

# **WATER QUALITY PREDICTION USING MACHINE LEARNING PROJECT REPORT**

*Submitted by*

**NIKHIL KUMAR JHA (310519106042)**

**PINTU KUMAR (310519106049)**

**RAJESH KUMAR (310519106093)**

*In partial fulfillment for the award of the degree*

*of*

**BACHELOR OF ENGINEERING**

*in*

***ELECTRONICS AND COMMUNICATION ENGINEERING***



**DHANALAKSHMI SRINIVASAN COLLEGE OF ENGINEERING AND  
TECHNOLOGY, MAMALLAPURAM. CHENNAI- 603 204.**

**MAY 2023**

**ANNA UNIVERSITY: CHENNAI 600 025**

## **BONAFIDE CERTIFICATE**

Certified that this project report entitled “**WATER QUALITY PREDICTION USING MACHINE LEARNING**” is a bonafide work of “**NIKHIL KUMAR JHA (310519106042), PINTU KUMAR (301519106049), RAJESH KUMAR (310519106093)** ” who carried out the project work under my supervision.

### **SIGNATURE**

**Mr. ANANDAN.R,**  
**ASSOCIATE PROFESSOR**  
**Department of ECE**

Dhanalakshmi Srinivasan  
College of Engineering &  
Technology, Mamallapuram  
Chennai - 603104

### **SIGNATURE**

**Mrs. N. BHUVNESWARI**  
**ASSISTANT PROFESSOR**  
**Department of ECE**

Dhanalakshmi Srinivasan  
College of Engineering &  
Technology, mamallapuram  
Chennai - 603104

Submitted for the Anna University project viva voce examination held at

**DHANALAKSHMI SRINIVASAN COLLEGE OF ENGINEERING  
AND TECHNOLOGY** on\_\_\_\_\_

**INTERNAL EXAMINER**

**EXTERNAL EXAMINAR**

## ACKNOWLEDGEMENT

First of all, we would like to thank, The **Almighty** for his blessings upon us to be strengthen our mind and soul to take up this project we owe a great many

Thanks to a great many people who helped and supported us in this project.

We would like to thank **THIRU. A. SRINIVASAN.M.** Our beloved **Chairman** who allowed us to do the project on the college campus.

We would like to express our sense of gratitude to our beloved **Principal Dr. R. SARAVANAN, M.E., Ph.D.**, for his constant support to do the Project.

I extend my grateful thanks to our beloved vice principal **Dr. V.JANAKIRAMAN, M.E., Ph.D.**, for providing his hands to us to successfully complete the project.

We are grateful to **Mr. R.ANANDAN, M.E, Ph.D.**, Our **HOD** ,Who expressed his interest in our work and supplied us with some useful ideas.

We would like to thank our guide **Mrs. N.BHUVANESWARI.**, for following our project with interest and forgiving us with constant support. she taught us not only how to do the project, but also how to enjoy the project.

We would like express our special thanks to the entire **TEACHING AND NON-TEACHING STAFF OF ELECTRONICS AND COMMUNICATION ENGINEERING DEPARTMENT** without whom this project would not see the light of the day.

Also we would like to thank our **Parents** and **Friends** for their blessing, support and Encouragement throughout our life.

## **ABSTRACT**

Freshwater is a critical resource for agriculture and industry's survival. Examination of water quality is a fundamental stage in the administration of freshwater assets. As indicated by the World Health Organization's yearly report, many individuals are getting sick or some are dead due to the lack of safe drinking water, especially pregnant ladies and kids. It is critical to test the quality of water prior to involving it for any reason, whether it is for animal watering, chemical spraying (Pesticides etc.), or drinking water. Water quality testing is a strategy for finding clean drinking water. Accordingly, appropriate water monitoring is basic for safe, clean, and sterile water. Water testing is fundamental for looking at the legitimate working of water sources, testing the safety of drinking water, identifying disease outbreaks, and approving methodology and safeguard activities. Water quality is a proportion of a water's readiness for a specific utilize in view of physical, chemical, and biological qualities. 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. The purpose of this work was to use machine learning techniques such as random forest, NN, MLR, SVM, and BTM to categorize a dataset of water quality in various places across India. 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.

## TABLE OF CONTENTS

CHAPTER NO	TITLE	PAGE NO.
	<b>BONAFIED CERTIFICATE</b>	i
	<b>ACKNOWLEDGEMENT</b>	ii
	<b>ABSTRACT</b>	iii
	<b>LIST OF FIGURES</b>	vi
<b>1</b>	<b>INTRODUCTION</b>	07
	1.1 OVERVIEW	08
	1.2 WATER QUALITY INDEX	08
	1.3 OBJECTIVE	09
<b>2</b>	<b>LITERATURE SURVEY</b>	10
	2.1 WATER QUALITY PREDICTION USING MACHINE LEARNING	10
	2.1.1 ABSTRACT	10
	2.1.2 INTRODUCTION	11
	2.1.3 METHODOLOGY	11
	2.1.4 CONCLUSION	13
	2.2 PROPOSITION OF NEW ENSEMBLE DATA-INTELLIGENCE MODELS FOR SURFACE WATER QUALITY PREDICTION	14
	2.2.1 ABSTRACT	13
	2.2.2 INTRODUCTION	14
	2.2.3 MODELING RESULTS AND ANALYSIS	18
	2.3 A COMPLETE PROPOSED FRAMEWORK FOR COASTAL WATER QUALITY MONITORING SYSTEM WITH ALGAE PREDICTIVE MODEL	21
	2.3.1 ABSTRACT	21
	2.3.2 INTRODUCTION	22
	2.3.3 RESEARCH FRAMEWORK AND METHOD	26
	2.3.4 CONCLUSION	27

	2.4 MACHINE LEARNING CLASSIFICATION, FEATURE RANKING AND REGRESSION FOR WATER QUALITY PARAMETERS RETRIEVAL IN VARIOUS OPTICAL WATER TYPES FROM HYPER-SPECTRAL OBSERVATIONS	28
	2.4.1 ABSTRACT	28
	2.4.2 INTRODUCTION	29
	2.4.3 RESULTS	31
	2.4.4 CONCLUSIONS AND FUTURE WORK	31
<b>3</b>	<b>METHODOLOGY</b>	<b>33</b>
	3.1 EXISTING SYSTEM	33
	3.1.1 DISADVANTAGES OF EXISTING SYSTEM	33
	3.2 PROPOSED SYSTEM	33
	3.2.1 ADVANTAGE OF PROPOSED SYSTEM	34
	3.3 ARCHITECTURE	34
	3.3.1 COLLECTING DATA	35
	3.3.2 DATA PREPROCESSING	35
	3.3.3 CHOOSING A MODEL	36
	3.3.4 TRAINING THE MODEL	36
	3.3.5 EVALUATING THE MODEL	36
	3.3.6 FLASK FRAMEWORK	37
	3.3.7 HTML	37
	3.4 DATA FLOW DIAGRAM	38
	3.4.1 MODULE NAME	38
	3.4.2 DATASET COLLECTION	38
	3.4.3 PRE-PROCESSING	39
	3.4.4 CHOOSE A MODEL	39
	3.4.5 TRAIN THE MODEL	39
	3.4.6 EVALUATE THE MODEL	39
	3.4.7 PARAMETER TUNING	40
	3.4.8 MAKE PREDICTIONS	40
	3.5 ALGORITHM USED	40
	3.5.1 EXTRA TREES CLASSIFIER	40
	3.6 UML DIAGRAM	41
	3.6.1 USE CASE DIAGRAM	41

	3.6.2 SEQUENCE DIAGRAM	42
	3.6.3 CLASS DIAGRAM	43
<b>4</b>	<b>SYSTEM REQUIREMENTS SPECIFICATION</b>	44
	4.1 FUNCTIONAL AND NON-FUNCTIONAL REQUIREMENTS	44
	4.1.1 FUNCTIONAL REQUIREMENTS	44
	4.1.2 EXAMPLES OF FUNCTIONAL REQUIREMENTS	44
	4.1.3 NON-FUNCTIONAL REQUIREMENTS	44
	4.1.4 EXAMPLES OF NON-FUNCTIONAL REQUIREMENTS	45
	4.2 SYSTEM SPECIFICATIONS	45
	4.2.1 HARDWARE REQUIREMENTS	45
	4.2.2 SOFTWARE REQUIREMENTS	45
<b>5</b>	<b>CONCLUSION AND FUTURE WORK</b>	46
	5.1 CONCLUSION	46
	5.2 SOURCE CODE	46
	5.2.1 APP.PY	46
	5.2.2 MACHINE LEARNING SOURCE CODE	48
	5.2.3 OUTPUT.PY	52
	5.3 SAMPLE OUTPUT	53
	5.2 REFERENCES	55

## LIST OF FIGURE

FIGURE NO.	TITLE	PAGE NO
Figure 3.1	Architecture of water quality prediction	34
Figure 3.2	quality prediction flow diagram	38
Figure 3.3	UML diagram	42
Figure 3.4	Sequence diagram	43
Figure 3.5	Class Diagram	43

# CHAPTER 1

## INTRODUCTION

### 1.1 overview

Water is the principle source for shipping energy to each cell in the body and is additionally the regulator of all body capacities. The cerebrum contains 80% of water. Extreme drying out may prompt mental hindrances and loss of capacity to obviously think. Water is one of the most fundamental regular assets for the endurance of the whole life on this planet. In light of the nature of water, it tends to be utilized for various purposes like drinking, washing, or water system. Plants and creatures likewise rely upon water for their endurance. To put it plainly, all living organic entities need an enormous amount and great nature of water for presence. Freshwater is a fundamental asset to horticulture and industry for its essential presence. Water quality observation is a key stage in the administration of freshwater assets. As indicated by the yearly report of WHO, many individuals are kicking the bucket because of the absence of unadulterated drinking water particularly pregnant ladies and youngsters. It is critical to check the nature of water for its expected reason, whether it be animals watering, compound showering, or drinking water.

Water quality testing is a device that can be utilized to find unadulterated drinking water. Consequently, the right checking of water is incredibly much significant for protecting unadulterated, and clean water. Water testing assumes a key part in breaking down the right activity of water supplies, testing the wellbeing of drinking water, perceiving sickness flare-ups, and approving cycles and precaution measures. Water quality is the proportion of the reasonableness



of water for a specific reason in view of explicit physical, substance, and organic attributes.

Testing the nature of a water body, both surface water, and groundwater, can assist us with responding to inquiries concerning whether the water is satisfactory for drinking, washing, or water system to give some examples of applications. It can utilize the consequences of water quality tests to look at the nature of water starting with one water body and then onto the next in a local, state, or across the entire country. Microbiological quality is for the most part the main pressing concern on the grounds that irresistible infections brought about by pathogenic microorganisms, infections, helminths, and so on are the most well-known and boundless wellbeing risk connected with drinking water. Overabundance amount of certain synthetic substances in drinking water prompts well-being risk. These synthetics incorporate fluoride, arsenic, and nitrate. Safe drinking (consumable) water should be passed on to the client for drinking, food game plan, individual neatness, and washing. The water ought to satisfy the normal quality rules for making it pure at the spot of supply to the clients.

## **1.2 WATER QUALITY INDEX**

WQI is the correlation of the sum with an erratic or logical norm or with a pre-determined base. In this way, the WQI observed and announced natural status and patterns on guidelines quantitatively. A water quality list is a way to sum up a lot of water quality information into straightforward terms (e.g., great) for answering to the board and the general population in a predictable way. Notwithstanding the nonattendance of a universally acknowledged composite file of water quality, a few nations have utilized and are involving collected water quality information in the improvement of water quality lists.

To calculate the water quality index(WQI) conventionally we take 10 features of water to reflect the quality of water like ph, chloride, conductance, etc. In this paper, we use all 10 parameters to calculate the WQI of the water. The general formula to calculate the water quality index.

### **1.3 OBJECTIVE**

The main objective of this research is to develop an Intelligent System to predict water quality using machine learning technique, namely, support vector classification. It is implemented as web-based application in this user answers.

## **CHAPTER 2**

### **LITERATURE SURVEY**

#### **2.1 WATER QUALITY PREDICTION USING MACHINE LEARNING -**

Sai Sreeja Kurra, Sambangi Geethika Naidu, Sravani Chowdala, Sree Chithra Yellanki, Dr. B. Esther Sunanda

##### **2.1.1 ABSTRACT :**

The major goal of this project is to use machine learning techniques to measure water quality. A potability is a numerical phrase that is used to assess the quality of a body of water. The following water quality parameters were utilised to assess the overall water quality in terms of potability in this study. ph, Hardness, Solids, Chloromines, Sulfate, Conductivity, Organic Carbon, Trihalomethanes, Turbidity were the parameters. To depict the water quality, these parameters are used as a feature vector. To estimate the water quality class, the paper used two types of classification algorithms: Decision Tree (DT) and K- Nearest Neighbor (KNN). Experiments were carried out utilising a real dataset containing information from various locations around Andhra Pradesh, as well as a synthetic dataset generated at random using parameters. Based on the results of two different types of classifiers, it was discovered that the KNN classifier outperforms other classifiers. According to the findings, machine learning approaches are capable of accurately predicting the potability. Potability, Water Quality Parameters, Data Mining, and Classification are all index terms.

Keywords: Machine Learning, Supervised Learning, K-Nearest Neighbour (KNN), Decision Tree, Hyper Parameter Tuning, Python Programming.

### **2.1.2 INTRODUCTION:**

Water quality analysis is a complex topic due to the different factors that influence it. This concept is inextricably linked to the various purposes for which water is used. Different needs necessitate different standards. There is a lot of study being done on water quality prediction. Water quality is normally determined by a set of physical and chemical parameters that are closely related to the water's intended usage. The acceptable and unacceptable values for each variable must then be established. Water that meets the predetermined parameters for a specific application is considered appropriate for that application. If the water does not fulfil these requirements, it must be treated before it may be used. Water quality can be assessed using a variety of physical and chemical properties. As a result, studying the behaviour of each individual variable independently is not possible in practise to accurately describe water quality on a spatial or temporal basis. The more challenging method is to combine the values of a group of physical and chemical variables into a single value. A quality value function (usually linear) represented the equivalence between the variable and its quality level was included in the index for each variable. These functions were created using direct measurements of a substance's concentration or the value of a physical variable derived from water sample studies. The major goal of this research is to examine how machine learning algorithms may be used to predict water quality.

### **2.1.3 METHODOLOGY:**

The proposed system is intended to determine potability. It is divided into two phases, one for training and the other for testing. The following procedures are carried out in both sections. Data on training pH and hardness testing data Solids, chloramines, sulphate, conductivity, organic carbon, trihalomethanes,

turbidity, and potability are all terms that can be used to describe something. The data set was chosen as follows: The collection of essential parameters that affect water quality, identification of the number of data samples, and definition of the class labels for each data sample present in the data are all factors that go into selecting the water quality data set, which is a prerequisite to model construction. Ten indicator parameters make up the data sets used in this study. pH value and hardness are examples of these factors. Solids, chloramines, sulfate, conductivity, organic carbon, trihalomethanes, turbidity, and potability are all terms that can be used to describe the properties of a substance. The proposed approach, however, is not constrained by the number of parameters or the selection of parameters. A k-fold cross validation technique is employed to set the learning and testing framework in this study, corresponding to each data sample in the data set. The dataset is separated into k-disjointed sets of equal size, each with roughly the same class distribution, using this technique. This division's subsets are utilised as the test set in turn, with the remaining subsets serving as the training set. These are Decision Tree (DT) and K-Nearest Neighbour (KNN) methods. In terms of the underlying relational structure between the indicator parameters and the class label, each of these strategies takes a different approach. As a result, each technique's performance for the same data set is likely to differ. Validating the performance of different classifiers on an unknown data set: Data mining provides several metrics for validating the performance of different classifiers on an unknown data set. A repeated cross-validation procedure in the MATLAB caret package was used to create the learning and testing environment. The following procedure was used to apply the classification algorithm:

1. The data set was split into two parts: training (80%) and testing (20%).
2. The training set was subjected to repeated cross-validation, with the number of iterations fixed to Classifiers were trained in this manner.

3. The model's optimal parameter configuration was selected, resulting in the maximum accuracy.
4. The model was scrutinized.

#### **2.1.4 CONCLUSION:**

Potability determines the quality of water, which is one of the most important resources for existence. Traditionally, testing water quality required an expensive and time-consuming lab analysis. This study looked into an alternative machine learning method for predicting water quality using only a few simple water quality criteria. To estimate, a set of representative supervised machine learning algorithms was used. It would detect water of bad quality before it was released for consumption and notify the appropriate authorities. It will hopefully reduce the number of individuals who drink low-quality water, lowering the risk of diseases like typhoid and diarrhea. In this case, using a prescriptive analysis based on projected values would result in future capabilities to assist decision and policy makers.

## **2.2 PROPOSITION OF NEW ENSEMBLE DATA-INTELLIGENCE MODELS FOR SURFACE WATER QUALITY PREDICTION**

### **2.2.1 ABSTRACT**

An accurate prediction of water quality (WQ) related parameters is considered as a pivotal decisive tool in sustainable water resources management. In this study, five different ensemble machine learning (ML) models including Quantile regression forest (QRF), Random Forest (RF), radial support vector machine (SVM), Stochastic Gradient Boosting (GBM) and Gradient Boosting Machines (GBM\_H2O) were developed to predict the monthly biochemical oxygen demand (BOD) values of the Euphrates River, Iraq. For this aim, monthly

average data of water temperature (T), Turbidity, pH, Electrical Conductivity (EC), Alkalinity (Alk), Calcium (Ca), chemical oxygen demand (COD), Sulfate (SO<sub>4</sub>), total dissolved solids (TDS), total suspended solids (TSS), and BOD measured for ten years period were used in this study. The performances of these standalone models were compared with integrative models developed by coupling the applied ML models with two different feature extraction algorithms i.e., Genetic Algorithm (GA) and Principal Components Analysis (PCA). The reliability of the applied models was evaluated based on the statistical performance criteria of determination coefficient ( $R^2$ ), root mean square error (RMSE), mean absolute error (MAE), Nash-Sutcliffe model efficiency coefficient (NSE), Willmott index (d), and percent bias (PBIAS). Results showed that among the developed models, QRF model attained the superior performance. The performance of the evaluated models presented in this study proved that the developed integrative PCA-QRF model presented much better performance compared with the standalone ones and with those integrated with GA. The statistical criteria of  $R^2$ , RMSE, MAE, NSE, d, and PBIAS of

### **2.2.2 INTRODUCTION**

**A. THE IMPORTANCE OF SURFACE WATER QUALITY MONITORING AND DETECTION** Human life is significantly reliant on the availability of water because humans depend on water for many activities such as for drinking, cooking, farming, personal hygiene, industrial and manufacturing purposes. Water is also important in other activities like biotransformation, electric power generation, etc. Owing to the reliance of human life on water availability, both surface and groundwater bodies are exposed to various levels of contamination from different contaminants. This has made the prediction of WQ a difficult task in recent times and many scholars have dedicated much effort to WQ assessment due to its importance to human life. A high level of stress has been experienced

over the last two decades in the area of water resources in the Iraqi region due to several reasons, such as the damming of Tigris and Euphrates Rivers, variations in global climate, and the decrease in the local annual rainfall precipitation rates. Water salinity is a critical issue in Iraq that affects WQ for domestic, agricultural, and industrial purposes. Poor drainage and irrigation practices have brought about low water table and soil salinization in the region; agricultural developments and other human activities have affected the quality of water in the Euphrates Basin. However, these impacts are not obvious at the point of water source for irrigation. Therefore, WQ management is necessary for the effective management of all water-related resource.

### MACHINE LEARNING MODELS LITERATURE REVIEW

The need for effective, dependable, accurate, and flexible prediction models has increased recently due to the acknowledgment of the issue of surface water pollution, coupled with the increasing interest in WQ assessment. It is expected that these models can precisely describe the mechanisms of WQ deterioration. Researchers have developed the idea of surface and underground WQ modeling using soft computing tools, such as ML models owing to their reliability and accuracy. However, the ML models demonstrated an inability of the generalization to handle the complicated and highly nonlinear relationship among the modeling parameters. Based on the reported literature (2014-2021), Scopus database indicated that there is a substantial attention on the BOD simulation using the feasibility of ML models. reported the major keywords occurrence clusters and the time span, used over the literature. Over 144 keywords were presented indicating the significant of this topic on modeling river water quality. The idea of the exploration of new ML models that are capable to solve environmental engineering problems is always going on and the research domain of modeling WQ using new sophisticated models are of interest of researchers and scientists. Although the literature revealed different version of ML models applied for surface WQ modeling such as artificial neural network, kernel models, fuzzy



logic, genetic programming, adaptive neuro-inference system models and several others [7]; however, there are several new versions of ML models are yet to be explored for modeling surface WQ phenomena. The efficiency of integrative intelligence models in WQ modeling has also been noted [9, 36-39]. Further, although ML models are the commonly used predictive models in surface WQ prediction, they are still facing several limitations, such as the need to tune their internal parameters, the need for time-consuming algorithms, poor generalization capability, and the need for human intervention during the modeling process. Hence, there is a need for models that are flexible enough to address the complicated nature of most environmental engineering problems.

### C. THE SIGNIFICANT OF THE SELECTED CASE STUDY

The accurate determination of BOD is necessary for water pollution control because it is an important index of good quality water. This parameter is delicate and tedious to analyze, especially BOD analysis. BOD presents the approximation of the biodegradable organic matter in the water and defines an essential indicator for water pollution. In addition, BOD is presented as the foremost parameter for the aquatic system health presentation and its proper quantification can contribute to development of strategic water resources protection and safety. Furthermore, for instance, the DO parameter, the analysis can be adopted in-situ instruments; however, BOD is recorded for at least five days. Accurate prediction of WQ parameters in a study area can save cost, energy, and time; this is why much effort is given to the modeling approaches when predicting these valuable parameters. The modeling approaches are more important in developing countries where the budget for environmental quality assessment and monitoring is low compared to the developed countries. The research is conducted on the base to predict monthly scale BOD for Euphrates river located in Iraq region. Five different ensemble ML models were developed for this purpose. The selection of those models was owing to their massive implementation received and confirming their potential in hydrological, climatological and environmental

researches. The obtained modeling results were compared with several well-established literature on river WQ prediction of diverse region all around the world.

#### D. RESEARCH MOTIVATION AND OBJECTIVES

Several review research articles presented lately on the progress of ML development for river WQ. The literature review emphasis on the exploration of new versions of ML models for modeling river WQ due the drawbacks of the associated limitations with the existed ML models. For instance, classical models such as artificial neural network (ANN), fuzzy logic (FL) and support vector machine (SVM) are associated with the drawbacks on tuning their internal parameters. Another issue reported in the previous studies on the importance selecting the significant and related predictors for the targeted predicted parameters. As the prediction matrix is highly influenced by the input feature selection, integrating a prior approach for the better understanding the predictors effects is an essential step in ML models development. The previous studies have shown an admirable trend for this point of view. For instance, the integration of improved Grey relational analysis (IGRA) algorithm with Long-Short term Memory (LSTM) predictive model, to simulate the DO concentration at the Tai Lake and Victoria Bay. In another study, water quality index (WQI) was predicted using the coupled Gaussian Naïve Bayes and several ML model at Rawal Lake [37]. Recently, some authors tested the capacity of the quantum teaching and learning based optimization as feature selection for WQI determination using weighted extreme learning machine model for groundwater samples collected at the Dharmapuri district in Tamil Nadu. Several other scholars adopted similar methodologies for surface WQ simulation. All those studies confirmed the significant of coupled ML models for modeling surface water quality for better understanding to the substantial correlation between the simulated WA parameters. Hence, the current research was prompted on the base to explore more reliable and robust soft computing predictive models. In addition, the investigation of the highly influential parameters on the prediction of BOD in river located with semi-arid

region. The objectives of the current research are (i) to explore the capacity of five ML models including Quantile regression forest (QRF), Random Forest

### **2.2.3 MODELING RESULTS AND ANALYSIS**

The modeling procedure adopted in this research was exhibited in a form of flowchart presented in Figure. A. Predictors selection In this study, the development of five different ensemble data-intelligence models (i.e., QRF, RF, SVM, GBM and GBM\_H2O) were established for surface water BOD prediction. In addition, the integration of the PCA and GA feature selection approaches was investigated as the second modeling scenario. The models' performances were compared based on multiple statistical criteria including determination coefficient ( $R^2$ ), root mean square error (RMSE), mean absolute error (MAE), Nash-Sutcliffe model efficiency coefficient (NSE), Willmott index (d), and percent bias (PBIAS) [73], [74], and graphical presentation. Owing to the fact that the wise selection of which predictor "water quality parameters" to be included in the prediction formula, it has more advantageous effects on overall performance than the choice of the modeling algorithm itself and thus the feature selection approaches were employed to identify the minimal subset of features for optimal learning. Overall, the applied ML models using the PCA approach performed better than GA approach by producing the lowest prediction error. However, QRF outperformed the GBM\_H2O in terms of the statistical performance metrics.  $R^2$ , RMSE, MAE, NSE, d, and PBIAS of QRF were 0.94, 0.12, 0.05, 0.93, 0.98, and 0.3, respectively. While those of GBM\_H2O were 0.89, 0.16, 0.09, 0.87, 0.97, and 0.3, respectively. The performance from QRF and GBM\_H2O was followed by SVM > GBM > RF. The boxplot of the results obtained from the evaluated the integrative PCA-ML modelling methods during the validation stage were analyzed as shown in Figure 12a. It can be confirmed that the distribution from QRF was the most similar to that from the observed.

The interquartile of the QRF model was almost the closest one to the observed values. Then followed by GBM\_H2O>GBM> SVM>RF. This fact was further confirmed by Taylor diagram (Figure) which prove that the optimal performance was from the QRF model while the RF and SVM were the worst, and the other evaluated methods in between. The correlation coefficient between the QRF and the observed data is greater than 0.95, and the centered pattern RMS difference between the two patterns is ~0.12. The subset selection using the PCA approach out performed that of the benchmark and GA-ML models. FIGURE . a) Boxplot and b) Taylor diagram for the developed PCAML models. v. discussion The redundant and irrelevant predictors significantly deteriorate the performances of regression models and causes overfitting problem in the prediction models. Therefore, extracting a smaller subset of predictors with most relevant predictors might be useful since it saves time in data collection and computation [76], [77]. In this study, two-feature selection were integrative with five different ensemble learning artificial intelligence models (i.e., QRF, RF, GBM\_H2O, GBM, SVM) in order to improve the surface BOD water quality prediction accuracy at the Euphrates River. These two-feature selections can be broadly categorized into filter methods (PCA) and wrapper methods (genetic algorithm) [78]. It was concluded that the performance from PCA outperforms the predictability performances of GA approach and the benchmark models. The GA works by searching the space of possible feature subsets and then evaluating a subset of features using a ML algorithm. This method is known as greedy algorithms owing to the fact that they aim to find the best possible combination of features, which result in the best performant algorithm model. This in turn would be computationally expensive, and impractical in the case of exhaustive search. While in PCA, each predictor is evaluated with a statistical performance metric and then ranked according to its performance indicator. Then after, the top-performing features is selected through the truncation selection before applying a ML models. Hence, the method is considered as a pre-processing step as it

doesn't consider the complex interactions between predictors and are independent of learning algorithms. As mentioned earlier, it is well identified that the PCA method is computationally efficient. However, one shortcoming was pointed when applying this method is being stuck in local optimum when the complex interactions among predictors are ignored. Many researchers argued that wrapper methods (the GA) take into consideration the interaction among predictors but they are not as computationally efficient as filter methods (the PCA) because of the larger space to search. It is well pointed out that the main drawback of applying GA is the necessity to be applied with a higher population size and larger number of generation, which are mostly time consuming. It is prevailed that the optimal features selection returns by GA and the better the network perform in prediction can be attained when there are a large population size and number of generations. Small data set for feature selection may cause the problem of overfitting which is why the performance of GA in this study was not superior in comparison to the baseline models. The combination of PCA with quantile regression forest model outperforms all the applies models in terms of the statistical performances criteria. In QRF model, the conditional quantiles can be inferred which was introduced by Munchausen [86] as a generalization form of random forests. The robustness of QRF method attributed to its non-parametric accurate way of estimating conditional quantiles for high-dimensional predictor parameters. The method is proved to be consistent when applied with multiple different scenarios, suggesting that the algorithm is competitive in terms of predictive power. It is worth to mention that span of the dataset used for the current study provided a satisfactory information for the ML models development and the learning process. It is true that several data span were adopted over the literature; however, in this study, the monthly scale of ten years observations were adequately construct the ML models. The current research modeling is associated with some limitations such as tuning the internal parameters of the SVM model with other advanced non-linear function. In

addition, using metaheuristic optimization algorithms can be another option to enhance the performance of the ML models learning process [88].

**VI. CONCLUSION** This study was proposed five relatively new explored ML models for BOD of surface WQ prediction. These models were considered in this work as a robust approach towards the prediction of WQ parameters rather than relying on laboratory analysis. Further enhancement, two feature selection approaches (GA and PCA) were integrated with the developed ML models to enhance their predictability performance. Various categories of water parameters, including physical, chemical, and biological parameters were used for the development of the proposed models as the input attributes. The data for the model construction was 10 years period laboratory information covering 2004-2013. The outcome of the research showed that PCA-QRF model provided a reliable performance of the BOD prediction compared to the other established models. Furthermore, the proposed model exhibited less approximation of the input parameters that are extremely for the catchments with less environmental or ecological information. Generally, the proposed ML models performed an accurate prediction of the WQ parameters of the Euphrates River. Future studies are aimed at the prediction of other WQ parameters, as well as the inclusion of more input attributes, such as climatological or hydrological factors.

## **2.3 A COMPLETE PROPOSED FRAMEWORK FOR COASTAL WATER QUALITY MONITORING SYSTEM WITH ALGAE PREDICTIVE MODEL**

### **2.3.1 ABSTRACT**

An end-to-end process to achieve a complete framework methodology for Harmful Algal Bloom (HAB) growth prediction is crucial for water management, especially in implementing robust predictive modelling of HAB to

prevent water pollution. Previous works have separately focused on the prediction part or the implementation of the water monitoring system that involves the integration of sensors through the Internet of Things (IoT). These studies lack in terms of discussion of both IoT with the algae ecological domain and prediction method. Therefore, this paper takes the initiative to provide a wider coverage on the end-to-end process including the assembly and integration of sensors, data acquisition and predictive modelling using data-driven approaches, for example, machine learning, deep learning and deep time series forecasting algorithm for future algal bloom outbreak mitigation. This paper believes that discussion in a complete framework perspective based on the execution of each phase is important besides providing a true understanding of the algae growth factors and prediction problems to achieve a robust prediction algorithm for algal growth. In the end, this paper presents proof that selecting the right features

and utilising time series with deep learning are much better for tackling the issues of highly non-linear and dynamic algae ecological data that are briefly introduced in this paper. Among all the algorithms selected, Long Short-term Memory (LSTM) is the best fit for the prediction method and has outperformed other basic machine learning methods in accurately predicting algal growth through the prediction of chlorophyll-a (Chl-a) as a strong indicator of algal presence for coastal studies.

### **2.3.2 INTRODUCTION**

Recently, water resources have been reported to be polluted by the increase of nutrients and minerals that consequently promote excessive algal growth. Harmful algal bloom (HAB) has long been a threat to water sources due to the rapid increase and accumulation of algae population that can cause harm. HAB toxin negatively affects human health, the environment, and the economy whilst non-toxic ones can damage fisheries resources and equipment. Harmful

The associate editor coordinating the review of this manuscript and approving it for publication was Nuno Garcia. toxins such as neurotoxins released by the algae encourage decomposers' growth that increases the biochemical oxygen demand (BOD) of the water. As the algae decompose, oxygen is removed from the water, which then starves the fish and plants of oxygen and damages the local ecology. BOD acts as a measure of the amount of dissolved oxygen (DO) that has been consumed. When BOD is high, less DO will be available for other organisms. This promotes competition for oxygen and finally, causes water pollution known as eutrophication [4]. With the current advancement of the Internet of Things (IoT), the use of various sensors has also increased, which further facilitates the process of monitoring and profiling water quality and eutrophication mitigation. The IoT is a network of physical objects that have evolved into a network of devices such as smartphones, cameras, etc. for homes and vehicles that all are connected, communicating and sharing information [5]. Due to the development of sensors as part of IoT for monitoring, fields such as ecological informatics [6] and bioinformatics have gained various benefits where manual profiling is automated. A sensor is a device that receives a signal (physical, chemical or biological) and converts it into an electric signal output such as current or voltage [7]. Since profiling processes are arduous, time-consuming and lack real-time outcomes to stimulate proactive response to water pollution, the use of sensors is considered a promising alternative for water quality control. To date, the key challenges in the study and management of HABs are species variety, life histories, ecosystems, and the impacts involved. For example, algae communities such as phytoplankton or cyanobacteria that are categorized as potentially harmful do not fit a sole, evolutionarily distinct group [8]. Since algae communities comprise various species and differ in nonlinear ways, they are complex and hard to analyze and are not well understood, resulting in unreliable predictive models [9]. The dynamic growth of algae, which can vary



on short timescales (e.g., hours to days) has made identifying the condition that favours HABs a major research effort.

In algae or HAB prediction, algal count and chlorophyll concentration, especially chlorophyll-a (Chl-a), have been widely used to indicate the presence or growth of algae. Algae concentration can change abruptly where the current chlorophyll content can sometimes increase or decrease up to 5 times than before, causing great difficulties in predicting accurately. Therefore, the prediction of algae remains difficult and unreliable due to the dynamic nature of the time series algae ecological data. Besides, this dynamic nature creates highly nonlinear data which results in randomness issues in model fitting. Randomness issues are rooted in anomalies that have made algal bloom predictions extremely complicated and not well understood. Various research works randomly selected factors for algal growth and depended only on the domain knowledge for feature selection by including all the factors that seemed to be important. This led to model fitting issues and caused fluctuating performance. This paper believes that if the dynamic issues can be tackled accurately, which considers from the data or features level until the algorithm level, all these mentioned strategies might improve the overall prediction performance more. Based on past literature on the prediction method, due to the success of the data-driven prediction method either with

or without considering temporal behaviour [16], researchers used historical data to predict algal blooms by incorporating machine learning technique. Machine learning provides a principled set of mathematical methods for extracting meaningful features from data into distinct and meaningful patterns that can be exploited for decision making, estimation and forecasting. The most applied machine

learning methods include Artificial Neural Network (ANN), Support Vector Machine (SVM), Decision Trees (DT), Random Forest (RF), and regressions. Even though only the input and output of the model are needed for data-driven

models, the prevailing data-driven models, especially those using basic machine learning techniques as mentioned above, are unable to effectively extract features of multi-factor timing data and solve the dynamic issues. Another issue concerns the implementation of the monitoring system. Inspired by the high cost of the commercialized sensor and the dynamic nature of algae that has complicated the prediction process, this article also discusses the development of a solar-powered and low-cost real-time monitoring system to profile the quality of water. Water quality data collected in the data acquisition phase will be used for the development of a predictive model. Our previous work [3] has managed to implement a solar-powered and low-cost water quality monitoring system (WQMS) for coastal studies. However, it was only a preliminary study focusing on simple data analysis of the parameter readings. This paper extends from that previous work and will discuss in detail the modelling phase for the predictive modelling, especially in tackling the dynamic problem of algae ecological data and will briefly mention the enhancement progress of the previous system. Later, the chosen predictive modelling will be applied to the data collected in the previous study [3]. Hence, a complete framework is presented in this paper, which covers the end-to-end process of developing a water monitoring system, deployment, installation, and prediction model development, which was missing in our previous work in this domain. Algae can damage fisheries equipment. This can further affect the algae ecological data with missing values and consequently, reduce the quality of the data. Not limited to algae ecology, in general, time series data are vaguely defined expert knowledge due to the existence of random variables, incomplete and inaccurate data, and approximate estimations rather than measurements, which rendered the understanding of data to remain elusive. which would later lead to missing data collection for the day. This is one

of the main reasons why early algae prediction remains crucial as more time is provided for facilities that use coastal water to shut down before their equipment is damaged. Existing ecological studies, especially those on the algae population are lacking in several aspects. To achieve robust predictive modelling of algal growth, several issues must be highlighted and addressed, for example, (i) the features must

### **2.3.3 RESEARCH FRAMEWORK AND METHODS**

As a revision to our previous work [3], this paper presents an enhanced and more detailed predictive modelling stage that has not been discussed yet. The complete framework is presented in In this stage, the problems of this research topic must be

identified first to find and collect the right data and determine the right method to tackle the problems. Previous studies have revealed a research gap in tackling the issues of highly non-linear, uncertainty and complexity due to the dynamic behaviour of algae aquatic ecosystems. Another gap concerns the way features or the parameters (factors of algal growth) are chosen. This paper opines that the method of selecting the features and the features themselves are important in improving the prediction performance as proven in past research. Furthermore, the dynamic problem itself originates from the features level. Despite only focusing on the algorithm level in tackling the dynamic issues, this paper will further investigate the preparation and designing of the dataset at the features level. This is because features selected in past works were mostly based on domain knowledge or were random [60] where most researchers considered all the features to be important and had no specific feature selection method. To address the issues, this paper has proposed a combination of knowledge based on the literature, and the features are then inspected using the feature selection method at the features level.

Next, at the algorithm level, there is a gap for the coastal dataset where the use of deep learning with time series has the least investigation performed, especially using LSTM. For coastal studies, only one study utilised deep time series using enhanced RNN. To the best of the authors' improved feature selection method in one study has never been applied for coastal studies. The question of whether LSTM can still outperform other algorithms using coastal datasets remains a grey area to be

investigated. Based on these gaps, the proposed method was compared and studied from the literature. After problem formulation, suitable and important parameters were identified using knowledge extracted from the literature and feature selection

method to further inspect and validate the importance of the predictor or features. Several parameters (features) that were reviewed could be categorized as biological factor (BF), physical factor (PF), chemical factor (CF), and meteorological factor (MF). The summary of analysis and categories extracted based on the domain knowledge discussed in past literature is provided in Table 2. After further analysis, features from the dataset were chosen based on these categories. For dataset,

the monthly/biweekly water quality monitoring data gathered by the Hong Kong Environmental Protection Department were utilised for modelling where the data were set up and designed according to certain guidelines. The data col

lection in this paper was based on individual indicators (water parameters) from the most weakly flushed monitoring station,. Nine water parameters or indicators were considered [13] as the target variables. The parameters were Chl-a ( $\mu\text{g/l}$ ), total

### **2.3.4 CONCLUSION**

Due to the issue of dynamics of algae that are highly nonlinear and uncertain, robust predictive modelling that tackles from the end-to-end process is

necessary. Selecting the right features are crucial in tackling the dynamic issues, and from the results, the algae ecology is dependent on the number and types of the features. Based on the discussion and analysis, it was observed that LSTM with the right features outperformed the other methods and grasped the temporal behaviour and tackled the dynamic issues. Besides, even though during this study the MF was excluded, and more CF and PF were included, this study outperformed the other studies. This indicates that the factors are dependent on the characteristics of the data to improve the prediction further. Additionally, this paper has concisely presented a complete framework that discusses in detail both IoT and predictive

modelling that consists of the main phases such as data acquisition, data management and lastly, predictive modelling. Later, the predictive model can be integrated into our system for future HABs prevention. Hence, with the suitable method that has been chosen during the predictive modelling stage, each phase has now been completed, which comprises all the phases in the framework, and overall has achieved all the objectives mentioned. For future work, the LSTM method can be improved further using the hybrid method with other suitable learning methods. Besides, the MF that might further improve the performance can be incorporated. To include MF, the discussion will be big in scope as it will include data segmentation, processing, and feature engineering. Finally, future research should investigate the relation of each feature that could enrich further the explanations.

## **2.4 MACHINE LEARNING CLASSIFICATION, FEATURE RANKING AND REGRESSION FOR WATER QUALITY PARAMETERS RETRIEVAL IN VARIOUS OPTICAL WATER TYPES FROM HYPER-SPECTRAL OBSERVATIONS**

### **2.4.1 ABSTRACT**

This work presents an approach how machine learning techniques can be utilized for upcoming hyper-spectral missions to improve water quality monitoring globally by using a combination of classification and regression methods, and at the same time understanding the spectral relevance for various Water Quality Parameters (WQPs) in different Optical Water Types (OWTs). Machine learning methods are studied to classify OWTs and assign relevance to spectral bands for the WQPs from hyper-spectral observations. The model takes hyper-spectral Level-2 Rrs to classify the data into OWTs by using a non-linear kernel Support Vector Machine (SVM). Then, for selected WQPs feature ranking is performed to assign relevance to the spectral bands in the OWTs. In this study, two WQPs were selected, Chlorophyll-a (Chl-a) content and absorption from Colored Dissolved Organic Matter (aCDOM). The results are presented in spectral relevance maps, showing how spectral relevance varies with increasing optical complexity. These spectral relevance maps are used to select the relevant hyper-spectral bands for a given WQP, which then can be used for regression. Two regression models are evaluated, the kernel Support Vector Regression and Neural Nets.

#### **2.4.2 INTRODUCTION**

Monitoring aquatic environments has been receiving increased focus due to the rapid changes in water quality. These changes might occur as a result of climate change, increasing anthropogenic activity and natural variations in water quality. Water quality can be continuously monitored by optical sensors onboard satellites. There are numerous operational sensors acquiring data about the world's water reservoirs on various spatial and temporal resolution. The spectral resolution usually varies between 3 to 11 spectral bands in the visible (VIS) range of the electromagnetic spectrum for most of these sensors. This spectral resolution often has several limitations with regard to Optical Water Types

(OWTs) and the type of Water Quality Parameters (WQPs). Hyper-spectral imaging might provide the possibility to overcome these limitations. Hyper-spectral sensors have a large number of spectral bands in the VIS and narrow bandwidth. This fine spectral resolution allows to retrieve detailed information about WQPs in different kind of OWTs. The upcoming Plankton, Aerosol, Cloud, ocean Ecosystem (PACE) mission of the National Aeronautics and Space Administration (NASA) will launch a hyper-spectral sensor, hence opening the possibility to have a more comprehensive understanding about aquatic ecosystems and environments. However, to be able to utilize hyper-spectral monitoring for instance PACE, it is required to detect the relevant spectral region for a WQP in interest in the given OWT. This also requires to be able to automatically discriminate between various OWTs, and accurately estimate the WQPs. This work presents a combination of machine learning approaches to map the relevant spectral region for various WQPs from hyper-spectral data. In the first step, classification is used to determine the OWT from above surface Remotely sensed reflectance ( $R_{rs}$ ). Then, feature ranking is applied to assign relevance to the spectral bands for a given WQP. Finally, regression is performed by selecting the most relevant spectral bands. The approach is presented on a synthesized hyper-spectral dataset [1]. In this study, for simplicity the dataset was split into two types of water bodies to train and evaluate the classifier. Then for each type, two WQPs were selected: Chlorophyll-a (Chl-a) concentration and absorption from Colored Dissolved Organic Matter (CDOM), and then feature ranking was applied. The results are presented on spectral relevance map. These maps are used to select the most relevant spectral bands for each WQPs for each OWTs, and use them as input to two machine learning regression methods, namely the kernel Support Vector Regression (SVR) and Neural Nets (NN). The computed statistical measures showed that selecting only the relevant spectral bands can

### **2.4.3 RESULTS**

Classification was carried out by using five-folds cross validation. The Gaussian kernel SVM performed with 98.1 % accuracy by using all the 35 spectral bands. (Note, several machine learning approaches were tested both for classification and feature ranking. In case of classification, it was found that the non-linear kernel SVM is the most suitable classifier. For feature ranking, the Sensitivity Analysis of Gaussian Processes (SA GPR) has also been studied in this work, and the results were consistent with the SA SVR.) The results of the feature ranking for Chl-a and aCDOM in OWT1 and OWT2 can be seen in Fig. 3. For all cases the values of Chl-a and aCDOM were sorted in an increasing order, then feature ranking was done for certain concentration/ absorption. The spectral relevance maps for OWT1 showed that bands centered at shorter wavelengths have relevance for both Chl-a and aCDOM. Since OWT1 represents clear open ocean like optical properties, the spectrum will also be mostly dominated by the optical properties of clear oceanic waters, phytoplankton and phytoplankton associated CDOM. Sea water has low absorption at these shorter wavelengths and high backscatter. The first absorption peak of the Chl-a is centered around 440 nm, and CDOM also absorbs in the shorter

### **2.4.4 CONCLUSIONS AND FUTURE WORK**

In this work, an approach of combining machine learning techniques for classifying OWTs, assigning relevance to spectral bands for WQPs for the given OWT, and estimating WQPs is presented for a synthesized hyper-spectral data. The results show that using the Gaussian kernel SVM classifier can successfully classify OWTs with high accuracy from the measured Rrs. For each OWT feature ranking was used to determine the relevance of spectral bands for two WQPs: Chl-a and aCDOM. For OWT1, the relevant spectral region for Chl-a



was in agreement with aCDOM. This might be due to the strong correlation between CDOM and Chl-a in clear waters, since CDOM in these waters are usually associated with phytoplankton. In case of OWT2, the spectral relevance maps showed deviations in the importance of spectral bands for the two WQPs above a certain domain. This might be a useful finding in the retrieval of biogeochemical parameters in waters with various degree of complexity. Chl-a content estimation is challenging in CDOM dominated turbid waters. Being able to isolate the relevant spectral bands for Chl-a (and CDOM) in these waters, might allow the improved monitoring of WQPs from hyper-spectral data to be acquired by for instance the upcoming PACE mission. Using machine learning for regression resulted in that the SVR and NN models performed almost equally well in the estimation of WQPs for OWT1. This was also valid for OWT2 in case of Chl-a content estimation. This indicates that using the spectral relevance maps, and including spectral bands centered over 700 nm might result in improvements in the estimation of Chl-a content in CDOM rich turbid waters. Estimating aCDOM in OWT2 showed poorer regression performance. However, considering the challenges in CDOM estimation in OWT2, these results might suggest to further investigate the methodology for aCDOM estimation. For future work, this case study will be extended to several OWTs and WQPs for in-situ observations. In addition, the presented models and other machine learning methods will be further studied to improve WQP retrieval by using the relevant spectral bands.

## **CHAPTER 3**

### **METHODOLOGY**

#### **3.1 EXISTING SYSTEM:**

- In Existing medical systems, It required training to operate.
- The structure of a neural network is disparate from the structure of microprocessors therefore required to be emulated.
- It needed high processing time for big neural networks.

##### **3.1.1 DISADVANTAGES OF EXISTING SYSTEM:**

- It required training to operate.
- The structure of a neural network is disparate from the structure of microprocessors therefore required to be emulated.
- It needed high processing time for big neural networks.

#### **3.2 PROPOSED SYSTEM:**

- This project proposed a performance analysis of the famous classification algorithms in the namely Decision Tree , random forest,neural network, multinomial Logistic regression, bagged Tree models, support vector machine, Extra Trees Classifier and using a real dataset retrieved from the water treatment station.

- The experimentation results gave us a good proof of the performances of the classification techniques.
- In addition it's found that best algorithm seems adequate for our water quality monitoring system.
- For further study, we are trying to integrate a new data aggregation algorithm to minimize the amount of the collected data to run the best classification algorithm

### 3.2.1 ADVANTAGE OF PROPOSED SYSTEM:

- It reduces overfitting problem
- It can also handle big data with numerous variables running into thousands.
- learn events and make decisions by commenting on similar events

### 3.3 ARCHITECTURE:

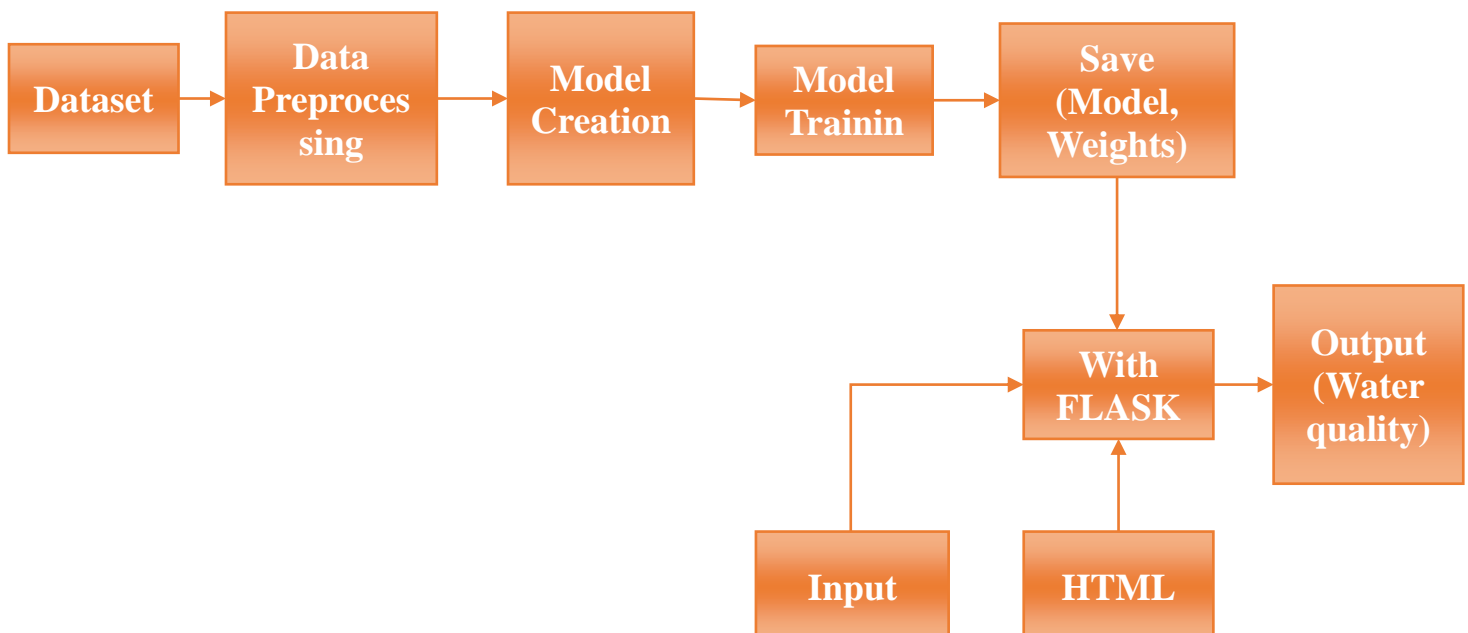


Figure 3.1 Architecture of water quality prediction

### **3.3.1 COLLECTING DATA:**

As you know, machines initially learn from the data that you give them. It is of the utmost importance to collect reliable data so that your machine learning model can find the correct patterns. The quality of the data that you feed to the machine will determine how accurate your model is. If you have incorrect or outdated data, you will have wrong outcomes or predictions which are not relevant.

Make sure you use data from a reliable source, as it will directly affect the outcome of your model. Good data is relevant, contains very few missing and repeated values, and has a good representation of the various subcategories/classes present.

### **3.3.2 DATA PREPROCESSING**

After you have your data, you have to prepare it. You can do this by :

- Putting together all the data you have and randomizing it. This helps make sure that data is evenly distributed, and the ordering does not affect the learning process.
- Cleaning the data to remove unwanted data, missing values, rows, and columns, duplicate values, data type conversion, etc. You might even have to restructure the dataset and change the rows and columns or index of rows and columns.
- Visualize the data to understand how it is structured and understand the relationship between various variables and classes present.

- Splitting the cleaned data into two sets - a training set and a testing set. The training set is the set your model learns from. A testing set is used to check the accuracy of your model after training.

### **3.3.3 CHOOSING A MODEL:**

A machine learning model determines the output you get after running a machine learning algorithm on the collected data. It is important to choose a model which is relevant to the task at hand. Over the years, scientists and engineers developed various models suited for different tasks like speech recognition, image recognition, prediction, etc. Apart from this, you also have to see if your model is suited for numerical or categorical data and choose accordingly.

### **3.3.4 TRAINING THE MODEL:**

Training is the most important step in machine learning. In training, you pass the prepared data to your machine learning model to find patterns and make predictions. It results in the model learning from the data so that it can accomplish the task set. Over time, with training, the model gets better at predicting.

### **3.3.5 EVALUATING THE MODEL:**

After training your model, you have to check to see how it's performing. This is done by testing the performance of the model on previously unseen data. The unseen data used is the testing set that you split our data into earlier. If testing was done on the same data which is used for training, you will not get an accurate measure, as the model is already used to the data, and finds the same patterns in it, as it previously did. This will give you disproportionately high

accuracy. When used on testing data, you get an accurate measure of how your model will perform and its speed.

### **3.3.6 FLASK FRAMEWORK:**

Flask is used for developing web applications using python, implemented on Werkzeug and Jinja2. Advantages of using Flask framework are:

- There is a built-in development server and a fast debugger provided.
- Lightweight
- Secure cookies are supported.
- Templating using Jinja2.
- Request dispatching using REST.
- Support for unit testing is built-in.

### **3.3.7 HTML:**

The Hyper Text Markup Language or HTML is the standard markup language for documents designed to be displayed in a web browser. It is frequently assisted by technologies such as Cascading Style Sheets (CSS) and scripting languages such as JavaScript.

Web browsers receive HTML documents from a web server or from local storage and render the documents into multimedia web pages. HTML describes the structure of a web page semantically and originally included cues for the appearance of the document.

### 3.4 DATA FLOW DIAGRAM:

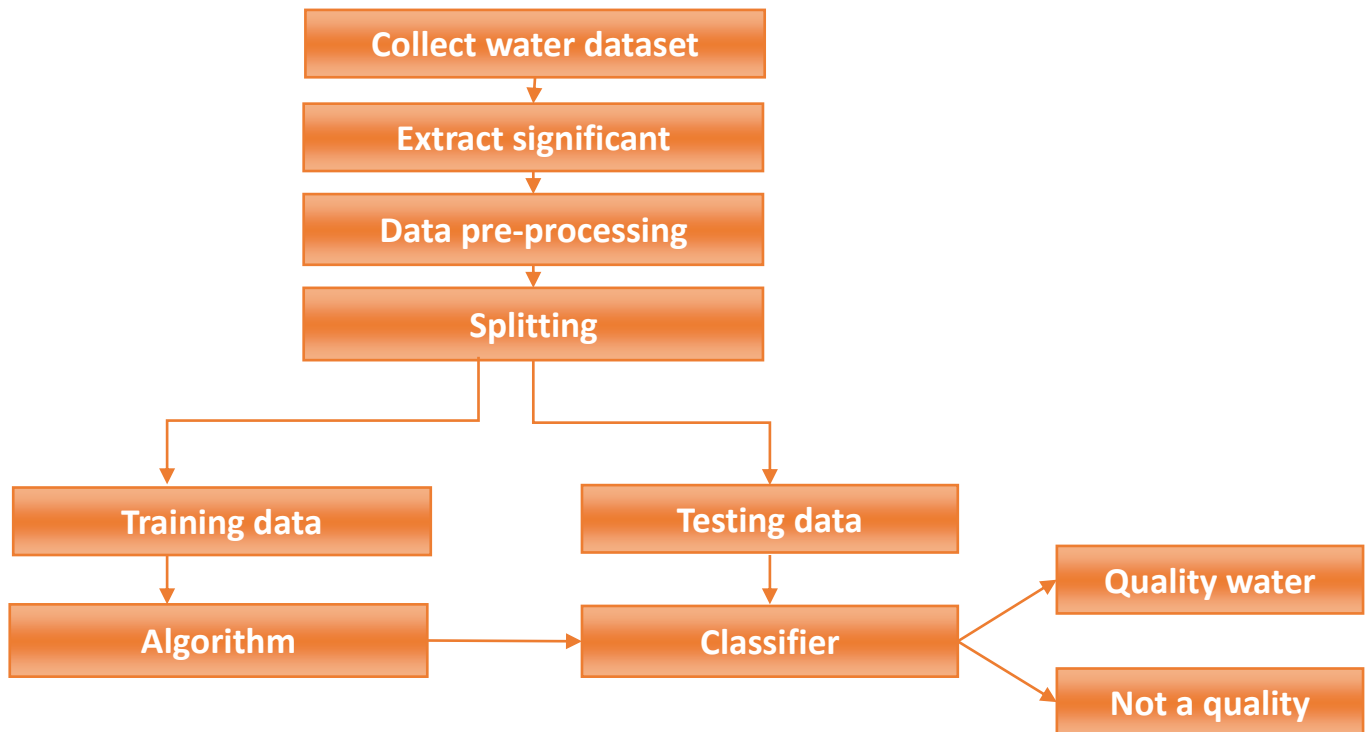


Figure 3.2 water quality prediction flow diagram

#### 3.4.1 MODULE NAME:

- Data collection
- Data pre-processing
- Choose a model
- Train the model
- Evaluate the model
- Parameter tuning
- Make prediction

#### 3.4.2 DATASET COLLECTION:

- The quantity & quality of your data dictate how accurate our model is
- The outcome of this step is generally a representation of data (Guo

simplifies to specifying a table) which we will use for training

- Using pre-collected data, by way of datasets from Kaggle, UCI, etc., still fits into this step

### **3.4.3 PRE-PROCESSING:**

- Wrangle data and prepare it for training
- Clean that which may require it (remove duplicates, correct errors, deal with missing values, normalization, data type conversions, etc.)
- Randomize data, which erases the effects of the particular order in which we collected and/or otherwise prepared our data
- Visualize data to help detect relevant relationships between variables or class imbalances (bias alert!), or perform other exploratory analysis
- Split into training and evaluation sets

### **3.4.4 CHOOSE A MODEL**

- Different algorithms are for different tasks; here we choose the Extra Trees Classifier
- 

### **3.4.5 TRAIN THE MODEL**

- The goal of training is to answer a question or make a prediction correctly as often as possible
- Linear regression example: algorithm would need to learn values for  $m$  (or  $W$ ) and  $b$  ( $x$  is input,  $y$  is output)
- Each iteration of process is a training step

### **3.4.6 EVALUATE THE MODEL**

- Uses some metric or combination of metrics to "measure" objective performance of model



- Test the model against previously unseen data
- This unseen data is meant to be somewhat representative of model performance in the real world, but still helps tune the model (as opposed to test data, which does not)
- Good train/eval split? 80/20, 70/30, or similar, depending on domain, data availability, dataset particulars, etc.
- 

### **3.4.7 PARAMETER TUNING**

- This step refers to hyperparameter tuning, which is an "artform" as opposed to a science
- Tune model parameters for improved performance
- Simple model hyperparameters may include: number of training steps, learning rate, initialization values and distribution, etc.

### **3.4.8 MAKE PREDICTIONS**

Using further (test set) data which have, until this point, been withheld from the model (and for which class labels are known), are used to test the model; a better approximation of how the model will perform in the real world

## **3.5 ALGORITHM USED:**

### **3.5.1 Extra Trees Classifier**

**Extremely Randomized Trees Classifier (Extra Trees Classifier)** is a type of ensemble learning technique which aggregates the results of multiple decorrelated decision trees collected in a “forest” to output its classification

result. In concept, it is very similar to a Random Forest Classifier and only differs from it in the manner of construction of the decision trees in the forest. Each Decision Tree in the Extra Trees Forest is constructed from the original training sample. Then, at each test node, Each tree is provided with a random sample of  $k$  features from the feature-set from which each decision tree must select the best feature to split the data based on some mathematical criteria (typically the Gini Index). This random sample of features leads to the creation of multiple de-correlated decision trees. To perform feature selection using the above forest structure, during the construction of the forest, for each feature, the normalized total reduction in the mathematical criteria used in the decision of feature of split (Gini Index if the Gini Index is used in the construction of the forest) is computed. This value is called the Gini Importance of the feature. To perform feature selection, each feature is ordered in descending order according to the Gini Importance of each feature and the user selects the top  $k$  features according to his/her choice.

### **3.6 UML DIAGRAM**

#### **3.6.1 Use case diagram:**

A use case is a methodology used in system analysis to identify, clarify and organize system requirements. The use case is made up of a set of possible sequences of interactions between systems and users in a particular environment and related to a particular goal. A use case document can help the development team identify and understand where errors may occur during a transaction so they can resolve them.

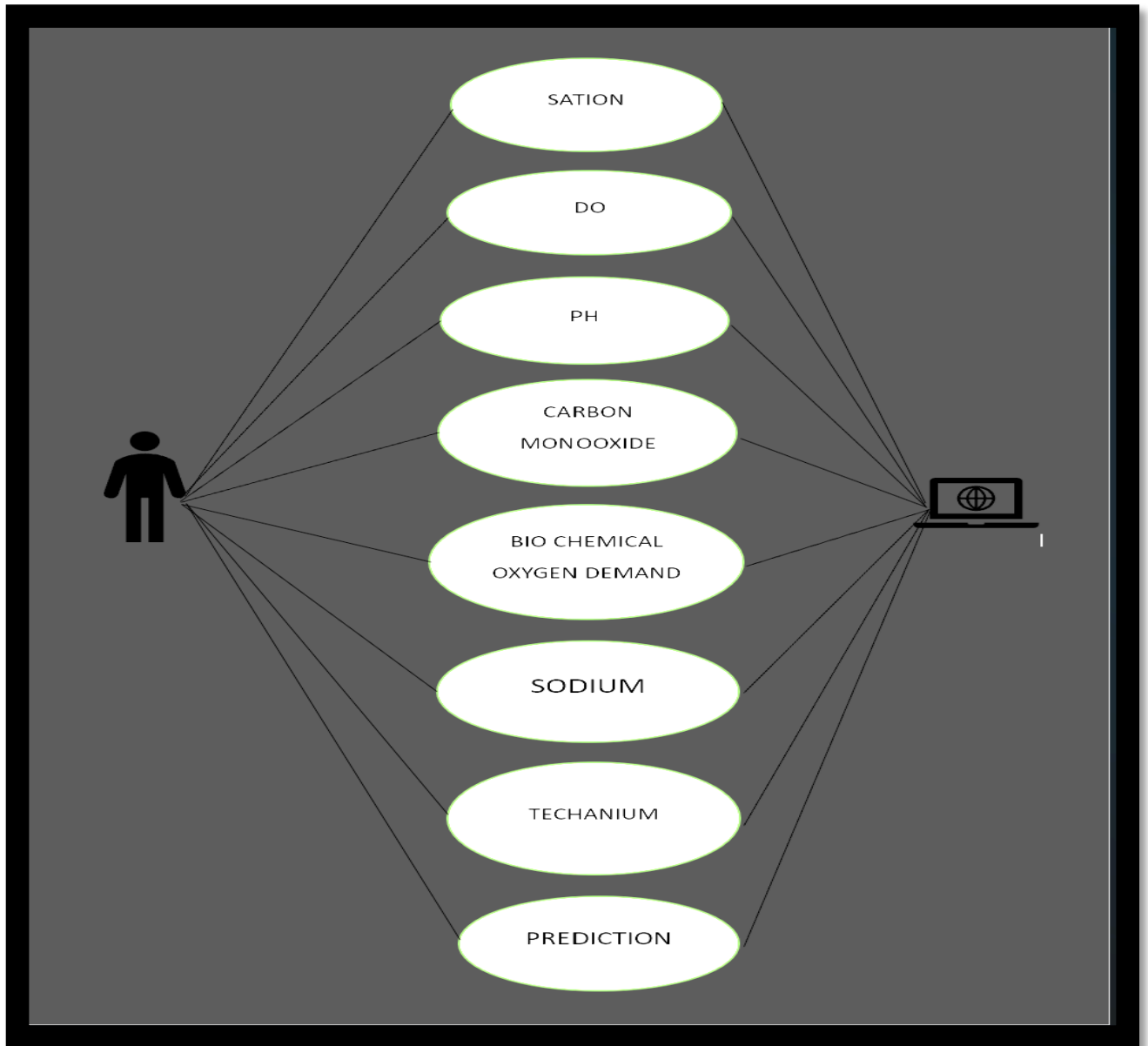


Figure 3.3 UML diagram

### 3.6.2 Sequence diagram:

A sequence diagram simply depicts interaction between objects in a sequential order i.e. the order in which these interactions take place. We can also use the terms event diagrams or event scenarios to refer to a sequence diagram. Sequence diagrams describe how and in what order the objects in a system function. These diagrams are widely used by businessmen and software

developers to document and understand requirements for new and existing systems.

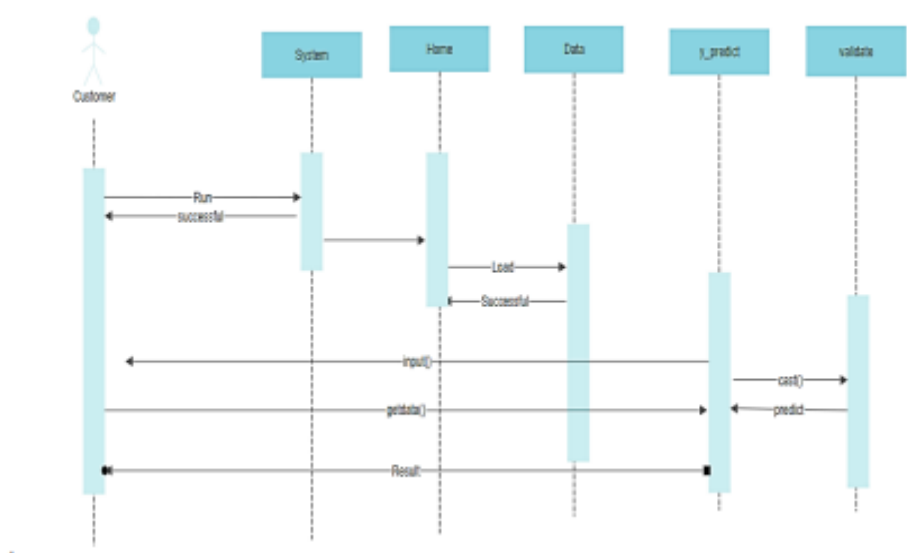


Figure 3.4 Sequence diagram

3.6.3 Class diagram:

Class diagram is a static diagram. It represents the static view of an application. Class diagrams are the only diagrams which can be directly mapped with object oriented languages and thus widely used at the time of construction.

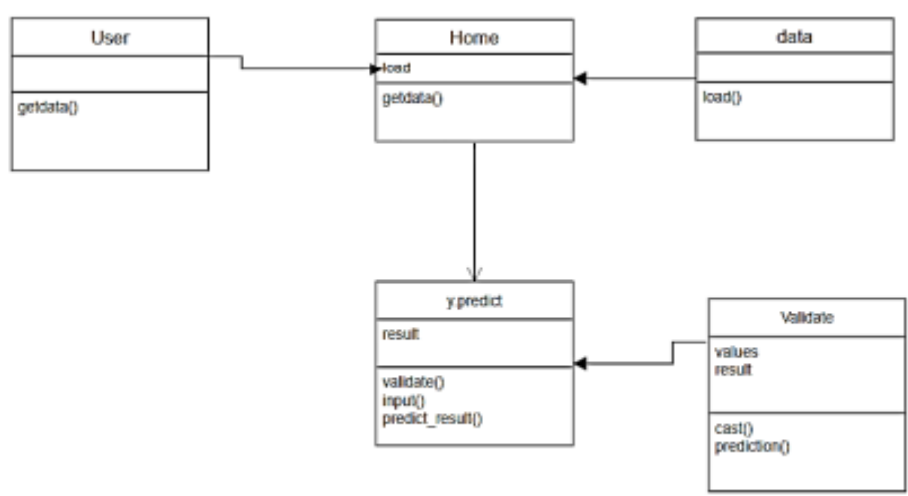


Figure 3.5 Class Diagram

## **CHAPTER 4**

### **SYSTEM REQUIREMENTS SPECIFICATION**

#### **4.1 Functional and non-functional requirements:**

Requirement's analysis is very critical process that enables the success of a system or software project to be assessed. Requirements are generally split into two types: Functional and non-functional requirements.

##### **4.1.1 Functional requirements:**

These are the requirements that the end user specifically demands as basic facilities that the system should offer. All these functionalities need to be necessarily incorporated into the system as a part of the contract. These are represented or stated in the form of Authentication of user whenever he/she logs into the system

- 1) System shutdown in case of a cyber-attack

A verification email is sent to user whenever input to be given to the system, the operation performed and the output expected. They are basically the requirements stated by the user which one can see directly in the final product, unlike the non-functional requirements.

##### **4.1.2 Examples of functional requirements:**

- 2) she register for the first time on some software system.

##### **4.1.3 Non-functional requirements:**

These are basically the quality constraints that the system must satisfy according to the project contract. The priority or extent to which these factors are implemented varies from one project to other. They are also called non-behavioral requirements.

They basically deal with issues like:

- Portability
- Security

- Maintainability
- Reliability
- Scalability
- Performance
- Reusability
- Flexibility

#### **4.1.4 Examples of non-functional requirements:**

- 1) Emails should be sent with a latency of no greater than 12 hours from such an activity.
- 2) The processing of each request should be done within 10 seconds
- 3) The site should load in 3 seconds whenever of simultaneous users are > 10000

## **4.2 SYSTEM SPECIFICATIONS:**

### **4.2.1 HARDWARE REQUIREMENTS**

- System : Pentium i3 Processor.
- Hard Disk : 500 GB.
- Input Devices : Keyboard, Mouse
- Ram : 2 GB

### **4.2.2 SOFTWARE REQUIREMENTS**

- Operating system : Windows 10.
- Coding Language : Python

## CHAPTER 5

### CONCLUSION AND FUTURE WORK

#### 5.1 CONCLUSION

As we all know the importance of water for the human body. So knowing the Quality of the water is very much necessary because if we drink water without knowing that it is safe for drinking we could get sick. There are plenty of water-borne diseases like Cholera, Typhoid, Giardia, E. Coli, Hepatitis A, and so on. These types of diseases happen if we drink non-drinkable water. So knowing the quality of the water is the most important thing. But the main problem lies here. For testing the water quality we have to conduct lab tests on the water which is costly and time-consuming as well. So, in this project, we propose an alternative approach using machine learning algorithms namely, support vector machine to predict water quality.

#### 5.2 SOURCE CODE

##### 5.2.1 App.py

```
# Main Page
```

```
@app.route('/')
```

```
@app.route('/home')
```

```
def home():
```

```
    quotes = ["It is health that is the real wealth, and not pieces of gold and  
silver",
```

```
    "The cheerful mind perseveres, and the strong mind hews its way  
through a thousand difficulties",
```

```
"I have chosen to be happy because it is good for my health",
"A sad soul can be just as lethal as a germ",
"Remain calm, because peace equals power",
"Healthy citizens are the greatest asset any country can have",
"Motivation is what gets you started. Habit is what keeps you going",
"The only bad workout is the one that didn't happen",
"Challenging yourself every day is one of the most exciting ways to live",
"When you feel like quitting, think about why you started",
"The same voice that says 'give up' can also be trained to say 'keep
going' "]
```

```
get_quotes = random.sample(quotes, 3)
```

```
return render_template('home.html', val1=get_quotes[0],
val2=get_quotes[1], val3=get_quotes[2])
```

```
@app.route('/waterpotability',methods=['GET','POST'])
```

```
def waterpotability():
```

```
    if request.method == "POST":
```

```
        to_predict_list = request.form.to_dict()
```

```
        to_predict_list = list(to_predict_list.values())
```

```
        result = output.model8(to_predict_list)
```

```
        if result == 1:
```

```
            body="portable water"
```

```
        elif prediction == 0:
```

```
            body="Non-portable water"
```

```
        else:
```

```
            body="Invalid Data Entered. Please Enter Numeric Data"
```

```
        msg = MIMEMultipart()
```

```
        msg['From'] = 'monikaa.adventure@gmail.com'
```



```

msg['To'] = 'monivijayababu@gmail.com'
msg['Subject'] = 'WATER QUALITY PREDICTION RESULT'

msg.attach(MIMEText(body))

server = smtplib.SMTP('smtp.gmail.com', 587)
server.starttls()
server.login('monikaa.adventure@gmail.com', 'wlcommusiramgkry')
text = msg.as_string()
server.sendmail('monikaa.adventure@gmail.com',
'monivijayababu@gmail.com', text)
server.quit()

return render_template('waterpotability.html',prediction=result)

return render_template('waterpotability.html')

if __name__=='__main__':
    app.run(debug=False)

```

## 5.2.2 MACHINE LEARNING SOURCE CODE

```

import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from imblearn.over_sampling import SMOTE
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.tree import DecisionTreeClassifier
from sklearn.ensemble import RandomForestClassifier, ExtraTreesClassifier,
GradientBoostingClassifier, AdaBoostClassifier
from sklearn.svm import SVC

```

```

from sklearn.neighbors import KNeighborsClassifier
from sklearn.naive_bayes import GaussianNB
from sklearn.metrics import roc_auc_score
import pickle

from sklearn.metrics import classification_report, confusion_matrix
import warnings
warnings.filterwarnings(action="ignore")
plt.style.use(["seaborn-bright", "dark_background"])

data = pd.read_csv("water_potability.csv")
data.head()

data.describe(include="all")

for i in data.columns:
    per = data[i].isnull().sum()/data.shape[0]
    print("Feature { } has { }% data missing".format(i,round(per*100,2)))

mean1 = data["ph"].mean()
mean2 = data["Sulfate"].mean()
mean3 = data["Trihalomethanes"].mean()

data["ph"] = data["ph"].fillna(mean1)
data["Sulfate"] = data["Sulfate"].fillna(mean2)
data["Trihalomethanes"] = data["Trihalomethanes"].fillna(mean3)

plt.figure(figsize=(8,8))
sns.heatmap(data.corr(),annot=True, cmap = "spring")
plt.show()

X = data.drop(columns=["Potability"])
y = data["Potability"]

smote = SMOTE()

```

```

X_sample, y_sample = smote.fit_resample(X, y)

print('Original dataset \n',y.value_counts())
print('Resample dataset \n', y_sample.value_counts())

x_train,          x_test,          y_train,          y_test          =
train_test_split(X_sample,y_sample,test_size=0.15, random_state=101)

model = LogisticRegression()
model.fit(x_train, y_train)
train_pred = model.predict(x_train)
test_pred = model.predict(x_test)

print(classification_report(y_train,train_pred))
print(classification_report(y_test,test_pred))

print(confusion_matrix(y_train,train_pred))
print(confusion_matrix(y_test,test_pred))

models = []
models.append(("LR", LogisticRegression()))
models.append(("DT", DecisionTreeClassifier()))
models.append(("RF", RandomForestClassifier()))
models.append(("ET", ExtraTreesClassifier()))
models.append(("GB", GradientBoostingClassifier()))
models.append(("SVC", SVC()))
models.append(("KNN", KNeighborsClassifier()))
models.append(("GNB", GaussianNB()))

for name, model in models:
    model.fit(x_train, y_train)
    train_pred = model.predict(x_train)
    test_pred = model.predict(x_test)
    print(name)

```

```

print(classification_report(y_train, train_pred))
print(classification_report(y_test, test_pred))

print(confusion_matrix(y_train, train_pred))
print(confusion_matrix(y_test, test_pred))
print("")

model1 = GradientBoostingClassifier()
model2 = DecisionTreeClassifier()
model3 = RandomForestClassifier()
model4 = ExtraTreesClassifier()

model1.fit(x_train,y_train)
model2.fit(x_train,y_train)
model3.fit(x_train,y_train)
model4.fit(x_train,y_train)

pred_prob1 = model1.predict_proba(x_test)
pred_prob2 = model2.predict_proba(x_test)
pred_prob3 = model3.predict_proba(x_test)
pred_prob4 = model4.predict_proba(x_test)
[19]
from sklearn.metrics import roc_curve

fpr1, tpr1, thresh1 = roc_curve(y_test, pred_prob1[:,1], pos_label=1)
fpr2, tpr2, thresh2 = roc_curve(y_test, pred_prob2[:,1], pos_label=1)
fpr3, tpr3, thresh3 = roc_curve(y_test, pred_prob3[:,1], pos_label=1)
fpr4, tpr4, thresh4 = roc_curve(y_test, pred_prob4[:,1], pos_label=1)

random_probs = [0 for i in range(len(y_test))]
p_fpr, p_tpr, _ = roc_curve(y_test, random_probs, pos_label=1)

auc_score1 = roc_auc_score(y_test, pred_prob1[:,1])
auc_score2 = roc_auc_score(y_test, pred_prob2[:,1])

```

```

auc_score3 = roc_auc_score(y_test, pred_prob3[:,1])
auc_score4 = roc_auc_score(y_test, pred_prob4[:,1])

print(auc_score1,",", auc_score2,",", auc_score3,",",auc_score4)

plt.plot(fpr1, tpr1, linestyle='--',color='r', label='Gradient Boosting')
plt.plot(fpr2, tpr2, linestyle='--',color='yellow', label='Decision Tree')
plt.plot(fpr3, tpr3, linestyle='--',color='c', label='Random Forest')
plt.plot(fpr4, tpr4, linestyle='--',color='lime', label='Extra Tree')
plt.plot(p_fpr, p_tpr, linestyle='-', color='blue')
plt.title('ROC curve')
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive rate')
plt.legend(loc='best')
plt.show();

model = ExtraTreesClassifier()
model.fit(x_train, y_train)

pickle_out = open("water_potability.pkl","wb")
pickle.dump(model,pickle_out)
loaded_model = pickle.load(open("water_potability.pkl","rb"))
result = loaded_model.score(x_test,y_test)
print(result)

```

### 5.2.3 OUTPUT.PY

```

import numpy as np
import pickle
def model8(dt):
    try:
        to_predict_list = list(map(float, dt))
        lst = np.array(to_predict_list).reshape(1, 9)

```


```

loaded_model = pickle.load(open("Models/water_potability.pkl", "rb"))
result = loaded_model.predict(1st)
except:
    result = -1
return result

```

### 5.3 SAMPLE OUTPUT

#### TEST CASE 1

 ENTER DETAILS

Ph (0-14)

Hardness (40-350 mg/L)

Solids (300-30000 ppm)

Chloramines (0-15 ppm)

Sulfate (180-500 mg/L)

Conductivity (180-700 µS/cm)

Organic Carbon (2-30 ppm)

Trihalomethanes (8-125 µg/L)


Turbidity (0-8 NTU)

**POTABLE WATER**

Predict

Figure 5.1 Potable Water Output

## TEST CASE II

 **ENTER DETAILS**

Ph (0-14) \_\_\_\_\_

Hardness (40-350 mg/L) \_\_\_\_\_

Solids (300-30000 ppm) \_\_\_\_\_

Chloramines (0-15 ppm) \_\_\_\_\_

Sulfate (180-500 mg/L) \_\_\_\_\_

Conductivity (180-700  $\mu\text{S}/\text{cm}$ ) \_\_\_\_\_

Organic Carbon (2-30 ppm) \_\_\_\_\_

Trihalomethanes (8-125  $\mu\text{g}/\text{L}$ ) \_\_\_\_\_

Turbidity (0-8 NTU) \_\_\_\_\_

**NON POTABLE WATER**

Predict

**Figure 5.2 Non Potable Water Output**

### 5.3 REFERENCES:

- [1] A Gollapalli, Mohammed; Ensemble Machine Learning Model to Predict the Waterborne Syndrome, 2022.
- [2] Umair Ahmed, Rafia Mumtaz, Hirra Anwar, Asad A. Shah, Rabia Irfan and José García-Nieto. Efficient Water Quality Prediction Using Supervised Machine Learning, 2019.
- [3] Ashwini K, D. Diviya, J.Janice Vedha, M. Deva Priya.Intelligent Model For Predicting Water Quality.
- [4] A.N.Prasad, K. A. Mamun, F. R. Islam, H. Haqva.Smart Water Quality Monitoring System, 2015
- [5] Hadi Mohammed, Ibrahim A. Hameed, Razak Seidu.Machine Learning: Based Detection of Water Contamination in Water Distribution systems,2018
- [6] Priya Singh,Pankaj Deep Kaur.Review on Data Mining Techniques for Prediction of Water Quality,2017
- [7] Manish Kumar Jha,Rajni Kumari Sah,M.S.Rashmitha,Rupam Sinha,B.Sujatha.Smart Water Monitoring System for Real-Time Water Quality and Usage Monitoring,2018
- [8] Water QualityMonitoring System using IoT and Machine Learning, in Proceedings of the IEEE International Conference onResearch in Intelligent and Computing in Engineering, pp.1-5, 2018
- [9] Amir Hamzeh Haghiabi, Ali Heidar Nasrolahi, Abbas Parsaie. Water quality prediction using machine learning methods, 2018
- [10] Singh, J.; Yadav, P.; Pal, A.K.; Mishra, V. Water pollutants: Origin and status. In Sensors in Water Pollutants Monitoring: Role of Material; Springer: Berlin/Heidelberg, Germany, 2020
- [11] S. Geetha, S. Gouthami. Internet of Things Enabled Real Time Water Quality Monitoring System ,2017.
- [12] Park, J.; Kim, K.T.; Lee, W.H. Recent advances in information and



Xcommunications technology (ICT) and sensor technology for monitoring water quality. *Water* 2020

[13] Jiang, J.; Tang, S.; Han, D.; Fu, G.; Solomatine, D.; Zheng, Y. A comprehensive review on the design and optimization of surface water quality monitoring networks. *Environ. Model. Softw.* 2020

[14] Kangabam, R.D.; Bhoominathan, S.D.; Kanagaraj, S.; Govindaraju, M. Development of a water quality index (WQI) for the Loktak Lake in India. *Appl. Water Sci.* 2017