Proceedings of the SMART–2022, IEEE Conference ID: 55829
11th International Conference on System Modeling & Advancement in Research Trends, 16th–17th, December, 2022
College of Computing Sciences & Information Technology, Teerthanker Mahaveer University, Moradabad, India

Water-Quality-Analysis using Machine Learning

Raavi Akshay[1], Gadiraju Tarun[2], Pinapothini Uday Kiran[3], K Durga Devi[4] and M.Vidhyalakshmi[5]
[1,2,3,4,5]*ECE, SRM Institute of Science and Technology, Ramapuram, Chennai, India*
Email: [1]akshay.raavi2001@gmail.com, [2]tarun176732@gmail.com, [3]puday223011@gmail.com,
[4]durgaece.0711@gmail.com, [5]m.vidhyalakshmi@gmail.com

2022 11th International Conference on System Modeling & Advancement in Research Trends (SMART) | 978-1-6654-8734-4/22/$31.00 ©2022 IEEE | DOI: 10.1109/SMART55829.2022.10047533

***Abstract*—One of the most serious and alarming problems facing humanity is the degradation of natural water resources such as lakes as well as rivers is one of the most serious and vexing problems we are facing today. The long-term effects of polluted water affect all areas of life. Therefore, it is essential to manage water resources if you want tomaximize the quality of your water. If data are examinedand water quality can be predicted, the effects of watercontamination can be dealt with effectively. The purpose of this study is to use machine learning to make a water qualityprediction model based on water quality measurements. Machine learning can be used for building models fromalgorithms with some data gathered from the sick ones. For a better examination of parametric findings, the acquired data will be pre-processed, separated into training and testing parts, and subjected to machine learning classification techniques. Decision tree, Naive Bayes, Random Forest, Support Vector Machine, and K-Nearest Neighbor are some of the classification-type algorithms employed in this work. All the model's performance indicators are calculated, and they change for each model. A technique for improving machine learning model performance metrics is hyperparameter tuning.**

Keywords: Dataset pre-processing, Data standardization, Performance metrics, Hyperparameter Tuning.

I. Introduction

In the scientific field of machine learning, it is investigated how computers learn via experience. Since the capacity to learn is the fundamental quality of an entity regarded as intelligent in the broadest meaning of the word, the words "Machine Learning" and "Artificial Intelligence" are frequently used synonymously in the minds of scientists. Building adaptable, experience-based computer systems is the goal of machine learning. It is now possible to discover a solution to this problem because of the development of machine learning methods. We have developed a technique that uses data mining to identify whether the water is portable or not. The enormous amount of data related to water quality can be mined for hidden knowledge. As a result, it now has a more significant role in the study. This research aims to develop a system that can predict water quality more precisely.

II. Literature Survey

An analysis of the use of soft computing techniques in water resources and tested support Vector Machine Algorithm in analysing water quality in clear streams [1,2]. [16] proposed significant opportunities exist for improving the classification and forecasting of water quality with artificial intelligence (AI). In this paper, they proposed a reliable framework for classifying water quality using machine learning methods (WQ). This study compares various AI algorithms to manage water quality data accumulated over time and offers a trustworthy approach for projecting water quality as precisely as feasible. [17] this study examines the effectiveness of artificial intelligence approaches such as artificial neural network (ANN), group method of data handling (GMDH), and support vector machine (SVM) for forecasting water quality using machine learning techniques. Different transfer and kernel function types were investigated to construct the ANN and SVM, respectively. [18] this work provides an intelligent wireless sensor network (WSN) system for water quality measurement utilizing machine learning that can assess the river water quality. This study's principal goal as a case study is to evaluate the river's status along its path by producing data reports into an interactive user interface.

Utilizing machine learning to analyse and forecast water quality: Our analysis of the literature suggests that a complete model is needed to comprehend what safe, potablewater is and to identify it from non-potable water using machine learning approaches. ANN has received widespread recognition as a tool for classifying complicated information, including those environmental dynamics. It describes the complex water quality dataset's non-linear connection.

[15] Predict irrigation water quality characteristics in a semi- arid environment using machine learning models. Conventional methods for evaluating water suitability for irrigation are typically expensive because they call for multiple characteristics, especially in underdeveloped nations. Therefore, creating precise and trustworthy models could be useful to solve this problem of managing the water utilized in agriculture. Eight Machine Learning (ML) models are utilized to do this. [19] The cutting-edge artificial intelligence algorithms NARNET and LSTM models were used to forecast the WQI utilizing the proposed technique. Additionally, the WQI data was categorized utilizingmachine learning techniques like SVM, KNN, and Naive Bayes. The proposed models were assessed and looked at using a few statistical factors. Water Quality Prediction Using Artificial Intelligence Algorithms. [20] For the conservation of the water environment, water

qualityprediction is very important. In this research, A method for predicting water quality based on IGRA and LSTM is suggested, considering the multivariate correlation and temporal sequence of the water quality information.

From the survey made [3-15], the major research gaps were listed below:

This research on the effectiveness of current or planned strategies for addressing WS&D needs a lot of information on these causal relationships, as well as the original problems with water scarcity and drought and their explanatory factors.

There is a great need for knowledge of these causal links, as well as the initial issues with water scarcity and drought and their explanatory factors, in this research on the success of present or proposed measures for treating WS&D.

They are frequently the least accurate and unreliable data regarding water resources (Gleick, 2006). Other investigations concluded that it is difficult to find accurate and comprehensive data on water availability and demand (Gleick et al., 2002).

Consequently, there is insufficient information regarding the motivators, constraints, and effects of the measures and support actions.

III. Proposed Method

Analyzing water quality is an important research area. Here shows the proposed step-by-step process to predictwater quality.

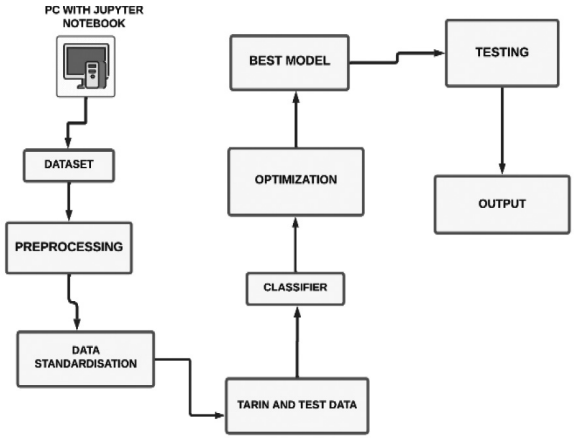


Fig. 1: Overall Architecture

The machine learning model is used to detect whether the water is potable or non-potable. Import relevant libraries to test and train our data set and required to install some packages related to nature-inspired algorithms. Split the data as training data set and testing dataset they should bein the ratio 80:20 respectively and perform the Model Selection. The Support Vector Classifier (SVC), Decision Tree, GaussianNB, Random Forest and XGBoost are these different classifiersthat are taken into consideration.

The model is assessed using several performance measures, suchas Test Accuracy and Train Accuracy. We got to get more Accuracy by using Hyperparameter Tuning in the Random Forest algorithm. Based on different performance metrics values the classifier with the greatest value is considered the best model.

Data Set - Dataset consists of observations of water quality for 3276 different sources of water:

pH - The water's pH (0 to 14). According to EPA recommendations, tap water's pH should range between (6.5 and 8.5). The pH level is a crucial factor in determining the acid-base nature of water. Additionally, it shows if the water is either alkaline or acidic. The present investigation's range fell between 6.52 to 6.83, which is within WHO guidelines.

Hardness - It is the amount of soap that may dissolve in one litre of water. Salts made of calcium and magnesium are the major causes of hardness. How long water is exposed to a hardness-producing substance influences how hard the water is while it is in the raw state. The ability of water to form soap due to calcium and magnesium precipitation was the original definition of hardness.

Total Dissolved Solids (TDS) - Water can dissolve a wide variety of chemicals and certain organic minerals or salts, including sodium, calcium, iron, zinc, bicarbonate ions, chloride ions, magnesium, and sulphates. These minerals affected the water's appearance and gave it foul smells. This is an important consideration while using water. A highTDS rating indicates that the water contains a lot of minerals. For drinking purposes, the maximum and desired TDS limits are 500 mg/l and 100 mg/l, respectively.

Sulfates - These are the organic substances that are found naturally in minerals, soil, and rocks. They are present in the air, groundwater, plants, and food in the area. Sulfate is mostly utilized for business purposes in the chemical sector. Around 2,700 mg/L of sulphate can be found in seawater. While certain places have significantly higher levels (1000 mg/L), most freshwater sources have values between 3 and 30 mg/L.

Conductivity - Pure water is great insulation of electrical current. By raising the ion concentration, the liquid's electric conductivity has been enhanced. The quantity of dissolved particles in the liquid often determines its conductivity. Electrical Conductivity measures how well they carry electricity through their ionic mechanism (EC). WHO recommendations state that the EC value shouldn't be higher than 400 S/cm.

Chloramines - The two primary disinfectants used in public water systems are chloride and chlorine. Ammonia is used in combination with chlorine to clean potable water.

Drinking water can include up to 4 mg/L of chlorine, which is regarded as a safe quantity.

Potability - It is a metric for determining whether water is fit for human consumption. Unpotable equals zero (0), while potable is one (1).

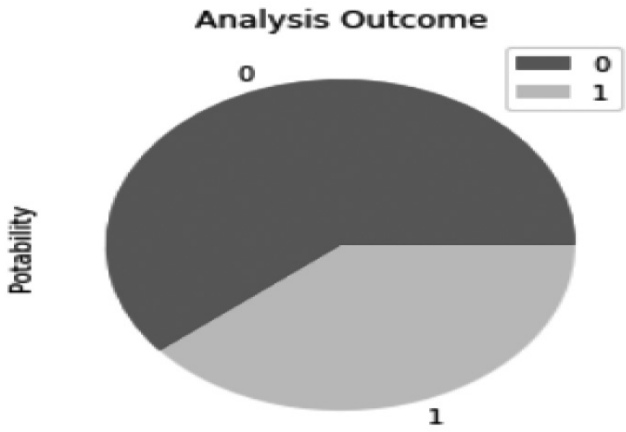
VI. Methodology & Evaluation

A. Data Description

The dataset consists of observations of water quality for 3276 different sources of water:

	ph	Hardness	Solids	Chloramines	Sulfate	Conductivity	Organic_carbon	Trihalomethanes	Turbidity	Potability
0	NaN	204.890455	20791.318981	7.300212	368.516441	564.308654	10.379783	86.990970	2.963135	0
1	3.716080	129.422921	18630.057858	6.635246	NaN	592.885359	15.180013	56.329076	4.500656	0
2	8.099124	224.236259	19909.541732	9.275884	NaN	418.606213	16.868637	66.420093	3.055934	0
3	8.316766	214.373394	22018.417441	8.059332	356.886136	363.266516	18.436524	100.341674	4.628771	0
4	9.092223	181.101509	17978.986339	6.546600	310.135738	398.410813	11.558279	31.997993	4.075075	0

```
0     1998
1     1278
Name: Potability, dtype: int64
```




```
0---------> Non Potable
1---------> Potable
```

Fig. 1: Data Description

B. Data Pre-processing

The data quality must be improved at the processing stage of the data analysis process. The Water quality index has been determined in this phase using the important dataset parameters. The act of converting collected data into something an algorithm for machine learning can use is known as data preparation. The most important and first stage in building an algorithm for machine learning is this one. Remove all instances where the value is 0. (zero). Zero is not a possible value. Therefore, this instance is terminated. The process of deciding on feature subsets, which decreases the dimension of the data and helps to work more quickly, involves removing irrelevant characteristics and instances.

```
1 df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3276 entries, 0 to 3275
Data columns (total 10 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   ph               2785 non-null   float64
 1   Hardness         3276 non-null   float64
 2   Solids           3276 non-null   float64
 3   Chloramines      3276 non-null   float64
 4   Sulfate          2495 non-null   float64
 5   Conductivity     3276 non-null   float64
 6   Organic_carbon   3276 non-null   float64
 7   Trihalomethanes  3114 non-null   float64
 8   Turbidity        3276 non-null   float64
 9   Potability       3276 non-null   int64
dtypes: float64(9), int64(1)
memory usage: 256.1 KB
```

Fig. 2 Data Processing

C. Histogram

Visualize each aspect of the provided data set as shown below; now examine the graphs. It demonstrates how each feature and label is dispersed across many ranges, further demonstrating the necessity of scalability. Second, categorical variables are indicated whenever there are discrete bars. Before using machine learning, we need to address these categorical data.

```
1 df.hist(figsize=(14,14))
2 plt.show()
```

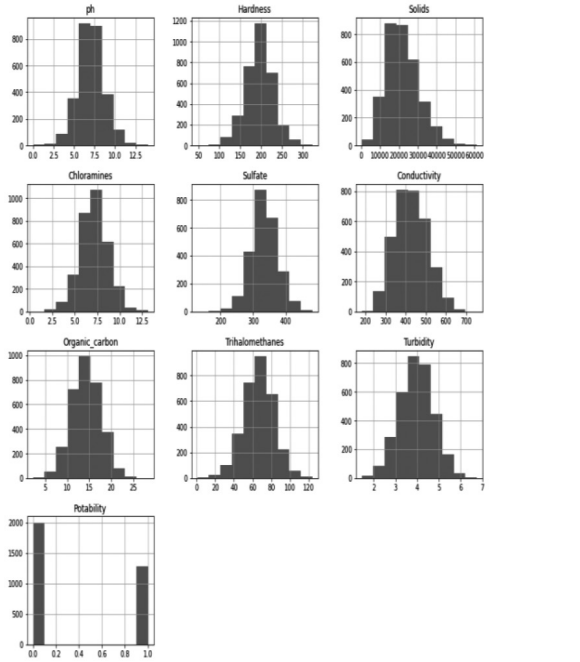


Fig. 3: Histogram

D. Correlation Matrix

By Visualizing the correlation of all characteristics using a thermal foot map function. But you can see from the heat map below that there is no correlation between any characteristic; this means that we cannot reduce the dimension.

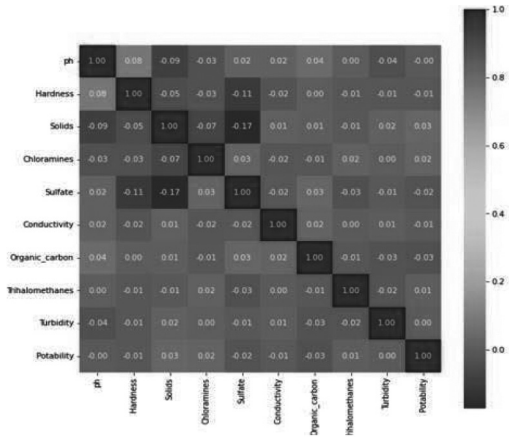


Fig. 4: Correlation Matrix

E. Data Standardization

In a scaling procedure called data standardization centers, the values on the mean of the unit standard deviation. Data standardization is the act of converting data into a commonly accepted format so that users may analyze and process it. We can import Data Standardization in many ways. It helps build unique, consistently defined components and features by providing acomprehensive catalogue of your data. No matter what insights or problems you're trying to address, a solid understanding of your data is a necessary first step. To get there, the information must be changed into a dependable format with clear meanings.

F. Training and Testing of Data

In machine learning, the model is instructed to perform a variety of tasks using a training set of data. The model is trained using certain features from the training set. Therefore, the prototype contains these structures. Words or word clusters are taken from tweets for sentiment analysis. They build connections, understand concepts, come to judgments, and assess their level of confidence using the training data. The quality and quantity of the Machine Learning training data, we use determines how well our data project performs, just as much as the algorithms they do. As a result, provided the training set is correctly labelled, the model will be able to learn about the features.

Divide the information into independent and dependent features.

```
from sklearn.preprocessing import StandardScaler
std=StandardScaler()
x_train_std=std.fit_transform(x_train)
x_test_std=std.transform(x_test)
x_train_std
x_test_std

```

```
array([[ 0.65628348,  0.62573575,  1.84574407, ..., -0.37768933,
         0.75434767,  0.59209683],
       [-0.18941828, -0.51794773,  0.20903809, ..., -0.21822116,
        -0.35736554,  0.22554529],
       [        nan, -0.48475986, -0.48327085, ..., -1.15081249,
        -2.11629388,  0.03685464],
       ...,
       [ 0.41989586,  0.86377001, -0.27955714, ...,  0.26260917,
         0.51874517, -0.34676108],
       [ 0.71002518,  0.19338311,  0.6535729 , ..., -0.0093478 ,
        -0.21324075, -0.78295779],
       [ 0.26754367,  0.4144899 ,  0.82993133, ...,  0.79369607,
         1.14237673,  1.95819472]])
```

Fig. 5: Data Standardization

G. Decision Tree

The decision tree is a Machine Learning algorithm, it is mostly focused on classification-related issues. The decision tree has a structured classifier in which the nodes within display the components of a particular dataset. Decisionnodes and leaf nodes are both types of nodes found in decision trees.

H. Logistic Regression

It is machine learning which is used for the binary classification model where one of two possible values is taken as output. Logistic regression is our output prediction that can take either of these values. Logistic regression is simpler to use, analyze, and train.

I. Support Vector Machine

The SVM is an algorithm which is used in machine learning to categorize the task. It is frequently used for classification problems. SVM separates the data into two classes by mapping the data points to a high-dimensional space and then locating the best hyperplane.

J. Random Forest Classifier

The popular learning algorithm Random Forest is a part of the supervised learning methodology. It may be applied to ML issues involving both classification and regression. It is built on the idea of ensemble learning, which is a method of integrating many classifiers to address difficult issues and enhance model performance. Random Forest, as the name implies, is a classifier that uses several decision trees on different subsets of the provided dataset and averages them to increase the dataset's prediction accuracy. Instead, then depending on a single decision tree, the random forest uses forecasts from each tree and predicts the result based on the votes of most predictions.

K. XGBoost

Extreme Gradient Boosting is a framework that can run on multiple languages. It is popular supervised learning which works on large datasets. It is implemented on top of the gradient boost. The way the XGBoost algorithm is designed to work uses the parallelization concept. It uses sequentially generated shallow decision trees and a highly scalable training method to minimize overfitting to deliver accurate results.

L. Performance Metrics

Accuracy - Accuracy is measured as the total count of actual predictions to the available predictions and it is multiplied by 100.

Precision - The ratio of actual positives to the total available positives is known as precision.

Recall - It mainly focuses on type-2 errors the ratio of true positives to false negatives is called recall.

Fl-score - The harmonic mean performance metric parameters precision with recall known as f1-score.

IV. Results and Discussion

1) *Support Vector Machine*

Support Vector Machine is giving an Accuracy of 62 per cent.

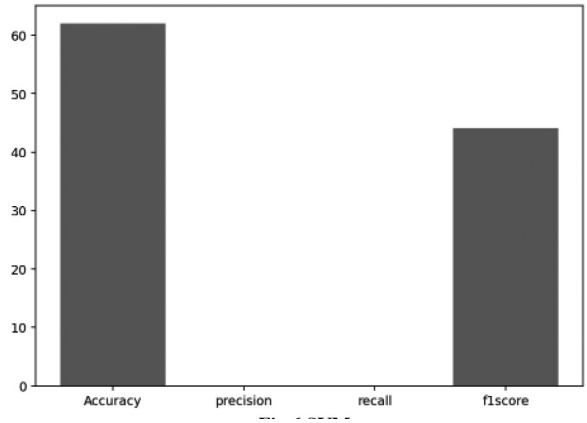


Fig. 6: SVM

2) *Random Forest Classifier*

Random Forest Classifier is giving an Accuracy of 70 per cent.

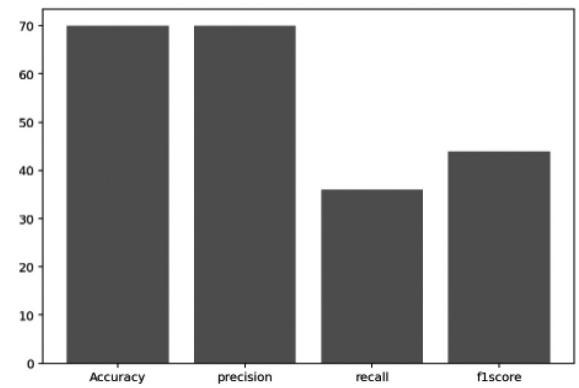


Fig. 7: Random Forest Classifier

3) *K Nearest Neighbour*

K Nearest Neighbour is giving an Accuracy of 65 per cent.

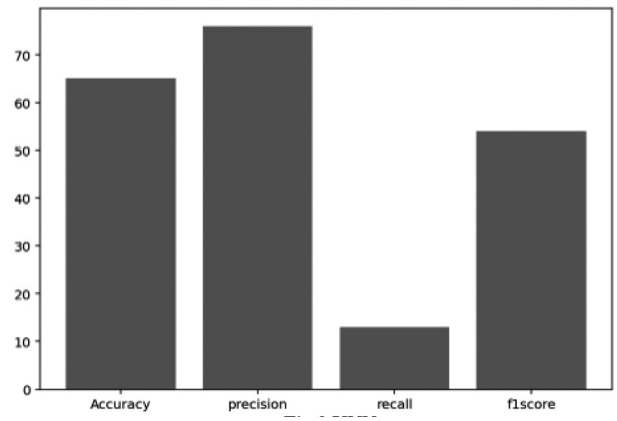


Fig. 8: KNN

4) *Decision Tree*

Decision Tree is giving an Accuracy of 62 per cent.

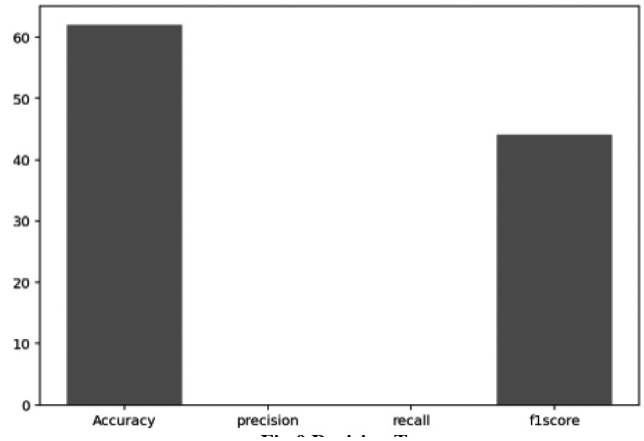


Fig. 9: Decision Tree

5) *GaussianNB*

GaussianNB is giving an Accuracy of 62 per cent.

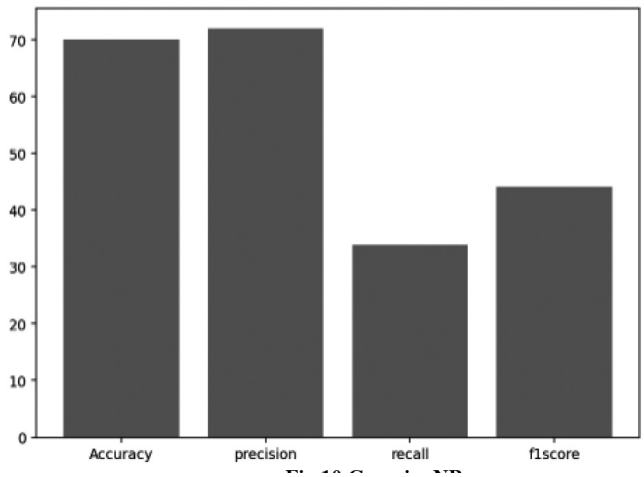


Fig. 10: GaussianNB

6) *XGBoost*

XGBoost is giving an Accuracy of 64 per cent.

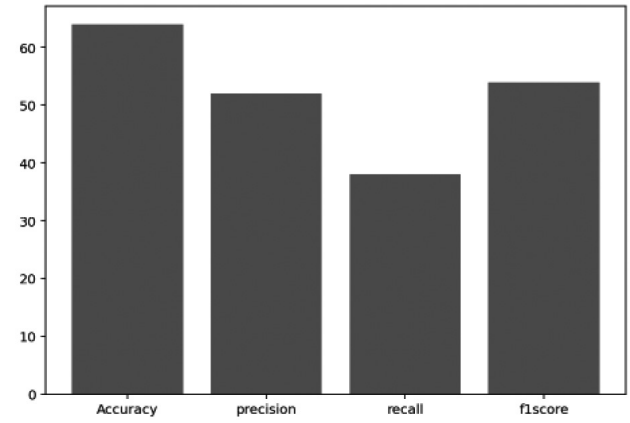


Fig. 11: XGBoost

Performance Metrics for machine learning algorithms with Hyperparameter Tuning.

	model	accuracy	precision	recall	f1score	rocauc	logloss	timetaken	confusionmatrix
0	Decision Tree	0.625000	0.000000	0.000000	0.000000	0.500000	12.952041	1.634679	[[205, 0], [123, 0]]
1	support vector machine	0.625000	0.000000	0.000000	0.000000	0.500000	12.952041	0.139009	[[205, 0], [123, 0]]
2	RandomForest	0.704268	0.724138	0.341463	0.464088	0.631707	10.214250	11.050999	[[189, 16], [61, 42]]
3	GassuianNB	0.628049	0.509434	0.219512	0.306818	0.546341	12.846803	0.093631	[[179, 26], [96, 27]]
4	KNN	0.658537	0.761905	0.130081	0.222222	0.552846	11.793741	50.114803	[[200, 5], [107, 16]]
5	XGB	0.640244	0.528090	0.382114	0.443396	0.588618	12.425638	4.709991	[[163, 42], [76, 47]]

Fig. 12: Performance Metrics

Random Forest worked the best to train the model, giving us an f1 score (Balanced with precision & recall) of around 70%.

V. Conclusion

Future cities would benefit from real-time monitoring and evaluation of water quality due to the advancement of machine learning techniques. This work presented the results of our most recent literature analysis and comparative recent studies on the assessment of water quality using big data analytics and machine learning models and methods. Finally, it offers a few insights into theproblems, demands, and needs of future studies. Environmental protection greatly benefits from the modelling and forecasting of water quality. The algorithm implemented in this work improves the performance of water quality classifiers. We previously examined the performance metrics of machine learning algorithms, andwe found that by utilizing Hyperparameter Tuning along with Random Forest Classifier, we delivered a better improvement in the execution of different performance metrics of the models using Hyperparameter Tuning. We have got better improvement in performance metrics. This strategy may be applied and improved for automated water quality monitoring.

References

[1] Azamathulla, H. M. 2013 2 – A Review on Application of Soft Computing Methods in Water Resources Engineering A2 – Yang, Xin-She. In: Metaheuristics in Water, Geotechnical and Transport Engineering (A. H. Gandomi, S. Talatahari & A. H. Alavi, eds). Elsevier, Oxford, pp. 27–41.

[2] Azamathulla, H. M. & Wu, F.-C. 2011 Support vector machine approach for longitudinal dispersion coefficients in natural streams. Appl. Soft Comput. 11 (2), 2902–2905

[3] World Health Organization, “Meeting the MDG drinking water andsanitation target: the urban and rural challenge of the decade”, Geneva, 2006

[4] L. Hu, C. Zhang, C. Hu, and G. Jiang, “Use of grey system for assessment of drinking water quality: a case study of Jiaozuo city, China”, Advances in Grey Systems Research, Springer Berlin Heidelberg, pp. 469-478, 2010.

[5] R. Rosly, M. Makhtar, M. K. Awang, M. N. A. Rahman, and M. M.Deris, “The Study on the Accuracy of Classifiers for Water Quality Application”, International Journal of u- and e- Service, Science and Technology, Vol. 8, No. 3, pp.145-154, 2015.

[6] D. Yang, L. Zheng, W. Song, S. Chen, and Y.Zhang, “Evaluationindexes and methods for water quality in ocean dumping areas”, Procedia Environmental Sciences: Proc. of the 7th International Conference on Waste Management and Technology, Vol. 16, pp.112-117, December 2012.

[7] Alley WM, Reilly TE, Franke OL (1999) Sustainability of ground- water resources: US Geological Survey Circular 1186, p79.

[8] Municipal Corporation of Tirupati, Draft City Sanitation Plan Volume I – Main report, (GIZ-ASEM, September 2011).

[9] APHA 2005. Standard methods for the examination of water and wastewater. American Public Health Association, Washington D.C.

[10] Nivruti T. Nirgude, Sanjay Shukla, and A.Venkatachalam. 2013. Physico- Chemical Analysis of Some Industrial Effluents from Vapi Industrial Area, Gujarat, India. Rasayan J. Chem, Vol. 6 | No.1 | 68-72.

[11] Heddam, S. 2016e Simultaneous modelling and forecasting of hourly dissolved oxygen concentration (DO) using radial basis function neural network (RBFNN) based approach: a case study fromthe Klamath River, Oregon, USA. Model. Earth Syst. Environ. 2 (3), 117–135.

[12] Herschy, R.1993 National and international standards in Streamflow measurement. Flow Meas. Instrum 4(1), 53-55.

[13] Ivakhnenko, A. G.1971 Polynomial theory of complex systems. IEEE Trans. Syst. Man Cybernet. 1 (4), 364–378.

[14] Jaddi, N. S. & Abdullah, S.2017 A cooperative- competitive masterslave global-best harmony search for ANN optimization and water-quality prediction. Appl. Soft Comput. 51, 209–224

[15] Jennings, G.2007 Water-based Tourism, Sport, Leisure, and Recreation Experiences. Elsevier, Oxford.

[16] Nasir, N., Kansal, A., Alshaltone, O., Barneih, F., Sameer, M., Shanableh, A. and Al-Shamma’a, A., 2022. Water quality classification

[17] Haghiabi, A.H., Nasrolahi, A.H. and Parsaie, A., 2018. Water quality prediction using machine learning methods. *Water Quality Research Journal*, 53(1), pp.3-13

[18] Rosero-Montalvo, P.D., López-Batista, V.F. and Peluffo- Ordóñez, D.H., 2022. A New Data-Preprocessing- Related Taxonomy of Sensors for IoT Applications. *Information*, 13(5), p.241

[19] Khan, Y. and See, C.S., 2016, April. Predicting and analyzing water quality using Machine Learning: a comprehensive model. In *2016 IEEE Long Island Systems, Applications and Technology Conference (LISAT)* (pp. 1-6). IEEE

[20] Zhou, J., Wang, Y., Xiao, F., Wang, Y. and Sun, L., 2018. Waterquality prediction method based on IGRA and LSTM.*Water*, 10(9),p.114