**A CLASSIFICATION MODEL FOR WATER POTABILITY ANALYSIS USING LOGISTIC REGRESSION**.

This is a Capstone project done in fulfillment of the Kadatemy Data Science programme. This project was done as a team with the following members Chika Njoku, Vivian Ossai, Kamdilichukwu Egenti, Christian Okoye

## INTRODUCTION

The quality of water needs to be examined in order to ascertain its safety to people, whether it is for drinking, domestic use, food production or recreational purposes. Quality water supply and sanitation can boost countries' economic growth and overall public health. According to the UN General Assembly in 2010, everyone has the right to sufficient, continuous, safe, acceptable, physically accessible and affordable water for personal and domestic use

### What is Potable Water?

Safe drinking-water or potable water is water of sufficiently high quality that can be consumed or used with low risk of immediate or long term harm. In most developed countries, the water supplied to households, commerce and industry is all of drinking water standard, even though only a very small proportion is actually consumed or used in food preparation. Typical uses include washing and landscape irrigation (Brisbane City Council Information, 2005).

Current position for drinking-water, over large parts of the world, humans have inadequate access to potable water and use sources contaminated with disease vectors, pathogens or unacceptable levels of toxins or suspended solids. Such water is not wholesome and drinking or using such water in food preparation leads to widespread acute and chronic illnesses and is a major cause of death and misery in many countries. Actually, over one sixth of the world's population lacks safe drinking-water sources. Unsafe water supplies, along with deficient sanitary infrastructure and inadequate personal hygiene, contribute substantially to the burden of 2.2 million annual deaths from diarrhoeal diseases. Although the definitive solution to the problem of access to safe drinking-water is the universal provision of piped and treated water, this option remains elusive because of the enormous expenditure of money and time that is required (WHO, UNICEF, 2005).

**Statement of the problem** Many citizens receive high quality drinking-water every day from public water systems (which may be publicly or privately owned). Nonetheless, drinking-water safety cannot be taken for granted. Lack of access to safe drinking-water and sanitation continues to be a major problem in both rural and urban communities. There are a number of threats to drinking-water, that is, improperly disposed of chemicals, animal wastes, pesticides, human threats, wastes injected underground and naturally occurring substances. Therefore, there was a need to determine whether the chemical content of the water exceeds the WHO standards laid down for safe drinking-water. However, even when

water is safe for drinking at the source, it is commonly re-contaminated during collection, storage and use at home.

**Objectives** The main objective is to assess water potability . The specific objectives are;

1. To use the WHO drinking-water benchmarks to establish whether the contaminants exceed the safe levels.

2.To use logistic regression model to determine;

a) the key contaminants of drinking-water.

b) the level of safety of drinking-water in particular sources

Supervised learning is a machine learning algorithm with powerful tools to classify and process data.The model can be used to recognise new patterns and assign a target to them. Applications of supervised learning include classification (e.g. classifying players according to their behaviour during a game) and regression (e.g. Predicting household prices according to features).

Classification is an instance of supervised learning that includes a training phase to create a model (Classifier). Its task is to predict the class of items in a data set using a certain model of a classifier. The model is constructed using already-labelled items of similar data sets. This step allows classification techniques to be considered as a supervised machine learning method. There are many classification techniques but the one used here is Logistic Regression.

## MATERIAL AND METHODOLOGY

Data analysis tool used in this study
The data analysis tool used in this study is Matplotlib, Scikit-learn, Jupyter etc
**Matplotlib** is a plotting and visualization library. It was used in generating graphs with the analysed data such as bar plot, scatter plot etc.
**Scikit-learn** was used for implementing machine learning algorithms.
**Pandas Library** for data preprocessing analysis.
**Statistics Module** for statistical analysis.
**Seaborn** for data visualisation.

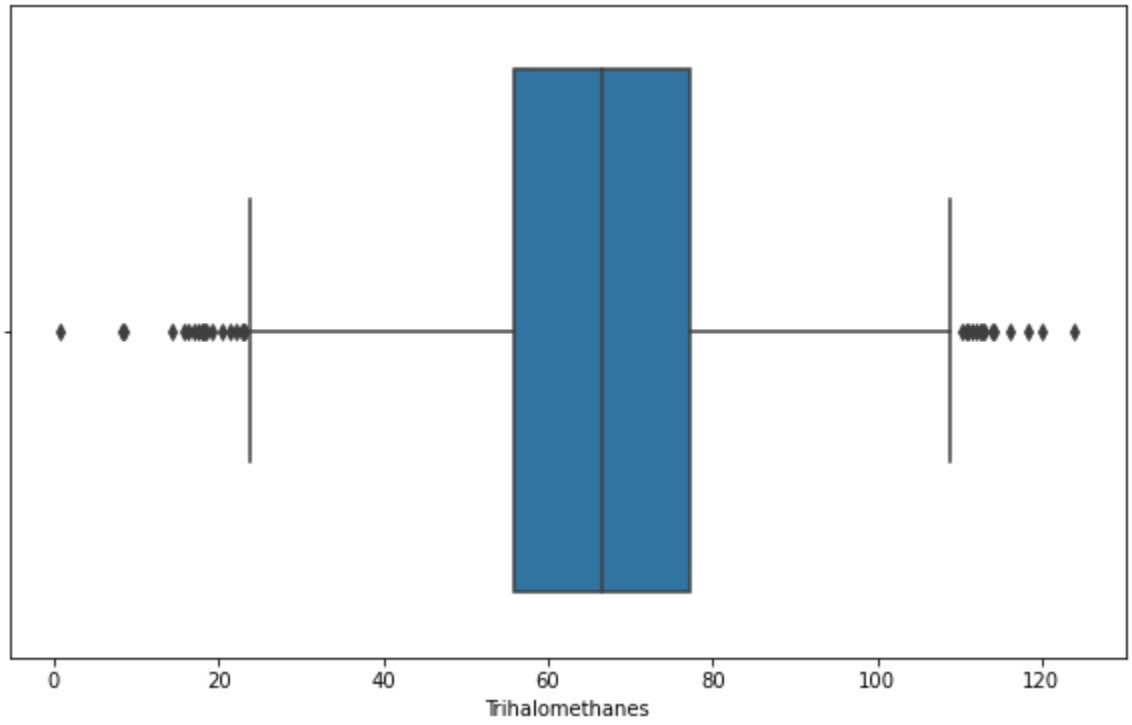
## DATASET

In this study, secondary data was used to build the logistic regression and this data was downloaded from https://www.kaggle.com/adityakadiwal/water-potability.
The analysis is carried out using publicly available data for water potability. The data is said to be synthetically generated. The dataset contains 3276 rows and 10 columns and we had a total of 1,434 missing values which we filled up with the median because we noticed a

significant number of outliers in the data set that resulted in unbalanced classes. We couldn't drop any columns because of the large presence of missing values that may invariably affect our model performance.

**Boxplot Analysis and Outlier Detection**

We chose box plot analysis for outlier detection because most of the parameters varied enough and were at the higher end of values and a box plot provides insightful visualization to decide outlier detection threshold values depending on the problem domain. Boxplot analysis shows most of the parameters lying outside the box deeming outliers normal so we adopted a strategy to fill out the missing values in the specific columns with median which is not easily affected by outliers and efficient for robust dataset like ours. For a specific column '**Trihalomethanes'** that had a total of 162 missing values we used the median figure which is 66.22485 so as not to bias the dataset.




**DATA ANALYSIS**

After all the data preprocessing and before applying machine learning algorithm there were some preliminary steps like data correlation analysis and data splitting to prepare the data to be given as input to actual machine learning algorithm

**Water quality parameters which is used for creating Model**

Several factors can interfere with surface water quality. Ten parameters were selected for the analysis. The parameters which were used to determine whether the water is clean is pH, Solids, Chloramines, Sulfate, Conductivity, Organic carbon, Trihalomethanes, Turbidity.

| Parameters Estimated | Who Standards |
|---|---|
| • **pH** | 6.5 - 8.5 |
| • **Solids** | 500mg/l -1000mg/l |
| • **Chloramines** | < 4mg/l |
| • **Sulfate** | 3-30 mg/L |
| • **Conductivity** | **<**400 µS/cm |
| • **Organic carbon**, | < 2 mg/L |
| • **Trihalomethanes** | <80 ppm |
| • **Turbidity** | **<**5.0 NTU |

**Classifiers Used**

For this research,Pandas libraries were selected for DataPreprocessing analysis. Matplotlib also has tools for preprocessing, basic statistics, data normalization, classification, Logistic regression.

The logistic Regression classifiers were run 10 times, the best result was added for this research, and the performance of the label evaluated, in compliance with the percentage of the correct classification. For the experiment, the water potability datasets were divided into two datasets: (1) training dataset with the 70% for training and testing dataset with 30% for validation.

**Evaluation process**

The confusion matrix helps practitioners to form a clear idea of whether the results have a high performance. The confusion matrix elements were: (1) true positive (TP), which tells that the water was potable and safe for drinking and was correctly predicted; (2) true negative (TN), where the water was not potable and unsafe for drinking was correctly predicted ; (3) false negative (FN), where the water was potable and safe for drinking but was wrongly falsely predicted as being not potable; and (4) false positive (FP), where the water was not potable and unsafe for drinking but was falsely predicted as being potable and safe for drinking water. In the medical field, false negatives are the most dangerous predictions. In this dataset classification we are more concerned with false positives (FP) . Our dataset was a binary classification. We have a 2 x 2 matrix with four values as shown below.

The different performance metrics were calculated using a confusion matrix. Accuracy (Acc) measured the properly classified instances [1]. The formula for calculating accuracy was given as the ratio between the number of correct predictions and the total number of predictions.

Accuracy simply measures how often the classifier makes the correct prediction.. The accuracy metric is not suited for unbalanced classes.

**Accuracy = TP + TN / TP + FP + FN + TN**

Accuracy has its own disadvantages, for imbalanced data, when the model predicts that each point belongs to the majority class label, the accuracy will be high. But, the dataset is not unbalanced. Our accuracy score was 0.65.

Our dataset has an unbalanced class so the precision score is more accurate and gives us better model performance.

Precision was the positive predictive value defined as the ratio of the total number of correctly classified positive classes divided by the total number of predicted positive classes. Which ideally should be 1 but we still had a good value of 0.75.

**Precision = TP / TP + FP OR Positive Predicted Value**

Recall identified the proportion of water being potable as the ratio of the total number of correctly classified positive classes divided by the total number of positive classes.

**Recall = TP / TP + FN**

The F1 score is considered a harmonic average between precision and recall.

**F1 Score = 2 * (Recall * Precision) / (Recall + Precision)**

The confusion matrix below shows 604 + 6 = 610 correct predictions and 371 + 6 = 377 as incorrect ones.

- **TP : True Positives (TP) : 6**
- **TN : True Negative (TN) : 604**
- **FP : False Positives (FP) : 2 (Type 1 error)**
- **FN : False Negative (FN) : 87 (Type 11 error)**

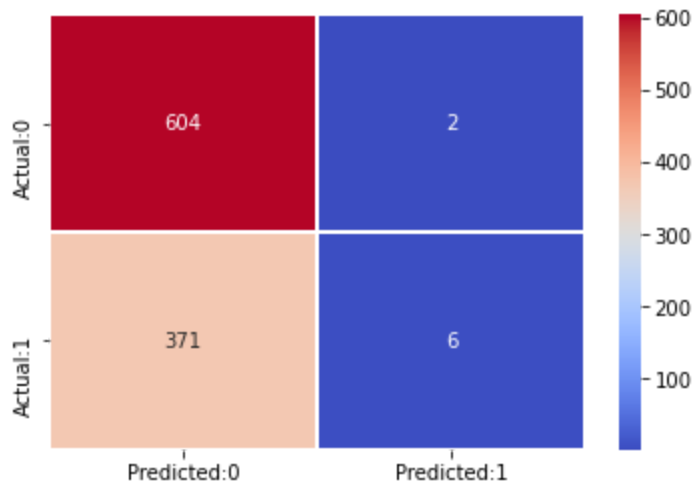



**Fig. 1: Confusion matrix of the dataset**

From the above figure using the statistics it is clear that the model is **more precise** than **accurate.**

**Exploratory Visual Analysis**

We also determined the ratio of water potability or not potable . We have the water potability count at a low rate of 1278 while the not potable has a high value count as 1998 from the fig below.

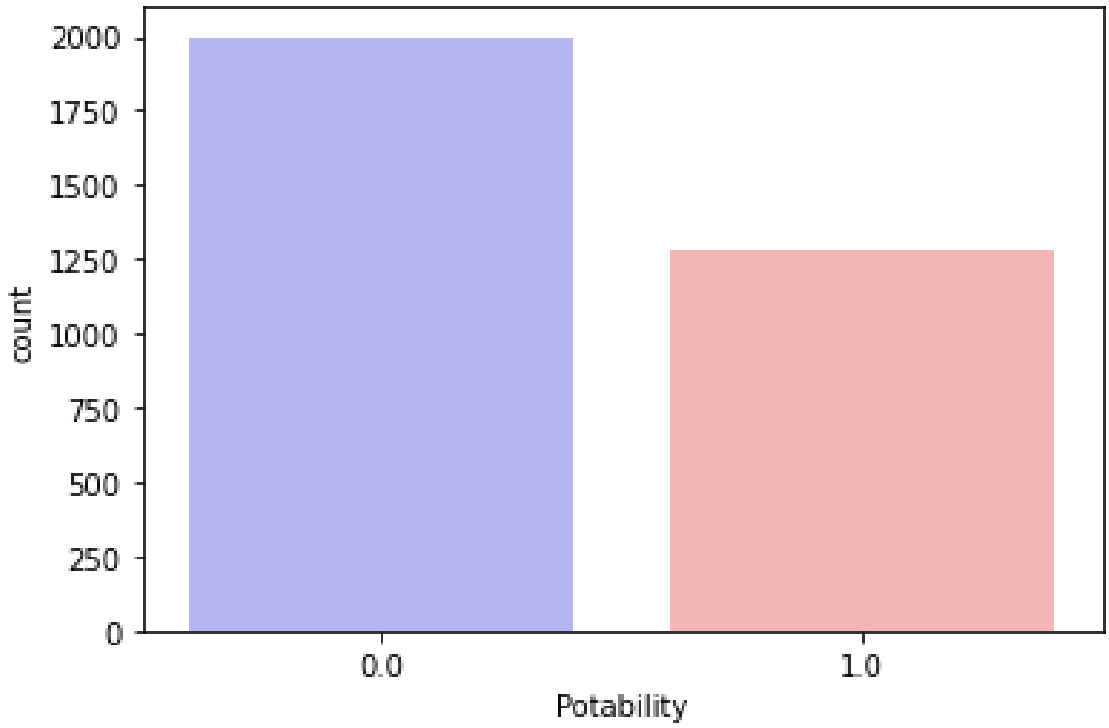



**Fig. 2: Potability distribution graph of the dataset**

This graph indicates the ratio of potable water to non potable water as contained in the dataset where 0.0 is an indicator for non potable whereas 1.0 indicates potable water.

The purpose of this step is to see the distribution of the data as shown below.

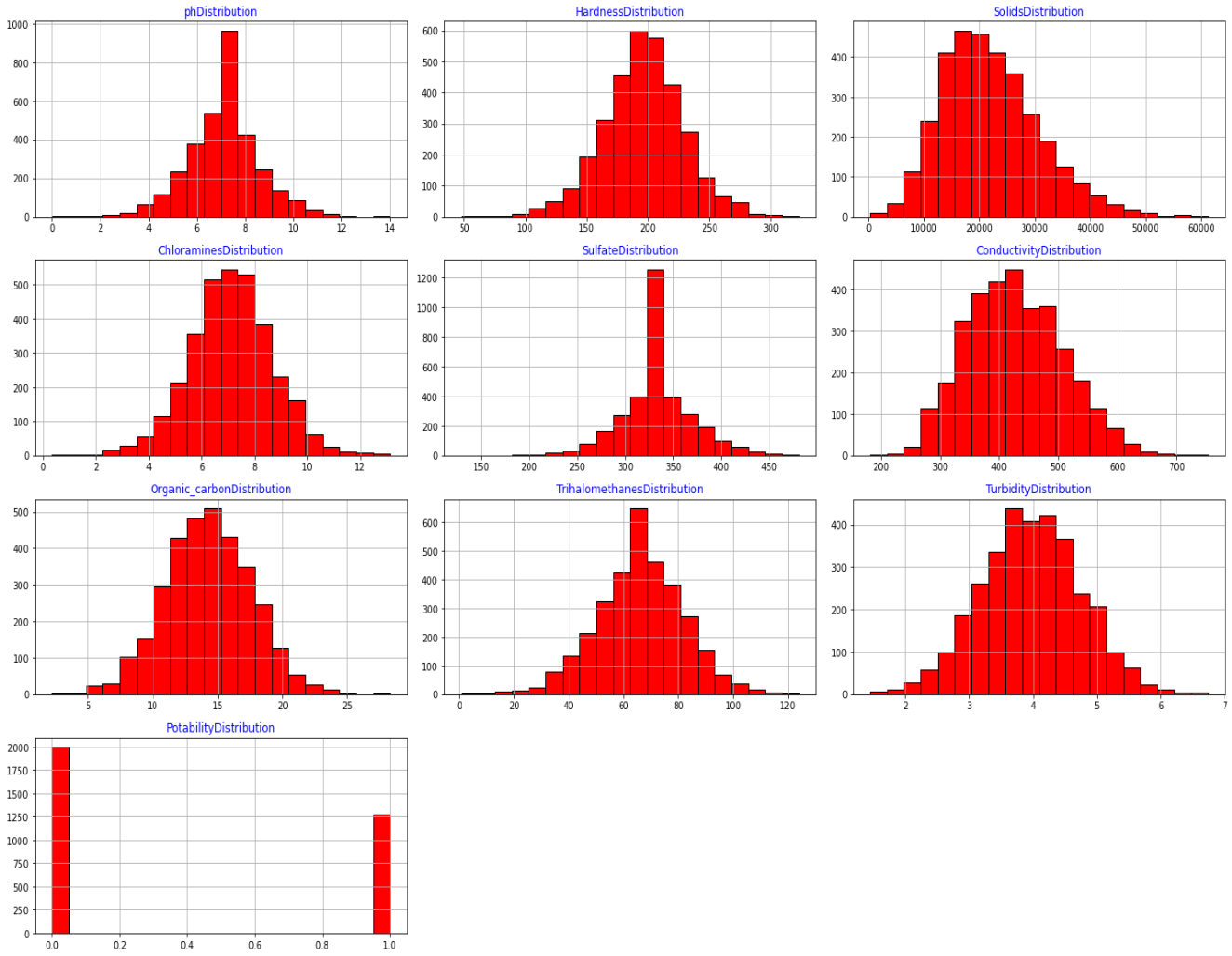



**Fig. 3: Distribution graphs of the features of the dataset**

**From the graphical representations above, we can deduce the following:**

**pH Distribution graph**: this shows the ph content of the different water sources in the dataset with the highest distributions having a ph in the range of 6-8 which are in the range of WHO standards. WHO has recommended a maximum permissible limit of pH from 6.5 to 8.5.

**Hardness Distribution graph:** Hardness is mainly caused by calcium and magnesium salts and the length of times these materials are in contact with water determines its degree of hardness. The graph shows that from 150 - 250 has a high degree of these chemicals hence producing a hard water.

**Solid Distribution graph:** The water with high solid value indicates that water is highly mineralized. The Desired limit for solid is 500 mg/l and maximum limit is 1000 mg/l which is

prescribed for drinking purpose. The graph shows the dataset contains water with high solid content values as high as 1000-4000.

**Chloramine Distribution graph:** The dataset contains water sources with high chloramines with the highest distribution being between 6 - 8 as against the reference range of 4 mg/l.

**Sulfate Distribution graph:** This shows a high concentration of sulfate between 250 - 350 in the distribution graph. Sulfates are naturally occurring substances that are found in minerals, soil, and rocks and range from 3-30 mg/l in most fresh water supplies although much higher concentration(1000 mg/l) can be found in some locations.

**Conductivity Distribution graph:** Conductivity actually measures the ionic process of a solution that enables it to transmit current. According to WHO standards, conductivity value should not exceed 400 µS/cm. From the graph, it is shown that the dataset contains water with high conductivity which is as a result of high solid content.

**Organic Carbon Distribution graph:** This is the organic content present in water and naturally, all forms of water contain organic carbon with a maximum acceptable limit of 2 mg/L for drinking water. From the distribution graph, over 99% of the samples are over the accepted limit. High organic carbon content signifies increase in growth of microorganisms in the water which results in a depletion of oxygen present in the water.

**Trihalomethanes Distribution graph:** Natural organic matter in water reacts with chlorine used in water purification to give trihalomethanes. At concentrated levels, they are known to be carcinogenic and impede natural reproduction. The maximum acceptable limit is 80 ppm (parts per million) and the distribution of the dataset shows the majority of the samples falling within the 0-80 ppm range with highest distribution between 63-67 ppm.

**Turbidity Distribution graph:** According to WHO, turbidity is defined as the amount of cloudiness in water. It is caused by the number of individual particles present in the water and these particles include dust, sand, organic materials, etc. Turbidity is an important test of water quality/potability. The lower the turbidity, the clearer the water and vice versa. The graph shows that a vast majority of the water sources are within the acceptable limit with highest distributions between 3.5-4.5 NTU and this is usually indicative of clear water.

**Potability Distribution graph:** Water is either drinkable or not. Potability is the ability of water to be safe for drinking. This is represented within the graph as "0.0" for non-potable (not fit for drinking) water and "1.0" for potable (fit for drinking) water. It is clear that the majority of the water sources are not fit for drinking with about 2000 of these sources (which accounts for roughly 61% of the samples) shown to be not potable with readings of 0.0 .

**CONCLUSION:**

In this study, water potability was implemented using data analysis tools. The analysis of water Chloramines, pH level and conductivity etc. can play a major role in assessing water quality. Modeling and prediction of water quality are very important for the protection of the environment. Several popular methods, such as Mann-Kendall, the seasonal Kendall test, and multiple regression methods,
are provided to detect and assess changes of various water portability parameters (Helsel,1992).

First, the present study explored an alternative method of predicting water quality by employing minimal and available water quality parameters. Datasets used for this analysis were obtained from Kaggle.

After examining the robustness and efficiency of the proposed model for predicting the water potability, in future work, the developed models will be implemented to predict the water quality.

**REFERENCES:**


Ahmed U. et al, (2019). Efficient Water Quality Prediction Using Supervised Machine Learning. *Published by Springer.* https://www.researchgate.net/publication/336808732_Efficient_Water_Quality_Prediction_Using_Supervised_Machine_Learning

Brisbane City Council, (2005). Retrieved from https://www.brisbane.qld.gov.au/planning-and-building/planning-guidelines-and-tools/brisbane-city-plan-2014/superseded-brisbane-city-plan-2000/superseded-subdivision-and-development-guidelines/2005-edition

Gakii C., Jepkoech J., (2019). A Classification Model For Water Quality Analysis Using Decision Tree. *Published by European Centre for Research Training and Development UK, Vol.7, No.3, pp.1-8.*

Helsel, D.R. and Hirsch, R.M. (1992) Statistical Methods in Water Resources. Published by *Elsevier, Vol 49.*

Nuzulul K. N., (2021). How to Predict Coronary Heart Disease Risk using Logistic Regression? Medium. https://medium.com/analytics-vidhya/how-to-predict-coronary-heart-disease-risk-using-logistic-regression-c069ab95cbec

Retrieved from https://en.m.wikipedia.org/wiki/Total_organic_carbon

Retrieved from https://en.m.wikipedia.org/wiki/Turbidity

Retrieved from https://www.kaggle.com/kanncaa1/water-quality-explanatory-data-analysis

USGS. Turbidity and Water. Retrieved from https://www.usgs.gov/special-topic/water-science-school/science/turbidity-and-water?qt-science_center_objects=0#qt-science_center_objects

World Health Organization, UNICEF, (2005). Water for life: Making it happen. Retrieved from http://www.who.int/water_sanitation_health/monitoring/jmp2005/en/