WATER POTABILITY PREDICTION USING MACHINE LEARNING

A. Prem Chand

ITCS-5156 Fall 2022 – Lee

1 Introduction

In recent years, water quality has been threatened for various reasons. So, water calibre modelling and prediction have become very important to control water pollution. This proposed system presents a Classification model using different Machine learning algorithms for the purpose of analysing water quality data from different water samples. There are certain limits of pollutants that water species can tolerate. Exceeding the limits affects the existence of creatures and threatens their lives. Utmost ambient water bodies similar as gutters, rivers, lakes, and streams have specific quality norms that indicate their quality. It is very important to have a new approach to analyse, predict and classify the calibre of water, based on the dimensions and properties of the water. This proposed system focuses on various statistical methods including sampling techniques like Over Sampling, Outlier detection and removal, correlation techniques and advanced machine learning algorithms in supervised learning which are used to predict the water potability classification. The water quality is veritably important in ensuring citizens get to drink clean water. Application of random forest algorithm as a data mining method to predict clean water sample based on the water quality parameters can ease the work of the laboratory technologist by prognosticating which water samples should proceed to the next step of analysis.

Keywords: Over Sampling, Outlier Detection, Water Potability, Random Forest, Logistic Regression, Support Vector Machine, Decision Tree

2 Related Work

2.1 Efficient Water Quality Prediction Using Supervised Machine Learning

This research explores the methodologies that have been employed to help solve problems related to water quality. Typically, conventional lab analysis is used in research to aid in determining water quality, while some analyses employ machine learning methodologies to assist in finding a solution for the water quality problem. Gathered water samples from different areas are tested them against different parameters using a manual lab analysis and found a high presence. The estimated water quality in their work is based on only three parameters: turbidity, temperature and pH, which are tested according to World Health Organization (WHO) standards. Using only three parameters and comparing them to standardized values is quite a limitation when predicting water quality.

In this predicted the WQI using 16 water quality parameters and ANN with Bayesian regularization. And also predicted the chemical oxygen demand (COD) and the biochemical oxygen demand (BOD) using two conventional machine learning methodologies namely, ANN and multivariate linear regression. However, they ignored the major parameters associated with WQI during the learning

process and they did not use any standardized water quality index to evaluate their predictions. They used 23 parameters to predict the WQI, which turns out to be quite expensive if one is to use it for an IoT system, given the prices of the sensors. They used 10 parameters to predict the DO, which again defeats the purpose if it has to be used for a real-time WQI estimation with an IoT system. Most of the research either employed manual lab analysis, not estimating the water quality index standard, or used too many parameters to be efficient enough.

2.2 Predictive Analysis of Water Quality Parameters using Deep Learning

In this paper they have used the Regression approach and Multi-Layer Perceptron approach to predict the water quality.

Regression approach: -

Linear regression is supervised leaning based approach in Linear regression calculates weights from the training dataset. Linear regression is well found venerable regression analysis the regression line of the predicted value of Y from X passes. Linear regression is a classical statistical method they have trained the first layer as an RBM that models the raw input x = h (0) as its visible layer. In step 2 they have used the first layer to obtain a representation of the input that will be used as data for the second layer. In next step they trained the second layer as an RBM, taking the transformed data (samples or mean activations) as training supervised predictions which comes under regression analysis. Regression analysis is this method, he weights are calculated from training data.

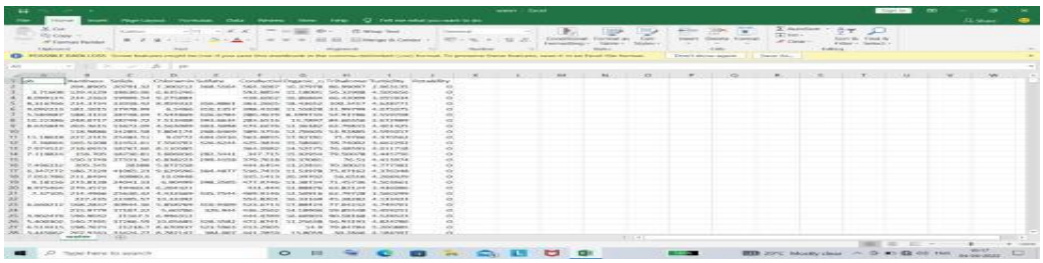
Multi-Layer Perceptron approach: -

Single-layer neural networks can only proceed linear data. Therefore, to overcome this limitation, Multiple layers are used. The perceptron is neural network that consists of input layer and hidden layer and exit layer. Each neuron is connected and weight to train the dataset using backpropagation uses a supervised learning approach. MLPs are the most generally used type of nonlinear neural network. Can Adapt and maintain change according to the environment Robustness. MLP uses a backpropagation algorithm for training the algorithm. Mainly used to overcome overfitting and underfitting limitations in regression analysis.

Methodology: --

Module-1: Data Collection.

Step-1:-Collection of data From Kaggle and retrieving the data in Jupyter note book using Pandas.



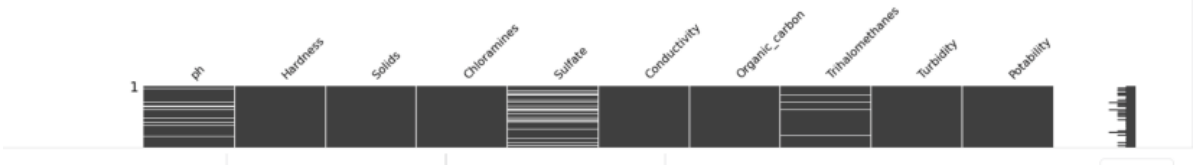
Step 2: Data gathering.

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

```
In [1]:
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.tree import DecisionTreeClassifier
from sklearn.metrics import accuracy_score,classification_report,confusion_matrix
import pickle
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestClassifier
import missingno as msno
import warnings
from imblearn.under_sampling import RandomUnderSampler
warnings.filterwarnings('ignore')

In [3]:
df=pd.read_csv(r'C:\Users\Lenovo\Desktop\water.csv')
msno.matrix(df)

Out[3]:
<AxesSubplot:>
```



MODULE-2: Performing Exploratory Data analysis.

Step 1: In ED analysis, firstly check the shape of the data set. Further, check that there are Null values present or not and we can find in the below image that ph., Sulfate, Trihalomethanes contain NULL values. Further, check the information of the data set.

Exploratory Data Analysis ¶

```
In [3]: df.shape
Out[3]: (3276, 10)

In [4]: df.isnull().sum()
Out[4]: ph                 491
        Hardness             0
        Solids               0
        Chloramines          0
        Sulfate            781
        Conductivity         0
        Organic_carbon       0
        Trihalomethanes    162
        Turbidity            0
        Potability           0
        dtype: int64

In [5]: df.info()
        <class 'pandas.core.frame.DataFrame'>
        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)
        memory usage: 256.1 KB
```

Step-2: Then describe the dataset which shows the minimum value, maximum value, mean value, count, standard deviation, etc

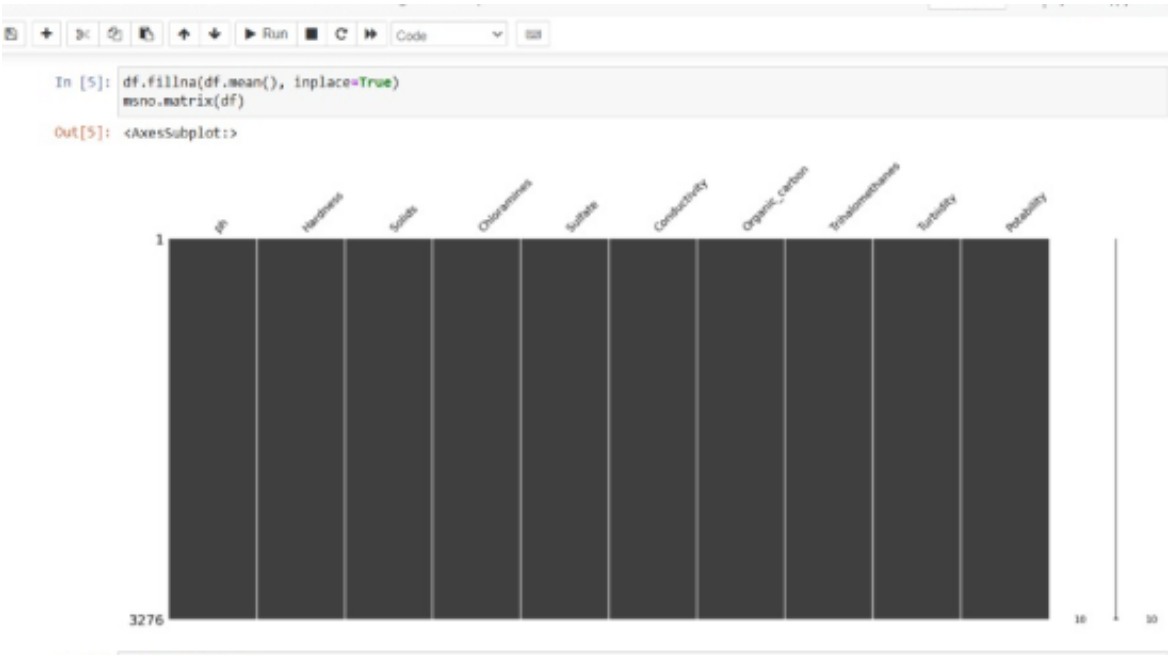
```
 5   Conductivity     3276 non-null   float64
 6   Organic_carbon   3276 non-null   float64
 7   Trihalomethanes  3114 non-null   float64
 8   Turbidity        3276 non-null   float64
 9   Potability       3276 non-null   int64
dtypes: float64(9), int64(1)
memory usage: 256.1 KB
```

```
In [6]: df.describe()
```

Out[6]:

	ph	Hardness	Solids	Chloramines	Sulfate	Conductivity	Organic_carbon	Trihalomethanes	Turbidity	Potability
count	2785.000000	3276.000000	3276.000000	3276.000000	2495.000000	3276.000000	3276.000000	3114.000000	3276.000000	3276.000000
mean	7.080795	196.369496	22014.092526	7.122277	333.775777	426.205111	14.284970	66.396293	3.966786	0.390110
std	1.594320	32.879761	8768.570828	1.583085	41.416840	80.824064	3.308162	16.175008	0.780382	0.487849
min	0.000000	47.432000	320.942611	0.352000	129.000000	181.483754	2.200000	0.738000	1.450000	0.000000
25%	6.093092	176.850538	15666.690297	6.127421	307.699498	365.734414	12.065801	55.844536	3.439711	0.000000
50%	7.036752	196.967627	20927.833607	7.130299	333.073546	421.884968	14.218338	66.622485	3.955028	0.000000
75%	8.062066	216.667456	27332.762127	8.114887	359.950170	481.792304	16.557652	77.337473	4.500320	1.000000
max	14.000000	323.124000	61227.196008	13.127000	481.030642	753.342620	28.300000	124.000000	6.739000	1.000000

Step-3: Then finally we handle the missing values. We filled the missing values in our features using a mean value of each feature which means we filled the mean value to handle missing data. Finally again check that there are null values present or not.

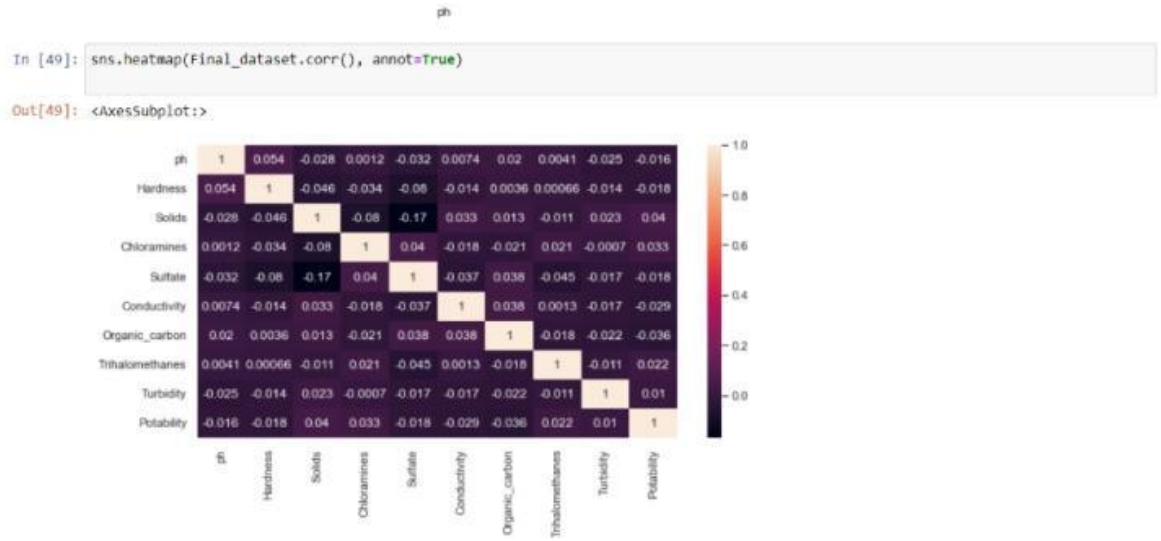


MODULE 3:

Applying Statistical methods like Over Sampling, Outlier Detection and removal.

As the data is imbalanced, we performed Over Sampling to equalize the data in respective of our dependent variable. Over sampling is a technique to balance Uneven datasets by keeping all of the data in the majority class and increasing the size of the minority class. It is a significant technique data scientist can use to extract more accurate information from originally imbalanced datasets. Having an outlier is not good to train the data.so using Empirical Formula technique we detected and removed the Outlier. It is a key tool in safeguarding data quality, as anomalous data and errors can be removed and analyzed once identified.

Outlier detection is a significant part of each stage of the machine learning process. The below figure is the correlation matrix of our data features.



Module 4:-

Finally, it's time to prepare the data set. Divide the data into the separate independent and dependent features. All the features are independent except Potability because Potability is our dependent feature. Now, Split the data set into the training and testing by using the train_test_split function which returns four data sets.

```
In [126…
X=Final_dataset.drop(['Potability'],axis=1)
Y=Final_dataset['Potability']
X_train, X_test, Y_train, Y_test = train_test_split(X,Y, test_size= 0.1)
model=RandomForestClassifier(n_estimators=92)
model.fit(X_train,Y_train)
prediction=model.predict(X_test)
```

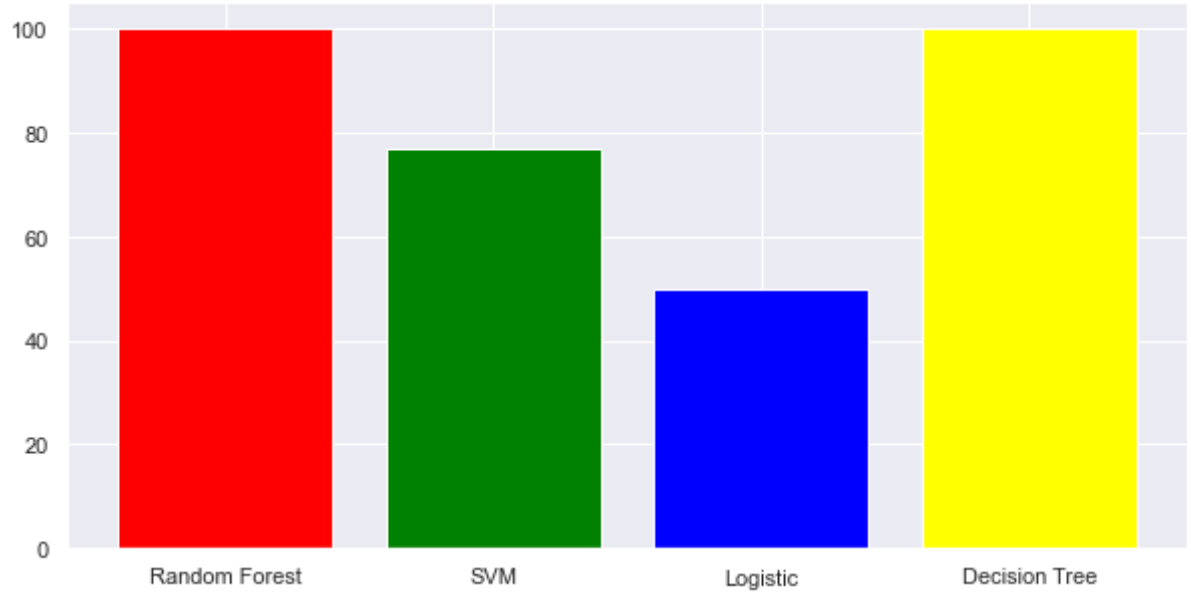
Likewise, define the different classifier models and train the model using the data set (X_train, Y_train). And eventually it's time to evaluate the models by using the accuracy score, confusion matrix and classification report. The Evaluation techniques take two parameters. one is called the actual data and the other one is a predicted data. And we can find that overall accuracy

Experiments: -

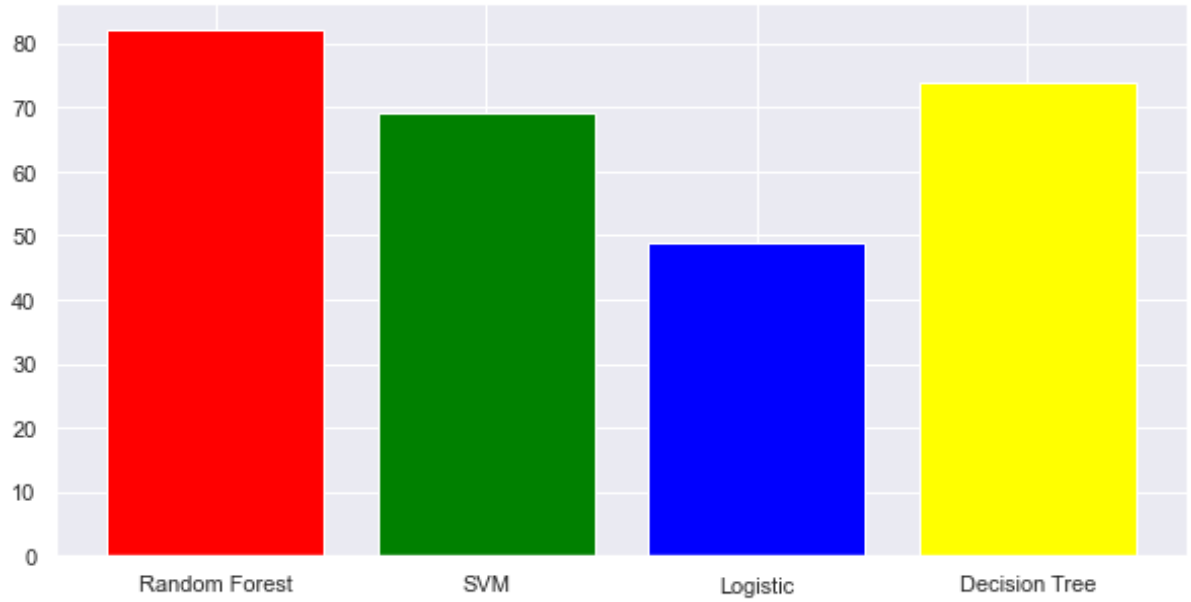
Data Flow Diagram:-

Description: - Description As show in the above data flow diagram first step includes the collecting of data from the Kaggle. In the data there exists missing values and outliers which effects our model accuracy. So, it is essential to remove all that Outliers using an Empirical Formula. our data size is decreased. So, with help of a statistical method that is Oversampling, normalizing the data size is productive method. After that training the model using some machine algorithms to get the effective results.

Training Accuracy: -



Testing Accuracy: -



Classification Report Of Random Forest :-

```
print(classification_report(Y_test,rf_pred_test))
```

	precision	recall	f1-score	support
0	0.84	0.81	0.82	125
1	0.80	0.83	0.82	114
accuracy			0.82	239
macro avg	0.82	0.82	0.82	239
weighted avg	0.82	0.82	0.82	239

Conclusion: -

The proposed model predicts whether the water is safe to drink or not using few parameters such as Ph value, conductivity, hardness, etc., The data were collected and trained using a machine learning algorithm, and the predicted data has been tested. Among the efficient algorithms used for water calibre prediction we found that random forest classification model is best with a good overall accuracy. Access to safe drinking water is essential factor to health, a basic human right and an effective element for health protection. It is important as a health factor and development issue at a national, regional and local level. In some regions, it has been found that investments in water supply and sanitation can yield a net economic benefit successfully. since the reductions in adverse health effects and health care costs outweigh the costs of undertaking the interventions. Therefore, the proposed model of using different algorithms together to predict water quality data has yield to better outcome of finding an efficient working model among all.

References & Citation: -

[1] Ahmed U, Mumtaz R, Anwar H, Shah AA, Irfan R, Garc´ıa-Nieto J. Efficient Water Quality Prediction Using Supervised Machine Learning. Water. 2019; 11(11):2210. https://doi.org/10.3390/w11112210

[2] Amir Hamzeh Haghiabi, Ali Heidar Nasrolahi, Abbas Parsaie; Water quality prediction using machine learning methods. Water Quality Research Journal 1 February 2018; 53(1):3–13.doi:https://doi.org/10.2166/wqrj.2018.025

[3] A. Solanki, H. Agrawal, and K. Khare, "Predictive analysis of water quality parameters using deep learning," International Journal of Computers and Applications, vol. 125, no. 9, pp. 29–34, 2015.

[4] H. Liao and W. Sun, "Forecasting and evaluating water quality of Chao Lake based on an improved decision tree method," Procedia Environmental Sciences, vol. 2, pp. 970–979, 2010.

[5] J. Liu, C. Yu, Z. Hu et al., "Accurate prediction scheme of water quality in smart mariculture with deep Bi-S-SRU learning network," IEEE Access, vol. 8, pp. 24784–24798, 2020.

[6] M. M. S. Cabral Pinto, C. M. Ordens, M. T. Condesso de Melo et al., "An interdisciplinary approach to evaluate human health risks due to long-term exposure to contacontaminated groundwater near a chemical complex," Exposure and Health, vol. 12, no. 2, pp. 199–214, 2020.

[7] P. Zeilhofer, L. V. A. C. Zeilhofer, E. L. Hardoim, Z. M.. Lima, and C. S. Oliveira, "GIS applications for mapping and spatial modeling of urban-use water quality: a case study in District of Cuiaba, Mato Grosso, Brazil," Cadernos ´ de Saude P ´ ublica, vol. 23, no. 4, pp. 875–884, 2007. ´

[8] S. Jaloree, A. Rajput, and G. Sanjeev, "Decision tree approach to build a model for water quality," Binary Journal of Data Mining Networking, vol. 4, pp. 25–28, 2014.

[9] S., Yogalakshmi A., Mahalakshmi. (2021). Efficient Water Quality Prediction for Indian Rivers Using Machine Learning. Asian Journal of Applied Science and Technology. 05. 100-109. 10.38177/ajast.2021.5111.

[10] Theyazn H. H Aldhyani, Mohammed Al-Yaari, Hasan Alkahtani, Mashael Maashi, "Water Quality Prediction Using Artificial Intelligence Algorithms", Applied Bionics and Biomechanics, vol. 2020, Article ID 6659314, 12 pages, 2020. https://doi.org/10.1155/2020/6659314