Counting Nulls in each Column:**
    - pH: 491,
    - Sulphate: 781
    - Trihalomethanes: 162
    - Hardness, Solids, Chloramines, Conductivity, Organic carbon and Turbidity : 0

- **Understanding the Label to be predicted [Potability Ratio]**

-



-

- ## Apply the required data pre-processing methods

1. Handling Missing Values (library: sklearn)
    1.1. Dropping rows with NAs
    1.2. Dropping rows having both pH and Sulphate NA
    1.3. Dropping pH column
    1.4. Dropping Sulphate column
    1.5. Hot Deck Imputation (special form of prediction using knn)
    1.6. Multivariate feature Imputation (functions: enable_iterative_imputer, IterativeImputer)

2. Handling Outliers
    2.1. First, find boundary values
    2.2. Trimming of Outliers
    2.3. Capping on Outliers

3. Sampling
    3.1. Random Sampling
    3.2. Stratified Sampling
        3.2.1. Stratum as pH
        3.2.2. Stratum as Hardness

4. Dropping duplicates

5. Data Standardisation (function: StandardScaler)

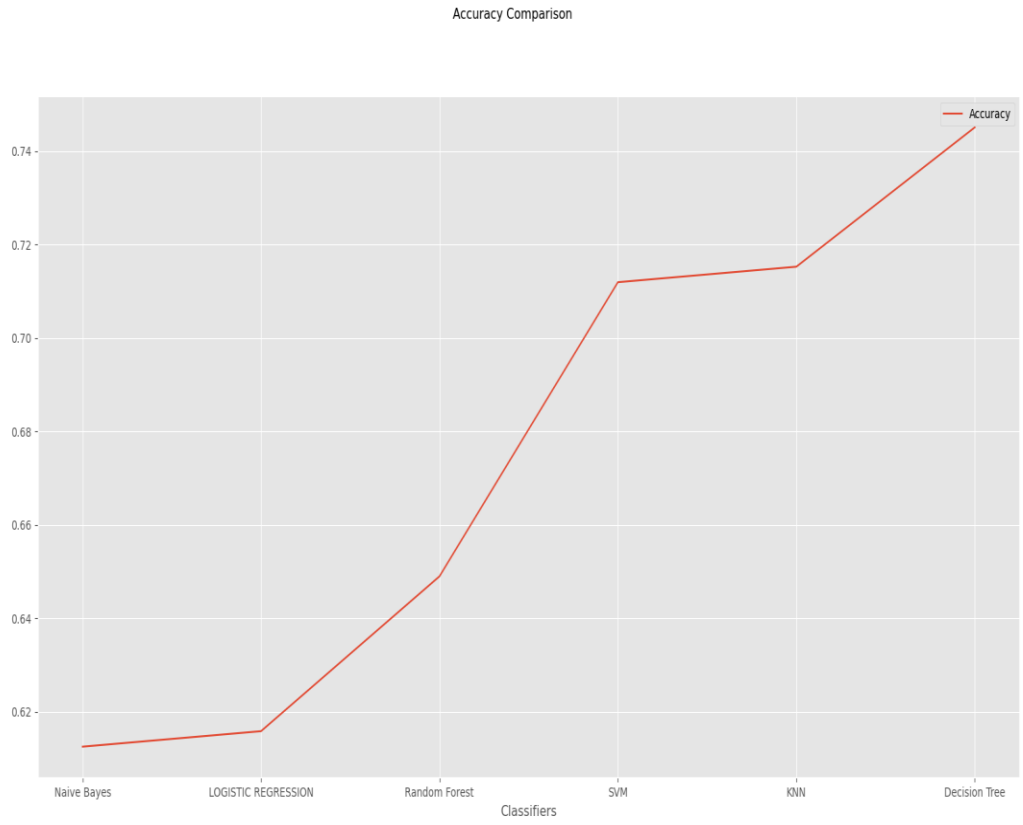
- ## Divide data set to training and testing data

  Normal Data Splitting (function: train_test_split)

- ## Implement the classification model that suits the project best

  Experimented all Classification Models: KNN, Naïve Bayes,Decision Tree, Random Forest, SVM and Logistic Regression

- Investigate the performance of your classifier on detecting the quality of water



- After building the classifier model. Use the testing data set to predict the water potability and calculate the error percentage

This step is done by Try and Error Criteria
That's to say, Applying different Pre-processing Flows [Methods and Techniques] and Pick out the best Paths

Number of experimented paths **18 Path**
Number of Picked Paths **5 Paths**
Top Path [Dropping rows with NAs, Dropping duplicates, Random Sampling, and Data Standardisation]
Highest Accuracy **76.8 %** goes to Decision tree.

- **Pre-processing scenarios, classifiers Performance and Error percentages**

1. 1st scenario: Dropping rows with NAs, Dropping duplicates, Random Sampling, and Data Standardisation.

	Accuracy	Error
KNN	73.841 %	26.476 %
Naïve Bayes	62.582 %	27.476 %
Decision Tree	76.827 %	24.178 %
SVM	72.182 %	27.807 %
Logistic Regression	59.602 %	40.410 %
Random Forest	70.198 %	29.099 %

2. 2nd scenario: Dropping Sulphate column, Dropping duplicates, Random Sampling, Hot Deck Imputation, and Data Standardisation.

	Accuracy	Error
KNN	69.308 %	30.691 %
Naïve Bayes	65.040 %	34.959 %
Decision Tree	70.528 %	29.471 %
SVM	68.902 %	31.097 %
Logistic Regression	64.430 %	35.569 %
Random Forest	65.447 %	34.552 %

3. 3rd scenario: Dropping duplicates, Random Sampling, Hot Deck Imputation, and Data Standardisation.

	Accuracy	Error
KNN	70.528 %	29.471 %
Naïve Bayes	64.227 %	35.772 %
Decision Tree	71.747 %	28.252 %
SVM	70.528 %	29.471 %
Logistic Regression	64.024 %	35.975 %
Random Forest	64.837 %	35.162 %

4. 4th scenario: Dropping rows with NAs, Dropping duplicates, Stratified Sampling Stratum as pH, and Data Standardisation.

	Accuracy	Error
KNN	69.918 %	30.081 %
Naïve Bayes	63.617 %	36.382 %
Decision Tree	72.764 %	27.235 %
SVM	70.528 %	29.471 %
Logistic Regression	64.634 %	35.365 %
Random Forest	63.821 %	36.178 %

5. 5th scenario: Iterative Imputation, Dropping duplicates, Random Sampling, and Data Standardisation.

	Accuracy	Error
KNN	68.543 %	31.456 %
Naïve Bayes	66.225 %	33.774 %
Decision Tree	58.948 %	41.059 %
SVM	74.834 %	25.165 %
Logistic Regression	65.231 %	34.768 %
Random Forest	66.887 %	33.112 %

- Mention your observations and study the parameters (features) that play a vital role in water potability.