POWER HACKS 2nd EDITION

ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING

WATER QUALITY DOCUMENTATION

AUTHOR: MUSHINDI RACHEL

BUSINESS UNDERSTANDING

Some water companies across the world are facing a hard time to conclude whether the water they are testing is potable or not. This is a threat to human beings who are at the consumption end.

Therefore coming up with a predictive model where the water quality parameters are inputted and the model is able to predict whether the water is potable will be a great help to the companies and also to the health of consumers.

The objective of the task

To develop a machine learning model that predicts the potability of water based on data values of water properties in order to improve peoples lives.

Data Acquisition

Source Systems

It was shared electronically in excel format.

Exploratory Data Analysis

The exploratory data analysis process

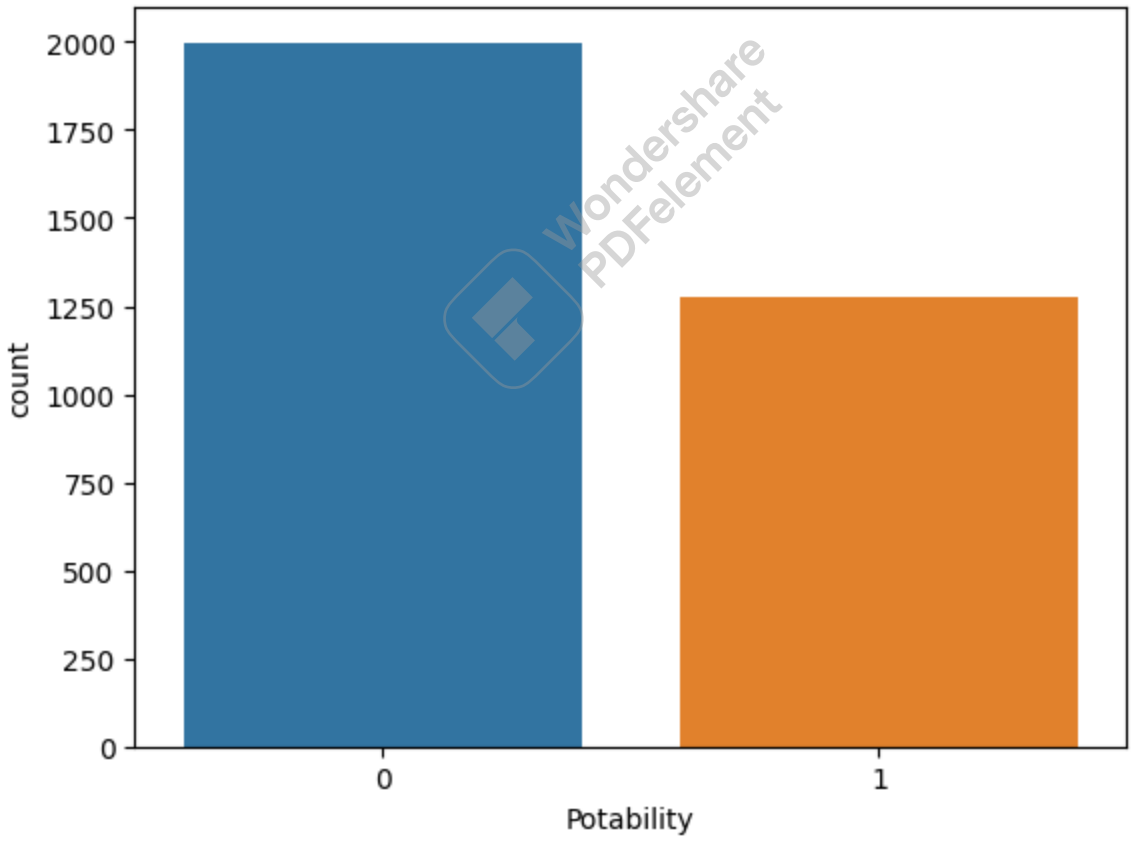
I used the python libraries tool for visualization, these were matplotlib and seaborn.

Under the libraries, I used bar plots to plot the graphs.

Different bar graphs have shown the relationships among variables in the data frame.

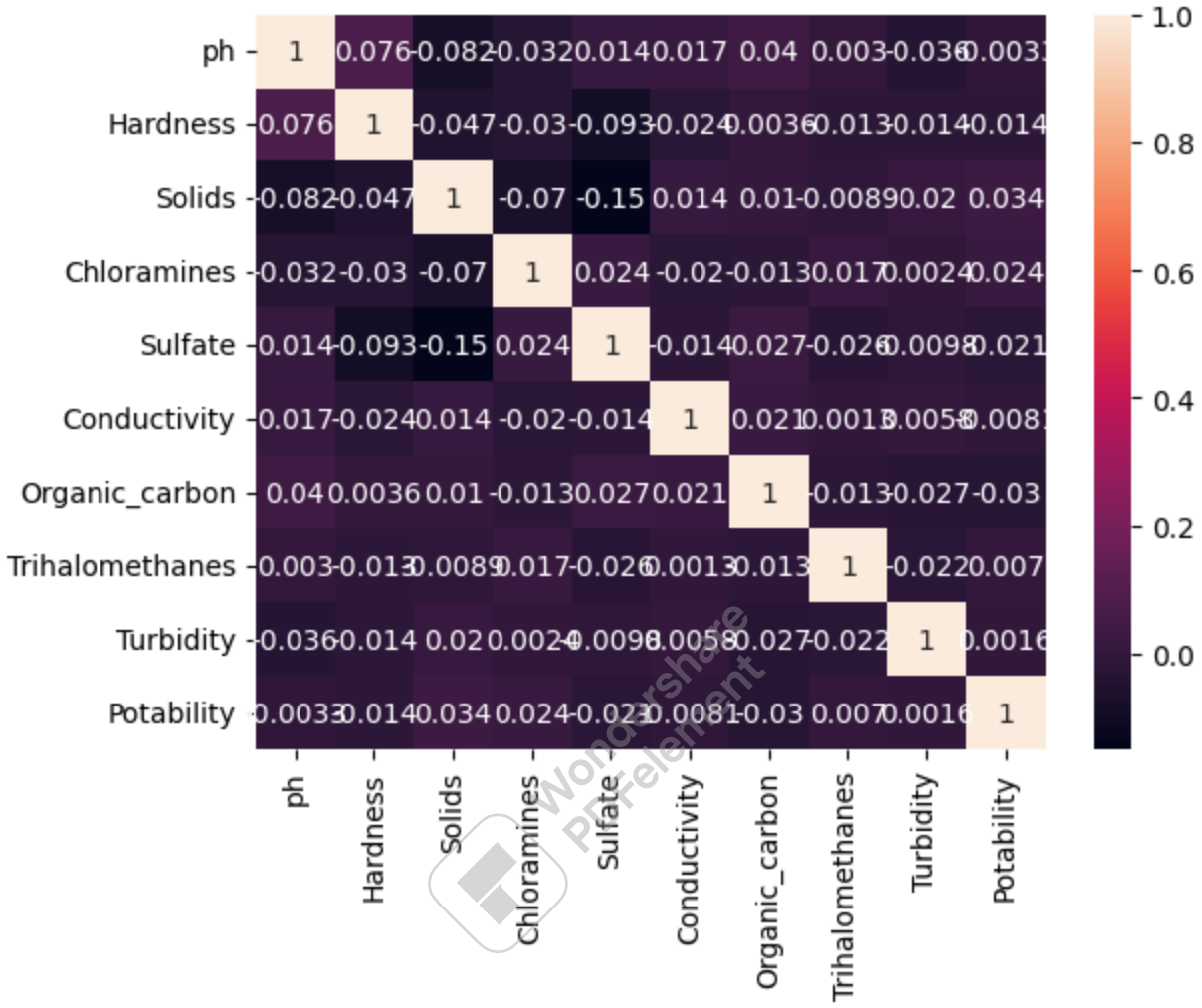
I was also able to show correlation among the features in the dataset.

1. **Bar Chart of values against potability .**



The above image shows the amount of potable water and of that which is not potable according to the dataset. The potable water is at 39.2% while that which is not potable is at 60.8%.

2. Correlation between the features in the dataset

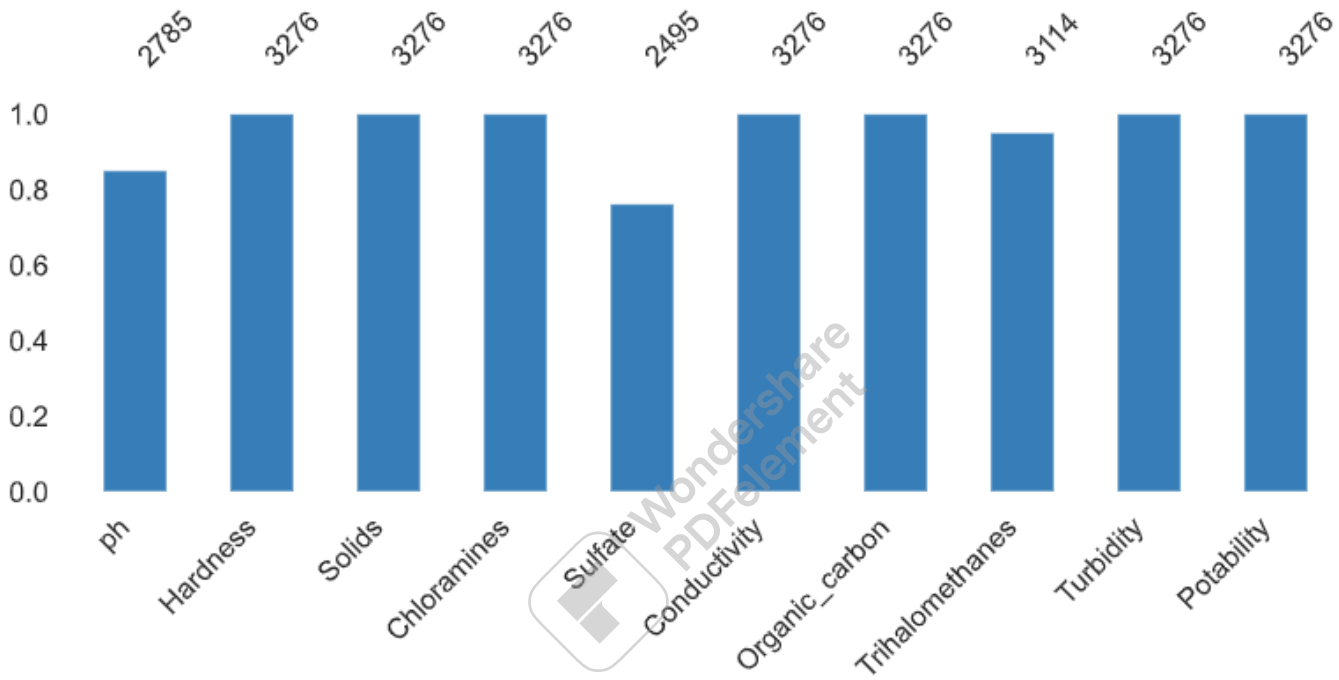


According to the above image, we can make conclusion that there is high correlation between Sulphates and Solids and low correlation between Conductivity and Trihalomethanes.

Data Cleaning

Data cleaning process

The initial data frame contained some empty cells. By use of the pandas .main() function I was able to replace the empty values with the mean. The empty values were 491 on ph, 781 on Sulphate and 162 on Trihalomethanes. As shown:

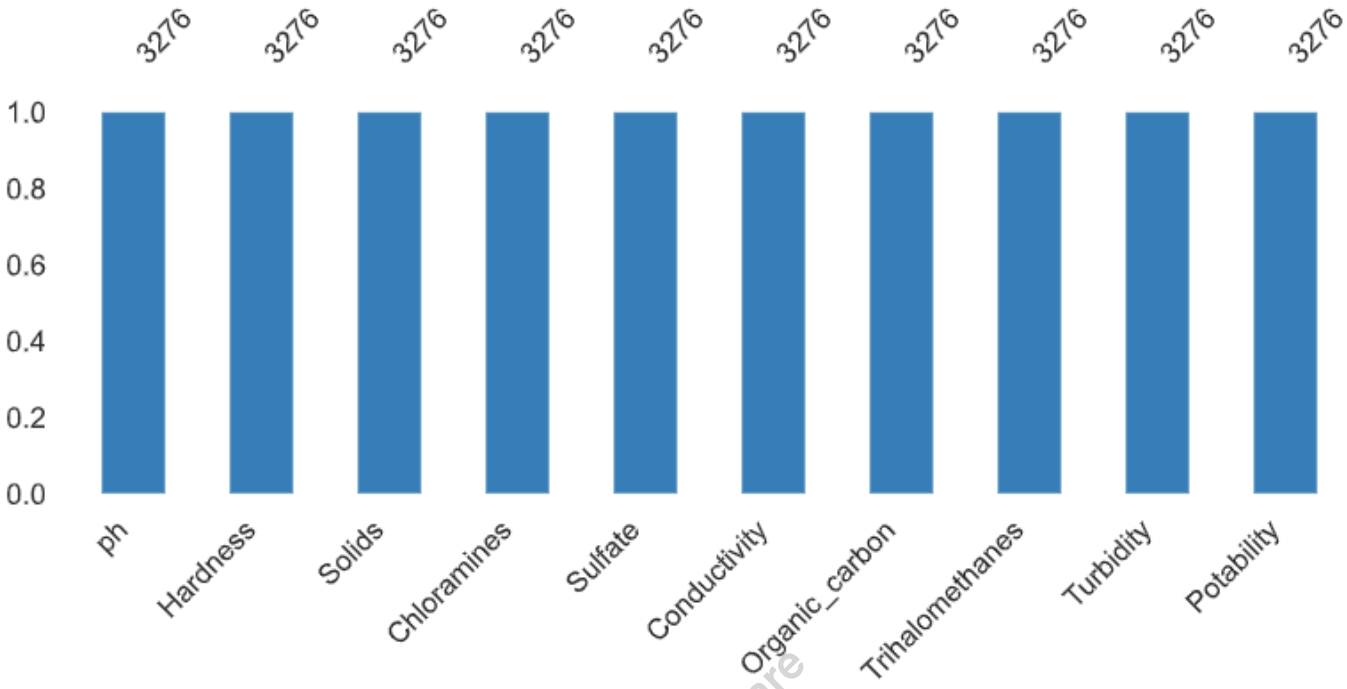


A simple visualization of nullity by column.

Data cleaning outcome

The data cleaning process resulted in a full dataset due to the replacement of cells having null values with the mean. The new data set became full, having all values . This should result to improved model performance in that the prediction model can be developed with a more accurate and reliable data

Filling the null values resulted in a more consistent dataset. The visualization of the new dataset is as shown:



A simple visualization of nullity by column.

Feature Engineering

In the feature engineering process, I used the first nine columns as the independent variables while the last column as the dependent variable which became the target variable.
The target variable was already in binary format , therefore I did not have to change that.
This helped in prediction of the water potability depending on the features , the water quality parameters.

Model Development

The model development approach

I trained the model from the training set. Here are the steps used to train each model:

a. **Decision Trees:** I trained the decision tree model on the training data using an appropriate criterion and hyper-parameter tuning. I evaluated the model's performance on the testing set using the accuracy metric.

Mean Absolute Error: 0.3567073170731707

Mean Squared Error: 0.3567073170731707

Accuracy: 0.6432926829268293

b. **Random forest regression**: I trained the random forest model on the training data using appropriate number of trees and hyper-parameter tuning. I evaluated the model's performance on the testing set using the accuracy metric.

Mean Absolute Error: 0.32164634146341464

Mean Squared Error: 0.32164634146341464

Accuracy: 0.6783536585365854

c. **KNeighborsClassifier:**

Mean Absolute Error: 0.37347560975609756

Mean Squared Error: 0.37347560975609756

Accuracy: 0.6265243902439024

d. **SVC**:

Accuracy: 0.6265243902439024

Mean Absolute Error: 0.37347560975609756

Mean Squared Error: 0.37347560975609756

The justification of the chosen model
The algorithm selected for use is the random forest classifier algorithm because it gives the highest accuracy among other algorithms used.

Model evaluation Metrics used
Justification for the choice of metric

Given the fact that the model developed is to be used to predict on the quality of water and whether it is potable or not based of some data given. The probability that the model would predict correctly with limited errors should be high to limit chances on making wrong decisions on water potability.

Model Deployment
The choice of model deployment platform was a web application : streamlit . This is because it's open-source and easier to implement and would take limited time to implement given the short duration of the entire project.

The process of deployment:

Install Streamlit if you have not yet installed it on your computer. I used 'pip', python package manager, by running the command 'pip install streamlit' in the terminal.

I then wrote the code for the streamlit app. This was done typically by importing necessary libraries, defining functions for data processing and then creating a user interface using streamlit's widgets.

I then tested our app code locally on my computer by running the 'streamlit run app.py' (where 'app.py' is the name of the app file). This launches a web server which I accessed in the web browser