Water quality prediction using machine learning models

1st Khanani Mathebula
School of Computer science and Applied mathematics
University of the Witwatersrand
Johanesburg, South Africa
1847799@students.wits.ac.za

2nd Dr Ritesh Ajoodha
School of Computer science and Applied mathematics
University of the Witwatersrand
Johanesburg, South Africa
Ritesh.Ajoodha@wits.ac.za

3rd Dr Ashwini Jadhav
Science Teaching and Learning Unit
University of the Witwatersrand
Johanesburg, South Africa
Ashwini.Jadhav@wits.ac.za

***Abstract*—Water is essential for the livelihood of all living organisms, thus it is deemed important to know if certain catchments can be used for consumption.This presents a need for an efficient methods for predicting water quality.This paper proposes machine learning models by comparing most investigated models in water prediction. This paper predict water quality over a dataset.Data processing is done on the dataset. subsequently water quality index (WQI) is calculated using water parameters .water quality classification(WQC) is obtained by grouping the results of WQI values into 3 classes(good ,poor,very poor).SMOTE is then used to balance data on Target column. Classifies investigated are ,Random forest, Naïve bayes, Multi linear regression, Decision tree and K-nearest neighbour.After using evaluation metrics, random forest(RF) was the best performing model in classifying water quality classes.**

***Index Terms*—Machine learning,water quality, classification, parameters, ,Prediction,Random forest, Naïve Bayes, Multi linear regression, Decision tree, K-nearest neighbour,SMOTE**

I. Introduction

Water is essential for the survival of all living organisms. Everyone has the human right to safe drinking water. This holds true in stability and in crisis, in urban and rural contexts, and in every country around the world [1]. Without water, humans can survive only for a days. Water comprises from 75 % body weight in infants to 55% in elderly and is essential for cellular homeostasis and life [11].It is thus important to measure the potability of water. Water quality Is the measurement of how suitable water is for a specific beneficial use. It is based on the biological, chemical and physical properties of water. There are certain factors that affect water quality ,the introduction of invasive aquatic plants , human pollution e.g industrial waste water and seasons just to mention are few have a direct impact on water quality. Water quality does not only refer to how clean or dirty water is but also on whether beneficial uses for things that rely on water can be fulfilled. Traditional water quality measuring and calculation are not efficient. This traditional testing has many shortcomings, but the safety of human water and the balance of aquatic ecosystems are such important and urgent issues that we need to solve [14].Most testing have to be done in laboratories and are labor-intensive and take time.Water quality is always a matter of agency for water is used daily and thus a real time solution is a better alternative.Machine learning proposes a solution to traditional methods. The use of Machine learning algorithms to solve the problem of labor-intensive and efficiency. The reduction of labor and time results in an inexpensive alternative which is preferable for underdeveloped and developing countries.

Our aim is to explore the classification of water parameters. We do so by investigating whether some of the top investigated machine learning models can classify water parameters as a method to predict water quality

This is deemed important to have a real time prediction of water quality for the health of aquatic life, and other living organisms. we believe this research will help in the management of water.

/

II. Related work

Conventional water quality assessment methods using WQI can be time-consuming and expensive, especially for complex datasets with multiple water quality parameters [12].Machine learning techniques are a better alternative to traditional laboratory water quality computation. As it can be deduced from the above statement, Machine learning models propose a cheaper and more efficient solution to the prediction and classification of water quality index. [9]The article generates WQI using the 8 best subset regression most used AI models. The article investigates how changing the sample size can affect

the performance of a classification model. The computation of the WQI tends to computationally expensive and gives errors when subindex is calculated. WQI estimate models such as ANN, MLR, SVM, M5P tree, RF, LWLR, RS, and AR were suggested. The author found that decreasing the sample size directly influences the classification performance of models, where the goal in the second scenario was to construct which model performed best under such settings. Out of all the suggested models MLR AND RF algorithms are seen to be more suitable for developing countries due to it being cost effective in computing WQI when given all the input.

[9] did a similar study on 7 traditional AI models and 3 newly ensemble learning models completely-random tree forest (CTF),[random 1 forest (RF), and deep cascade forest(DCF)) [5]. Traditional algorithms increased performance when data set was increased while the 3 emerging learning remained the same. DCF and CRF showed better prediction of water quality when weight f1-score were compared.The finding of the study was that big data could elevate the overall predictive analysis of both learning models investigated in the study [8] [2] [13] [6] [10] Authors investigated 5 of the models discussed in this paper separately.] [2] [13] [6] found that RF and DT are very efficient in predicting water quality with a confidence score of 93.75 % and 95.4545% respectively and that MLR is very cost efficient when compared to others which are the same conclusions [5] made. [7] compiled a comparative analysis on the prediction and classification of water quality. The dataset used include important parameters: DO, biological oxygen demand,temp, nitrate,conductivity, fecal coliform , total coliform and pH [4]. neuro-fuzzy inference system (ANFIS) algorithm was used as a prediction algorithm for WQI ,predicted the water quality index with a 96.17 % regression coefficient. K-nearest neighbors and Feed-forward neural network (FFNN) were used as classification models when FFNN produced 100 % accuracy.

III. Methodology

A. preferred approach

-perform data pre-processing to account for missing values and outliers -use water parameters to calculate WQI(water quality index) and determine the WQC - Train 5 machine learning algorithms to accuracy predict WQI and classify WQC(potability) -Evaluate the performance of the models.

B. Study area and collection:

The dataset used to conduct this research was obtained from Kaggle. The data was sampled from different 666 water catchment(rivers and lakes) across India over a period of 9 years between 2005to 2014. Dataset contains 1679 samples and overall 8 water parameters. Namely temp, pH, DO, biological oxygen demand , fecal coliform, nitrate, biological oxygen demand and total coliform.

The dataset has most of parameters known to affect water quality.The dataset is seen as a reliable source.

C. Data pre-processing:

- calculate the mean and median of each value and use them to account for missing values.
- z-score normalization: data is normalised by considering the mean and standard deviation of given data. Z-score normalization is carried out to account for outliers.
- Data prepossessing is deemed crucial for the quality of data to improve [3].
-From the reprocessed Dataset, Significant parameters are used in calculating WQI . Then, water samples have been classified on the basis of the WQI values [3]

z score is stastistical defined as:

$$Z = \frac{x - \mu}{\sigma} \tag{1}$$

D. Class imbalance:

From fig.1 we can see that the dataset has class imbalances. Synthetic Minority Oversampling technique is used to address it. The approach is to balance the sample size of minority and majority class.This is achieved by iterative process of setting total number of samples to a an integer, randomly picking an instance of a minority class.close neighbor of the instance is then found.This iteration is repeated until more samples are classed into respective classes.

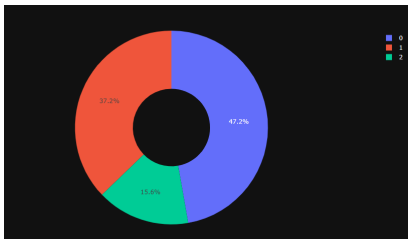

Fig. 1: Example of a figure caption.

Water quality index calculation:

Below is the mathematical equation that describes water quality

$$WQI = \frac{\sum_{i=1}^{N} q_i w_i}{\sum_{i=1}^{N} w_i} \tag{2}$$

Where q_i is the quality rating of i parameters .N is the sum of all parameters used to calculated the wqi and w_i is the unit weight of each i parameter.

$$q_i = 100 \frac{V_i - V_{ideal}}{S_i - V_{ideal}} \tag{3}$$

V_i is the measured value parameter ,whilst V_ideal is the tested value.S_i is the standard value of each parameter.

$$w_i = \frac{K}{S_i} \tag{4}$$

Where K is the Proportionality constant calculated as below

$$K = \frac{1}{\sum_{i=1}^{N} S_i} \tag{5}$$

E. Parameters:

1) Temperature:: Influences water quality due to its properties when the temperature is increased or decreased. Chemical reaction tends to increase with higher temperatures and dissolved oxygen levels are observed to be higher in lower temperature.

2) PH: defined by the concentration of hydrogen ion in a solution. PH measures how acidic or alkaline water is.

3) Nitrate: Nitrite in high concentration is harmful and acts as a catalyst in increasing aquatic plant growth(e.g. hyacinth) due to it's contribution on eutrophication. When the levels are low it serves as an aquatic plant nutrient.

4) Dissolved oxygen: Indicated the level of oxygen found in water catchments. The higher the level(indicated in values) of oxygen the better the water quality measured.

5) Conductivity: The conductivity is used as tool to indicate the presence of salt, sulphides, chlorides and etc. The measurement is conducted by measuring electrical current that water allows to pass through it.

6) Biological oxygen demand: The measurement of the amount of oxygen needed during the process of decomposition by microbes. This parameter is essential for it helps eliminate waste from water.

7) Fecal coliform: This parameter only looks at one factor with coliform bacteria that affect water quality. It regards only the bacteria produced by warm blooded animals.

8) Total coliform: This is a bacteria that is found in most living organisms e.g. the digestive tract of animals. This parameter totals the coliform bacteria present in water due to humans, animals, surface water and soil material.

TABLE I: parameter constants

Parameters	Si(limits)	weights (wi)	Videal
Total coliform/100 mL	1000	0.0022	0
pH	8.5	02604	7.0
Nitrate, mg/L	45	0.0492	0
Conductivity, µS/cm	1000	0.0022	0
Biological oxygen demand, mg/L	5	0.4426	0
Dissolved oxygen, mg/L	10	02213	14.6
Fecal coliform/100 mL	100	0.0221	0

Water quality classification (WQC) is determined using the above calculated WQI. WQI values are divided into classes using range in table II. Normally there is 4 classes but due to our dataset limitations that will be discussed later in the paper , the classes are scaled down to 3.

F. classification models:

The machine learning algorithms listed below as discussed in II were investigated by other authors and found to be the best in classification, to be specific in the classification of water quality. Some of the factors that were looked at were sample size, efficiency and cost.

TABLE II: classification classes

Water quality index	Range Classification
0-25	excellent
26-59	Good
51-75	poor
< 75	Very poor

MLR: Multi linear regression is derived from linear regression but differ in the evaluation.MLR assumes that there is a linear relationship amongst variables and attempt to model the relation. Water quality as the dependent variable is affected by multiple water parameters in given dataset thus MLR is used as one of the classification models in this research. RF:

Random forest is a supervised learning algorithm that merges decision trees in an effect to produce more accurate results. The fundamental of this algorithm is based on randomness as namesake. The sample at the beginning is selected at random and to construct a decision tree, several,attributes are chosen at random. The above 2 random process are repeated until a number of decision trees are constructed. The results are given by taking the average of the constructed decision trees.

DT: A real life analogy of decision trees , would be a tree that is upside down. A decision trees comprises of the top node as a root node, decision nodes(features) and Leaf nodes(possible outcomes all dataset) that are held together by branches. This particular algorithm takes a greedy approach to search for points and classify them recursively into desired class labels.

KNN:

Supervised learning classifier.Can be used for regression and classification.The aim of this algorithm is to assign a class label to a point by identifying the closest neighbor of that query point The algorithms for classification assumes that similar points a likely to be found near one another. Distance metrics (e.g Euclidean distance and Manhattan distance) are used to calculate the distance between query point and other points by grouping this points into different regions. NB: Takes

a probabilistic approach to classify water parameters. This model is based on the Bayes theorem. The algorithm assumes that all features are independent of one another.

$$p(A|B) = \frac{P(A|B)P(A)}{P(B)} \tag{6}$$

Description of above Bayes theorem: A and B are events. P(A),P(B) the probability of independent A or B occurring , P(A—B) the probability if A given be.

The naïve Bayes allows us to calculate posterior probability given below:

$$p(c|x) = \frac{P(x|c)P(c)}{P(x)} \tag{7}$$

It is clear that the above equation is derived from Bayes theorem [6] Thus, the Bayes theorem definition can be used to understand the naive Bayes equation and algorithm as a whole.

G. *Evaluation metrics:*

confusion matrix to record a predicted model's performance during classification.s. They encode the complete specification of misclassifications: the numbers of misclassified items for each pair original class in which items should be classified, incorrect class in which items are erroneously classified [4].Confusion matrix groups true positive(TP),true negative(TN), false positive (FP) ,and false negative(FN)in a matrix for easy visualisation.

From the matrix the accuracy , precision, recall and f-score are calculated.Below are equations used to calculate aspects of the matrix.

$$Precision = \frac{TP}{TP + FP} \tag{8}$$

$$Recall = \frac{TP}{TP + FN} \tag{9}$$

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{10}$$

$$F1 - score = \frac{2}{\frac{1}{Recall} + \frac{1}{Precision}} \tag{11}$$

H. *Ethics clearance :*

No ethical clearance is required for this research

IV. Results and Discussion:

This section presents results that were obtained from the preferred approach.

1) Data processing and data imbalances:: The data Processing performed evaluated the quality of the data. Median and mean were used to fill missing values since the dataset used for this paper is not large enough to let go of data entries. Z score was used to remove outliers , dataset stringed from (1988,12) to (1862,12).

To obtained the WQI the discussed parameters and equation in III were used. From the results of the WQI columns, wqi values were used against the wqi range to determine the Water quality classification(WQC) as "C" as seen in fig.2 .

The dataset is split in ratio of 8:2 for training and testing respectively. SMOTE was then utilised to balance the WQC classes.the (0-25) class was only made up of 3% of the target column, it gave an error when passing it through SMOTE. We then decided to merge it with the (26-59) classes resulting in 3 classes as displayed in fig 1. SMOTE successfully balanced the data as observed in fig.3.

confusion matrix was used to measure the performance of the 5 classification models.We found that Naive Bayes performed the list with an accuracy score of 58% and random forest out performed all models with an accuracy score of

	Temp	D.O. (mg/l)	PH	CONDUCTIVITY (µmhos/cm)	B.O.D. (mg/l)	NITRATENAN N+ NITRITENANN (mg/l)	FECAL COLIFORM (MPN/100ml)	TOTAL COLIFORM (MPN/100ml)Mean	WQI	C
1857	27.0	7.9	738.0	7.2	2.7	0.518	0.518	202.0	12746.407333	0
1858	29.0	7.5	585.0	6.3	2.6	0.155	0.155	315.0	10091.343432	0
1859	28.0	7.6	98.0	6.2	1.2	0.516	221.000	570.0	1629.125767	0
1860	28.0	7.7	91.0	6.5	1.3	0.516	221.000	582.0	1508.008186	0
1861	29.0	7.6	110.0	5.7	1.1	0.516	221.000	546.0	1836.555177	0

Fig. 2: obtained WQI and WQC values.

```
Balancing the data by SMOTE - Oversampling of Minority level

Before SMOTE Counter({0: 669, 1: 498, 2: 225})

After SMOTE Counter({1: 669, 0: 669, 2: 669})
```

Fig. 3: data imbalances before and after.

91%.The rest of the models scores can be seen in table 4. Fig 4 is a confusion matrix visual representation of the best performing model RF and 195 is the True positive(TP) obtained from the model.

Table 3:Score %

Model	Train$_{A}ccuracy$	Test$_{a}ccuracy$
MLR	0.749868	0.759140
RF	0940692	0.909677
DT	0.895341	0.877419
KNN	0.737930	0.692473
NB	0.0630699	0.0.582796

Table 4:Model performance %

Model	Accuracy	Precision	Recall	F-score
MLR	0.76	0.92	0.87	0.89
RF	0.91	0.94	0.94	0.94
DT	0.87	0.89	0.93	0.91
KNN	0.71	0.87	0.77	0.82
NB	0.58	0.96	0.62	0.75

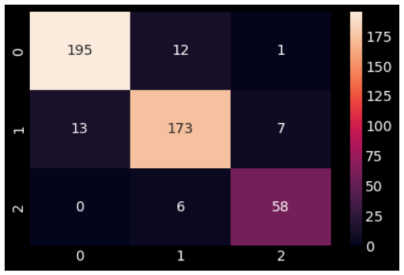

Fig. 4: data imbalances before and after.

V. Conclusions and Recommendations

In this paper water quality was predicted by firstly calculating WQI using water quality parameters. Additionally, WQI values obtained from the calculations were divided into 3 ranges that were used in the classification of water quality classes. After, 5 machine learning classification models (RF, MLR, DT, NB and KNN) were investigated. The top 2 out

of the 5 were random forest (RF) and Decision trees. They achieved overall 91 % and 87% accuracy score respectively.

Limitations experienced on this paper are the dataset sample size and validity of the data. We resorted to changing water quality range from(0-25) and (26- 59) to (0-59) due to the (0-25) class having less samples for data processing e.g data imbalances using SMOTE. The source of the data is used by most data scientist but anyone can compile a data set and upload it on kaggle.
furthermore as observed in section II machine learning models perform better with a greater sample data thus researchers should look or compile datasets that a much larger than the one explore on this paper. Obtaining water parameter datasets that a verified should be considered. Automation of the calculation of WQI is be considered , for the calculation method used on this paper is prone to human area subsequently if error exist it affect the validity of the research and when applied to the real world it could affect life if there's a misclassification of water quality and water is consumed. Lastly Machine learning models were seen to accurately predict water quality using water parameters

VI. Acknowledgements

I would like to thank both my supervisors Dr Ritesh Ajoodha and Dr Ashwini Jadhav for the support and guidance in this research

VII. References

References

[1] Water. https://www.unicef.org/wash/water. [Online; accessed 2022-11-22].

[2] Fowzia Akhter, Hasin Reza Siddiquei, Md Eshrat E. Alahi, and Subhas C. Mukhopadhyay. Recent Advancement of the Sensors for Monitoring the Water Quality Parameters in Smart Fisheries Farming. *Computers*, 10(3):26, feb 27 2021.

[3] Theyazn H. H Aldhyani, Mohammed Al-Yaari, Hasan Alkahtani, and Mashael Maashi. Water Quality Prediction Using Artificial Intelligence Algorithms. *Applied Bionics and Biomechanics*, 2020:1–12, dec 29 2020.

[4] Emma M.A.L. Beauxis-Aussalet, Joost Van Doorn, and Lynda Hardman. Supporting End-User Understanding of Classification Errors: Visualization and Usability Issues. *The Journal of Interaction Science*, 7:29, oct 9 2019.

[5] Kangyang Chen, Hexia Chen, Chuanlong Zhou, Yichao Huang, Xiangyang Qi, Ruqin Shen, Fengrui Liu, Min Zuo, Xinyi Zou, Jinfeng Wang, Yan Zhang, Da Chen, Xingguo Chen, Yongfeng Deng, and Hongqiang Ren. Comparative analysis of surface water quality prediction performance and identification of key water parameters using different machine learning models based on big data. *Water Research*, 171:115454, 3 2020.

[6] Sanjeev Gour and Mamta Gour. Study on Water quality of Narmada River by analyzing physicochemical and biological parameters using random forest model. *International Journal of Computer Sciences and Engineering*, 7(1):826–831, jan 31 2019.

[7] Mosleh Hmoud Al-Adhaileh and Fawaz Waselallah Alsaade. Modelling and Prediction of Water Quality by Using Artificial Intelligence. *Sustainability*, 13(8):4259, apr 12 2021.

[8] M. Ilić, Z. Srdjević, and B. Srdjević. Water quality prediction based on Naïve Bayes algorithm. *Water Science and Technology*, 85(4):1027–1039, jan 10 2022.

[9] Jiping Jiang, Sijie Tang, Dawei Han, Guangtao Fu, Dimitri Solomatine, and Yi Zheng. A comprehensive review on the design and optimization of surface water quality monitoring networks. *Environmental Modelling Software*, 132:104792, 10 2020.

[10] Hossin M and Sulaiman M.N. A Review on Evaluation Metrics for Data Classification Evaluations. *International Journal of Data Mining Knowledge Management Process*, 5(2):01–11, mar 31 2015.

[11] Barry M Popkin, Kristen E D'Anci, and Irwin H Rosenberg. Water, hydration, and health. *Nutrition Reviews*, 68(8):439–458, jul 20 2010.

[12] Illa Iza Suhana Shamsuddin, Zalinda Othman, and Nor Samsiah Sani. Water Quality Index Classification Based on Machine Learning: A Case from the Langat River Basin Model. *Water*, 14(19):2939, sep 20 2022.

[13] Marlon Valentini, Gabriel Borges dos Santos, and Bruno Muller Vieira. Multiple linear regression analysis (MLR) applied for modeling a new WQI equation for monitoring the water quality of Mirim Lagoon, in the state of Rio Grande do Sul—Brazil. *SN Applied Sciences*, 3(1), 1 2021.

[14] Lei Xin and Tianyu Mou. Research on the Application of Multimodal-Based Machine Learning Algorithms to Water Quality Classification. *Wireless Communications and Mobile Computing*, 2022:1–13, jul 26 2022.