Model On Water Quality Dataset:

# 

B-Tech, Data Science Engineering

Abstract:

The quality of water needs to be examined 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 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. 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.

Features:

A picture containing crossword puzzle, shoji, text, scoreboard

Description automatically generated

### pH value: PH is an important parameter in evaluating the acid–base balance of water. It is also the indicator of acidic or alkaline condition of water status. WHO(World Health Organisation) 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.

* **Hardness:** Hardness is mainly caused by calcium and magnesium salts. These salts are dissolved from geologic deposits through which water travels. The length of time water is in contact with hardness producing material helps determine how much hardness there is in raw water. Hardness was originally defined as the capacity of water to precipitate soap caused by Calcium and Magnesium.
* **Solids (Total dissolved solids - TDS):** Water can dissolve a wide range of inorganic and some organic minerals or salts such as potassium, calcium, sodium, bicarbonates, chlorides, magnesium, sulphates etc. These minerals produced un-wanted taste and diluted colour in appearance of water. This is the important parameter for the use of water. The water with high TDS value indicates that water is highly mineralized. Desirable limit for TDS is 500 mg/l and maximum limit is 1000 mg/l which prescribed for drinking purpose.
* **Chloramines:** Chlorine and chloramine are the major disinfectants used in public water systems. Chloramines are most formed when ammonia is added to chlorine to treat drinking water. Chlorine levels up to 4 milligrams per litre (mg/L or 4 parts per million (ppm)) are considered safe in drinking water.
* **Sulphate:** Sulphates are naturally occurring substances that are found in minerals, soil, and rocks. They are present in ambient air, groundwater, plants, and food. The principal commercial use of sulphate is in the chemical industry. Sulphate concentration in seawater is about 2,700 milligrams per litre (mg/L). It ranges from 3 to 30 mg/L in most freshwater supplies, although much higher concentrations (1000 mg/L) are found in some geographic locations.
* **Conductivity:** Pure water is not a good conductor of electric current rather a good insulator. Increase in ions concentration enhances the electrical conductivity of water. Generally, the amount of dissolved solids in water determines the electrical conductivity. Electrical conductivity (EC) measures the ionic process of a solution that enables it to transmit current. According to WHO standards, EC value should not exceed 400 μS/cm.
* **Organic carbon:** Total Organic Carbon (TOC) in source waters comes from decaying natural organic matter (NOM) as well as synthetic sources. TOC is a measure of the total amount of carbon in organic compounds in pure water. According to US EPA < 2 mg/L as TOC in treated / drinking water, and < 4 mg/Lit in source water which is use for treatment.
* **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, the amount of chlorine required to treat the water, and the temperature of the water that is being treated. THM levels up to 80 ppm is considered safe in drinking water.
* **Turbidity:** The turbidity of water depends on the quantity of solid matter present in the suspended state. It is a measure of light emitting properties of water and the test is used to indicate the quality of waste discharge with respect to colloidal matter. The mean turbidity value obtained for Wondo Genet Campus (0.98 NTU) is lower than the WHO recommended value of 5.00 NTU.
* **Potability:** Indicates if water is safe for human consumption where 1 means Potable and 0 means Not potable.

Diagram

Description automatically generated

**Data Preparation:**

|  |  |
| --- | --- |
| Data Manipulation | No |
| Data Blending | No |
| Missing Values Handling | Yes |
| Dimensionality Reduction | No |
| Outlier Removal | No |
| Label Encoding | No |

**Corelation matrix:** We can observe and say that no chance of dimensionality reduction since we cannot pick out any attribute.

**Teams

Description automatically generated with medium confidence**

Algorithms Usage:

### Logistic Regression: Logistic Regression is used to estimate discrete values (usually binary values like 0/1) from a set of independent variables. It helps predict the probability of an event by fitting data to a logit function. It is also called logit regression.

### These methods listed below are often used to help improve logistic regression models:

1. Include interaction terms.
2. Eliminate features.
3. Regularize techniques.
4. Use a non-linear model.

* Accuracy = 62.25

Chart, treemap chart

Description automatically generated

### Decision Tree: Decision Tree algorithm in machine learning is one of the most popular algorithms in use today; this is a supervised learning algorithm that is used for classifying problems. It works well classifying for both categorical and continuous dependent variables. In this algorithm, we split the population into two or more homogeneous sets based on the most significant attributes/ independent variables.

### Chart, treemap chart Description automatically generatedA picture containing graphical user interface Description automatically generated

### **Accuracy = 57.98**

* **SVM (Support Vector Machine) Algorithm:** SVM algorithm is a method of classification algorithm in which you plot raw data as points in an n-dimensional space (where n is the number of features you have). The value of each feature is then tied to a particular coordinate, making it easy to classify the data. Lines called classifiers can be used to split the data and plot them on a graph.

**Confusion matrix
**

* Accuracy = 68.15
* **Naive Bayes Algorithm:** A Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature. Even if these features are related to each other, a Naive Bayes classifier would consider all these properties independently when calculating the probability of a particular outcome. A Naive Bayesian model is easy to build and useful for massive datasets. It's simple and is known to outperform even highly sophisticated classification methods.

**Chart, treemap chart

Description automatically generated**

* Accuracy = 61.24
* **KNN (K- Nearest Neighbours) Algorithm:** This algorithm can be applied to both classification and regression problems. Apparently, within the Data Science industry, it's more widely used to solve classification problems. It’s a simple algorithm that stores all available cases and classifies any new cases by taking a majority vote of its k neighbours. The case is then assigned to the class with which it has the most in common. A distance function performs this measurement. KNN can be easily understood by comparing it to real life.

For example, if you want information about a person, it makes sense to talk to his or her friends and colleagues.

Things to consider before selecting K Nearest Neighbours Algorithm:

1. KNN is computationally expensive.
2. Variables should be normalized, or else higher range variables can bias the algorithm.
3. Data still needs to be pre-processed.

Chart, treemap chart

Description automatically generated

* Accuracy = 66.32
* **K-Means:** It is an unsupervised learning algorithm that solves clustering problems. Data sets are classified into a particular number of clusters (let's call that number K) in such a way that all the data points within a cluster are homogenous and heterogeneous from the data in other clusters.

How K-means forms clusters:

1. The K-Means Algorithm picks k number of points, called centroids, for each cluster.
2. Each data point forms a cluster with the closest centroids, i.e., K clusters.
3. It now creates new centroids based on the existing cluster members.
4. With these new centroids, the closest distance for each data point is determined. This process is repeated until the centroids do not change.

Elbow Method: In cluster analysis, the elbow method is a heuristic used in determining the number of clusters in a data set. The method consists of plotting the explained variation as a function of the number of clusters and picking the elbow of the curve as the number of clusters to use.

* Accuracy = 62.05

Chart, treemap chart

Description automatically generatedShape

Description automatically generated

### Random Forest Algorithm: A collective of decision trees is called a Random Forest. To classify a new object based on its attributes, each tree is classified, and the tree “votes” for that class. The forest chooses the classification having the most votes (over all the trees in the forest).

### Each tree is planted & grown as follows:

### If the number of cases in the training set is N, then a sample of N cases is taken at random. This sample will be the training set for growing the tree.

### If there are M input variables, a number m<<M is specified such that at each node, m variables are selected at random out of the M, and the best split on this m is used to split the node. The value of m is held constant during this process.

### Each tree is grown to the most substantial extent possible. There is no pruning.

### Accuracy = 67.14

### Chart, treemap chart Description automatically generatedA picture containing night Description automatically generated

**LIMITATIONS:**

We would have loved for the dataset to contain geographical locations and well as date of collections of data. It would have helped in getting a more precise result.

**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. Modelling and prediction of water quality are very important for the protection of the environment.

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.

### A screenshot of a computer Description automatically generated with low confidenceChart, bar chart Description automatically generated