# Efficient Water Quality Analysis and Prediction using Machine Learning

## Problem Statement:

To develop a web application that efficiently analyses and predicts water quality using machine learning algorithm.

## Introduction:

Water constitutes 70% of the earth's surface. It is considered as one of the most important resources for sustaining life. Rapid Industrialization and urbanization have led to contamination of water at an alarming rate which causes grieving diseases. Water quality degradation is one of the most critical problems currently faced by India. India is ranked 120$^{th}$ among 122 countries in water quality index, with nearly 70% of water body contaminated. So, there rise a necessity for water quality treatment and analysis. Owing to the dearth of real time water quality assessment and decision support systems in India, water quality assessment is currently carried out only in research laboratories where data is processed in non-real time. For this purpose, data analysis with a number of machine learning algorithms have been applied to predict the quality of water.

## Literature Survey:

### Surface Water Pollution Detection using Internet of Things

In the paper "Surface Water Pollution Detection using Internet of Things" proposed an Internet of Things based water quality system capable of measuring the quality of water in near real time. The solution proposed in the paper considers water quality metrics defined by World Health Organization. The architecture proposed in this paper combines the hardware and software solution. The microcontroller – ATMega328 forms the main part of the hardware architecture. The software solution consists of mobile application and data-analytics module. The mobile app comprises dashboard to remotely monitor the water quality. The data-analytics module in the cloud uses machine learning algorithms to predict the water quality.

The hardware solution sends data to the cloud for real-time storage and processing. The predictive analysis is done on the collected data. The models considered for analysis are:

- Support Vector Machine (SVM):

  SVM is a supervised machine learning algorithm which constructs hyperplane to classify data-points into classes. SVMs are different from other classification algorithms because of the way they choose the decision boundary that maximizes the distance from the nearest data points of all the classes. The decision boundary created by SVMs is called the maximum margin classifier or the maximum margin hyper plane.
- Neural Network:

  Neural Networks are considered as the heart of deep learning. Neural Networks are designed in such a way that a computer system modelled on the human brain and nervous system.

- K- Nearest Neighbors (KNN):

  The k-nearest neighbors algorithm is a non-parametric supervised learning classifier which uses the concept of similarity to group data points. It groups the input-data points to the group which contains similar data-points.

Then the processed data can be remotely monitored and water flow can be controlled using our developed software solution comprising of mobile app and a dashboard.

The estimated water quality in their work is based on only three parameters: turbidity, temperature and pH, which are tested according to World Health Organization (WHO) standards. Using only three parameters and comparing them to standardized values is quite a limitation when predicting water quality.

## Efficient Prediction of Water Quality Index Using Machine Learning Algorithms:

The models proposed in the paper **"Efficient Prediction of Water Quality Index (WQI) Using Machine Learning Algorithms"** to predict the water quality index were Random Forest, Neural Network, Multinomial Logistic Regression, Bagged Tree Model and Support Vector Machine.

**Random Forest:**

Random Forest belongs to the supervised learning technique that is used in classification and regression problems. This method works by training a large number of decision trees. Decision trees are built from different samples and takes their majority vote for classification and average in case of regression. Random forests are faster than decision trees as they work with subsets of data and can solve a huge amount of features without any complications. Random subset of given proportion is generated based on the dividing features set.

**Neural Network:**

Neural network reflects the behavior of the human brain which finds the hidden relationships in data. They contain an input layer, one or more hidden layers and an output layer. Each node connects to another node and has a weight and threshold associated with it. If the output of a node is above the threshold, then the node sends data to the next layer of the network.

The output of neuron Y is:

$$Y = f(W_1.X_1 + W_2.X_2 + b)$$

where, $X_1$ and $X_2$ = Numerical input, $W_1$ and $W_2$ = Weights associated with the inputs, b = bias weight, and f is known as the activation function which is non-linear.

**Multinomial Logistic Regression:**

Multinomial Logistic Regression is a classification technique that helps the logistic regression algorithm to solve multiclass problems. It helps to identify categorical data in various fields. A response variable which measures the rate of relative significance of independents, assesses interrelationships and signifies the effect of correlation control variables. Response variable is predicted using categorical explanatory variables.

**Bagged Tree Model:**

Bagging is a procedure for reducing the variance of a statistical learning method. This helps in improving the accuracy and stability of ML techniques used in statistical classification and regression. Bagging is a model averaging subset. It builds trees on the bootstrap samples taken from the training dataset and then aggregates the output from all the trees and predicts the output.

**Support Vector Machines:**

Support vector machines is a supervised machine learning algorithm used for both classification and regression. SVM classifies data points based on the hyperplane in an N – dimensional space. The separation function in support vector classification is a linear combination of kernels linked to the support vector.

## Using Machine Learning Models for Predicting the Water Quality Index in the La Buong River, Vietnam:

The paper "Using Machine Learning Models for Predicting the Water Quality Index in the La Buong River, Vietnam" authored by Dao Nguyen Khoi, Nguyen Trong Quan, Do Quang Linh, Pham Thi Thao Nhi and Nguyen Thi Diem Thuy aims to evaluate the performance of twelve machine learning models in estimating the surface water quality

**Models:**

1. **Boosting based algorithms:**

   Boosting algorithm is an ensemble meta-algorithm strategy that seeks to improve the predictive performance of several weaker algorithms by primarily reducing bias and variance in supervised learning problems thereby including a bias-variance tradeoff. The basic idea behind the boosting method starts by creating a model from the training data, and then a second model based on the previous one by reducing the bias error that arises when the first model could not infer the relevant patterns in the given data. Every time a new learning algorithm is added, the weights of data are readjusted, which is called as "re-weighting". These models are added sequentially till the training data is reasonably predicted or the number of learners has been reached an end of adding further. Five types of boosting-based algorithms including adaptive boosting (AdaBoost), gradient boosting (GBM), histogram-based gradient boosting (HGBM), light gradient boosting (LightGBM), and extreme gradient boosting (XGBoost) has been used for the study.

2. **Decision Tree-Based Algorithms:**

   The decision tree and its variants are the other learning algorithms that divide the input space into regions and has separate parameters for each region. They are classified as non-parametric supervised learning method which is widely used in classification and regression, as well as in representing decisions and decision making. The structure of a decision tree is a tree-like flowchart, in which each internal node represents a "test" on an attribute, each branch represents the outcome of the test, and each leaf node represents a class label. Besides, the paths from root to leaf represent classification rules. Three decision tree-based models, including

decision tree (DT), extra trees (ExT), and random forest, were evaluated in relation to various learning methods.

3. **ANN-based algorithms:**

ANN-based models have been used widely nowadays due to its robustness and capability to handle nonlinear data even with its typically structured, single hidden layer, or advanced-structured, multiple hidden layers. ANN includes three layers: input, hidden, and output layers. In case of increasing complexity of the problem, the number of layers will rise and the computational resources will consequently also rise. Here, both the mentioned structures of the ANN-based models were utilized for predicting the water quality index viz., multilayer perceptron (MLP), radial basis function (RBF), deep feed-forward neural network (DFNN), and convolutional neural network (CNN).

**Dataset and features:**

The measure used to determine the water quality is Water Quality Index (WQI). WQI is a feature by which water quality data is summarized for reporting to the public. It is similar to the UV index or an air quality index, which helps us to understand in simple terms, what the quality of drinking water is from a drinking water supply. Eight years of bimonthly WQ data at four WQ monitoring stations alongside the La Buong River was collected for the dataset. It consists of 10 different qualities examined from the water viz., temperature (T), pH, DO, BOD, COD, turbidity (TUR), total suspended solid (TSS), coliform, ammonium ($NH^{4+}$), and phosphate ($PO_4^{3-}$).

WQI can be defined as:

$$WQI = \frac{WQI_{pH}}{100}\left[\frac{1}{5}\sum_{a=1}^{5} WQI_a \times \frac{1}{2}\sum_{b=1}^{2} WQI_b \times WQI_c\right]^{1/3}$$

where,

- $WQI_a$ is the WQI values for DO, BOD, COD, $NH^{4+}$ and $PO_4^{3-}$
- $WQI_b$ is the WQI values for TSS and TUR
- $WQI_c$ is the WQI values for Coliform
- $WQI_{pH}$ is the WQI value for pH

The WQI calculated would vary from 0 to 100. For example, a very low quality (WQI=3.02) and an excellent quality (WQI=98.30).

**Performance Evaluation:**

Two model efficiency statistics has been used for evaluation, namely, root mean square error (RMSE) and coefficient determination ($R^2$). RMSE determines the deviation between the observed and predicted values, and $R^2$ measures the degree of correlation between the observed and predicted data.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(O_i - P_i)^2}{n}} \qquad R^2 = 1 - \frac{\sum_{i=1}^{n}(O_i - P_i)^2}{\sum_{i=1}^{n}(O_i - \overline{O})^2}$$

where n is the total number of predicted values, Oi is the observed value, $\overline{O}$ is the mean of observed values, and $P_i$ is the predicted value.

**REFERENCES:**


- Shafi, U.; Mumtaz, R.; Anwar, H.; Qamar, A.M.; Khurshid, H. Surface Water Pollution Detection using Internet of Things.
- Daud, M.K.; Nafees, M.; Ali, S.; Rizwan, M.; Bajwa, R.A.; Shakoor, M.B.; Arshad, M.U.; Chatha, S.A.S.; Deeba, F.; Murad, W.; et al. Drinking water quality status and contamination in Pakistan. BioMed Res. Int. 2017, 2017, 7908183.
- Vietnam Dao Nguyen Khoi, Nguyen Trong Quan, Do Quang Linh, Pham Thi Thao Nhi and Nguyen Thi Diem Thuy; Using Machine Learning Models for Predicting the Water Quality Index in the La Buong River.
- Md. Mehedi Hassan, Laboni Akter, Md. Mushfiqur Rahman, Sadika Zaman, Khan Md. Hasib, Nusrat Jahan, Raisun Nasa Smrity, Jerin Farhana, M. Raihan, Swarnali Mollick; Efficient Prediction of Water Quality Index (WQI) Using Machine Learning Algorithms