River Water Quality Prediction and index classification using Machine Learning

River Water Quality Prediction and index classification using Machine Learning

Jitha P Nair [1], Vijaya M S [2]

[1] Research Scholar, Department of Computer Science, PSGR Krishnammal College for Women, Peelamedu, Coimbatore, India.

[2] Associate Professor, Department of Computer Science, PSGR Krishnammal College for Women, Peelamedu, Coimbatore, India.

jithapnair20@gmail.com

Abstract: Various pollutants have had a substantial impact on the quality of water in recent years. The quality of water directly impacts human health and the environment. The water quality index (WQI) is an indicator of effective water management. Water quality modelling and prediction have become essential in the fight against water pollution. The research aims to build an efficient prediction model for river water quality and to categorize the index value according to the water quality standards. The data has been collected from eleven sampling stations located in various locations across the Bhavani River, which flows through Kerala and Tamilnadu. The water quality index is determined by 27different parameters affecting water quality like dissolved oxygen, temperature, pH, alkalinity, hardness, chloride, coliform, etc. Data normalization and feature selection are done to construct the dataset to develop machine learning models. Machine learning algorithms such as linear regression, MLP regressor, support vector regressor and random forest has been employed to build a water quality prediction model. Support vector machines (SVM), naïve bayes, decision trees, MLP classifiers, have been used to develop a classification model for classifying water quality index. The experimental results revealed that the MLP regressor efficiently predicts the water Quality index with root mean squared error as 2.432, MLP classifier classifies the water quality index with 81% accuracy. The developed models show promising output concerning water quality index prediction and classification.

Keywords: River water quality, Prediction model, Classification model, Exploratory data analysis, Machine learning algorithms.

1. Introduction

Water is the most vital resource for life, as it is required for the survival of all living things, including humans. Better quality and quantity of water is essential for life on earth. Some pollution levels are satisfactory to aquatic species, but when the level increases the oxygen content in the water decreases and lead to disasters. A large percentage of environmental water sources, such as lakes, rivers, and streams, have

quality standards that demonstrate their worth. Guidelines are applied to all types of water bodies for all applications and uses.

Irrigation water does not need to be either too saline or harmful to the plant or soil, thus ruining the ecosystem. Water quality also requires different qualities based on certain various processes for industrial applications. Natural water resources are among the cheapest options for freshwater, such as ground and surface water. Human and industrial activity and other natural processes can pollute natural resources. So, rapid industrial growth has led to a significant decline in water quality. The quality of drinking water is significantly affected by the infrastructure, lacking public awareness, and poor hygiene standards. The effects of contaminated drinking water are quite severe health issues, the environment, and infrastructure. According to UN research, around 1.5 million people die every year from water-borne diseases. 80% of health problems have been reported to be caused by polluted water in developing countries. Annual reports include 2.5 billion humans affected by water-borne diseases and five million deaths. The death rate of humans mainly focuses on crimes, accidents and terrorist attacks.

Novel approaches to analyzing and forecasting water quality (WQ) are critical. It is recommended that the temporal dimension of predicting water quality patterns be studied to monitor the seasonal shift of the WQ. However, using a specific model variation to forecast water quality outcomes performs better than using a single model. Several approaches to predicting and simulating water quality are being proposed. Statistical techniques, visual modelling, algorithm analysis, and predictive algorithms are commonly used. Multivariate statistical techniques were used to determine the correlation and relationship between different water quality parameters. For transitional probability, regression analysis, multivariate interpolation, and geostatistical approaches were used.

Massive population growth, the use of fertilizers and pesticides, the industrial revolution, seem to have serious consequences for water quality environments. The models for predicting water quality are extremely useful for monitoring water contamination. Modelling and predicting water quality are employed with mechanism oriented and no-mechanism-oriented models. The mechanism model is sophisticated and it simulates the water quality using advanced system structure data, it is regarded as a multifunctional model that can be applied to any water body.

This paper aims to develop an accurate model for forecasting river water quality using a developed framework. River water quality data has been collected from eleven sampling stations across Bhavani River which flows through Kerala and Tamilnadu and analyzed to understand feature distribution and correlation. Machine learning algorithms are used to predict water quality index values and to classify water quality indexes. Mean squared error, mean absolute error and root mean squared error is used for performance evaluation of prediction algorithms. Water quality index classification models are evaluated using the metrics such as accuracy, precision, recall, and F1-score.

2. Literature review

This study analyzes the approaches that were used to effectively address water quality challenges. In most studies, traditional statistical analysis and lab analysis is implemented to determine water quality, but other studies apply machine learning approaches to find an optimal solution to the water quality problem.

Sillberg et al. [1] have developed a machine learning-based technique integrating attribute-realization (AR) and support vector machine (SVM) to classify the Chao Phraya River water quality. Using the linear

function, the AR has identified the most important elements for improving river quality. NH3-N, TCB, FCB, BOD, DO, and Sal was the most contributing characteristics in the categorization, with contributed values in the range of 0.80–0.98, compared to 0.25–0.64 for TDS, Turb, TN, SS, NO3-N, and conductivity. The best classification results were achieved using the SVM linear approach, which had an accuracy of 0.94, a precision average of 0.84, a recall average of 0.84, and an F1-score average of 0.84. When applied to three to six parameters, the validation revealed that AR-SVM was a powerful method for identifying river water quality with 0.86–0.95 accuracy.

Yilma et al. [2] used an artificial neural network to simulate the Akaki river water quality index. The index was calculated using twelve water quality indicators from 27 dry and wet season sample locations. All projected findings, except one upstream location, have revealed low water quality. Through 12 inputs and one output, the neural network model was trained and verified using the number of hidden layers (2–20), hidden layer neurons (5, 10, 15, 20, and 25), transfer training, and learning functions. According to their research, artificial neural network with eight hidden layers and 15 hidden neurons accurately predicted the WQI with an accuracy of 0.93.

Ding et al. [3] have designed a hybrid intelligent method that incorporates Principal Component Analysis (PCA), Genetic Algorithm (GA), and Back Propagation Neural Network (BPNN) techniques for predicting river water quality. This research included 23 different water quality indicator variables, each of which has a complex non-linear relationship with water quality. PCA boosted the training pace of follow-up algorithms substantially, while GA optimized the BPNN parameters. According to the findings, the average prediction rates for non-polluted and polluted water quality were 88.9% and 93.1%, respectively, while the overall prediction rate was about 91%.

Ahmed et al. [4] assessed the water quality index (WQI) using supervised machine learning algorithms, where an individual index was used to summarise the overall quality of water and water quality class. The proposed techniques, as well as gradient boosting with a learning rate of 0.1 and polynomial regression with a degree of 2, most effectively predicted the WQI, and that WQI was later evaluated with a mean absolute error (MAE) of 1.9642 and 2.7273. In this case, the MLP with the configuration (3, 7) has the highest accuracy for classification with 85.07%.

Zhang, Zhu, Yue, and Wong [5] presented an anomaly detection algorithm for water quality data that employs dual time-moving windows to identify anomaly data from historical patterns in real time. The algorithm was built on statistical models, specifically the auto-regressive linear combination model. The authors tested the algorithm with three month PH data from a river basin and analysed using real water quality monitoring station data. The experimental results depicted that the algorithms have a lower rate of false positives and better anomaly detection performance than the AD and ADAM algorithms.

Sakizadeh [6] used 16 water quality metrics and Bayesian regularization to predict the WQI. The research work found correlation coefficients of 0.94 and 0.77 between observed and predicted values.

The majority of the studies used manual lab analysis, failed to calculate the water quality index standard, and even included so many parameters. As seen, machine learning can produce good results for detecting anomalies in water quality, and the existing work is inspired by related work. Machine learning algorithms have the potential to significantly reduce the number of incorrect predictions.

3. Overview of Study Area and Dataset

Bhavani river flows through Tamilnadu and Kerala, India. The river originates from Nilgiri hills, then enter silent valley national park, Kerala and flows through Tamilnadu. Bhavani river is a long perennial long with 217 km and is fed by both southwest and northeast monsoons. Its watershed drains 0.62 million hectares spread across Tamil Nadu (87%), Kerala (9%), and Karnataka (4%). Fig. 1. depicts the flow of the Bhavani River, which flows primarily through the Attappady Plateau of Palakkad district and then travels to the Tamil Nadu districts of Coimbatore and Erode. Approximately 90% of the river water is used for agricultural irrigation. Data collected from different stations of the Bhavani River includes Kottathara, Thavalam, Chalayur, Karathur, Cheerakuzhi, Elachivazhi, Badrakaliamman kovil, Sirumugai, Bhavanisagar, Bhavani, Sathyamangalam.

Fig.1. Map of Bhavani River

3.1 Data Collection

The main stations in the Bhavani River basin include Kottathara, Thavalam, Chalayur, Karathur, Cheerakuzhi, Elachivazhi, Badrakaliamman kovil, Sirumugai, Bhavanisagar, Bhavani, Sathyamangalam. Temperature, pH, dissolved oxygen, turbidity, chloride, and other parameters which determine the water quality index are collected from the eleven-sampling station of the Bhavani River. The 10560 data samples with 31 attributes used in this research work were collected from the water quality monitoring stations for the period from January 1, 2016, to December 31, 2020. The time fluctuation trend of water quality indicators of the Bhavani River's water source was examined for these samples. The values of each water quality parameter over sometime were represented as time-series data. Annual trends in temperature, pH, conductivity, BOD, COD, Nitrate-N, fluoride, potassium, TC, FDS, chloride, sodium, TDS, hardness and sulphate are shown in Fig.2. The sample data of river water quality monitored from eleven sampling stations is shown in Table 1.

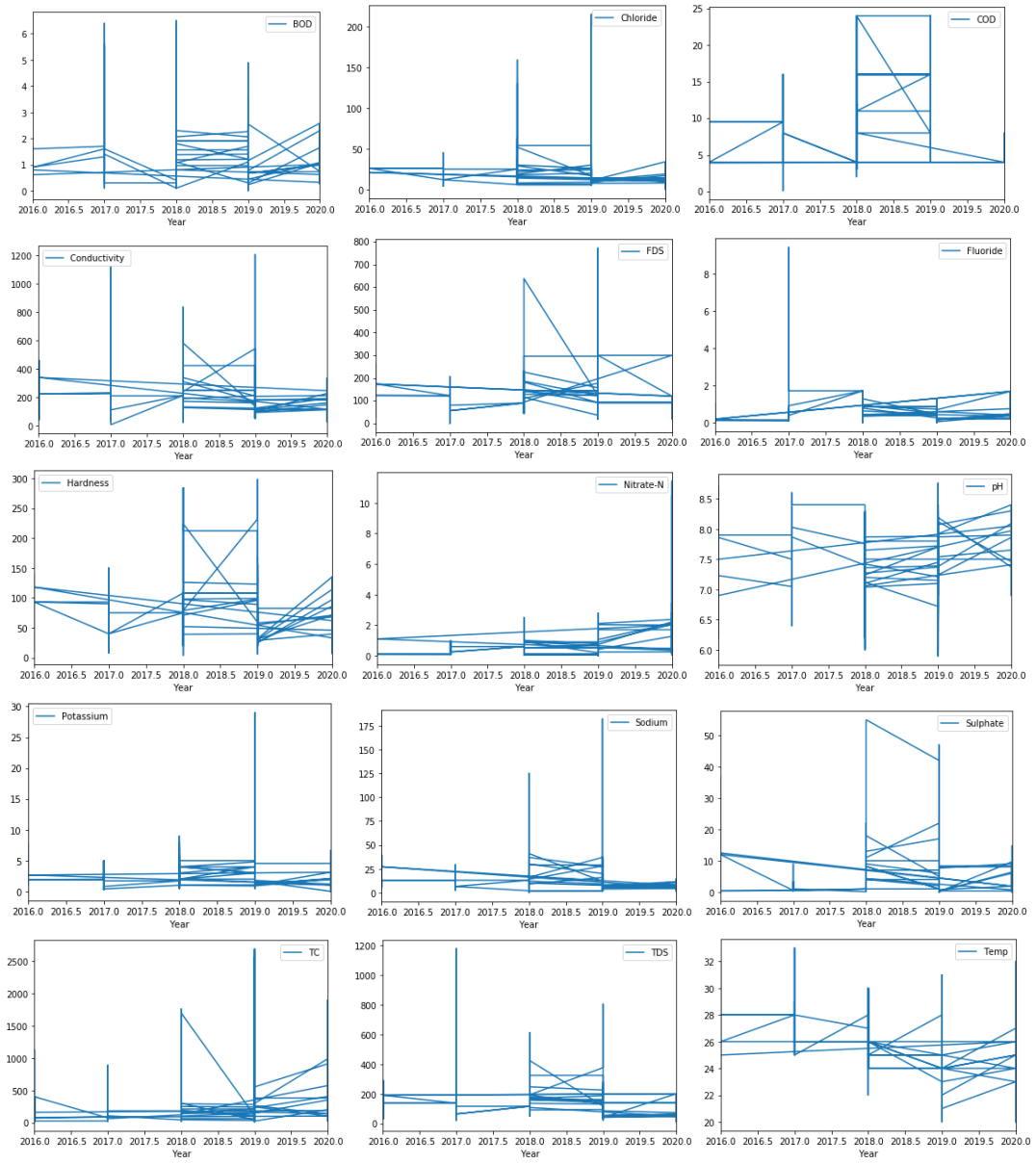

Fig. 2. Variation trend of Bhavani River from 2016 to 2020 of some parameters

Table1: Data Collected from Sampling Stations

Date	02/06/2016	05/02/2016	05/09/2016	03/03/2019	04/06/2019	04/07/2019	05/04/2018	08/03/2018	10/02/2018	06/05/2018`	12/03/2018
Temp	25	26	25	24	23	24	26	26	24	24	26
pH	7.15	7.5	6.9	7.87	8.12	8.07	7.06	7.82	7.27	7.2	7.24
Conductivity	340	340	340	116	95	102	132	316	189	583	240
Chloride	21	21	21	11	13	9	14	33	22	52	19
COD	4	3.9	4	4	4	4	8	8	16	24	11
Sodium	27.1	27.1	27.1	5.26	4.21	5.13	10	38	9	41	16
TSS	300	300	300	69	56	61	4	6	18	4	1
TDS	190	190	190	41	47	52	108	213	177	424	193
FDS	174	174	174	300	300	300	94	144	140	638	115
DO	6.99	7.51	7.4	7.24	7.97	7.58	6.8	6.93	5.43	5.4	7.47
Nitrate	1.1	1.1	1.1	0.243	0.734	1.062	0.9	0.87	0.9	1	0.912
TC	88	79.414	160	147	163	295	130	192	155	1700	191
FC	80	80	80	39	51	18	27	109	56	430	150
Station	1	2	3	4	5	6	7	8	9	10	11
Latitude	76.591123E	76.72055E	76.657335E	76.657335E	76.72055E	76.683688E	77.408333E	77.001587E	77.114998E	77.682777E	77.23003E
Longitude	11.093055N	11.188917N	11.147534N	11.147534N	11.188917N	11.163695N	11.4975N	11.329323N	11.474167N	11.433611N	11.506945N
Year	2016	2016	2016	2019	2019	2019	2018	2018	2018	2018	2018

3.2 Calculation of the WQI: The water quality index is measured using the weighted arithmetic water quality index method. The most commonly measured water quality parameters like pH, BOD, COD, DO, Nitrate, sodium, sulphate, chloride, faecal coliform, etc. are calculated according to the following formula (1):

$$WQI = \frac{\Sigma wiqi}{\Sigma wi} \tag{1}$$

qi is a relative value of water quality that is specific to each parameter and i rep the number of parameters taken into consideration. Wi is a factor that measures recent the importance of a parameter in the calculation of the WQI index is the relative weight. Qi is calculated by applying formula 2 below.

$$qi = 100 * \frac{vi - vo}{si - vo} \tag{2}$$

where: v_i represents the value experimentally and determined the analyzed parameter, v_o represents the ideal value of that parameter whereas the ideal value is zero for all other parameters except DO = 14.6 mg/l and pH = 7.0. Si represents the standard, legally accepted, value for the water category in which the analyzed water sample was included. The w_i factor is calculated by using formula 3.

$$wi = \frac{K}{si} \tag{3}$$

where K is a constant, which can result from applying the formula 4,

$$K = \frac{1}{\Sigma\left(\frac{1}{si}\right)} \tag{4}$$

Tables 2 represent the unit weight of each parameter and their permissible limits for finding WQI. Based on the value obtained for the weighted arithmetic WQI method, the water ecological status can be determined, as illustrated by Table 3.

Table 2: Parameters with Permissible limits and Unit Weights

Parameters	**Permissible Limits**	**Weights**
Temp(oC)	28	0.035714286
pH	8.5	0.117647059
Conductivity	150	0.006666667
Turbidity	5	0.2
Phenolpth Alkalinity	20	0.05
Total Alkalinity	200	0.005
Chloride	250	0.004
COD	10	0.1
TKN	100	0.01
Ammonia	50	0.02
Hardness	100	0.01
Ca. Hardness	75	0.013333333
Mg. Hardness	30	0.033333333
Sulphate	200	0.005
Sodium	200	0.005
TSS	300	0.003333333
TDS	1000	0.001
FDS	200	0.005
Phosphate	0.3	3.333333333
Boron	1	1
Potassium	2.5	0.4
BOD	3	0.333333333
Fluoride	1.5	0.666666667
DO	7.5	0.133333333
Nitrate-N	0.503	1.988071571
TC	100	0.01
FC	60	0.016666667

Table 3: Weight Arithmetic Water Quality Index standards for Water Quality

WQI	Water Quality Class	Water Quality
0-30	A	Excellent
31-60	B	Good
61-90	C	Poor
91-120	D	Very Poor
>121	E	Unsuitable

Calculation of the water quality index has been done using the above formulas and follows the standards of weighted arithmetic water quality index mentioned in Table 3. Water quality index calculation requires permissible values and unit weight of each parameter as shown in Table 2.

After calculating the water quality index value for each sample, it was added to the respective instance. Water quality index class for each instance is identified using the existing standards and added as class labels to the respective instance. Finally, the river water quality dataset is developed with 33 attributes and 10560 labelled instances.

3.3 Exploratory Data Analysis

Exploratory data analysis was applied to the Bhavani River water quality dataset obtained from eleven sampling stations with 27 water quality parameters from 2016 to 2020. Unsupervised pattern recognition and display algorithms were used to extract correlations and similarities between variables and categorise river water quality samples data. A water quality dataset with 33 attributes and 10560 instances is used for analyzing and visualizing the importance of each attribute in the work. Univariate and multivariate analysis was performed using various interesting statistical graphs such as heatmaps, histograms, pair plots, and box plots. The heatmap analysis shows the correlation of the parameters with the water quality index, some parameters such as boron and TSS were negatively correlated and had duplicate instances. The boxplot and histogram analysis shows that the parameters such as conductivity, total coliform have a wide range of values, and hence data normalization is required for these parameters to set a smaller range. Thus, exploratory data analysis provided insights that led to data preprocessing.

3.4 Data Preprocessing

Preprocessing data helps to increase data quality and efficiency. The quality of raw data is harmed by its inconsistency and noise. The exploratory data analysis performed in this research work provided a proper understanding of the data that the dataset contains duplicates, negatively correlations among the parameters and distribution of values.

Data Cleaning: The practice of rectifying duplicate, incorrect, or incomplete data from a dataset is known as data cleaning. The river water quality dataset developed in this work contains 33 attributes and 10560 instances. The two attributes such as boron and TSS are negatively correlated and these two attributes are not contributive to water quality index prediction. Hence these two attributes are removed from the dataset for building an efficient model.

Normalization Method: Normalization is a technique used for standardizing the attribute values in the dataset. The result of EDA shows that the two parameters of the river quality dataset such as conductivity

and total coliform n to be normalized to a specific range as the range of values of the parameters is wide. Total coliform has a minimum value of 10 and a maximum value of 2500 but most of its values lie in the range of 10 to 300. Similarly, conductivity has a minimum value of 1 and a maximum value of 1200, most of the values lie between 60 and 210. Here Z-score normalisation is applied to these two parameters for standardizing the parameter values by using the mean and standard deviation. It is done using the following formula:

$$v' = \frac{v - \bar{A}}{\sigma A} \quad (5)$$

4. Methodology

The proposed framework of the WQI prediction model consists of various building blocks such as data collection and data exploration, data preprocessing, construction of WQI prediction and classification models, performance evaluation. Machine learning approaches have been used to build the prediction of water quality index value and classification. Various metrics such as mean squared error, root mean squared error, R^2 is used for evaluating the performance of the prediction model, and the WQI classifier is evaluated using accuracy, precision, recall, F1 score. The architecture of the proposed WQI prediction model is shown in Fig.3. and described below.

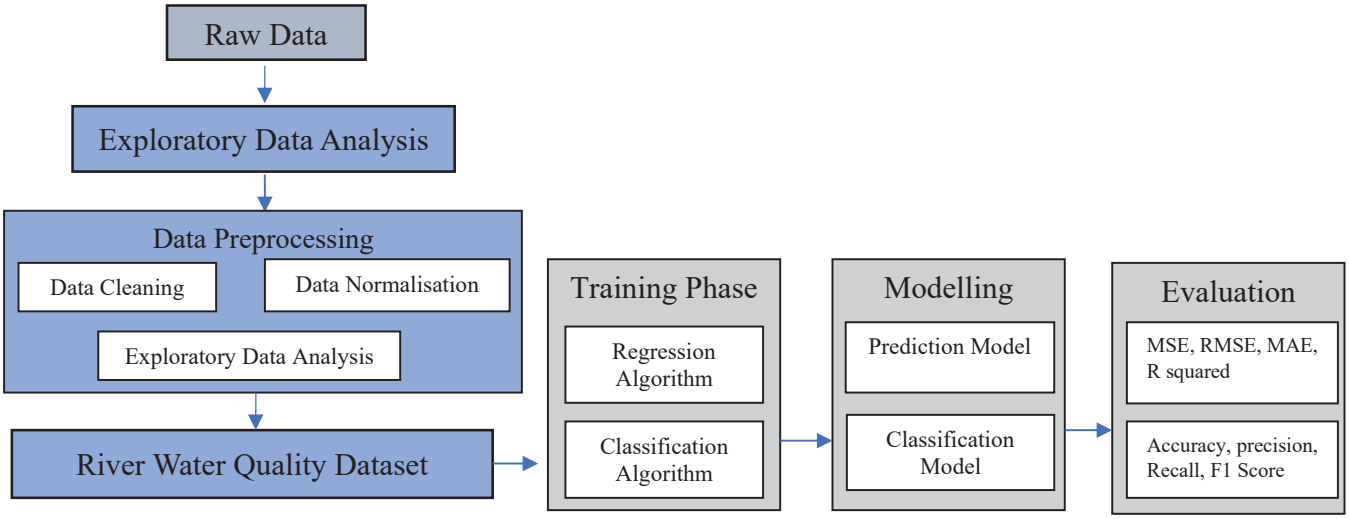

Fig 3. A framework of the proposed work

4.1 Feature Selection

The goal of the feature selection method is to exclude features that do not make a significant contribution to defining the water quality index. Feature selection has an impact on the measurement of data dimensions, whether for training data or testing data. Select K best, is applied to the developed dataset to find the best contributive attributes in determining the water quality index. Parameters such as phenolpth alkalinity, ammonia, TKN, and phosphate have less importance in predicting the water quality index. The parameters which are having high importance in developing the water quality index are mainly conductivity, alkalinity,

chloride, hardness, sulphate, sodium, phosphate, potassium, BOD, fluoride, DO, nitrate, coliform and so on. This feature selection process yielded a better river water quality dataset which contains 10560 parameters and 27 attributes for building the prediction and classification model.

4.2 Building WQI Prediction model

The water quality prediction model is developed by learning the trends in the river water quality dataset using linear regression, random forest, support vector regression, and multilayer perceptron regression. Here water quality index is the target variable and all other parameters in the dataset are independent variables in modelling the regression process. Regression can be used to forecast a response using a new set of predictors.

Linear regression is used to find the best-predicted weights, that is, the weights with the smallest residuals. They usually minimise the sum of squared residuals (SSR) for all observations to get the best weights. This method is known as the method of ordinary least squares. The regression model is evaluated to ensure that it matches the data required to forecast and predict true or false occurrences.

Support Vector Regression is a supervised algorithm that is derived from wide margin kernel methods for analysis and is used to predict discrete values. It has all of the properties of support vector machine maximal margin algorithms, such as duality, sparseness, kernel, and convexity. It has evolved into a powerful predictive data analysis technique with a wide range of applications. The basic concept of SVR is to find the best fit line and the best fit line in SVR is the hyperplane with the maximum number of points.

Random forest is a supervised machine learning algorithm used for regression problems by handling continuous variables containing datasets. Random forests are built from subsets of data, and the final output is based on average or majority ranking, which eliminates the problem of overfitting. The random forest chooses observations at random, builds a decision tree, and uses the average result.

An MLP is a type of feed-forward artificial neural network that has at least three layers of nodes (neurons), including an input layer, one or more hidden layers, and an output layer. From the input to the output, the nodes are fully coupled in the form of a directed graph. Except for the input nodes, all nodes have an associated activation function that is used to compute the node output using weighted inputs from other nodes. The weights between the nodes are updated iteratively for reducing the error function in an MLP model, which is trained to utilize a backpropagation mechanism with gradient-descent as an optimization algorithm. For all hidden layer nodes in regression, a relu activation function is employed, and for the output layer nodes, a linear activation function is used. The result is a predicted real-valued quantity based on the input sample $x \epsilon X$.

4.3 Building WQI classification model.

The water quality index classification model is developed by learning the patterns in the river water quality dataset. The water quality index is used as a class label to categorise the water quality into five standards-based water quality parameters. Machine learning algorithms such as Support vector machine (SVM), Decision Tree, Naive Bayes, and Neural Networks, are used to construct classification models.

SVM is a supervised algorithm that can be used to classify and predict data. Each data point in n-dimensional space is displayed independently, allowing the two classes to be identified easily. In the fields of technology, pattern recognition, and learning classification, SVMs are gaining traction. A linear or non-linear separation surface can be used to classify the input region. In support vector classification, the separation function is a linear combination of kernels related to the support vector.

A decision tree classifier has a simple structure that can be kept in a small amount of space and efficiently classifies fresh data. Decision tree classifiers can automatically choose features and reduce complexity, and their tree structure provides easily understood and interpretable information about the classification's predictive or generalization capacity. The main goal of the decision tree growth algorithm is to determine which characteristic to test at each node in the tree**.**

The Naive Bayes algorithm is a form of classification method based on Bayes' theorem when the target value is set then the other attributes are independent variables. The Bayesian technique employs probability and statistical knowledge to predict and classify datasets. Using the Bayesian approach with prior and posterior probability, the bias and overfitting concepts of applying sample information can be avoided.
MLP Classifier is a machine-learning-based classification system. For all hidden layer nodes, a relu activation function is utilized, and for the output layer nodes, a softmax activation function is employed. The result is a vector that contains the odds that sample x∈X belongs to each class, which is the same as a categorical probability distribution. The class with the highest probability is the final result. For each class label per input, a distinct loss is calculated, and the outcome is the total of all those losses.

4.4 Performance Evaluation
The performance of the developed models in predicting the water quality index and classifying the water quality index is evaluated to find the best algorithm. The efficient prediction algorithms are found when the RMSE value is very less and the best classification model is evaluated by checking the accuracy. The following are the statistical parameters that were used:

1. Mean Squared Error (MSE) :

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_a - Y_b)^2 \quad (6)$$

2. Mean Absolute Error (MAE)

$$MAE = abs(Y_a - Y_b) \quad (7)$$

3. Root Mean Squared Error (RMSE)

$$RMSE = \sqrt{(Y_a} - Y_b)^2/_n \quad (8)$$

where Y_a and Y_b are the actual responses and the predicted value, respectively, and n is the total number of variables.

4. Accuracy

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \quad (9)$$

5. Precision

$$Precision = \frac{TP}{Tp+FP} \tag{10}$$

6. Recall

$$Recall = \frac{TP}{TP+FN} \tag{11}$$

7. F1-Score

$$F1 - Score = \frac{2*precision*recall}{Precision+recall} * 100 \tag{12}$$

where TP, TN, FP, and FN are the true positive, true negative, false positive, and false negative, respectively. Evaluation of machine learning algorithms for developing water quality index in prediction and classification are evaluated using the above equations to check the efficiency of algorithm using river water data.

5. Experiment

Here the experiments have been carried out to develop an accurate water quality index prediction model and the corresponding classification model using the Bhavani River water quality dataset. A dataset with 10560 instances and 27 attributes has been split into training and testing sets with 80%of instances for training and 20% of instances for testing. The support vector regressor, linear regression, random forest, and MLP Regressor algorithms were implemented to build water quality index prediction model and were implemented to build water quality index classification model using support vector machine, Naive Bayes, decision tree, MLP classifier. Machine learning algorithms developed a relationship with the independent and dependent parameters while modelling and then the test data is used to determine whether the models are effective with respect to various performance evaluation metrics.

5.1 Water Quality Prediction

In this section regression models are developed by training river water quality dataset through implementing MLP regressor, linear regression, support vector regressor, and random forest using python libraries. The performance of the prediction models is evaluated for its efficiency to forecast the WQ using mean squared error, mean absolute error, root mean squared error, and R-squared values.

From the prediction results, it was observed that mean absolute error value of support vector regressor based prediction model is 3.623, whereas prediction models based on linear regression, random forest and MLP regressor gives 2.85911, 2.015, and 1.9143 respectively. Thus, high error rate was produced by support vector regressor and less error by MLP regressor as illustrated in Fig.4a.

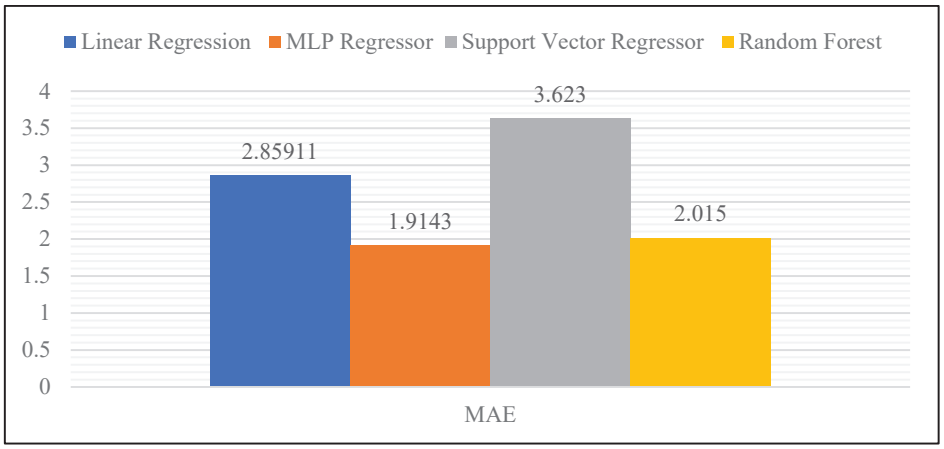

Fig.4a. Mean Absolute Error of Prediction Models

Similarly, it was found that the root mean squared error value of support vector regressor based prediction model is 4.3281, whereas prediction models based on linear regression, random forest and MLP regressor gives 3.8428, 3.79783, and 2.432 respectively. Thus, high error rate was produced by support vector regressor and less error by MLP regressor as shown in Fig.4b.

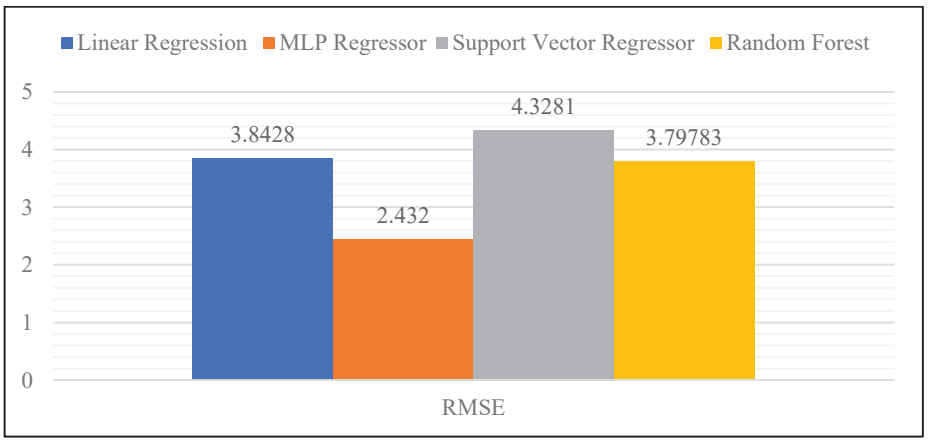

Fig. 4b. Root Mean Squared Error of prediction models

In the next case, the mean squared error value of support vector regressor based prediction model is 20.237, whereas linear regression, random forest and MLP regressor shows the respective metric values for prediction as 11.3318, 9.8278, and 7.1032. Thus, high error rate was produced by support vector regressor and less error by MLP regressor as depicted in Fig.4c.

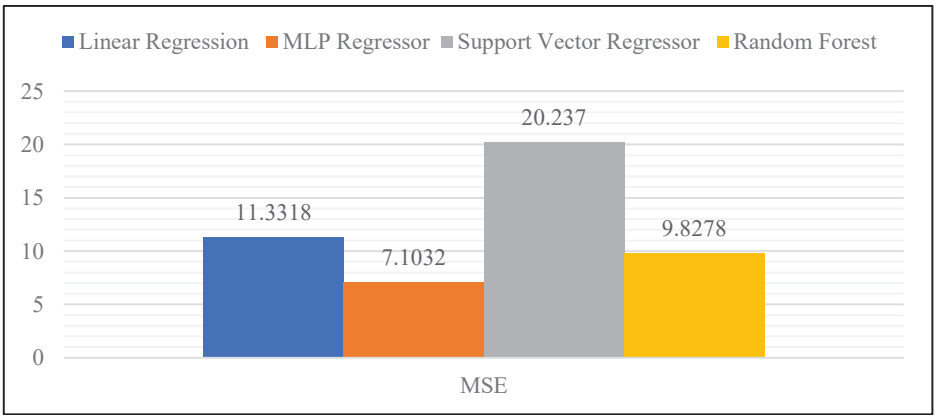

Fig. 4c. Mean Squared Error of Prediction Models

It was found that the R-squared value of support vector regressor based prediction model is -2.7132, whereas prediction models based on linear regression, random forest and MLP regressor yields 0.6375, 0.6923, and 0.7342 correspondingly. Thus, R-squared value is negatively produced by support vector regressor which means it has less accuracy and have high accuracy for MLP regressor as shown in Fig.4d.

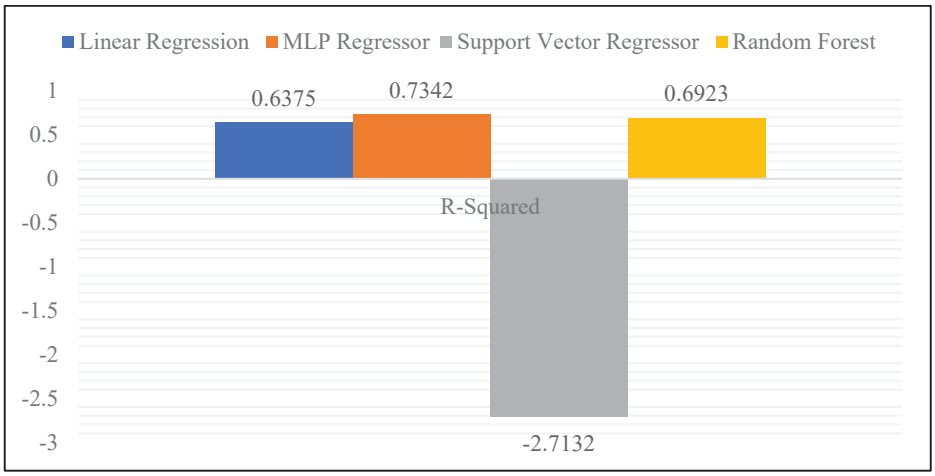

Fig. 4d. R-Squared value of Prediction Models

The comparative performance results of prediction algorithms such as linear regression, support vector regressor, random forest, MLP regressor with respect to mean absolute error, root mean squared error, mean squared error and R squared value is depicted in Table 4. It is confirmed that the MLP regressor based water quality index prediction model produced high accuracy with less error rate whereas support vector regressor based model show less accuracy and high error rate as illustrated in Fig.4e.

Table 4: Comparitive Performance results of water quality index prediction models

Models	MAE	RMSE	MSE	R-Squared
Linear Regression	2.85911	3.8428	11.3318	0.6375
MLP Regressor	1.9143	2.432	7.1032	0.7342
Support Vector Regressor	3.623	4.3281	20.237	-2.7132
Random Forest	2.015	3.79783	9.8278	0.6923

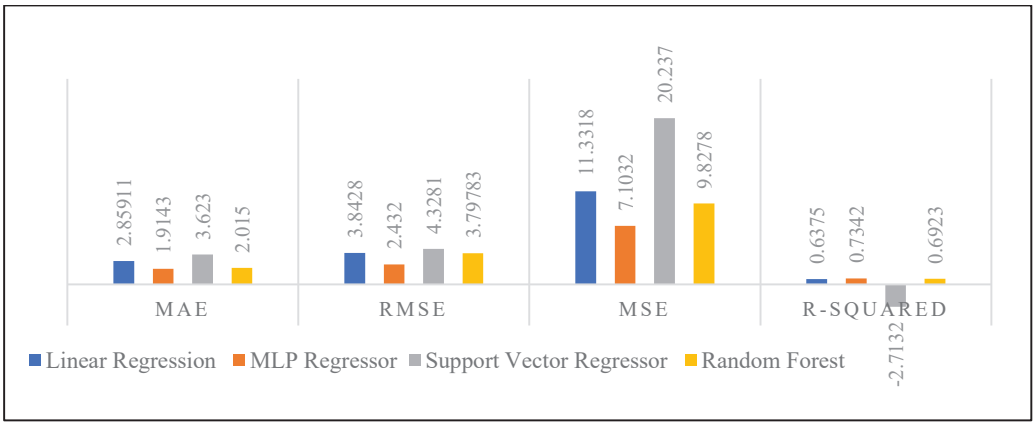

Fig. 4e. Comparitive Performance analysis of water quality index prediction models

5.2 Classification of water quality index

In this section classification models are constructed using the river water quality dataset by applying MLP classifier, support vector machine, naïve bayes, and decision tree using python libraries. The performance of the classification models is evaluated for checking the efficiency in classifying the WQI using metrics such as accuracy, precision, recall and F1 score.

From the experimental results it was found that the accuracy of MLP classifier based classification model is 0.8132, whereas classification models based on naïve bayes, decision tree and support vector machine yield 0.7738, 0.74, and 0.61 correspondingly. Thus, accuracy of MLP classifier is higher as compared to other classifiers whereas support vector machine has low accuracy as shown in Fig.5a.

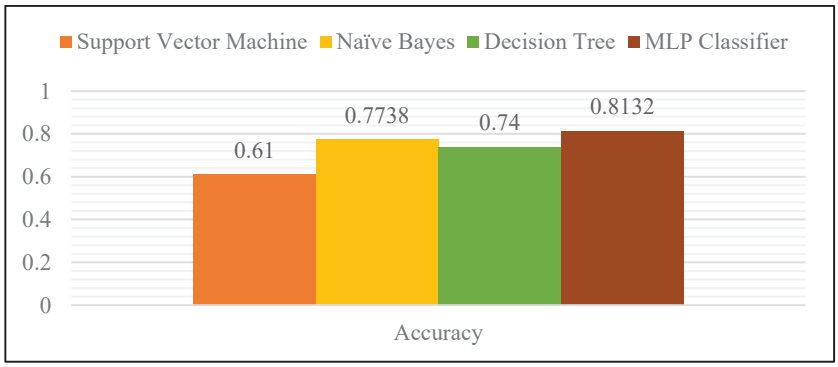

Fig. 5a. Accuracy of Classification Models

Similarly, it is observed that the results of classification model show the precision of MLP classifier based classification model is 0.632, whereas classification models based on naïve bayes, decision tree and support vector machine yield 0.5421, 0.512, and 0.429 correspondingly. Thus, precision of MLP classifier is higher as compared to other classifiers whereas support vector machine has low precision as depicted in Fig.5b.

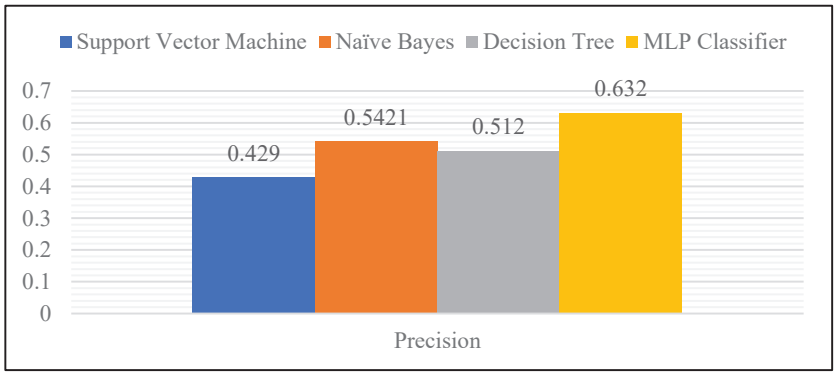

Fig. 5b. Precision analysis of Classification Models

In next case, the results of various classifiers shows that the recall value of MLP classifier based classification model is 0.6123, whereas classification models based on naïve bayes, decision tree and support vector machine gives 0.5523, 0.512, and 0.429 correspondingly. Thus, recall value of MLP classifier is higher as compared to other classifiers whereas support vector machine has low precision as illustrated in Fig.5c.

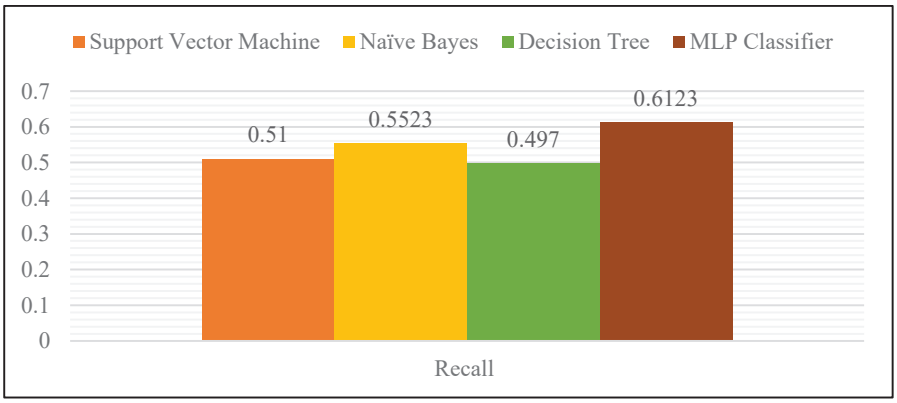

Fig. 5c. Recall evaluation of Classification Models

It was found that the results of classifiers with respect to F1 score confirms that MLP classifier based classification model is 0.5913, whereas classification models based on naïve bayes, decision tree and support vector machine yields 0.5031,0.507, and 0.42 respectively. Thus, F1score value of MLP classifier is higher as compared to other classifiers whereas support vector machine has low precision as illustrated in Fig.5d.

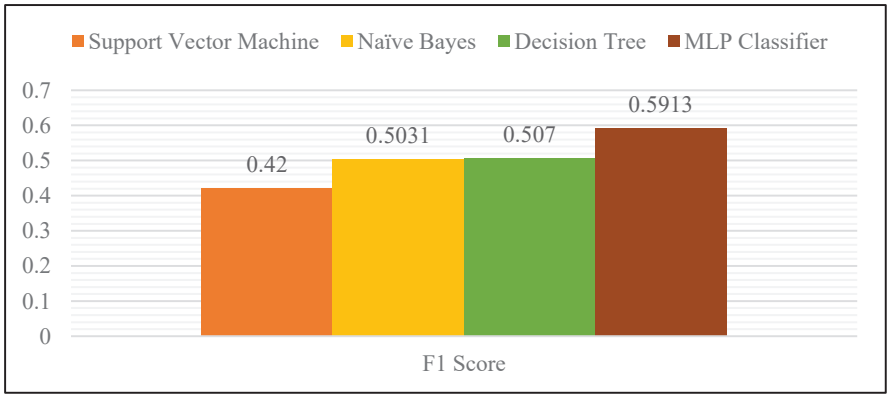

Fig. 5d. F1 score evaluation of Classification Models

The comparative performance results of classification algorithms such as support vector machine, naïve bayes, decision tree, MLP classifier with respect to accuracy, precision, recall and F1score, is illustrated in Table 5. It is proved that the MLP classifier employed in water quality index classification yield high accuracy whereas support vector classifier based model show less accuracy as depicted in Fig.5e.

Table 5. Performance results of water quality index classification models

Algorithms	Accuracy	Precision	Recall	F1 Score
Support Vector Machine	0.61	0.429	0.51	0.42
Naïve Bayes	0.7738	0.5421	0.5523	0.5031
Decision Tree	0.74	0.512	0.497	0.507
MLP Classifier	0.8132	0.632	0.6123	0.5913

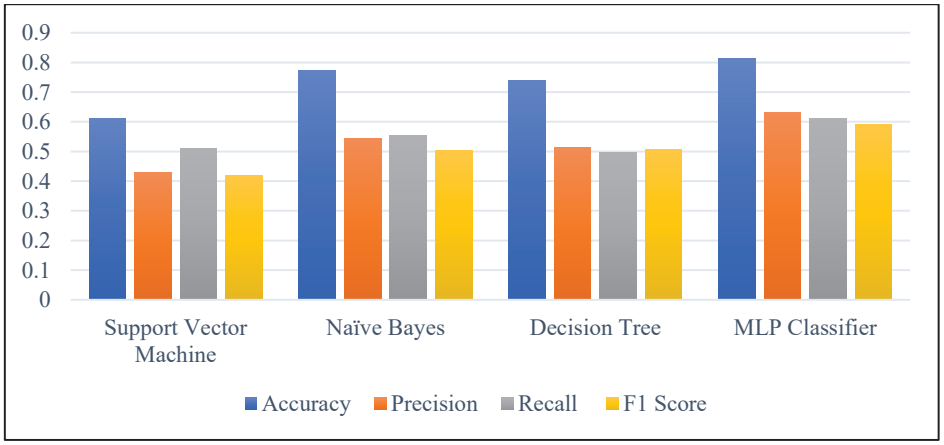

Fig.5e. Performance analysis of water quality index classification models

6. Conclusion

This study investigated the efficiency of machine learning algorithms in predicting river water quality and to classify the water quality index. Machine learning techniques like the random forest, MLP regressor, linear regression, and support vector regressor, were implemented to build WQI prediction models and decision tree, MLP classifier, naive Bayes, and support vector machine algorithms were implemented to build WQI classifiers. The Bhavani River water data with parameters like BOD, DO, TC, nitrate, pH, temperature, and so on, were collected, modelled and employed in building models. The performance of the models in forecasting river water quality index was evaluated using performance metrics. The MLP regressor shows less RMSE value while predicting the water quality index, and the MLP classifier gives high accuracy for classifying the water quality class. It is concluded that, MLP regressor and MLP classifier out performs than the other models in forecasting water quality index. In future, hybrid models with deep learning algorithm can be built to improve the efficiency of the water quality prediction.

References

[1] C.V. Sillberg, P. Kullavanijaya, O. Chavalparit, Water quality classification by integration of attribute-realization and support vector machine for the chao phraya river, Journal of Ecological Engineering 22 (2021), 70–86.
[2] M. Yilma, Z. Kiflie, A. Windsperger, N. Gessese, Application of arti- ficial neural network in water quality index prediction: a case study in little Akaki River, Addis Ababa, Ethiopia, Modeling Earth Systems and Environment 4 (2018), 175–187.
[3] Y.R. Ding, Y.J. Cai, P.D. Sun, B. Chen, The use of combined neural networks and genetic algorithms for prediction of river water quality, Journal of Applied Research and Technology 12 (2014), 493–499.
[4] U. Ahmed, R. Mumtaz, H. Anwar, A.A. Shah, R. Irfan, J. García-Nieto, Efficient water quality prediction using supervised machine learning, Water 11 (2019), 2210.
[5] Zhang, J., Zhu, X., Yue, Y., & Wong, P. W. (2017). A real-time anomaly detection algorithm/or water quality data using dual time-moving windows. 2017 Seventh international conference on innovative computing technology (INTECH) (pp. 36–41). IEEE.
[6] Sakizadeh, M. Artificial intelligence for the prediction of water quality index in groundwater systems. Model. Earth Syst. Environ. 2016, 2, 8.
[7] Fitore Muharemi, Doina Logofătu & Florin Leon (2019) Machine learning approaches for anomaly detection of water quality on a real-world data set, Journal of Information and Telecommunication, 3:3, 294-307
[8] Tejas Subramanya, Davit Harutyunyan, Roberto Riggio, Machine learning-driven service function chain placement and scaling in MEC-enabled 5G networks, Computer Networks, Volume 166, 2020,106980, ISSN 1389-1286
[9] J. P. Nair and M. S. Vijaya, "Predictive Models for River Water Quality using Machine Learning and Big Data Techniques - A Survey," 2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS), 2021, pp. 1747-1753
[10] Kalimur Rahman, Saurav Barua, H.M. Imran, Assessment of water quality and apportionment of pollution sources of an urban lake using multivariate statistical analysis, Cleaner Engineering and Technology, Volume 5, 2021, 100309, ISSN 2666-7908
[11] Arunkumar, R & Thambusamy, Velmurugan. (2021). An Exploratory Data Analysis Process on Groundwater Quality Data. 54. 41-48
[12] Marisol Vega, Rafael Pardo, Enrique Barrado, Luis Debán, Assessment of seasonal and polluting effects on the quality of river water by exploratory data analysis, Water Research, Volume 32, Issue 12, 1998, Pages 3581-3592, ISSN 0043-1354
[13] Stroomberg G. J., Freriks I. L., Smedes F. and Co®no W. P. (1995) In Quality Assurance in Environmental Monitoring, ed. P. Quevauviller. VCH, Weinheim.
[14] G Tan, J Yan, C Gao, and S Yang, Prediction of water quality time series data based on least squares support vector machine, Procedia Engineering, Vol. 31, 2012, pp. 1194-1199.
[15] WC Leong, A Bahadori, J Zhang, and Z Ahmad, Prediction of water quality index (WQI) using support vector machine (SVM) and least square-support vector machine (LS-SVM), International Journal of River Basin Management, Vol. 19, 2021, pp. 149-156.