JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue

JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)

An International Scholarly Open Access, Peer-reviewed, Refereed Journal

MACHINE LEARNING BASED WATER QUALITY PREDICTION MODEL

[1] **Ms.V.S.Esther Pushpam, [2]P.Nandha Praveen Kumar, [3]Pooja Eshwaramurthy, [4] N.Sowmiya**

[1]Assistant Professor, [2]Student, [3]Student, [4]Student

[1,2,3,4]Computer Science and Engineering,

[1,2,3,4]Sri Ramakrishna Institute of Technology, Coimbatore, India

Abstract: Now a days many people are suffering from dangerous diseases which are caused due to impure water. In our project we are doing analysis for quality of water monitoring system, it gives data about the quality of water. We are about the water quality prediction using the machine learning algorithm. The deteriorating quality of natural water resources like lakes and streams , is one of the direst and most worrisome issues faced by humanity. The effects of un-clean water are far-reaching, impacting every aspect of life. Therefore, management of water resources is very crucial in order to optimize the quality of water. The effects of water contamination can be tackled efficiently if data is analyzed and water quality is predicted beforehand. This issue has been addressed in many previous researches, however, more work needs to be done in terms of effectiveness, reliability, accuracy as well as usability of the current water quality management methodologies. The goal of this study is to develop a water quality prediction model with the help of water quality factors using Artificial Neural Network (ANN) and time-series analysis. This research uses the water quality historical data of the year of 2014, with 6-minutes time interval. Data is obtained from the kaggle online resource called National Water Information System (NWIS). For this paper, the data includes the measurements of parameters which affect and influence water quality. For the purpose of evaluating the performance of model, the performance evaluation measures used are Mean-Squared Error (MSE), Root Mean-Squared Error (RMSE) and Regression Analysis. Previous works about Water Quality prediction have also been analyzed and future improvements have been proposed in this paper**.**

Keywords **- Parameter, Accuracy, Heatmap Generation, Time series analysis**

I .INTRODUCTION

Water is the most significant resource of life, crucial for supporting the life of most existing creatures and human beings. Living organisms need water with enough quality to continue their lives. There are certain limits of pollutions that water species can tolerate. Exceeding these limits affects the existence of these creatures and threatens their lives. Most ambient water bodies such as rivers, lakes, and streams have specific quality standards that indicate their quality. Moreover, water specifications for other applications/usages possess their standards. For example, irrigation water must be neither too saline nor contain toxic materials that can be transferred to plants or soil and thus destroying the ecosystems. Water quality for industrial uses also requires different properties based on the specific industrial processes. Some of the low-priced resources of fresh water, such as ground and surface water, are natural water resources. However, such resources can be polluted by human/industrial activities and other natural processes. Hence, rapid industrial development has prompted the decay of water quality at a disturbing rate. Furthermore, infrastructures, with the absence of public awareness, and less hygienic qualities, significantly affect the quality of drinking water . In fact, the consequences of polluted drinking water are so dangerous and can badly affect health, the environment, and infrastructures. As per the World report, about 1.5 million people die each year because of contaminated water-driven diseases. In developing countries, it is announced that 80% of health problems are caused by contaminated water. Five million deaths and 2.5 billion illnesses are reported annually .. It is recommended to consider the temporal dimension for forecasting the Water Quality (WQ) patterns to ensure the monitoring of the seasonal change of the WQ. However, using a special variation of models together to predict the WQ grants better results than using a single model . There are several methodologies proposed for the prediction and modeling of the WQ. These methodologies include statistical approaches, visual modeling, analyzing algorithms, and predictive algorithms. For the sake of the determination of the correlation and relationship among different water quality parameters, multivariate statistical techniques have been employed. The geostatistical approaches were used for transitional probability, multivariate interpolation, and regression analysis.

1.1.Machine learning

Machine learning is important because it gives enterprises a view of trends in customer behavior and business operational patterns, as well as supports the development of new products. Many of today's leading companies, such as Facebook, Google and

Uber, make machine learning a central part of their operations. Machine learning has become a significant competitive differentiable for many companies. Classical machine learning is often categorized by how an algorithm learns to become more accurate in its predictions. There are four basic approaches: supervised learning, unsupervised learning, semi-supervised learning and reinforcement learning.

The type of algorithm data scientists choose to use depends on what type of data they want to predict.

- Development of new machine learning algorithms that learn more accurately, utilize data from dramatically more diverse data sources available over the Internet and intranets, and incorporate more human input as they work;
- Integration of these algorithms into standard database management systems; and
- An increasing awareness of data mining technology within many organizations and An attendant increase in efforts to capture, warehouse, and utilize historical data to support evidence-based decision making.

In this type of machine learning, data scientists supply algorithms with labelled training data and define the variables they want the algorithm to assess for correlations. Both the input and the output of the algorithm is specified.

Unsupervised learning: This type of machine learning involves algorithms that train on unlabelled data. The algorithm scans through data sets looking for any 11 meaningful connections. The data that algorithms train on as well as the predictions or recommendations they output are predetermined.

Semi-supervised learning: This approach to machine learning involves a mix of the two preceding types. Data scientists may feed an algorithm mostly labeled training data, but the model is free to explore the data on its own and develop its own understanding of the data set.

Reinforcement learning: Data scientists typically use reinforcement learning to teach a machine to complete a multi-step process for which there are clearly defined rules. Data scientists program an algorithm to complete a task and give it positive or negative cues as it works out how to complete a task. But for the most part, the algorithm decides on its own what steps to take along the way

1.2.PROBLEM STATEMENT

Water is a free energy source however it is highly unpredictable which is significant problem for prediction of water Quality. Water data need to be collected is a time series manner to process the data with its accuracy. Data need to cleaned and explored to identify its insights because the water characteristic varies based on water quality parameters. To make a prediction it is necessary to identifying appropriate machine learning algorithm which can be obtained based on literature survey. With a collected data of Kaggle, the model need to build and trained well to achieve greater accuracy.

1.3.Problem Objective

The main objective of the project is to predict the water quality using machine learning algorithms with the help of input parameters like; PH, Solids, Hardness, Chloramine, sulphate, conductivity, Organic Carbon, Trihalomethane, Turbidity. The Machine learning algorithm we are going to use in this project are Logistic Regression, Decision Tree, Random Forest, XG Boost, KNN, SVM, Ada Boost by calculating MSE and RMSE to find the which algorithm has greater efficiency

1.4.Applications

water quality has been threatened by various pollutants. Therefore, modelling and predicting water quality have become very important in controlling water pollution. Water quality has a direct impact on public health and the environment. Water is used for various practices, such as drinking, agriculture, and industry. A first step in setting water quality standards for a water body is to identify the intended uses of that water body, whether a lake, a section of a stream, or areas of an estuary. The most restrictive of the specific desired uses of a water body is termed a designated use. Barriers to achieving the designated use are the presence of pollutants or hydrologic and geomorphic changes that impact the quality of the water body. The designated use dictates the appropriate type of water quality standard. For example, the standards of a water body whose designated use involves human contact recreation should protect humans from exposure to microbial pathogens while swimming, wading, or boating. Other uses might require standards to protect humans and aquatic life including fish, shellfish, and other wildlife from consuming harmful substances.

1.5.Scope Of The Project

Water quality prediction is used to alert to current, ongoing and emerging problems to determine with drinking water standards and to protect other beneficial uses of water. To developing a model by using machine learning, we can be used to predict the feature water quality.

1.6.Existing System

The quality of water has a direct influence on both human health and the environment. Water is utilized for a variety of purposes, including drinking, agriculture, and industrial use. The water quality index (WQI) is a critical indication for proper water management. Water quality is dictated by features such as dissolved oxygen (DO), total coliform (TC), biological oxygen demand (BOD), Nitrate, pH, and electric conductivity (EC). These features are handled in five steps: data pre-processing using min-max normalization and missing data management using RF, feature correlation, applied machine learning classification, and model's feature importance. The highest accuracy Kappa, Accuracy Lower, and Accuracy Upper findings. In this case, the model stacking technique was applied to three different beaches around and compared to all five basis models. After analysis, the model stacking strategy performed better than all of the base models. Year-over-year, stacking model accuracy scores were constantly at or near the top of the rankings, with a year-on-year accuracy average of 78%, 81%, and 82.3% at the three tested beaches. Developed a machine learning-based approach integrating attribute-realization (AR) and support vector machine (SVM) algorithm to classify the Chao Phraya River's water quality. The AR has determined the most significant factors to improve the river's quality using the linear function. Here, the optimal input combinations have differed across algorithms but the variables with poor correlations have performed worse. The Hybrid algorithms have improved their prediction power of several of the standalone models. An ensemble learning approach for regression, classification, and other complications that works by training a large number of DT. It builds decision trees from samples and uses majority voting for classification and regression. Because random forests work with subsets of data, they are faster than decision trees.

1.6.1.Limitation

Modelling and prediction of water quality are very important for the protection of the environment.

Developing a model by using advanced artificial intelligence algorithms can be used to measure the future water quality.

1.7.Proposed System- change>..

The proposed algorithm uses the particular attributes such as the logistic regression , decision tree classification , XG boost classifier , k –Nearest Neighbors and SVM. The classification of the algorithm uses the attributes such as the PH , HARDNESS, SOLIDS , CHORAMINIS, etc.,

1.7.1Advantages

- It is simple to implement.
- It is robust to the noisy training data
- It can be more effective if the training data is large.

1.8.Environmental Modelling

Environmental modelling is the creation and use of mathematical models of the environment. Environmental modelling may be used purely for research purposes, and improved understanding of environmental systems, or for providing an interdisciplinary analysis that can inform decision making and policy.

1.8.1TIME-SERIES ANALYSIS

In mathematics, a time series is a series of data points indexed (or listed or graphed) in time order. Most commonly, a time series is a sequence taken at successive equally spaced points in time. Thus it is a sequence of discrete-time data. Examples of time series are heights of ocean tides, counts of sunspots, and the daily closing value of the Dow Jones Industrial Average.

A Time series is very frequently plotted via a run chart (which is a temporal line chart). Time series are used in statistics, signal processing, pattern recognition, econometrics, mathematical finance, weather forecasting, earthquake prediction, electroencephalography, control engineering, astronomy, communications engineering, and largely in any domain of applied science and engineering which involves temporal measurements.

Time series analysis comprises methods for analysing time series data in order to extract meaningful statistics and other characteristics of the data. Time series forecasting is the use of a model to predict future values based on previously observed values. While regression analysis is often employed in such a way as to test relationships between one or more different time series, this type of analysis is not usually called "time series analysis", which refers in particular to relationships between different points in time within a single series. Interrupted time series analysis is used to detect changes in the evolution of a time series from before to after some intervention which may affect the underlying variable.

II. LITERATURE SURVEY

[1] Water Quality Prediction Using Artificial Intelligence Algorithms by Theyazn H, H Aldhyani, Mohammed AI-Yaari, Hasan Alkahtani, and Mashael Maashi, this paper describes the modelling and predicting water quality. As now in our advance technology the Water Quality Index (WQI) and Water Quality Classification (WQC) algorithm has been used in this paper. The deep learning algorithms like Nonlinear Autoregressive Neural Network (NARNET) and Long Short-Term Memory (LSTM) learning algorithms used. This also includes machine learning algorithms such as SVM, KNN, and Naïve Bayes are used to classify the WQI. The dataset has 7 models to predict the water quality according to superior robustness. This technique had achieved similar accuracy for testing phase with a slight difference in the regression coefficient. The future work can be developed models that will implement to predict the water quality for different types of water.

[2] A COMPARATIVE STUDY OF AUTOREGRESSIVE NEURAL NETWORK HYBRIDS Tugba Taskaya-Temizel et.al., has proposed. In this paper Many researchers have argued that combining many models for forecasting gives better estimates than single time series models. For example, a hybrid architecture comprising an autoregressive integrated moving average model (ARIMA) and a neural network is a well-known technique that has recently been shown to give better forecasts by taking advantage of each model’s capabilities. However, this assumption carries the danger of underestimating the relationship between the model’s linear and non-linear components, particularly by assuming that individual forecasting techniques are appropriate, say, for modeling the residuals. In this paper, we show that such combinations do not necessarily outperform individual forecasts. On the contrary, we show that the combined forecast can underperform significantly compared to its constituents' performances. We demonstrate this using nine data sets, autoregressive linear and time-delay neural network models. If the cyclic patterns are not of direct interest, one can remove them by seasonal differencing conditional on the stochastic variation present in the data. Trend and seasonality removal processes are referred as pre-whitening methods. If the cyclic patterns are of interest, one can apply seasonal models. In the case of cycles that are symmetric, linear AR model variants can be employed, whereas a time series that exhibits multiplicative seasonality can be transformed into additive form using functional transformations such as logarithms. Cyclic patterns are oscillations that generally have a fixed period. Seasonality is regarded as a special case of cycles whose periods are calendar fixed. In economic data, there is increasing evidence that business cycles are not symmetric (Chatfield, 2004). Asymmetric cyclic behaviors in the economy can be explained as the rate of change in recession, being different to the rate of change in emerging from recession. Well-known data sets such as the sunspot and Canadian lynx series (Rao & Sabr, 1984) show evidence of asymmetric cycles, with such behavior difficult to model with linear techniques. percentage performance improvement for the mean and best fit of the TDNN, best fit of the AR neural network hybrid and AR single models as compared to the mean of hybrid architecture. For four out of the nine data sets, the mean hybrid outperforms the single model. However, for five of the data sets, either the linear AR or TDNN model outperforms the hybrid. Of these improved single models, three significantly outperform the hybrid. These improvements appear to be related to model configuration, where selection for generalization performance allows for better results.

[3] TIME SERIES FORECASTING USING HYBRID ARIMA AND ANN MODELS BASED ON DWT DECOMPOSITION Ina Khandelwa et.al., has proposed. In this paper Recently Discrete Wavelet Transform (DWT) has led to a tremendous surge in many domains of science and engineering. In this study, we present the advantage of DWT to improve time series forecasting precision. This article suggests a novel technique of forecasting by segregating a time series dataset into linear and nonlinear components through DWT. At first, DWT is used to decompose the in-sample training dataset of the time series into linear (detailed) and non-linear (approximate) parts. Then, the Autoregressive Integrated Moving Average (ARIMA) and Artificial Neural Network (ANN) models are used to separately recognize and predict the reconstructed detailed and approximate components, respectively. In this manner, the proposed approach tactically utilizes the unique strengths of DWT, ARIMA, and ANN to improve the forecasting accuracy. Our hybrid method is tested on four real-world time series and its forecasting results are compared with those of ARIMA, ANN, and Zhang's hybrid models. Results clearly show that the proposed method achieves best forecasting accuracies for each series.Achieving reasonably accurate forecasts of a time series is a very important yet challenging task. ARIMA and ANN are two widely popular and effective forecasting models. ARIMA assumes linear data generations function, whereas ANN is most suitable for nonlinearly generated time series. But, it is almost impossible to establish the exact nature of a series and a real-world time series most often contains both linear as well as nonlinear correlation structures. As such, in this paper, we have proposed a hybrid forecasting method that applies ARIMA and ANN separately to model linear and nonlinear components, respectively after a prior decomposition of the series into low and high frequency signals through DWT. The final combined forecasts are obtained as the averages of the forecasts through harr, db2, and db4 wavelets. The empirical results with four real-world time series clearly demonstrate that the proposed method has yielded notably better forecasts than ARIMA, ANN, and Zhang's hybrid model

[4] Ground Water Quality Prediction using Machine Learning Algorithms in R by S.Vijay and Dr.K.Kamaraj , the paper describes the bore wells from which the samples were collected are extensively used for drinking purpose. The water quality parameters such as PH, TDS, EC, Chloride, Sulphate, Nitrate, Carbonate, Bicarbonate, metal ions, trace elements have been estimated. There are two major classifications like High , Low level of water contamination observed in Vellore district. This paper focus on predicting water quality by using Machine Learning classifier algorithm C5.0, Naïve Bayes and Random forest as leaner for water quality prediction with high accuracy and efficiency.

[5] APPLICATION WASP MODEL ON VALIDATION OF RESERVOIR-DRINKING WATER SOURCE PROTECTION AREAS DELINEATION Jianping Huang et.al., has proposed. In this paper Applied the OTOXI module and EUTRO module of WASP7.2 model to carrying out forecast analysis on attenuation condition of the main water quality control factors within secondlevel protection area which delineated by experience value, took the reality water volume of 2007 and forecast water volume of 2010, 2020 as the validating condition, demonstrated the rationality of the second-level protection area which delineated by experience value method. The results indicated that the second-level protection area could satisfy industry and life water request in 2007 and 2010, and could also satisfy the request of living in 2020, the protection area delineation has been proved reasonable. Further obtained the allowable maximum taking water volume of the protection area is 145 million m3 /a through the computation, it is suggested that the industry intake moves when water volume which taken from the first-level and secondlevel protection areas is bigger than the maximum volume, to ensuring the quality and stability of the intake drinking water. According to validating study, the results indicated that the second-level protection area which delineated by experience value could satisfy industry and life water request in 2005 and 2010, and could also satisfy the

request of living in 2020, the protection area delineation has been proved reasonable; The allowable maximum taking water volume of the second-level protection area is 145 million m3 /a, it is suggested that the industry intake moves when water volume which taken from the first-level and second-level protection area is bigger than the maximum volume, to ensuring the quality and stability of the intake drinking water; The use of WASP7.2 model in simulation of water quality should base on a lot of parameters. The following work should enhance the monitoring work of reservoir and provide available data for the model, making the model more suitable to actual situation

[6] Water quality prediction and classification based on principal component regression and gradient boosting classifier approach by Author Md. Saikat Islam Khan, Nazrul Islam, Jia Uddin sifatul Islam, Mostofa Kamal Nasir this paper describes a water quality prediction model utilizing the principal component regression technique. Firstly, the water quality index (WQI) is calculated using the weighted arithmetic index method. Secondly, the principal component analysis (PCA) is applied to the dataset, and the most dominant WQI parameters have been extracted. Thirdly, to predict the WQI, different regression algorithms are used to the PCA output. Finally, the Gradient Boosting Classifier is utilized to classify the water quality status. The proposed system is experimentally evaluated on a Gulshan Lake-related dataset. The results demonstrate 95% prediction accuracy for the principal component regression method and 100% classification accuracy for the Gradient Boosting Classifier method, which show credible performance compared with the state-of-art models.

[7] EVALUATION OF THE SUITABILITY OF SURFACE WATER FROM RIYADH MAINSTREAM SAUDI ARABIA FOR A VARIETY OF USES A. A. Al-Othman et.al has proposed **,** An evaluation of the suitability of surface water for different uses from the mainstream running through Riyadh, Saudi Arabia, was conducted using three indices namely: water quality index (WQI), percentage of sodium (%Na) and Nemerow’s pollution index (NPI). The water samples were collected on monthly basis from June 2009 to March 2010 from 10 sites. The WQI was calculated based on different water standards for different uses. Excellent water has WQI value of 50–100, where WQI is calculated based on 29 parameters. The permissible range of %Na in water used for irrigation purpose should be between 40% and 60%. On the other hand, if $0 < NPI < 1.0$, then the water is regarded as being in a good condition. In this study, WQI ranged from 34 to 513 with an average of 282, thereby indicating mild pollution at some sites. The %Na value ranged between 19 and 66, with an average value of 46 indicating that the water is suitable for irrigation at most of the sites. The NPI ranged between 1 and 11 with an average value of 5, which indicates that water in most of the sites is in a good condition. It is recommended that surface water from these sites should not be used for human activities. The formulas used to calculate the water quality indices are easy to use thus providing a valuable tool for the accurate monitoring of water pollution. [a] 2015 Production and hosting by Elsevier B.V. on behalf of King Saud University**.** The surface water quality in a region largely depends on the nature and extent of the industrial, agricultural and other anthropogenic activities which exist within the catchments (Banejad and Olyaie, 2011). The control of water quality has become very important in maintaining the sustainability of water resources. However, the main cause of water pollution is human activities (Ashraf et al., 2010). Two sources of water pollution can be recognized, namely point sources and nonpoint sources. Point sources discharge pollutants at specific locations (e.g. landfills and industrial wastes) through inlets into surface water, while non-point sources (e.g. acid rain, agriculture, construction and domestic pollutants) cannot be traced to a single discharge for example, fish survival, agricultural usage, including irrigation and livestock watering are affected by the physical, chemical, biological, and microbiological conditions existing in a watercourses . Industrial effluents as well as domestic sewage/wastes are disposed into it, either with partial or no pretreatment and hence, increasing concentration of different kinds of pollutants reaches this watercourse. The watershed area of the stream is estimated to be about 4400 km2 and the mainstream flood-channel is located slightly east of the centre of the catchment area and flows northwest to southeast. The source of the water in the main channel of the stream is seasonal rainfall. Water samples from the specified locations were collected during June, July, August, September, October, November and December of the year 2009 and during January, February and March of the year 2010. The water samples were taken according to standard method (APHA, 1992.). Samples were collected from the number of distribution sites and selection of the sample was performed depending on stream characteristics, study objectives, availability of equipment, and few other required factors. The integrated sample was taken from top to bottom in the middle of the main channel of the stream and from side to side at mid-depth, preferably they were taken at various points of equal distance across the stream. After that, the samples were transferred to the laboratory of Saudi Berkefeled Filters Co., in Riyadh for the chemical analysis. All chemical analyses were achieved according to standard methods.

[8] Predictive Models for River Water Quality using Machine Learning and Big Data Techniques by Jitha P Nair, M S Vijaya this paper describes the water resources become more polluted. Waste disposal from industry, human wastes, automobile wastes, agricultural runoff from farmlands containing chemical factors, unwanted nutrients, and other wastes from point and non-point source flow to water bodies, which affects the quality of the water resources. etc. The increase in pollution influences the quantity and quality of water, which results high risk on health and other issues for human as well as for living organisms on the planet. Hence, evaluating and monitoring the quality of water, and its prediction become crucial and applicable area for research in the current scenario. In various researchers they have used traditional approaches; Now, they are using technologies like ML, big data analytics for evaluation and prediction of water quality.

[9]Application of Predictive Intelligence in Water Quality Forecasting of the River Ganga Using Support Vector Machines by Anil Kumar Bisht, Ravendra Singh, Rakesh Bhutiani and Ashutosh Bhatt this paper describes A precise prediction of river water quality may benefit the water management bodies. However, due to the complex relationship existing among various factors, the prediction is a challenging job. Here, the authors attempted to develop a model for forecasting or predicting the water quality of the river Ganga using application of predictive intelligence based on machine learning approach called support vector machine (SVM). The monthly data sets of five water quality parameters from 2001 to 2015 were taken from five sampling stations from Dev prayag to Roorkee in the Uttarakhand state of India. The experiments are conducted in Python 2.7.13 language using the radial basis function (RBF) as a kernel for developing the non-linear SVM-based classifier as a model for water quality prediction. The results indicated a prediction performance of 96.66% for best parameter combination which proved the significance of predictive intelligence in water quality forecasting.

[10] DEVELOPMENT OF A WATER QUALITY INDEX (WQI) FOR THE LOKTAK LAKE IN INDIA R. Das Kangabam, S. D. Bhoominathan, S. Kanagaraj, and M. Govindaraju et.al has proposed, The present work was carried out to assess a water quality index (WQI) of the Loktak Lake, an important wetland which has been under pressure due to the increasing anthropogenic activities. Physicochemical parameters like temperature (Tem), potential hydrogen (pH), electrical conductivity (EC), turbidity (T), dissolved oxygen (DO), total hardness (TH), calcium (Ca), chloride (Cl), fluoride (F), sulphate (SO2 4), magnesium (Mg), phosphate (PO3 4), sodium (Na), potassium (K), nitrite (NO2), nitrate (NO3), total dissolved solids (TDS), total carbon (TC), biochemical oxygen demand (BOD), and chemical oxygen demand (COD) were analyzed using standard procedures. The values obtained were compared with the guidelines for drinking purpose suggested by the World Health Organization and Bureau of Indian Standard. The result shows the higher concentration of nitrite in all the location which is beyond the permissible limit. Eleven parameters were selected to derive the WQI for the estimation of water potential for five sampling sites. A relative weight was assigned to each parameter range from 1.46 to 4.09 based on its importance. The WQI values range from 64 to 77 indicating that the Loktak Lake water is not fit for drinking, including both human and animals, even though the people living inside the Lake are using it for drinking purposes. The implementation of WQI is necessary for proper management of the Loktak Lake and it will be a very helpful tool for the public and decision makers to evaluate the water quality of the Loktak Lake for sustainable management. Water quality is an important contributor touching on all aspects of ecosystems and human well-being and a significant tool in determining the human poverty, wealth, and education levels (UN Water 2010). The ecosystem services of water from rivers and lakes are directly or indirectly contribute to both human welfare and aquatic ecosystem. The increase in pollution of water sources like lakes and rivers is a major concern for the global scenario as most of the water bodies around the world are the source for water supply including human consumption and domestic purposes . The health of the aquatic ecosystem is determined by the water quality parameter which includes the physical, chemical, and biological characteristics. Therefore, a particular problem with water quality monitoring is a complex issue associated with analyzing a large number of associate measures of variables and the high variability among the variables is due to increase in anthropogenic activities including natural influences . Environmental pollution of water resources has become a major global issue, including developing countries which have been suffering from the impact of pollution due to poor socio economic growth associated with the exploitation of natural resources. As a result of it, water is considered as the highest risk to the world for future due to increase in demand as well as increase in pollution.

III.SYSTEM SPECIFICATION AND REQUIREMENTS

The requirements specifications states what are the components and materials that are needed to carry out the project in a successful manner. The requirements can be categorized as hardware requirements and software requirements. In this, the hardware requirements are physical devices needed to run the project., whereas the software requirements are platforms/tools used to build the project.

3.1 Hardware Requirements

• Processor type	:	i5 or Above
• RAM	:	4GB or Above
• Hard disk	:	1TB
• Keyboard	:	101/102 standard keyboards
• Mouse	:	Optical mouse

3.2 Software Requirements

• Operating system	:	Windows 10 or Above
• Front end	:	Jupiter notebook / Anaconda tools
• Coding language	:	Python
• Type	:	Machine learning

Jupiter Notebook

The Jupiter Notebook is an open-source web application that allows you to create and share documents that contain live code, equations, visualizations and narrative text. Uses include: data cleaning and transformation, numerical simulation, statistical modelling, data visualization, machine learning, and much more. The software requirement specification is created at the end of the analysis task. The function and performance allocated to software as part of system engineering are developed by establishing a complete information report as functional representation, a representation of system behaviour, an indication of performance requirements and design constraints, appropriate validation criteria.

Python

Python is a high-level programming language designed to be easy to read and simple to implement. It is open source, which means it is free to use, even for commercial applications. Python can run on Mac, Windows, and Unix systems and has also been ported to Java and .NET virtual machines. Python is considered a scripting language, like Ruby or Perl and is often used for creating Web applications and dynamic Web content. It is also supported by a number of 2D and 3D imaging programs, enabling users to create custom plug-ins and extensions with Python. Examples of applications that support a Python API include GIMP, Inkscape, Blender, and Autodesk Maya. Scripts written in Python (.PY files) can be parsed and run immediately. They can also be saved as a compiled programs (.PYC files), which are often used as programming modules that can be referenced by other Python programs.

PYTHON LIBRARIES

NUMPY

NumPy is a very popular python library for large multi-dimensional array and matrix processing, with the help of a large collection of high-level mathematical functions. It is very useful for fundamental scientific computations in Machine Learning. It is particularly useful for linear algebra, Fourier transform, and random number capabilities. High-end libraries like TensorFlow uses NumPy internally for manipulation of Tensors.

SCIPY

SciPy is a very popular library among Machine Learning enthusiasts as it contains different modules for optimization, linear algebra, integration and statistics. There is a difference between the SciPy library and the SciPy stack. The SciPy is one of the core packages that make up the SciPy stack. SciPy is also very useful for image manipulation.

SCIKIT-LEARN

Scikit-learn is one of the most popular ML libraries for classical ML algorithms. It is built on top of two basic Python libraries, viz., NumPy and SciPy. Scikit-learn supports most of the supervised and unsupervised learning algorithms. it can also be used for data-mining and data-analysis, which makes it a great tool who is starting out with ML.

PANDAS

Pandas is a popular Python library for data analysis. It is not directly related to Machine Learning. As we know that the dataset must be prepared before training. In this case, Pandas comes handy as it was developed specifically for data extraction and preparation. It provides high-level data structures and wide variety tools for data analysis. It provides many inbuilt methods for groping, combining and filtering data.

MATPLOTLIB

It is a very popular Python library for data visualization. Like Pandas, it is not directly related to Machine Learning. It particularly comes in handy when a programmer wants to visualize the patterns in the data. It is a 2D plotting library used for creating 2D graphs and plots. A module named pyplot makes it easy for programmers for plotting as it provides features to control line styles, font properties, etc. It provides various kinds of graphs and plots for data visualization, viz., histogram, error charts, bar chats, etc.

IV.METHODOLOGY

4.1 SYSTEM ARCHITECTURE

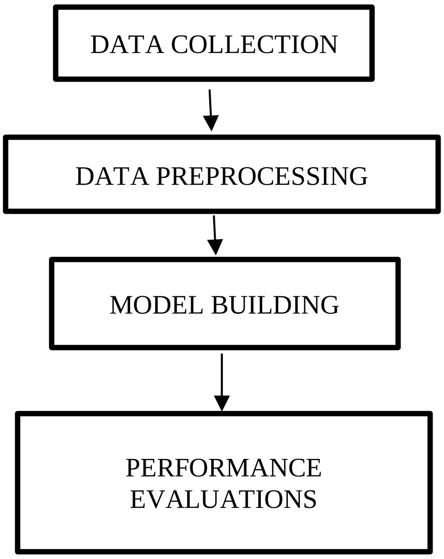


FIGURE 4.1 BLOCK DIAGRAM

DATA COLLECTION

The Dataset used in the proposed system consists of 3277 rows* 10 columns and it is in the form of CSV. It consists of parameters like PH, Solids, Hardness, Chloramine, sulphate, conductivity, Organic Carbon, Trihalomethane, Turbidity.

PH

PH is a measure of how acidic or basic (alkaline) the water is (the term pH comes from the French: "puissance dihydrogen" which means strength of the hydrogen). It is defined as the negative log of the hydrogen ion concentration.
The pH scale is logarithmic and goes from 0 to 14. For each whole number increase (i.e. 1 to 2) the hydrogen ion concentration decreases ten fold and the water becomes less acidic.

As the pH decreases, water becomes more acidic. As water becomes more basic, the pH increases

- Many chemical reactions inside aquatic organisms (cellular metabolism) that are necessary for survival and growth of organisms require a narrow pH range.
- At the extreme ends of the pH scale, (2 or 13) physical damage to gills, exoskeleton, fins, occurs.
- Changes in pH may alter the concentrations of other substances in water to a more toxic form. Examples: a decrease in pH (below 6) may increase the amount of mercury soluble in water. An increase in pH (above 8.5) enhances the conversion of nontoxic ammonia (ammonium ion) to a toxic form of ammonia (un-ionized ammonia).

CONDUCTIVITY

Solids can be found in nature in a dissolved form. Salts that dissolve in water break into positively and negatively charged ions. Conductivity is the ability of water to conduct an electrical current, and the dissolved ions are the conductors. The major positively charged ions are sodium, (Na+) calcium (Ca+2), potassium (K+) and magnesium (Mg+2). The major negatively charged ions are chloride (Cl-), sulfate (SO4-2), carbonate (CO3-2), and bicarbonate (HCO3-). Nitrates (NO3-2) and phosphates (PO4-3) are minor contributors to conductivity, although they are very important biologically.

Salinity is a measure of the amount of salts in the water. Because dissolved ions increase salinity as well as conductivity, the two measures are related. The salts in sea water are primarily sodium chloride (NaCl). However, other saline waters, such as Mono Lake, owe their high salinity to a combination of dissolved ions including sodium, chloride, carbonate and sulfate.

Salts and other substances affect the quality of water used for irrigation or drinking. They also have a critical influence on aquatic biota, and every kind of organism has a typical salinity range that it can tolerate. Moreover, the ionic composition of the water can be critical. For example, cladocerans (water fleas) are far more sensitive to potassium chloride than sodium chloride at the same concentration.

Conductivity will vary with water source: ground water, water drained from agricultural fields, municipal waste water, rainfall. Therefore, conductivity can indicate groundwater seepage or a sewage leak.

TEMPERATURE

Temperature is a measure of the average energy (kinetic) of water molecules. It is measured on a linear scale of degrees Celsius or degrees Fahrenheit.

It is one of the most important water quality parameters. Temperature affects water chemistry and the functions of aquatic organisms. It influences the:

- amount of oxygen that can be dissolved in water,
- rate of photosynthesis by algae and other aquatic plants,
- metabolic rates of organisms,
- sensitivity of organisms to toxic wastes, parasites and diseases, and timing of reproduction, migration, and aestivation of aquatic organisms.

TURBIDITY

Turbidity is a measure of the amount of suspended particles in the water. Algae, suspended sediment, and organic matter particles can cloud the water making it more turbid.

Suspended particles diffuse sunlight and absorb heat. This can increase temperature and reduce light available for algal photosynthesis. If the turbidity is caused by suspended sediment, it can be an indicator of erosion, either natural or man-made. Suspended sediments can clog the gills of fish. Once the sediment settles, it can foul gravel beds and smother fish eggs and benthic insects. The sediment can also carry pathogens, pollutants and nutrients.

DATA PREPROCESSING

The dataset is taken as the input which consist of the attributes and values The input dataset is given as the input with the attributes such as the ph, hardness, solids, chloramines, sulfate etc., Large number of records collected in the dataset are given as the input. Data processing occurs when data is collected and translated into usable information. Usually performed by a data scientist or team of data scientists, it is important for data processing to be done correctly as not to negatively affect the end product, or data output. Data processing starts with data in its raw form and converts it into a more readable format (graphs, documents, etc.), giving it the form and context necessary to be interpreted by computers and utilized by employees throughout an organization.

MODEL BUILDING

ALGORITHMS

KNN ALGORITHM

K-Nearest Neighbour is one of the simplest Machine Learning algorithms based on Supervised Learning technique. K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories. K-NN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using K- NN algorithm. K-NN algorithm can be used for Regression as well as for Classification but mostly it is used for the Classification problems. K-NN is a non-parametric algorithm, which means it does not make any assumption on underlying data. It is also called a lazy learner algorithm because it does not learn from the training set immediately instead it stores the dataset and at the time of classification, it performs an action on the dataset. KNN algorithm at the training phase just stores the dataset and when it gets new data, then it classifies that data into a category that is much similar to the new data.

Working

Step 1 − For implementing any algorithm, we need dataset. So during the first step of KNN, we must load the training as well as test data.

Step 2 − Next, we need to choose the value of K i.e. the nearest data points. ...

Step 3 − For each point in the test data do the following −

Step 4 − End.

FORMULA

$d=\sqrt{((x2-x1)^2+(y2-y1)^2)}$

XG BOOST

From the project description, it aims to provide a "Scalable, Portable and Distributed Gradient Boosting (GBM, GBRT, GBDT) Library". It runs on a single machine, as well as the distributed processing frameworks. This algorithm goes by lots of different names such as gradient boosting, multiple additive regression trees, stochastic gradient boosting or gradient boosting machines. Boosting is an ensemble technique where new models are added to correct the errors made by existing models. Models are added sequentially until no further improvements can be made. Gradient boosting is an approach where new models are created that predict the residuals or errors of prior models and then added together to make the final prediction. It is called gradient boosting because it uses a gradient descent algorithm to minimize the loss when adding new models.

Working

Step 1: Create a Jupiter Notebook.

Step 2: Download, Explore, and Transform Data.

Step 3: Train a Model.

Step 4: Deploy the Model.

Step 5: Evaluate the Model.

Step 6: Clean Up.

FORMULA

F2(x) =H0(x)+eta(H1(x)) +eta(H2(x))

LOGISTIC REGRESSION

Logistic regression is a process of modeling the probability of a discrete outcome given an input variable. Logistic regression is a useful analysis method for classification problems, where you are trying to determine if a new sample fits best into a category. As aspects of cyber security are classification problems, such as attack detection, logistic regression is a useful analytic technique. Logistic regression, despite its name, is a classification model rather than regression model. Logistic regression is a simple and more efficient method for binary and linear classification problems. It is a classification model, which is very easy to realize and achieves very good performance with linearly separable classes. It is an extensively employed algorithm for classification in industry

WORKING

Step 1: Data Pre-processing

Step 2: Fitting Logistic Regression to the Training set.

Step 3: Predicting the test result.

Step 4: Test accuracy of the result

Step 5: Visualizing the test set result.

FORMULA

log(p/1-p) is the link function. Logarithmic transformation on the outcome variable allows us to model a non-linear association in a linear way. This is the equation used in Logistic Regression

DECISION TREE

Decision tree is the most powerful and popular tool for classification and prediction. A Decision tree is a flowchart like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (terminal node) holds a class label. A tree can be "learned" by splitting the source set into subsets based on an attribute value test. This process is repeated on each derived subset in a recursive manner called recursive partitioning. The recursion is completed when the subset at a node all has the same value of the target variable, or when splitting no longer adds value to the predictions. The construction of decision tree classifier does not require any domain knowledge or parameter setting, and therefore is appropriate for exploratory knowledge discovery. Decision trees can handle high dimensional data. In general decision tree classifier has good accuracy. Decision tree induction is a typical inductive approach to learn knowledge on classification.

SVM

Classifying data is a common task in machine learning. Suppose some given data points each belong to one of two classes, and the goal is to decide which class a new data point will be in. In the case of support-vector machines, a data point is viewed as a p-dimensional vector (a list of and we want to know whether we can separate such points with a dimensional hyper plane. This is called a linear classifier. There are many hyper planes that might classify the data. One reasonable choice as the best hyper plane is the one that represents the largest separation, or margin, between the two classes. So we choose the hyper plane so that the distance from it to the nearest data point on each side is maximized. SVMs are helpful in text and hypertext categorization, as their application can significantly reduce the need for labelled training instances in both the standard inductive and transudative settings. Some methods for shallow semantic parsing are based on support vector machines. What makes the SVM algorithm stand out compared to other algorithms is that it can deal with classification problems using an SVM classifier and regression problems using an SVM regressor. However, one must remember that the SVM classifier is the backbone of the support vector machine concept and, in general, is the attest algorithm to solve classification problems.

EVALUTION

PREDICTION

Traditionally, machine learning models have not included insight into why or how they arrived at an outcome. This makes it difficult to objectively explain the decisions made and actions taken based on these models. Prediction Explanations avoid the "black box" syndrome by describing which characteristics, or feature variables, have the greatest impact on a model's outcomes. When the reasons behind a model's outcomes are as important as the outcomes themselves, Prediction Explanations can uncover the factors that most contribute to those outcomes. For example, banks using models to determine whether or not they should approve a loan can use Prediction Explanations to gain insight into why an application was accepted or rejected. With that insight they can develop models that comply with regulations, easily explain model outcomes to stakeholders, and identify high-impact factors to help focus their business strategies.

In statistics, prediction is a part of statistical inference. One particular approach to such inference is known as predictive inference, but the prediction can be undertaken within any of the several approaches to statistical inference. Indeed, one possible description of statistics is that it provides a means of transferring knowledge about a sample of a population to the whole population, and to other related populations, which is not necessarily the same as prediction over time. When information is transferred across time, often to specific points in time, the process is known as forecasting. Forecasting usually requires time series methods, while prediction is often performed on cross-sectional data.

A prediction or forecast, is a statement about a future event or data. They are often, but not always, based upon experience or knowledge. There is no universal agreement about the exact difference from "estimation"; different authors and disciplines ascribe different connotations.

Future events are necessarily uncertain, so guaranteed accurate information about the future is impossible. Prediction can be useful to assist in making plans about possible developments; Howard H. Stevenson writes that prediction in business "is at least two things: Important and hard.

Statistical techniques used for prediction include regression analysis and its various sub-categories such as linear regression, generalized linear models (logistic regression, passion regression, Probitregression),etc.Incaseof forecasting, autoregressive moving average models and vector autoregression models can be utilized. When these and/or related, generalized set of regression or machine learning methods are deployed in commercial usage, the field is known as predictive analytics.

In many applications, such as time series analysis, it is possible to estimate the models that generate the observations. If models can be expressed as transfer functions or in terms of state-space parameters then smoothed, filtered and predicted data estimates can be calculated.[citation needed] If the underlying generating models are linear then a minimum-variance Kalman filter and a minimum-variance smoother may be used to recover data of interest from noisy measurements. These techniques rely on one-step-ahead predictors (which minimise the variance of the prediction error). When the generating models are nonlinear then stepwise linearizations may be applied within Extended Kalman Filter and smoother recursions. However, in nonlinear cases, optimum minimum-variance performance guarantees no longer apply.

To use regression analysis for prediction, data are collected on the variable that is to be predicted, called the dependent variable or response variable, and on one or more variables whose values are hypothesized to influence it, called independent variables or explanatory variables. A functional form, often linear, is hypothesized for the postulated causal relationship, and the parameters of the function are estimated from the data—that is, are chosen so as to optimize is some way the fit of the function, thus parameterized, to the data. That is the estimation step. For the prediction step, explanatory variable values that are deemed relevant to future (or current but not yet observed) values of the dependent variable are input to the parameterized function to generate predictions for the dependent variable.

Prediction in the non-economic social sciences differs from the natural sciences and includes multiple alternative methods such as trend projection, forecasting, scenario-building and Delphi surveys. The oil company Shell is particularly well known for its scenario-building activities.

One reason for the peculiarity of societal prediction is that in the social sciences, "predictors are part of the social context about which they are trying to make a prediction and may influence that context in the process".As a consequence, societal predictions can become self-destructing. For example, a forecast that a large percentage of a population will become HIV infected based on existing trends may cause more people to avoid risky behavior and thus reduce the HIV infection rate, invalidating the forecast (which might have remained correct if it had not been publicly known). Or, a prediction that cybersecurity will become a major issue may cause organizations to implement more security cybersecurity measures, thus limiting the issue

V.SYSTEM DESIGN

5.1 INTRODUCTION

A technology that enables a machine to stimulate human behaviour to help in solving complex problems is known as Artificial Intelligence. Machine learning is a subset of AI and allows machines to learn from past data and provide an accurate output. In our proposed system, the prediction of Water quality has been made on Time series data. The data will be obtained from Kaggle. In Prepossessing, the data need to be cleaned and explore to understand the insights, some cases normalization also required to minimize the computation. Then the required parameters are scaling using Min-Max scalar techniques. The time series cleaned data can be trained with Logistic Regression, Decision Tree, Random Forest, XG Boost, KNN, SVM, Ada Boost build model will be tested for its performance, it can be measured with MSE and RMSE.

5.2 FLOW CHART

A flowchart is a type of diagram that represents a workflow or process. A flowchart can also be defined as a diagrammatic representation of an algorithm, a step by-step approach to solving a task.

The flowchart shows the steps as boxes of various kinds, and their order by connecting the boxes with arrows. This diagrammatic representation illustrates a solution model to a given problem. Flowcharts are used in analysing, designing, documenting or managing a process or program in various fields.

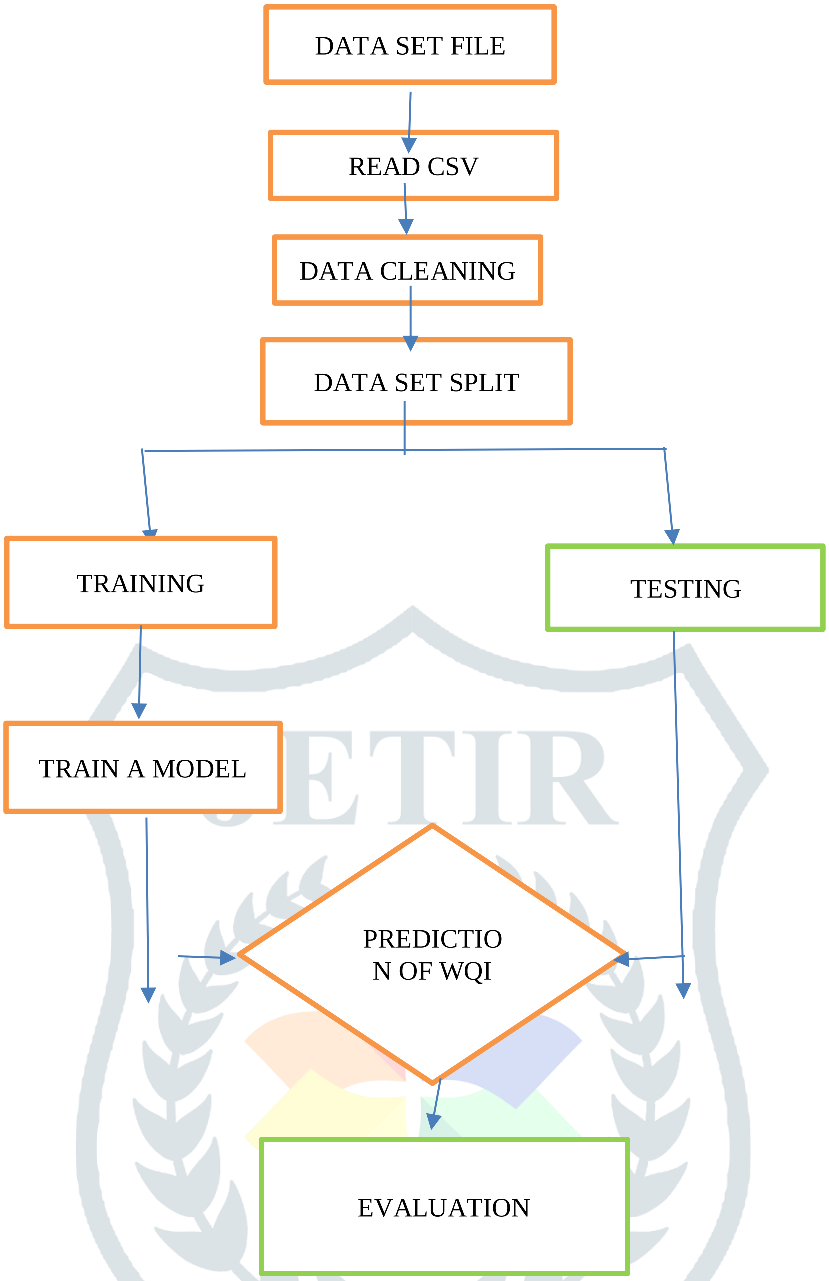
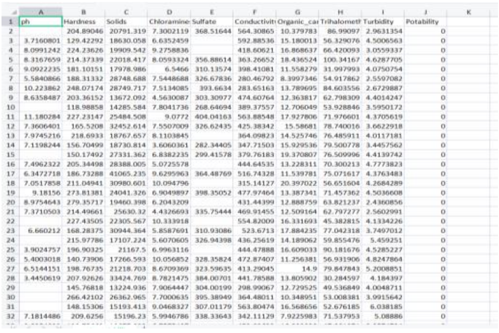


FIGURE 5.1 FLOW CHART

DATASET FILE

The Dataset consists of 3277 rows* 10 columns and it is in the form of CSV. It consists of parameters like PH, Solids, Hardness, Chloramine, sulphate, conductivity, Organic Carbon, Trihalomethane, Turbidity.



	A	B	C	D	E	F	G	H	I	J	K
1	ph	Hardness	Solids	Chloramine	Sulfate	Conductivit	Organic_car	Trihalometh	Turbidity	Potability	
2		204.89046	20791.319	7.3002119	368.51644	564.30865	10.379783	86.99097	2.9631354	0	
3	3.7160801	129.42292	18630.058	6.6352459		592.88536	15.180013	56.329076	4.5006563	0	
4	8.0991242	224.23626	19909.542	9.2758836		418.60621	16.868637	66.420093	3.0559337	0	
5	8.3167659	214.37339	22018.417	8.0593324	356.88614	363.26652	18.436524	100.34167	4.6287705	0	
6	9.0922235	181.10151	17978.986	6.5466	310.13574	398.41081	11.558279	31.997993	4.0750754	0	
7	5.5840866	188.31332	28748.688	7.5448688	326.67836	280.46792	8.3997346	54.917862	2.5597082	0	
8	10.223862	248.07174	28749.717	7.5134085	393.6634	283.65163	13.789695	84.603556	2.6729887	0	
9	8.6358487	203.36152	13672.092	4.5630087	303.30977	474.60764	12.363817	62.798309	4.4014247	0	
10		118.98858	14285.584	7.8041736	268.64694	389.37557	12.706049	53.928846	3.5950172	0	
11	11.180284	227.23147	25484.508	9.0772	404.04163	563.88548	17.927806	71.976601	4.3705619	0	
12	7.3606401	165.5208	32452.614	7.5507009	326.62435	425.38342	15.58681	78.740016	3.6622918	0	
13	7.9745216	218.6933	18767.657	8.1103845		364.09823	14.525746	76.485911	4.0117181	0	
14	7.1198244	156.70499	18730.814	3.6060361	282.34405	347.71503	15.929536	79.500778	3.4457562	0	
15		150.17492	27331.362	6.8382235	299.41578	379.76183	19.370807	76.509996	4.4139742	0	
16	7.4962322	205.34498	28388.005	5.0725578		444.64535	13.228311	70.300213	4.7773823	0	
17	6.3472718	186.73288	41065.235	9.6295963	364.48769	516.74328	11.539781	75.071617	4.3763483	0	
18	7.0517858	211.04941	30980.601	10.094796		315.14127	20.397022	56.651604	4.2684289	0	
19	9.18156	273.81381	24041.326	6.9049897	398.35052	477.97464	13.387341	71.457362	4.5036608	0	
20	8.9754643	279.35717	19460.398	6.2043209		431.44399	12.888759	63.821237	2.4360856	0	
21	7.3710503	214.49661	25630.32	4.4326693	335.75444	469.91455	12.509164	62.797277	2.5602991	0	
22		227.43505	22305.567	10.333918		554.82009	16.331693	45.382815	4.1334226	0	
23	6.660212	168.28375	30944.364	5.8587691	310.93086	523.6713	17.884235	77.042318	3.7497012	0	
24		215.97786	17107.224	5.6070605	326.94398	436.25619	14.189062	59.855476	5.459251	0	
25	3.9024757	196.90325	21167.5	6.9963116		444.47888	16.609033	90.181676	4.5285227	0	
26	5.4003018	140.73906	17266.593	10.056852	328.35824	472.87407	11.256381	56.931906	4.8247864	0	
27	6.5144151	198.76735	21218.703	8.6709369	323.59635	413.29045	14.9	79.847843	5.2008851	0	
28	3.4450619	207.92626	33424.769	8.7821475	384.00701	441.78588	13.805902	30.284597	4.184397	0	
29		145.76818	13224.936	7.9064447	304.00199	298.99067	12.729525	49.536849	4.0048711	0	
30		266.42102	26362.965	7.7000635	395.38949	364.48011	10.348951	53.008381	3.9915642	0	
31		148.15306	15193.413	9.0468327	307.01179	563.80474	16.568656	52.676185	6.038185	0	
32	7.1814486	209.6256	15196.23	5.9946786	338.33643	342.11129	7.9225983	71.537953	5.08886	0	

FIGURE 5.1 DATA SET

DATA CLEANING

Data cleaning is the process of fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data with in a dataset. When combining multiple data sources, there are many opportunities for data to be duplicated or mislabelled**.**

DATASET SPLIT

In machine learning, data splitting is typically done to avoid overfitting. That is an instance where a machine learning model fits its training data too well and fails to reliably fit additional data. The original data in a machine learning model is typically taken and split into three or four sets.

TRAINING AND TESTING

Split the data set into two data sets: a training and a testing set.
The training set should be a random selection of 80% of the original data.
The testing set should be the remaining 20%.

Steps to training a model:
Step 1: Begin with existing data
Step 2: Analyze data to identify
Step 3 : Make prediction

SYSTEM TESTING

System testing makes a logical assumption that, if all parts of the systems are correct, system testing is its utility as a user oriented vehicle before implementation. The best program is worthless if it does not meet user needs. System testing identifies the errors, presenting the proposal to the administrator and changes the modification and also checks the reliability of output. Before implementation, the system is tested whether the required software and hardware are available to run this project.
This project has undergone the following testing procedures to ensure its correctness.

- **Unit testing**
- **Integration testing**
- **Validation testing**

Unit Testing

The primary goal of unit testing is to take the smallest piece of testable software in the application, isolate it from the remainder of the code, and determine whether it behaves exactly as you expect. Each unit is tested separately before integrating them into modules totes the interfaces before modules. Unit testing has proven its value in that a large percentage of defects are identified during its use.
The procedure level testing is made first. By giving improper inputs, the errors occurred are noted and eliminated. Then the form level testing is made. For example ,storage of data to the table is in the correct manner. In this system, each form is considered as a separate unit and tested for errors. Every user input is unit tested for a valid accepted range.

Integration Testing

Integration testing sometimes called integration and Testing, abbreviated" I&T" is the phase in software testing in which individual software modules are combined and tested as a group. It occurs after unit testing and before system testing. Integration testing takes as its input modules that have been unit tested, groups them in larger aggregates, applies tests defined in an integration test plan to those aggregates, and delivers as its output the integrates system ready for system testing.

The purpose of integration testing is to verify functional performance, and reliability requirements placed on major design items. These "design items ",ie. assemblages (or groups of units), are exercised through their interfaces using Black box testing ,success and error cases being simulated via appropriate parameter and data inputs.

Testing is done for each module .After testing all the modules, the modules are integrated and testing of the final system is done with the test data, specially designed to show that the system will operate successfully in all its aspects conditions. Thus the system testing is a confirmation that all is correct and an opportunity to show the user that the system works.

Validation Testing

Validation can be defined in many ways, but a simple definition is that can be reasonable expected by the clients, which is defined in the software requirement specification, a document that describes all user visible attribute of the software.

PREDICTION OF WQI

5.3 INPUT DESIGN

The input design process is to design the input needs into a machine-oriented format. The object of input design is to create an input layout that is easy to follow user friendly and to avoid operator errors.

In accurate data cause most common errors in data processing made by data entry operators. The help of error message can enter the required and formatted date by the user. So you are design the inputted design to simply entered format.

The Formatted input entries such as edit mask, radio button, drop down data window help the user to enter the data very easily without much knowledge of the product.

Here also much care is taken to have standardization over the GUI based development with same standard set & rules.

The Menu based product helps even the native user work with the product. The success are designs in such a way to help the user to get the information whenever necessary.

5.4 OUTPUT DESIGN

The Output designs are displayed some different report formats. Different output design will improve the clarity and performing of output. The output designs are classified into individuals and group of tables is possible.

And also display the reports are in lab tests, cross matching, issue details is available in my project. It is used to check the collection of particular time of the period.

VI.SYSTEM IMPLEMENTATION

6.1 SYSTEM MAINTAINANCE

The objectives of this maintenance work are to make sure that the system gets into work all time without any bug. Provision must be for environmental changes which may affect the computer or software system. This is called the maintenance of the system. Nowadays there is the rapid change in the software world. Due to this rapid change, the system should be capable of adapting these changes. In this project the process can be added without affecting other parts of the system.

Maintenance plays a vital role. The system is liable to accept any modification after its implementation. This system has been designed to favor all new changes. Doing this will not affect the system's performance or its accuracy.

Maintenance is necessary to eliminate errors in the system during its working life and to tune the system to any variations in its working environment. It has been seen that there are always some errors found in the system that must be noted and corrected. It also means the review of the system from time to time.

The review of the system is done for:

- Knowing the full capabilities of the system.
- Knowing the required changes or the additional requirements.
- Studying the performance.

6.2 SYSTEM IMPLEMENTATION

After proper testing and validation, the question arises whether the system can be implemented or not. Implementation includes all those activities that place to convert from old system or new. The new system may be totally new replacing an existing or automated system, or it may be a major modification to an existing system. In other case, proper implementation is essential to provide a reliable system to meet organization requirements.

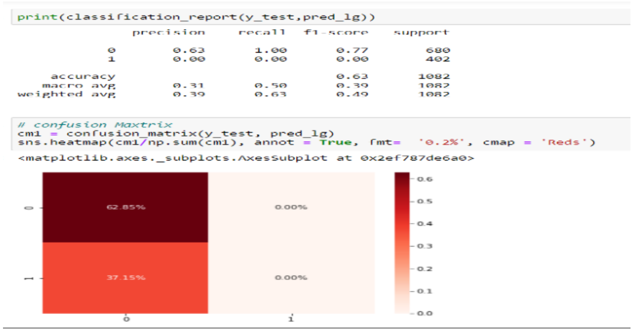
Implementation is the stage where the theoretical design is turned into a working system. The most crucial stage is to achieve a new successful system and to make user confident that the new system will work effectively. All information given by a user is successful stored in a database for future referenc

After having the user acceptance of the new system developed, the implementation phase begins. Implementation is the stage of a project during which theory is turned into practice. During this phase, all the programs of the system are loaded onto the user's computer. After loading the system, training of the user starts. Main topics of such type of training are:

- How to execute the package
- How to enter the data
- How to process the data (processing details)
- How to take out the reports

VII.RESULT

Modeling and the prediction of water quality have played a pivotal and significant role in saving time and consumption in lab analysis. Artificial intelligence algorithms were explored as an alternative method to estimate and predict water quality. This study suggests that the combined approach of the artificial intelligence techniques proposed in the current study should be applied as a promising tool to accurately simulate water level and quality. The developed model can be used easily and inexpensively to predict water quality and index and thus water quality classification with high accuracy. In this study the SVM shows the better results in predict the accuracy. The machine learning algorithms gives overall better performance accuracy.

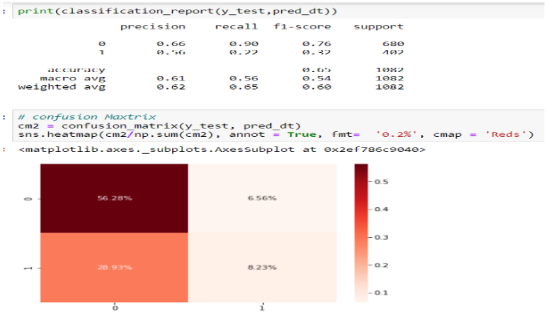


```
print(classification_report(y_test,pred_lg))
              precision    recall  f1-score   support

           0       0.63      1.00      0.77       680
           1       0.00      0.00      0.00       402

    accuracy                           0.63      1082
   macro avg       0.31      0.50      0.39      1082
weighted avg       0.39      0.63      0.49      1082
```

```
// confusion Maxtrix
cm1 = confusion_matrix(y_test, pred_lg)
sns.heatmap(cm1/np.sum(cm1), annot = True, fmt=  '0.2%', cmap = 'Reds')
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x2ef787de6a0>
```

F IGURER 7.1.LOGISTIC REGRESSION

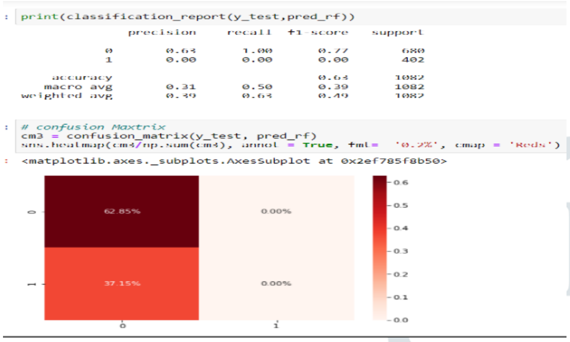


```
print(classification_report(y_test,pred_dt))
              precision    recall  f1-score   support

           0       0.66      0.90      0.76       680
           1       0.56      0.22      0.32       402

    accuracy                           0.65      1082
   macro avg       0.61      0.56      0.54      1082
weighted avg       0.62      0.65      0.60      1082
```

```
# confusion Maxtrix
cm2 = confusion_matrix(y_test, pred_dt)
sns.heatmap(cm2/np.sum(cm2), annot = True, fmt=  '0.2%', cmap = 'Reds')
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x2ef786c9040>
```

F IGURE 11.2 DECISION TREE

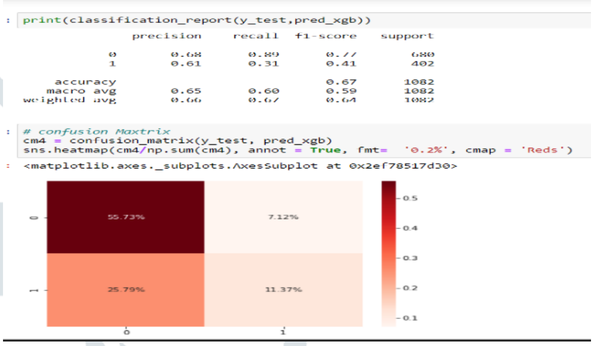


```
print(classification_report(y_test,pred_rf))
              precision    recall  f1-score   support

           0       0.63      1.00      0.77       680
           1       0.00      0.00      0.00       402

    accuracy                           0.63      1082
   macro avg       0.31      0.50      0.39      1082
weighted avg       0.39      0.63      0.49      1082
```

```
# confusion Maxtrix
cm3 = confusion_matrix(y_test, pred_rf)
sns.heatmap(cm3/np.sum(cm3), annot = True, fmt=  '0.2%', cmap = 'Reds')
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x2ef785f8b50>
```

FIGURE 7.3 RANDOM FOREST

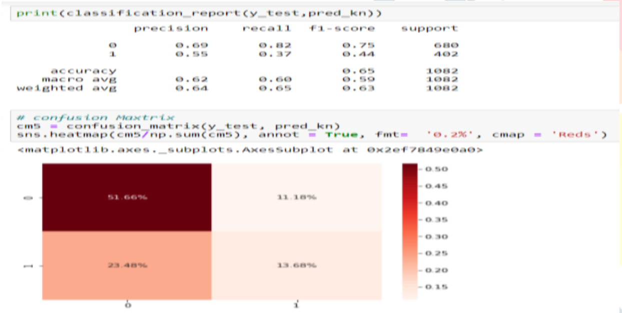


```
print(classification_report(y_test,pred_xgb))
              precision    recall  f1-score   support

           0       0.68      0.89      0.77       680
           1       0.61      0.31      0.41       402

    accuracy                           0.67      1082
   macro avg       0.65      0.60      0.59      1082
weighted avg       0.66      0.67      0.64      1082
```

```
# confusion Maxtrix
cm4 = confusion_matrix(y_test, pred_xgb)
sns.heatmap(cm4/np.sum(cm4), annot = True, fmt=  '0.2%', cmap = 'Reds')
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x2ef78517d30>
```

FIGURE 7.4 XG BOOST

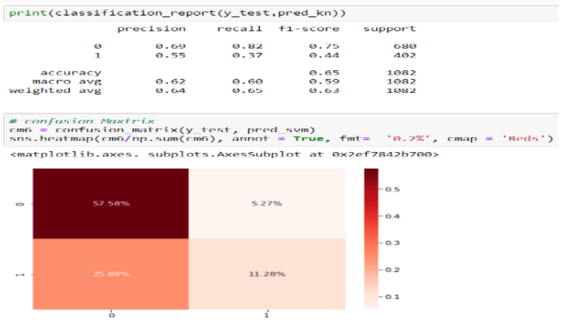


```
print(classification_report(y_test,pred_kn))
              precision    recall  f1-score   support

           0       0.69      0.82      0.75       680
           1       0.55      0.37      0.44       402

    accuracy                           0.65      1082
   macro avg       0.62      0.60      0.59      1082
weighted avg       0.64      0.65      0.63      1082
```

```
# confusion Maxtrix
cm5 = confusion_matrix(y_test, pred_kn)
sns.heatmap(cm5/np.sum(cm5), annot = True, fmt=  '0.2%', cmap = 'Reds')
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x2ef7849e0a0>
```

FIGURE 7.5 KNN

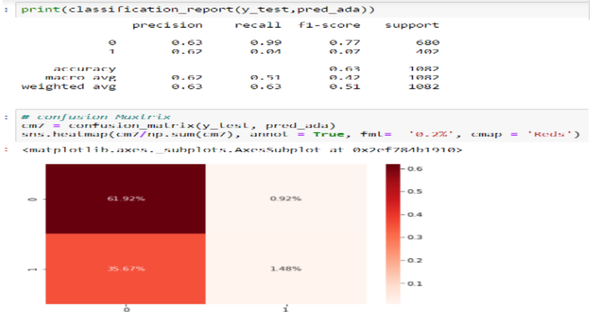


```
print(classification_report(y_test,pred_kn))
              precision    recall  f1-score   support

           0       0.69      0.82      0.75       680
           1       0.55      0.37      0.44       402

    accuracy                           0.65      1082
   macro avg       0.62      0.60      0.59      1082
weighted avg       0.64      0.65      0.63      1082
```

```
# confusion Maxtrix
cm6 = confusion_matrix(y_test, pred_svm)
sns.heatmap(cm6/np.sum(cm6), annot = True, fmt=  '0.2%', cmap = 'Reds')
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x2ef7842b700>
```

FIGURE 7.6 SVM

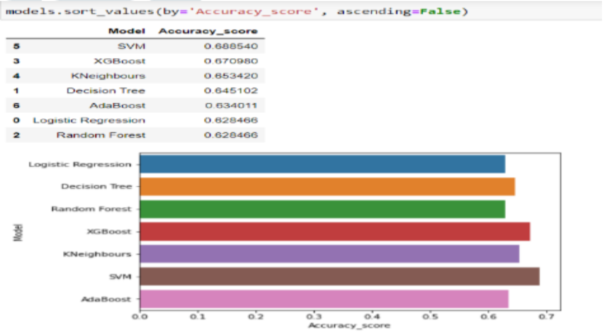


```
print(classification_report(y_test,pred_ada))
              precision    recall  f1-score   support

           0       0.63      0.99      0.77       680
           1       0.62      0.04      0.07       402

    accuracy                           0.63      1082
   macro avg       0.62      0.51      0.42      1082
weighted avg       0.63      0.63      0.51      1082
```

```
# confusion Maxtrix
cm7 = confusion_matrix(y_test, pred_ada)
sns.heatmap(cm7/np.sum(cm7), annot = True, fmt=  '0.2%', cmap = 'Reds')
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x2ef784b1910>
```

FIGURE 7.7 ADABOOST



```
models.sort_values(by='Accuracy_score', ascending=False)
```

	Model	Accuracy_score
5	SVM	0.688540
3	XGBoost	0.670980
4	KNeighbours	0.653420
1	Decision Tree	0.645102
6	AdaBoost	0.634011
0	Logistic Regression	0.628466
2	Random Forest	0.628466

FIGURE 7.8 OVER ALL RESULT

VIII. Conclusion and Future work

Modeling and prediction of water quality are very important for the protection of the environment. Developing a model by using advanced artificial intelligence algorithms can be used to measure the future water quality. In this proposed methodology, the advanced artificial intelligence algorithms, namely KNN ALGORITHM , XG BOOST, LOGISTIC REGRESSION, RANDOM FOREST, DECISION TREE and SVM. The system plan is reasonable, the structure is rigorous, and the functions are perfect. A new type of low-cost, energy-saving, low-power consumption, flexible, easy-to-expand, and convenient operation and management monitoring system has been implemented.

Future work

In future works, we propose integrating the findings of this research in a large-scale IoT-based online monitoring system using only the sensors of the required parameters. The tested algorithms would predict the water quality immediately based on the real-time data fed from the IoT system. The proposed IoT system would employ the parameter sensors of pH, turbidity, temperature and TDS for parameter readings and communicate those readings using an Arduino microcontroller and ZigBee transceiver. It would identify poor quality water before it is released for consumption and alert concerned authorities. It will hopefully result in curtailment of people consuming poor quality water and consequently de-escalate harrowing diseases like typhoid and diarrhea. In this regard, the application of a prescriptive analysis from the expected values would lead to future facilities to support decision and policy makers

REFERENCES

1. P. Zeilhofer, L. V. A. C. Zeilhofer, E. L. Hardoim, Z. M. . Lima, and C. S. Oliveira, “GIS applications for mapping and spatial modeling of urban-use water quality: a case study in District of Cuiabá, Mato Grosso, Brazil,” Cadernos de Saúde Pública, vol. 23, no. 4, pp. 875–884, 2007.
2. UN water, “Clean water for a healthy world,” Tech. Rep., Development, 2010.
3. K. Farrell-Poe, W. Payne, and R. Emanuel, Water Quality & Monitoring, University of Arizona Repository, 2000, http://hdl.handle.net/10150/146901.
4. T. Taskaya-Temizel and M. C. Casey, “A comparative study of autoregressive neural network hybrids,” Neural Networks, vol. 18, no. 5–6, pp. 781–789, 2005.View at: Publisher Site | Google Scholar
5. C. N. Babu and B. E. Reddy, “A moving-average filter based hybrid ARIMA-ANN model for forecasting time series data,” Applied Soft Computing, vol. 23, pp. 27–38, 2014.
6. Y. C. Lai, C. P. Yang, C. Y. Hsieh, C. Y. Wu, and C. M. Kao, “Evaluation of non-point source pollution and river water quality using a multimedia two-model system,” Journal of Hydrology, vol. 409, no. 3-4, pp. 583–595, 2011.
7. E. Batur and D. Maktav, “Assessment of surface water quality by using satellite images fusion based on PCA method in the Lake Gala, Turkey,” IEEE Transactions on Geoscience and Remote Sensing, vol. 57, no. 5, pp. 2983–2989, 2019.
8. A. A. M. Ahmed and S. M. A. Shah, “Application of adaptive neuro-fuzzy inference system (ANFIS) to estimate the biochemical oxygen demand (BOD) of Surma River,” Journal of King Saud University - Engineering Sciences, vol. 29, no. 3, pp. 237–243, 2017.
9. A. A. Al-Othman, “Evaluation of the suitability of surface water from Riyadh Mainstream Saudi Arabia for a variety of uses,” Arabian Journal of Chemistry, vol. 12, no. 8, pp. 2104–2110, 2019.View at: Publisher Site | Google Scholar
10. T. H. H. Aldhyani, M. Alrasheedi, A. A. Alqarni, M. Y. Alzahrani, and A. M. Bamhdi, “Intelligent hybrid model to enhance time series models for predicting network traffic,” IEEE Access, vol. 8, pp. 130431–130451, 2020