Project Report

**A Machine Learning Based -Water Potability Prediction Model**



Project Report submitted on the fulfilment of the requirements of Post graduate Diploma in Big data Analytics

Authors: Co-Ordinator:

1) PATWARDHAN SAYUJA SUMUKH    (220960925033) Ms.Roopa Panikar(CDAC)  
2) TALELE TUSHAR SANJAY                   (220960925049) Ms.M Soorya (CDAC)  
3) THOPATE GIRISH PRAKASH               (220960925050) Ms.Divya Das (CDAC)  
4) WAGH ROHIT VILAS                            (220960925051)  
5) ZOPE JAYESH PRAKASH                     (220960925052)

**STDC**

**CDAC Thiruvananthapuram**

**Trivandrum, Kerala 695581**

**Title Page**

**Contents**

**1. Abstract**………………………………………………………………. 3

**2. Introduction**………………………………………………………….. 4

**3. Literature Survey**……………………………………………………. 5

**4. Proposed System**…………………………………………………….. 6

**a. Dataset pre-processing**…………………………………… 8

**b. Feature Extraction**……………………………………….. 11

**c. Models used**……………………………………………….. 13

**5. Discussion and Results**……………………………………………….15

**a. Output obtained**…………………………………………… 15

**b. Evaluation measures used**………………………………… 16

**6. Conclusion**……………………………………………………………. 18

**7. References**……………………………………………………………. 19

**Abstract**

We cannot imagine our life without water. It is one of the most important basic need for sustaining. Now a days due to industrialization and numerous other activities there is a threat caused to pure water resources. There is a lot of increase in water Pollution due to above mentioned factors. Polluted water causes serious waterborne illnesses and poses a threat to human health. Therefore it is very crucial to check the water Portability before human consumption.Potability means the check if the water is suitable for drinking or not. Predicting the quality of drinkable water may reduce the incidence of water-related diseases. The latest machine learning approach has shown promising predictive accuracy for water quality. This research uses nine different learning algorithms to determine drinking water quality. These are various classification and boosting algorithms implemented using PYTHON in-built libraries.

**Introduction**

Water, Food and Shelter are the 3 basic needs for a human being to survive. Due to vast advancement and due to urbanization we all are exploding the natural resources carelessly. Natural Water resources are becoming extinct. There are some places where there are resources but they are found to be highly impure due to numerous activities and actions. So it is very much necessary to see to it that we get good quality water for consumption and usage. We must see to it that the water which we are drinking is of good quality and safe for us. The Drinkable water quality prediction is essential to ensure safe public health. It is a very much serious issue for a person to survive healthy life. Polluted drinking water can cause various kinds of health diseases. According to the survey, almost 3,575,000 people are died every year due to water-related diseases. Predicting drinkable water is difficult for those countries that have limited drinkable water sources. In the industrial revolution, chemical dust causes the most water pollution.

Our project aims to predict water potability by taking into consideration various factors associated with it like pH, hardness, sulfate, organic carbon, turbidity, and a few more. Machine learning techniques show significant prediction results in water quality prediction. Logistic Regression, Decision Tree, Random Forest, XGBoost Classifier, KNeighbours (KNN), Support Vector Machine(SVM), AdaBoost Classifier, LGBM, Artificial Neural Network(ANN)are the most popular machine learning algorithm for prediction. We tried fitting the model to various algorithms and thereby check and comparing the accuracies of the same. The conventional laboratory technique for assessing water quality is time-consuming and sometimes costly. The algorithms proposed in this work can predict drinking water quality within a short period of time. ANN has 67 percent accuracy with a training error of 0.61 percent during the training period. XGBoost has an F1 score of 79% and a prediction accuracy of 68%.Using an extended data set could improve how well predictions are made and help stop waterborne diseases in the long run.

**Literature Survey**

This research explores the methodologies that have been employed to help solve problems related to water quality.

* Our **Focus** is Build Model which identifiers whether Water is Drinkable or not
* Dinesh kumar[1] carried out conventional lab analysis and statistical analysis in research to aid in determining water quality, while some analyses employ machine learning methodologies to assist in finding an optimized solution for the water quality problem.
* Chidanand Patil[2] Analysed water samples, The parameters analysed during the study period were pH, Total dissolved solids (TDS),Total Hardness, Nitrate, Most Probable Number (MPN) and heavy metal such as Lead using standard laboratory procedures.
* Mona A. Hagras et.al [03] to assess the quality of groundwater and to characterize the hadrochemical characteristics of the groundwater in Punjab, groundwater samples were collected from different cities of Punjab Province and analysed for 28 water quality parameters Groundwater suitability for domestic and irrigation purposes was assessed by using WHO and USDA standards.

**Proposed System**

* **Dataset Overview:**

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:

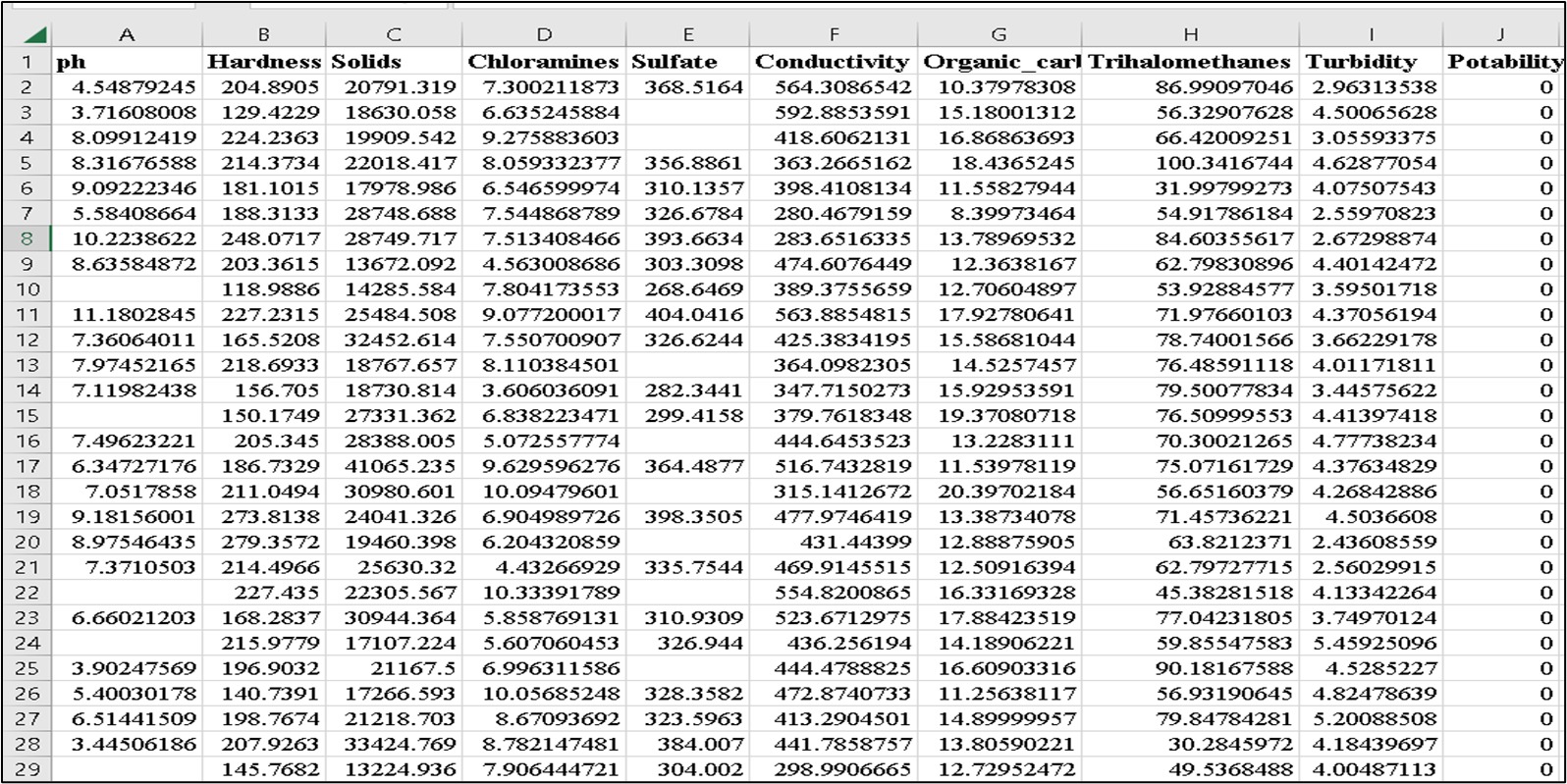
THMs are chemicals which may be found in water treated with chlorine. The concentration of THMs in drinking water varies according to the level of organic material in the water, the amount of chlorine required to treat the water, and the temperature of the water that is being treated. THM levels up to 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.



Pal et al. International Journal of Advanced Engineering Research and Science, 9(6)-2022

www.ijaers.com Page | 1

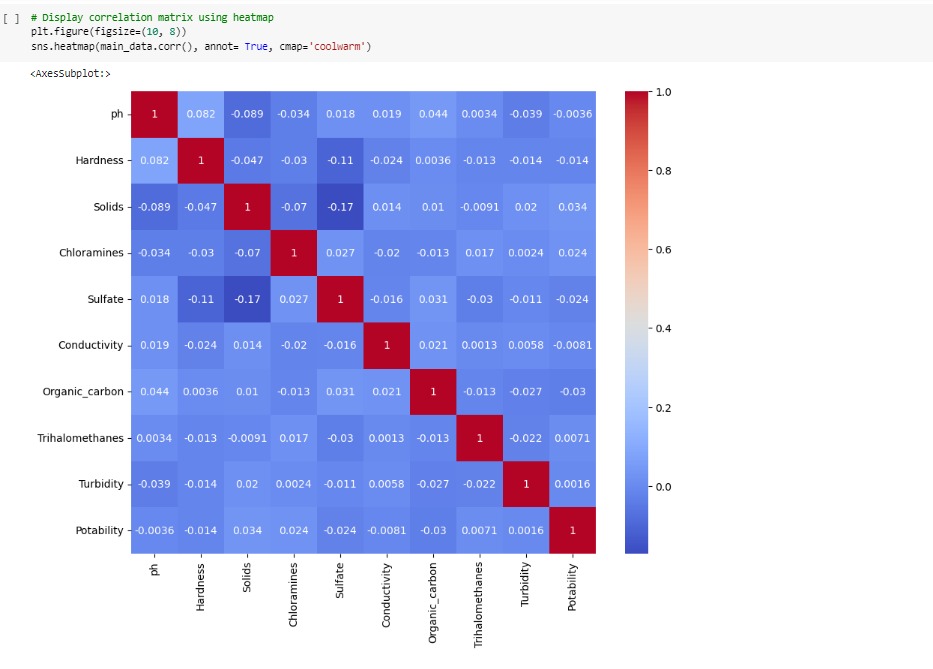
* **Data Pre-Processing:**

1. Exploratory Data Analysis:

In this step we tried to acquire the basic information about our dataset. Number of rows and columns along with the missing values. We collected information about the mean values, the median values and the maximum and minimum values for each column.

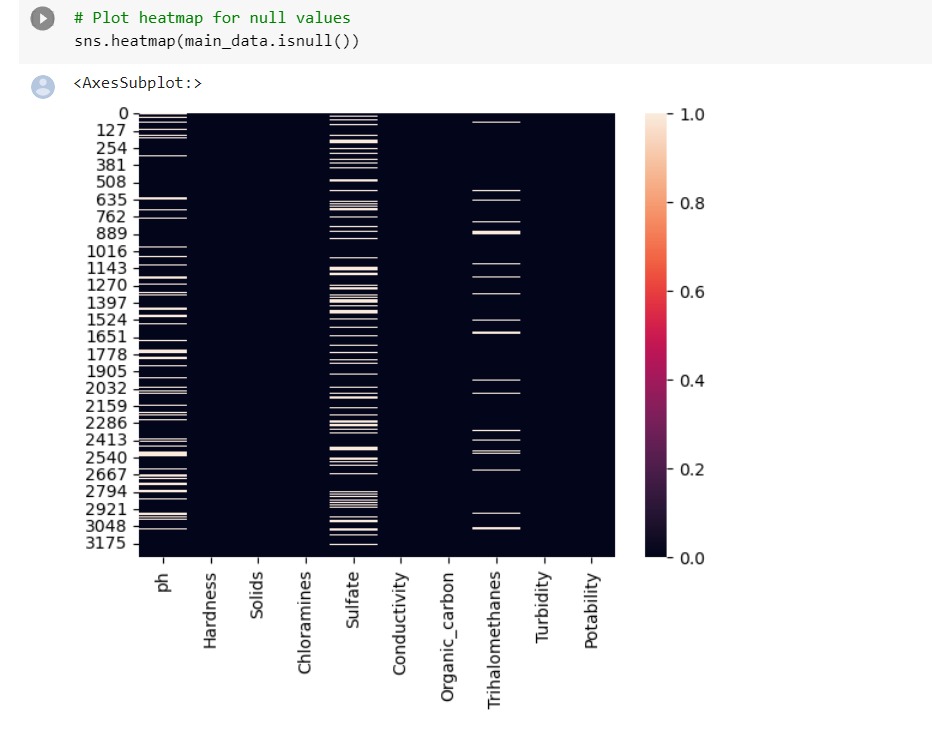
1. Visualizing Correlation Matrix using Heatmap:

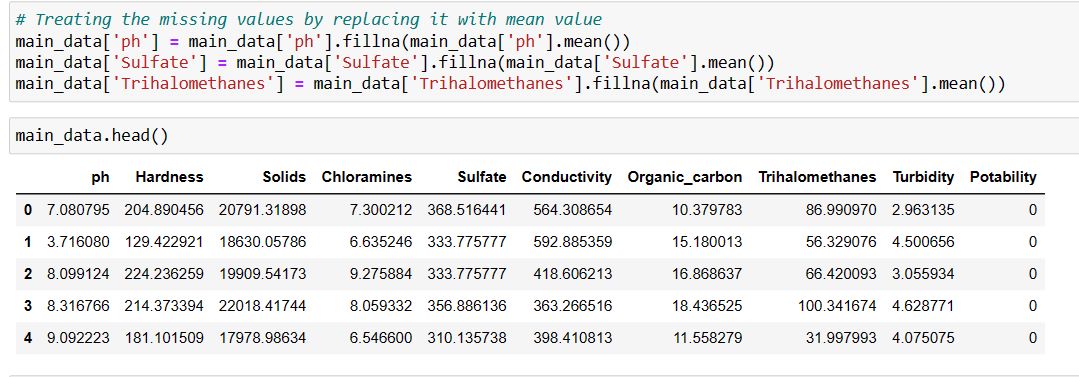
A correlation matrix is simply a table which displays the correlation coefficients for different variables. The matrix depicts the correlation between all the possible pairs of values in a table. It is a powerful tool to summarize a large dataset and to identify and visualize patterns in the given data. So we analyzed the dataset and tried finding correlation and presented it in the form of a heat map.

****

1. Treating Missing Values by replacing them with Mean:

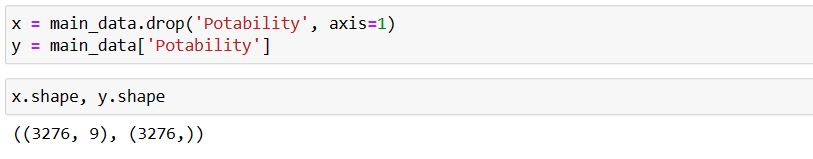
Our dataset had a lot of missing to Nan values. So before analyzing further it was very much important to treat these missing values or drop these values. Dropping off the missing values was not a suitable option so we decided to fill that values. We tried filling those values with median and mean. But after demo calculations we concluded that filling Missing values with Mean Values was the best option.



****

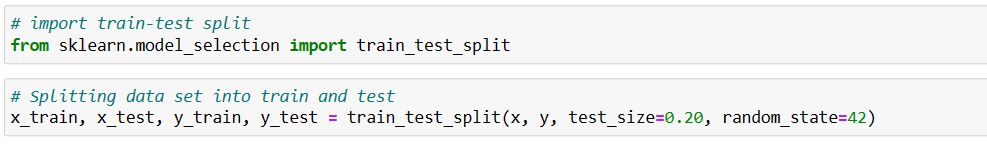
1. Setting a Target Variable:

Since our project is a classification based model so the result should be either YES or No. So in accordance to that we analyzed our dataset and found that the Column 'POTABILITY' should be our target variable.



1. Splitting the Data into training and testing:

In order to train the model, the data must be split, tested with a subset of the data, and computed with accuracy measures to determine the model's performance in the final stage before applying the machine learning model. Training data and test data were created from the dataset. The training data contained 70% of the total dataset and the testing data only contained 30% of the complete dataset. The ML builds a link with the independent and dependent parameters in order to forecast or choose an alternative, and then the test data is taken to determine if the machine learning technique is effective or not.



* **Feature Extraction:**

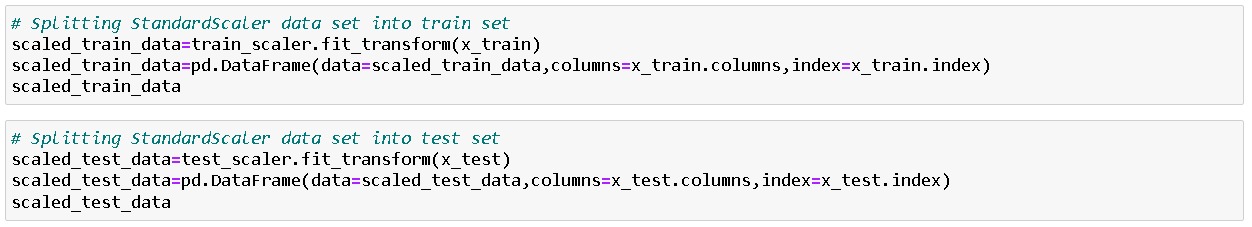
1. StandardScaler:

In Machine Learning, StandardScaler is used to resize the distribution of values ​​so that the mean of the observed values ​​is 0 and the standard deviation is 1.To bring a uniformity in our values we decided to implement Standard Scaling.This helped in easy analysis of our values and all the values ranged between -1 to 1.



1. Splitting Scaled Dataset:

After Scaling we again split the dataset into training and testing for achieving better results. This is done by using train test split and the ratio was 80% training dataset and 20% testing dataset.



1. Mutual\_info\_classif.:

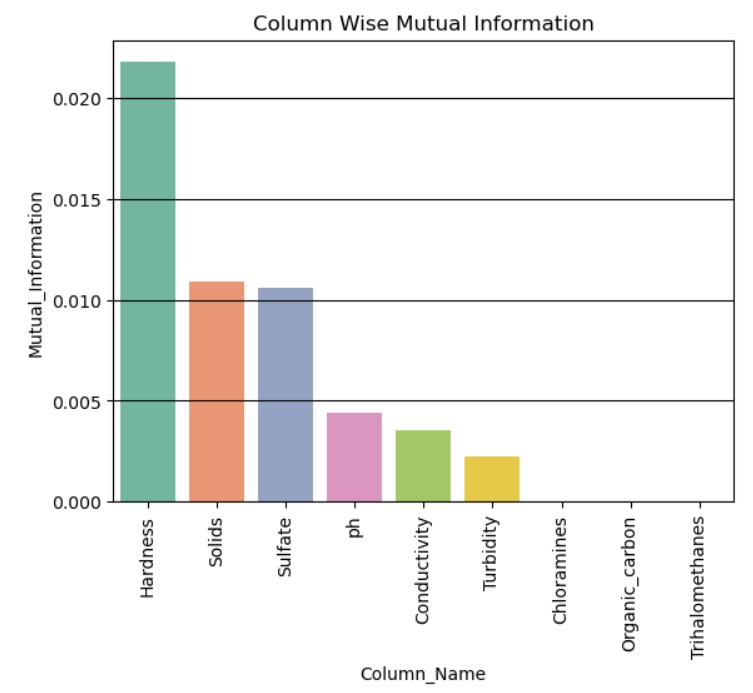
This method basically utilize the mutual information. It calculates mutual information value for each of independent variables with respect to dependent variable, and selects the ones which has most information gain.



1. SelectKBest method:

Feature selection is a technique where we choose those features in our data that contribute most to the target variable. In other words we choose the best predictors for the target variable. The classes in the sklearn.





* **Models used:**

1. Logistic Regression:

Logistic regression is a classification algorithm. It is based on the logistic function or the sigmoid function, hence the name. It is the most common algorithm used in the case of binary classification.

in this algorithm we are getting 62.8% Accuracy.

1. Decision Tree:

A decision tree is a simple self-explanatory algorithm, which can be used for both classification and regression. The decision tree, after training,

makes decisions based on values of all the relevant input parameters. It uses entropy to select the root variable, and, based on this, it looks toward

the other parameters’ values. It has all the parameter decisions arranged

in a top-to-down tree and projects the decision based on different values of different parameters. in this algorithm we are getting 63.2% Accuracy.

1. Random Forest:

Random forest is a model that uses multiple base models on subsets of

the given data and makes decisions based on all the models. In random

forest, the base model is a decision tree, carrying all the pros of a

decision tree with the additional efficiency of using multiple models.

in this algorithm we are getting 62.8% Accuracy.

1. Support Vector Machine:

Support vector machines (SVMs) are mostly used for classification but

they can be used for regression as well. Visualizing data points plotted

on a plane, SVMs define a hyper plane between the classes and extend the margin in order to maximize the distinction between two classes, which results in fewer close miscalculations.

in this algorithm we are getting 67.22% Accuracy.

1. K-nearest Neighbors:

The K nearest neighbor algorithm classifies by finding the given points

nearest N neighbors and assigns the class of majority of n neighbors to it.

in this algorithm we are getting 62% Accuracy.

1. XGBoost Classifier:

XGBoost is an optimized distributed gradient boosting library designed for efficient and scalable training of machine learning models. It is an ensemble learning method that combines the predictions of multiple weak models to produce a stronger prediction. XGBoost stands for “Extreme Gradient Boosting” and it has become one of the most popular and widely used machine learning algorithms due to its ability to handle large datasets and its ability to achieve state-of-the-art performance in many machine learning tasks such as classification and regression

in this algorithm we are getting 68.9% Accuracy.

1. Adaboost Classifier:

By turning a number of poor learners into strong learners, these methods boost prediction power. Boosting algorithms work on the idea of first

building a model on the training dataset and then building a second model to correct the faults in the first model.

in this algorithm we are getting 63.8% Accuracy.

1. Light Gradient Boosting Machine:

In contradiction with AdaBoost, the training context loads are not improved; however, every estimator is prepared by using presidency’s errors as symbols. Gradient Boost is a technique that includes Classification and Regression Tree (CART) as the concealer trainee.

in this algorithm we are getting 65.3% Accuracy.

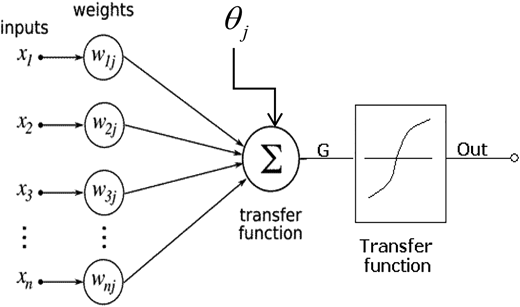
1. Artificial Neural Network:

Neural nets are loosely based on the structure of neurons. They contain multiple layers with interconnected nodes. They contain an input layer and output layer, and hidden layers in between these two mandatory layers. The input layer takes in the predicting parameters and the output layer shows the prediction based on the input. They iterate through each training data point and generalize the model by giving and updating the weight on each node of each layer. The trained model then uses those weights to decide what units to activate based on the input. Multi-layer perceptron (MLP) is a conventional model of neural net, which is mostly used for

classification, but it can be used for regression as well. We used it for

classification with the configuration of (3, 7) running for a maximum of 100 epochs .

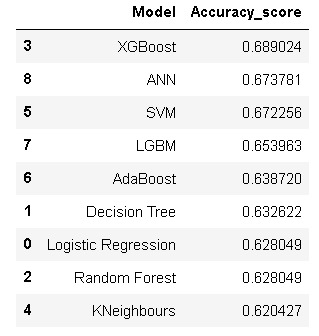
in this algorithm we are getting 67.3% Accuracy.

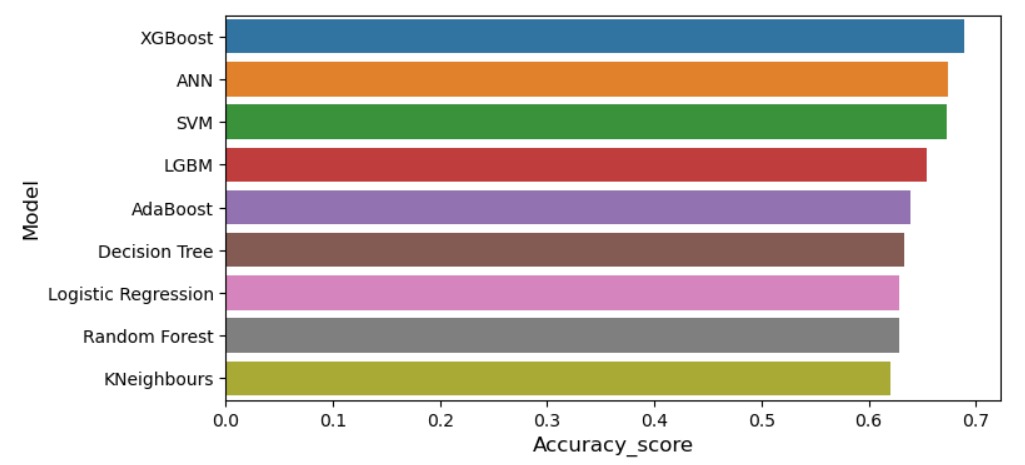


**Discussion and Results**

* **Output obtained:**

For creating our classifier and regression model based on the dataset, we used all of the algorithms stated above. However, we were just employing nine classifiers, which are the most accurate of all the systems. Logistics Regression, Random Forest Classifier, Light Gradient Boosting Classifier(LGBM), Decision Tree, XGBoost, AdaBoost Classifier, KNeighbours, Support Vector Machine(SVMs), Artificial Neural Network(ANN), are some of the algorithms we used. To evaluate our model, we used on these nine classifiers as shows in Table.

****

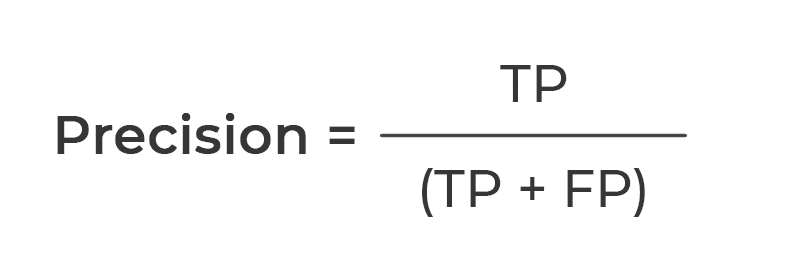
****

* **Evaluation measures used:**

**1.Precision:**

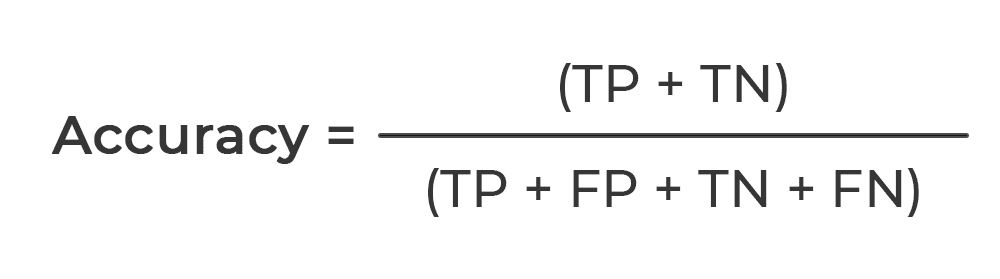
The proportion of accurately categorized occurrences as in a classifier among all the interpreted contexts is known as precision. Equation is used to compute TP (denoting positive class) while FP is about false

alarm in precision.

****

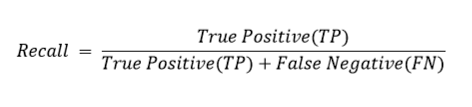
**2. Accuracy:**

The proportion of valid simulation provided across all confidence intervals according to the variant is known as accuracy. Equation is used to calculate accuracy, TP conveys true positive, TN signifies true negative, FP reflects false positive, and FN specifies false negative.



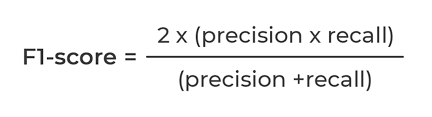
3. **Recall:**

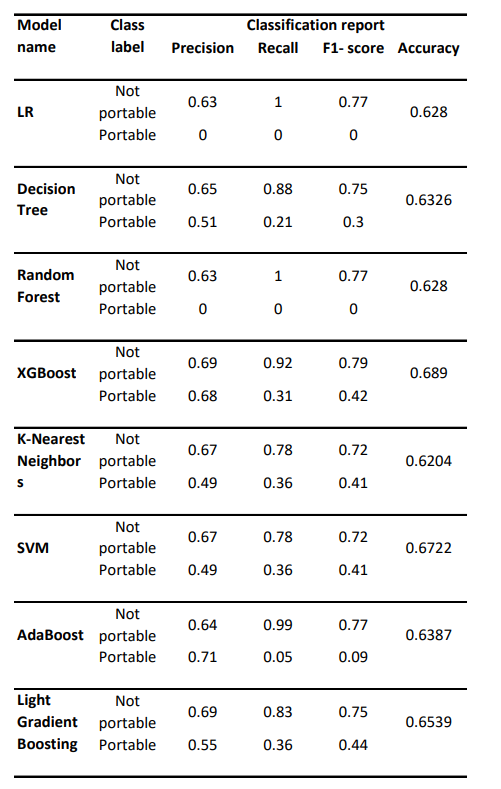
The margin for jurisdictions having a certain strong group of individuals willing properly categorized is known as recall. In the formula illustrated in to determine recall, TP accounts as true positive and FN refers to false negative.



**4. F1 Score:**

Because not everything is enclosed under efficiency and recall elements of validation on their own, as per the formula, we preferred a harmonized average to depict F1 score, 15, which thus encompasses either characteristic and more accurately depicts the total reliability metric. It has a range of 0 to 1. The greater the score is, the more accurate it is.





**Conclusion**

Predicting drinkable water is essential for environmental preservation and pollution prevention. It is necessary to provide clean drinking water in order to maintain excellent public health. Drinking water from safe sources can ensure the potability of the water. It becomes difficult to predict drinkable water accurately. The ideal learning algorithm is needed to prevent prediction errors. An intelligent model based on nine different machine learning algorithms may be used to predict the potability of drinking water based on 9 standard parameters such as pH, hardness, organic carbon, and other factors. In this current work, XGBoost Classifier achieved highest 69 percent accuracy. ANN has 67 percent accuracy with a training error of 0.61 percent .In future, the proposed model will be implemented to predict and analysis of different region drinking water along with IoT based quality detection model.

**References**

1. Dinesh kumar tank and c. p. Singh chandel, “Analysis of the major ion constituents in groundwater of Jaipur city”, Nature and Science,8(10), 1-7
2. Chidanand Patil, Shreekant Narayanakar and Arjun Virupakshi, “Assessment of Groundwater Quality Around Solid Waste Landfill Area - A Case Study, International Journal of Innovative Research in Science, Engineering and Technology, Vol. 2, pp. 3131-3136.
3. Mona A. Hagras Assistant Professor, Irrigation &Hydraulics Department, Faculty of Engineering, Ain Shams University, Cairo, Egypt
4. Reference taken from Kaggle website.
5. https://www.kaggle.com/datasets/adityakadiwal/water-potability.