Water Quality Prediction using Machine Learning

Introduction

In this tutorial, we are going to implement a water quality prediction using machine learning techniques. In this technique, our model predicts that the water is safe to drink or not using some parameters like Ph value, conductivity, hardness, etc.

About dataset :-

Context

Access to safe drinking-water is essential to health, a basic human right and a component of effective policy for health protection. This is important as a health and development issue at a national, regional and local level. In some regions, it has been shown that investments in water supply and sanitation can yield a net economic benefit, since the reductions in adverse health effects and health care costs outweigh the costs of undertaking the interventions.

Content

The water_potability.csv file contains water quality metrics for 3276 different water bodies.

1. 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 has recommended a 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. 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.

3. Solids (Total dissolved solids - TDS):

Water has the ability to dissolve a wide range of inorganic and some organic minerals or salts such as potassium, calcium, sodium, bicarbonates, chlorides, magnesium, sulfates etc. These minerals produced an unwanted taste and diluted color in the 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. The Desired limit for TDS is 500 mg/l and maximum limit is 1000 mg/l which is prescribed for drinking purposes.

4. Chloramines:

Chlorine and chloramine are the major disinfectants used in public water systems. Chloramines are most commonly formed when ammonia is added to chlorine to treat drinking water. Chlorine levels up to 4 milligrams per liter (mg/L or 4 parts per million (ppm)) are considered safe in drinking water.

5. Sulfate:

Sulfates 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 sulfate is in the chemical industry. Sulfate concentration in seawater is about 2,700 milligrams per liter (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.

6. Conductivity:

Pure water is not a good conductor of electric current rather it's 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) actually 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.

7. 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 the US EPA < 2 mg/L as TOC in treated / drinking water, and < 4 mg/Lit in source water which is used for treatment.

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, 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.

9. 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.

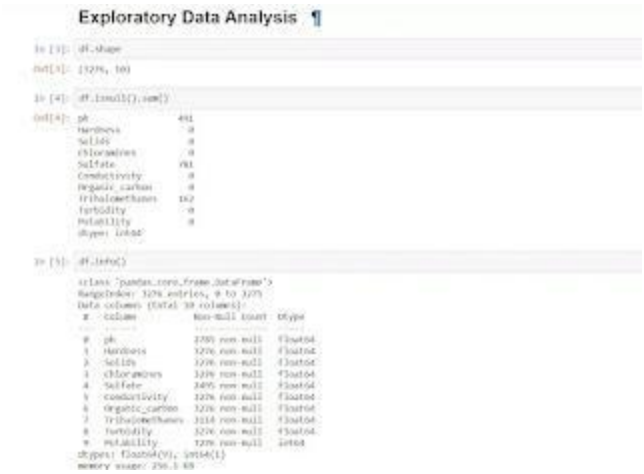
10. Potability:

Indicates if water is safe for human consumption where 1 means Potable and 0 means Not potable. (0) Water is not safe to drink and (1) Water is safe to drink.

Let's start :-

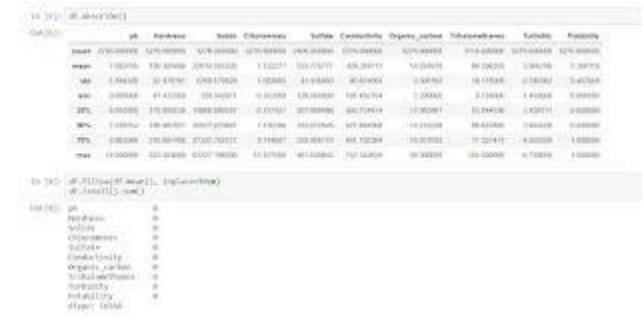
Import all the required libraries which are used to train the model or visualise the data. Then load the data set using a Pandas's function read_csv() and display the top five rows of the data set.

Then perform Exploratory Data Analysis. In EDA,Firstly check the shape of the data set. Then check that there are Null values or not and you can see in the below image that ph, Sulfate, Trihalomethanes contain NULL values. Then check the information of the data set.



Now describe the dataset which shows the minimum value, maximum value, mean value, count, standard deviation, etc.

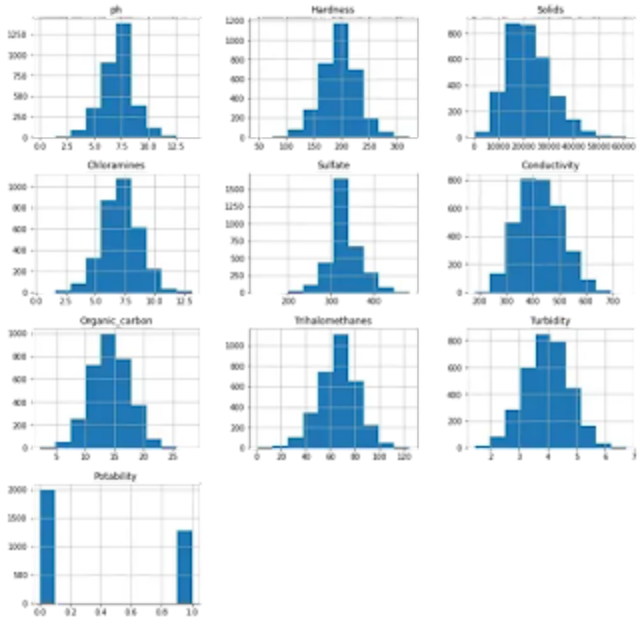
Then finally we handle the missing values. We filled the missing values in our feachers using a mean value of each feature which means we filled the mean value to handle missing data. Then again check that there are null values present or not.



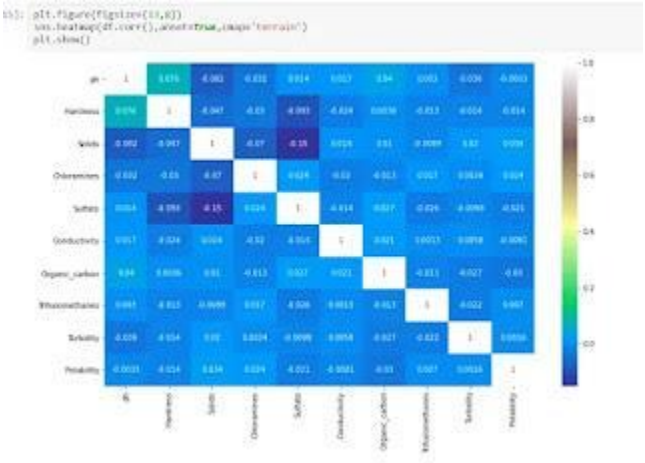
Check the value counts of our target feature Potability. Then visualize the portability using a countplot function of seaborn.

Now visualize the pH value using a distplot function to check that it contains a normal distribution or not. So, you can see that it is a normal distribution.

Visualize all the features of the data set as you can see below.

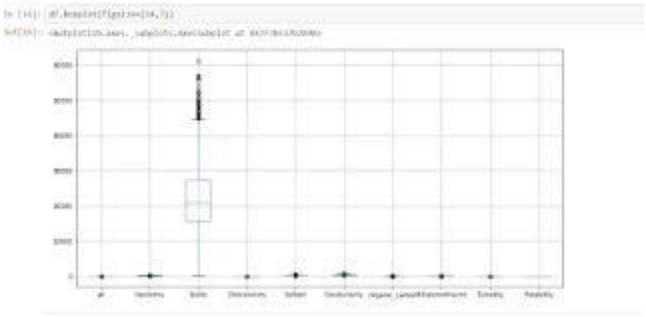


Visualize the correlation of all the features using a heat map function of seaborn. but you can see in the below heatmap that there is no correlation between any feature; it means that we can’t reduce the dimension.



Now see the outlier using a boxplot function. So, you can see that the Solid feature contains outliers but we can't remove the outliers from it because if we remove the outliers from the Solid feature. So, water will be safe to drink every time. It contains an

outlier to make the water impure which means it will tell us that water is safe or not. Solid may be high to make the water unsafe to drink.



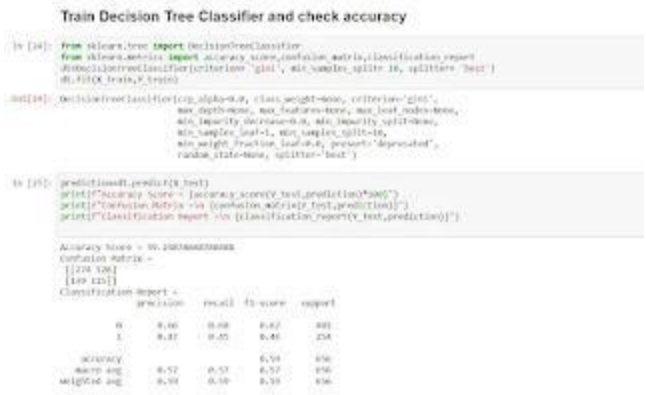
Now it’s time to prepare the data set. Divide the data into the independent and dependent features. All are independent features except Potability because Potability is our dependent feature.

Split the data set into the training and testing using the train_test_split function which returns four data sets.

Now define the decision tree classifier model and train the model using the data set (X_train, Y_train).

Then test the model using the test data set (X_test).

Now it’s time to evaluate the model using the accuracy score, confusion matrix and classification report. Evaluation techniques take two parameters; one is the actual data and the other one is a predicted data. And You can see that overall accuracy is 59%.



Then test the model on a custom data set and you can also see the result in the below image.

Now apply a hyper parameter tuning on the decision tree classifier using a GridSearchCV. Only three parameters to tune the model as you can see in the image below.

Now see how much time the model has trained with different parameters. Also check the training and testing accuracy as you can see that the training accuracy is 90% and testing accuracy is 60% which is okay.