A Machine Learning Based Approach To Analyse And Predict Water Quality

Dr R Pushpalakshmi[1], Dharshini K[2], Divya Dharshini P[3], Harishma K[4]
[1, 2, 3, 4] Dept of Information Technology
[1, 2, 3, 4] PSNA College of Engineering and technology, Dindigul, Tamilnadu

Abstract- *Water is a need for all living things. Water is required for numerous daily activities such as drinking, washing, bathing, and cooking. If the water is of poor quality, it is unsuitable for drinking and other activities. Water quality is often defined by its physical, chemical, and biological properties. As a result, it is required to determine the appropriateness of water for drinking, irrigation, and industrial purposes. The groundwater quality based on Sodium percent, Sodium Absorption Ratio, and Residual Sodium Carbonate will aid in determining water suitability for irrigation. Increasing industrialisation and the use of chemical fertilisers and pesticides in agriculture are causing water quality to deteriorate and aquatic biota to become depleted.Water-borne illnesses affect the human population as a result of the usage of polluted water. Temperature, pH, turbidity, salinity, nitrates, TDS, Cations, Anions, and phosphates are some of the parameters that may be measured.*

Keywords- Water Quality,Machine Learning, Random Forest Algorithm,Water Borne Diseases

I. INTRODUCTION

The major sources of water available to people in general are ground water, surface water (rivers, streams, and ponds), atmospheric water (rain, snow, and hail), and springs. The quality of these bodies of water varies greatly based on their location and environmental conditions. Precipitation infiltrates the ground and flows through the soil and pore spaces of rocks as the primary source of ground water. Water infiltration from lakes and streams, recharge ponds, and waste-water treatment systems are among the other sources. Several contaminants, such as disease-causing microbes, are filtered out of ground water as it passes through soil, silt, and rocks.
Many developing-country water supplies are unsafe because they include dangerous physical, chemical, and biological contaminants.Yet, water needs to be safe to drink and uphold regional and worldwide standards for flavour, odour, and appearance in order to sustain good health. National and international norms and recommendations for water quality standards are being utilised to monitor the water resource and assure sustainability. (WHO1993; 2005).

Water's chemistry is extremely dynamic and is heavily influenced by the medium with which it comes into contact. Since the chemistry of water directly predicts its quality for a variety of uses, monitoring and evaluating it have become increasingly important in the twenty-first century.

The load on both surface and groundwater rose dramatically as a result of the population growth. Because of the aquifer's filtration capabilities, it is thought that groundwater was the most reliable source of drinking water during the dawn of human civilization. Nonetheless, it is difficult to consume water straight from the source in the modern world without sufficient purification. For research on public health, the physical and chemical characteristics of groundwater are highly crucial. These investigations are a key component of research on environmental contamination.

The dissolved particles in groundwater provide it physical qualities including odour, taste, and temperature.The physical environment, the source, and the direction of the water all affect the natural quality of groundwater. By interactions with soil, rock, and organic matter, the hydrological cycle causes the water to undergo a number of chemical, physical, and biological processes that alter its initial characteristics. Changes in groundwater quality are caused by both natural and human processes, either directly or indirectly. About 80% of all human diseases, according to the WHO, are brought on by water.

BACKGROUND:

A. Machine Learning

The rapidly expanding discipline of data science includes machine learning as a key element. Algorithms are taught using statistical techniques to produce classifications or predictions and to find important insights in data mining projects. The decisions made as a result of these insights influence key growth indicators in applications and enterprises, ideally.

Data scientists will be more in demand as big data continues to develop and flourish. They will be expected to

assist in determining the most pertinent business issues and the information needed to address them.

The majority of the time, machine learning algorithms are developed utilising accelerated solution development frameworks like TensorFlow and PyTorch.

B. Supervised Learning

A subset of machine learning and artificial intelligence is supervised learning, commonly referred to as supervised machine learning. It is distinguished by the way it trains computers to properly categorise data or predict outcomes using labelled datasets. The model modifies its weights as input data is fed into it until the model has been properly fitted, which takes place as part of the cross validation process. For as categorising spam in a different folder from your email, supervised learning assists enterprises in finding scalable solutions to a number of real-world issues.

C. Classification Algorithm

On the basis of training data, the Classification algorithm is a Supervised Learning approach that is used to categorise fresh observations. In classification, a programme makes use of the dataset or observations that are provided to learn how to categorise new observations into various classes or groups. For instance, cat or dog, yes or no, 0 or 1, spam or not spam, etc. Targets, labels, or categories can all be used to describe classes.

In contrast to regression, classification's output variable is a category rather than a value, such as "Green or Blue," "fruit or animal," etc. The Classification method uses labelled input data since it is a supervised learning approach, therefore it comprises input and output information.

D.Regression Algorithm

A supervised machine learning approach called regression is used to forecast continuous output values based on input. Simple linear regression, multiple linear regression, and polynomial regression are the three primary categories of regression algorithms. Let's examine each of them using some instances.

- A mapping function is used in simple linear regression to model the linear relationship between an independent variable and a dependant variable that must be predicted. Consider, for instance, that a locality's housing costs are only influenced by its geography. Hence, basing the trained model on historical data, one may forecast home costs given the area of any new region.

- Regression analysis using many independent variables and one dependent variable is known as multiple linear regression. For instance, a restaurant's evaluations are influenced by the calibre of the cuisine as well as the ambiance, service, and location. So, in this situation, several independent variable impact the dependent parameter linearly (rating of a restaurant).

- The non-linear relationship between the independent and dependent variables is mapped using the polynomial regression procedure. The equation is not linear since the mapping consists of many powers of an independent variable. This kind of algorithm, for instance, may be used to forecast the number of Covid-19 cases. Polynomial regression can map the non-linearity and forecast since the growth or reduction in instances is not linearly connected to the number of persons wearing masks.

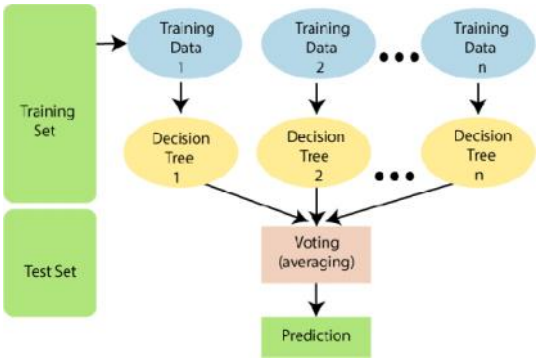
E. Random Forest Algorithm

Popular machine learning algorithm Random Forest is a part of the supervised learning methodology. It may be applied to ML issues involving both classification and regression. It is built on the idea of ensemble learning, which is a method of integrating many classifiers to address difficult issues and enhance model performance.

Random Forest, as the name implies, is a classifier that uses a number of decision trees on different subsets of the provided dataset and averages them to increase the dataset's prediction accuracy. Instead than depending on a single decision tree, the random forest uses forecasts from each tree and predicts the result based on the votes of the majority of predictions.

Higher accuracy and overfitting are prevented by the larger number of trees in the forest.

The Random Forest method is illustrated in the picture below:



Key Benefits:

Less chance of overfitting: Decision trees have a propensity to closely match all the samples contained in training data, which increases the possibility of overfitting. The classifier won't, however, overfit the model when there are a large number of decision trees in a random forest since the averaging of uncorrelated trees reduces the total variance and prediction error.

Allows for flexibility: Data scientists frequently use random forest because it is highly accurate at handling both regression and classification problems. The random forest classifier benefits from feature bagging by maintaining accuracy even when some of the data is missing, which makes it a useful tool for guessing missing values.

Simple evaluation of feature contribution: Random forest makes it simple to assess variable contribution. There are several methods for determining feature relevance. To gauge how much the model's accuracy declines when a particular variable is removed, the gini importance and mean drop in impurity (MDI) are frequently utilised. A different significance metric is permutation importance, often known as mean decrease accuracy (MDA). By randomly permuting the feature values in oob samples, MDA can determine the average reduction in accuracy.

Key Challenges:

Process that takes a long time: Because random forest methods can handle big data sets, they can make predictions that are more accurate. But, because they must compute data for each individual decision tree, they can take a long time to process data.

More resources are needed: Because random forests analyse bigger data sets, more resources are needed to store that data.
More complex: As compared to a forest of decision trees, a single one's prediction is simpler to understand.

F. Python Implementation of Random Forest Algorithm:

Now let's implement Random Forest in scikit-learn.

1. Import the libraries.
2. Import the dataset.
3. Define the features and the target.
4. Split the dataset into train and test sets.
5. Build the random forest regression model with random forest regressor function.
6. Evaluate the random forest regression model.

G.Google Colaboratory

A free Jupyter notebook called Google Colab enables the use of Python in the browser without the need for intricate settings. It already has Python installed, along with all of the essential Python libraries. Moreover, free GPUs are embedded into it. The quickest method to start using Python on any computer is through Google Colab.

The pip command and the exclamation point (!) make it simple to install a Python package in Google Colab. A shell command may be executed by using the exclamation point at the beginning of a cell, and pip is the Python package installer that enables the installation of Python libraries.

Through Google Colab, you may link to Google Drive to access previously stored files or even to store the output of your scripts. You must first mount your disc in order to use the files from Google Drive in Google Colab. You will be prompted via an overlay to grant the laptop access to Google Drive files. To grant access to your Google Drive, click "Connect to Google Drive" and then follow the on-screen instructions.

Your files will be listed in the following directory: "content > drive >MyDrive".

For accelerating the training of your machine learning models, Google Colab offers complimentary GPUs. Machine learning benefits from the usage of GPUs, or graphics processing units. While constructing complex machine learning models, they enable many simultaneous processing of calculations.

To enable free GPUs in Colab, go to:

Runtime > Change Runtime Type and select the right Hardware accelerator.

H. Libraries

NumPy

Nearly all branches of research and engineering utilise the free source Python library known as NumPy (Numerical Python). It is the de facto standard for manipulating numerical data in Python and forms the basis of both the PyData and scientific Python ecosystems. Users of NumPy range from novice programmers to seasoned academics working on cutting-edge academic and commercial research and development. The majority of other Python data science and scientific programmes, including Pandas, SciPy, Matplotlib, scikit-learn, and scikit-image, make substantial use of the NumPy API.

Sklearn

Scikit-learn is mostly written in Python and significantly makes use of the NumPy module for computations involving arrays and linear algebra. To further the effectiveness of this library, certain basic algorithms are also written in Cython. Using wrappers created in Cython for LIBSVM and LIBLINEAR, support vector machines, logistic regression, and linear SVMs are done. In certain conditions, expanding these functions with Python might not be practical. Other additional Python tools, including as SciPy, Pandas data frames, NumPy for array vectorization, Matplotlib, seaborn, and plotly for showing graphs, as well as many more, work well with Scikit-learn.

Matplotlib

For Python and its numerical extension NumPy, Matplotlib is a cross-platform data visualisation and graphical charting package. As a result, it presents a strong open source substitute for MATLAB. The APIs (Application Programming Interfaces) for matplotlib allow programmers to include graphs into GUI applications.

The way a Python matplotlib script is written makes it possible to create a visual data plot in the majority of cases with just a few lines of code. Two APIs are covered by the Matplotlib scripting layer:

- A hierarchy of Python code objects with matplotlib at its top makes up the Pyplot API.
- A set of OO (Object-Oriented) API items that can be put together more easily than Pyplot. Direct access to Matplotlib's backend layers is made possible using this API.

II. LITERATURE REVIEW

1.Performance analysis of machine learning algorithms for water quality monitoring system

Author:Dziri Jalal and Tahar Ezzedine
Year: 2019

Problem Identified and Objective

Every local government has a dilemma with the availability of clean drinking water due to the rapidly expanding urbanisation.

Any moment can lead to water pollution. So, the water we store in the water tank on our apartment's roof or in the basement of our building may not be secure. In India, the majority of people still utilise basic water purifiers, which are insufficient to guarantee the purity of the water. Occasionally the water contains harmful contaminants or chemicals that general-purpose water purifiers cannot remove. Also, it is impossible to physically examine the water's purity every time. In order to keep track of the condition of the water stored in our water tank for the community or flat, an autonomous real-time monitoring system is necessary.

In this study, we trained various machine learning algorithms using a real dataset. In this study, we examined the precision and accuracy of several methods. A difficult machine learning problem is correctly categorising the quality of drinking water based on its physicochemical and microbiological parameters. The current contribution belongs to this lineage.

Methodology

This is a real-time Internet of Things-based system for monitoring water quality. The proposed system provides remote monitoring of water quality assessment and water flow management via a mobile application. Four machine learning algorithms—Support Vector Machine (SVM), KNearestNeighbor (KNN), single-layer neural network, and deep neural network—have been employed to categorise water quality. Similar to the last case study, this one describes how to monitor water contamination using SVM based on Color Layout Descriptor (CLD) and Fast Fourier Transform (FFT). The trained SVM passed a successful preliminary test using the FFT and CLD combination.

Findings

The performance of Decision Trees and SVMs, two well-known classification techniques, was examined in this study. We have strong proof of the efficiency of the classification methods thanks to the test results. Also, we found that linear SVM seems to work well with our system for monitoring water quality. Our method has a number of drawbacks; in reality, the monitoring system for water quality is reactive. It is unable to anticipate interferences that can lower the quality of the water. We are seeking to add a new data aggregation methodology for further study in order to decrease the amount of data obtained to run the SVM classification algorithm.

2. Predicting and Analysing Water Quality using Machine Learning: A Comprehensive Model

Author: Yafra Khan and Chai Soo See
Year: 2016

Problem Identified and Objective

The main goal of this research is to develop a thorough approach that uses certain water quality metrics to examine and forecast the water quality of specific places. These characteristics encompass aspects of water quality that are physical, biological, or chemical in nature. The World Health Organization (WHO) and the Environmental Protection Agency (EPA) have established a number of quality standards that serve as a guideline for judging the calibre of water. The Environmental Protection Agency (EPA) lists a total of 101 factors in its publication "Parameters of Water Quality" that in some way affect water quality. Yet certain factors have a more significant and obvious impact on water quality than others.

By proposing a model based on machine learning techniques to forecast the future trends in water quality of a specific area with the use of existing water quality data, this research seeks to solve this issue. To provide a thorough technique for effective water quality prediction and analysis, Artificial Neural Networks (ANN) with Nonlinear Autoregressive (NAR) time series model is employed. In this investigation, four specific water quality parameters—chlorophyll, specific conductance, dissolved oxygen, and turbidity—were employed. The purpose of this study is to create effective models that can forecast water quality parameter values based on their current values.

Methodology

The methodologies include statistical methods, visual modelling, algorithms for analysis and prediction, and decision-making. Principal Component Analysis (PCA), a multivariate statistical approach, has been used to establish relationships between several water quality measures. Kriging, transitional probability, multivariate interpolation, regression analysis, and other geo-statistical methods have all been employed. Artificial intelligence (AI) methods including Bayesian networks (BN), artificial neural networks (ANN), neurofuzzy inference, support vector regression (SVR), decision support systems (DSS), and auto-regressive moving average may be used in the algorithms for analysis and prediction (ARMA). Yet, mapping input-output data and forecasting future water quality are highly difficult tasks due to the non-linear structure of water quality data, as in this study.

Findings

In order to evaluate the concentration of chlorophyll, dissolved oxygen, turbidity, and specific conductance, this study examines and predicts the values of water quality indicators. The findings are then analysed. Time series data from the year 2014 were obtained from the USGS National Water Information System (NWIS).

A channel located in the State of New York serves as the designated monitoring station. For easier data management, the measurements of the water quality parameters were scaled between 0 and 1. Scaled Conjugate Gradient (SCG) has been utilised in conjunction with Artificial Neural Network (ANN) with Nonlinear Autoregressive (NAR) time series as the training technique.

3. Predictive Models for River Water Quality using Machine Learning and Big Data Techniques – A Survey

Author: Jitha P Nair and Vijaya M S
Year: 2021

Problem Identified and Objective

Even with the use of machine learning techniques and big data models, it is challenging to anticipate and evaluate the quality of moving water in comparison to still water.

The major goals of this work are to analyse the performance of various prediction models, to offer a brief overview of time series analysis, machine learning, and deep learning approaches in water quality prediction and assessment, and to identify potential research challenges and difficulties.

Methodology

The Water Quality Index, a standardised measure of water quality, is created by calculating the quality of water using several characteristics (WQI). The usual number of water quality parameters is 14, however WQI can also be determined with less parameters. pH, coliform, temperature, total dissolved solids, turbidity, biological oxygen demand, alkalinity, and other variables are some of those used to calculate WQI. Nitrites' relative relevance can be used as a weight for determining WQI. To determine the WQI of each sample, the parameters are given weightages.

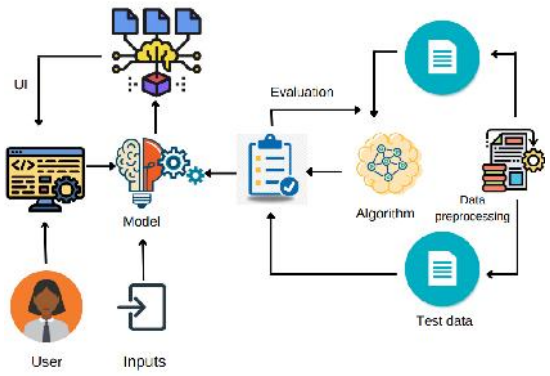
Findings

Manual data collection and standard methods-based data analysis were used to evaluate the water quality. The use of GRA and its straightforward procedure, which has the finest operability and physical importance, produced the best results. Using a mathematical model with a known input and an accurate predictor of the output, the quality of river water was predicted without consideration of chemical or physical parameters.

III. PROPOSED SYSTEM

Water is seen as a crucial resource that has an impact on many elements of human health and existence. Those who live in metropolitan areas are often concerned about the quality of the water. The cornerstone for the prevention and management of waterborne infections is the quality of the water, which is a significant environmental influence.

As a result, this project aims to develop a Machine Learning (ML) model to Predict Water Quality by taking into account all water quality standard indicators. However, this is a difficult task because the water quality varies nonlinearly in urban spaces and depends on numerous factors, such as meteorology, water usage patterns, and land uses.



To overcome this problem, we propose the solution using Random Forest Algorithm. First of all using google colab the dataset, which we are going to use for training and testing will be uploaded with the help of google drive. The required libraries for the process are imported. Null values in the dataset were handled. The Water Quality Indicator(WQI) is estimated based on the features which arepresent in the database. Using sklearn library the data is splitted into training set and testing set based on test_size and random_state . Now the random forest algorithm is imported using RandomForestRegressor from sklearn.ensemble . Based on the collected data, the model is being evaluated. The variables like MAE (Mean Absolute Error), MSE (Mean Squared Error), RMSE (Root Mean Squared Error) are calculated. Using these variables the accuracy of the model is estimated.

By comparing the accuracy rate between the previous and current data sets, the answer is deduced from the data sets. It is possible to ascertain whether the water can be recycled or used again. Eco-friendly and kind to users, positively affecting human health. Good and healthful water is given by analysing the water's quality. The programme is first tested on a small group of users. Afterwards, it enters the image so that everyone may view through networking. By engaging in various activities that highlight the significance of water qualityaids in obtaining water in its entirety.

Parameters

Temperature

One of water's most fundamental characteristics is temperature, and many other metrics rely on it to be accurate. By tracking temperature variations in the thermocline, which have an impact on the wellbeing of aquatic animals and creatures, we can monitor thermal loading or discharge. The high temperatures can be harmful to many aquatic creatures. Warmer water limits oxygen delivery because oxygen is less soluble in it.

Dissolved Oxygen

Gaseous, molecular oxygen (DO) is an end product of photosynthesis or comes from the atmosphere in the form of O2. After it has been dissolved in water, it may be utilised by living things and can have a big impact on a lot of chemical reactions that take place in the aquatic environment. This oxygen is identical to the oxygen we breathe except from being dissolved in water.

pH

In almost every application involving water quality, pH measurement is a crucial component. Several treatment procedures for wastewater are pH-dependent, and pH regulation in wastewater treatment is a requirement for discharge permits. High or low pH readings in environmental sampling and monitoring might be a sign of contamination.

Conductivity

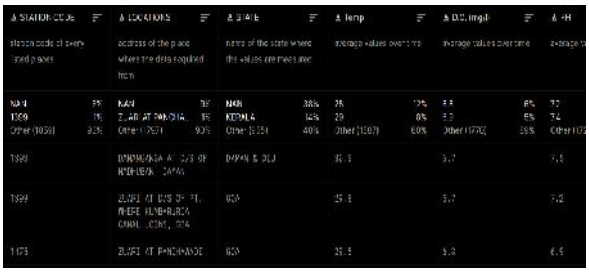
A measure of water quality is electrical conductivity. Data on conductivity may be used to assess water cleanliness, identify pollutants, and assess solution concentration. YSI conductivity sensors use nickel electrodes and an AC voltage to detect conductivity. The current passes between the electrodes and the sample when these electrodes are submerged in a water sample (or other liquid). The link between the conductivity of the solution and the current level is direct.

BOD

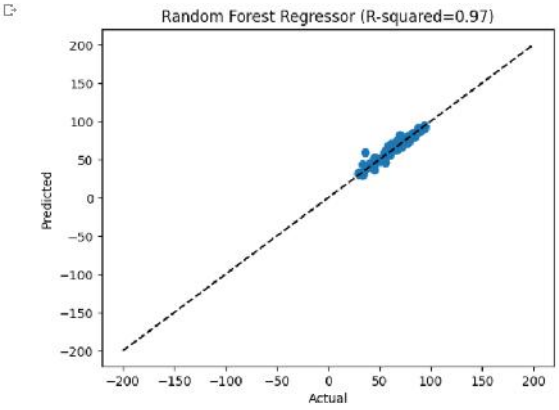
BOD, or biochemical oxygen demand, is a chemical method for calculating how much dissolved oxygen aerobic biological organisms in a body of water require to decompose organic material present in a given water sample at a particular temperature during a certain time period.

IV. EXPERIMENTAL RESULTS AND DISCUSSION

This model is implemented to check the water quality using the dataset holding parameters like temperature, pH, conductivity, dissolved oxygen (DO), biochemical oxygen demand (BOD), Nitratenan, Total Coliform, Fecal Coliform.The Accuracy of the model is **96.971918125809**.



Random Forest Algorithm is being used to calculate the accuracy of the model.



V. CONCLUSION

The water quality crisis is an emerging, natural problem in our country. Usage of water and the quality of it has raised and decreased respectively. Considering the basic crisis of water we have proposed and trained the model which analyses the quality control of water using **Random Forest Algorithm** in Machine Learning. The result of our model using various parameters is at accuracy **96.97**.

REFERENCES

[1] C.Sadashivaia1,C.R.Ramakrishnaiah and G. Ranganna, “Hydrochemical Analysis and Evaluation of Groundwater Quality in TumkurTaluk, Karnataka State, India, International Journal ofEnvironmental Research and Public Health, 2008, 5(3) 158-164.

[2] Adetunde L.A, Glover R.L.K &Oguntola G.O, “assessment of the ground water quality in ogbomoso township of oyo state of nigeria”, IJRRAS8 (1) july 2011, 115-122.

[3] Shimaa M. Ghoraba&A.D.Khan, “ hydrochemistry and groundwater quality assessment in balochistan province, Pakistan”, IJRRAS 17 (2) November 2013, 185-199

[4] M. R. G. Sayyed1, G. S. Wagh2, A. Supekar3, “Assessment of impact on the groundwater quality due to urbanization by hydrogeochemical facies analysis in SE part of Pune city, India”, Proceedings of the International Academy of Ecology and Environmental Sciences, 2013, 3(2): 148-15.

[5] Dinesh kumar tank and c. p. Singh chandel, “Analysis of the major ion constituents in groundwater of Jaipur city”, Nature and Science, 2010;8(10), 1-7

[6] VikasTomar, Kamra S.K, Kumar S, Kumar Ajay and Vishal Khajuria, “Hydro-chemical analysis and evaluation of groundwater quality for irrigation in Karnal district of Haryana state, India”, International Journal of Environmental Sciences,
Volume 3, No 2, 2012, pp.756-766.

[7] Mona A. Hagras Assistant Professor, Irrigation &Hydraulics Department, Faculty of Engineering, Ain Shams University, Cairo, Egypt
[8] S. Prabaharan, R. Manonmani, M. Ramalingam and T. Subramani, "Groundwater Contamination due to Municipal Solid Waste Disposal in Salem City using GIS", International Journal of Earth Science and Engineering, Volume 05, No. 04, pp- 696-702.
[9] ManeT.T. and HinganeHemalataN."Existing Situation of Solid Waste Management in Pune City, India", Research Journal of Recent Sciences , Vol. 1, 2012, pp.348-351
[10] S.S. Castaneda , R.J. Sucgang , R.V. Almoneda , N.D.S. Mendoza and C.P.C. David, "Environmental isotopes and major ions for tracing leachate contamination from a municipal landfill in Metro Manila, Philippines", Journal of Environmental Radioactivity 110 (2012), pp.30-37.
[11] GunjanBhalla, Swamee, P.K, Arvind Kumar, Ajay Bansal , "Assessment of groundwater quality near municipal solid waste landfill byan Aggregate Index Method", International Journal of Environmental Sciences Volume 2, No 2, 2012, pp. 1492- 1503.
[12] P.I. Agber, A. Ali and N. A. Tsaku, "Assessment of Ground Water Quality, Soil Propertiesand Nutrient Content of Soil in Areas Close to Municipal Refuse Dump Sites in Makurdi, Nigeria", J. Biol. Chem. Research. Vol. 30, No. 1, 2013, pp.88-97.