Enhancing Groundwater Quality Evaluation Using Associative Rule Mining Technique with Random Forest Split Gini Indexing Algorithm for Nitrate Concentration Analysis

[1]Siddthan R and [2]Shanthi PM
[1,2] Department of Computer Science, J. J. College of Arts & Science (Autonomous)
(Affiliated to Bharathidasan University), Pudukkotai, Tamil Nadu, India.
[1]sithan314@gmail.com, [2]shantisuman28@gmail.com

Correspondence should be addressed to Siddthan R : sithan314@gmail.com

Article Info
Journal of Machine and Computing (http://anapub.co.ke/journals/jmc/jmc.html)
Doi: https://doi.org/10.53759/7669/jmc202404067
Received 21 November 2023; Revised from 30 December 2023; Accepted 14 June 2024
Available online 05 July 2024.

Abstract – Human actions and changing weather patterns are contributing to the growing demand for groundwater resources. Nevertheless, evaluating the quality of groundwater is crucial. Nitrate is a significant water contaminant that can lead to blue-baby syndrome or methemoglobinemia. Therefore, it is necessary to assess the level of nitrate in groundwater. Current methods involve evaluating the quality of groundwater and integrating it into the models. The inappropriate datasets, lack of performance, and other constraints are limitations of current methods. Ground water dataset is used and pre-processed the data's. Selected data's are feature extracted and associated with the rule ranking. In the suggested model, the use of associative rule mining technique has been implemented to address these challenges and assess nitrate levels in groundwater. The method of rule ranking is carried out using association rule mining technique to divide the datasets. The split gini indexing algorithm is introduced in the proposed model for data classification. The Split Gini Indexing algorithm is a decision tree induction algorithm that is used to build decision trees for classification tasks. It is based on the Gini impurity measure, which measures the heterogeneity of a dataset. The quality of groundwater has been classified using Naïve Bayes, SVM, and KNN algorithms. The proposed approach's efficiency is evaluated by calculating performance metrics such as precision, accuracy, F1-score, and recall values. The suggested method in the current research attains an improved accuracy of 0.99, demonstrating enhanced performance.

Keywords – Nitrate Concentration, Groundwater, Contamination, Rule Ranking, Classification.

I. INTRODUCTION

Groundwater is considered as the most critical freshwater resource in various regions of the world. However, the population growth and various economic actions have negatively impacted the quality of ground water. Most people around the globe depend on groundwater as an essential freshwater source. Actually, groundwater is considered as an indispensable freshwater source, and it also gives around 20% for irrigation purpose [1]. Numerous fertilizers are used in the agricultural lands in rural areas in order to improve their productivity. Fertilizers contain nitrate compounds, which lead to the contamination of groundwater resources and thus affect the consumer's health [2]. Due to human inhabitation, nitrate pollution in groundwater is deliberated as one of the significant problems among different environmental problems. Typically, groundwater contamination occurs in plains and basins, where urbanization and agriculture are well-developed [3].

Nitrate is considered as the most general chemical contaminant of groundwater after pesticides, which has attracted significant importance. Because, nitrate may enter the body of the human through exogenous and endogenous pathways, which becomes more toxic while nitrate compound is reduced into nitrite compound in the oral cavity [4]. Excessive amount of nitrate in drinking groundwater is a risk to human health resulting in many health issues like decreased functioning of the thyroid gland, increased risk of methemoglobinemia, cancer (lymphatic, gastric and esophageal cancers), and even abortions and reproductive damage for women [5]. So, in recent years, the public concern over the groundwater quality deterioration due to nitrate contamination has significantly increases in many nations. As a result, the evaluation of nitrate contamination in groundwater gets increased. This evaluation helps water management authorities to perform the suitable nitrate risk elimination procedures. For that reason, it is important to know the groundwater vulnerability effects in public water supply regions. So, several techniques are used to measure the vulnerability level and groundwater contamination around the world. The techniques include interpolation and index approaches, statistical and process-based models [6].

GIS (Geographic Information System) and WQI (Water Quality Index) approaches are the widely used techniques for visualizing and assessing groundwater quality differences [7]. GIS revolutionized the way to understand and interact with the world. GIS is a cutting-edge technology that integrates spatial data with other relevant information to create interactive maps, visualizations, and analyses that empower users to make informed decisions. It has a wide range of applications across various industries and sectors which includes land use planning, natural resource management, public health, transportation planning and retail and marketing. WQI is one of the evaluation tools, which are used for measuring the quality of groundwater. Whereas, WQI is an important tool for assessing groundwater quality, including biological, physical and chemical properties and also assess how it is managed in a specific region. This tool also helps to select an economically possible treatments, purification or desalination methods in order to resolve the problems in water quality easily. WQI also helps the decision-makers in building sound legislation and incorporating the water quality programs of the government [8]. WQI concept is a valuable tool for assessing and communicating the overall quality of water. By combining multiple water quality parameters into a single numerical score, WQI provides a comprehensive evaluation of water quality and facilitates informed decision-making for water resource management. Detection of groundwater nitrate contamination is very essential for water resource management as well as pollution control [9].

In previous days, there are few manual techniques used for analysing the concentration of nitrate in ground water. One of the techniques are Cadmium Reduction method, in which the nitrate ions are reduced to nitrite ions using a cadmium reagent. The resulting nitrite ions can be assessed calorimetrically using a spectrophotometer or a colorimeter. This method is widely used due to its simplicity and accuracy in determining the nitrate levels in ground water. Manual techniques may be prone the human error, causing inaccuracies in the measurement, and these methods are generally time consuming in nature. To advance nitrate evaluation method, ML (Machine Learning) methods like RF (Random Forest) have verified its efficiency for evaluating the quality of groundwater according to spatial environmental predictors [10]. Though, this technique gives better performance than the manual techniques, it has some limitations. It needs large amount of data for training, which is more complicated to find the nitrate concentration level in ground water.

Excessive nitrate concentration in ground water is a serious issue. Hence it needs to be addressed as soon as possible. However, three extensively used ML models, namely, XGB (eXtreme Gradient Boosting), ANN (Artificial Neural Networks) and SVM (Support Vector Machines), which are utilized to determine the nitrate concentration in ground water [11]. Though, existing studies confer the better results, there is a lack of accuracy in its performance that needs to be addressed. Hence, there is a need to overcome those challenges in order to improve the efficiency. Further, Random forest split gini indexing algorithm is used with associative rule mining technique in the proposed model for analysing nitrate concentration in ground water.

XGBoost

It is a sophisticated version of the gradient boosting algorithm. A greatly effective and expandable form of gradient boosting, it has become well-liked in machine learning competitions because of its fast speed and performance. XGBoost functions by progressively incorporating predictors into a group, with each one adjusting the errors made by the previous one. In contrast to gradient boosting's approach of correcting residual errors in predictions, XGBoost employs a more constrained model structure to tackle over fitting, ultimately resulting in improved performance.

ANN

ANNs form the basis of deep learning algorithms. The artificial neural networks are influenced by the biological neural networks found in animal brains. An ANN consists of numerous interconnected neurons that generate a series of real-valued activations. Input neurons are activated by sensors detecting the surroundings, while other neurons are activated by connections with weight from previously active neurons. The output of every neuron is determined by applying a nonlinear function to the total sum of its inputs. The network is created by linking the input layer with at least one hidden layer that ultimately connects to an output layer.

SVM

A robust supervised machine learning algorithm that is utilized for classification and regression tasks. Nevertheless, it is primarily utilized in classification tasks. In the SVM algorithm, we represent every data point as a point in an n-dimensional space (where n is the number of features) with each feature value being a specific coordinate value. Next, classification happens by identifying the hyper-plane that most effectively separates the two classes.

Objectives of the Study

The major objectives of the study are discussed below.

- To determine the nitrate concentration level in groundwater by using the given sample obtained from multilevel spatial database.
- To classify the predictor variables such as hydrogeology and hydrology using the associative rule mining algorithm with Random Forest classification approach.
- To compare the obtained results with the existing studies in order to evaluate the proposed study.

Paper Organisation

The forthcoming section 2 discussed about the existing techniques that have been done for the evaluation of nitrate contamination in groundwater with the problem identification. The proposed work includes the novel random forest split gini indexing algorithm using associative rule mining algorithm, which is depicted in section 3. The results of the discussed system are deliberated in section 4. The entire proposed system is concluded in section 5.

II. LITERATURE REVIEW

The below section enumerated review analysis of conventional researchers, discussing about the different approaches associated with the detection of groundwater nitrate concentration.

Detection of the nitrate concentration in groundwater is an essential part for water resource management and pollution control. In the Marvdasht watershed, Iran, the study [9] has been aimed to model the nitrate concentration in spatial groundwater. The detection based on various AI (Artificial Intelligence) techniques of SVM, RF, Baysia-ANN (Bayesian artificial neural network) and Cubist ML models. However, the outcomes of the groundwater nitrate concentration region have been shown the northern regions of the case study have the greatest quantity of nitrate that is very much greater than the other regions [12]. Similarly, data in this study have been divided into two groups of testing (30%) and training (70%) for modelling purpose. However, the outcomes of the modelling has been shown the vulnerability of the nitrate concentration in groundwater as RF model better than the other models such as Cubist, Bayesian-ANN and SVM.

However, the evaluation of groundwater vulnerability index and the accurate mapping are played a significant part in the prevention of groundwater resources from the pollution. The study [13] has been used Cl/Br and Cl/NO3 ratio to find the source of nitrate in groundwater. Novel intelligent predictive ML regression models of KNN (k-Neighborhood), BA (Bagging Regression) and ERT (ensemble Extremely Randomized Trees) at two stages of modelling have been used in the study [14] to enhance the model of DRASTIC-LU. Hence, two stage ML modelling has been an excellent method for the proactive management of groundwater resources over pollution. However, the precise assessment of groundwater pollution vulnerability, which is very important for the protection and management of the pollution of groundwater in the watershed. Further, the study [15] has been utilized the advanced ML models of RBNN (Radial Basis Neural Networks), RFR (ensemble Random Forest Regression) and SVR (Support Vector Regression) to find the precise performance for the assessment of groundwater pollution vulnerability. The results have been predicted that the ensemble RFR is a vigorous tool to improve the groundwater pollution vulnerability pollution that has been contributed to the protection of environmental over groundwater pollution.

To advance the process, Bayesian approaches like BGLM (Bayesian generalized linear model), BART (Bayesian Additive Regression Tree), BRR (Bayesian Ridge Regression) and BRNN (Bayesian Regularized Neural Network) have been used to model groundwater nitrate pollution in a semiarid area. The results have been revealed that the all Bayesian models utilized in the considered study has been competent to groundwater nitrate and the BART models with $R2 = 0.83$ that has been more efficient than the other models. Further, the study [16] has been constructed a HELM (Hybrid Machine Learning Model) for modelling nitrate concentration in resources of water. In addition, the SFFS (Sequential Forward Floating Selection) approach has also been incorporated to choose the ideal input parameters to simulate the nitrate content in every cluster. In existing studies, there are various ML techniques have been used to evaluate the groundwater nitrate concentration [17].

In addition, three ML models like SVM, ANN and XGB have been utilized [11] to determine the nitrate pollution from pesticides in groundwater. The models have been assessed for their efficiency in determining the pollution levels by utilizing sparse data with non-linear relationships. In US, the analytical capability of models has also been evaluated utilizing a dataset comprising of 303 wells through 12 Midwestern states. Lastly, the study has been evaluated the significance of properties using the values of game-theoretic Shapley to rank properties constantly. The execution of nine various ML algorithms has been assessed in [18] in order to determine phosphorus and nitrate concentration for five various watersheds with various land-use practices. The results and methodology has been evaluated in the study that guide policymakers to determine the concentration of phosphorus and nitrate accurately, which will be involved in making a proper plan to deal with water pollution problem.

Detection of groundwater contamination due to different chemical components is important for policymaking, planning and groundwater resources management. The application of ML methods for GWQ (Ground Water Quality) modelling has widely increased in the last twenty years. The study [19] has been evaluated all semi-supervised, implemented ensemble ML, unsupervised and supervised models to determine any parameters of groundwater quality. The threat evaluation of groundwater nitrate pollution has been achieved in [20]. As a result, nitrate concentration data in 130 wells has been utilized to make pollution mapping. In addition, four ML models have been utilized to evaluate the groundwater pollution probability, including GLM, SVM and BRT. AUC characteristic curve has been utilized to assess the validity of each model. By considering the risk map of groundwater pollution because of probability mapping, contamination and vulnerability, the south-western and western regions of aquifer are at high. The expose of major risks to nitrate pollution, which is reliable with the land use map of urban and agricultural regions.

A novel framework approach has been constructed in [21] in order to detect and map nitrate concentration vulnerability in the Bangladesh's coastal multi-aquifers. This has been done by combining the K-fold cross evaluation technique as well as novel EL (Ensemble Learning) procedures, comprising Bagging, RF and Boosting. Hence, the dependability of EL modelling has been verified the applicability and effectiveness of the suggested novel technique for decision-makers in groundwater contamination at the regional and local levels. Further, the study [22] has been aimed to find the nitrate vulnerability regions of Easter India's coastal districts by three data mining approaches such as bagging, RF and boosting approaches. From this, the results have been shown that boosting model is more efficient than the bagging ad RF models. The study [23] has been dealt with the performance assessment of application of ML algorithms like DNN, GBM and XGBoost in order to assess groundwater. Two water quality indices have been used to examine the applicability of these models like EWQI and WQI. For WQI, the values of RMSE (Root Mean Square Error), NSE (Nash–Sutcliffe Efficiency), CC (Correlation Coefficient), Index of agreement for EWQI and CC, NSE and RMSE for WQI have been achieved.

The performance of AI techniques has been investigated in [24], which has been included PSO (Particle Swarm Optimization), SVM and a NBC (Naive Bayes classifier) in order to determine the water quality index. Further, the outcomes of the study have been suggested that the ensemble ML algorithms has been utilized to estimate and determine the WQI with accuracy. A ML-based framework has been developed in [25], to map the quality of groundwater in an released aquifer in North Iran. In this study, groundwater samples have been conferred from 248 monitoring wells through the area of North Iran. Additionally, six ML classifiers, which includes SVM, XGB, RF, ANN, GCM (Gaussian Classifier Model) and KNN, which have been utilized to make relationships among GWQI and its controlling factors. Moreover, the study [26] has been proposed three ML methods including XGB, DNN and MLR (Multiple Linear Regression) have been utilized to determine the nitrate pollution in groundwater in North Iran. The results have been shown that the evaporation rates, groundwater depth and population density are the significant factors influencing groundwater nitrate pollution.

Problem Identification

From the evaluation of above existing works, core concerns are emphasized as explored below.

- By discovering intellectual hyperparameter tuning approaches, particularly for ANNs, which also improve their predictive efficiency. Similarly, using a bigger dataset may enhance the execution for all models by conferring more knowledge to evaluate other methods [11].
- Limited number of independent variables have been used in the study [18], so these approaches may be applied to determine concentration of nitrate with limited data that maximizes the use of developed models.
- Many parameters like anthropological impacts or long-term climate conditions, which are difficult to obtain or estimate for a dataset. The exclusion of powerful indicators in the dataset may nullify the model [19].
- Using a larger field dataset in [24], is often a difficult to provide the painstaking approach of sample accumulation and laboratory analysis for all the water quality parameters.

III. PROPOSED METHODOLOGY

The proposed method uses the spatial dataset for determining the nitrate concentration in groundwater. Though, existing studies have been provided considerable results, still it has been lacked with better accuracy in determining the nitrate contamination in groundwater. However, the projected model has attained better accuracy than the conventional models by implementing various methods in the model. The overall flow to predict the nitrate concentration from groundwater using the associative rule mining with Random forest classification as exposed in **Fig 1.**

Ground water dataset are selected as the input source. The collected data is then pre-processed to ensure its quality and suitability for analysis. Pre-processing steps include:

- Missing Value Handling: Neglecting records with missing values to avoid bias.
- Normalization: Scaling values to a common range to facilitate comparison.

Pre-processed data's are predicted by feature selection. This process identifies the most relevant features (input variables) for the classification task. Features are selected based on their ability to discriminate between different classes. Association rule is carried out to select the data's by rule ranking. During feature selection, dimensionality is eliminated because of computational efficiency, noise reduction, improved visualization, feature selection and data compression. Association rule learning is applied to discover relationships between features and class labels. Rules are generated based on the frequency of co-occurrences of features and classes. Based on the rule ranking of the data's, it is classified. Rules generated by association rule learning are ranked based on their confidence and support. Confidence measures the strength of the association between features and classes, while support indicates the frequency of the association. It is classified by split gini indexing. Based on the rule ranking, data is classified into different classes. Split Gini Indexing is used to determine the optimal split point for each rule. This metric measures the impurity within a dataset and helps identify the best split for classification. Based on the rule ranked data's, it is predicted by performance metrics. The performance of the classification model is evaluated using performance metrics such as accuracy, precision, recall, and F1-score. These metrics quantify the model's ability to correctly classify data points.

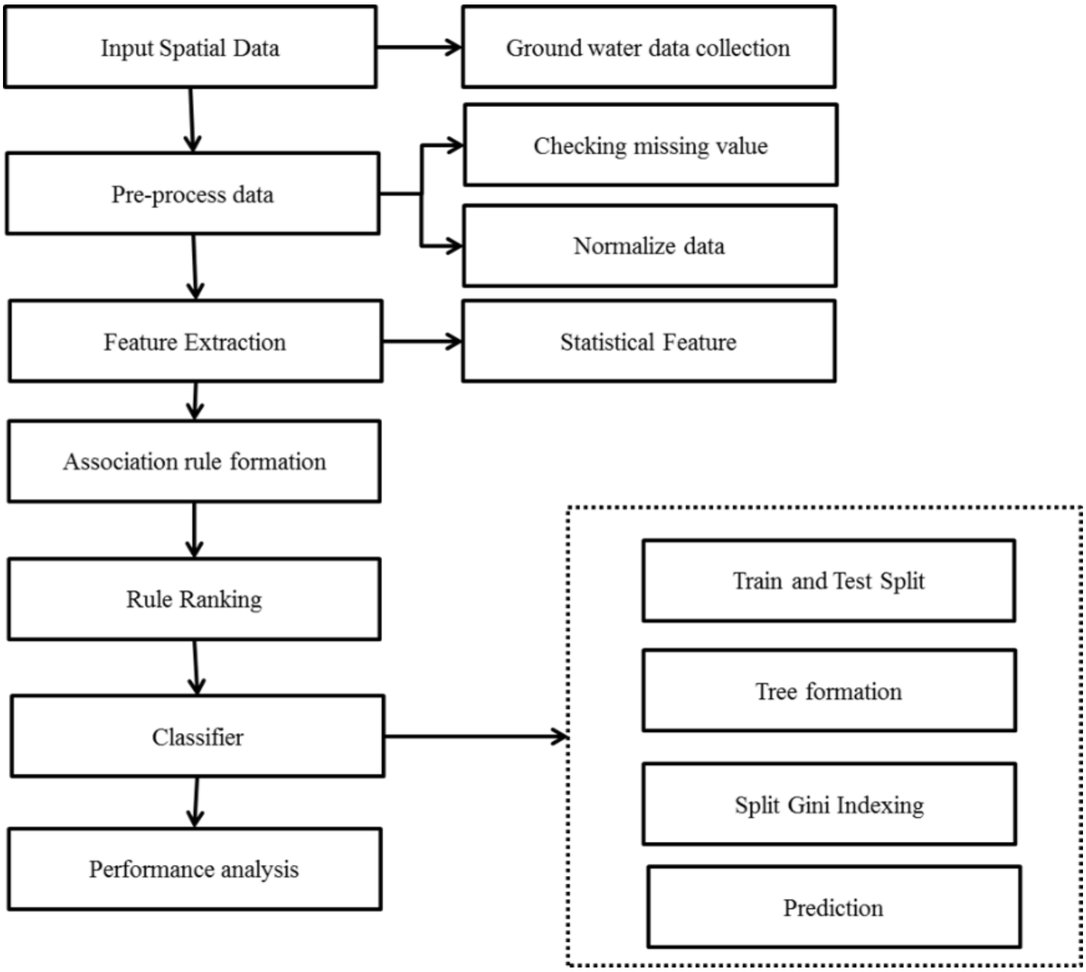


Fig 1. Overall Proposed Framework.

Dataset Description

The quality of water is crucial for maintaining a healthy ecosystem. Pure water sustains a variety of plants and animals. Although they may appear unrelated initially, our activities on land impact the quality of our water. Pollutants, too many nutrients from fertilizers, and sediment are often transported into nearby lakes and rivers through runoff from urban areas or agricultural fields.

Researchers assess different characteristics to evaluate the quality of water. Temperature, pH, specific conductance, turbidity, dissolved oxygen, hardness, and suspended sediment are among the factors considered. Each one shows a distinct aspect of the condition of a water system. The corresponding section deals with the data used in the corresponding study with their appropriate data link for reference.

Dataset Link

https://www.kaggle.com/datasets/balabaskar/water-quality-data-india

In a healthy ecosystem, water quality is one of the significant factors. Contaminants and excessive nutrients from pesticides pollute groundwater with its excessive nitrate content. Following are the properties of water, which determines water quality.

Temperature

For fish and aquatic plants, water temperature is very important. Because, it affects the oxygen level and the ability of organisms that resist particular contaminants.

Acidity – pH

It helps to measure the quantity of hydrogen in a water molecule that permits to determine whether the water molecule is acidic, basic or neutral.

Dissolved Oxygen

About ten molecules of water is dissolved in water per million molecules of water. Microscopic organisms and fish require dissolved oxygen to live.

Turbidity

It creates water opaque or cloudy. It is defined as the quantity of particulates like silt, clay, microscopic organisms or plankton, which are suspended in water.

Specific Conductance

It is defined as the measure of the capacity of water to conduct an electric current. Generally, it is based on the amount of dissolved solids like salt in the water.

BOD (Biochemical Oxygen Demand)

It is an evaluation of the amount of oxygen needed to eradicate waste organic matter from water in decomposition procedure by aerobic bacteria. Hence, the higher BOD denotes the lower water quality, whereas the lower BOD indicates the higher water quality.

Nitrate and Nitrite

These are soluble compounds comprising of oxygen and nitrogen. Normally, nitrite changes to nitrate in the environment, which means nitrite occurs seldom in groundwater. Nitrate is presented in all grains and vegetables, whereas nitrite is used for manufacturing explosives, curing meat and industrial boilers maintenance.

Pre-processing

The model gets complicated in processing all the features of data retrieved from the input raw data of every dataset. Generally, the collected data comprises of either numeric or non-numeric data. So, data pre-processing phase is involved in the suggested system in order to accelerate the training of the input data to produce optimal solutions and performance. In this study, pre-processing phase involves two steps like checking the missing values and normalization.

Checking Missing Values

It is the process of checking missing values in data pre-processing phase that includes identifying and handling any instances, where data is missing or incomplete. This is a significant step to assure the reliability and quality of dataset before executing any modelling or analysis.

Normalisation

This process eliminates the efficiency of original feature values. Therefore, the properties of numerical data are scaled to a certain range in data normalisation process. This is done to assure that the characteristics are on the same scale thus helps in accelerating the speed of training technique.

Feature Selection

The statistical feature is used for feature selection process in proposed study. The statistical feature extraction is a technique used to extract relevant information from data by implementing statistical characteristics of the dataset to determine and choose properties that are most informative than other techniques. Common statistical feature extraction process such as measures of central tendency (eg., meridian or mean), hypothesis testing, measures of dispersion (eg, standard deviation, variation) and correlation analysis are used for the feature selection process. These methods help in determining the properties that are statistically important ad have a strong relationship with the target variable.

After feature selection process, associative rule formation and rule ranking are performed.

Associative Rule Formation

This finds the interesting connections and relationships among large set of data items. This rule formation is applicable in analyzing datasets. This helps to reveal the possibility of relationships among data items, within bigger datasets in different database kinds.

- Associative rule mining is developed to produce the set of rules or to build classifiers which are inspired to different researchers to implement regarding ML and data mining concepts.
- The class attribute is calculated in associative rule mining. In a rule of A□ B considered as if-then rule, the B is the class attribute. However, more accurate classification systems are generated by the associative rule mining.
- Moreover, the proposed model follows the three basic stages.

STEP 1 – Initialize the selected data's from the dataset
STEP 2 - Association rule is carried out for selecting the data's based on rule ranking
STEP 3 – Rule ranked data's are classified and performance metrics are calculated.

Rule Ranking

It is the process in data mining and ML, where association rules are generated from a dataset that are assessed and ranked based on certain criteria. This helps in determining the most relevant and significant rules that can confer valuable insights and determinations.

Data Splitting

The most important features that are chosen though the proposed feature selection approach are then passed into train and test phases. The input data are classified into two splits in this phase namely train split and test split, which is in the ratio 80:20. Among this ratio, 80% of data are used to train the classifier, whereas 20% of data are used to test the actual and predicted values. This assures that both the datasets are representative of the entire dataset and confers the effective measure of accuracy detection of the trained classifier and the unseen testing set.

Classification

Classifier is an procedure that automatically classifies data into one or more set of classes. In the proposed model, the rule ranking data is classified into test and train splits. Similarly, the data is classified into the formation of trees. SVM, KNN and NB are the classification algorithms that are used in the proposed study. The novel random forest with split gini indexing algorithm is proposed in the study. Finally, performance metrics of the proposed study is evaluated.

SVM

SVM is a ML algorithm that utilizes supervised learning models to resolve complicated regression, classification and external prediction problems by executing optimal data transformations that predict boundaries among data points regarding predefined labels, outputs and classes. The formula for the SVM output is

$$z = \underline{w}.\underline{d} - q \tag{1}$$

where the regular vector to the hyperplane is w and d is input vector. Whereas, the splitting hyperplane is the plane z=0. And, the closest points lie on the planes z = ±1. Then, the margin m is expressed as

$$m = \frac{1}{||w||_2} \tag{2}$$

Whereas, maximizing margin can be stated through the corresponding optimization limitation.

$$min \frac{1}{2} \ ||w|^2 \ y_i \ (\underline{w}.\underline{di} - q) \geq 1, \forall i \tag{3}$$

Here, d_i is the i^{th} training example, and y_i is the exact SVM's output for the i^{th} training example. And, the value y_i is –1 for negative examples and +1 for positive examples in a class.

The optimization limitation is changed into a dual form using a Lagrangian, where the objective operation Ψ is completely based on a group of Lagrange multipliers α_i,

$$\psi(\alpha \rightarrow \ ) = \frac{1}{2} \ \sum_{i=1}^{N} \square \sum_{j=1}^{N} \square \, y_i y_j \ (\overrightarrow{d_i}, \overrightarrow{d_j}) \alpha_i \alpha_j - \sum_{i=1}^{N} \square \ \alpha_i \tag{4}$$

In equation (4), N is defined as the total amount of training examples subject to the inequality limitations,

$$\alpha_i \geq 0, \forall i, \tag{5}$$

and for one linear equality constraint,

$$\sum_{i=1}^{N} \square \ \alpha_i \ y_i \ = 0 \tag{6}$$

Among every Lagrange multiplier and training example, there is one-to-one relationship. The threshold b and the normal vector w^{-} are determined from Lagrange multipliers, once the Lagrange multipliers are derived.

$$\underline{\ \ w} = \sum_{i=1}^{N} \square \, y_i \ \alpha_i \, \overrightarrow{d_i} \ , q = \ \underline{w} \, \overrightarrow{d_k} - y_k \ \ for \ some \ \alpha_k > 0 \tag{7}$$

The quantity of execution needed to assess a linear SVM is constant in the total amount of non-zero support vectors, as $\bar{w}$ can be evaluated through equation (7) from the training data before use.

All the datasets are not linearly independent. Additionally, there is no hyperplane that partitions positive examples from negative examples. The equation (7) suggests the alteration to the original optimization statement (3) and the modification is

$$min\,\frac{1}{2}\;||w|^2\;+\;C\,\sum_{i=1}^{N}\square\,\zeta_i\;subject\;to\;y_i\;\left(\underline{w}.\underline{d}-q\right)\geq 1\;-\zeta_i\;,\forall i \tag{8}$$

Where, ζ_i are slack variable that allows the margin failure and C is a parameter which trades off margin with a small amount of margin failures. The new optimization problem is changed into the dual form, it simply alters the limitation (5) into a box limitation.

Moreover, SVMs can be generalized to non-linear classifiers. Whereas, the output of a non-linear SVM is executed from the Lagrange multipliers explicitly.

$$z=\;\sum_{i=1}^{N}\square\,y_j\;a_j\;K\left(\overrightarrow{d_j}\,\underline{d}\right)-q \tag{9}$$

Here, K is a kernel function that evaluates the connection or distance among the input vector $\vec{x}$ and the stored training vector $\vec{x}_j$. The examples of K include polynomials, neural network non-linearities and Gaussians. If K is linear, then the equation (1) for SVM is improved.

Through quadratic program, the Lagrange multipliers α_i are still executed. Whereas, the non- linearities modify the quadratic form, and the dual objective function Ψ is still quadratic in α as follows,

$$\begin{gathered}\psi(\,\alpha\rightarrow\;)=\frac{1}{2}\;\sum_{i=1}^{N}\square\sum_{j=1}^{N}\square\,y_i y_j\;K\left(\overrightarrow{d_i},\overrightarrow{d_j}\right)\alpha_i\alpha_j-\sum_{i=1}^{N}\square\,\alpha_i\\ 0\leq\;\alpha_i\;\leq C,\forall i\\ \sum_{j=1}^{N}\square\,y_i\alpha_i\;=0.\end{gathered} \tag{10}$$

KNN

It is a non-parametric, supervised leaning classifier, which utilizes proximity to make determinations or classifications about the grouping of an individual data point. KNN is one of the familiar and easiest classification that is used in ML algorithms. The algorithm for KNN classification is shown below.

Algorithm I: KNN

```
Step 1: Transform the data into statistical data
Step 2 : Normalise all data utilizing:
        For every feature
        For every feature − value b
        Normalized value = (b − min)/range
        End for
        End for
Step 3: Evaluate the distance of test instances (with feature
                 − value $b_v, b_v$) from every training instance using:
For each training instance with feature − values  $b_x, b_x$
   Euclidean distance = $\sqrt{(b_v * b_x)^2} + (d_v * d_{x)}{}^2)$
Step 4: Assign the majority class label between K nearest instances to the test instance.
```

NB

It is a general probabilistic classifier, which has strong feature independence assumption. This algorithm is dependent on Bayes' theory from which PROB (M) and the PROB (M/L), which are evaluated for an instance observed before and after the evidence correspondingly. The algorithm for Naïve Baiyes classification is shown.

Algorithm II: NB

```
Step 1: Training:
    For every feature
    For every feature − value M
    For every class − label L
    $PROB\left(\frac{M}{L}\right) = (total\ quantity\ of\ occurrences\ of\ feature$
        $-\ value\ with\ class\ label)/(total\ quantity\ of\ occurences\ of\ class\ label)$
    End for
    End for
    End for
Step 2: Testing:
    For every instance in test data
    Measure posterior probability using
        $PROB\left(\frac{M}{L}\right) = \left(PROB\left(\frac{M}{L}\right) * PROB(M)\right)/PROB(L)$
    End for
Step 3: Assign the class label with higher posterior probability to the test instance.
```

Tree Formation

The process of forming a decision tree involves recursively splitting the data based on the values of various attributes. Basically, the algorithm chooses the best attribute to partition the data at every node, regarding the certain criteria. RF is one of the type of EL hat has the all constructor classifiers are of identical type (i.e, decision tree). Because, RF is a homogeneous EL technique. In the proposed study, random forest is used, which are described below.

RF

RF is one of the types of EL, which is a generally used ML algorithm. RF algorithm is an extension of bagging process, as it uses both the feature and bagging randomness to make an uncorrelated forest of DTs (Decision Trees). The procedure used for RF classification, which are shown below.

Algorithm III: Random Forest

```
Let $D = \{(s_1, m_1), (s_2, m_2), \ldots\ldots (s_N, m_N)\}$ denote the training data with $s_i = (s_{i1}, s_{i2} \ldots . s_{ip})^T$
For $j = 1$ to $J$:
Take a bootstrap sample D of size N from D.
Utilizing the bootstrap sample, Dj as the training data fit a tree.
(a) Initiate with all observations in a single node
(b) Loop the following steps recursively for each node until the stopping criterion is met:
        (i) Choose m predictors at random from the p available predictors
Find the best binary split between all binary splits in the predictors from step
        (i) Partition the node into two descendant nodes by the split from step
        (ii) To create a prediction at a new point s
$\tilde{f}(s) = argmax\ s_m\ \sum_{j=1}^{b} \square\ b(\hat{h}_j(s))$
Where $\hat{h}_j(s)$ is the detection of the response variable at x by the jth tree .
```

Split Gini Indexing

The Split Gini Index measures the impurity of a node in a decision tree. It represents the probability that two randomly selected samples from the node belong to different classes. A lower Split Gini Index indicates higher purity and better class separation. In decision tree induction, the goal is to find the split point that maximizes the reduction in impurity. Split Gini indexing evaluates each possible split point and selects the one that results in the greatest decrease in the Gini Index. Split Gini indexing is a powerful technique that significantly enhances the performance of decision tree algorithms. By identifying the most discriminant split points, it leads to more accurate and efficient models. Its robustness, interpretability, and wide applicability make it an invaluable tool for machine learning practitioners. The proposed model introduces random forest with split gini indexing algorithm for determining nitrate concentration in groundwater. Generally, RF is an effective combined ML method integrated by decision tress. However, the RF identification technique is appropriate for high-dimensional data, and it runs fast. The proposed method uses random forest ranking algorithm that has been described below.

Random Forest Ranking

RF is an ensemble classifier that utilizes many decision tree models. However, the pseudo code of RF ranking is shown below.

Algorithm IV: Random Forest Ranking

```
Input: training data sets $n_{N*p}$ and Number of trees $(V)$
        For every variable $i \in$
 P do
        For $y = 1$ to $V$:
          1. Draw a sample $B*$ of size $O$ from the training data.
2. Grow a RF tree $F_y$ to the $\frac{2}{3}$ of data.
3. Determine classification of the leftover $\frac{1}{5}$ by the tree, and measure the classification rate $=$
accuracy rate (OOB), namely $accuracy_y$.
          4. For variable $i$, transform the value of variable and execute the accuracy $(accuracy_b)$, subtract to the origina
$=$ $accuracy_y - e_y$), the increase is an indication of the variable's importance.
          End for
          Collect total accuracy from all trees and calculate variance.
                 $\hat{\imath} = \frac{1}{A}\sum_{k=1}^{V} □\, i_k$ and $s_i^2 = \frac{1}{V-1}\sum_{k=1}^{V} □\, (i_k - \hat{\imath})^2$
          calculate importance of variable $i$: $e_i = \hat{\imath}/s_i$
          End for
```

Split Gini Indexing Algorithm

The novel split gini indexing algorithm is used in the proposed study in order to partition the data into subsets that are basically as pure as could be expected. To decide the best split in each node, the impurity index (like split gini index) is used. Split gini indexing is a proportion of impurity or inequality in statistical and monetary settings. In ML, it is used as an impurity measure in decision algorithms for classification purposes. **Fig 2** reveals the overall architecture of the proposed study.

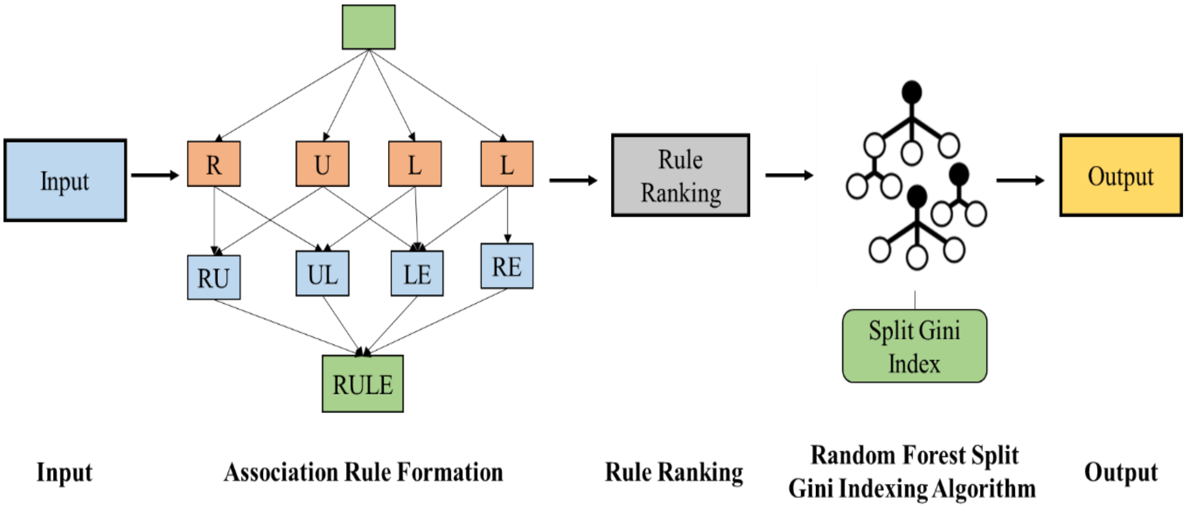


Fig 2. Overall Architecture.

From **Fig 2** it infers that when the input data of ground water samples is given, the association rule formation is performed to the given input data. Set of rules, which illustrate the target variable is extracted using association rules formation technique. Similarly, the rule ranking process is performed, where the association rules are produced from a dataset. These datasets are then evaluated and ranked regarding certain rules. After that, the ranking data is entered into the random forest spilt gini indexing algorithm phase. In this phase, the ranked data are partitioned into subsets to decide the best split in each node. Finally, the nitrate concentration of groundwater is predicted.

IV. RESULTS AND DISCUSSION

The results generated by execution of proposed split gini indexing algorithm is discussed in this section. Besides, performance metrics, EDA and performance analysi are also deliberated. To exhibit the effectiveness of the propsoed model, the comparision of existing approches with the proposed model is also elaborated.

Performance Metrics

The efficacy of the suggested model is assessed by the performance metrics such as recall, F1score, precision and accuracy values. The performance metrics can be explained in the following equations. In equations, where TN $(True_{negative})$ represents the quantity of negative samples, which are determined accurately. TP $(True_{positive})$ denotes the total number of accurately identified samples, FP $(False_{positive})$ represents the total number of positive samples, which are inaccurately identified and FN $(False_{negative})$ signifies the total amount of negative samples, which are incorrectly predicted.

Recall

It is defined as the ratio of the accurately identified outcomes to overall findings, which is given by equation (11),

$$Recall = \frac{True_{positive}}{False_{negative} + True_{positive}} \tag{11}$$

F1-score

This performance metric is derived by the mean evaluation of recall and precision values. It also denotes, if F1 score is greater, the classifier quality is also enhanced, which is given by the equation (12),

$$F1 - Score = 2 * \frac{Recall * Precision}{Recall + Precision} \tag{12}$$

Precision

It is given by the ratio of values of true positive to the combination of false positives and true positives, which is given by equation (13),

$$Precision = \frac{True_{positive}}{True_{positive} + False_{positive}} \tag{13}$$

Accuracy

It is defined as the ratio of the accurate identification of samples to the overall identification of the classifier. The accuracy formula is explained in equation (14),

$$Accuracy = \frac{True_{positive} + True_{negative}}{True_{positive} + False_{positive} + True_{negative} + False_{negative}} \tag{14}$$

EDA (Exploratory Data Analysis)

EDA is a technique of utilizing the descriptive statistics and graphical tools that assist in comprehending the data in a better way. However, it is also utilized to maximize the insights of dataset and find the anomalies, outliers and then evalaute the underlying assumptions.

Generally, box plot is a graph, which confers a visual indication on how the minimum, maximum and outlier values of the dataset, which are spread and compare each other. Whereas, **Fig 3** shows the values of properties of the groundwater dataset are correlated each other. The box plot shows that the depth to groundwater in the region is generally between 15 and 35 feet. However, there are a few outliers that are either shallower or deeper than this range. The box plot also shows that the data is skewed towards the deeper end. The box plot shows that the groundwater levels in the dataset are normally distributed. The median groundwater level is 100 feet, and the majority of the data is within 25 feet of the median. There are two outliers in the dataset, one at 50 feet and one at 150 feet. The box plot can be used to compare the groundwater levels in different areas.

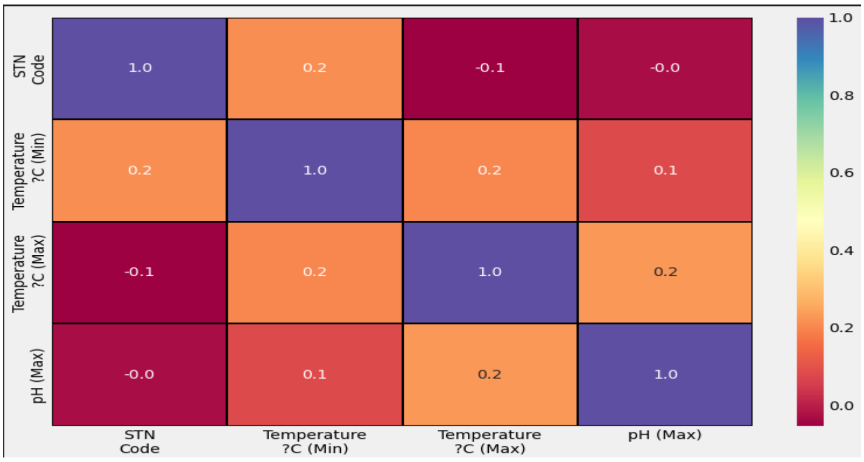


Fig 3. Box Plot of the Dataset.

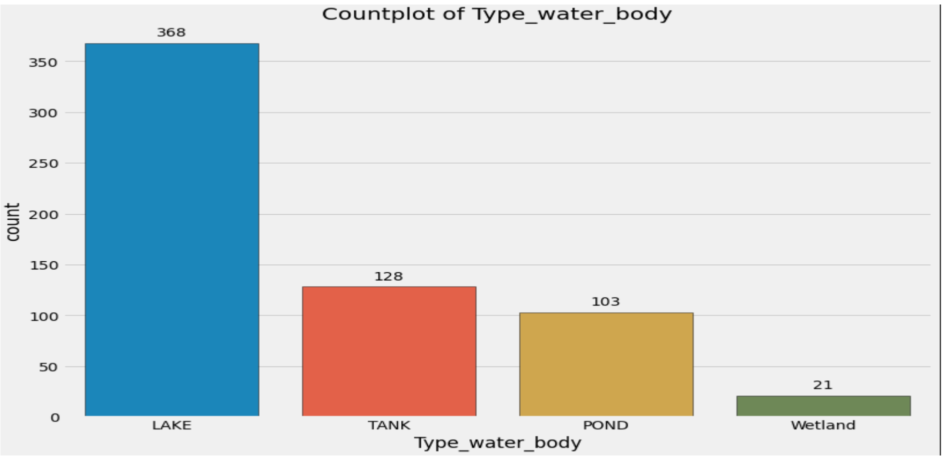


Fig 4. Count Plot of Type Water Body.

In **Fig 4**, the bar graph displays the frequency of different water body types, revealing that lakes have a higher count compared to other water bodies. The wetland type is smaller than the other type of water bodies. Nitrate is the primary form of nitrogen present in lakes and streams. The nitrate elements which includes KNO3 and KH2PO4. It readily dissolves in water and is usually the most prevalent form of nitrogen present in lakes. High levels of nitrate can cause serious issues with water quality, even though it is essential for plant growth. The addition of phosphorus to nitrates can lead to eutrophication, causing a noticeable increase in aquatic plants and changing the species composition of organisms in the stream.

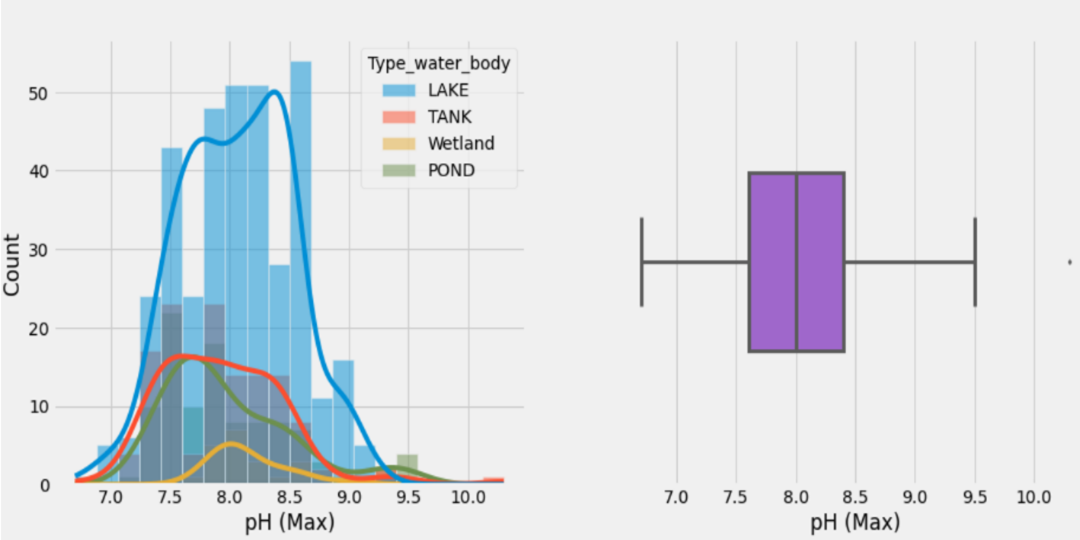


Fig 5. Histogram and Boxplot of pH (Max).

Fig 5 illustrates the histogram and boxplot of pH (Max). It clearly shows that the lake water body has maximum count with pH value. Whereas, the wetland water body has minimum count with minimum pH value. The histogram depicts the frequency distribution of pH values. It reveals that lake water bodies exhibit a wider range of pH values compared to wetlands. The lake water body has a significant peak at a pH value of 7, indicating that a substantial portion of the samples fall within a neutral pH range. The boxplot further elaborates on the pH distribution. The median pH value for the lake water body is slightly higher than that of the wetland water body, confirming the trend observed in the histogram. The interquartile range (IQR), represented by the box, indicates that the pH values in the lake water body are more dispersed, while those in the wetland water body are more tightly clustered.

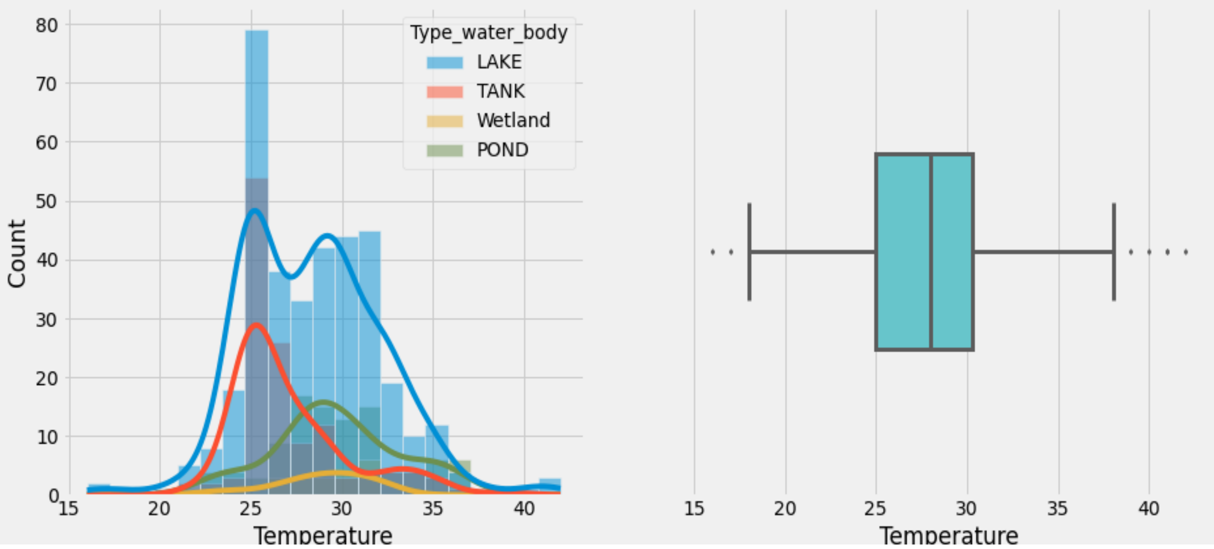


Fig 6. Histogram and Boxplot of Temperature (Max).

Fig 6 exposes the histogram and boxplot of Temperature (Max). The boxplot value of temperature (max) lies between 25 and 30. All water bodies have low count when its temperature (max) are high. High temperature (max) values are associated with a lower count of water bodies, suggesting that extreme temperatures may have negative consequences for water resources. The relationship between temperature extremes and water body availability has important implications for water resource management. As climate change continues to increase the frequency and intensity of extreme temperature events, it is essential to consider the potential impacts on water bodies and develop strategies to mitigate these effects.

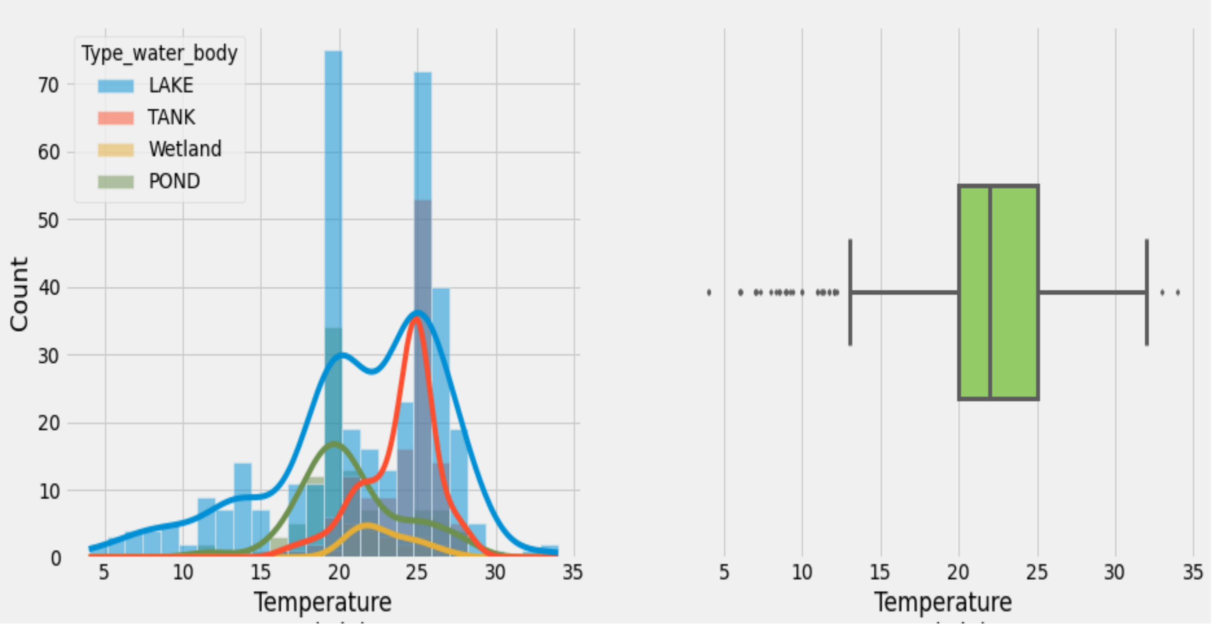


Fig 7. Histogram and Boxplot of Temperature (Min).

Fig 7 reveals the histogram and boxplot of Temperature (Min). Types of water body which is mentioned are lake, tank, wetland and pond. The histogram shows the frequency distribution of Temperature (Min). It indicates that the majority of water bodies have Temperature (Min) values between 20 and 25 degrees Celsius. There are relatively fewer water bodies with very low or very high Temperature (Min) values. The boxplot value of temperature (min) lies between 20 and 25. All water bodies have low count when its temperature (min) are high. The boxplot provides a summary of the Temperature (Min) distribution for each type of water body. The median (middle value) of Temperature (Min) is represented by a line within the box. The box itself represents the interquartile range (IQR), which encompasses the middle 50% of the data. The whiskers extend from the edges of the box to the most extreme data points that are not considered outliers.The boxplot shows that Temperature (Min) values are generally higher in tanks and wetlands compared to lakes and ponds. The IQR is largest for wetlands, indicating greater variability in Temperature (Min) among this type of water body.

Table 1. Factors Of the Nitrate Elements in Lake, Pond, Tank, And Wetland

Factors	**Lake**	**Tank**	**Pond**	**Wetland**
PO_4^-	0.853	0.752	0.801	0.716
Depth	0.884	0.754	0.737	0.643
Temp	0.627	0.623	0.612	0.565
$SO_4$2-	0.683	0.534	0.331	0.447
K+	0.476	0.195	0.337	0.267
Ca2+	0.496	0.259	0.321	0.165
Na+	0.577	0.168	0.258	0.401
EC	0.282	0.179	0.223	0.103
HCO3-	0.354	0.261	0.254	0.231
Cl^-	0.235	0.158	0.177	0.219
Mg2+	0.039	0.026	0.002	0.021
Salinity	0.199	0.041	0.024	0.109
pH	0.065	0.039	0.044	0.052
AS	0.069	0.063	0.027	0.046
F-	0.972	0.917	0.922	0.869

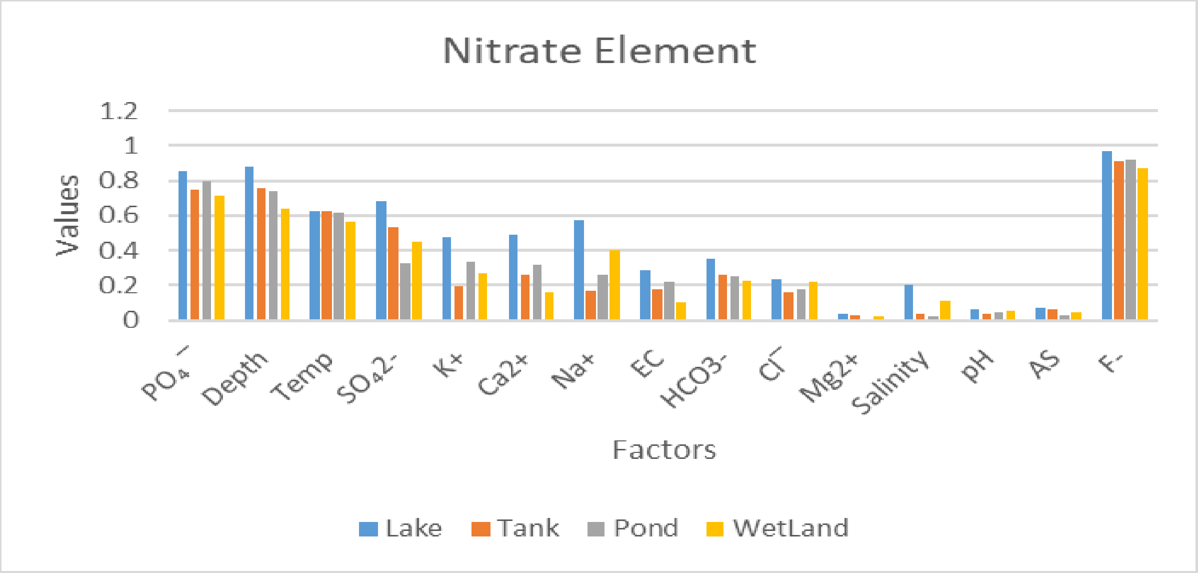
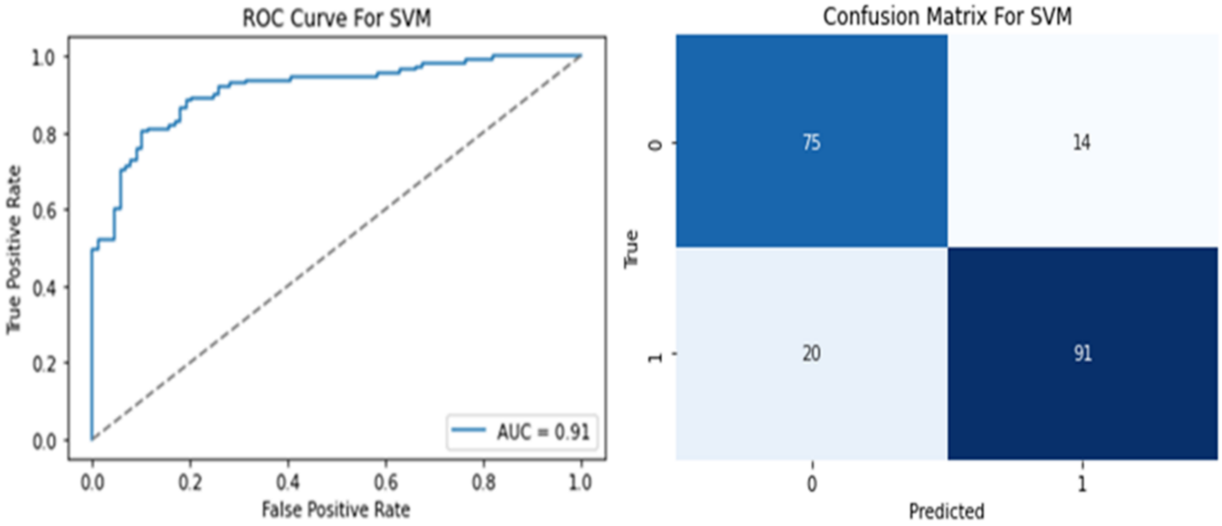


Fig 8. Nitrate elements present in lake, pond, tank, Wetland.

Fig 8 and **Table 1** explains about the measurement of various nitrate elements which is present in the lake, pond, tank and wetland. Totally, 15 factors are examined which describes the elements present in various places and it is represented as the graphical presentation. **Table 1** provides a detailed breakdown of the nitrate elements present in each of the aquatic environments studied. The table lists the 15 factors examined, along with their corresponding nitrate concentrations in each environment. It is evident th at the nitrate concentrations in lakes and are generally higher than those in tanks, ponds and wetlands. This could be due to the larger surface area of lakes and ponds, which allows for greater interaction with the atmosphere and the surrounding environment.

Performance Analysis

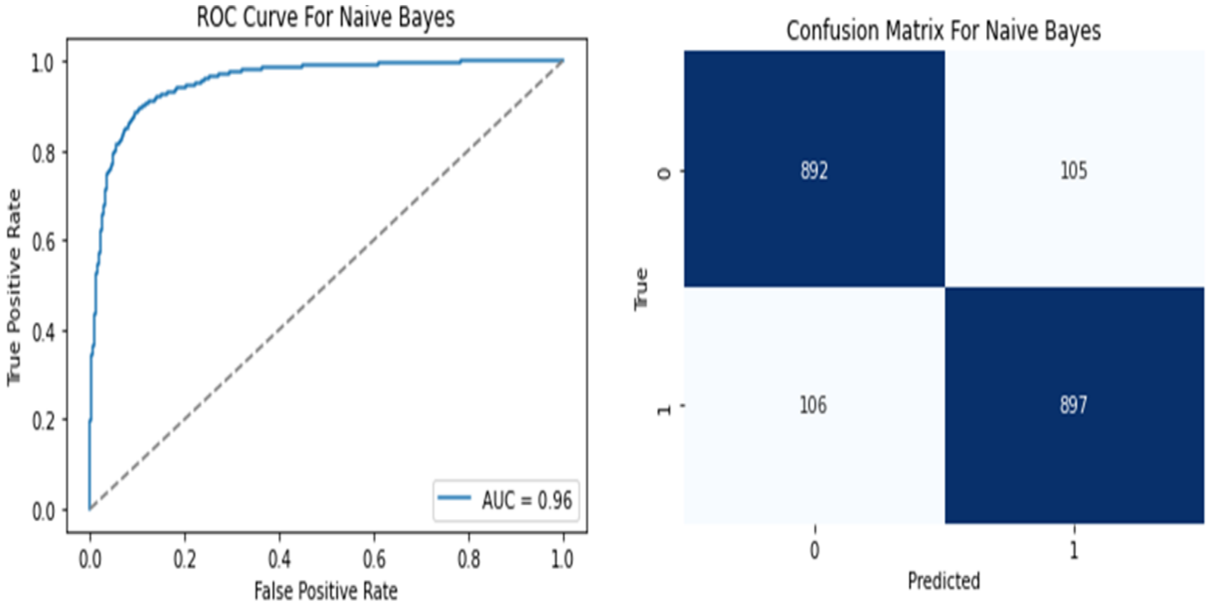
It is the process of analysing the performance of the suggested algorithms with the existing algorithms. However, the performance analysis is evaluated using ROC curve and confusion matrix of the proposed and existing algorithms. In the proposed study, the performance of the novel split gini indexing algorithm with SVM, KNN and NB algorithms.



(a) **(b)**

Fig 9. (a) ROC Curve for SVM (b) Confusion Matrix for SVM.

The ROC curve for SVM signifies the graphical representation of the execution of the introduced model. It also explains the trade-off among FPR and TPR at various classification thresholds. However, it is generated by plotting FPR on x-axis and TP on y-axis. ROC curve of the introduced model is show in the **Fig 9** (a), which confers the accuracy value of 0.91 representing the classifier's execution. **Fig 9** (b) represents the confusion matrix, which evaluates the effectiveness of the classifier regarding the target classes.



(a) **(b)**

Fig 10. (a) ROC Curve for Naïve Bayes (b) Confusion Matrix for Naïve Bayes.

The ROC curve for NB represents a graphical depiction that illustrates the execution of the suggested system. However, the ROC curve of the introduced model is shown in **Fig 10** (a), which provides the accuracy of value 0.96 denoting the improved performance of the classifier. **Fig 10** (b) represents the confusion matrix, which evaluates the effectiveness of the classifier regarding the target classes. A confusion matrix is a square matrix that evaluates the performance of a classifier regarding its target classes. The values in actual and predicted labels are almost similar, which shows its enhanced performance. The results of the NB model is 892 TP, 105 TN, 106 FN, and 897 FP. The confusion matrix indicates that the NB model has a high true positive rate (892 TP) and a low false negative rate (106 FN), suggesting that it is effective in identifying positive instances. However, it also has a relatively high false positive rate (897 FP), indicating that it may be prone to misclassifying negative instances as positive.

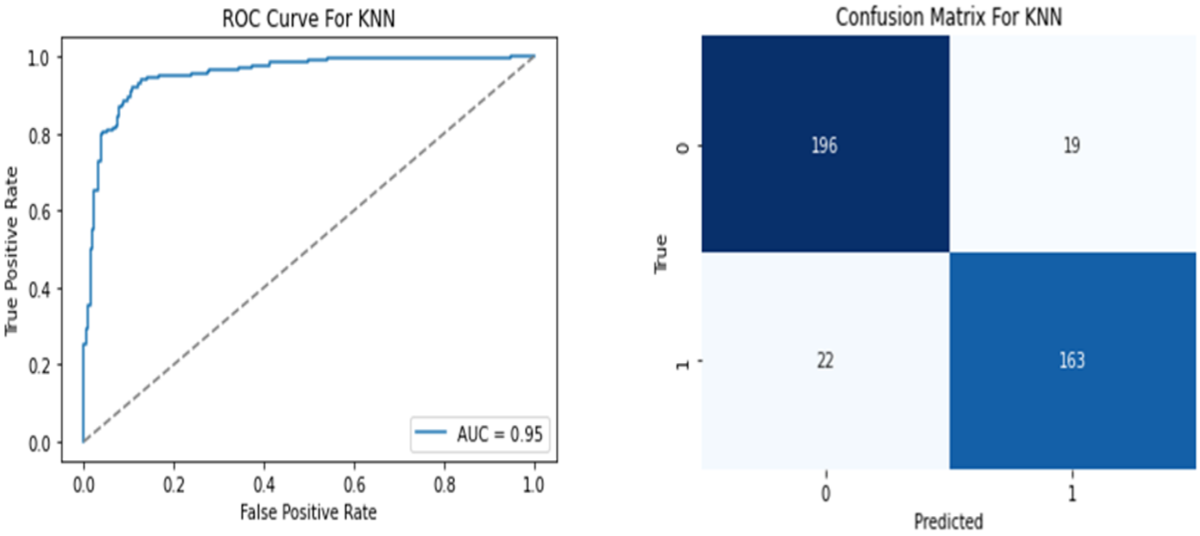
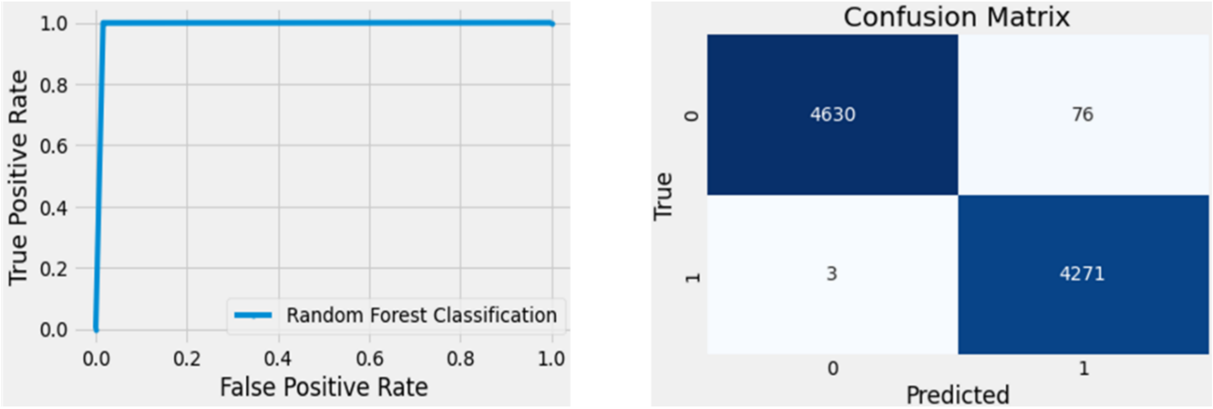


Fig 11. (a) ROC Curve for KNN (b) Confusion Matrix for KNN.

ROC curve for KNN denotes a graphical representation that explains the performance of the introduced method. In **Fig 11** (a), the accuracy value is 0.95, which represents the enhanced performance of the classifier. Whereas, **Fig 11** (b) denotes the confusion matrix that assesses the effectiveness of the classifier based on the target classes. Results are 196 TP, 19 TF, 22 FN, and 163 FP. The confusion matrix provides additional insights into the classifier's performance beyond the ROC curve. For example, it shows that the classifier has a high sensitivity (recall) of 196 / (196 + 22) = 0.9, indicating that it correctly identifies a large proportion of positive instances. However, it also has a relatively high false positive rate of 19 / (163 + 19) = 0.11, indicating that it incorrectly classifies a small proportion of negative instances as positive.



(a) **(b)**

Fig 12. (a) ROC Curve of RF (b) Confusion Matrix.

ROC curve for proposed model denotes a graphical representation that explains the enhanced performance of the introduced method. In **Fig 12** (a), the accuracy value is 1, which represents the enhanced performance of the classifier. Whereas, **Fig 12** (b) denotes the confusion matrix that assesses the effectiveness of the classifier based on the target classes. The outcome of the confusion matrix is 4630 TP, 76 TF, 3 FP, and 4271 FN. The ROC curve and confusion matrix presented which provides evidence of enhanced performance of the proposed model. The ROC curve shows an AUC value of 1, indicating perfect discrimination between target classes. The confusion matrix demonstrates the model's ability to accurately identify a large number of true positives and true negatives, while minimizing false positives and false negatives. These results highlight the effectiveness of the proposed model for the task at hand.

Comparative Analysis

Comparative analysis is the process of comparing the proposed model with the existing algorithms to access the overall execution of the suggested model, which are shown in **Table 1**. Similarly, the corresponding performance analysis is projected in the form of graphical representation in **Fig 7.** In this, class 0 denotes the performance without nitrate and class 1 is for performance with nitrate.

Table 2. Performance Metrics of RF

Classes	**A**	**P**	**R**	**F1**
0	0.99	1	0.98	0.99
1	0.99	0.99	1	0.99

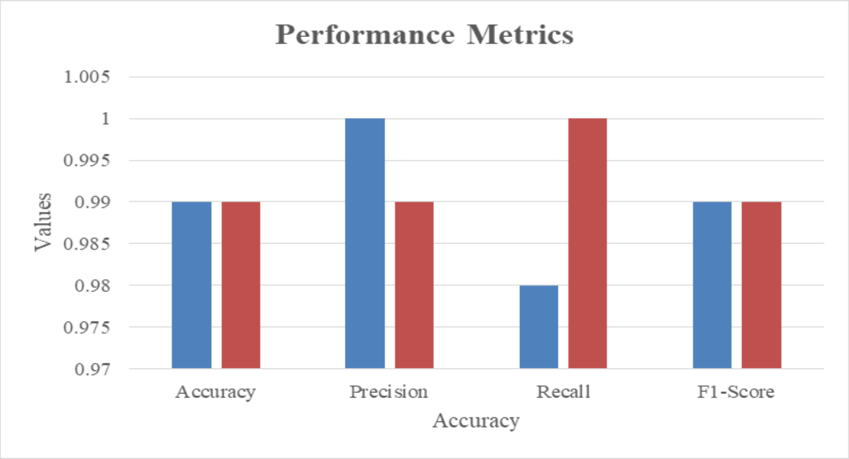


Fig 13. Performance Metrics of RF – Graphical Representation.

The performance metrics of RF model is shown in **Table 2** and **Fig 13.** The A value for class 0 and class 1 is 0.99, and the P value for class 0 is 1, which is greater than class 0. However, the R values of the class 1 is 1, which is greater than the recall value of class 0. Finally, the value of F1 for classes 0 and 1 is 0.99.

Table 3. Comparative Analysis of NB with the Proposed Model

Classes	**A**	**P**	**R**	**F1**
0	0.96	0.94	0.95	0.93
1	0.96	0.93	0.95	0.94
Proposed	**0.99**	**0.99**	**1**	**0.99**

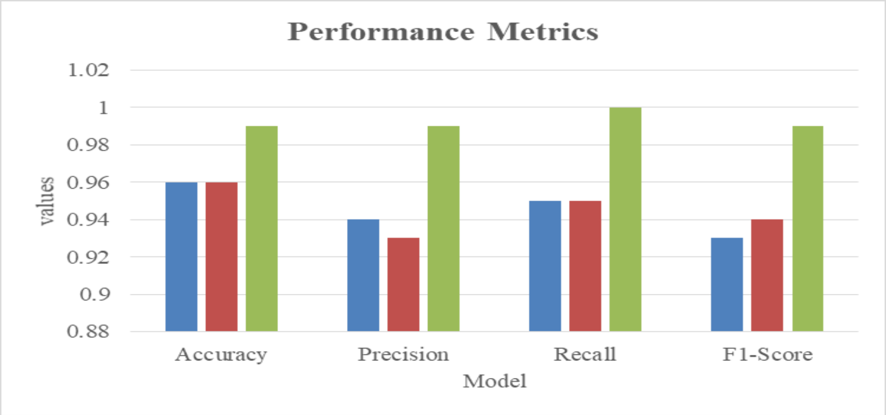


Fig 14. Comparative Analysis of NB with the Proposed Model – Graphical Representation.

Fig 14 and **Table 3** reveal the comparative analysis of the proposed model over NB. The performance metrics like A, P, R and F1 values of the suggested model are higher than the NB classes. The accuracy, precision and F1score values of the suggested model is 0.99, whereas the recall value of the introduced model is 1. These values of the introduced model shows that the projected model has the better efficiency than the NB model.

Table 4. Comparative Analysis of SVM with the Proposed Model

Classes	**A**	**P**	**R**	**F1**
0	0.91	0.83	0.81	0.83
1	0.91	0.82	0.83	0.84
Proposed	**0.99**	**0.99**	**1**	**0.99**

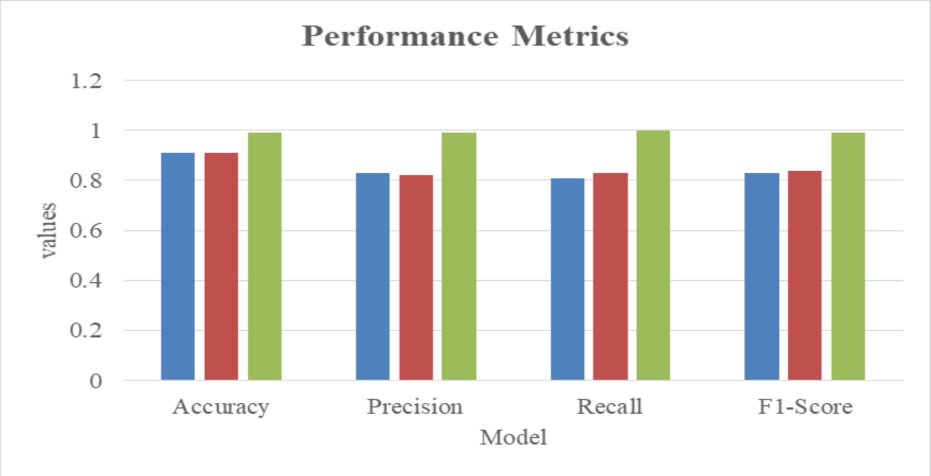


Fig 15. Comparative Analysis of SVM with the Proposed Model - Graphical Representation.

The comparative analysis of the proposed model with SVM is shown in the **Fig 15** and **Table 4**. The proposed model has high performance metrics than SVM. The recall value of the proposed system is 1, whereas the other performance metrics remain 0.99. This indicates that the proposed model is highly efficient than SVM.

Table 5. Comparative Analysis of KNN with the Proposed Model

Classes	**A**	**P**	**R**	**F1**
0	0.95	0.92	0.92	0.93
1	0.95	0.93	0.94	0.92
Proposed	**0.99**	**0.99**	**1**	**0.99**

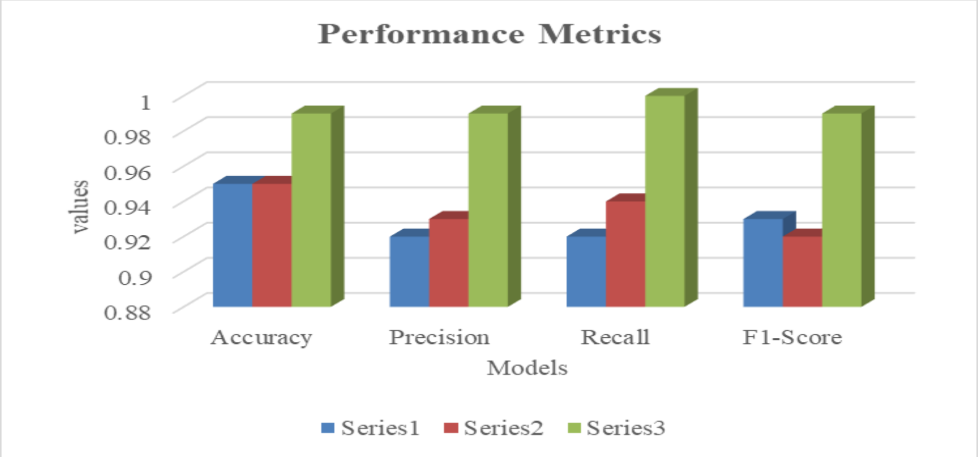


Fig 16. Comparative Analysis of KNN with the Proposed Model - Graphical Representation.

Table 5 and **Fig 16** show that the comparative analysis of the projected model with KNN. Whereas, the proposed model has high performance metrics than KNN. The values of A, P and F1 score is higher than the KNN. The R value of the suggested model is 1, which shows the better efficiency of the model than KNN. Hence, the evaluation of the groundwater quality is enhanced by the proposed model than the conventional techniques.

V. CONCLUSION

In recent days, groundwater nitrate contamination is one of the crucial issues, which led to various health issues. In this study, the RF classifier is used to evaluate the groundwater quality for determining the nitrate concentration. The proposed model uses the associative rule mining for rule ranking the datasets that are pre-processed from the loaded dataset. The novel split gini indexing algorithm is used with RF in order to split the data into subsets. Classification algorithms like NB, SVM and KNN is used in order to classify the dataset. Whereas, the performance measures like precision, accuracy, F1 score and recall are calculated from the proposed model in order to evaluate its efficiency. Similarly, performance metrics of suggested system is compared with the performance metrics of the classification algorithms SVM, KNN and NB. The values of accuracy of the proposed model are higher, which are 0.99 over NB, SVM and KNN. Similarly, the values of precision of the proposed model are greater, which are 0.99 over NB, SVM and KNN. And, the recall values of the proposed model are 1 over NB, SVM and KNN and the F1 score values of the proposed model are 0.99 over NB, SVM and KNN. This indicates that the suggested method confers the better efficiency in evaluating the groundwater quality for nitrate concentration than the other models. In future, the research work will be enhanced by implementing more classification approaches to improve the accuracy of the projected model.

Declarations

All the authors mentioned in the manuscript have agreed for authorship, read and approved the manuscript, and given consent for submission and subsequent publication of the manuscript.

Data Availability

No data was used to support this study.

Conflicts of Interests

The author(s) declare(s) that they have no conflicts of interest.

Funding

No funding agency is associated with this research.

Competing Interests

There are no competing interests.

Author contributions:

All authors have contributed equally.

References

[1]. T. G. Nguyen, K. A. Phan, and T. H. N. Huynh, “Application of Integrated-Weight Water Quality Index in Groundwater Quality Evaluation,” Civil Engineering Journal, vol. 8, no. 11, pp. 2661–2674, Nov. 2022, doi: 10.28991/cej-2022-08-11-020.

[2]. H. Soleimani et al., “Groundwater quality evaluation and risk assessment of nitrate using monte carlo simulation and sensitivity analysis in rural areas of Divandarreh County, Kurdistan province, Iran,” International Journal of Environmental Analytical Chemistry, vol. 102, no. 10, pp. 2213–2231, Apr. 2020, doi: 10.1080/03067319.2020.1751147.

[3]. S. He, J. Wu, D. Wang, and X. He, “Predictive modeling of groundwater nitrate pollution and evaluating its main impact factors using random forest,” Chemosphere, vol. 290, p. 133388, Mar. 2022, doi: 10.1016/j.chemosphere.2021.133388.

[4]. S. Shukla and A. Saxena, “Sources and Leaching of Nitrate Contamination in Groundwater,” Current Science, vol. 118, no. 6, p. 883, Mar. 2020, doi: 10.18520/cs/v118/i6/883-891.

[5]. Q. Zhang, H. Qian, P. Xu, W. Li, W. Feng, and R. Liu, “Effect of hydrogeological conditions on groundwater nitrate pollution and human health risk assessment of nitrate in Jiaokou Irrigation District,” Journal of Cleaner Production, vol. 298, p. 126783, May 2021, doi: 10.1016/j.jclepro.2021.126783.

[6]. L. Knoll, L. Breuer, and M. J. E. R. L. Bach, "Nation-wide estimation of groundwater redox conditions and nitrate concentrations through machine learning," vol. 15, no. 6, p. 064004, 2020.

[7]. K. Fatah, “Evaluation Groundwater Quality By Using Gis And Water Quality Index Techniques For Wells In Bardarash Area, Northern Iraq,” Iraqi Geological Journal, vol. 53, no. 2C, pp. 87–104, Sep. 2020, doi: 10.46717/igj.53.2c.7rs-2020-09.07.

[8]. K. A. Ahmed, M. El-Rawy, A. M. Ibraheem, N. Al-Arifi, and M. K. Abd-Ellah, “Forecasting of Groundwater Quality by Using Deep Learning Time Series Techniques in an Arid Region,” Sustainability, vol. 15, no. 8, p. 6529, Apr. 2023, doi: 10.3390/su15086529.

[9]. S. S. Band et al., “Comparative Analysis of Artificial Intelligence Models for Accurate Estimation of Groundwater Nitrate Concentration,” Sensors, vol. 20, no. 20, p. 5763, Oct. 2020, doi: 10.3390/s20205763.

[10]. L. Knoll, L. Breuer, and M. Bach, “Nation-wide estimation of groundwater redox conditions and nitrate concentrations through machine learning,” Environmental Research Letters, vol. 15, no. 6, p. 064004, May 2020, doi: 10.1088/1748-9326/ab7d5c.

[11]. S. Bedi, A. Samal, C. Ray, and D. Snow, “Comparative evaluation of machine learning models for groundwater quality assessment,” Environmental Monitoring and Assessment, vol. 192, no. 12, Nov. 2020, doi: 10.1007/s10661-020-08695-3.

[12]. Y. Teng, R. Zuo, Y. Xiong, J. Wu, Y. Zhai, and J. Su, “Risk assessment framework for nitrate contamination in groundwater for regional management,” Science of The Total Environment, vol. 697, p. 134102, Dec. 2019, doi: 10.1016/j.scitotenv.2019.134102.

[13]. S. D. S. Putro and W. Wilopo, “Assessment of nitrate contamination and its factors in the urban area of Yogyakarta, Indonesia,” Journal of Degraded and Mining Lands Management, vol. 9, no. 4, p. 3643, Jul. 2022, doi: 10.15243/jdmlm.2022.094.3643.

[14]. H. E. Elzain et al., “Novel machine learning algorithms to predict the groundwater vulnerability index to nitrate pollution at two levels of modeling,” Chemosphere, vol. 314, p. 137671, Feb. 2023, doi: 10.1016/j.chemosphere.2022.137671.

[15]. R. Haggerty, J. Sun, H. Yu, and Y. J. W. R. Li, "Application of machine learning in groundwater quality modeling-A comprehensive review," vol. 233, p. 119745, 2023.

[16]. Mazraeh, M. Bagherifar, S. Shabanlou, and R. Ekhlasmand, “A Hybrid Machine Learning Model for Modeling Nitrate Concentration in Water Sources,” Water, Air, &amp; Soil Pollution, vol. 234, no. 11, Nov. 2023, doi: 10.1007/s11270-023-06745-3.

[17]. M. Ghaderzadeh, M. Aria, and F. Asadi, “X-Ray Equipped with Artificial Intelligence: Changing the COVID-19 Diagnostic Paradigm during the Pandemic,” BioMed Research International, vol. 2021, pp. 1–16, Aug. 2021, doi: 10.1155/2021/9942873.

[18]. Bhattarai, S. Dhakal, Y. Gautam, and R. Bhattarai, “Prediction of Nitrate and Phosphorus Concentrations Using Machine Learning Algorithms in Watersheds with Different Landuse,” Water, vol. 13, no. 21, p. 3096, Nov. 2021, doi: 10.3390/w13213096.

[19]. R. Haggerty, J. Sun, H. Yu, and Y. Li, “Application of machine learning in groundwater quality modeling - A comprehensive review,” Water Research, vol. 233, p. 119745, Apr. 2023, doi: 10.1016/j.watres.2023.119745.

[20]. S. Sahour et al., "Evaluation of machine learning algorithms for groundwater quality modeling," vol. 30, no. 16, pp. 46004-46021, 2023.

[21]. R. Md. T. Islam et al., “Application of novel framework approach for prediction of nitrate concentration susceptibility in coastal multi-aquifers, Bangladesh,” Science of The Total Environment, vol. 801, p. 149811, Dec. 2021, doi: 10.1016/j.scitotenv.2021.149811.

[22]. S. C. Pal, D. Ruidas, A. Saha, A. R. Md. T. Islam, and I. Chowdhuri, “Application of novel data-mining technique based nitrate concentration susceptibility prediction approach for coastal aquifers in India,” Journal of Cleaner Production, vol. 346, p. 131205, Apr. 2022, doi: 10.1016/j.jclepro.2022.131205.

[23]. H. Raheja, A. Goel, and M. Pal, “Prediction of groundwater quality indices using machine learning algorithms,” Water Practice and Technology, vol. 17, no. 1, pp. 336–351, Dec. 2021, doi: 10.2166/wpt.2021.120.

[24]. P. Agrawal et al., “Exploring Artificial Intelligence Techniques for Groundwater Quality Assessment,” Water, vol. 13, no. 9, p. 1172, Apr. 2021, doi: 10.3390/w13091172.

[25]. S. Sahour, M. Khanbeyki, V. Gholami, H. Sahour, I. Kahvazade, and H. Karimi, “Evaluation of machine learning algorithms for groundwater quality modeling,” Environmental Science and Pollution Research, vol. 30, no. 16, pp. 46004–46021, Jan. 2023, doi: 10.1007/s11356-023-25596-3.

[26]. S. Sahour, M. Khanbeyki, V. Gholami, H. Sahour, H. Karimi, and M. Mohammadi, “Particle swarm and grey wolf optimization: enhancing groundwater quality models through artificial neural networks,” Stochastic Environmental Research and Risk Assessment, vol. 38, no. 3, pp. 993–1007, Nov. 2023, doi: 10.1007/s00477-023-02610-1.