Water Quality Prediction Using Machine Learning

Y.Chandana[1],Baddepudi Deepthi[2],Shaik Farzana[3],Kattula Yamini Saisri[4],Madhavabotla Bharani[5]

[1]Professor,[2,3,4,5]Student

1.ychandana@gmail.com, 2.deepthi01902@gmail.com, 3.skfarzana2244@gmail.com, 4.kysaisri@gmail.com, 5.madhavabotlabharani201@gmail.com

Department of Computer Science and Engineering,

Narasaraopeta Engineering College, Narasaraopet, Andhra Pradesh, India

ABSTRACT: Water is the most important resource for every human being in their daily life. Mainly water is used for agricultural purpose and drinking purpose to use the water in both the aspects it is more important to check the quality of water whether the water is safe to use/drink either it may contain any harmful chemical substances, whether the water may contains the harmful chemicals then the water is determined as unsafe to drink/use.

The application of ML in water quality prediction is vast and includes real time a rigorous analysis, we explore the application of diverse machine learning algorithms, including Decision Trees, XGBoost, Support Vector Machines (SVM), and Ensemble Methods like Random Forest and KNN, in forecasting key water quality metrics such as Temperature, PH, and Conductivity.

In result, our study underscores the transformative impact of ML on environmental science, offering a novel approach to addressing the pressing challenge of maintaining water quality in an era of unprecedented environmental change.

KEYWORDS: WaterQualityPrediction, Machine Learning, Support Vector Machine, Random Forest, Decision Tree, KNN, XGBoost.

1.INTRODUCTION:

The primary objective of water quality prediction is to provide timely insights into potential issues, such as contamination [1] events or ecological disruptions, enabling proactive management and intervention strategies. By leveraging machine learning techniques, predictions [2] can be made with greater accuracy and efficiency compared to traditional methods.

2.LITERATURE SURVEY:

Performance analysis of the water quality index model for predicting water state using machine learning techniques.

In order to find the optimal solution for the water quality we used the algorithms including XGBoost, KNN, SVM, Navie Bayes, Decision Tree, Random Forest.

Existing water quality index models assess water quality using a range of classification schemes. Consequently, different methods to provide a number of interpretations for the same water properties that contribute toa considerable amount of uncertainty in the correct classification of water quality.

Water quality has a direct impact on public health and the environment. Water is used for various practices, such as drinking, agriculture, and industry. Recently, development of water sports and entertainment has greatly helped to attract tourists. Among various sources of water supply, due to easy access, rivers have been used more frequently for the development of human societies.

Water makes up about 70% of the earth’s surface and is one of the most important sources vital to sustaining life. Rapid urbanization and industrialization have led to deterioration of water quality of an alarming rate, resulting in harrowing diseases. Water quality has been conventionally estimated through expensive and time consuming lab and statistical analysis.

The method we propose is based on parameters like Temperature, PH, Turbidity, Solids, Hardness etc.. The use of multiple regression algortihms has proven to be important and effective in predicting the water quality.

Water quality prediction is an essential work in water environment management. Accurate forecasting value will undoubtedly improve the management level of water resources. At present, many water resource management departments have set monitoring points to observe water quality changes.

Methedology:

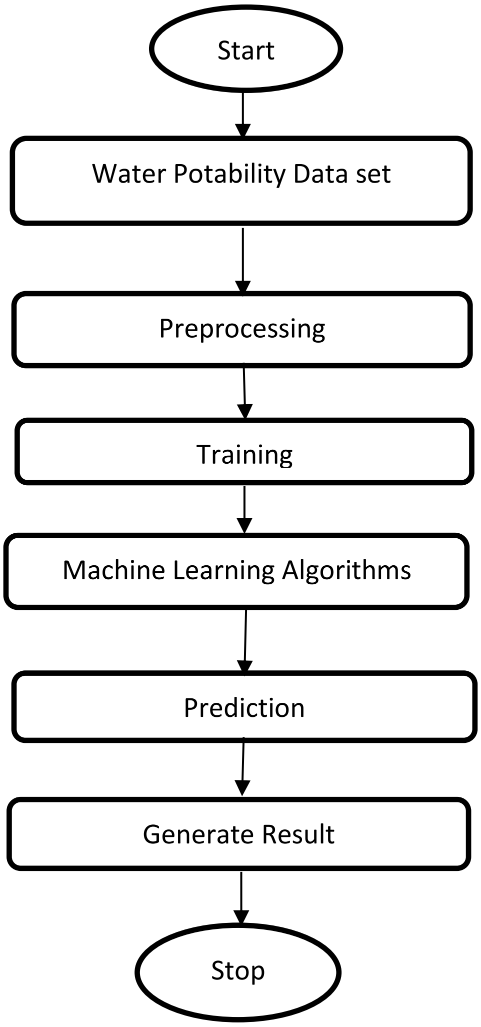


Fig 1: Overview of water Quality prediction

Archiving WQI score for coastal water Quality using recently developed improved WQI approaches and then determine the water quality classes was utilized the coastal water quality classification scheme.

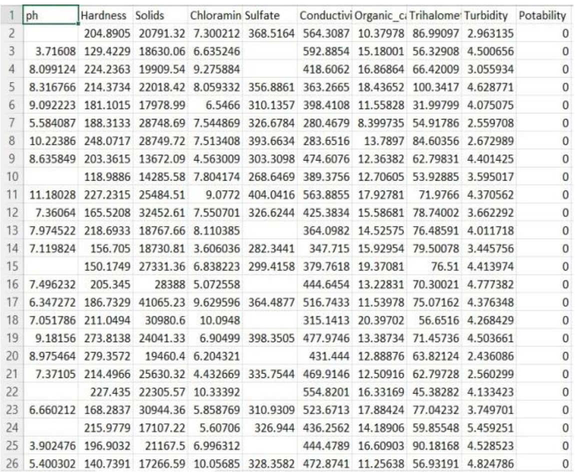
In the Fig.1 overview of this water Quality Prediction has been done through loading the data pre processing into the model inputs and after the training and testing set algorithms. ROC curve analysis is done. Develop unique classification scheme based on the best cut-point of ROC.

3.PROPOSED SYSTEM:

Our model is proposed is based in the following criteria:

3.1. Dataset Analysis
3.2. Data Visualization
3.3. Preprocessing Techniques
3.4. Model creation and Evaluation
3.5. Accuracy

3.1.Dataset Analysis: We are taken the dataset from the Kaggle website to predict the quality of water in an optimal way. For that we have the Fig.2 water_potability.csv(2022) dataset with 10 columns, those are Temperature,pH,Solids,Hardness,Chloramines,Sufates,Organic_carbon,Turbidity,Trihalomethanes, Potability. Among all those attributes Potability is considered as the target attribute.



1	ph	Hardness	Solids	Chloramin	Sulfate	Conductivi	Organic_ca	Trihalome	Turbidity	Potability
2		204.8905	20791.32	7.300212	368.5164	564.3087	10.37978	86.99097	2.963135	0
3	3.71608	129.4229	18630.06	6.635246		592.8854	15.18001	56.32908	4.500656	0
4	8.099124	224.2363	19909.54	9.275884		418.6062	16.86864	66.42009	3.055934	0
5	8.316766	214.3734	22018.42	8.059332	356.8861	363.2665	18.43652	100.3417	4.628771	0
6	9.092223	181.1015	17978.99	6.5466	310.1357	398.4108	11.55828	31.99799	4.075075	0
7	5.584087	188.3133	28748.69	7.544869	326.6784	280.4679	8.399735	54.91786	2.559708	0
8	10.22386	248.0717	28749.72	7.513408	393.6634	283.6516	13.7897	84.60356	2.672989	0
9	8.635849	203.3615	13672.09	4.563009	303.3098	474.6076	12.36382	62.79831	4.401425	0
10		118.9886	14285.58	7.804174	268.6469	389.3756	12.70605	53.92885	3.595017	0
11	11.18028	227.2315	25484.51	9.0772	404.0416	563.8855	17.92781	71.9766	4.370562	0
12	7.36064	165.5208	32452.61	7.550701	326.6244	425.3834	15.58681	78.74002	3.662292	0
13	7.974522	218.6933	18767.66	8.110385		364.0982	14.52575	76.48591	4.011718	0
14	7.119824	156.705	18730.81	3.606036	282.3441	347.715	15.92954	79.50078	3.445756	0
15		150.1749	27331.36	6.838223	299.4158	379.7618	19.37081	76.51	4.413974	0
16	7.496232	205.345	28388	5.072558		444.6454	13.22831	70.30021	4.777382	0
17	6.347272	186.7329	41065.23	9.629596	364.4877	516.7433	11.53978	75.07162	4.376348	0
18	7.051786	211.0494	30980.6	10.0948		315.1413	20.39702	56.6516	4.268429	0
19	9.18156	273.8138	24041.33	6.90499	398.3505	477.9746	13.38734	71.45736	4.503661	0
20	8.975464	279.3572	19460.4	6.204321		431.444	12.88876	63.82124	2.436086	0
21	7.37105	214.4966	25630.32	4.432669	335.7544	469.9146	12.50916	62.79728	2.560299	0
22		227.435	22305.57	10.33392		554.8201	16.33169	45.38282	4.133423	0
23	6.660212	168.2837	30944.36	5.858769	310.9309	523.6713	17.88424	77.04232	3.749701	0
24		215.9779	17107.22	5.60706	326.944	436.2562	14.18906	59.85548	5.459251	0
25	3.902476	196.9032	21167.5	6.996312		444.4789	16.60903	90.18168	4.528523	0
26	5.400302	140.7391	17266.59	10.05685	328.3582	472.8741	11.25638	56.93191	4.824786	0

Fig 2: water potability Dataset

3.2. Data Visualization: To visualize the data in the form of the graphs we used numpy, pandas, seaborn, pyplotlib, sklearn libraries. each and every single library is used to represent the graph in detailed manner. This helps us to recognize potability of water among all the combined compounds.

Data visualization techniques such as histograms, scatter plots and box plot will allow us to explore the distribution of data, identify patterns, correlations, outliers and potential relation between variables.

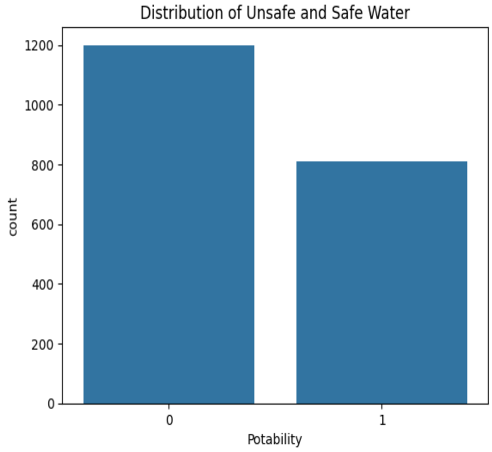
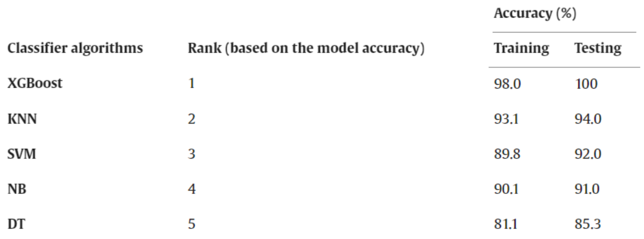


Fig 3: Distribution of safe & unsafe water.

This Fig.3 shows us the distribution of safe and unsafe water, count is taken on y-axis and potability is taken on x-axis it explains the count of occurrences for each category of potability which represents whether water is potable or not.



Classifier algorithms	Rank (based on the model accuracy)	Accuracy (%)	
		Training	Testing
XGBoost	1	98.0	100
KNN	2	93.1	94.0
SVM	3	89.8	92.0
NB	4	90.1	91.0
DT	5	81.1	85.3

Fig 4: Model performance on different classifiers

Water Quality Index (WQI)

In the above Fig.4 we can observe the models that are performed on different classifiers for model prediction based on accuracy. For the purpose of reducing model uncertainty [11] recently have proposed an enhanced and comprehensive WQI approach for computing WQI scores in order to assess the costal and transitional water quality. This approach is shown to be more reliable than that used in existing methods because it is the most up to date method for computing WQI, and also it may be an effective tool to avoid model ambiguity.

3.3.Preprocessing Techniques:

Correlation: Feature selection is an effective strategy to reduce dimensionality, remove irrelevant data and increase learning accuracy[12] it is the stastical summary of the relationship between two sets of variables it tells about how two variables move in relation to one another. correlation analysis refers to methods that estimates the impulse response of a linear model.

By applying the correlation we have observed that two attributes are correlated. Hence for the further process we didn’t make any changes to our dataset.

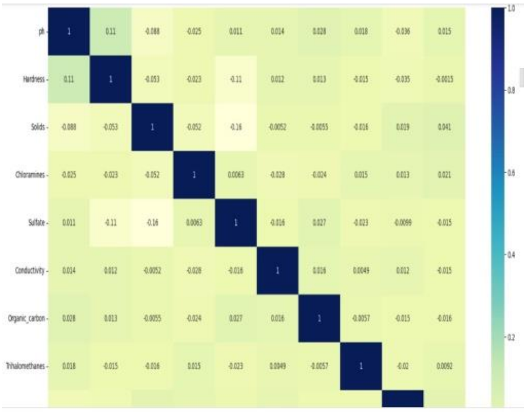


Fig 5: correlation matrix

Outliers are datapoints that significantly refer from other obsevations from the dataset they (Fig.5) can arise due to measurement error, experimental variability, genuine deviations in the data, It is essential to identify outliers in the existing data using visualizations like box plots.in some cases outliers can be removed from the dataset if they are irrelevant to the analysis. Outliers are infrequent observations between a dataset the problem handling is essential to ensure accurate reasons.

The data points may arise due to various reasons such as measurement errors, data correction, or genuine rare events. Handling outliers is important because they can skew statistical analysis and machine learning models, leading to inaccurate results and decreased model performances.

The results of the classifiers were evaluated using four validation metrics (accuracy, precision, Recall, and F1 score) for the imbalanced dataset, the confusion matrices is one of them.

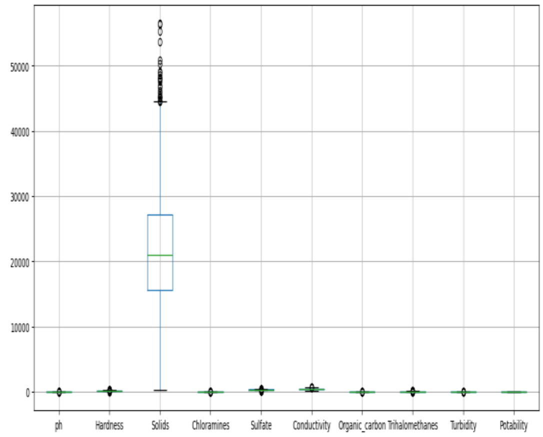


Fig 6: Boxplot of the dataset.

Null values: In water potability.csv dataset we have observed many null values present in the dataset For that we have removed all the null values which will affect to our further preprocessing techniques as shown in Fig.6.

And in these preprocessing techniques we have observed the factors that are affecting to the main attribute (Potability). We have clearly explained the factors that affect to each and every attribute to understand clearly about the factors to know the relation between dependent and independent attributes.

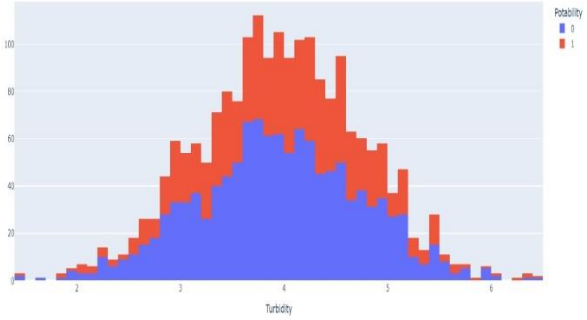
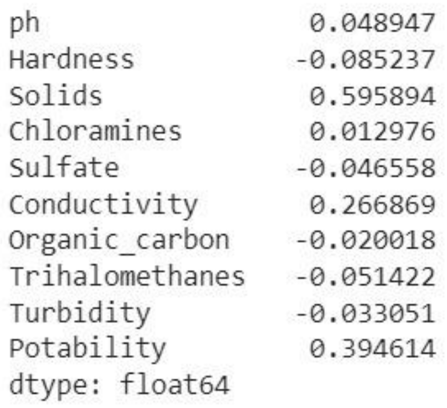


Fig 7: Factors affecting target attribute

We used the function SKEW which calls and appears to be related to calculating Fig.7 the skewness of data along the specified axis. The skew() method measures the asymmetry of the distribution of values in the dataset.



```
ph                  0.048947
Hardness           -0.085237
Solids              0.595894
Chloramines         0.012976
Sulfate            -0.046558
Conductivity        0.266869
Organic_carbon     -0.020018
Trihalomethanes    -0.051422
Turbidity          -0.033051
Potability          0.394614
dtype: float64
```

Fig 8: correlation coefficients

Skewness is a measure of the asymmetry of the probability distribution of a real-valued random variable as shown in Fig 8. In simpler terms, it tells us how much and in which direction a distribution deviates from a normal distribution.

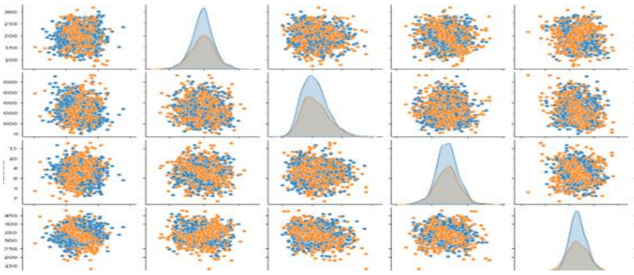


Fig 9: Pair plot of the dataset

A pair plot is a versatile and powerful visualization tool used to understand the relationships between multiple variables within a dataset. It presents a comprehensive overview by plotting pairwise relationships across all the variables as represented in Fig.9. This can help in identifying patterns, correlations, and potential hypotheses for more detailed analysis.

Diagonal plots are usually histograms or density plots showing the distribution of a single variable.

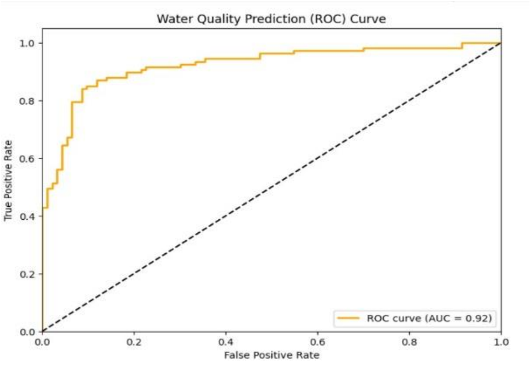


Fig 10: ROC curve

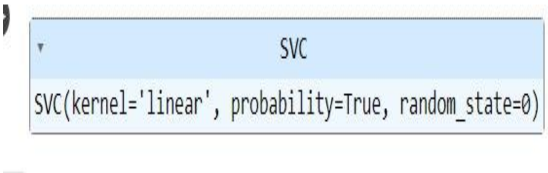
Roc curve applies probability confidence interval or ranking to each prediction models such as naïve bayes and SVM ranks as part of their algorithm. Which is represented in Fig.9. Basically, prediction ranking is employed in ROC algorithm to achieve distinct decision.

3.4. Model Creation and Evaluation:

XGBoost: It is a machine learning algorithm used for specially the gradient boosting frame work. It uses decision tree as base learners for model generalization. XGBoost is commonly used to perform tasks such as ranking, classification and regression [5] . In order to increase the predictive accuracy XGBoost has out performed the other models and improved the performance prediction.

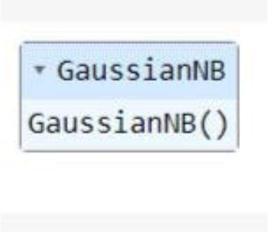
K-Nearest Neighbour (KNN): It is classified as to identify unlabeled observations by allocating them to the class to the similar labeled examples. Characteristics of KNN classifier are collected for both training and test data set [6].KNN is a non-parametric algorithm it will not obtain parameters for the model. K is the most important parameter in the KNN algorithm to identify the difference of the diagnostic accuracy of KNN model.

SupportVectorMachine(SVM):SVM is a binary classifier it attempts to generate an another hyper plane in actual space of endcordinates between two different classes. Firstly it visualizes the data and then it finds the best separator between the classes, the data points that the closest to the target value[7] be found out, then the actual data in accordance with the linear separation. Basically SVM draws the line between the different points of data considers in the dataset.



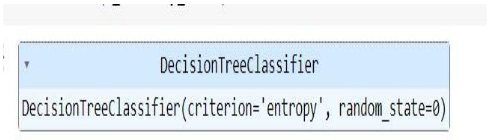
```
SVC
SVC(kernel='linear', probability=True, random_state=0)
```

Navie Bayes(NB): It is a scalable and it requires set of parameters which are proportional to the number of variables in a learning problem. It makes a probability decision by likelihood of two features which are independent and equal significance. Navie Bayes theorem generally works on the phases known as; [8] probability and independence, training phase, feature probability, class probability, prediction, class selection.



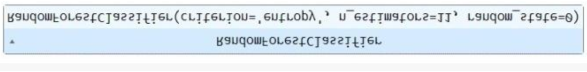
```
GaussianNB
GaussianNB()
```

Decision Tree(DT): Decision tree contains root node, branches and leaf nodes. Testing an attribute is done on every internal node, the outcome of the test will results on leaf node. Decision tree is easy to understand because it is similar to human decision making process, it can solve continuous data as input.[9].The main advantage is that it can be able to select the most biased feature and comprehensibility nature.



```
DecisionTreeClassifier
DecisionTreeClassifier(criterion='entropy', random_state=0)
```

Random Forest(RF): Random Forest is a new combination algorithm which is a combination of series of structure classifiers like tree the application scope of random forest is very extensive it is widely used for prediction, classification and regression [10] compared with other traditional algorithms random forest has many good virtues. During the training process random forest can operate learning method by constructing a large number of decision trees.



```
RandomForestClassifier(criterion='entropy', n_estimators=11, random_state=0)
▸ RandomForestClassifier
```

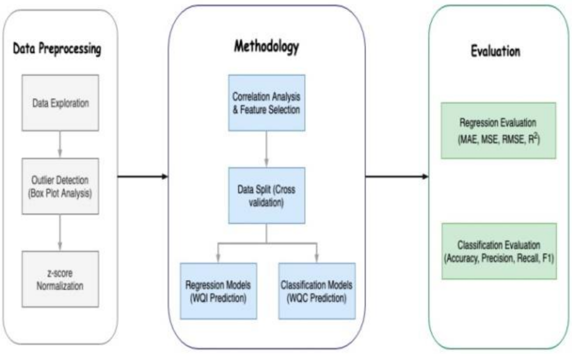

Data Preprocessing
Data Exploration
Outlier Detection (Box Plot Analysis)
z-score Normalization
Methodology
Correlation Analysis & Feature Selection
Data Split (Cross validation)
Regression Models (WQI Prediction)
Classification Models (WQC Prediction)
Evaluation
Regression Evaluation (MAE, MSE, RMSE, R²)
Classification Evaluation (Accuracy, Precision, Recall, F1)

Fig 11: Water Quality Prediction using Supervised Machine Learning

In the above figure.11 We can observe the methodology used to predict the quality of water using Supervised machine learning. In this study we can see the Evaluation and data preprocessing.

3.5. Accuracy

Accuracy is used to evaluate the classification in machine learning [13]

Accuracy=TP+TN TP+TN+FP+FN

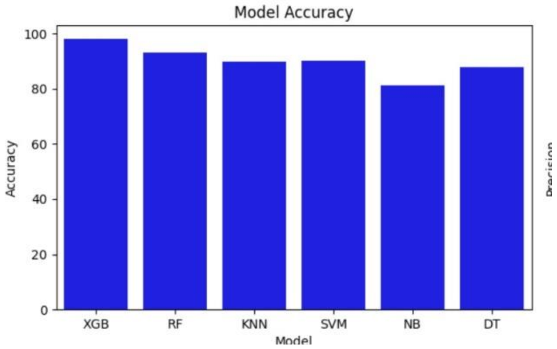


Fig 12: Model Accuracy score

Precision refers to close to the measurments between the algorithm predictions and observation of the same classification are[11] to each other. Precision is determined Fig.13 as follows:

Precision = TP / TP+FP

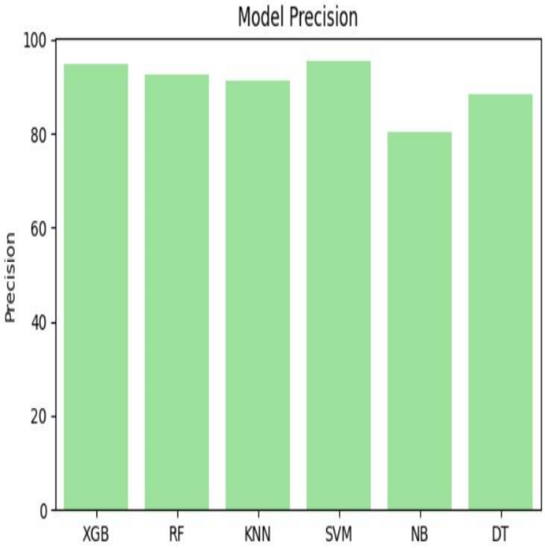


Fig 13: Model precision Score

Recall measures how frequently the algorithm detects the correct classification from the given data whereas the actual correct classification has occurred in dataset. False negative are labeled as negative whereas the observation classes are actually positive. Recall is determined Fig.14 as follows:

Recall = TP / TP+FN

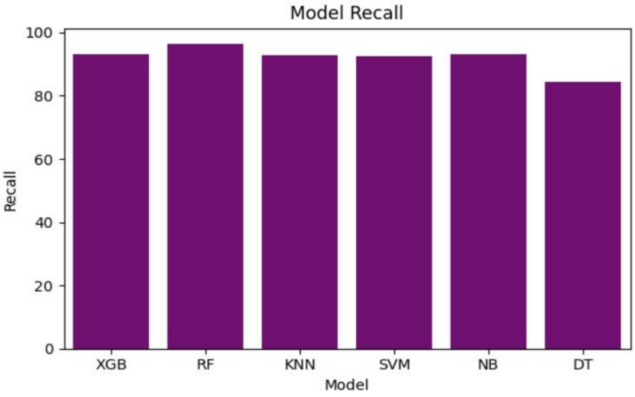


Fig 14: Model recall score

F1 score is evaluate the multi class classification and also it is an approach to harmonizing the precision and recall of the predictive model. F1 score is obtained Fig.15 as follows:

F1score=2*(precision*recall/precision+recall)

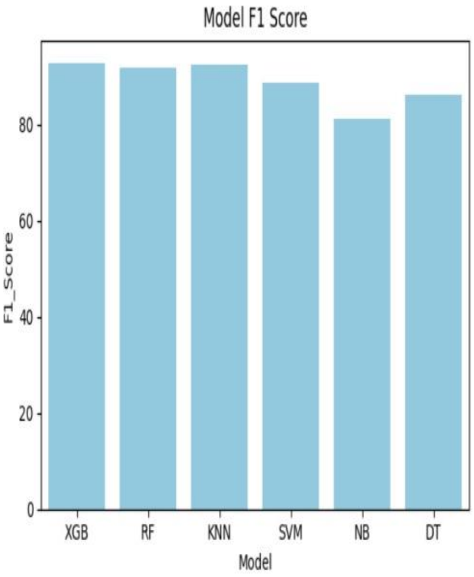


Fig 15: Model F1 score

TP: The actual observation indicates that water quality classes has classified accurately and the model predicted correct classification of water quality from the given data.

TN: The actual observation that indicates the water quality classes has classified accurately but the model detected in correct classification from the given data.

FP: The actual observation refers that water quality classes has not classified accurately whereas the model also detects the incorrect classification of water quality from the given data.

FN: The actual observation reveals the water quality classes has not classified accurately although the model predicted correct classification for water quality from the given dataset.

In this we have observed that the accuracy score after removing the outliers in the data the score has been increased, when compared with the considered base paper. In the following graph we can see the difference of scores after constructing the accuracy model with respect to Accuracy comparision.

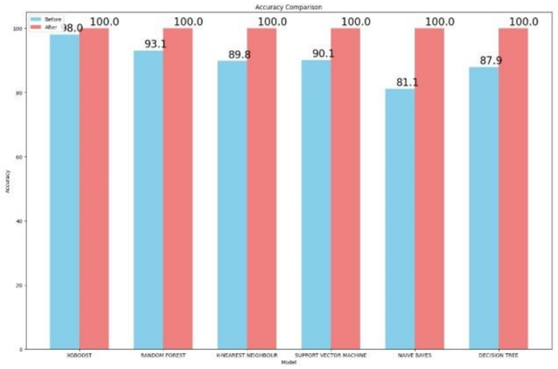


Fig 16: Accuracy Comparision

In the precision we have observed that the precision score also have the high score compared to the scores of the dataset that we have taken. Precision determines the exact values of the scores of the data.

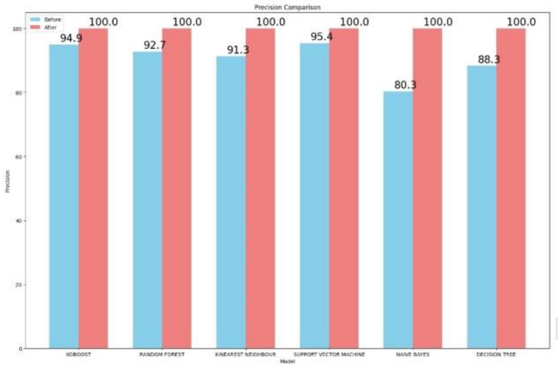


Fig 17: Precision Comparision

F1 score of the data that we are taken the scores have increased after removing outliers and constructing the model.

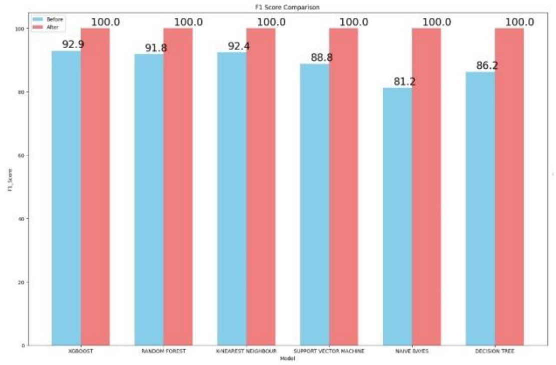


Fig 18: Recall Comparision

When all the models has applied the recall score of the data has increased as compared to the existing data.

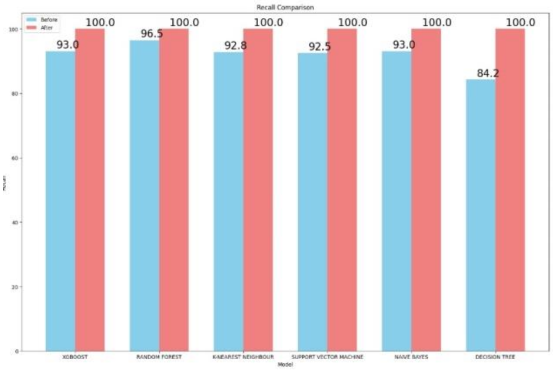


Fig 19: F1 score Comparision

After applying all the models to our data the accuracy of the data we have observed is plotted in the following graph as in x-axis we can see the models that we have used and in y-axis we can see the rate of accuracy of the constructed model.

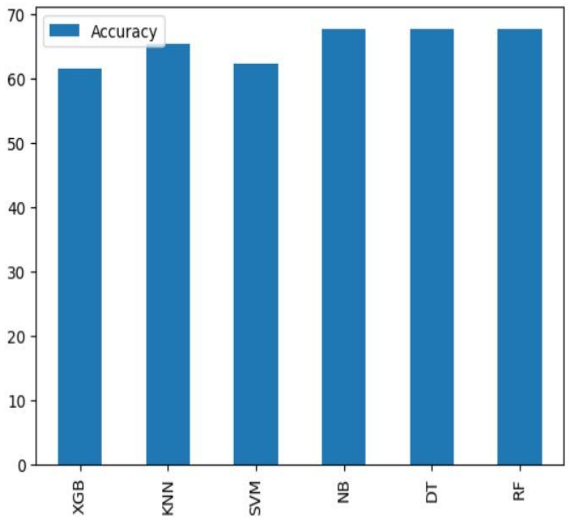


Fig 20: Accuracy

Conclusion:

The main aim of our study were to develop a framework for assessing performance of WQI model in order to correct classification of water quality. After removing all the missing values in the data we have observed the accuracy of the algorithms have increased.

However, as best of our knowledge, this study provides the first comprehensive approach to evaluate the performance of WQI model adopting new classification scheme for multi-class classification of coastal water quality.

Moreover, the results of this study could be effective in obtaining the proper classification of water quality. Which might be useful to improve the WQI model accuracy.

References:

[1] Wang, Xianhe, Ying Li, Qian Qiao, Adriano Tavares, and Yanchun Liang. 2023. "Water Quality Prediction Based on Machine Learning and Comprehensive Weighting Methods" Entropy 25, no. 8: 1186.

[2] Integrating multisensor satellite data merging and image reconstruction in support of machine learning for better water quality management
J. Environ. Manag.
(2017)

[3] P. Wickramasinghe proposed a methodology to predict the performance of batsman for the previous five years using different machine learning algorithms,PP.64- 81,March 2018.

[4] P. Wickramasinghe proposed a methodology to predict the performance of batsman for the previous five years using different machine learning algorithms,PP.64- 81,March 2018 .

[5] Asselman, A., Khaldi, M., & Aammou, S. (2020). Evaluating the impact of prior required scaffolding items on the improvement of student performance prediction. *Education and Information Technologies*, 25, 3227–3249. https://doi.org/10.1007/s10639-019-10077-3

[6] Zhang Z. Introduction to machine learning: k-nearest neighbors. Ann Transl Med. 2016 Jun;4(11):218. doi: 10.21037/atm.2016.03.37. PMID: 27386492; PMCID: PMC4916348.

[7] K.-T. Chang *et al* .Modeling typhoon- and earthquake-induced landslides in a mountainous watershed using logistic regression
Geomorphology (2021)

[8] T.H.H. Aldhyani, M. Al-Yaari, H. Alkahtani, M. Maashi Water Quality Prediction Using Artificial Intelligence Algorithms Appl. Bionics Biomech. (2020)

[9] Patel HH, Prajapati P. Study and analysis of decision tree based classification algorithms. International Journal of Computer Sciences and Engineering. 2018 Oct 31;6(10):74-8.

[10] Kalyankar, G.D., Poojara, S.R., Dharwadkar, N.V.: Predictive analysis of diabetic patient data using machine learning and hadoop. In: International Conference On I-SMAC (2017). ISBN 978-1-5090-3243-3

[11] River water quality index prediction and uncertainty analysis: A comparative study of machine learning models

J. Environ. Chem. Eng., 9 (2021),
Article 104599, 10.1016/j.jece.2020.104599