Nature Environment and Pollution Technology
An International Quarterly Scientific Journal

p-ISSN: 0972-6268
(Print copies up to 2016)
e-ISSN: 2395-3454

Vol. 21 | No. 2 | pp. 433-446 | 2022

Original Research Paper

https://doi.org/10.46488/NEPT.2022.v21i02.002

Open Access Journal

Estimating the Water Quality Class of a Major Irrigation Canal in Odisha, India: A Supervised Machine Learning Approach

S. K. Bhoi*, C. Mallick and C. R. Mohanty***†**

*Department of Computer Science and Engineering, Parala Maharaja Engineering College, Berhampur-761003, Odisha, India

**Department of Basic Science, Parala Maharaja Engineering College, Berhampur-761003, Odisha, India

***Department of Civil Engineering, Parala Maharaja Engineering College, Berhampur-761003, Odisha, India

†Corresponding author: C. R. Mohanty: chitta123@yahoo.com

Nat. Env. & Poll. Tech.
Website: www.neptjournal.com

Received: 26-06-2021
Revised: 24-07-2021
Accepted: 20-08-2021

Key Words:
Taladanda canal
Water quality class
Supervised machine learning
Prediction model
Classification accuracy

ABSTRACT

Contamination of surface water by rapid industrialization, natural and anthropogenic activities is of great concern over the last few decades. Nowadays, canal water systems are no exception to this form of contamination, which results in water quality degradation. To classify the canal water based on the Bureau of Indian Standards (BIS), it was thought to develop a quick and inexpensive approach as an alternative to the time-consuming analysis approach. With this motivation, the present study explores building a machine learning model for water quality classification of a major canal namely the Talaldanda canal operating in the state of Odisha, India. The water quality class is predicted using supervised machine learning (ML) prediction models for the new canal water input parameters. The water quality parameters such as pH, dissolved oxygen (DO), biochemical oxygen demand (BOD), and total coliform (TC) at six strategic locations of the canal from the year 2013-2020 were collected from Odisha State Pollution Control Board for the training phase. The supervised ML models used in the study are Decision Tree (DT), Neural Network (NN), k-NN (k-Nearest Neighbor), Naïve Bayes (NV), Support Vector Machine (SVM), and Random Forest (RF). The predictions of the models are evaluated using the Orange-3.29.3 data analytics tool. When analyzing the performance parameters by sampling the training data into training and testing using cross-validation, the results show that DT has a higher classification accuracy (CA) of 96.6 percent than other ML models. In addition, the likelihood of DT correctly predicting water quality class for the testing dataset is higher than that of other prediction models.

INTRODUCTION

Water is one of the most important renewables and finite natural resources on earth. Over the years, demand for freshwater for households, agriculture, and industrial use has led to the degradation of water quantity and quality of water bodies. Water pollution has, therefore, emerged as an important issue in India. The quality of water in the canals has also deteriorated drastically over the years due to the letting of sewage/sullage effluents, agricultural runoff carrying toxic chemicals, dumping of garbage and dead animals, and human defecation along the canal banks, etc. (Solanki et al. 2007, Shankar &Balasubramanya 2008, Rincy &Tessy 2010, Guru Prasad 2003). Many studies have found that natural and anthropogenic activities, as well as their physicochemical properties, are causing canal water quality to deteriorate.

Water quality parameters are classified into five categories based on their permissible limits and purposes - Class A, Class B, Class C, Class D, and Class E (BIS 1982). Due to imbalance in these parameters the water quality reduces and also the class level changes. This seriously affects human life. Therefore, regular monitoring of water quality is needed as it's a basic need of consumption (Prati 1971, Schaeffer &Konnanur 1977). Regular monitoring increases the size of data as the parameter size also increases. Therefore, some prediction models need to be implemented in this monitoring so that the time for manual evaluation is saved (Prati1971, Schaeffer &Konnanur 1977).

The delta area of the Mahanadi river basin in India primarily depends on Taladanda canal water as a source of drinking water for livestock, intensive agriculture, intensive aquaculture, poultry farms, and other purposes, with irrigation by far the largest user. Taladanda canal was dug in 1862 by the East India Company for irrigation purposes as well as to serve as a waterway in the coastal part of Odisha, India, and completed by the British government in 1869. This canal scheme is situated in the Mahanadi river basin having a command area in the two coastal districts of Odisha. The

canal is nearly 85 km long starting from Cuttack in Odisha, India, and ending at Athrabanki, Paradeep of India. The details of sampling locations and study maps are illustrated in Fig. 1.

Previous studies have revealed that DO concentrations in the Taladanda canal are influenced by environmental conditions upstream points and along the sections of the canal (Prusty& Biswal 2017, Das & Acharya 2003). The poor water quality in respect of pH, DO, BOD, and FC (Fecal Coliforms) in Taladanda Canal at Paradeep area is due to human activities and industrialization (Samantray et al. 2009, Prusty& Biswal 2020a, 2020b, Das &Panda 2010, Mishra 2012).

Several studies have been conducted to develop an effective machine learning-based model for water sample prediction and quality analysis. ML is a branch of Artificial Intelligence (AI) (Patro et al. 2020, Panda et al. 2020, Nayak et al. 2018) that deals with the problems of automation, optimization, etc. ML is divided into four types, 1. Supervised Learning, 2. Unsupervised Learning, 3. Semi-Supervised Learning, 4. Reinforcement Learning. A deep learning (DL) model with random forest, XGBoost, and ANN (artificial neural network)was used for the prediction of groundwater at Arang of Raipur district, India with an observation that DL was found to be better with higher classification accuracy (Singha et al. 2021). Bisht et al. (2019) employed prediction intelligence to predict water quality in the Ganga River in India using SVM, with a prediction accuracy of 96.66%. Other researchers such as Ahmed et al. (2019) predicted the water quality of different River Basins in India using Adaptive Neuro-Fuzzy Inference System (ANFIS), RBF-ANN (Radial Basis Function), and MLPNN (Multilayer Perceptron) (Ahmed et al. 2019, Aldhyani et al. 2020)

From the previous studies, it is observed that no studies are available to classify the Taladanda canal water quality by developing a quick and inexpensive technique as an alternative to the time-consuming analysis approach (Ross 1977). Therefore, these facts have motivated the investigators to conduct the study to classify the canal water using a supervised ML model.

MATERIALS AND METHODS

Study Area Description

The canal has many stations with the starting point as Jobra, then Ranihat, Chatrabazaar, Nuabazaar, Biribati, and Atharabanki as shown in Fig. 2. It is a nearly 150 years old canal built for irrigation, navigation, drinking, bathing, industrial water supply, municipality water supply, etc. However, the canal is contaminated and the water quality has degraded in recent years.

The main cause of this is the pollution of water, air, and soil. Industry wastes, medical wastes, plastics, carbon emis-

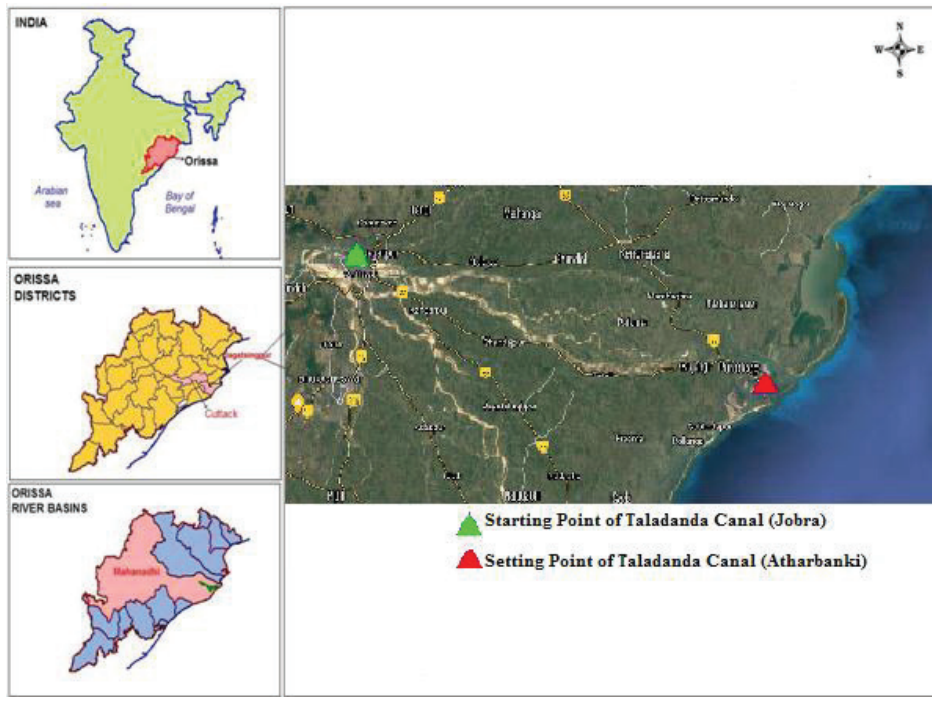

Fig. 1: Study area showing Taladanda Canal with starting and ending point.

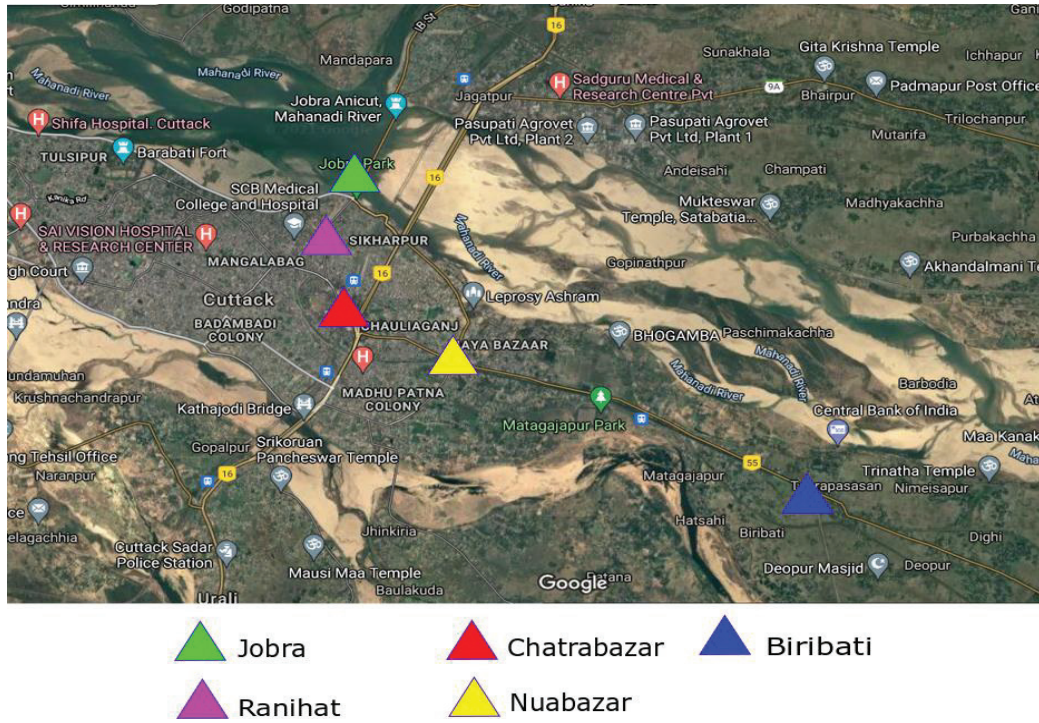

Fig. 2: Location of sample collection stations in the canal.

sions from industries, human wastes, hazardous chemicals waste from industries, are some of the main pollutants that reduce the water quality.

Methodology

In this section, methodologies adopted to classify the water quality of Indian rivers have been discussed.

Data collection: The data collected for the training phase is taken from State Pollution Control Board, Odisha, India from the year 2013 to 2020 (OPCB 2020). This data was collected year wise and it is the average of the observations taken per year. The data consists of 6 stations data of the canal. Due to the smaller dataset for training, we have also generated a synthetic dataset as per the permissible limits for Class A, Class B, and Class C only (Jayalakhmidevi & Belagadi 2005, Meenakumari & Hosmati 2003).

Data pre-processing: Raw data are always noisy and inconsistent. Data pre-processing helps in enhancing the quality of the model. As this data contains various missing values, the mean imputation technique is used to fill these missing values.

Training dataset: The training dataset from the year 2013 to 2020 was collected from Odisha State Pollution Control Board, India (OPCB 2020). The water quality parameters were pH, DO ($mg.L^{-1}$), BOD ($mg.L^{-1}$), TC ($MPN.100mL^{-1}$) and Existing Water Class (target class). Table 1 shows the water quality class and its intended use, whereas Table 2 shows the parameter value for pH, DO, BOD, and TC tolerance level for classes A to E.

Table 1: Water quality class.

Water Class	Purpose
A	Drinking after proper disinfection without conventional treatment
B	Outdoor Bathing
C	Drinking after proper disinfection with conventional treatment
D	Fishing and other animal activities
E	Irrigation and Industry cooling, etc.

Table 2: Water quality parameters of the classes.

Class	pH	DO	BOD	TC
A	6.5-8.5	6 and above	2 or less	50 or less
B	6.5-8.5	5 and above	3 or less	500 or less
C	6.5-8.5	4 and above	3 or less	5000 or less
D	6.5-8.5	4 and above		
E	6.5-8.5			

A sample of training data for the year 2013 is shown as follows for 6 stations of Taladanda Canal in Table 3: Sample average data of the year 2013.

For training, we have considered 2013-2020 water quality data. So, the number of instances is 6×8 = 48 instances (rows), where 6 is the number of stations per year and 8 is the number of years (2013-2020). So, it is a smaller dataset for training, therefore, we have generated data or instances for increasing the input size. For that, we have considered Table 2 tolerance levels and used pseudorandom number distribution to generate 100 instances for each Class A, Class B, and Class C except Classes D and E because they have other parameters for validation(BIS1982). So, the total numbers of instances generated are 48+100+100+100= 348 instances. These 348 instances are now used as input for training in a prediction model.

Testing dataset: For testing, we have collected data with 4 observations at each station in a year-wise manner from the year 2014 to 2018. The average of the observations is taken for each parameter. The data contains the same set of parameters for each year. So, the total instances used for testing are 6×5=30 instances, where six is the number of stations and five is the number of years for which the data is recorded.

Machine Learning and Prediction of Water Quality Class

In this work, we have considered six supervised machine learning prediction models for predicting the output or target class (water quality class) for the testing dataset. The models considered are discussed as follows (Aldhyani et al. 2020):

Neural network: NN solves the multiclass classification problem which will best suit for our work to predict the class as per the water quality parameters. It has more than one neuron or N neurons in the output layer which facilitates it to solve the multiclass classification problem. Mostly the last layer of the network is the softmax function that is an algebraic simplification of N number of logistic classification.

Table 3. It is the average data of the number of observations taken.

Taladanda Canal water monitoring station	pH	DO	BOD	TC	Existing water class
Jobra	7.9	7.6	3.7	58475	Other class (E)
Ranihat	7.7	8.3	8.9	106750	Other class (E)
Chatrabazar	7.6	5.3	7.8	116250	Other class (E)
Nuabazar	7.6	5.5	5.0	99000	Other class (E)
Biribati	7.8	7.4	4.8	66250	Other class (E)
Atharabanki	7.8	5.4	4.9	113317	Other class (E)

k-NN: It also solves the multiclass classification problem. It can use classification on regression. In this method, the target or output is a membership class. A member is classified based on the votes of its neighbors, with the member assigned to that class which is the most common from its k nearest neighbors.In this classification, the function is approximated in a local maximum, and all other computations are ignored until the function is evaluated. If k=1, that unknown member will be allocated to that one class.

Naïve bayes: It also solves the multiclass classification problem. It is mainly used for the construction of classifiers: by modeling, class labels for assignment of class labels to problems represented as vectors, where class labels are taken from sets with finite data. It assumes the value of a selected feature as independent of other features. The parameter estimation mainly uses the maximum likelihood method.

Decision tree: It also solves the multiclass classification problem. In this method, the non-leaf nodes are labeled with input features. The values of target characteristics are labeled on the arc from the indicated nodes. A class or probability distribution is assigned to a leaf node. The tree is built by separating the source set into root nodes and subsets like successors children. The categorization features are used to separate the data.This process is repeated recursively and called recursive partitioning. This process stops when the subset at a root node has the same values of target output or when no values are added to the prediction after splitting.

SVM: It also solves the multiclass classification problem. It mainly aims to maximize the margin by maximizing the minimum distance from the hyperplane to the nearest example. In this method, for the multiclass problem, additional parameters and constraints are implemented to efficiently classify or predict the classes.

Random forest: It also solves the multiclass classification problem. In this method, several decision trees are ensembled for classification purposes. Each tree in this forest outputs a prediction and the majority of votes for the class is called the output class. It is a faster and a flexible method to implement with some constraints.

Steps for Prediction of Water Class

The steps for predicting the water class are shown in Fig. 3 and discussed as follows:

Step 1: In the first step, the input is taken as the training dataset and fed into the ML model.

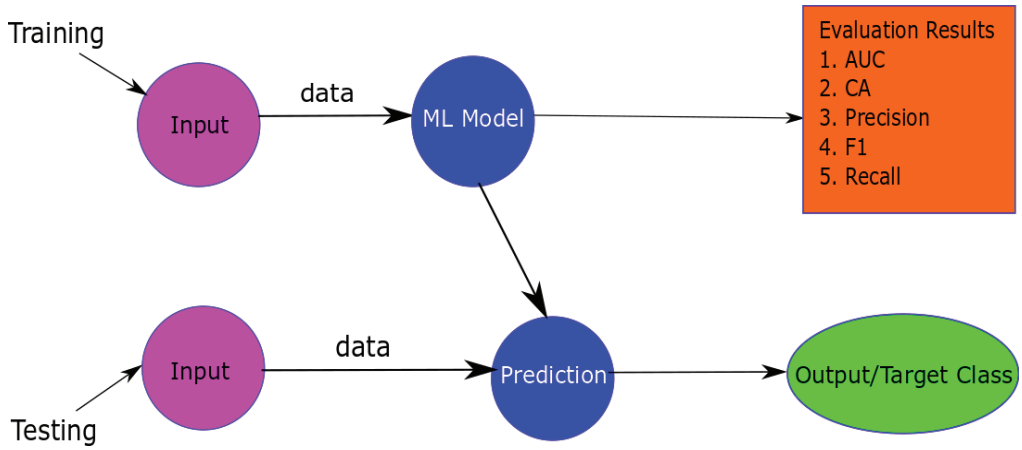

Fig. 3: Flowchart of a prediction model to estimate the water class.

Step 2: The ML model uses cross-validation as the sampling technique for training and testing.

Step 3: The evaluation results are then analyzed to know the model which has a higher classification accuracy (CA).

Step 4: That ML model with higher CA is selected for predicting the water class. However, we can take all ML models for testing.

Step 5: The testing is done by taking the testing dataset as input and fed into the prediction module to get the target class with a higher probability.

Step 6: The prediction model which shows a higher probability of estimation of water class for the input parameters is considered for taking the class data.

RESULTS AND DISCUSSION

The performance of the methodology is evaluated using Orange-3.29.3 data analytics tool installed in a Core-i3 machine with 8 Gb RAM, 2.4 GHz processor, and 64-bit Windows 10 OS platform.

From Fig. 4, it is observed that the input file of training

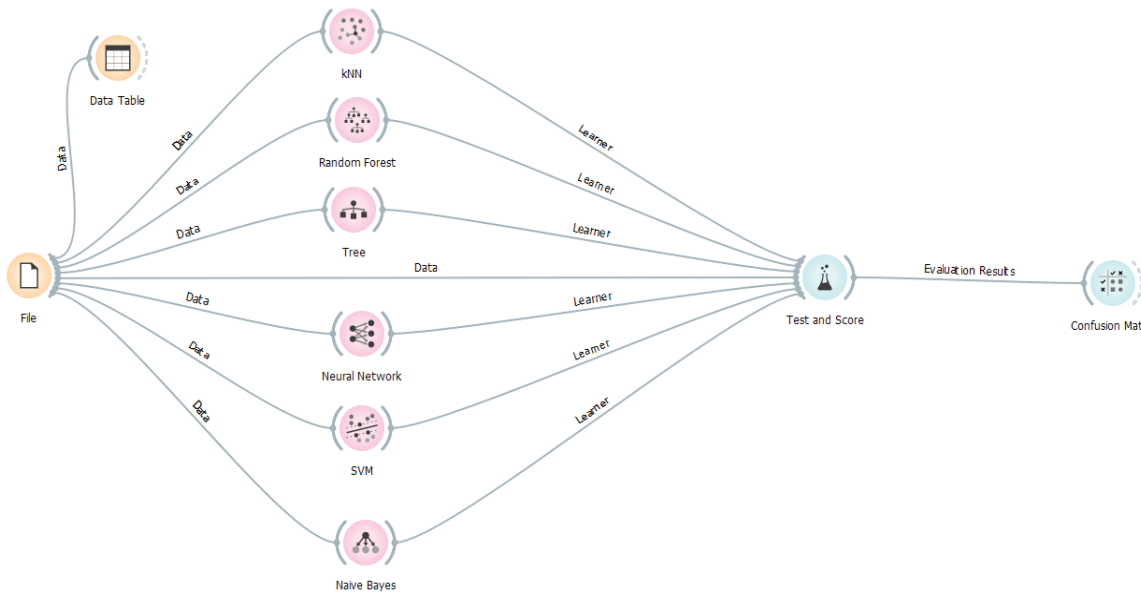

Fig.4: Orange workflow for evaluating the results of different ML algorithms.

is fed into the ML models for training using the cross-validation sampling technique with 10 folds. The test and score module shows the different performance parameters of the ML models and the confusion matrix shows how accurately the instances are predicted from the actual. The performance parameters taken are:

1. AUC (area under the curve): It describes how much the ML model classifies the classes well. The model with 100% accuracy of prediction has an AUC of 1.0.
2. CA (classification accuracy): The number of predictions made correct from the observed values is called CA. Eq. (1) shows the formula for CA:

$$CA = (TP+TN)/(TP+FP+TN+FN) \quad \ldots(1)$$

Where TP is the true positive, TN is the true negative, FP is the false positive, and FN is the false negative.

3. F1: The F1 score shows the harmonic mean of precision and recall to better understand accuracy. It is shown as follows in Eq. (2).

$$F1= (2*Precision*Recall)/(Precision+Recall) \quad \ldots(2)$$

4. Precision: It shows how many examples of a positive class are correctly classified out of the entire number of instances classified in that class? Eq. (3) shows the formula for precision:

$$Precision = TP/(TP+FP) \quad \ldots(3)$$

5. Recall: Recall means the proportion of instances correctly classified for a particular class. Eq. (4) shows the formula for Recall:

$$Recall = TP/(TP+FN) \quad \ldots(4)$$

From Table 4, it is observed that the CA of DT is greater than other ML models. So, we can conclude that this model will be better for prediction. It is also seen that the AUC of RF is better, F1 is better in DT, Precision is better in RF, and Recall is better in DT. Those results can be visualized from Fig. 5 to 9 respectively. However, we have taken the main parameter as CA for prediction.

The confusion matrix (CM) mainly shows the actual number of instances predicted accurately. As we know, we have taken 348 instances in the training set. The diagonal

Table 4: Performance parameters generated from Orange tool for evaluation.

Model	AUC	CA	F1	Precision	Recall
kNN	0.988	0.951	0.951	0.954	0.951
Tree	0.979	0.966	0.966	0.966	0.966
SVM	0.859	0.477	0.430	0.438	0.477
RF	0.996	0.966	0.965	0.967	0.966
NN	0.846	0.598	0.571	0.566	0.598
Naïve Bayes	0.929	0.799	0.796	0.804	0.799

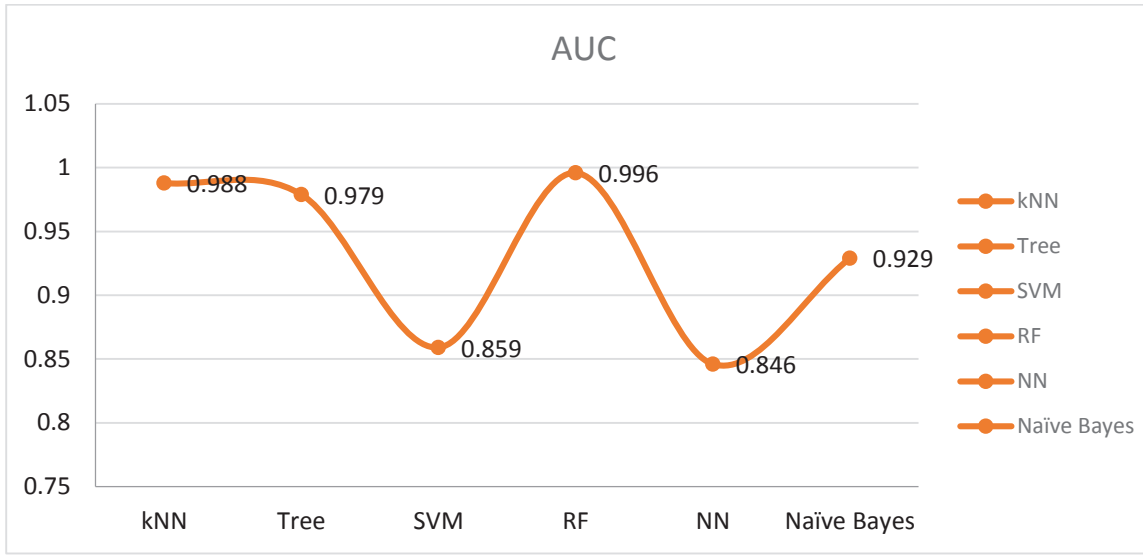

Fig. 5: Performance of AUC of all ML models.

matrix of a CM shows the number of instances accurately predicted for a particular class. Therefore, Fig. 10 (a-e) shows the confusion matrix of different ML models. The left part is marked with "Actual" which means the actual instances and the top part is marked with "Predicted" which shows the actual instances to be predicted accurately.

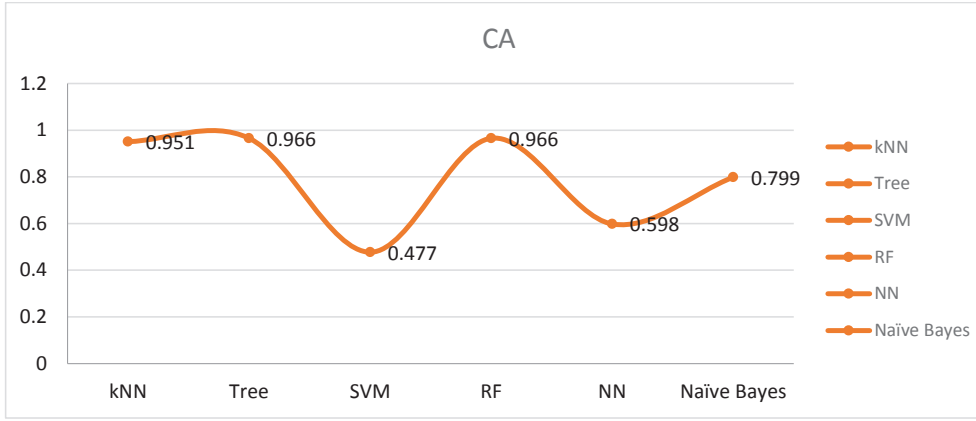

Fig. 6: Performance of CA of all ML models.

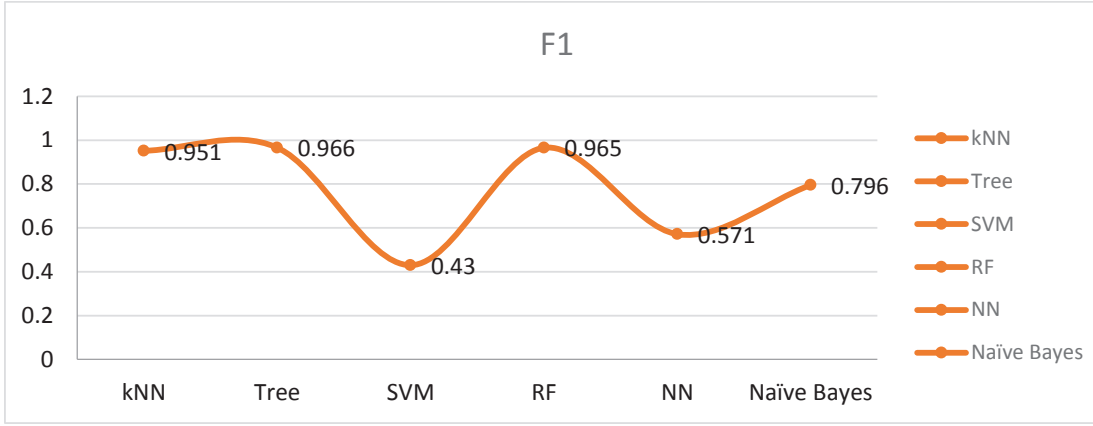

Fig. 7: Performance of F1 of all ML models.

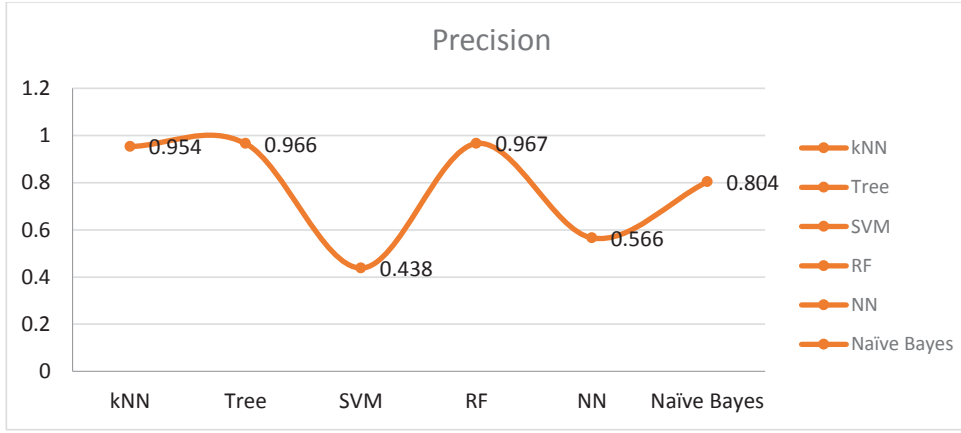

Fig. 8: Performance of precision of all ML models.

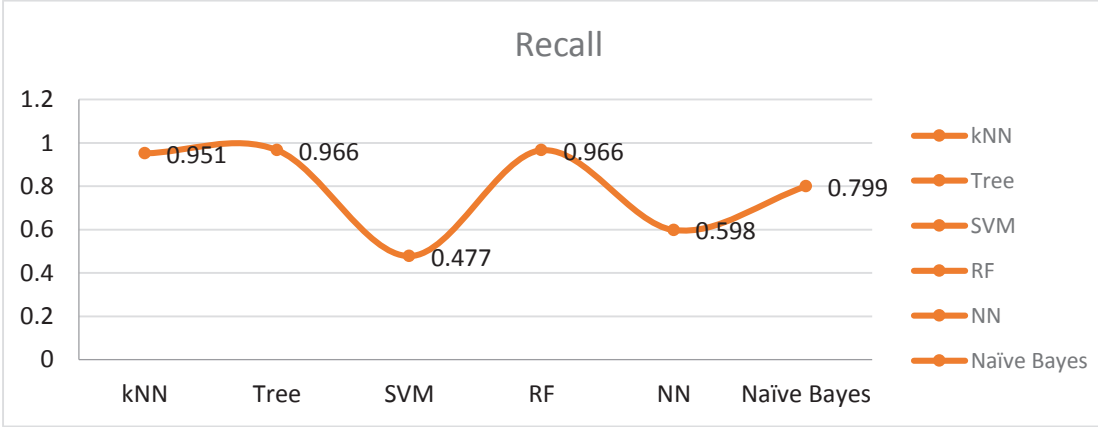

Fig. 9: Performance of Recall of all ML models.

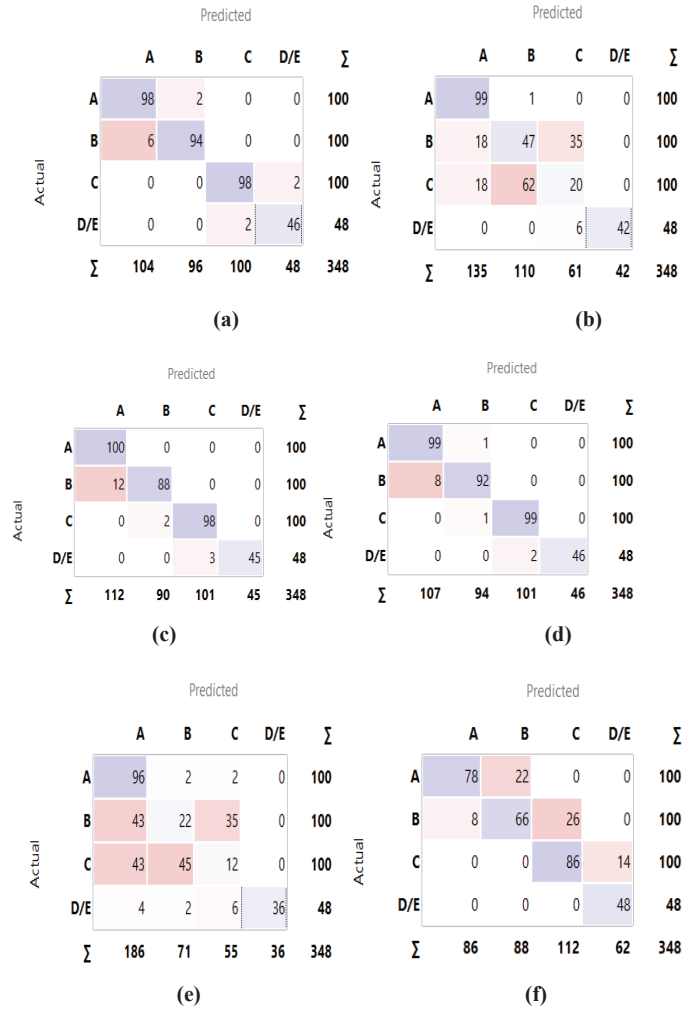

Fig. 10: Confusion matrix for 348 instances (a) Tree (b) NN (c) k-NN (d) RF (e) SVM (f) NB

Fig. 11 shows the prediction model design using the Orange workflow. The prediction module for each ML model is forecasted with a probability value for a certain class for each case (testing set) (0.00 to 1.00). Therefore, our main goal is to find that water class that has a higher probability of prediction for the ML model. From the results shown in Fig. 11 to 14, it is observed that for water class D/E the probability is higher using the DT ML model. So, we can use that data as the prediction data (water class) for each instance (input testing set with four parameters). Fig. 12 shows that for class A, the probabilities of prediction are very low for each ML model. Fig. 13 shows the probabilities of prediction for each ML model for class B. Fig. 14 shows the probabilities of prediction for each ML model for class C. Fig. 15 shows the probabilities of prediction for each ML model for class D/E.From this figure, it is observed that DT predicts the classes with a probability of 1.00 for every 30 instances, and the average probability is 1.0. For other ML models, the average of these probabilities of 30 instances is smaller than DTs average probability. Therefore, we consider the DT column of Fig. 15 as the classes predicted with a high probability for each instance (input).

CONCLUSION

Using supervised machine learning (ML) prediction models, the water quality class is predicted for the new input parameters of the water samples taken in this study. The input parameters considered for the study are pH, dissolved oxygen (DO), biochemical oxygen demand (BOD), total coliform (TC), and target water class. The predictions of the models are evaluated using the Orange-3.29.3 data analytics tool. From the results, it was found that DT shows a higher classification accuracy (CA) of 0.966 than other ML models. Also, for the testing dataset, the average probability of prediction of water quality class D/E for the DT ML model is 1.00, which is greater than other prediction models. So, for these types of datasets, DT will be a better prediction model to categorize the water class well. In the future, the model may be implemented in predicting the water quality for other sources of water like rivers, ponds, groundwater in different locations. The main research challenge is the improvement of prediction accuracy of water pollution levels in larger or smaller datasets in different water sources. These research areas need to be explored.

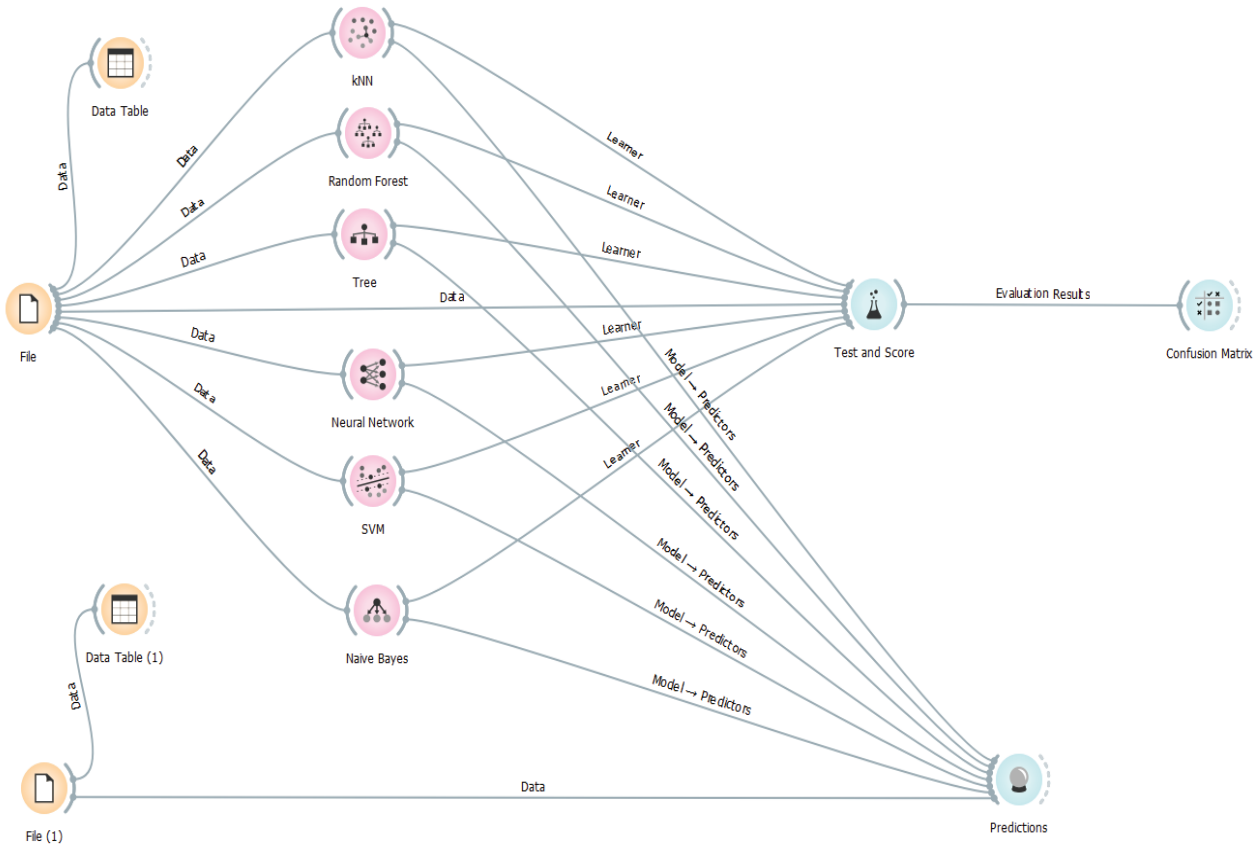

Fig. 11: Orange workflow for prediction of water class for the testing input.

	kNN	Random Forest	Tree	Neural Network	SVM	Naive Bayes
1	0.00 → D/E	0.00 → C	0.00 → D/E	0.00 → D/E	0.00 → D/E	0.01 → C
2	0.00 → D/E	0.00 → C	0.00 → D/E	0.00 → D/E	0.00 → D/E	0.01 → C
3	0.00 → D/E	0.00 → C	0.00 → D/E	0.00 → D/E	0.00 → D/E	0.01 → C
4	0.00 → D/E	0.00 → C	0.00 → D/E	0.00 → D/E	0.00 → D/E	0.01 → C
5	0.00 → D/E	0.00 → C	0.00 → D/E	0.00 → D/E	0.00 → D/E	0.01 → C
6	0.00 → D/E	0.00 → C	0.00 → D/E	0.06 → C	0.06 → C	0.01 → C
7	0.00 → D/E	0.00 → D/E	0.00 → D/E	0.00 → D/E	0.00 → D/E	0.01 → C
8	0.00 → D/E	0.00 → C	0.00 → D/E	0.00 → D/E	0.00 → D/E	0.01 → C
9	0.00 → D/E	0.00 → C	0.00 → D/E	0.00 → D/E	0.00 → D/E	0.01 → C
10	0.00 → D/E	0.00 → C	0.00 → D/E	0.00 → D/E	0.00 → D/E	0.01 → C
11	0.00 → D/E	0.00 → D/E	0.00 → D/E	0.00 → D/E	0.00 → D/E	0.01 → C
12	0.00 → D/E	0.00 → C	0.00 → D/E	0.00 → D/E	0.00 → D/E	0.01 → C
13	0.00 → D/E	0.00 → D/E	0.00 → D/E	0.04 → C	0.04 → C	0.01 → C
14	0.00 → D/E	0.00 → C	0.00 → D/E	0.00 → D/E	0.00 → D/E	0.01 → C
15	0.00 → D/E	0.00 → C	0.00 → D/E	0.00 → D/E	0.00 → D/E	0.01 → C
16	0.00 → D/E	0.00 → C	0.00 → D/E	0.00 → D/E	0.00 → D/E	0.01 → C
17	0.00 → D/E	0.00 → C	0.00 → D/E	0.00 → D/E	0.00 → D/E	0.01 → C
18	0.00 → D/E	0.00 → D/E	0.00 → D/E	0.00 → D/E	0.00 → D/E	0.01 → C
19	0.00 → D/E	0.00 → D/E	0.00 → D/E	0.04 → C	0.05 → C	0.01 → C
20	0.00 → D/E	0.00 → D/E	0.00 → D/E	0.00 → D/E	0.00 → D/E	0.01 → C
21	0.00 → D/E	0.00 → D/E	0.00 → D/E	0.00 → D/E	0.00 → D/E	0.01 → C
22	0.00 → D/E	0.00 → D/E	0.00 → D/E	0.00 → D/E	0.00 → D/E	0.01 → C
23	0.00 → D/E	0.00 → D/E	0.00 → D/E	0.01 → D/E	0.00 → D/E	0.01 → C
24	0.00 → D/E	0.00 → C	0.00 → D/E	0.00 → D/E	0.00 → D/E	0.01 → C
25	0.00 → D/E	0.00 → C	0.00 → D/E	0.07 → C	0.03 → C	0.01 → C
26	0.00 → D/E	0.00 → C	0.00 → D/E	0.00 → D/E	0.00 → D/E	0.01 → C
27	0.00 → D/E	0.00 → C	0.00 → D/E	0.00 → D/E	0.00 → D/E	0.01 → C
28	0.00 → D/E	0.00 → C	0.00 → D/E	0.00 → D/E	0.00 → D/E	0.01 → C
29	0.00 → D/E	0.00 → C	0.00 → D/E	0.00 → D/E	0.00 → D/E	0.01 → C
30	0.00 → D/E	0.00 → C	0.00 → D/E	0.05 → C	0.01 → C	0.01 → C

Fig. 12**:** Prediction of class A using different ML models for the input instances with probability values (0.00-1.00).

	kNN	Random Forest	Tree	Neural Network	SVM	Naive Bayes
1	0.00 → D/E	0.00 → C	0.00 → D/E	0.01 → D/E	0.00 → D/E	0.02 → C
2	0.00 → D/E	0.00 → C	0.00 → D/E	0.00 → D/E	0.00 → D/E	0.02 → C
3	0.00 → D/E	0.00 → C	0.00 → D/E	0.00 → D/E	0.00 → D/E	0.02 → C
4	0.00 → D/E	0.00 → C	0.00 → D/E	0.00 → D/E	0.00 → D/E	0.02 → C
5	0.00 → D/E	0.00 → C	0.00 → D/E	0.01 → D/E	0.00 → D/E	0.02 → C
6	0.00 → D/E	0.00 → C	0.00 → D/E	0.32 → C	0.24 → C	0.02 → C
7	0.00 → D/E	0.00 → D/E	0.00 → D/E	0.02 → D/E	0.00 → D/E	0.02 → C
8	0.00 → D/E	0.00 → C	0.00 → D/E	0.00 → D/E	0.00 → D/E	0.02 → C
9	0.00 → D/E	0.00 → C	0.00 → D/E	0.00 → D/E	0.00 → D/E	0.02 → C
10	0.00 → D/E	0.00 → C	0.00 → D/E	0.00 → D/E	0.00 → D/E	0.02 → C
11	0.00 → D/E	0.00 → D/E	0.00 → D/E	0.00 → D/E	0.00 → D/E	0.02 → C
12	0.00 → D/E	0.00 → C	0.00 → D/E	0.00 → D/E	0.00 → D/E	0.02 → C
13	0.00 → D/E	0.00 → D/E	0.00 → D/E	0.36 → C	0.34 → C	0.02 → C
14	0.00 → D/E	0.00 → C	0.00 → D/E	0.00 → D/E	0.00 → D/E	0.02 → C
15	0.00 → D/E	0.00 → C	0.00 → D/E	0.00 → D/E	0.00 → D/E	0.02 → C
16	0.00 → D/E	0.00 → C	0.00 → D/E	0.00 → D/E	0.00 → D/E	0.02 → C
17	0.00 → D/E	0.00 → C	0.00 → D/E	0.00 → D/E	0.00 → D/E	0.02 → C
18	0.00 → D/E	0.00 → D/E	0.00 → D/E	0.00 → D/E	0.00 → D/E	0.02 → C
19	0.00 → D/E	0.00 → D/E	0.00 → D/E	0.37 → C	0.35 → C	0.02 → C
20	0.00 → D/E	0.00 → D/E	0.00 → D/E	0.00 → D/E	0.00 → D/E	0.02 → C
21	0.00 → D/E	0.00 → D/E	0.00 → D/E	0.00 → D/E	0.00 → D/E	0.02 → C
22	0.00 → D/E	0.00 → D/E	0.00 → D/E	0.00 → D/E	0.00 → D/E	0.02 → C
23	0.00 → D/E	0.00 → D/E	0.00 → D/E	0.05 → D/E	0.00 → D/E	0.02 → C
24	0.00 → D/E	0.00 → C	0.00 → D/E	0.01 → D/E	0.00 → D/E	0.02 → C
25	0.00 → D/E	0.00 → C	0.00 → D/E	0.24 → C	0.09 → C	0.02 → C
26	0.00 → D/E	0.00 → C	0.00 → D/E	0.00 → D/E	0.00 → D/E	0.02 → C
27	0.00 → D/E	0.00 → C	0.00 → D/E	0.00 → D/E	0.00 → D/E	0.02 → C
28	0.00 → D/E	0.00 → C	0.00 → D/E	0.00 → D/E	0.00 → D/E	0.02 → C
29	0.00 → D/E	0.00 → C	0.00 → D/E	0.00 → D/E	0.00 → D/E	0.02 → C
30	0.00 → D/E	0.00 → C	0.00 → D/E	0.22 → C	0.05 → C	0.02 → C

Fig. 13: Prediction of class B using different ML models for the input instances with probability values (0.00-1.00).

	kNN	Random Forest	Tree	Neural Network	SVM	Naive Bayes
1	0.00 → D/E	0.50 → C	0.00 → D/E	0.05 → D/E	0.00 → D/E	0.66 → C
2	0.00 → D/E	0.50 → C	0.00 → D/E	0.01 → D/E	0.00 → D/E	0.66 → C
3	0.00 → D/E	0.50 → C	0.00 → D/E	0.00 → D/E	0.00 → D/E	0.66 → C
4	0.00 → D/E	0.50 → C	0.00 → D/E	0.01 → D/E	0.00 → D/E	0.66 → C
5	0.00 → D/E	0.50 → C	0.00 → D/E	0.03 → D/E	0.00 → D/E	0.66 → C
6	0.00 → D/E	0.50 → C	0.00 → D/E	0.45 → C	0.27 → C	0.66 → C
7	0.00 → D/E	0.40 → D/E	0.00 → D/E	0.09 → D/E	0.00 → D/E	0.85 → C
8	0.00 → D/E	0.50 → C	0.00 → D/E	0.00 → D/E	0.00 → D/E	0.66 → C
9	0.00 → D/E	0.50 → C	0.00 → D/E	0.00 → D/E	0.00 → D/E	0.66 → C
10	0.00 → D/E	0.50 → C	0.00 → D/E	0.00 → D/E	0.00 → D/E	0.66 → C
11	0.00 → D/E	0.40 → D/E	0.00 → D/E	0.02 → D/E	0.00 → D/E	0.85 → C
12	0.00 → D/E	0.50 → C	0.00 → D/E	0.01 → D/E	0.00 → D/E	0.66 → C
13	0.20 → D/E	0.40 → D/E	0.00 → D/E	0.48 → C	0.34 → C	0.85 → C
14	0.00 → D/E	0.50 → C	0.00 → D/E	0.00 → D/E	0.00 → D/E	0.66 → C
15	0.00 → D/E	0.50 → C	0.00 → D/E	0.00 → D/E	0.00 → D/E	0.66 → C
16	0.00 → D/E	0.50 → C	0.00 → D/E	0.00 → D/E	0.00 → D/E	0.66 → C
17	0.00 → D/E	0.50 → C	0.00 → D/E	0.00 → D/E	0.00 → D/E	0.66 → C
18	0.00 → D/E	0.40 → D/E	0.00 → D/E	0.01 → D/E	0.00 → D/E	0.85 → C
19	0.20 → D/E	0.40 → D/E	0.00 → D/E	0.48 → C	0.34 → C	0.85 → C
20	0.00 → D/E	0.40 → D/E	0.00 → D/E	0.00 → D/E	0.00 → D/E	0.85 → C
21	0.00 → D/E	0.40 → D/E	0.00 → D/E	0.00 → D/E	0.00 → D/E	0.85 → C
22	0.00 → D/E	0.40 → D/E	0.00 → D/E	0.02 → D/E	0.00 → D/E	0.85 → C
23	0.00 → D/E	0.40 → D/E	0.00 → D/E	0.19 → D/E	0.00 → D/E	0.85 → C
24	0.00 → D/E	0.50 → C	0.00 → D/E	0.04 → D/E	0.00 → D/E	0.66 → C
25	0.00 → D/E	0.50 → C	0.00 → D/E	0.43 → C	0.12 → C	0.66 → C
26	0.00 → D/E	0.50 → C	0.00 → D/E	0.00 → D/E	0.00 → D/E	0.66 → C
27	0.00 → D/E	0.50 → C	0.00 → D/E	0.00 → D/E	0.00 → D/E	0.66 → C
28	0.00 → D/E	0.50 → C	0.00 → D/E	0.01 → D/E	0.00 → D/E	0.88 → C
29	0.00 → D/E	0.50 → C	0.00 → D/E	0.01 → D/E	0.00 → D/E	0.66 → C
30	0.00 → D/E	0.50 → C	0.00 → D/E	0.41 → C	0.06 → C	0.66 → C

Fig. 14: Prediction of class C using different ML models for the input instances with probability values (0.00-1.00).

	kNN	Random Forest	Tree	Neural Network	SVM	Naive Bayes
1	1.00 → D/E	0.50 → C	1.00 → D/E	0.93 → D/E	1.00 → D/E	0.31 → C
2	1.00 → D/E	0.50 → C	1.00 → D/E	0.99 → D/E	1.00 → D/E	0.31 → C
3	1.00 → D/E	0.50 → C	1.00 → D/E	1.00 → D/E	1.00 → D/E	0.31 → C
4	1.00 → D/E	0.50 → C	1.00 → D/E	0.99 → D/E	1.00 → D/E	0.31 → C
5	1.00 → D/E	0.50 → C	1.00 → D/E	0.96 → D/E	1.00 → D/E	0.31 → C
6	1.00 → D/E	0.50 → C	1.00 → D/E	0.17 → C	0.43 → C	0.31 → C
7	1.00 → D/E	0.60 → D/E	1.00 → D/E	0.88 → D/E	1.00 → D/E	0.12 → C
8	1.00 → D/E	0.50 → C	1.00 → D/E	1.00 → D/E	1.00 → D/E	0.31 → C
9	1.00 → D/E	0.50 → C	1.00 → D/E	1.00 → D/E	1.00 → D/E	0.31 → C
10	1.00 → D/E	0.50 → C	1.00 → D/E	1.00 → D/E	1.00 → D/E	0.31 → C
11	1.00 → D/E	0.60 → D/E	1.00 → D/E	0.97 → D/E	1.00 → D/E	0.12 → C
12	1.00 → D/E	0.50 → C	1.00 → D/E	0.99 → D/E	1.00 → D/E	0.31 → C
13	0.80 → D/E	0.60 → D/E	1.00 → D/E	0.12 → C	0.27 → C	0.12 → C
14	1.00 → D/E	0.50 → C	1.00 → D/E	1.00 → D/E	1.00 → D/E	0.31 → C
15	1.00 → D/E	0.50 → C	1.00 → D/E	1.00 → D/E	1.00 → D/E	0.31 → C
16	1.00 → D/E	0.50 → C	1.00 → D/E	1.00 → D/E	1.00 → D/E	0.31 → C
17	1.00 → D/E	0.50 → C	1.00 → D/E	1.00 → D/E	1.00 → D/E	0.31 → C
18	1.00 → D/E	0.60 → D/E	1.00 → D/E	0.99 → D/E	1.00 → D/E	0.12 → C
19	0.80 → D/E	0.60 → D/E	1.00 → D/E	0.12 → C	0.26 → C	0.12 → C
20	1.00 → D/E	0.60 → D/E	1.00 → D/E	1.00 → D/E	1.00 → D/E	0.12 → C
21	1.00 → D/E	0.60 → D/E	1.00 → D/E	1.00 → D/E	1.00 → D/E	0.12 → C
22	1.00 → D/E	0.60 → D/E	1.00 → D/E	0.97 → D/E	1.00 → D/E	0.12 → C
23	1.00 → D/E	0.60 → D/E	1.00 → D/E	0.75 → D/E	1.00 → D/E	0.12 → C
24	1.00 → D/E	0.50 → C	1.00 → D/E	0.96 → D/E	1.00 → D/E	0.31 → C
25	1.00 → D/E	0.50 → C	1.00 → D/E	0.26 → C	0.76 → C	0.31 → C
26	1.00 → D/E	0.50 → C	1.00 → D/E	1.00 → D/E	1.00 → D/E	0.31 → C
27	1.00 → D/E	0.50 → C	1.00 → D/E	1.00 → D/E	1.00 → D/E	0.31 → C
28	1.00 → D/E	0.50 → C	1.00 → D/E	0.99 → D/E	1.00 → D/E	0.09 → C
29	1.00 → D/E	0.50 → C	1.00 → D/E	0.99 → D/E	1.00 → D/E	0.31 → C
30	1.00 → D/E	0.50 → C	1.00 → D/E	0.32 → C	0.88 → C	0.31 → C

Fig. 15: Prediction of class D/E using different ML models for the input instances with probability values (0.00-1.00).

ACKNOWLEDGMENTS

The authors acknowledge the support provided by Parala Maharaja Engineering College, Berhampur, Odisha, India.

REFERENCES

Ahmed, A.N., Faridah B.O., Haitham, A.A., Rusul, K.I., Chow, M.F., Md Shabbir, H., Mohammad, E. and Ahmed, E. 2019. Machine learning methods for better water quality prediction. J. Hydrol., 578: 124084.

Aldhyani, T.H., Al-Yaari, M., Alkahtani, H. and Maashi, M. 2020. Water quality prediction using artificial intelligence algorithms. Appl. Bion. Biomech., 5: 1-12.

Bisht, A.K., Singh, R., Bhutiani, R. and Bhatt, A. 2019. Application of Predictive Intelligence in Water Quality Forecasting of the River Ganga Using Support Vector Machines. In Predictive Intelligence Using Big Data and the Internet of Things, IGI Global, Pennsylvania, USA, pp. 206-218.

Bureau of Indian Standards (BIS). 1982. Tolerance Limits of Parameters for Use. https://www.indiawaterportal.org/articles/indian-standard-drinking-water-bis-specifications-10500-1991

Das, J. and Acharya, B.C. 2003. Hydrology and assessment of lotic water quality in Cuttack City, India. Water, Air Soil Pollut., 150(1): 163-175.

Das, M. and Panda, T. 2010. Water quality and phytoplankton population in sewage fed river of Mahanadi, Orissa, India. J. Life Sci., 2(2): 81-85.

Guru Prasad, B. 2003. Evaluation of water quality in Tadepallimandal of Guntur district, AP. Nat. Environ. Pollut. Technol., 2(3): 273-276

Jayalakhmidevi, O. and Belagadi, S.L. 2005. Water quality assessment from different districts of southern Karnataka. Nat. Environ. Pollut. Technol., 4: 589-596

Meenakumari, H.R. and Hosmati, S.P. 2003. Bacteriological examination of groundwater samples in and around Mysore city, Karnataka, India. Nat. Environ. Pollut. Technol., 2(2): 213-215

Mishra, L. 2012. Water quality assessment and modeling of Cuttack city. Doctoral Thesis, National Institute of Technology, Rourkela, Odisha, pp. 1-42.

Nayak, S.K. and Panda, S.K. 2018. A user-oriented collaborative filtering algorithm for recommender systems. In 2018 Fifth International Conference on Parallel, Distributed and Grid Computing (PDGC), 20-22 December 2018, Jaypee University of Information Technology, India, IEEE, India, pp. 374-380.

Odisha Pollution Control Board (OPCB). 2019. Annual Report. http://ospcboard.org/publications-reports/

Panda, S.K., Bhoi, S.K. and Singh, M. 2020. A collaborative filtering recommendation algorithm based on a normalization approach. J.Amb. Intell. Hum. Comput., 11: 1-23.

Patro, S.G.K., Mishra, B.K., Panda, S.K., Kumar, R. and Apoorva, A. 2020. Hybrid Social Recommender Systems for Electronic Commerce: A Review. In 2020 International Conference on Computer Science, Engineering and Applications (ICCSEA), 13-14 March 2020, Gunupur, Indian, IEEE, India, pp. 1-6.

Patro, S.G.K., Mishra, B.K., Panda, S.K., Kumar, R., Long, H.V., Taniar, D. and Priyadarshini, I., 2020. A hybrid action-related K-nearest neighbor (HAR-KNN) approach for recommendation systems. IEEE Access, 8: 90978-90991.

Prati, L. 1971. Assessment of water quality by a single index of pollution. Water Res., 5: 741-751

Prusty, R. and Biswal, T. 2017. Water quality assessment of Taladanda Canal in the command area of Cuttack city. Int. J. Adv. Agric. Sci. Technol., 4: 40-48.

Prusty, R. and Biswal, T. 2020a. Assessment of Pollution Load in Terms of Water Quality Index and Modelling of Taladanda Canal and Mahanadi River in Paradip Area, Odisha, India. Asian J. Water, Environ. Pollut., 17(4), 59-72.

Prusty, R. and Biswal, T. 2020b. Physico-chemical, bacteriological, and health hazard effect analysis of the water in Taladanda Canal, Paradip area, Odisha, India. J. Groundw. Sci. Eng., 8(4): 338-348.

Rincy, J. and Tessy, P.P. 2010. Water quality and pollution status of Chalakudy river at KathiKudam, Thrissur District, Kerala, India. Nat. Environ. Pollut. Technol., 9(1): 113-118.

Ross, S.L. 1977. An index system for classifying river water quality. Water Pollut. Control, 76(1): 113-122

Samantray, P., Mishra, B.K., Panda, C.R. and Rout, S.P. 2009. Assessment of water quality index in Mahanadi and Atharabanki Rivers and Taldanda Canal in Paradip area, India. J.Hum. Ecol., 26(3): 153-161.

Schaeffer, D.J. and Konnanur, G.J. 1977. Communicating environmental information to the public: A water quality index. J. Environ. Edu., 8: 16-26

Shankar, B.S. and Balasubramanya, N. 2008. Evaluation of quality indices for the groundwaters of an industrial area in Bangalore, India. Nat. Environ. Pollut. Technol., 7(4): 663-666

Singha, S., Pasupuleti, S., Singha, S.S., Singh, R. and Kumar, S. 2021. Prediction of groundwater quality using efficient machine learning technique. Chemosphere, 276: 130265.

Solanki, V.R., Murthy, S.S., Kour, A. and Sabita Raja, S. 2007. Variation in dissolved oxygen and biochemical oxygen demand in two freshwater lakes of Bodhan, Andhra Pradesh, India. Nat. Environ. Pollut. Technol., 6(4): 623-628.