H515- Data Analytics

Project Report

Project Title: Water Quality Prediction using ML Algorithm
Group 8
Group Members: Palak Jain, Akshaya Ramesh, Vijay Mittal, Raj Zaveri
Advisor: Prof. William Fadel

1 Introduction

Water has always been one of most essential element in the world as it has a direct impact on public health as well as environment. This essential element is provided by the mother nature in abundance in the form of water bodies like rivers, lakes,oceans,streams etc but every source has specific water quality standard which help us figuring out its usage and application in different sectors. So water quality is very important taking into consideration its uses for various practices, such as drinking,industrial,agricultural,households use and many more.
Due to its wide usage based on quality, modeling techniques in predicting water quality are becoming more and more popular with their importance in controlling water pollution and knowing its significance.This kind of promising research can contribute significantly to water management.These highly efficient models can be generalized and used to forecast the water pollution process which will help the decision-makers to make the right decisions at the right time [1].

2 Background Work and Research

There has been a lot of Background work and research as well going on this.Researchers have mostly emphasize enhancing the applicability and reliability of water quality prediction/modelling by using a variety of new technologies such as Fuzzy logic, stochastic, ANN, and deep learning [2,3].
Shafi et al. [5] has proposed four machine learning algorithms, which are Support Vector Machines (SVM), Neural Networks (NN), Deep Neural Networks, and k-Nearest Neighbors (kNN), for the predicting the quality of water and also using single feed-forward neural networks to classify water quality.
Ranković et al. [6] estimated the dissolved oxygen using the ANN model.
Gazzaz et al. [7] calculated the water quality index by using an ANN model, and the Internet of Things technology was applied to collect the dataset from water resources.
Abyaneh [8] used machine learning methods like ANN and regression to predict the chemical oxygen demand .
In addition to this , it has been reported that deep learning techniques give a much better performance in predicting the quality of water when compared to the traditional methods.

3 Problem Description

From the past few years, water quality has been degrading and threatened because of various pollutants, industrial waste, release of non-compostable materials and chemicals in the water bodies and many other human intervention and this in turn has led to lack of availability of clean water becoming a more and more serious problem. There is a huge possibility of getting contaminated water if we directly take it from water bodies.For example, water used for irrigation purposes should be clean with no chemicals so that they don't harm the plants or soil.Industrial use requires different properties based on the specific industrial processes. On the Other hand , effects of polluted drinking water can affect human health, the environment,and infrastructures.In a report by United Nations (UN),it says almost one and half million people die each year because of contaminated water-driven diseases[4].
Therefore, there is a necessity to check water quality standards before use. Everyone should get safe water to drink in order to stay healthy. In this project, we have determined water potability based on various other features such as ph, hardness,amount of solids, sulfate, chloramines and few others. We have tried to solve this water potability problem using Machine Learning classification models. We have basically tried to determine if the water is potable or not based on various other factors such as water pH value, sulphate level in water etc.

4 About the Dataset

The dataset is obtained from kaggle community website:
https://www.kaggle.com/datasets/adityakadiwal/water-potability
Measurements: ppm: parts per million g/L: microgram per litre mg/L: milligram per litre
Column descriptions:
1. ph: pH of 1. water (0 to 14).
2. Hardness: Capacity of water to precipitate soap in mg/L.
3. Solids: Total dissolved solids in ppm.
4. Chloramines: Amount of Chloramines in ppm.
5. Sulfate: Amount of Sulfates dissolved in mg/L.
6. Conductivity: Electrical conductivity of water in S/cm.
7. Organic carbon: Amount of organic carbon in ppm.
8. Trihalomethanes: Amount of Trihalomethanes in g/L.
9. Turbidity: Measure of light emiting property of water in NTU.
10. Potability: Indicates if water is safe for human consumption. Potable -1 and Not potable -0

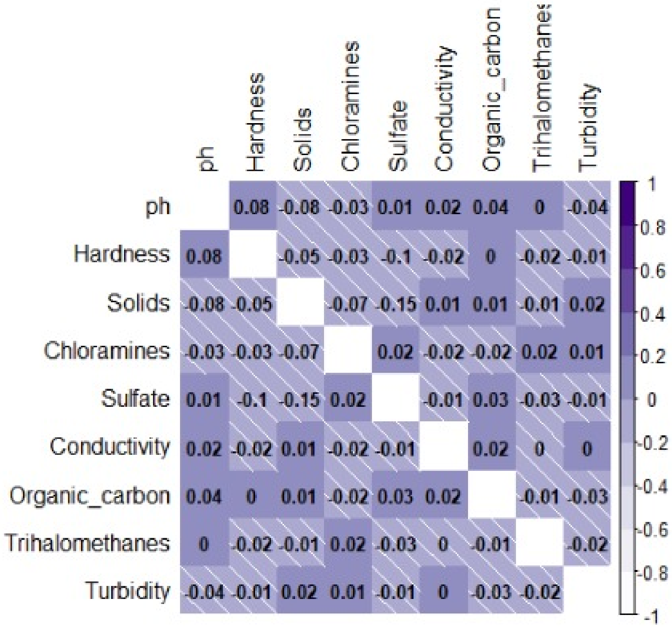
5 Methods

5.1 Pre-Processing

The Following PreProcessing / cleaning has been done after initial exploratory data analysis.
1. Oversampling done for a class with lower number of records to balance the dataset.
2. Removed records having more than one Null value.
3. Replaced values with mean for records having Null value only one column.

5.2 Exploratory Data Analysis

We did a initial exploratory data analysis to see the correlation between different features available to us as a part of the data set. We have tried representing the analysis results in the matrix below in Figure1. We can see that none of the variables have significant correlation between them



	ph	Hardness	Solids	Chloramines	Sulfate	Conductivity	Organic_carbon	Trihalomethanes	Turbidity
ph		0.08	-0.08	-0.03	0.01	0.02	0.04	0	-0.04
Hardness	0.08		-0.05	-0.03	-0.1	-0.02	0	-0.02	-0.01
Solids	-0.08	-0.05		-0.07	-0.15	0.01	0.01	-0.01	0.02
Chloramines	-0.03	-0.03	-0.07		0.02	-0.02	-0.02	0.02	0.01
Sulfate	0.01	-0.1	-0.15	0.02		-0.01	0.03	-0.03	-0.01
Conductivity	0.02	-0.02	0.01	-0.02	-0.01		0.02	0	0
Organic_carbon	0.04	0	0.01	-0.02	0.03	0.02		-0.01	-0.03
Trihalomethanes	0	-0.02	-0.01	0.02	-0.03	0	-0.01		-0.02
Turbidity	-0.04	-0.01	0.02	0.01	-0.01	0	-0.03	-0.02	

Figure 1: Feature Correlation matrix

Also, we created a pie chart showing the distribution of data and check if we have enough number of records for each class to understand our data is balanced or not. We found that we have around 39 percent of records with class 1 (in our case potability =1) and 61 percent with class 0 (potability =0). so clearly it depicts that the data is quite unbalanced and there is a high chance that any model would be biased and will predict the dominating class

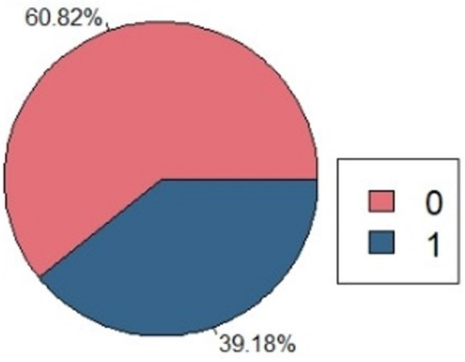


Figure 2: Potability Imbalance

5.3 Modelling Methods

To determine potability from diverse features, we used different machine learning classification models starting from Logistic regression, SVM, QDA, K-nearest classifier to ensemble methods. As our data was imbalanced, we implemented all the above mentioned models with the original data and also with the over sampled balanced data to check the effect of oversampling on model accuracy.

Model 1 : Logistic Regression

Logistic Regression is a classification model which is used to predict the probability of categorical function. Let's say our input for training this model is (y1, x1), . . . (yn, xn) where y1, y2...yn are either 0 or 1 for potability and x1, x2... xn is the combination of all other features. If our test input is x' , the output from logistic regression will be P(1/x') i.e. the probability of getting 1 if input is x' . If probability is greater than 0.5, final output will be class 1 (in our case 'Potable') and if probability is less than 0.5, final output will be 0 (in our case 'Non-potable'). This is the first approach we experimented on and in most of the cases, this method works good as well. But in our project, data was imbalanced in terms of water potability. Therefore, this method did not work for our data as it was predicting only class 0 for any given input and therefore, precision is 1 and recall is very low.
On the other hand, when we used oversampled data, precision and recall were balanced but not good.

Model 2 : QDA

Quadratic Discriminant Analysis (QDA) is very similar to Linear discriminant analysis to find separation boundaries for the data. The only difference is that in QDA, we assume that the covariance matrix can be different for each class and therefore, we will estimate the covariance matrix separately for each class k, k =1, 2, ... , K. Decision boundaries are quadratic for QDA. For classification, we just have to find the class k which maximizes the quadratic discriminant function which is represented by-

$$\delta_k(x) = -\frac{1}{2}\log|\Sigma_k| - \frac{1}{2}(x-\mu_k)^T\Sigma_k^{-1}(x-\mu_k) + \log\pi_k$$

For QDA, we did get somewhat balanced precision and recall for both unsampled and oversampled training data.

Model 3 : SVM

Support vector machines are very useful when we want to deal with high dimensional spaces and non linear decision boundaries for classification. It has also been proved that in some cases, this model works very well even if number of dimensions is greater than number of samples. So, after logistic regression and QDA, we tried SVM for our unbalanced input data. But it was no better than the logistic regression.

Model 4 : K-nearest classifier

Another method we used for determining potability of water is K-nearest neighbour. Sometimes, we just make our model so complicated unnecessarily and it can be solved just by looking at it's neighbours.
The KNN, K-nearest neighbour algorithm assumes that all similar things are near to each other. We can classify any target point by looking at it's k nearest points and decide classify target as 1 if majority of it's k neighbours are 1 and 0 if majority is for 0. We tried with different values of k and found that k=15 is giving the best model among all given values of k. Therefore, we considered k=15 for our final k-nearest classifier.

Model 5 : Random Forest

Decision trees are built in the form of tree with nodes where branches depend on a number of factors. It takes one feature as node(based on Gini impurity value) and divide the whole data into branches and this process continue till it achieves a threshold value. This threshold can be length of tree or impurity value or any other tree parameter. Although decision trees are very sensitive and has high variance. To reduce this problem, ensemble methods have been used many times which is basically combining more than one model together in parallel or series and then take the combined output as the final output from the model.
Random forest is also an ensemble technique which consists of a large number of individual decision trees in parallel. It gives a part of training data to each decision tree and then combine their output (consider majority in case of classification problem). Figure 3 shows the feature importance of each variable in random forest and based on that, e considered top 5 features for our final random forest model.
One of the biggest advantage of ensemble method is that it can handle imbalanced data easily. Therefore, we were also hopeful with this model and yes, it worked pretty well in comparison to all other models. At the end, we implemented 10 fold cross validation in random forest model to improve it.

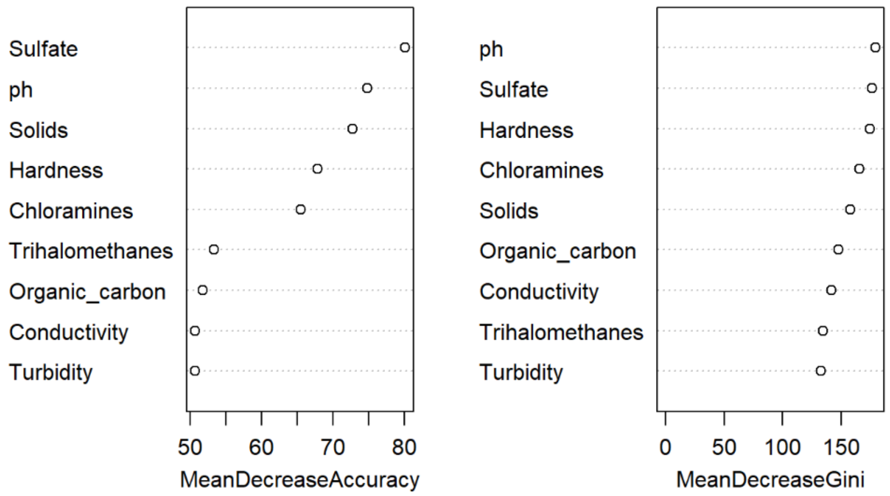
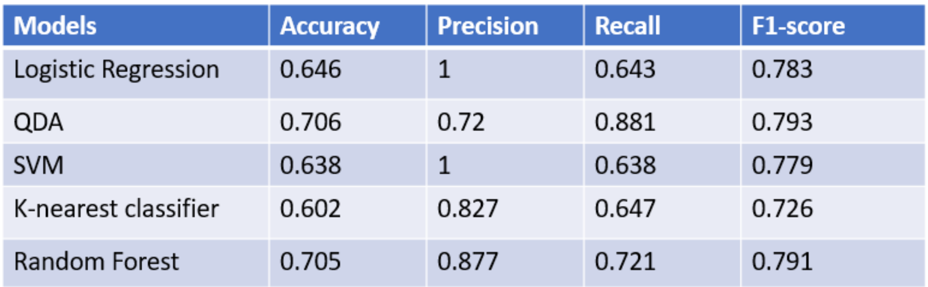


Figure 3: Feature Importance for Random Forest

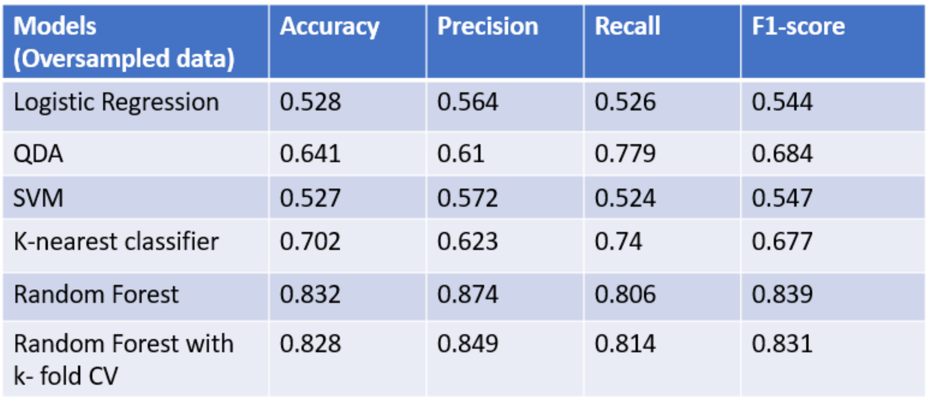
6 Results

On comparing various models, we found that the random forest algorithm worked best to determine potability of water even if the data is unbalanced. In case of random forest, precision and recall, both were balanced and high for unbalanced data [table 1] as well as over sampled data [table 2].
On the top of that, it is a good practice to use cross validation techniques to deal with unbalanced data. So, we implemented k-fold cross validation for the random forest to verify our oversampling method. Precision and recall, both are nearly same for random forest with cross validation and without cross validation. So, we can conclude that our oversampling method has created a proper balance in the data and therefore, we have 'Random Forest with over sampled data' as our final model with accuracy of 0.832, precision 0.874, recall 0.806 and f1 score of 0.839.



Models	Accuracy	Precision	Recall	F1-score
Logistic Regression	0.646	1	0.643	0.783
QDA	0.706	0.72	0.881	0.793
SVM	0.638	1	0.638	0.779
K-nearest classifier	0.602	0.827	0.647	0.726
Random Forest	0.705	0.877	0.721	0.791

Table 1: Evaluation parameters for unbalanced data



Models (Oversampled data)	Accuracy	Precision	Recall	F1-score
Logistic Regression	0.528	0.564	0.526	0.544
QDA	0.641	0.61	0.779	0.684
SVM	0.527	0.572	0.524	0.547
K-nearest classifier	0.702	0.623	0.74	0.677
Random Forest	0.832	0.874	0.806	0.839
Random Forest with k- fold CV	0.828	0.849	0.814	0.831

Table 2: Evaluation parameters for oversampled data

The below figure (fig 3) shows the visual comparison using bar chart of all the different models that we have used and we can see that precision,accuracy,f1 score and recall are higher for the random forest in comparison to all other classification models.

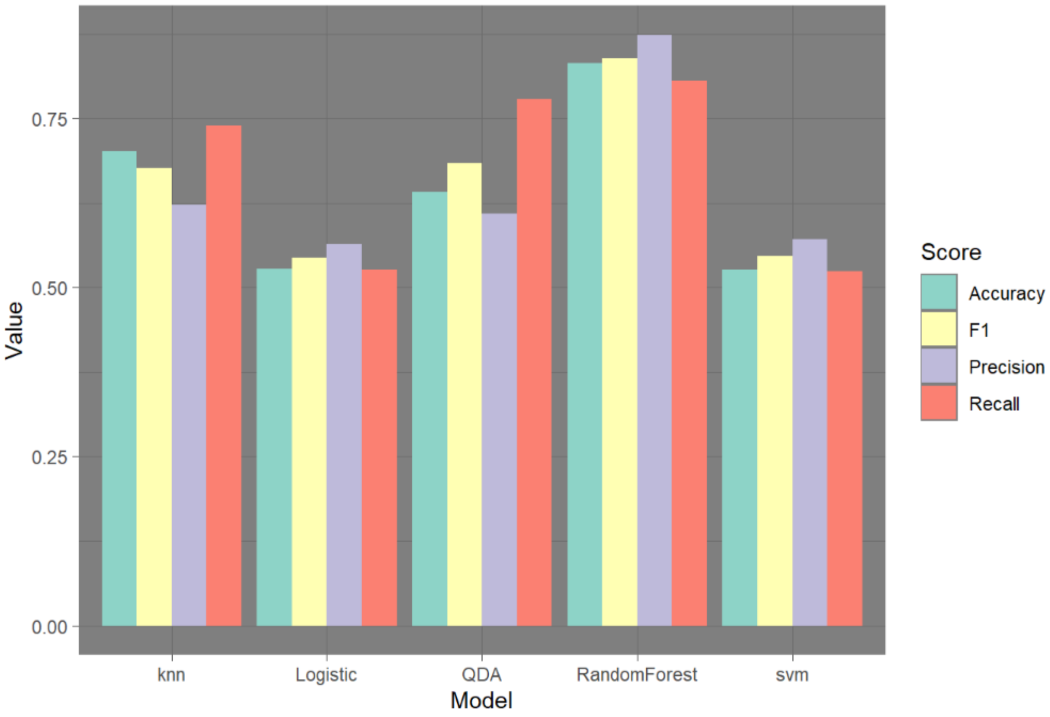


Figure 4: Comparison of different models

7 Discussions and Future work

We come to the conclusion that modelling techniques for predicting water quality are very important tools and are the need of the hour to protect our environment. For future work , using advanced artificial intelligence algorithms ,we can even measure the future water quality. These modeling techniques to predict the quality of water should be promoted and practiced worldwide.

In addition to this, we have built our models on a very small set of data, so there is definitely a scope of much better and larger data set to create more accurate models.

To see the combined effect of different features, we can even use deep learning methods in the future with more data and features. Currently we have considered only 9 features while building the prediction and the performance of these models can greatly enhanced by adding more feature to data set.

8 References

[1] 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

[2] H. R. Maier, A. Jain, G. C. Dandy, and K. P. Sudheer, "Methods used for the development of neural networks for the prediction of water resource variables in river systems: current status and future directions," Environmental Modelling Software, vol. 25, no. 8, pp. 891–909, 2010.

[3] S. Lee and D. Lee, "Improved prediction of harmful algal blooms in four major South Korea's rivers using deep learning models," International Journal of Environmental Research

and Public Health, vol. 15, no. 7, p. 1322, 2018.

[4] UN water, “Clean water for a healthy world,” Development, 2010, https://www.undp.org/content/undp/en/home/ presscenter/articles/2010/03/22/clean-water-for-a-healthyworld.html.

[5] U. Shafi, R. Mumtaz, H. Anwar, A. M. Qamar, and H. Khurshid, “Surface water pollution detection using internet of things,” in 2018 15th International Conference on Smart Cities: Improving Quality of Life Using ICT IoT (HONET-ICT), pp. 92–96, Islamabad, Pakistan, October 2018.

[7] V. Ranković, J. Radulović, I. Radojević, A. Ostojić, and L. Čomić, “Neural network modeling of dissolved oxygen in the Gruža reservoir, Serbia,” Ecological Modelling, vol. 221, no. 8, pp. 1239–1244, 2010.

[8] N. M. Gazzaz, M. K. Yusoff, A. Z. Aris, H. Juahir, and M. F. Ramli, “Artificial neural network modeling of the water quality index for Kinta River (Malaysia) using water quality variables as predictors,” Marine Pollution Bulletin, vol. 64, no. 11, pp. 2409–2420, 2012.

[9] H. Z. Abyaneh, “Evaluation of multivariate linear regression and artificial neural networks in prediction of water quality parameters,” Journal of Environmental Health Science and Engineering, vol. 12, no. 1, p. 40, 2014.

[10] Kadiwal, A. (2021, April 25). Water quality. Kaggle. Retrieved May 5, 2022, from https://www.kaggle.com/datasets/adityakadiwal/water-potability

[11] A. Liaw and M. Wiener (2002). Classification and Regression by randomForest. R News 2(3), 18–22

[12] H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016

[13] Taiyun Wei and Viliam Simko (2021). R package ’corrplot’: Visualization of a Correlation Matrix (Version 0.92). Available from https://github.com/taiyun/corrplot

[14] Greg Snow (2020). TeachingDemos: Demonstrations for Teaching and Learning. R package version 2.12. https://CRAN.R-project.org/package=TeachingDemos

[15] David Meyer, Evgenia Dimitriadou, Kurt Hornik, Andreas Weingessel and Friedrich Leisch (2021). e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien. R package version 1.7-9. https://CRAN.R-project.org/package=e1071

[16] Max Kuhn (2022). caret: Classification and Regression Training. R package version 6.0-92. https://CRAN.R-project.org/package=caret

[17] Venables, W. N. Ripley, B. D. (2002) Modern Applied Statistics with S. Fourth Edition. Springer, New York. ISBN 0-387-95457-0

[18] Nicola Lunardon, Giovanna Menardi, and Nicola Torelli (2014). ROSE: a Package for Binary Imbalanced Learning. R Journal, 6(1), 82-92.

[19] Stephen Milborrow (2021). rpart.plot: Plot ’rpart’ Models: An Enhanced Version of ’plot.rpart’. R package version 3.1.0. https://CRAN.R-project.org/package=rpart.plot

[20] Hadley Wickham, Romain François, Lionel Henry and Kirill Müller (2022). dplyr: A Grammar of Data Manipulation. R package version 1.0.8. https://CRAN.R-project.org/package=dplyr

[21] Tierney N (2017). ”visdat: Visualising Whole Data Frames.” JOSS 2(16), 355. doi: 10.21105/ joss.00355