Water Quality Analysis

MD. ANISUR RAHMAN[1], MOSAROF HOSSAIN[1], ARMANUL ISLAM[1], TORIKUL ISLAM[1], and RAJIB HOSSAIN[1]

[1]Daffodil International University

August 2021

Abstract

Water is perhaps the most precious natural resource after air. Though the surface of the earth is mostly consisting of water, only a small part of it is usable, which makes this resource very limited. This precious and limited resource, therefore, must be used with prudence. As water is required for different purposes, the suitability of it must be checked before use. Also, sources of water must be monitored regularly to determine whether they are in sound health or not. Poor condition of water bodies is not only the indicator of environmental degradation, it is also a threat to the ecosystem. In industries, improper quality of water may cause hazards and severe economic loss. Thus, the Potability of water is very important in both environmental and economic aspects. Thus, water potability analysis is essential for using it in any purpose. After days of research, water potability analysis is now consists of some standard protocols. There are guidelines for sampling, preservation and analysis of the samples. Here the standard chain of action is discussed briefly so that it may be useful to the analysts and researchers. The objective of this paper is to predict the water quality from the given parameters of the data set using multiple machine learning approaches and compare the performance of those approaches. The following models are approached to water quality data set, Linear regression(LN), Support vector machine(SVM), Decision tree(DT) and Random forest(RF). The performance evaluation is conducted by using F1-score. The results shows that using random forest and by tuning and balancing the data we can get a higher accuracy.

1 Introduction

71% of the Earth's surface is being covered by water. This means that the overall quality of the water has a serious impact on the health of humanity and its environment. Scientists, governments and many dedicated researchers are committed to carrying out water quality monitoring projects all around.[11] Hence, water bodies as a domain for planning and management has been accepted all over the world. Unfortunately, these useful resources are deteriorated by humans. Contamination of chemicals from industries, sediments etc. are the factors that affect the quality of the water directly. Due to increase in urbanization, the exploitation deterioration of water bodies has been increased. Contaminated water resources can cause serious effects on human as well as aquatic life. Along with these, the quality of water is influenced by numerous factors like human activities, soils, geology, indiscriminate disposal of sewage, human activities.[12] Also, contaminated water can lead to some waterborne diseases and also influences child mortality. On the other hand we also can see floods happening in the monsoon season. he monsoon floods of the year 2020 has an overall impact on the Northern, North-Eastern and South-Eastern region of Bangladesh. The floods have impacted 21 districts of Bangladesh with moderate to severe impact on 16 Districts. As of 22 July, 2020, 102 upazila and 654 unions have been inundated in flood, affecting 3.3 million people and leaving 7,31,958 people water logged. people became sick of 400 villages of 56 unions in nine upazilas are living misery as they face an acute shortage of pure water one of the countries with the highest quality of water in the world. In Bangladesh, it is officially recognized by the government of Bangladesh that 50% of the country's approximately 150 million people, are at risk of arsenic poisoning from groundwater used for drinking.

But we all know water is one of the major natural resources for people. In 2012 it was declared that a safe water supply for every person is a crucially important task worldwide. There are special water sustainability guides issued by the World Health Organization (WHO) and regulated water quality standards.

Recently, the government of Bangladesh, in its Action Plan for Poverty Reduction, stated its desire to ensure 100% access to pure drinking water across the region within the shortest possible time frame [3]. This is also consistent with key goals of the Millennium Development Goal "Eradication of extreme poverty and hunger" and "Halving by 2015, the proportion of people without sustainable access to safe drinking water". Whether this is achievable within the stated time is debatable, but it clearly delineates the state of the world we live in. - Abul Hussam, in Monitoring Water Quality, 2013.[3][7]

Contaminated water and poor sanitation are linked to transmission of diseases such as cholera, diarrhoea, dysentery, hepatitis A, typhoid, and polio. Absent, inadequate, or inappropriately managed water and sanitation services expose

individuals to preventable health risks. This is particularly the case in health care facilities where both patients and staff are placed at additional risk of infection and disease when water, sanitation, and hygiene services are lacking. Globally, 15% of patients develop an infection during a hospital stay, with the proportion much greater in low-income countries.

So,our model was encouraged based on the problems, where we use Water Potability analytical system which helps to find and measure the water purity. Water Potability can be defined as the chemical, physical and biological characteristics of water, usually in respect to its suitability for a designated use. Water can be used for recreation, drinking, fisheries, agriculture or industry. Each of these designated uses has different defined chemical, physical and biological standards necessary to fulfil the respective purpose. The reason behind using this system for surety or some measurement that how much surface elements are mixed with raw water. In raw water there have to maintain pH value that is an important in evaluating the acid-base balance of water. Hardness which mainly caused by calcium and magnesium salts. These salts are dissolved from geologic deposits through which water travels. Chloramines, Sulfate and Organic carbon are naturally occurring substance that are found in minerals, soil, and rocks. Some minerals produced un-wanted taste and diluted color in appearance of water. such as potassium, calcium, sodium, bicarbonates, chlorides, magnesium, sulfates etc. We need to measure Trihalomethanes, Conductivity, Turbidity.

Safe and readily available water is important for public health, whether it is used for drinking, domestic use, food production or recreational purposes. Improved water supply and sanitation, and better management of water resources, can boost countries' economic growth and can contribute greatly to poverty reduction. Water Potability analysis is required mainly for monitoring purpose. (i) To check whether the water Potability is in compliance with the standards, and hence, suitable or not for the designated use. (ii) To monitor the efficiency of a system, working for water quality maintenance (iii) To check whether upgradation change of an existing system is required and to decide what changes should take place (iv) To monitor whether water Potability is in compliance with rules and regulations.[12]

Recent report of United Nations reveals that the study can be helpful to limit the waterborne diseases and also it can help people to get pure and safe water for their daily needs. 3 million people in the world die of water related diseases due to contaminated water each year, including 1.2 million children.[8] Predictive analysis can help to capture relationships among many factors that can help to assess risk with a particular set of conditions. Predictive analysis includes data processing techniques, statistical analysis and modelling. In this process, defining the objectives, deliverables and identifying the datasets is done at first. After defining the project, various data processing techniques are used to discover the useful information from the dataset. These extracted information is applied various data analysis and statistical analysis techniques. Finally, the predictive model is created in order to get the predictions. Multiple models are applied to the same dataset and the model which best fit is chosen [1].

Our contribution in this paper is as follows:

- Studying the different attributes and parameters such as pH, Chloramines and turbidity.
- Making the dataset as clean and usefull as possible.
- Further analysis using machine learning techniques.
- Performance analysis using precision, recall, confusion matrix and F1-score.

The remainder of this paper is organized as follows: Section 2 provides a literature review regarding this model. In Section 3, we review the dataset and perform further processing as needed. In Section 4, we employ various machine learning methodologies to predict water quality and discuss the results of regression and classification algorithms, in terms of accuracy and classification precision. In Section 5,we conclude the paper and provide future lines of work.

2 Literature Review

One of the similar study, Benchtop analysis and online monitoring equipment were used to measure pH, chlorine residual, turbidity, and total organic carbon values before and after the introduction of these contaminants. Results indicate that all four contaminants can be detected at relatively low concentrations. Three of the four contaminants were detected below a concentration that would cause significant health effects [5].

Another study, Zhang, Zhu, Yue, and Wong (2017) propose a novel anomaly detection algorithm for water quality data using dual time-moving windows, which can identify anomaly data from historical patterns in real-time. The algorithm is based on statistical models, autoregressive linear combination model. They have tested the algorithm using 3-month water quality data of PH from a real water quality monitoring station in a river system. Experimental results show that their algorithms can significantly decrease the rate of false positive and has better anomaly detection performance than AD and ADAM algorithms.[14]

In 2014, some researchers of Malaysia conducted a study on two lakes named Chini and Bera. They collected samples from 2005 to 2009. The data sample consisted of 11 parameters which were used to predicate DO concentration. The DO concentration was refered into High, Medium and Low. They ranked the input parameters and they used forward selection method to determine the optimum parameters that yield the lowest errors and highest accuracy. The initial results showed

that pH, temperature and conductivity significantly affect the prediction of DO. Then, they applied SVM model using the Anova kernel with those parameters providing 74% accuracy rate.[10]

In 2014, some researchers conducted a study on managing the water quality of watersheds. The study was conducted at Rawal watershed in Pakistan. They collected monthly data from the watershed in order to analyze the quality of the water as per WHO standards. Regression model was applied whereas the combination of supervised and unsupervised machine learning techniques was applied to test the quality indices. The study found that unsupervised learning techniques like average linkage method of hierarchical clustering using Euclidean distance is an accurate method to find out the water quality index and for classifications, they found MLP, a supervised learning technique which can give accurate results.[4]

Familiar researchers in Pakistan explored research employing machine learning methodologies in the realm of water quality estimated water quality using classical machine learning algorithms namely, Support Vector Machines (SVM), Neural Networks (NN), Deep Neural Networks (Deep NN) and k Nearest Neighbors (kNN), with the highest accuracy of 93% with Deep NN. The estimated water quality in their work is based on only three parameters which are turbidity, temperature and pH.[1]

Kang, Gao, and Xie have developed a model where best result was achieved using Artificial Neural Network with Nonlinear Autoregressive. In 2009, Xiang and Jiang applied least squares support vector machine (LS-SVM) with particle swarm optimization methods to predict the water quality. They discovered that through simulation testing, the model shows high proficiency in estimating the water quality of the Liuxi River Xiang and Jiang.[13]

Employee of Ahmad et al. used forward neural networks and a combination of multiple neural networks to estimate the WQI. They used 25 water quality parameters as the input. Using a combination of backward elimination and forward selection selective combination methods, they achieved an R2 and MSE of 0.9270, 0.9390 and 0.1200, 0.1158, respectively. The use of 25 parameters makes their solution a quite tough one.[2]

There was a study where, Liu et al. introduced a water quality prediction model leveraging LSTM deep neural networks. They collected data by using third-party water monitoring stations. The results demonstrated that the model can predict water quality over a 6 months period. Even though the results are of relevance, the main limitations of the proposed prediction model are related to 1-dimensional inputs and limited dataset.[9]

Another study, Gazzaz et al. used ANN to predict the WQI with a model explaining almost 99.5% of variation in the data. They used 23 parameters to predict the WQI, which turns out to be quite expensive if one the WQI with a model explaining almost 99.5% of variation in the data.[6]

As we have seen, machine learning can achieve good results for anomaly detection of water quality, and our work is inspired by these related works. We used 4 types of algorithm in our project. Linear regression, Decision tree, Support vector machine(SVM), and Random forest. This random Forest estimated water potability in our work is based on only three parameters. Default parameter, Hyper parameter tuning and Balancing of our data. This overall parameter helps to get accurate predictions which is 88%. Machine learning algorithms can significantly decrease the number of false predictions. The attributes are described below,

3 Methodology

Methodologies adopted to get predictions for water quality has been discussed in this section. Our working process for building our model is mainly divided into four parts or subsection. First we have investigated the parameters in our dataset, then we analysed and processed the dataset, a brief description of the algorithm we used and lastly the implementation procedures. Further explanation is given below,

3.1 Dataset Description

The dataset of our research was collected from kaggle, where many results were combined to make better analysis. The dataset gives us water metrics of 3276 records of water potability records. We have 10 attributes where one of them is the targeted attribute which is 'Potability'. The other attributes are mainly independent which are necessary for predicting the water quality.

- pH value: PH is an important parameter in evaluating the acid–base balance of water. Maximum permissible limit of pH is from 6.5 to 8.5 recommended by WHO. In our dataset, the range was mainly from 6.52-6.83.
- Hardness: Hardness is mainly caused by calcium and magnesium salts. The length of time water is in contact with hardness producing material helps determine how much hardness there is in raw water.
- 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 unwanted taste and diluted color in appearance of water. Desirable limit for TDS is 500 mg/l and maximum limit is 1000 mg/l for drinking purpose
- Chloramines: Chlorine and chloramine are the major disinfectants used in public water systems. Chlorine levels up to 4 milligrams per liter (mg/L or 4 parts per million (ppm)) is considered safe in drinking water.

- Sulfate: Sulfates are naturally occurring substances that are found in minerals, soil, and rocks. 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.
- Conductivity: Pure water is not a good conductor of electric current rather's a good insulator. Increase in ions concentration enhances the electrical conductivity of water. According to WHO standards, EC value should not exceeded 400 μS/cm.
- Organic_carbon: TOC is a measure of the total amount of carbon in organic compounds in pure water. According to US EPA $<$ 2 mg/L as TOC in treated / drinking water, and $<$ 4 mg/Lit in source water which is use for treatment.
- Trihalomethanes: THMs are chemicals which may be found in water treated with chlorine. THM levels up to 80 ppm is considered safe in drinking water.
- Turbidity: The turbidity of water depends on the quantity of solid matter present in the suspended state. The mean turbidity value obtained for Wondo Genet Campus (0.98 NTU) is lower than the WHO recommended value of 5.00 NTU.
- Potability: This is our targeted attribute where 1 means Potable and 0 means Not potable.

3.2 Data analysis and processing

After knowing the parameter that are stated in the dataset, we have started exploring it. We got to know about the full overview of the dataset which includes median, standard deviation, maximum and minimum value 25-75% of the data. Then we observed, how many null values are presented in the dataset and which attribute holds how many.

Table 1: Overall Information

	ph	Hardness	Solids	Chloramines	Sulfate	Conductivity	Organic$_c arbon$	Trihalomethanes	Turbidity
count	2785.00	3276.00	3276.00	3276.00	2495.00	3276.00	3276.00	3114.00	3276.00
mean	7.08	196.36	22014.09	7.122	333.775	426.205	14.289	66.396	3.966
std	1.59	32.879	8768.57	1.58	41.416	80.82	3.308	16.175	0.78
min	0.00	47.43	320.94	0.35	129.00	181.48	2.20	0.734	1.45
25%	6.09	176.85	15666.69	6.13	307.70	365.73	12.07	55.84	3.44
50%	7.04	196.97	20927.83	7.13	333.07	421.88	14.22	66.62	3.96
75%	8.06	216.67	27332.76	8.11	359.95	481.79	16.56	77.34	4.50
max	14.00	323.12	61227.20	13.13	481.03	753.34	28.30	124.00	6.74

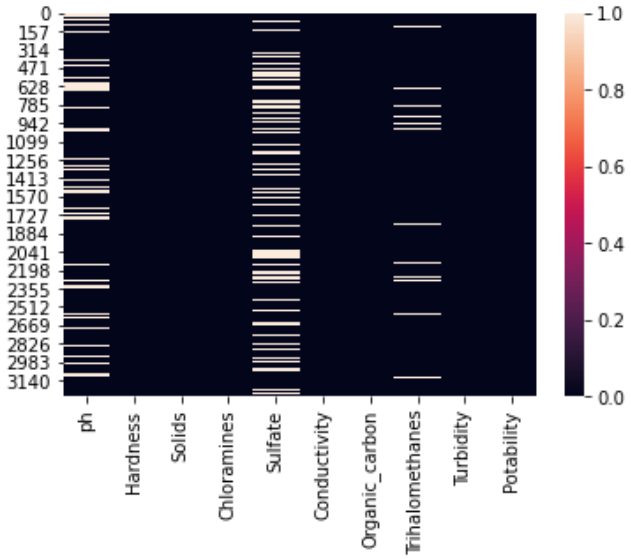

Figure 1: Heatmap of Null variables

From the figure we can see that, there are three parameters or variables which has quite a few null values. ph, Sulfate and Trihalomethanes are the selctive parameters.

Then we saw the co relation between the variables. To create a model, we need to know how strong the relation is between the variables. So that, we can drop out one value and select only the one value from the strong relation variables. This eases out to train the model for further processing. From the co relation report we can see that, the correlation between

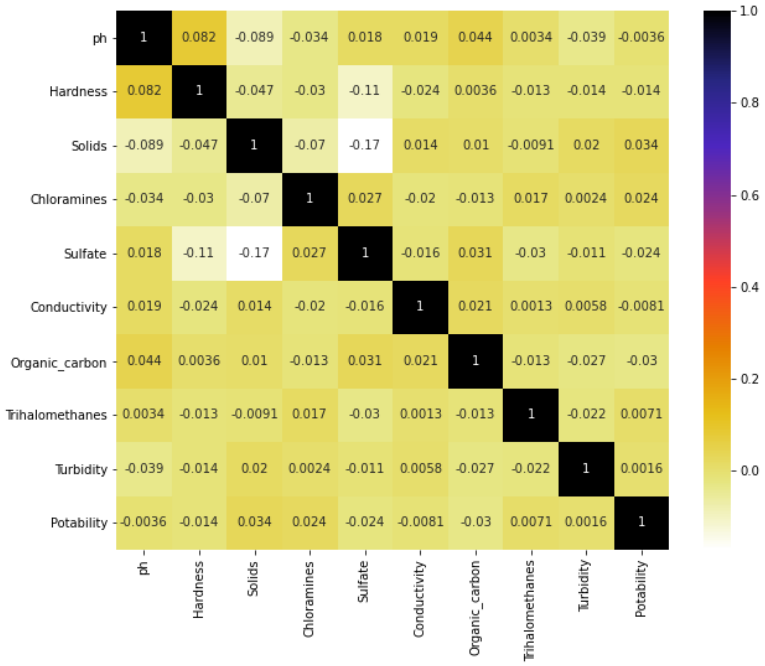

Figure 2: Co relation of the variables

the attributes are fairly low and this suggests the attributes are independent. As we didn't saw any strong commonalities we took all the attributes for preparing our model. There is no multicollinearity.

As we have multiple null variables in ph, Sulfate and Trihalomethanes, we wanted to fill them with the appropriate values. In order to that, we have used kernel density estimate (KDE) plot.

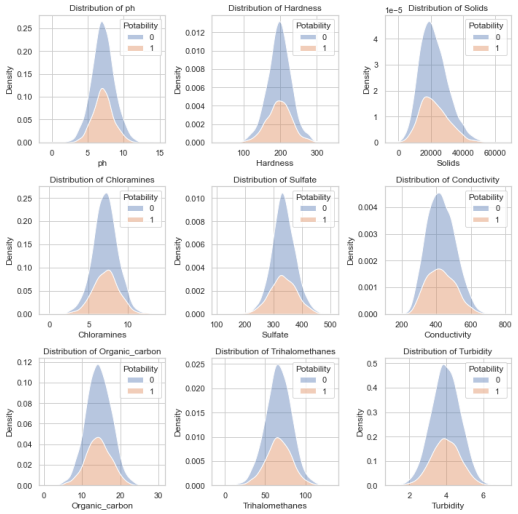

Figure 3: Distribution of variables regarding Potability

By using the kdeplot, we have a visual of the distribution of observations of the variables where hue is our targeted parameter Potability. It is representing the data using a continuous probability density curve in one or more dimensions. From the plot, we can see that all the distributions are fairly normal and distributed around the mean. So, we filled the null values with the mean values of the selective parameter corresponding to Potability.

3.3 Algorithm Description

In our model we have use 4 different algorithms to fit and train our data. We have used Logistic Regression, Support Vector Machine(SVM), Decision Tree and Random Forest. The selected algorithms are described below,

1. Logistic Regression: It is a generalized from of linear regression. Supervised learning classification algorithm is used to predict the probability of a target variable. For a non linear relationships among the independent variables logistic regression is needed. Our model falls into binary or binomial logistic reaction because if we look at our target attribute we can see that there are only two possible types either 1 or 0. Whee 1 means potable and 0 means not potable. If we consider the following equation,[15]

$$y = \alpha_0 + \alpha_1 Z_1 + \alpha_2 Z_2 + ... + \alpha_n Z_n \tag{1}$$

   Here, y is the response variable and Z1;Z2;Z3;...Zn are the predictor variables. By applying the sigmoid function on the equation, we can get the logistic function.

$$y = 1/[1 + e^{-(\alpha_0 + \alpha_1 Z_1 + \alpha_2 Z_2 + ... + \alpha_n Z_n)}] \tag{2}$$

2. Support Vector Function(SVM): It is another supervised learning algorithm we used for our model. a representation of different classes in a hyperplane in multidimensional space. In order to minimize the error hyperplane works in a iterative way. SVM divides the dataset into classes, then generate hyperplane iterative and lastly it chooses the correct hyperplane to separate the classes. To separate two classes, let's assume we are given a training data set, D (x1;C1)(x2;C2);...(xN;CN); where Xi denotes input vector and Ci refers to the class label of the vector which could be specified as either positive or negative [15]. For specifying any unspecified vector X, the condition is as follows:

$$f(X) = \sum_{i=1}^{N} \alpha C_i (x_i^T X) + b \tag{3}$$

3. Decision Tree: It is a supervised learning technique which uses a tree structured classifier. It is called a decision tree because it starts with a root and then creates branches to a number of solutions just like a tree and the number of branches increases with the increasing number of decisions or conditions. A tree grows from its roots and then creates branches as it gets bigger and bigger. This algorithm compares the values of root attribute with the record attribute and, based on the comparison, follows the branch and jumps to the next node. Decision tree uses an important technique Attribute selection measure or ASM in order to find the best nodes of a tree. Information gain and Gini index are the two techniques for ASM which helps to think about all the possible outcomes from a certain record. The decision of splitting the data is controlled by entropy and which can be defined by the equation below, where pj is the probability of the jth class,[15]

$$E(S) = \sum_{j=1}^{e} -P_1 \log_2 P_1 \tag{4}$$

4. Random Forest: It is based on ensemble learning which combines multiple classifiers to solve a complex problem and tries to improve the performance of that model. It contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset. It is capable of handling large datasets with high dimensionality and enhances the accuracy of the model and prevents the overfitting issue. For b = 1, ..., B:

   Sample, with replacement, n training examples from X, Y; call these Xb, Yb. Train a classification or regression tree fb on Xb, Yb.After training, predictions for unseen samples x' can be made by averaging the predictions from all the individual regression trees on x':[15]

$$f' = \frac{1}{B} \sum_{b=1}^{B} f_b(x') \tag{5}$$

3.4 Implementation Process

The implementation of our model was done with some useful machine learning libraries. We have used numpy, seaborn, pandas, matplotlib, plotly and other libraries for statistical analysis and visualizing our dataset.

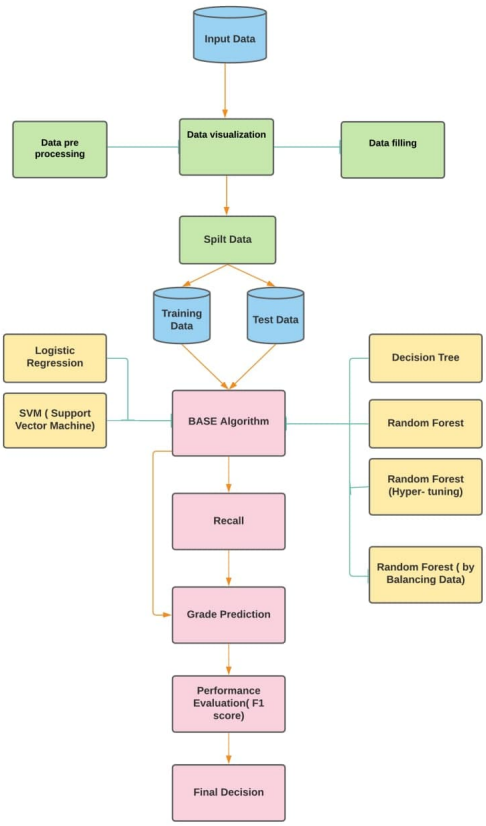

Figure 4: Step by step building our model

Here, we have collected the data from kaggle which contains the combination of many water parameter records. The dataset consist of 3276 records with 9 independent parameter and one target variable.

Then we started preprocessing our data for the model. We have visualized the data and saw the overall statistical analysis of our dataset. We have used correlation in our dataset to see the relations between the parameters. We were searching for a strong correlation so that we can neglect some parameters which could have easier for the model to learn. But the relation was fairly and we took all the parameters in action. There were many null values present at our dataset, So, for filling them with appropriate values we viewed many graphical plots where the hue was our target attribute. As a result we filled the missing values with the mean value of each parameter with respective to our target attribute. For increasing our model accuracy we have balanced the data of our target variable. By this our data was processed and was ready for training and testing.

Before fitting our model, we have split the data into two parts. Training data and Test data. We have taken 70 percent data for training and the rest for testing. Then we standardized the training and test data to make the the data scale free. The entire data set scales with a zero mean and unit variance, altogether.

After splitting the data , we fit the training data to our selective algorithms. For our model, we have used, logistic regression, support vector machine, decision tree and random forest algorithms. From that we can see that random forest algorithm gives a better accuracy than the other algorithms with the recall and precision and F1 score. So for increasing our accuracy we have used hyper tuning and saw the models accuracy was increased. Then again, it was not satisfactory to us. So now, by balancing the data of our target attribute we saw a good accuracy with precision, recall and F1 score. And our model was trained.

4 Performance Result

In this section we will see the performance analysis of the machine learning algorithms we used to build our model. We will find out the the best algorithm suited for our model. We will the precision, recall and the f1 score to measure our algorithms performance. Besides that, confusing matrix will help us to understand how good or bad the model is working. First, we will visualize the confusion matrix for every algorithm,

- Logistic Regression:

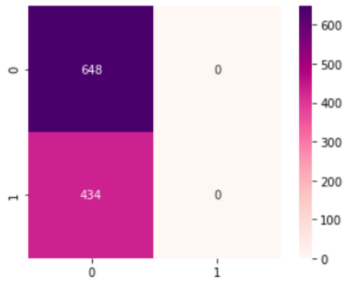

Figure 5: Confusion matrix of logistic regression

- Support Vector Machine:

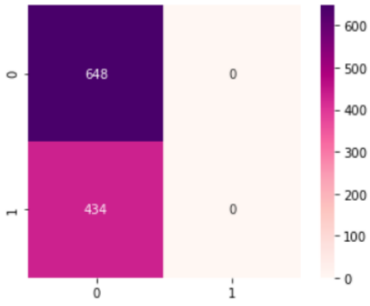

Figure 6: Confusion matrix of SVM

- Decision Tree:

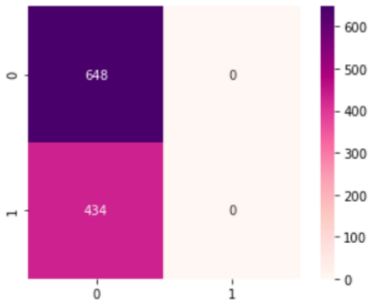

Figure 7: Confusion matrix of Decision Tree

- Random Forest:

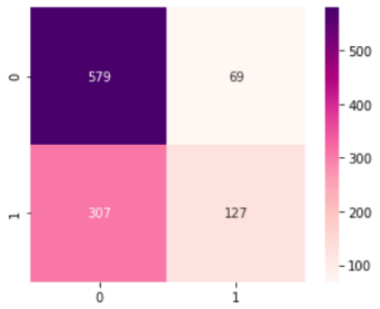

Figure 8: Confusion matrix of Random Forest

- Random Forest(Hyper Tuning):

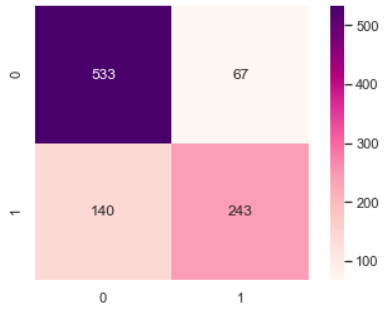

Figure 9: Confusion matrix of Random Forest(Hyper Tuning)

- Random Forest(Balancing data):

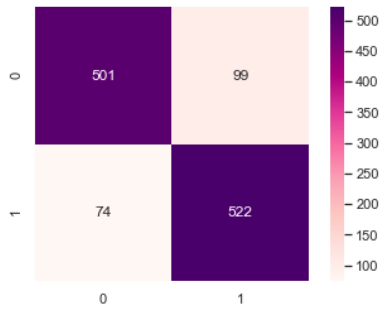

Figure 10: Confusion matrix of Random Forest(Balancing data)

We also evaluated the performance by noticing the precision , recall, f1 score of the algorithms. We saw that among all the algorithms random forest gives us the highest accuracy as well as the F1 score. But, in order to get that accuracy we had to balance the records of target attribute and also do the hyper tuning where after many cross validations it should us a better accuracy than the other modes based algorithms.

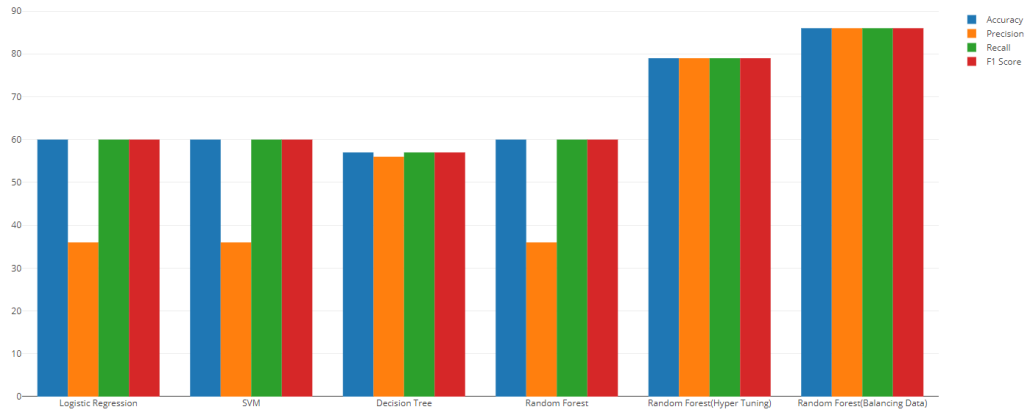

Figure 11: Overall performance of the ML algorithms

From the above figure we can see that, the precision , recall and the F1 score of the last two that means random forest with hyper tuning and random forest by balancing the data we found a good accuracy for our model.

We can also measure the model's performance by calculating True Positive Rate(TPR), True Negative Rate(TNR), False Positive Rate(FPR) and False Negative Rate(FNR). If we see the model is giving the TPR and TNR high and FPR and FNR low we can say that it has a better performance rate. In the case of logistic regression , decision tree and support vector machine, TNR and TPR is very low which indicates its bad performance. On the other hand, if we see the rate of the three random forest algorithm,

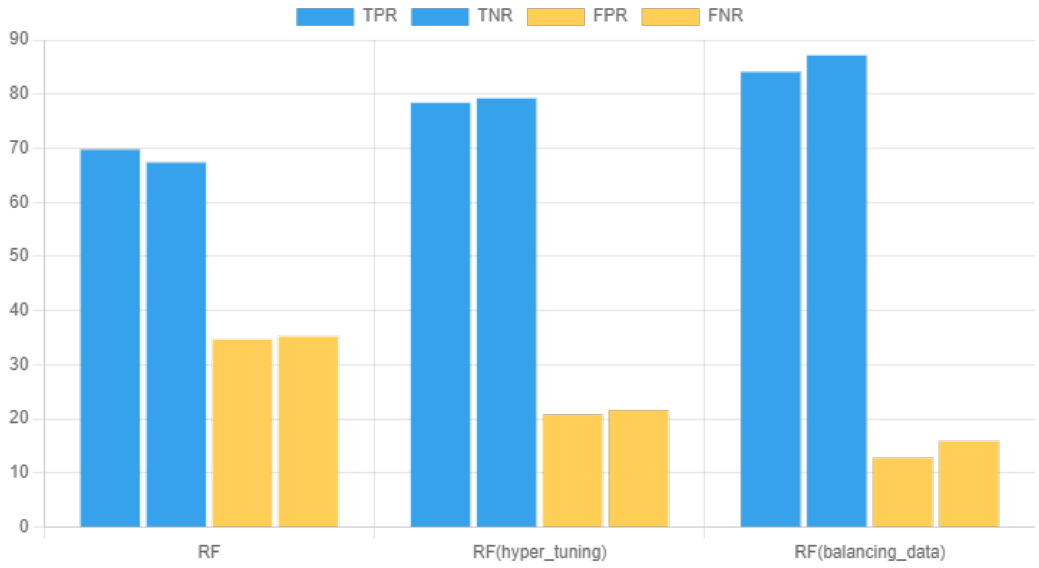

Figure 12: Relative rate of TPR,TNR , FPR, FNR

From the above chart we can also ensures that the random forest by balancing the data gives us a good accuracy competitively than others.

5 Conclusion

In this study, we have analysed the water quality indexes to find the potable and not potable water in our surroundings. We saw that water has many features and matters which changes its appearance. By analysing those matters and parameters we have build a model which give a 85 percent of accuracy in identifying potable and not potable water. As water is an essential part of our life, we need to make sure we are drinking safe water. Water potability analysis is useful test for detecting water contamination by harmful chemicals or pathogens in order to take preventive measures for the safe drinking water among population. Our work represents a brief study and its main aim is to detect water quality. The results carried out from the study concludes that using unsupervised learning, data with variation can be predicted at acceptable accuracy rate. To be exact, random forest gives us higher accuracy than any other algorithm that we used if we just tune the data, cross validate the data and also by balancing the data. Our model can also help preventing many waterborne disease. Our limitation can be that many algorithm with perfect sensor accessibility can produce more efficient accuracy in detecting the water quality. So, our future work will be to increase the performance measure more than 95 percent which will make sure the correct analysis of detecting water quality and also it will be beneficial for the people.

References

[1] Umair Ahmed, Rafia Mumtaz, Hirra Anwar, Asad A Shah, Rabia Irfan, and José García-Nieto. Efficient water quality prediction using supervised machine learning. *Water*, 11(11):2210, 2019.

[2] Umair Ahmed, Rafia Mumtaz, Hirra Anwar, Asad A Shah, Rabia Irfan, and José García-Nieto. Efficient water quality prediction using supervised machine learning. *Water*, 11(11):2210, 2019.

[3] Satinder Ahuja. *Monitoring water quality: Pollution assessment, analysis, and remediation*. Newnes, 2013.

[4] Maqbool Ali and Ali Mustafa Qamar. Data analysis, quality indexing and prediction of water quality for the management of rawal watershed in pakistan. In *Eighth International Conference on Digital Information Management (ICDIM 2013)*, pages 108–113. IEEE, 2013.

[5] David Byer and Kenneth H Carlson. Real-time detection of intentional chemical contamination in the distribution system. *Journal-American Water Works Association*, 97(7), 2005.

[6] Nabeel M Gazzaz, Mohd Kamil Yusoff, Ahmad Zaharin Aris, Hafizan Juahir, and Mohammad Firuz Ramli. Artificial neural network modeling of the water quality index for kinta river (malaysia) using water quality variables as predictors. *Marine pollution bulletin*, 64(11):2409–2420, 2012.

[7] Abul Hussam. Potable water: Nature and purification. *Monitoring Water Quality*, pages 261–283, 2013.

[8] Azra Jabeen, Xisheng Huang, Muhammad Aamir, et al. The challenges of water pollution, threat to public health, flaws of water laws and policies in pakistan. *Journal of Water Resource and Protection*, 7(17):1516, 2015.

[9] Ping Liu, Jin Wang, Arun Kumar Sangaiah, Yang Xie, and Xinchun Yin. Analysis and prediction of water quality using lstm deep neural networks in iot environment. *Sustainability*, 11(7):2058, 2019.

[10] Sorayya Malek, Mogeeb Mosleh, and Sharifah M Syed. Dissolved oxygen prediction using support vector machine. *Int. J. Bioeng. Life Sci*, 8:46–50, 2014.

[11] Fitore Muharemi, Doina Logofătu, and Florin Leon. Machine learning approaches for anomaly detection of water quality on a real-world data set. *Journal of Information and Telecommunication*, 3(3):294–307, 2019.

[12] Archana Solanki, Himanshu Agrawal, and Kanchan Khare. Predictive analysis of water quality parameters using deep learning. *International Journal of Computer Applications*, 125(9):0975–8887, 2015.

[13] Yunrong Xiang and Liangzhong Jiang. Water quality prediction using ls-svm and particle swarm optimization. In *2009 Second International Workshop on Knowledge Discovery and Data Mining*, pages 900–904. IEEE, 2009.

[14] Jin Zhang, Xiaohui Zhu, Yong Yue, and Prudence WH Wong. A real-time anomaly detection algorithm/or water quality data using dual time-moving windows. In *2017 Seventh international conference on innovative computing technology (INTECH)*, pages 36–41. IEEE, 2017.

[15] Md Sabab Zulfiker, Nasrin Kabir, AA Biswas, P Chakraborty, and Md M Rahman. Predicting students' performance of the private universities of bangladesh using machine learning approaches. *International Journal of Advanced Computer Science and Applications*, 11(3):672–679, 2020.