INDIAN INSTITUTE OF INFORMATION
TECHNOLOGY ALLAHABAD

Prayagraj, UP 211015, India

Data Mining Project Report On

GANGA WATER QUALITY ASSESSMENT USING DEMPSTER SHAFER THEORY

July, 2020

Submitted By:

Shivani Chauhan

B. Tech (CSE), 2nd Year Student

GLA University, Mathura

DECLARATION

I, the undersigned, solemnly declare that this project report is based on my own work carried out during the course of my study under the supervision of **Dr. Manish Kumar**, Associate Professor, IIITA.

I assert the statements made and conclusions drawn are an outcome of my research work. I further certify that

I. The work contained in the report is original and has been done by me under the general supervision of my supervisor.
II. I have followed the guidelines provided by the university in writing the report.
III. Whenever I have used materials (data, theoretical analysis, and text) from other sources, I have given due credit to them in the text of the report and have given their details in the references.

Shivani Chauhan

B. Tech (CSE), 2nd Year,

GLA University, Mathura

CONTENTS

INTRODUCTION

India is blessed with rich water resources. The Indian riverine system extends over approximately 45,000 km. Among the 12 major river basins, the Ganga river basin is the largest. It extends over 11 states of India, emerging from the Himalayan Mountains and ending at the Bay of Bengal.

In this work, we will study the water samples from the holy river of Ganga which flows in the city of Prayagraj, situated at the confluence of the Rivers Ganga, Yamuna and Saraswati.

The determination of water quality has been an important environmental problem. This study proposes an idea to develop a machine learning model which can accurately assess the quality of river water using data mining techniques.

The parameters of the time-series dataset collected from the river Ganga using daily measure of sensors include: Electrical Conductivity, Dissolved Oxygen (DO), Oxidation Reduction Potential (ORP), pH and Temperature of water.

The conductivity of water is measured by its capability to pass electrical flow. Conductivity is quite an essential parameter to assess water quality as it is dependent on the concentration of conductive ions in the water.

The amount of oxygen dissolved in the water can tell a lot about its quality. As the dissolved oxygen (DO) increases, the temperature of water decreases. It is usually measured in both milligrams per litre and percentage saturation.

Another quality parameter is Oxidation Reduction Potential, which is also known as REDOX. It is a measurement that reflects the ability of a molecule to oxidise or reduce another molecule.

The relative amount of free hydrogen and hydroxyl ions determines the pH of water. In other words, it is a measure of how acidic or basic is the given water sample. pH ranges from 0-14, with 7 being neutral.

Temperature is yet another important factor to consider when assessing water quality. Temperature can alter the physical and chemical properties of water.

Using these physicochemical parameters, a machine learning model can be made to effectively assess the quality of Ganga river water.

LITERATURE SURVEY

Several studies have been conducted to assess the quality and health status of the Ganga river basin at various stations throughout India. A methodology has been developed to integrate the water quality index (WQI) with the geographic information system (GIS) [1]. The determination of WQI is done using the overall index of pollution.

In 2017, a study was carried out to assess the Ganga water quality at Haridwar, using various physicochemical parameters such as turbidity, total solids, hardness, free $CO_{2,}$ nitrate, nitrite, COD, and phosphate [2]. The values of these parameters are compared with WHO and ISI standards.

In Rishikesh, samples of Ganga water were collected to assess its suitability for drinking, irrigation and industrial usages using various indices & were classified based on the values obtained [3].

Another study was conducted in Kanpur [4] using Pearson's correlation coefficient value which is determined using correlation matrix to identify highly correlated and inter-related quality parameters.

It has been found that Allahabad, present day – Prayagraj, is the most polluted station of Ganga River Basin [1].

A study of satellite data was done at Allahabad using Multiple Linear Regression [5], which showed that the parameters were significantly correlated with the radiance values of ETM+ image.

Ganga River has been a major recipient of industrial wastes. The occurrence of five heavy metals for sediment [6] from Ganga river was studied to assess the metal contamination in the sediment.

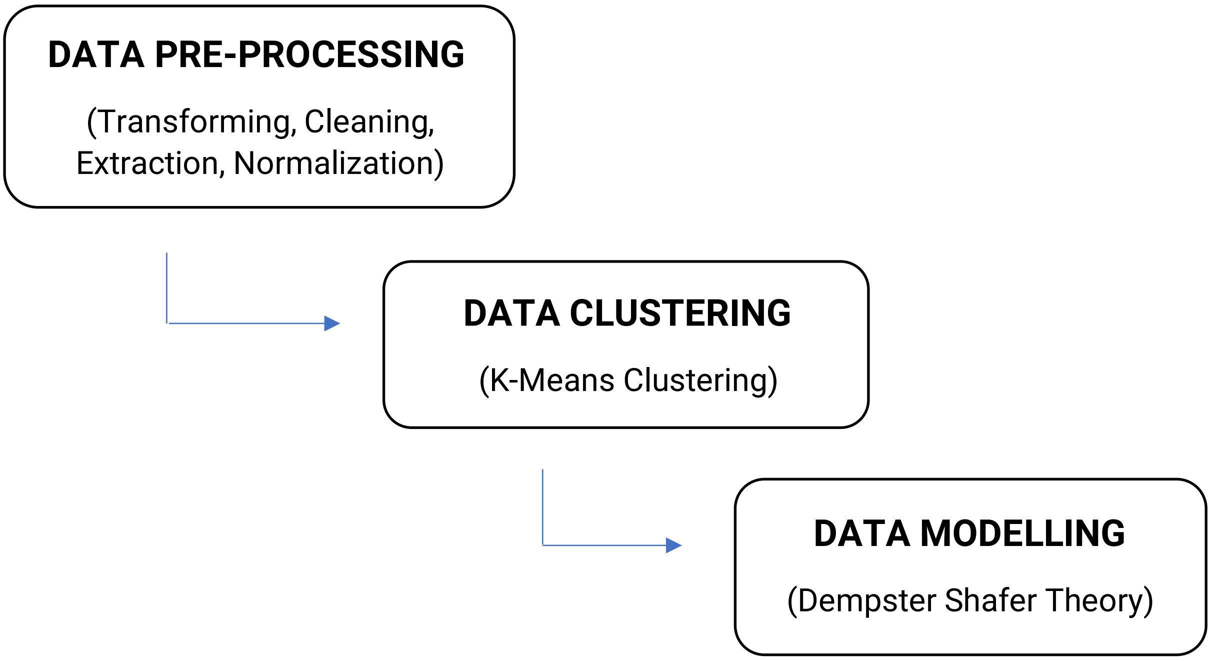
PROBLEM STATEMENT

Complete survey surmised that the water quality is measured from spatial-temporal changes of water quality, statistical analysis of quality parameter and satellite data of the Ganga River basin.

Also, the traditional water quality monitoring approaches give precise measurements, but their costs are expensive and they are time-consuming. Moreover, these methods did not complete regional needs (Yao et al. 2010). Also, it requires large travelling and laboratory expenses, especially for a large area.

So, it is very challenging to analyse water quality in real-time. For solving these problems, we are trying to make a water quality assessment model based on the intrinsic nature of the collected IoT data.

METHODOLOGY



DATA PRE-PROCESSING

(Transforming, Cleaning, Extraction, Normalization)

DATA CLUSTERING

(K-Means Clustering)

DATA MODELLING

(Dempster Shafer Theory)

DATA PRE-PROCESSING:

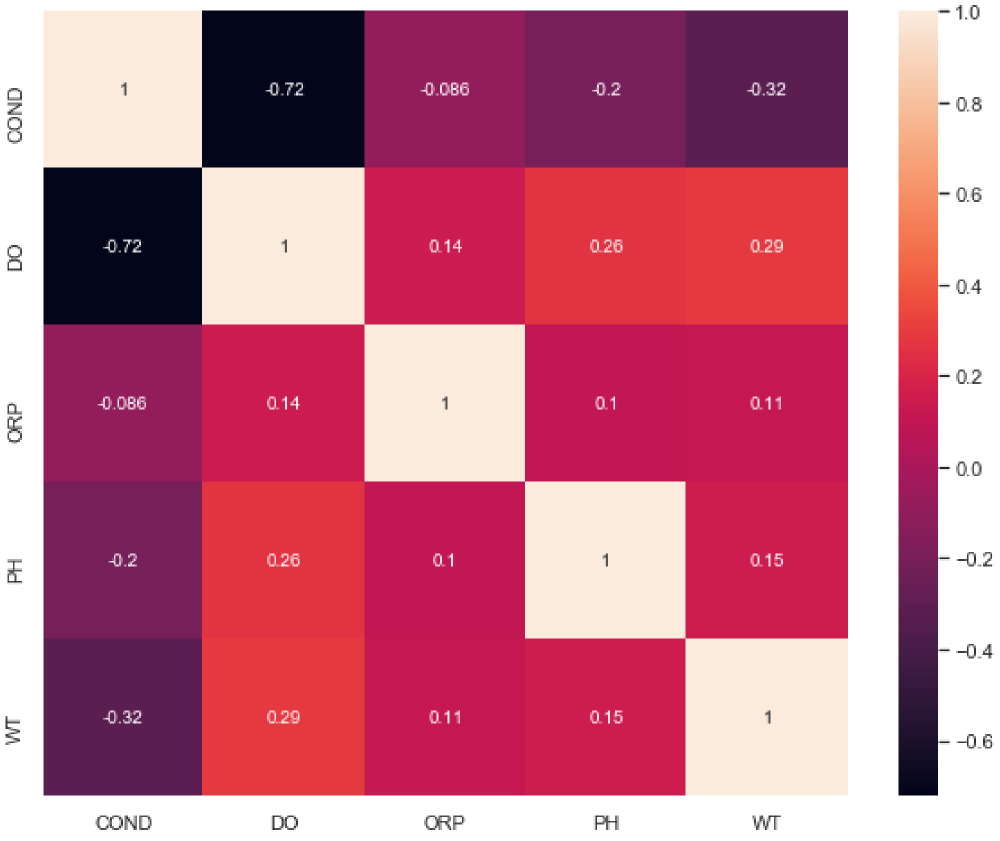
Since the dataset collected is noisy in nature, it is pre-processed so that it can be easily modelled. A number of pre-processing steps were applied to the dataset. The transformations performed on the data using NumPy and Pandas libraries include:

- Changing the data types of Date and Time attributes
- Dropping unnecessary attributes
- Sorting the time-series data based on the Date attribute
- Transforming the resulting data using pivot tables so that each sensor becomes an attribute

The sensing data collected includes data of Ganga as well as Sangam river. Based on the sensing details, the details of Ganga river have been extracted and the negative values are replaced by NaN values, to handle the missing data. The missing data in each attribute is then replaced by the mean of the last five values of the particular attribute using the forward fill technique.

The details of Ganga river have been extracted into samples for further processing. Negative values are dropped using the python libraries. As per the report on sensing details, the data is extracted into 21 buckets using the Pandas library.

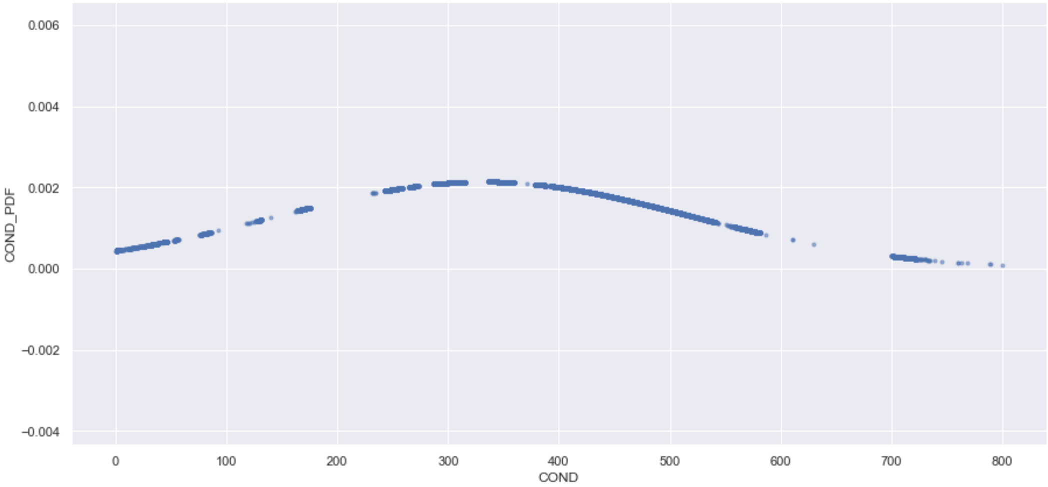
The data is then visualized in a Heatmap by plotting the correlation between the quality parameters used as follows:

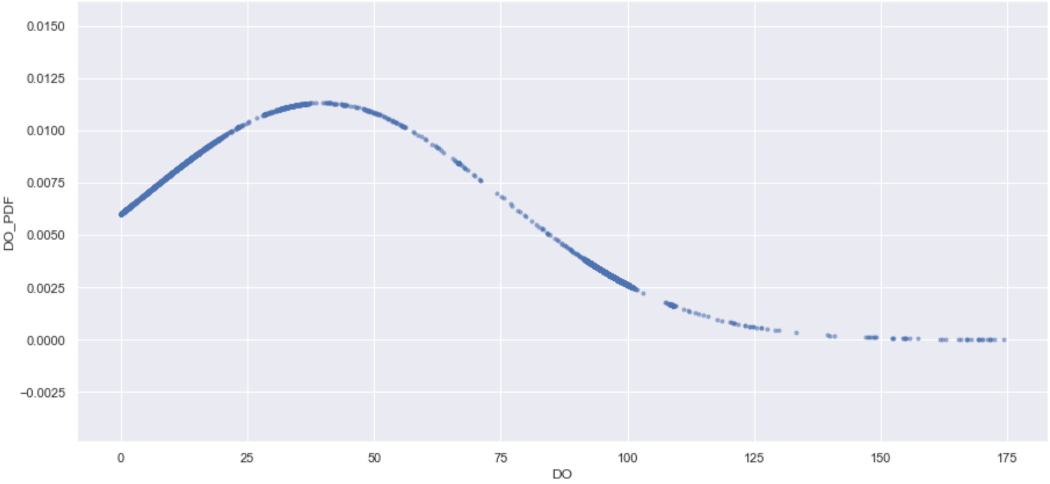


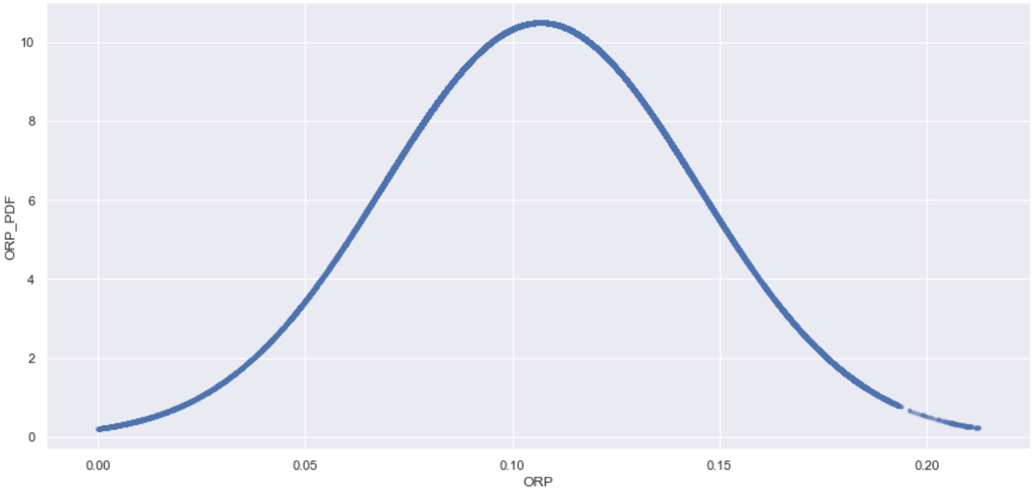
In order to understand the underlying behaviour of our attributes, we need to apply the Gaussian distribution formula to check whether the data is normally distributed. On plotting the probability distribution if a bell-shaped curve is formed and the mean, mode, and median are equal then the variable is normally distributed. Normal distribution is dependent on two parameters which are mean and the standard deviation. If the data exhibits normal distribution, it is feasible to be forecasted with higher accuracy.

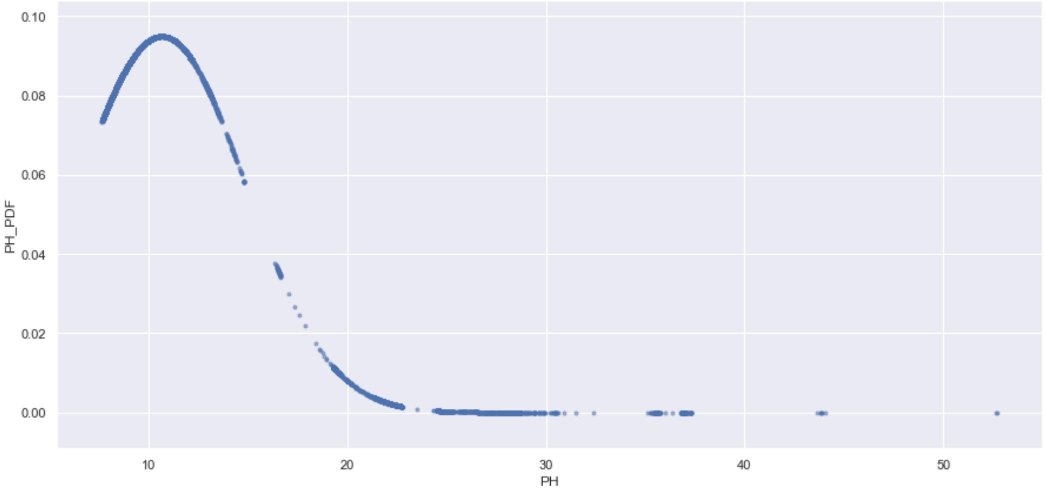
$$y = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

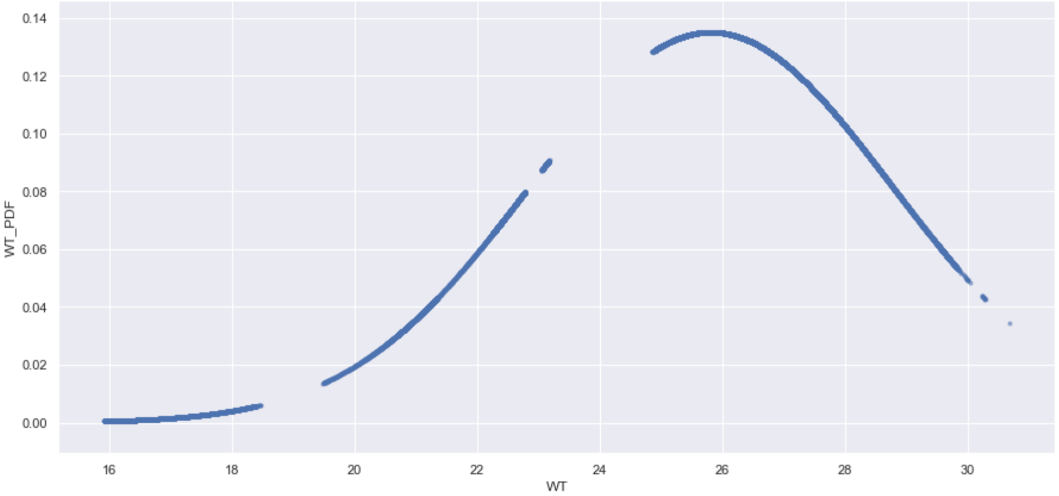
$$\begin{aligned} \mu &= \text{ Mean} \\ \sigma &= \text{ Standard Deviation} \end{aligned}$$







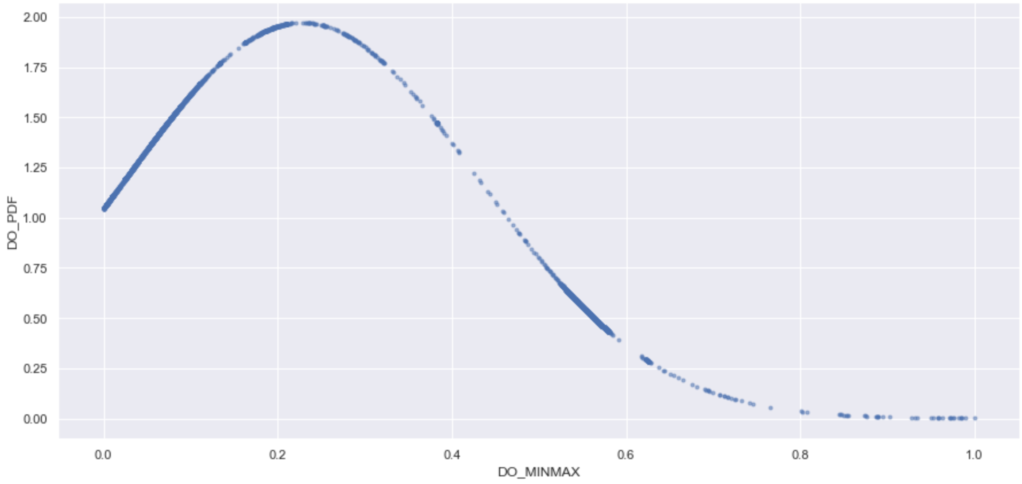


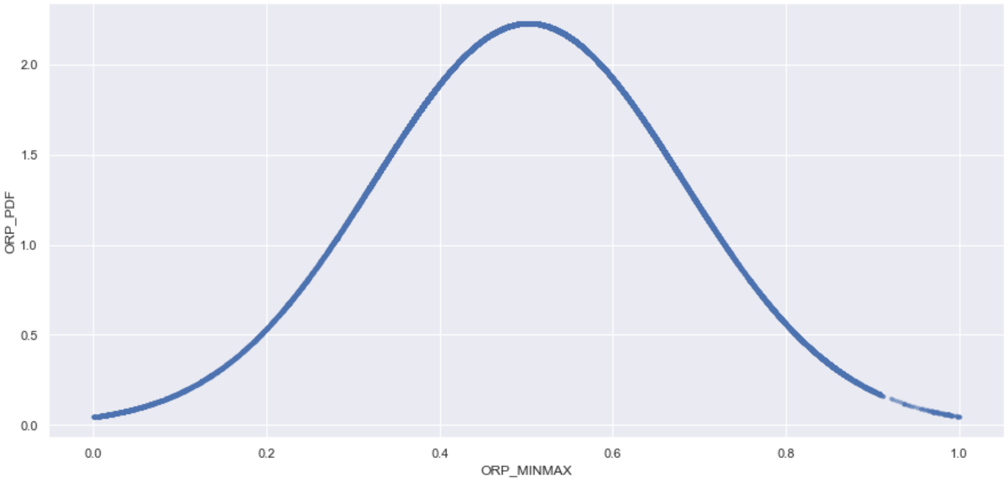


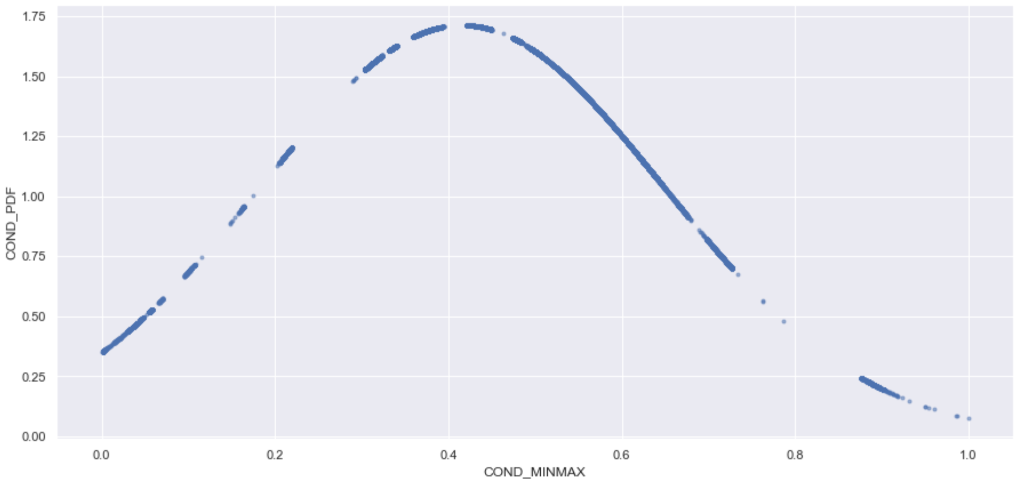
From the plots, we find that the collected data for oxidation-reduction potential of Ganga river water does not follow normal distribution as its mean is greater than 1.

We apply the Central Limit Theorem, which states that as the sample size is increased, the sampling distribution of the mean reaches a normal distribution. The random samples from the extracted data, each of sample size greater than equal to 30, are then checked for normal distribution using the Gaussian distribution formula. We find that the bell-shaped curve for oxidation-reduction potential data has a high mean while others have mean less than 1.

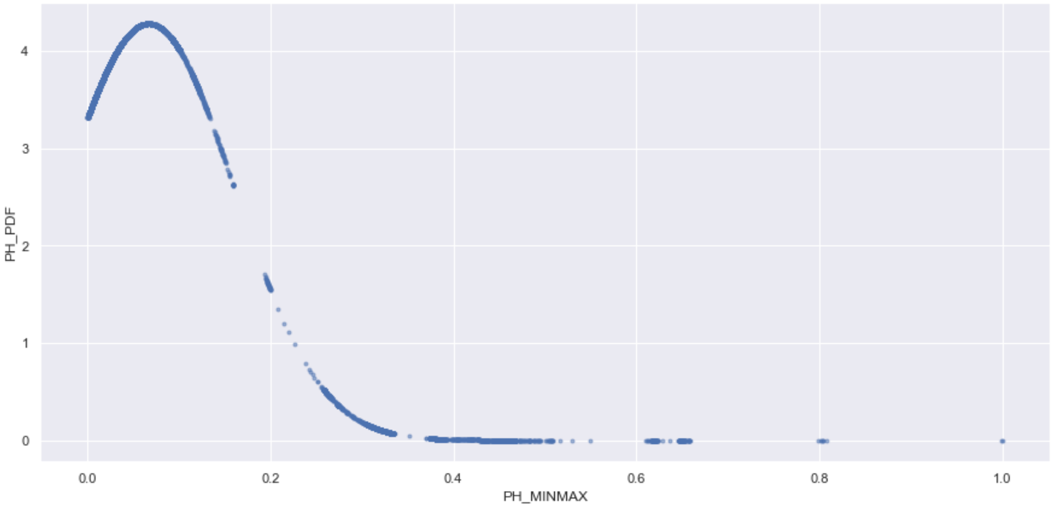
We normalize the data as variables that are measured at different scales do not contribute equally to the model fitting. Thus, to deal with this potential problem feature-wise normalization is usually used before model fitting. So, the data is normalized using MinMaxScaler() from scikit-learn library. The visualizations done are shown below:



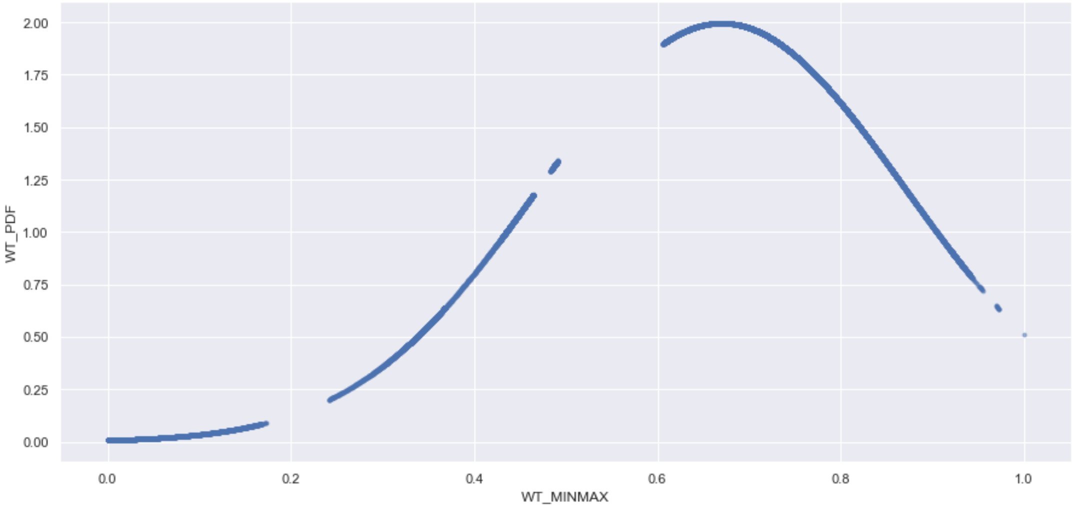




COND_PDF
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
0.0
0.2
0.4
0.6
0.8
1.0
COND_MINMAX

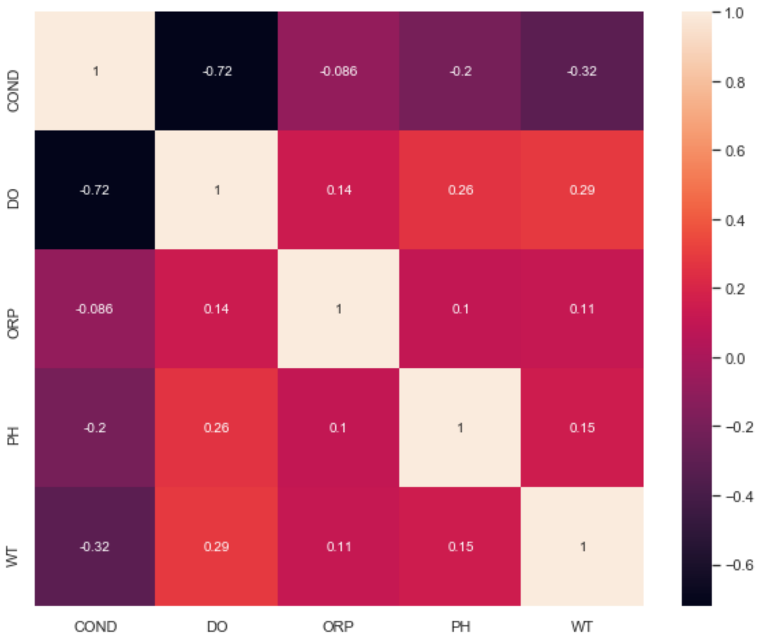


PH_PDF
4
3
2
1
0
0.0
0.2
0.4
0.6
0.8
1.0
PH_MINMAX



WT_PDF
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
0.0
0.2
0.4
0.6
0.8
1.0
WT_MINMAX

The heatmap obtained from plotting the normalized data using correlation of quality parameters is:

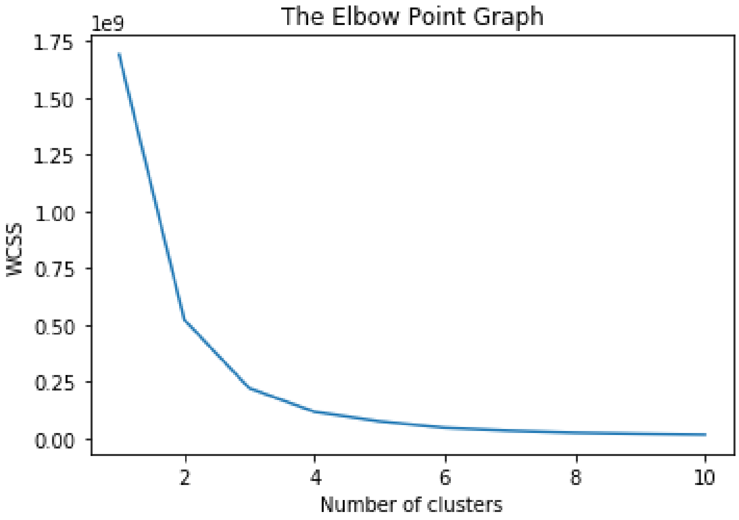


DATA CLUSTERING:

Before fitting the model, we need to classify the data using K-Means Clustering, which is an extensively used technique for data clustering.

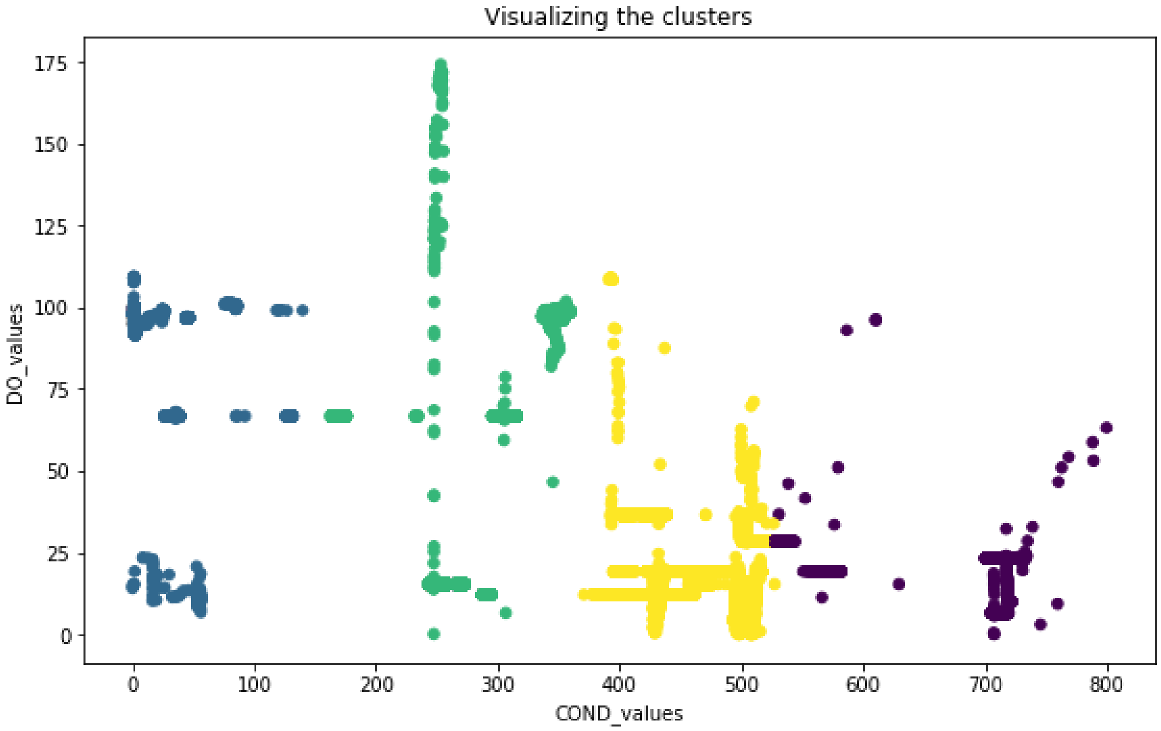
The K-means algorithm begins with a first group of randomly selected centroids. This cluster is used as the beginning points for every cluster. Then iterative calculations are done to optimize the positions of the centroids. It stops when there is no change in the values of centroids.

The elbow method computes an average score for all clusters. It runs k-means clustering for a range of values for k (say from 1-10).



From the graph, we can see, the appropriate value for k is 4.

So, on plotting the curve for k=4 clusters, we get the following:



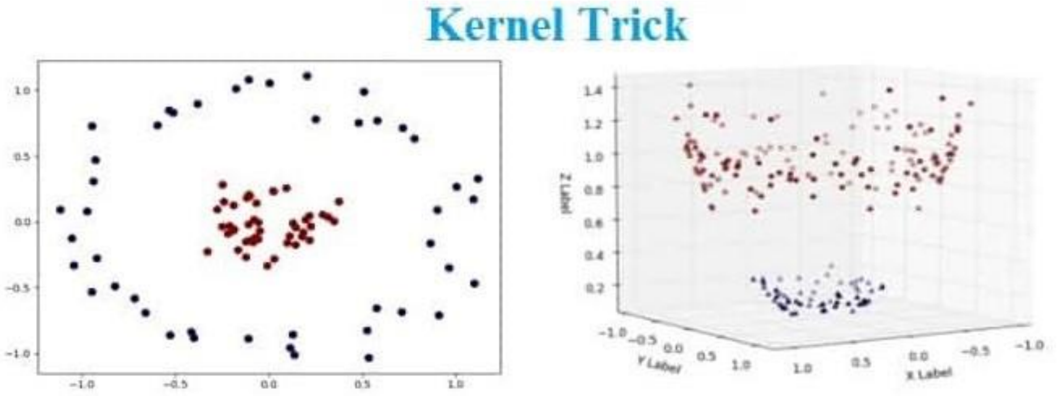
From the above plot for conductivity and dissolved oxygen attributes, we can see a few points of one cluster overlaps with the points of another cluster. To get better results, we apply SVM Kernel Trick to classify & model the data and visualize them by plotting their Receiver Operating Characteristic (ROC) Curves.

DATA MODELLING:

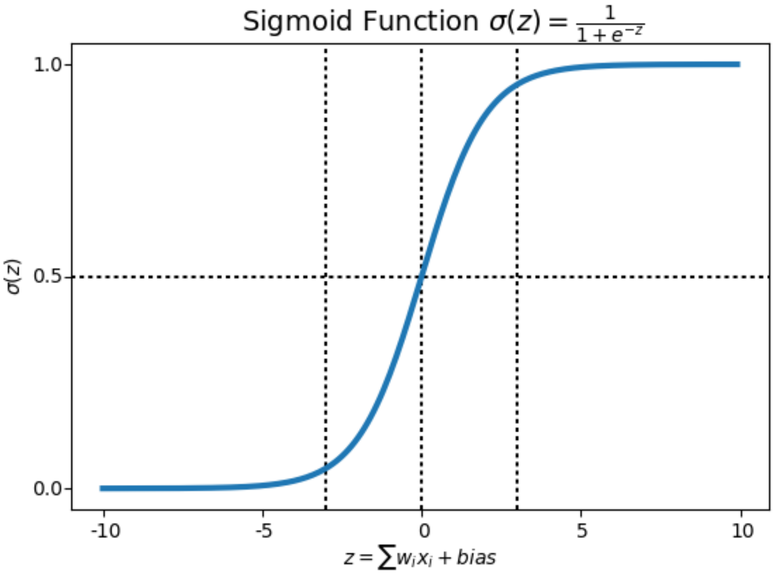
The data is modelled using the Dempster Shafer Theory with the help of machine learning models like SVM, Logistic Regression, Naïve Bayes and Decision Tree.

Support Vector Machine (SVM) is a supervised machine learning algorithm which can be used for both classification and regression problems. A line or a hyperplane in a high or infinite dimensional space is constructed.

A Kernel Trick is a simple method where a non-linear data is projected onto a higher dimension space so as to make it easier to classify the data where it could be linearly divided by a plane.



Logistic regression is a supervised learning classification algorithm used to predict the probability of a target variable. Based on the number of categories, it can be classified as: Binomial, Multinomial and Ordinal. Logistic regression is a powerful algorithm that utilizes a sigmoid function or a logistic function.



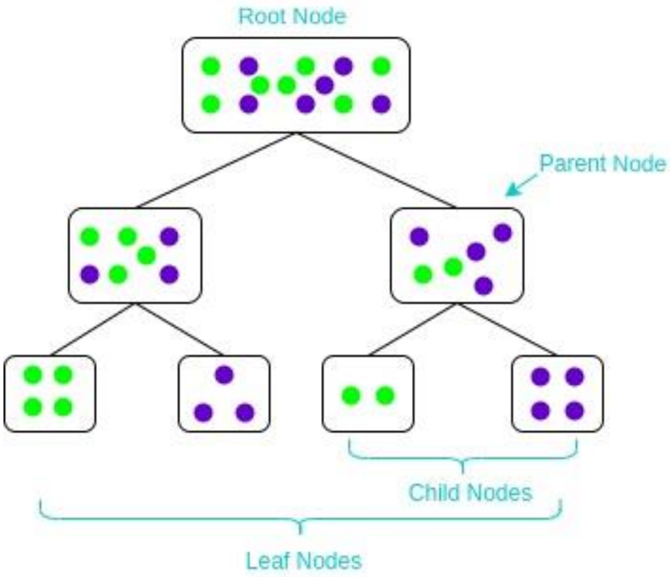
Naive Bayes Classifier is a supervised machine learning algorithm that uses the Bayes' Theorem, which relies on the naive assumption that input variables are independent of each other.

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

Gaussian Naive Bayes supports continuous data. When working with continuous data, an assumption often taken is that the continuous values associated with each class are distributed according to a normal (or Gaussian) distribution. The likelihood of the features is assumed to be –

$$P(x_i \mid y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{(x_i - \mu_y)^2}{2\sigma_y^2}\right)$$

Decision tree is a type of supervised learning algorithm that is mostly used in classification problems. A decision tree makes decisions by splitting nodes into sub-nodes. This process is performed multiple times during the training process until only homogenous nodes are left.



The Dempster Shafer Theory, also known as the theory of belief functions, was designed to mathematically model and validate the uncertainty involved in statistical inferences. It combines a set of representations and model data when there is a lack of information.

DST is mostly known to represent uncertainties or imprecision in a hypothesis. The hypotheses characterize all the possible states of the system. These hypotheses are assigned a probability mass assignment (PMA) which when combined leads to a decision [7].

The process of forming mass assignment function and combining the same is thus crucial for accurate prediction. The high-level features are converted into Dempster-Shafer mass functions by aggregating them using Dempster's rule of combination.

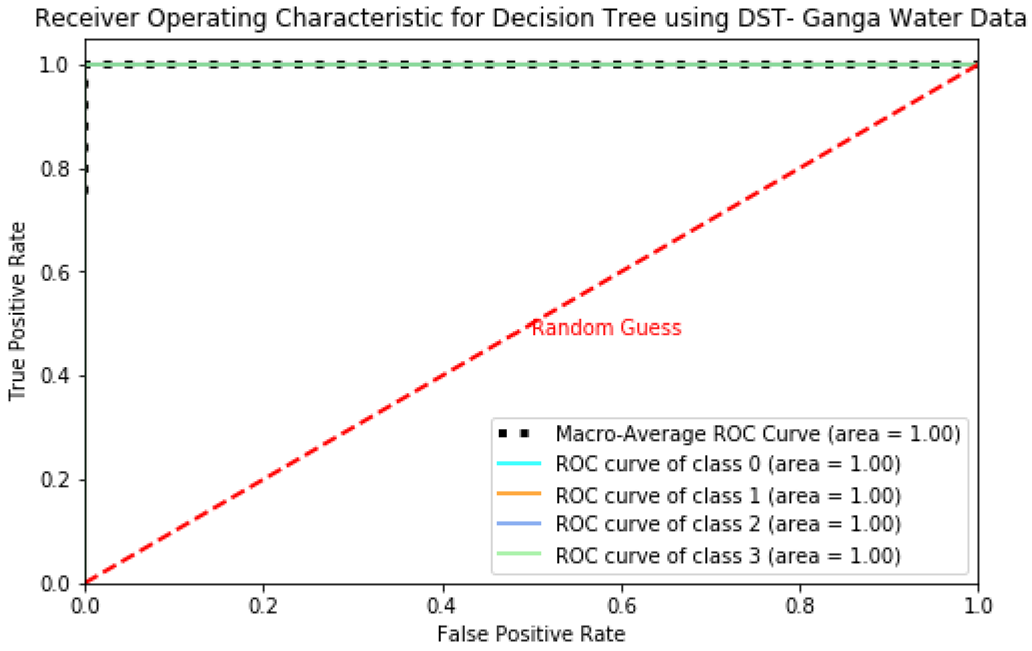
$$m_{1,2}(A) = (m_1 \oplus m_2)(A) = \frac{1}{1-K} \sum_{B \cap C = A \neq \emptyset} m_1(B) m_2(C)$$

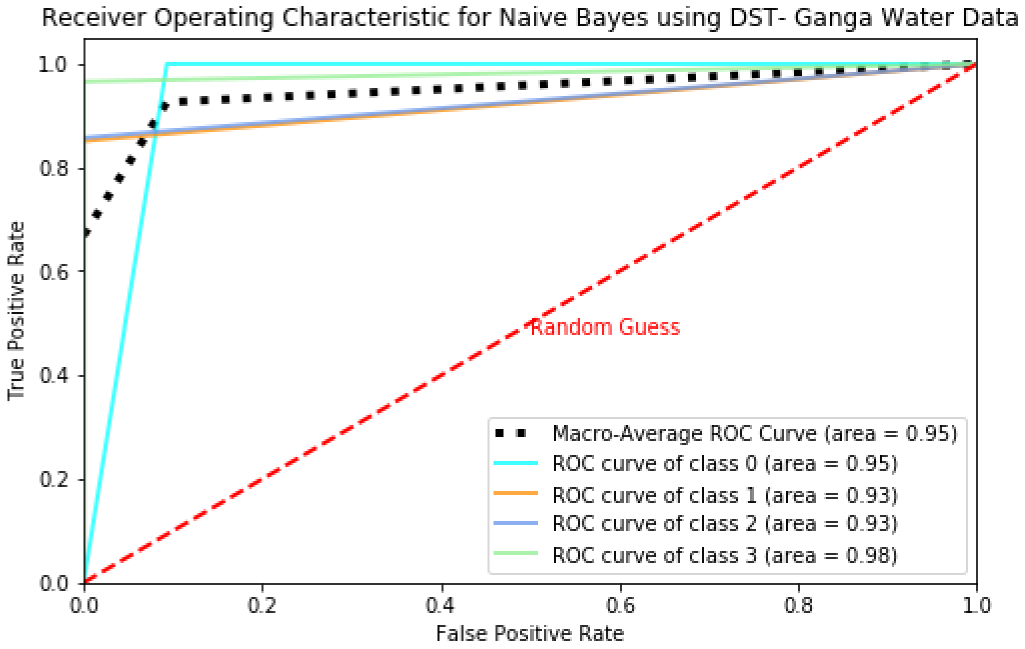
Where

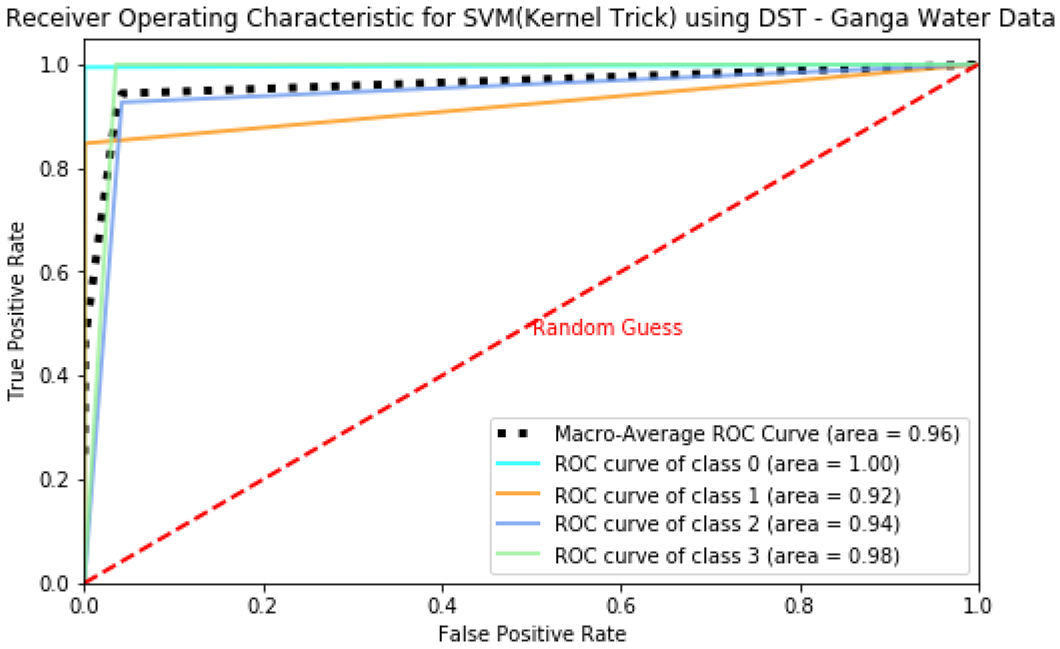
$$K = \sum_{B \cap C = \emptyset} m_1(B) m_2(C).$$

EXPERIMENTAL RESULTS

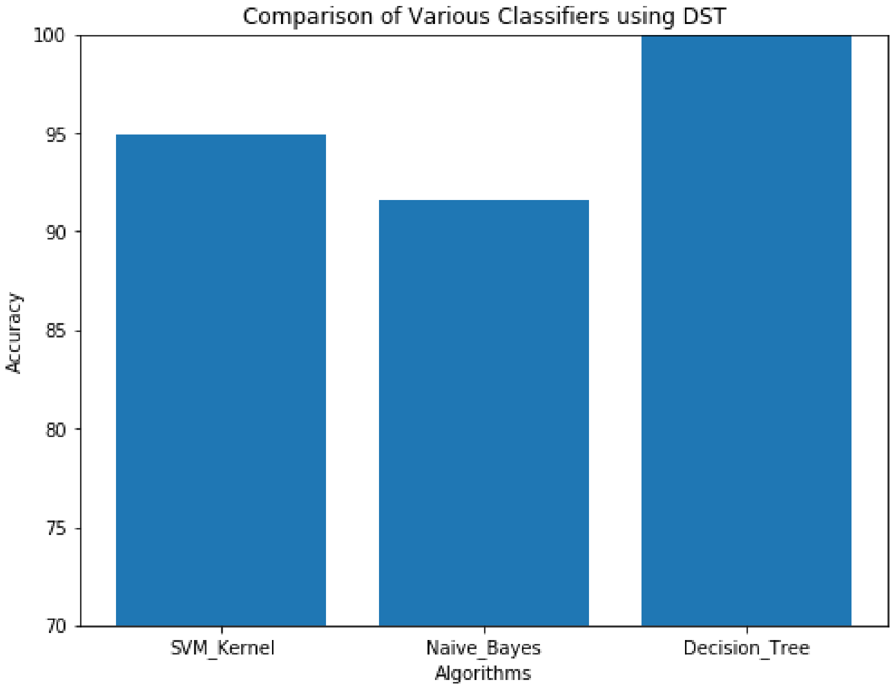
The Dempster Shafer models are visualized by plotting the following ROC Curves for each of the machine learning algorithm:



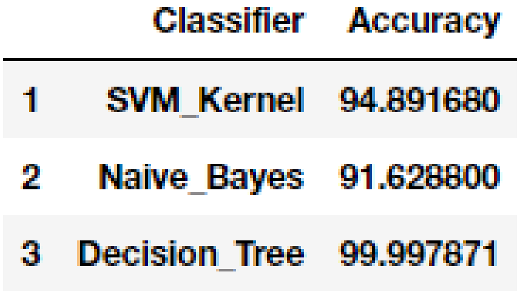




During the study, we found that the Logistic Regression model could not handle the data generated using Dempster Shafer Theory as Logistic Regression is generally used for binary classifications. Now, let us visualize the results of the algorithms used to find the best machine learning model which can assess the quality of water more accurately. The following bar-plot compares the models based on their accuracies:



The accuracies of the classifiers are:



	Classifier	Accuracy
1	SVM_Kernel	94.891680
2	Naive_Bayes	91.628800
3	Decision_Tree	99.997871

We can see, the **Decision Tree Classifier** has the highest accuracy.

CONCLUSION & FUTURE WORK

Determining the quality of water using data mining techniques is an interesting area of research. The study involved the details of water samples collected from the Ganga River.

We classified the samples into four clusters using K-Means clustering algorithm based on five parameters namely Electrical Conductivity, Dissolved Oxygen, Oxidation Reduction Potential, pH and Temperature. The models are then trained on various algorithms like SVM, Logistic Regression, Naïve Bayes and Decision Tree Classifier using Dempster Shafer Theory.

Among the classifiers, Decision Tree obtained highest accuracy, which indicates that the model can accurately assess the quality of water samples of River Ganga.

In future, we can use extended dataset and some other machine learning techniques to assess the quality of water.

Also, with extended dataset, we would be able to extract the pattern of the river water at a particular interval of time over a year by using predictive analysis. The forecasting of river flow has great importance in water resources.

REFERENCES

1) A. K. Shukla, C. S. P. Ojha and R. D. Garg “Surface water quality assessment of Ganga River Basin, India using Index mapping”

2) Gagan Matta, Sachin Srivastava, R. R. Pandey, K. K. Saini “Assessment of physicochemical characteristics of Ganga canal water quality in Uttarakhand”

3) A. K. Haritash, Shalini Gaur, Sakshi Garg “Assessment of water quality & suitability analysis of River Ganga in Rishikesh, India”

4) Naseema Khatoon, A. H. Khan, M. Rehman, Vinay Pathak “Correlation study for the assessment of water quality & it’s parameters of Ganga River, Kanpur, Uttar Pradesh, India”

5) B. Sharma, Mukesh Kumar, D. M. Denis, S. K. Singh “Appraisal of river water quality using open-access earth observation data set: A study of river Ganga at Allahabad”

6) Ruby Pandey, Divya Raghuvanshi, P. K. Sharma, Harendra Singh, Beena Tripathi, Shubham Bajpai, Anupam Dixit “Evaluation of the status of heavy metal contamination in sediment of the River Ganga at Allahabad, India”

7) Nashreen Nesa, Indrajeet Banerjee “IoT-Based sensor data fusion for occupancy sensing using Dempster–Shafer Evidence Theory for smart buildings”