WATER QUALITY ANALYSIS
USING RANDOM FOREST REGRESSION

ABSTRACT

A Regression algorithm is used to assign predefined classes to test instances for evaluation (or) future instances to an application. This study presents a Regression model using Random Forest Algorithm to analyze water quality. Water quality is very important in ensuring citizens can get to drink clean water. Application of Random Forest as an Ensemble Techniques to predict clean water based on the water quality parameters can ease the work of the laboratory technologist by predicting which water samples should proceed to the next step of the analysis. Regression using Random Forest was applied to predict the clean and not clean water. The analysis of water Hardness, solids, Turbidity, pH level, Sulfate, and conductivity can play a major role in assessing water quality. Nowadays Most Diseases are caused by Using Water To avoid Those diseases We are Implementing This Model.

KEYWORDS: Random Forest Regression, Ensemble Techniques, Etc.

INTRODUCTION

Supervised learning is a machine learning algorithm that receives a feature vector and the target pattern as an input to build a model. The model can be used to recognize new patterns and assign a target to them. Applications of supervised learning include classification and Regression, Unsupervised learning is a machine learning algorithm that only receives the feature vector as an input, and its task is to find similar groups of items with comparable features. The essential application of unsupervised learning is clustering, such as determining the distribution of data items within a multidimensional space of given data.

Regression is an instance of supervised learning that includes a training phase to create a model (Regressor). Its task is to predict the class of items in a data set using a certain model of a Regression.

The model is constructed using already-labeled. The model is constructed using already-labeled items of similar data sets. This step allows Regression techniques to be considered as a supervised machine learning method. **Ensemble Techniques** are used to predict High accuracy using **hyperparameter tuning** and not only the regressor, but the Classifier algorithm also uses ensemble Techniques.

Different Regression models are created by using different Regression algorithms, which can be divided into three main categories: **Linear Regression**, **Logistic Regression**, and **Random Forest Regression**. These classifiers are discussed in the following subsections, considering Random Forest Regressor used for the experiment in this research.

Water quality analysis is required mainly for Some importance of such required assessments include:

(i) To check whether the water quality complies with the standards, and hence, suitable or not for the designated use.

(ii) To monitor the efficiency of a system, working for water quality maintenance

(iii) To check whether-gradation / change of an existing system is required and to decide what changes should take place

(iv) To monitor whether water quality complies with rules and regulations.

(v) Water quality analysis is extremely necessary for the sectors of:

- Public Health (especially for drinking water)
- Industrial Use

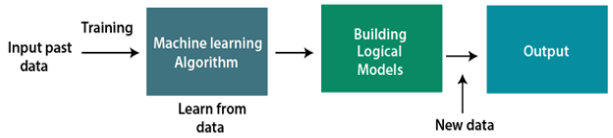
LITERATURE REVIEW

The Neural Network of Machine Learning is based on Random Forest Regression to obtain a proper solution to address the problem of changes in the quality of drinking water.

- Regression is an important problem in machine learning. It has been widely applied in many real-world applications examples as Food testing, loan prediction, and checking online fraud payments. To build a Regression, a user first needs to collect a set of training examples/instances that are labeled with predefined classes. A Regression algorithm is then applied to the training data to build a Regressor that is subsequently employed to assign the predefined classes to test instances (for evaluation) or future instances (for application).
- Random forest is a tree-based algorithm that involves the building of several trees and combining them with the output to improve the generalization ability of the model. This method of combining trees is known as an ensemble Technique where we can increase the accuracy of the model. The ensemble is nothing but a combination of weak learners (individual trees) to produce a strong learner of the given data.
- The bagging Algorithm is used to create random samples. Take data set D1 is given for n rows and m columns, and new data set D2 is created for sampling n cases at random with replacement from the original data to Predict a Machine Learning model. From dataset D1,1/3rd of rows are left out and are known as Out of Bag samples. Then, a new dataset D2 is trained to this model, and Out of Bag samples is used to determine an unbiased estimate of the error. Out of m columns, M << m columns are selected at each node in the data set. TheM columns are selected at random. Usually, the default choiceof M is m/3 for the regression tree and M is sqrt(m) for the classification tree. Unlike a tree, no pruning takes place in a random forest i.e; each tree is grown fully. For indecision trees, pruning is a method to avoid overfitting. Pruning meansselecting a subtree that leads to the lowest test error rate. Cross-validation is used to determine the test error rate of a subtree. Several trees are grown and the final prediction is obtained by averaging or voting.

EXISTING SYSTEM

This research was based on unsupervised learning. The significance of this paper was to find new methods for Water Analysis and to increase the accuracy of results. The data set for this paper is based on real-life transactional data by a large European company and personal details in data are kept confidential and safe. The accuracy of an algorithm is around 70%. Thus, the accuracy of the results obtained from these methods is less when compared with the proposed system. A comprehensive understanding of the quality of a water sample can be helpful for us to solve the problem of Water Quality Analysis. The work provides a comprehensive discussion of the challenges and problems of water.



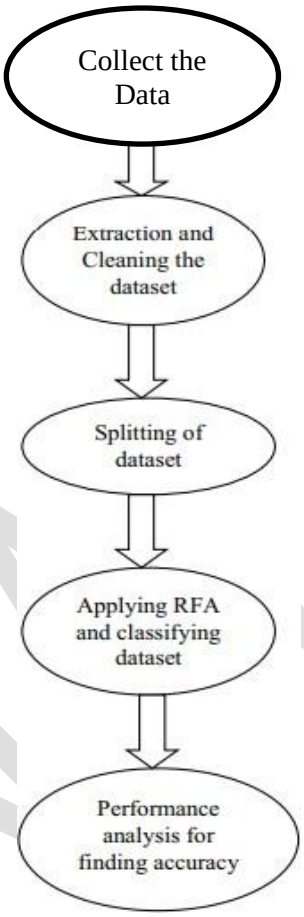
PROPOSED SYSTEM

In the proposed system, we use RFA for the classification and regression of the dataset. First, we will collect the Water Quality dataset and analysis will be done on the collected dataset. After the analysis of the dataset then cleaning of the dataset is required. Generally, in any dataset there will be many duplicates and null values will be present, so to remove all those duplicates and null values cleaning process is required. Then we must split the dataset into two categories a trained dataset and a testing dataset for comparing and analyzing the dataset. After dividing the dataset, we must apply the RFA where this algorithm will give us better accuracy about the Water Quality Measures. By applying the RFA, the dataset will be classified into four categories which will be obtained in the form of a confusion matrix. In this analysis, the accuracy of Water Quality Analysis can be obtained which will be finally represented in the form of a graphical representation.

RFA

A random forest is also called a random decision forest which is used for classification, regression, and other tasks that are performed by constructing multiple decision trees. This RFA is based on supervised learning and the major advantage of this algorithm is that it can be used for both classification and regression. RFA gives you better accuracy when compared with all other existing systems and this is the most used algorithm. In this paper, the use of RFA in credit card fraud detection can give you an accuracy of about 90 to 95%.

VI.MODULES



MODULE 1: Exploratory Data Analysis In this module we will first collect all the Water Quality analysis dataset and store it in a database. Then we will perform some descriptive analysis of the dataset.

MODULE 2: Data Cleaning Is the next step, after analyzing the dataset then we have to clean the data. In this cleaning process, all the duplicate values and null values that are present in the dataset will be removed and moved to further process.

MODULE 3: Preprocessing of dataset In this module the cleaned dataset will be preprocessed where the dataset will be divided based on Given Data

MODULE 4: Dataset Partition In this module first the dataset will be divided into two partitions a trained dataset and a testing dataset. After the data partitions, the Random Forest Algorithm is applied. After applying RFA finally, a confusion matrix is obtained.

MODULE 5: Evaluation Now the resultant data obtained in the form of a confusion matrix can be evaluated by using graphical representation which gives better accuracy.

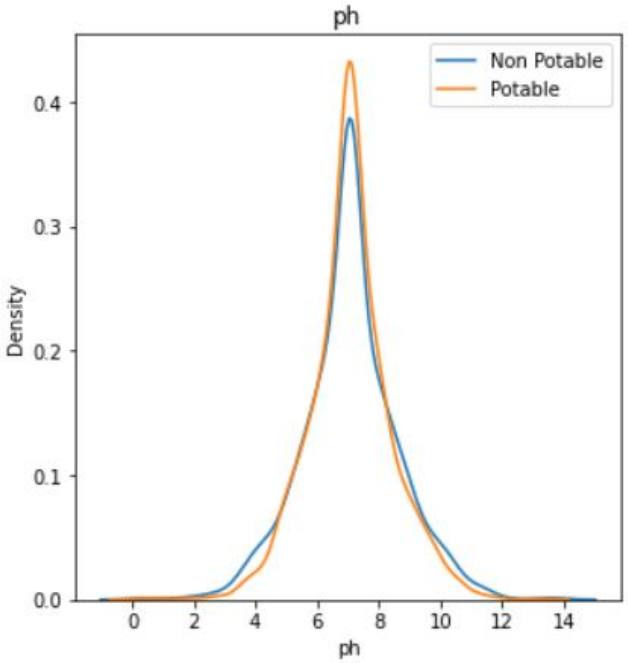
VII.PERFORMANCE METRICS AND EXPERIMENTAL RESULTS

PERFORMANCE METRIX

The basic performance equations are derived from the confusion matrix. The confusion matrix is a 2 by 2 matrix table that contains four outcomes produced by the binary classifier. Various measures such as sensitivity, specificity, accuracy, and error rate are derived from the confusion matrix.

Accuracy:

Accuracy is calculated as the total number of two correct predictions(A+B) divided by the total number of the dataset(C+D). It is calculated as (1-error rate).



Accuracy=A+B/C+D

Whereas,

A=True Positive

B=True Negative

C=Positive

D=Negative

Error rate:

The error rate is calculated as the total number of two incorrect predictions(F+E) divided by the total number of the dataset(C+D).

Error rate=F+E/C+D

Whereas:

E = False Positive F = False Negative
C = Positive D = Negative

Several positives(C).

Sensitivity=A/C

Specificity:

Specificity is calculated as per the number of correct negative predictions(B) divided by the

Sensitivity:

Sensitivity is calculated as the number of correct positivepredictions(A) divided by the total

total number of negatives(D). total number of negatives(D).

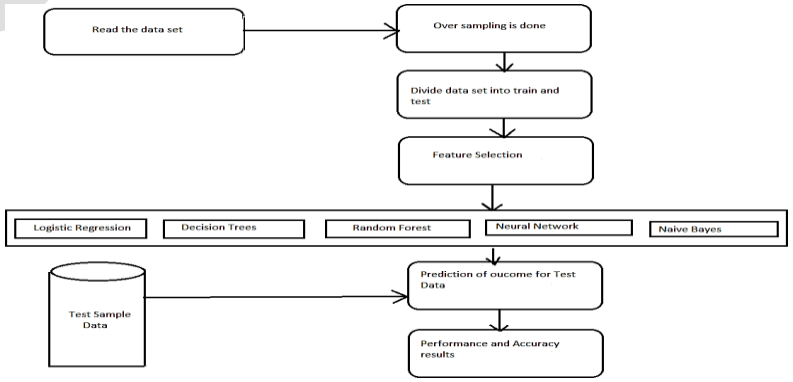
Specificity=B/D correct negativepredictions(B) divided by the total number of negatives(D).

Specificity=B/D

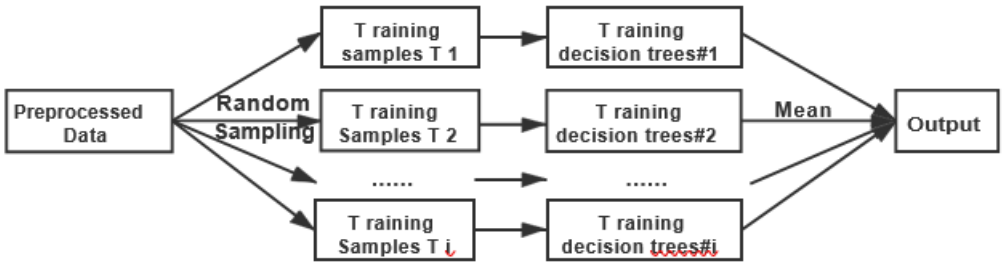
ALGORITHM STEPS

Step 1: Read the dataset.
Step 2: Random Sampling is done on the data set to make it balanced.
Step 3: Divide the dataset into two parts i.e., the Train dataset and the test dataset.
Step 4: Feature selection is applied to the proposed models.
Step 5: Accuracy and performance metrics have been calculated to know the efficiency of different algorithms.
Step6: Then retrieve the best algorithm based on efficiency for the given dataset.

ARCHITECTURE DIAGRAM



VIII.2 TRAINING PROCESS OF RANDOM FOREST

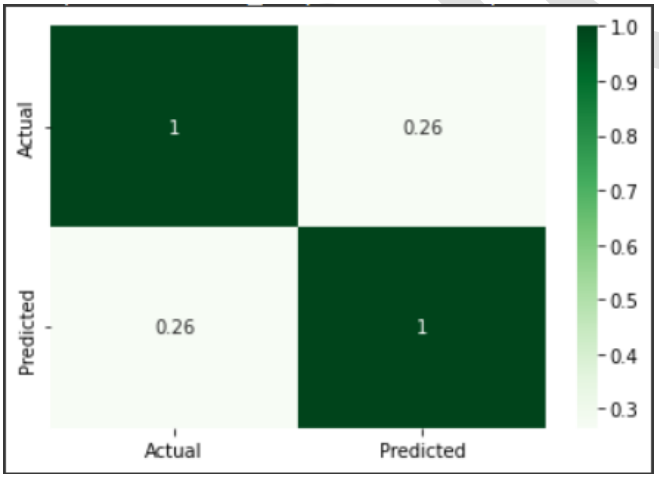


IX.WATER QUALITY PARAMETER WHICH IS USED FOR CREATING MODEL

Many parameters can influence the surface water quality. In this study, four parameters are selected for the investigation. The parameters which were used to determine whether the water is clean are PH level, Alkalinity, conductivity, and color.

X.EXPERIMENTAL RESULTS

This section shows the details and results of the experiments. Firstly, a performance comparison is made on the same subset. Then we explore the relation between a model's performance and the ratio of actual value and predicted value in a subset. Finally, it shows the performances of models on a much bigger dataset, which is more closed to the actual result.



XI.CONCLUSION

In this study, the water quality model was implemented using the Random Forest technique. The analysis of water Alkalinity, pH level, and conductivity can play a major role in assessing water quality. Although random forest obtains good results on small set data, there are still some problems such as imbalanced data. Our future work will focus on solving these problems. The algorithm of the random forest itself should be improved. For example, the voting mechanism assumes that each of the base classifiers has equal weight, but some of them may be more important than others. Therefore, we also try to make some improvements to this algorithm. By using the Random Forest algorithm we got an accuracy of 70% and to enhance that accuracy we have used ensemble techniques. Finally, we have predicted high accuracy when compared to other algorithm models.

REFERENCES

1. Jiang, J.; Tang, S.; Han, D.; Fu, G.; Solomatin, D.; Zheng, Y. A comprehensive review on the design and optimization of surface water quality monitoring networks. Environ. Model. Soft. 2020, 132, 104792.

2. Zhou, Y.; Yu, D.; Yang, Q.; Pan, S.; Gai, Y.; Cheng, W.; Liu, X.; Tang, S. Variations of Water Transparency and Impact Factors in the Bohai and Yellow Seas from Satellite Observations. Remote Sens. 2021, 13, 514.

3.Mateo Pérez, V.; Mesa Fernández, J.M.; Villanueva Balsera, J.; Alonso Álvarez, C. A Random Forest Model for the Prediction of FOG Content in Inlet Wastewater from Urban WWTPs. Water 2021, 13, 1237.

4. Zhou, Z.H. Ensemble learning. In Machine Learning; Springer: Berlin/Heidelberg, Germany, 2021; pp. 181–210

5. Karami, J.; Alimohammadi, A.; Seifouri, T. Water quality analysis using a variable consistency dominance-based rough set approach. Comput. Environ. Urban Syst. 2014, 43, 25–33.