EFFICIENT WATER ANALYSIS AND PREDICTION USING MACHINE LEARNING

A PROJECT REPORT

Submitted by

vidhyavarshini D (111619104166)

vunnam mahitha (111619104171)

vallaru sai prudvila chowdary (111619104160)

sangaragu dharani (111619104130)

sandireddy bhavana (111619104129)

in partial fulfillment for the award of the degree

of

BACHELOR OF ENGINEERING

in

COMPUTER SCIENCE AND ENGINEERING

R. M. K. COLLEGE OF ENGINEERING AND TECHNOLOGY

(An Autonomous Institution)

PUDUVOYAL

ANNA UNIVERSITY: CHENNAI 600 025

DEC 2022

ANNA UNIVERSITY : CHENNAI 600 025

BONAFIDE CERTIFICATE

Certified that this project report "**EFFICIENT WATER ANALYSIS AND PREDICTION USING MACHINE LEARNING**" is the bonafide work of "**Vidhyavarshini D,Mahitha V,Prudvilla V,Dharani S,Bhavana S**" who carried out the project work under my supervision.

SIGNATURE	**SIGNATURE**
Dr. D. PAULRAJ	**MR. NARESH SAMMETA**
PROFESSOR AND HEAD	**ASST. PROF & SUPERVISOR**
Department of CSE	Department of CSE
R.M.K College of Engg. and Tech,	R.M.K College of Engg. and Tech
Puduvoyal – 601 026.	Puduvoyal – 601 026.

Submitted for the Project viva-voce held on ___________

INTERNAL EXAMINER	**EXTERNAL EXAMINER**

ACKNOWLEDGEMENT

Support on demand, encouragement at the needed moment and guidance in the right direction are in dispensable for the success of any project. We have received these in excess from all corners from various people, we are glad to submit our gratitude to them.

We thank **Thiru. R. S. Munirathinam**, Founder and Chairman - RMK group of Institutions, and **Thiru. R. M. Kishore**, Vice Chairman - RMK group of Institutions for extending a generous hand in providing the best of resources to the college. **Dr. K. Ramar**, the esteemed Head of our Institution has been a source of motivation to all the staffs and students of our college. We are so much thankful to him.

Our sincere thanks to **Dr. D. Paulraj**, the Head of the Department for his continuous support and motivation throughout our project.

We extend our profound gratitude to our Project Coordinator Mr.REJIN PAULand our Guide Mr. NARESH SAMMETA for their guidance. We thank them for giving us full support to complete the project successfully.

Last, but not the least, we take this opportunity to thank all the staff members of the Department of Computer Science and Engineering. Regards to our family, classmates and friends who offered an unflinching moral support for completion of this project.

TABLE OF CONTENTS

1.INTRODUCTION

1.1 project overview

The deteriorating quality of natural water resources like lakes, streams and estuaries, 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.

1.2 Purpose

Data is obtained from the United States Geological Survey (USGS) online resource called National Water Information System (NWIS). For this paper, the data includes the measurements of 4 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.

2. Literature survey

2.1 Existing problem

Modeling the quality of water resources is vitally important for water scheduling and management. In the past, scientists regularly sampled the water in water quality monitoring stations and assessed the components in the water sample in a lab. However, this process takes a long time, and thus, the detected results are not timely. With the emergence of artificial intelligence (AI) techniques since the last decade, researchers have begun to adopt multivariate linear regression (MLR), artificial neural networks (ANN), adaptive neuro-fuzzy inference system (ANFIS), and Fuzzy time series (FTS) model to predict water quality by exploring the linear and non-linear relationships residing in water quality datasets. In addition, the wavelet denoising method and intelligent algorithms are also proposed to combine with machine learning techniques to enhance the prediction accuracy. In the following, we will review these related work in four categories of machine learning methods.

1. MLR

MLR is a kind of statistical analysis method which is used to estimate the target value based on given values collected from a set of independent variables. It is adopted to predict the water quality because of its speed and simplicity. In [3], the MLR model is used to predict biochemical oxygen demand (BOD) and chemical oxygen demand base on four independent variables, temperature, pH, total suspended solid, and total suspended. The system quickly receives relatively good result in BOD prediction with a correlation coefficient value of 0.5. MLR model has also been used to predict the water quality index in [10] and found to be reliable in formulating the relationship excluding the parameter chloride. However, the MLR model can only be used to formulate linear relationship. It is likely to have a large prediction error if the
MLR model is used to predict non-linear relationship. .

2. ANN

Various ANN models have been designed to predict water and wastewater discharge quality based on previous existing datasets. A two-layer ANN model has been applied to predict the DO concentration in the Mathura River [11], and the experimental result showed that the ANN model worked well. In [12], various neural network types are compared in predicting water temperatures in streams. A radial basis function neural network has also been proposed to describe the water quality parameters in [13]. The summary of the experiment result shows the model outperforms the linear regression model in conductivity, turbidity, and total dissolved solids prediction. A time series prediction model, namely the autoregressive integrated moving average, was integrated with the ANN model to improve the prediction performance. The experimental results showed that the hybrid model provided better accuracy than ARIMA and ANN models [14]. Additionally, a comprehensive comparison between ANN and MLR models in biochemical oxygen demand and chemical oxygen demand prediction has been performed [3]. The experimental results show that a three-layer neural network model outperforms an MLR model. The other comparison between ANN and MLR models in water quality index prediction furtherly proves that the ANN model is a better option [10]. Although ANN models can effectively improve the prediction accuracy of water quality parameters, shortcomings still exist. Especially in some scenarios where the input parameters are ambiguous, neural networks struggle to formulate a non-linear relationship. In [15], wavelet transformation was applied to the ANN model to improve the prediction accuracy of a variety of ocean water quality parameters. An integration of a particle swarm optimization algorithm with ANN models has also been investigated to improve the forecasting performance [16]. In [17], 120 data samples, collected from 2002 to 2012, are used to verify whether the integration of fuzzy logic and ANN models can improve the water quality prediction performance. The experimental results confirm that the proposed method works.

3 .ANFIS

Many studies have proven that ANFIS, which can integrate linear and non-linear relationships hidden in the dataset, is a better option in this scenario [5]. The experimental results in [6] show that an ANFIS model works much better than an ANN model in predicting dissolved oxygen, even though there are only 45 data samples available. An ANFIS model with eight input parameters is used to predict total phosphorus and total nitrogen, the experiment result based on 120 water samples shows the proposed model is reliable [18]. The ANFIS model has also been applied to estimate the biochemical oxygen demand in the Surma River [19]. The testing results from 36 water samples confirmed that the ANFIS model could accurately formulate the hidden relationship and correlation analysis can improve the prediction accuracy. Two different kinds of ANFIS model, fuzzy c-means and subtractive clustering-based was compared in [20], the experiment result shows the ANFIS model built by fuzzy c-means provides more accurate prediction result. In [21], the ensemble models of wavelet ANNs are found to be superior to the best single model for forecasting chlorophyll and salinity concentrations in coastal water. An ensemble of ANN and ANFIS is proposed in [22] to improve the prediction performance of the ANN and ANFIS model, the test result shows there is a significant improvement in the Ensemble ANN-ANFIS model. According to the developer of the ANFIS model, the size of the training dataset should be no less than the number of training parameters [23]. In the aforementioned papers, though the ANFIS models have received higher prediction accuracy, the sizes of the training datasets are data have a large value range and there exist some extreme data value points, an out-of-range error is likely to occur, which happens when the testing dataset cannot find any insight from the training model. A few out-of-range errors can cause a very large prediction error, even though the model can accurately predict most of the data samples. In [24], a dataset collected from 122 wells in Mashhad plain (Iran) is used to investigate the performance of ANFIS, ANN, and geostatistical models in groundwater quality prediction. The experimental result shows that the ANFIS model has poor performance in the testing stage because the limited training dataset cannot build a robust or reliable model. Recently, a few researchers have tried to integrate a machine learning model with a wavelet de-noising technique to improve prediction accuracy. Wavelet support vector regression and wavelet artificial neural networks have been proposed to model monthly pan evaporation [25]. The experimental results confirm that wavelet artificial neural networks outperform other models. An integrated wavelet de-noising ANFIS model was proposed to predict electrical conductivity (EC) and total dissolved solids (TDS) in [26]. Although the size of the dataset was smaller than the requirement, the model still achieved good prediction performance. In [27], eight different kinds of membership functions, with different wavelet de-noising schemes, were investigated to improve the performance of an ANFIS model. Based on the above two research studies, an optimized wavelet-ANFIS model is proposed in [28] and the experimental results show that a bell-shaped membership function with random sampling has the best prediction performance. In [29], a wavelet-ANFIS model is proposed to predict the groundwater level. Compared to ARIMA and ANFIS, the proposed model provides a more precise prediction result. A comparative study of different wavelet-based ANN models to predict sewage sludge quantity is given in [30], the experiment result also proves wavelet can improve the accuracy to the ANN models. On the other hand, many researchers have also tried to integrate intelligence algorithms with the ANFIS model to improve the performance of the proposed model. An application of genetic algorithm (GA), ant colony optimization for continuous domains, and differential evolution is introduced in [31] to improve the performance of the ANFIS model in predicting parameter electrical conductivity, sodium absorption ratio, and total hardness. The experiment result confirms that the proposed model can improve the performance of the ANFIS model for predicting EC and pH and the root mean square error (RMSE) value of the proposed model in the testing stage is 73.03 and 49.55, respectively. In [32], the genetic algorithm and particle swarm optimization (PSO) algorithm are integrated with the ANFIS model to optimize the threshold bank profile prediction. This method is also used in precipitation modeling. The experimental result indicates that the integrated ANFIS models with hybrid GA/PSO achieve better accuracy than the simple ANFIS model [33].

4 FTS

A water quality data is a kind of time series dataset which is likely to have complicated linear and nonlinear relationships. The Fuzzy time series (FTS) model was first proposed by Song and Chissom in 1993 to address an enrollment prediction problem [34]. Chen improved this model by replacing complicated max-min composition operations with simplified arithmetic operations [35]. In [8], a Heuristic Gaussian cloud transformation was integrated with an FTS model to forecast water quality. The experimental results showed that the proposed model significantly improved the prediction accuracy. However, there were only 520 water quality samples available to build the cloud, and thus, the model was not reliable or robust. Time series analysis is alsoproposed to address dissolve oxygen prediction, and the experimental results show that the proposed analysis method can find out valuable knowledge from water quality historical timeseries data [36]. In this dissertation, MLR, ANN, ANFIS, and FTS models are integrated with statistical analysis, wavelet denoising, and intelligence algorithm to explore the prediction of water quality.

2.2 REFRENCES

[1] L. Metcalf and H. P. Eddy, Wastewater Engineering: Treatment and Resource Recovery, 5th Edition.

[2]McGraw Hill, Chapter 3, 2014. Journal of Water, vol. 8, no. 2, pp. 1-15, Jan. 2016

[3] A. H. Zare, "Evaluation of multivariate linear regression and artificial neural networks in prediction of water quality parameters," Journal of Environmental Health Science & Engineering, vol. 12, no. 1, pp. 1-8, Jan. 2014.

[4] L. Li, P. Jiang, H. Guang, L. Dong, G. Wu, and H. Wu " Water quality prediction based on recurrent neural network and improved evidence theory: a case study of Qiantang River, China," Environmental Science and Pollution Research, vol. 26, no. 19, pp. 19879-19896, Mar. 2019.

[5] J.-S. R. Jang, "ANFIS: Adaptive-Network-Based Fuzzy Inference Systems," IEEE Transactions on Systems, Man, and Cybernetics, vol. 23, no. 3, pp. 665-685, May. 1993.

[6] A. Najah, A. El-Shafie, O. A. Karim, and A. H. El-Shafie, "Performance of ANFIS versus MLP-NN dissolved oxygen prediction models in water quality monitoring," Environmental Science and Pollution Research, vol. 21, no. 3, pp. 1658 1670, Aug. 2013.

[7] J. Wan, M. Huang, Y. Ma, W. Guo, Y. Wang, H. Zhang, W. Li, and X. Sun, "Prediction of effluent quality of a paper mill wastewater treatment using an adaptive network-based fuzzy inference system," Applied Soft Computing, vol. 11, no. 3, pp. 3238-3246, Apr. 2011.
[8] W. Deng, G. Wang, and X. Zhang, "A novel hybrid water quality time series prediction method based on cloud model and fuzzy forecasting," Chemometrics and Intelligent Laboratory Systems, vol. 149, pp. 39-49, Dec. 2015.
[9] P. Singh, "Rainfall and financial forecasting using fuzzy time series and neural networks based model," International Journal of Machine Learning and Cybernetics,vol. 9, no. 3, pp. 491- 506, May. 2018.
[10] A. K. Kadam, V. M. Wagh, A. A. Muley, B. N. Umrikar, and R. N. Sankhua, "Prediction of water quality index using artificial neural network and multiple linear regression modelling approach in Shivganga River basin, India," Modeling Earth Systems and Environment, vol. 5, no. 3, pp. 951-96, Mar. 2019.
[11] A. Sarkar and P. Pandey, "River Water Quality Modelling Using Artificial Neural Network Technique," Aquatic Procedia, vol. 4, pp. 1070-1077, 2015.
[12] A. P. Piotrowski, M. J. Napiorkowski, J. J. Napiorkowski, and M. Osuch, "Comparing various artificial neural network types for water temperature prediction in rivers," Journal of Hydrology, vol. 529, no. 1, pp. 302-315, Oct. 2015.
[13] A. Najah, A. El-Shafie, O. A. Karim, and A. H. El-Shafie, "Application of artificial neural networks for water quality prediction," Neural Computing and Applications, vol. 22, no. 1, pp. 187-201, Apr. 2012.
[14] L. Zhang, G. X. Zhang, and R. R. Li, "Water quality analysis and prediction using hybrid time series and neural network models," Journal of Agricultural Science and Technology, vol. 18, no. 4, pp. 975-983, Dec. 2015.
[15] M. J. Alizadeh and M. R. Kavianpour, "Development of wavelet-ANN models to predict water quality parameters in Hilo Bay, Pacific Ocean," Marine Pollution Bulletin, vol. 98, no. 1, pp. 171-178, 2015.
[16] A. G. Mohammad, K. Reza, C. Kwok-Wing, S. Shahaboddin, and T. G. Pezhman,
[17] A. A. Nadiri, S. Shokri, F. T.Tsai, and A. A. Moghaddam, "Prediction of effluent quality parameters of a wastewater treatment plant using a supervised committee fuzzy logic model," Journal of Cleaner Production, vol. 180, no. 7, pp. 539-549, Jan. 2018.
[18] M. Khadr and M. Elshemy, "Data-driven modeling for water quality prediction case study: The drains system associated with Manzala Lake, Egypt," Ain Shams Engineering Journal, pp. 1-9, Sep. 2016.
[19] 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, Jul. 2017.

[20] S. Tiwari, R. Babbar, and G. Kaur, "Performance evaluation of two ANFIS models for predicting water quality Index of River Satluj (India)," Advances in Civil Engineering, vol. 2018, pp. 1-10, Mar. 2018.
[21] S. Shahaboddin, J. N. Ehsan, E. A. Jason, A. M. Azizah, M. Amir, and C. Kwok-wing, -day ahead forecasting of chlorophyll-a Engineering Applications of Computational Fluid Mechanics, vol. 13, no. 1, pp. 91-101, Dec. 2018.
[22] Y. Khan and S. Chai, "Ensemble of ANN and ANFIS for water quality prediction and analysis - a data driven approach," Journal of Telecommunication, Electronic and Computer Engineering , vol. 9, no. 2, pp. 117-122, 2017.
[23] J.-S. R. Jang, "Frequently Asked Questions - ANFIS in the Fuzzy Logic Toolbox," [Online]. Available: http://www.cs.nthu.edu.tw/~jang/anfisfaq.htm.
[24] A. Khashei-Siuki and M. Sarbazi, "Evaluation of ANFIS, ANN, and geostatistical models to spatial distribution of groundwater quality," Arabian Journal of Geosciences, vol. 8, no. 2, pp. 903-912, Nov. 2013.
[25] N. Q. Sultan, S. Saeed, K. Salar, J. Salar, K. Ozgur, and S. monthly pan evaporation using wavelet support vector regression and wavelet artificial neural Engineering Applications of Computational Fluid Mechanics, vol. 13, no. 1, pp. 177-187, Jan. 2019.
[26] A. A. Najah, A. El-Shafie, O. A. Karim and O. Jaafar, "Water quality prediction model utilizing integrated wavelet-ANFIS model with cross-validation," Neural Computing and Applications, vol. 21, no. 5, pp. 833-841, Nov. 2010.
[27] R. Barzegar, J. Adamowski, and A. A. Moghaddam, "Application of wavelet-artificial intelligence hybrid models for water quality prediction: a case study in Aji-Chay River, Iran," Stochastic Environmental Research and Risk Assessment, vol. 30, no. 7, pp. 1797 1819, Jan. 2016.
[28] Z. Fu, J. Cheng, M. Yang, and J. Batista, "Prediction of industrial wastewater quality parameters based on wavelet-ANFIS model," in Proc. IEEE 8th Annual Computing and Communication Workshop and Conference (CCWC), Las Vegas, 2018.
[30] -based neural Environmental Monitoring and Assessment, vol. 191, no. 3, pp. 163-191, Feb. 2019. 78
[31] A. Azad, H. Karami, S. Farzin, A. Saeedian, H. Kashi, and F. Sayyahi, "Prediction of Water Quality Parameters Using ANFIS Optimized by Intelligence Algorithms (Case Study: Gorganrood River)," KSCE Journal of Civil Engineering, vol. 22, no. 7, pp. 2206-2213, Sep. 2017.
[32] A. Gholami, H. Bonakdari, I. Ebtehaj, M. Mohammadian, B. Gharabaghi, and S. R. swarm optimization with ANFIS to predict threshold bank profile shape based on digital laser Journal of the International Measurement Confederation, vol. 121, pp. 294- 303, Jun. 2018.
[33] A. Azad, M. ManOochehri, H. Kashi, S. Farzin, H. Karami, V. Nourani, and J. Shirie, -fuzzy inference Journal of Hydrology, vol. 571 pp. 214-224, Apr. 2019. [34] Q.

Song and B. S. Chissom, "Fuzzy time series and its models," Fuzzy Sets and Systems, vol. 54, no. 3, pp. 269-277, Mar. 1993.
[35] S.-M. Chen, "Forecasting enrollments based on fuzzy time series," Fuzzy Sets and Systems, vol. 81, no. 3, pp. 311-319, Aug. 1996.
[36]International Journal of Nonlinear Science, vol. 17, no.3, pp. 234-240, Dec. 2013.

2.3 Problem statement definition

Water is considered as a vital resource that affects various aspects of human health and lives. The quality of water is a major concern for people living in urban areas. The quality of water serves as a powerful environmental determinant and a foundation for the prevention and control of waterborne diseases. However predicting the urban water quality is a challenging task since the water quality varies in urban spaces non-linearly and depends on multiple factors, such as meteorology, water usage patterns, and land uses, so this project aims at building a Machine Learning (ML) model to Predict Water Quality by considering all water quality standard indicators.

EMPATHY MAP

SAYS

I EXPECT DIFFERENT

WHERE SHOULD I START?

I NEED SOMETHING RELIABLE?

THINKS

WHAT IS TOO HARD

WHAT ELSE I AM MISSING

USERS

DOES

TESTING DATA

COLLECT DATA

FEELS

FEAR

EXICTED

3.2 IDEATION AND BRAINSTORMING

The deteriorating quality of natural water resources like lakes, streams and estuaries, 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 United States Geological Survey (USGS) online resource called National Water Information System (NWIS). For this paper, the data includes the measurements of 4 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.

3.IDEATION & PROPOSED SOLUTION

3.3 PROPOSED SOLUTION

s.no	Parameter	review
1	Problem statement	Efficient water analysis using machine learning
2	idea	Water is considered as a vital resource that affects various aspects of human health and lives. The quality of water is a major concern for people living in urban areas. This project aims at building a Machine Learning (ML) model to Predict Water Quality by considering all water quality standard indicators.
3	novelty	One of the biggest advantages of using deep learning approach is **its ability to execute feature engineering by itself**. In this approach, an algorithm scans the data to identify features which correlate and then combine them to

		promote faster learning without being told to do so explicitly.
4	Social impact	ML **helps to predict demand better and can be cutting-edge technology for supply change management**. It helps in accurate market segregation and plans marketing strategies accordingly this surely improves ROI on marketing budget
5	Machine learning	**the capability of a machine to imitate intelligent human behavior**. Artificial intelligence systems are used to perform complex tasks in a way that is similar to how humans solve problems.
6	Scalability of solution	The latest machine learning approach has shown promising **predictive accuracy** for water quality.

Project design phase 1`

Customer segment

The aim of this study is the prediction of water quality component Artificial intelligence(AI) techniques

Problem

1. Understanding Which Processes Need Automation
2. Lack of Quality Data.
3. Inadequate Infrastructure
4. Implementation.
5. Lack of Skilled Resources.

TRIGGERS TO ACT

The supervised learning, which trains a model on known input and output data so that it can predict future outputs, and unsupervised learning, which finds hidden patterns or intrinsic structures in input data.

Available solutions

pros

It is automatic,used in various fields

cons

Data Acquisition

Time and Resources

Your solution

The solution is based on supervised learning algorithms and to analysis and predict eficiency of water using the given data sets.

4. REQUIREMENT ANALYSIS

SOLUTION REQUIREMENTS (FUNCTIONAL & NON FUNCTIONAL)

4.1 FUNCTIONAL REQUIREMENTS

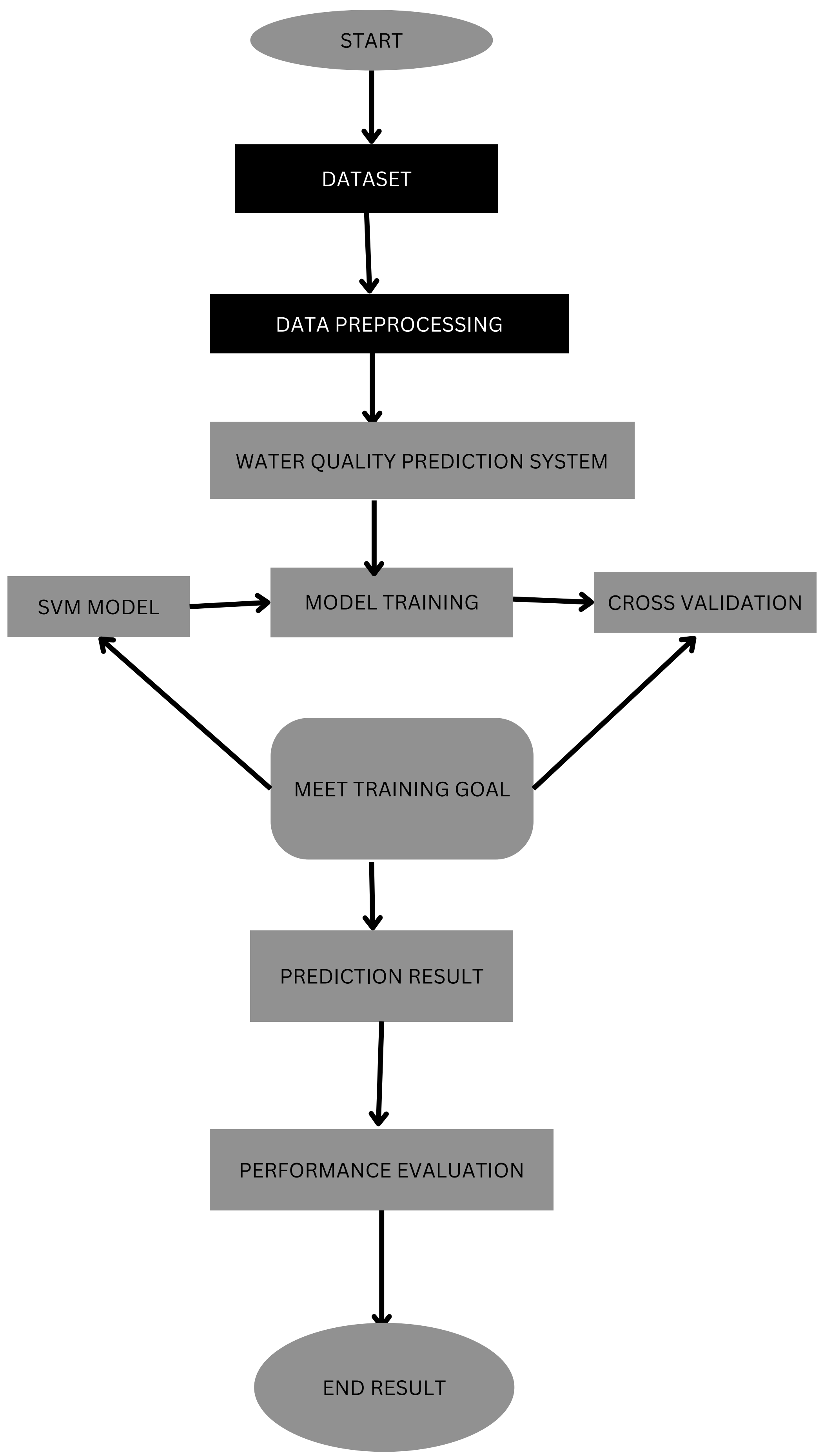
Following are the functional requirements of proposed solution

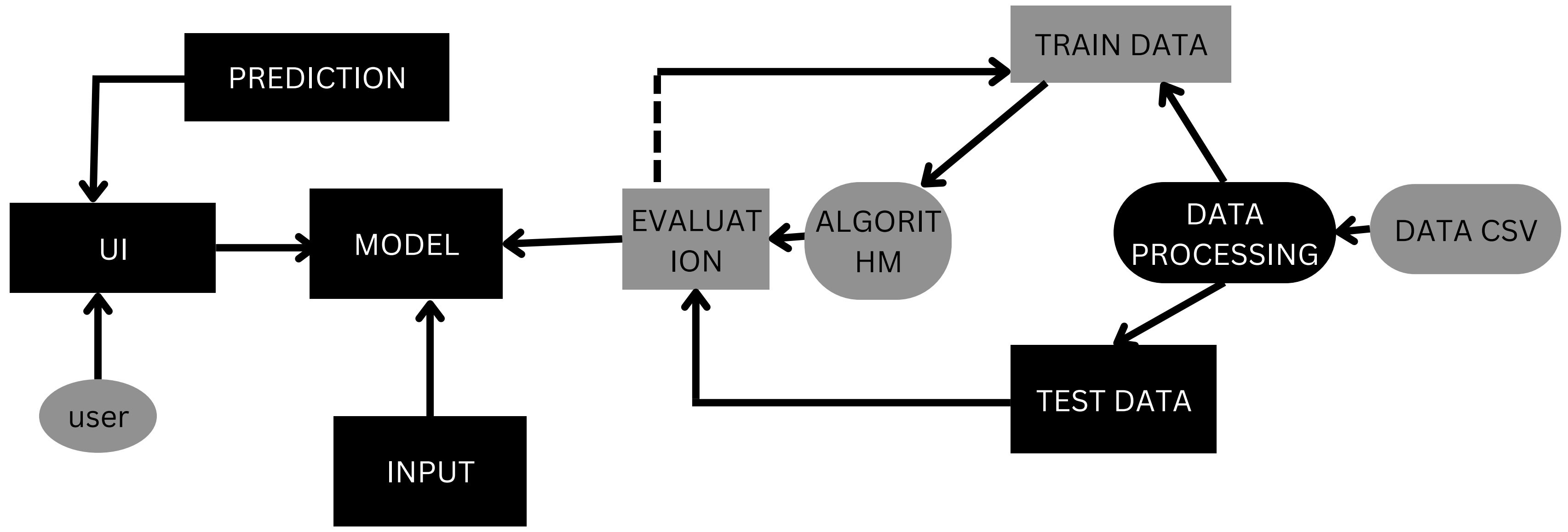
FR .No.	Functional requirement	Description
FR-1	External interface	**The supported set of interactions and behavioural properties of a (provided or required) service. The behavioural description is more abstract than the one of component-interfaces (e.g., it does not include information on the local state)**
FR-2	Authentication	Authentication technology provides access control for systems by checking to see if a user's credentials match the credentials in a database of authorized users or in a data authentication server.
FR-3	Authorization	Authorization is the process of granting someone to do something. It means it a way to check if the user has permission to use a resource or not. It defines that what data and information one user can access.
FR-4	Business rules	A business rule is, at the most basic level,a specific directive that constraint or defines the activities of business.

4.2 NON - FUNCTIONAL REQUIREMENTS

Following are non functional requirements of proposed solution.

NFR. No	Non-functional requirements	Description
NFR-1	Usability	Usability is the degree of ease with which products such as software and web applications can be used to achieve required goals effectively and efficiently.
NFR-2	Security	Machine learning can be applied in various ways in security, for instance,in malware analysis to make prediction and clustering events.
NFR-3	Reliability	Reliability is application of data analytics include ml to predict when asset will fail so that it can be serviced or replaced before failing.
NFR-4	Performance	Model performance is an assessment of the model's ability to perform a task accurately not only with training data but also in real-time with runtime data when the model is actually deployed through a website or an app.


START
DATASET
DATA PREPROCESSING
WATER QUALITY PREDICTION SYSTEM
SVM MODEL
MODEL TRAINING
CROSS VALIDATION
MEET TRAINING GOAL
PREDICTION RESULT
PERFORMANCE EVALUATION
END RESULT


PREDICTION
UI
MODEL
user
INPUT
EVALUAT
ION
ALGORIT
HM
TRAIN DATA
DATA
PROCESSING
DATA CSV
TEST DATA

5.3 USER STORIES

sprint	User story number	User story/task	Story points	Priority and team mebers
Sprint 1	USN-1	As a user , i can able to understand ml.	2	high
Sprint 3	USN-2	As a user, i can able to think extra about AI.	1	Low
Sprint 2	USN-3	As a user, i can able to have clear idea.	1	medium

Sprint 4	USN-4	As a user, i can grasp whole content easily	2	high

6. PROJECT PLANNING & SCHEDULING

6.1 Sprint planning and estimation

phase	**week1**	**week2**	**week3**	**Week 4**	**week5**
Ideation phase	29aug-3sep	4sep-10sep	11sep-17-sep		
Project design phase 1	19sept-25sep	26sep-1oct			
Project design phase 2	3oct-9oct	10-15oct			

Project planning phase	17oct-22oct				
Project development phase	24oct-30oct	1-7nov	8-14nov	15-19nov	

6.2 Sprint delivery schedule

sprint	User story number	User story/task	Story points	Priority and team mebers
Sprint 1	USN-1	As a user , i can able to understand ml.	2	high
Sprint 3	USN-2	As a user, i can able to think extra about AI.	1	Low
Sprint 2	USN-3	As a user, i can able to have clear idea.	1	medium

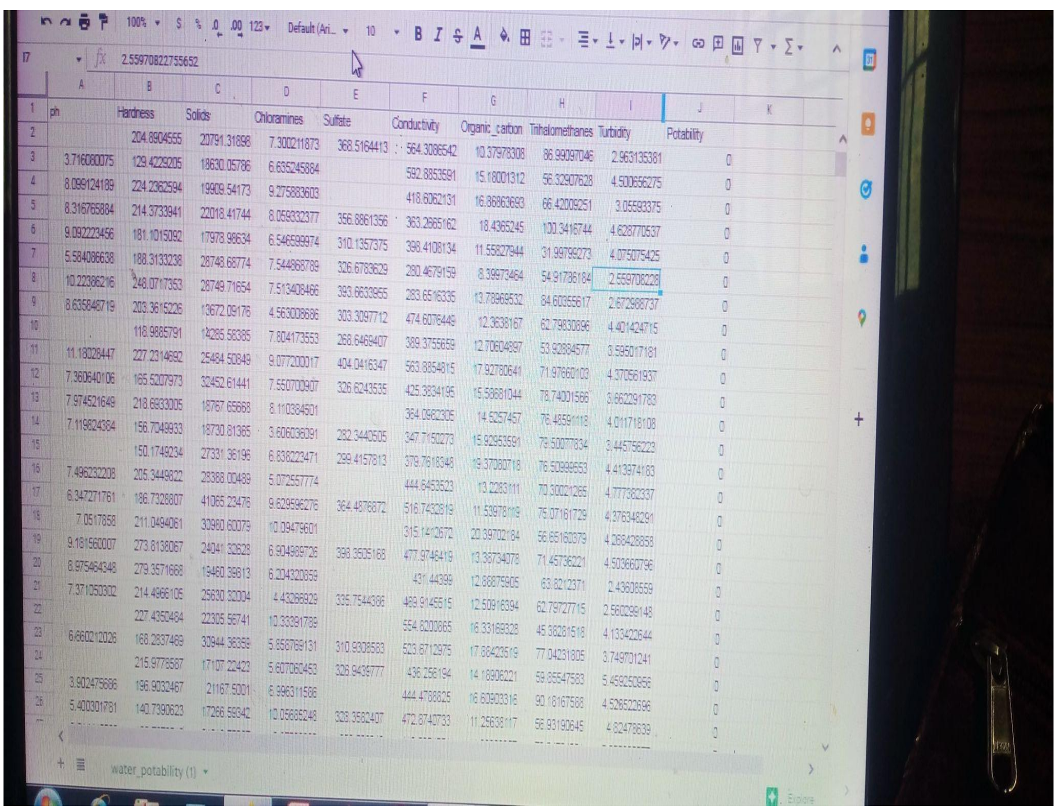
Sprint 4	USN-4	As a user, i can grasp whole content easily	2	high

7. CODING AND SOLUTIONING

7.1 Feature 1 and 7.2 feature 2:

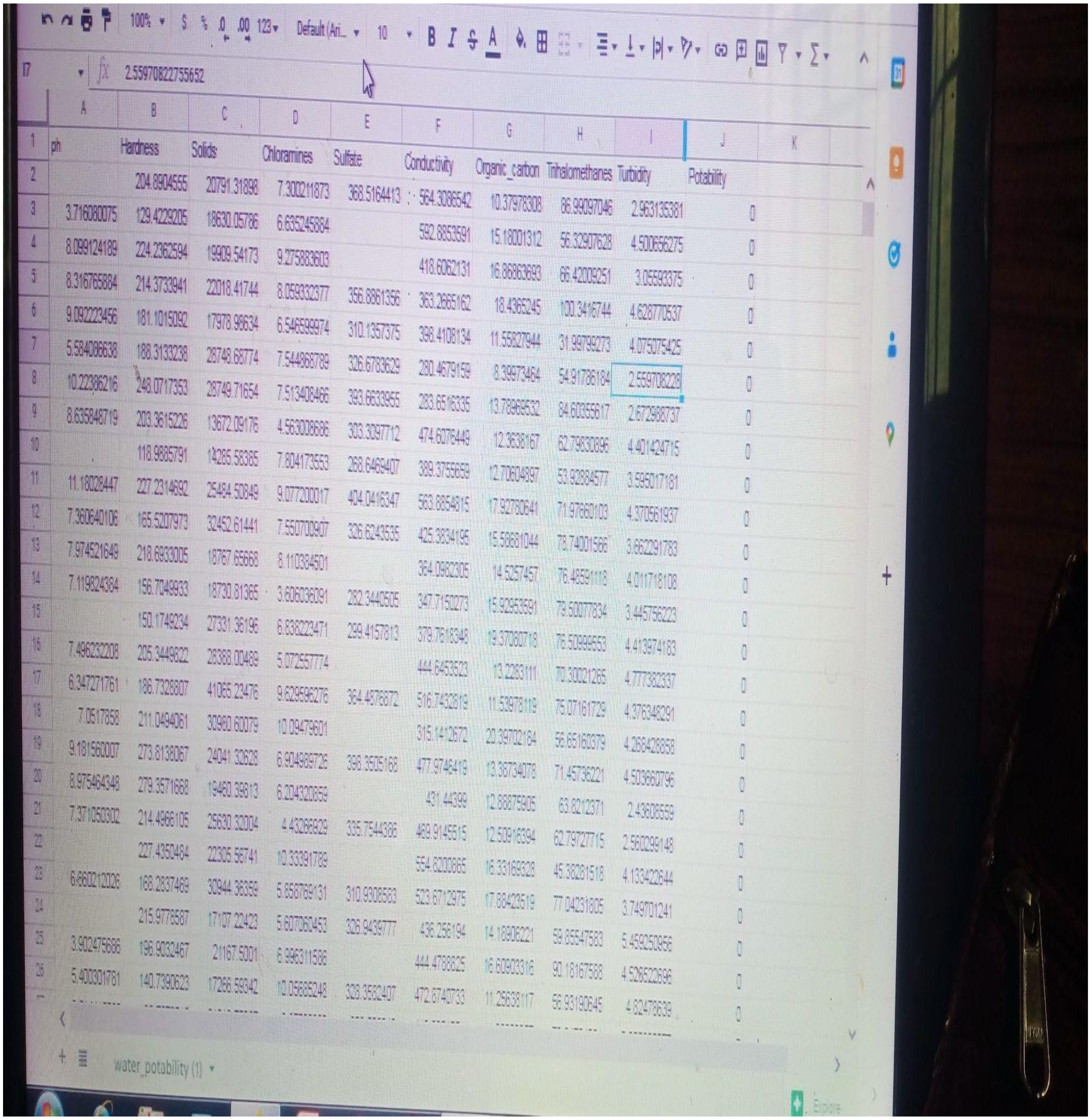
Preprocessing of Data
Feature Engineering
Diverse Algorithms
Algorithm Selection
Training and Tuning
Ensembling
.   Head-to-Head Model Competitions
Human-Friendly Insights.

7.3 Database schema



8.TESTING

8.1 TEST CASES



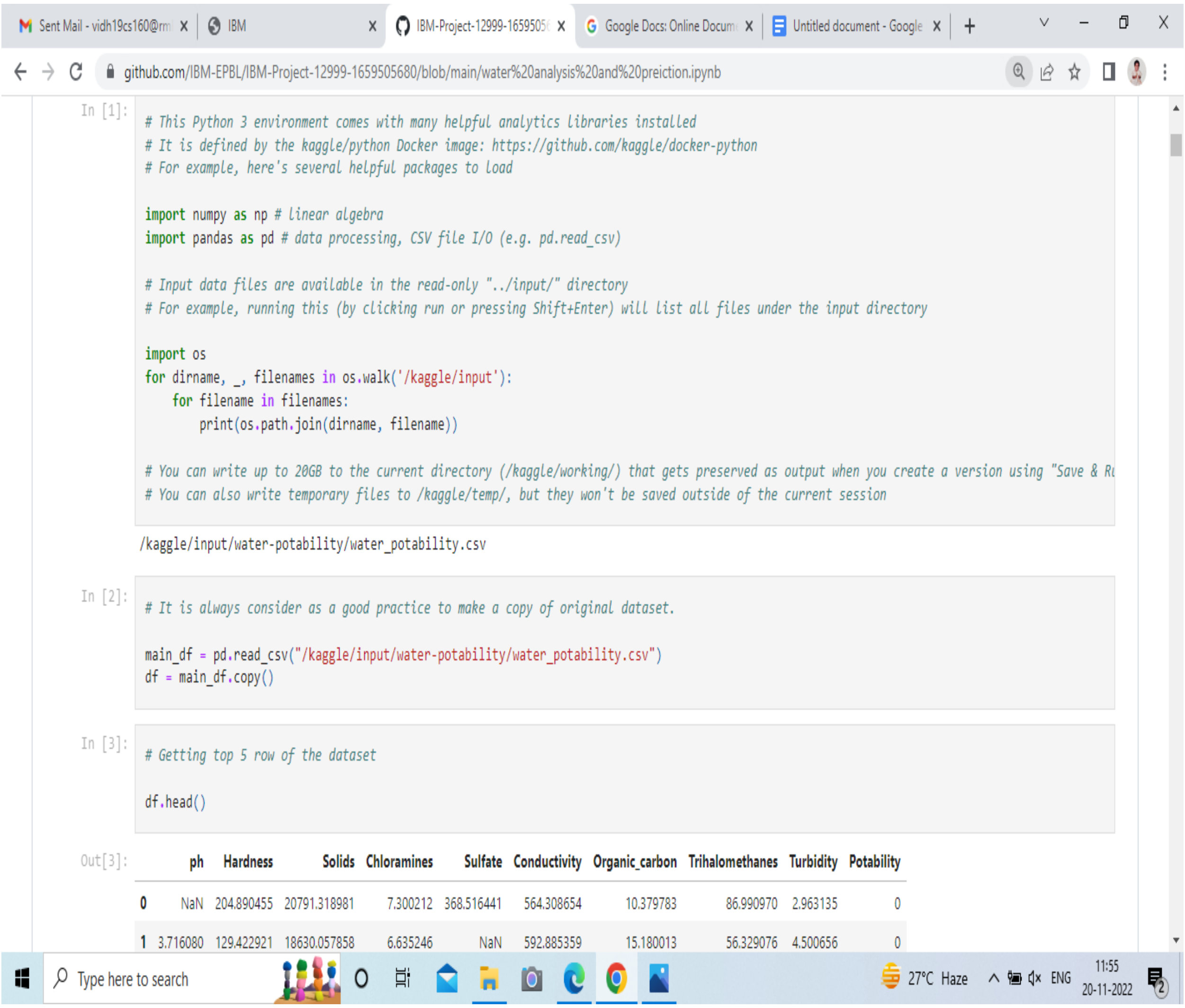
I7 fx 2.55970822755652

	A	B	C	D	E	F	G	H	I	J
1	ph	Hardness	Solids	Chloramines	Sulfate	Conductivity	Organic_carbon	Trihalomethanes	Turbidity	Potability
2		204.8904555	20791.31898	7.300211873	368.5164413	564.3086542	10.37978308	86.99097046	2.963135381	0
3	3.716080075	129.4229205	18630.05786	6.635245884		592.8853591	15.18001312	56.32907628	4.500656275	0
4	8.099124189	224.2362594	19909.54173	9.275883603		418.6062131	16.86863693	66.42009251	3.05593375	0
5	8.316765884	214.3733941	22018.41744	8.059332377	356.8861356	363.2665162	18.4365245	100.3416744	4.628770537	0
6	9.092223456	181.1015092	17978.98634	6.546599974	310.1357375	398.4108134	11.55827944	31.99799273	4.075075425	0
7	5.584086638	188.3133238	28748.68774	7.544868789	326.6783629	280.4679159	8.39973464	54.91786184	2.559708228	0
8	10.22386216	248.0717353	28749.71654	7.513408466	393.6633955	283.6516335	13.78969532	84.60355617	2.672988737	0
9	8.635848719	203.3615226	13672.09176	4.563008686	303.3097712	474.6076449	12.3638167	62.79830896	4.401424715	0
10		118.9885791	14285.58385	7.804173553	268.6469407	389.3755659	12.70604897	53.92884577	3.595017181	0
11	11.18028447	227.2314692	25484.50849	9.077200017	404.0416347	563.8854815	17.92780641	71.97860103	4.370561937	0
12	7.360640106	165.5207973	32452.61441	7.550700907	326.6243535	425.3834195	15.58681044	78.74001566	3.662291783	0
13	7.974521649	218.6933005	18767.65668	8.110384501		364.0982305	14.5257457	76.4859118	4.011718108	0
14	7.119824384	156.7049933	18730.81365	3.606036091	282.3440505	347.7150273	15.92953591	79.50077834	3.445756223	0
15		150.1749234	27331.36196	6.838223471	299.4157813	379.7618348	19.37080718	76.50999553	4.413974183	0
16	7.496232208	205.3449822	28388.00489	5.072557774		444.6453523	13.2283111	70.30021265	4.777382337	0
17	6.347271761	186.7328807	41065.23476	9.629596276	364.4876872	516.7432819	11.53978119	75.07161729	4.376348291	0
18	7.0517858	211.0494061	30980.60079	10.09479601		315.1412672	20.39702184	56.65160379	4.268428858	0
19	9.181560007	273.8138067	24041.32628	6.904989726	398.3505168	477.9746419	13.36734078	71.45736221	4.503660796	0
20	8.975464348	279.3571668	19460.39813	6.204320859		431.44399	12.88875905	63.8212371	2.43608559	0
21	7.371050302	214.4966105	25630.32004	4.43266829	335.7544386	469.9145515	12.50916394	62.79727715	2.560299148	0
22		227.4350484	22305.56741	10.33391789		554.8200865	16.33169328	45.38281518	4.133422644	0
23	6.660212026	168.2837469	30944.36359	5.858769131	310.9308583	523.6712975	17.88423519	77.04231805	3.749701241	0
24		215.9778587	17107.22423	5.607060453	326.9439777	436.256194	14.18906221	59.85547583	5.459250956	0
25	3.902475686	196.9032467	21167.5001	6.996311586		444.4788825	16.60903316	90.18167588	4.528522896	0
26	5.400301781	140.7390623	17266.59642	10.05685248	328.3582407	472.8740733	11.25638117	56.93190645	4.82478639	0

water_potability (1)

8.2 USER ACCEPTANCE TESTING

Unit testing was performed in jupyter notebook and output is displayed.



In [1]:

```python
# This Python 3 environment comes with many helpful analytics libraries installed
# It is defined by the kaggle/python Docker image: https://github.com/kaggle/docker-python
# For example, here's several helpful packages to load

import numpy as np # linear algebra
import pandas as pd # data processing, CSV file I/O (e.g. pd.read_csv)

# Input data files are available in the read-only "../input/" directory
# For example, running this (by clicking run or pressing Shift+Enter) will list all files under the input directory

import os
for dirname, _, filenames in os.walk('/kaggle/input'):
    for filename in filenames:
        print(os.path.join(dirname, filename))

# You can write up to 20GB to the current directory (/kaggle/working/) that gets preserved as output when you create a version using "Save & Ru
# You can also write temporary files to /kaggle/temp/, but they won't be saved outside of the current session
```

/kaggle/input/water-potability/water_potability.csv

In [2]:

```python
# It is always consider as a good practice to make a copy of original dataset.

main_df = pd.read_csv("/kaggle/input/water-potability/water_potability.csv")
df = main_df.copy()
```

In [3]:

```python
# Getting top 5 row of the dataset

df.head()
```

Out[3]:

	ph	Hardness	Solids	Chloramines	Sulfate	Conductivity	Organic_carbon	Trihalomethanes	Turbidity	Potability
0	NaN	204.890455	20791.318981	7.300212	368.516441	564.308654	10.379783	86.990970	2.963135	0
1	3.716080	129.422921	18630.057858	6.635246	NaN	592.885359	15.180013	56.329076	4.500656	0

Sent Mail - vidh19cs160@rm | IBM | IBM-Project-12999-1659505 | Google Docs: Online Docum | Untitled document - Google

github.com/IBM-EPBL/IBM-Project-12999-1659505680/blob/main/water%20analysis%20and%20preiction.ipynb

```
main_df = pd.read_csv("/kaggle/input/water-potability/water_potability.csv")
df = main_df.copy()
```

In [3]:

```
# Getting top 5 row of the dataset

df.head()
```

Out[3]:

	ph	Hardness	Solids	Chloramines	Sulfate	Conductivity	Organic_carbon	Trihalomethanes	Turbidity	Potability
0	NaN	204.890455	20791.318981	7.300212	368.516441	564.308654	10.379783	86.990970	2.963135	0
1	3.716080	129.422921	18630.057858	6.635246	NaN	592.885359	15.180013	56.329076	4.500656	0
2	8.099124	224.236259	19909.541732	9.275884	NaN	418.606213	16.868637	66.420093	3.055934	0
3	8.316766	214.373394	22018.417441	8.059332	356.886136	363.266516	18.436524	100.341674	4.628771	0
4	9.092223	181.101509	17978.986339	6.546600	310.135738	398.410813	11.558279	31.997993	4.075075	0

In [4]:

```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
%matplotlib inline
import plotly.express as px
import warnings
warnings.filterwarnings('ignore')
```

Following are the list of algorithms that are used in this notebook.

Algorithm

Type here to search

27°C Haze ENG 11:56 20-11-2022

9. RESULT

9.1 PERFORMANCE METRICES

In this project we use following performance metrics:

1.ph
2.Solid
3.Hardness
4.Chloramines
5.Sulfate
6.Conductivity
7.Organic carbon
8.Trihalomethane
9.Turbity
10.Potability

1.ph:PH stands for **Hydrogen potentials**. It refers to the concentration of the hydrogen ions in a solution. This is the indicator of a solution's acidity or alkalinity. The pH value on a pH-scale varies from 0 to 14.

2.Hardness:the quality or condition of being hard.

3.solid:hard and firm; not in the form of liquid or gas.

4.chloramines:any of a group of antiseptics and disinfectants which are sulphonamide derivatives containing chlorine bonded to nitrogen.

5.sulfate:a salt or ester of sulphuric acid, containing the anion SO_4^{2-} or the divalent group —OSO_2O—.

6.conductivity:the degree to which a specified material conducts electricity, calculated as the ratio of the current density in the material to the electric field which causes the flow of current.

7.organic carbon:Organic matter makes up just 2–10% of most soil's mass and has an important role in the physical, chemical and biological function of agricultural soils.

8.trihalmethanes:In chemistry, trihalomethanes (THMs) **are chemical compounds in which three of the four hydrogen atoms of methane (CH 4) are replaced by halogen atoms**

9.turbidity:the quality of being cloudy, opaque, or thick with suspended matter.

10.potability:Potable water, also known as drinking water.

10 MERITS & DEMERITS

Advantages of Machine Learning	Disadvantages of Machine Learning
It is automatic	Chances of error or fault are more
It is used in various fields	Data requirement is more
It can handle varieties of data	Time-consuming and more resources required

11.CONCLUSION

In conclusion, machine learning methods can identify the types of seawater pollutants, determine the concentration and distribution of pollutants, and provide a relevant analysis of the status of marine organisms.

12. FUTURE SCOPE:

In future, the designed system with used machine learning classification system algorithms can be used to predict water quality. The work can be extended and improved for automation of water analysis including some othe machine learning algorithm.

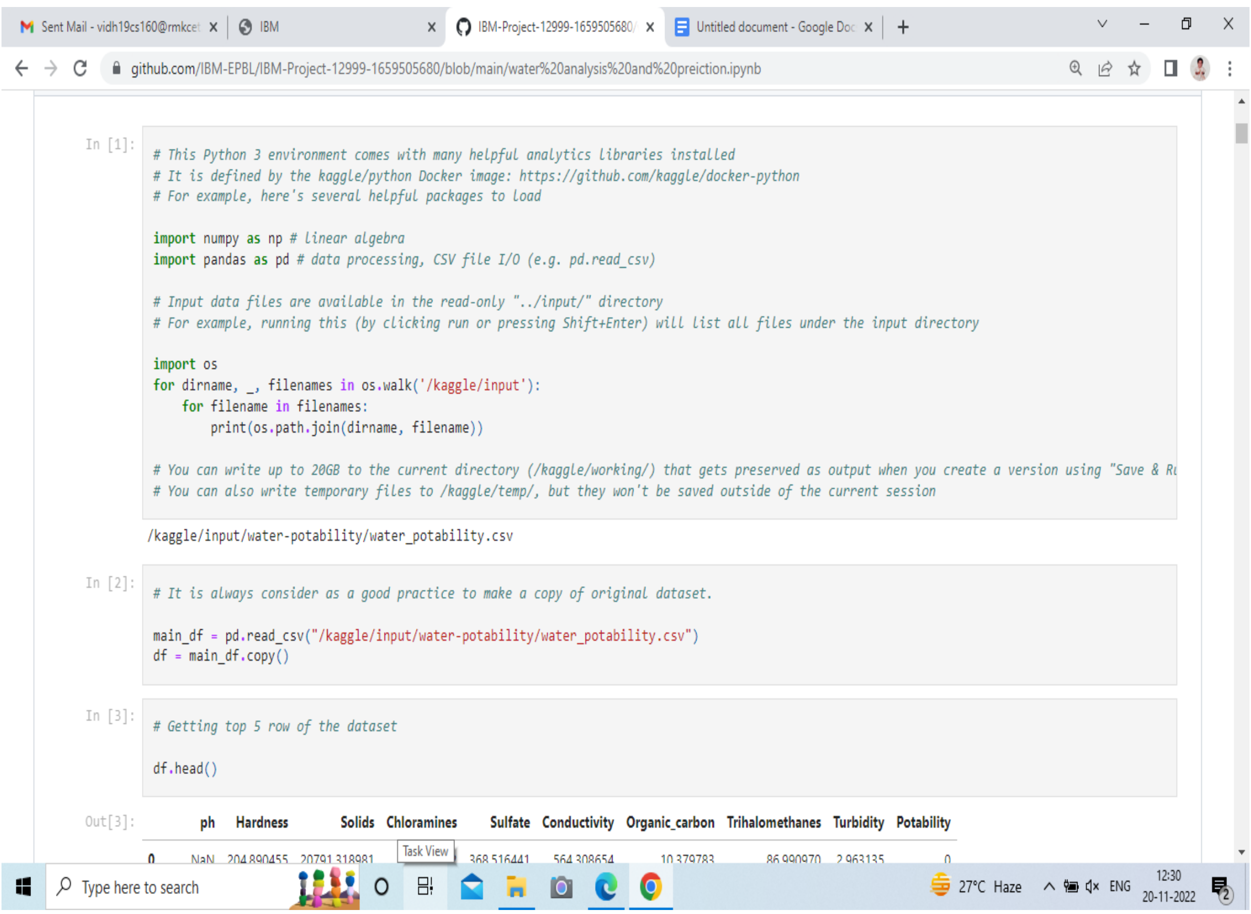
13 .APPENDIX 1

SAMPLE CODE

```
import numpy as np # linear algebra
import pandas as pd # data processing, CSV file I/O (e.g. pd.read_csv)
import os
for dirname, _, filenames in os.walk('/kaggle/input'):
    for filename in filenames:
        print(os.path.join(dirname, filename))
main_df = pd.read_csv("/kaggle/input/water-potability/water_potability.csv")
df = main_df.copy()
df.head()
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
%matplotlib inline
import plotly.express as px
import warnings
warnings.filterwarnings('ignore')
plt.figure(figsize=(10, 8))
sns.heatmap(df.corr(), annot= True, cmap='coolwarm')
# Unstacking the correlation matrix to see the values more clearly.
corr = df.corr()
c1 = corr.abs().unstack()
c1.sort_values(ascending = False)[12:24:2]
ax = sns.countplot(x = "Potability",data= df, saturation=0.8)
plt.xticks(ticks=[0, 1], labels = ["Not Potable", "Potable"])
plt.show()
x = df.Potability.value_counts()
labels = [0,1]
print(x)
x = df.Potability.value_counts()
labels = [0,1]
print(x)
fig, ax = plt.subplots(ncols = 5, nrows = 2, figsize = (20, 10))
index = 0
ax = ax.flatten()

for col, value in df.items():
    sns.boxplot(y=col, data=df, ax=ax[index])
```

```
    index += 1
plt.tight_layout(pad = 0.5, w_pad=0.7, h_pad=5.0)
models = pd.DataFrame({
    'Model':['Logistic Regression', 'Decision Tree', 'Random Forest', 'XGBoost',
'KNeighbours', 'SVM', 'AdaBoost'],
    'Accuracy_score' :[lg, dt, rf, xgb, kn, sv, ada]
})
models
sns.barplot(x='Accuracy_score', y='Model', data=models)

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

APPENDIX 2
PROGRAM SCREENSHOTS



Out[3]:

	ph	Hardness	Solids	Chloramines	Sulfate	Conductivity	Organic_carbon	Trihalomethanes	Turbidity	Potability
0	NaN	204.890455	20791.318981	7.300212	368.516441	564.308654	10.379783	86.990970	2.963135	0
1	3.716080	129.422921	18630.057858	6.635246	NaN	592.885359	15.180013	56.329076	4.500656	0
2	8.099124	224.236259	19909.541732	9.275884	NaN	418.606213	16.868637	66.420093	3.055934	0
3	8.316766	214.373394	22018.417441	8.059332	356.886136	363.266516	18.436524	100.341674	4.628771	0
4	9.092223	181.101509	17978.986339	6.546600	310.135738	398.410813	11.558279	31.997993	4.075075	0

In [4]:

```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
%matplotlib inline
import plotly.express as px
import warnings
warnings.filterwarnings('ignore')
```

Following are the list of algorithms that are used in this notebook.

Algorithm
Logistic Regression
Decision Tree
Random Forest
XGBoost
KNeighbours
SVM

In [6]:

```
print(df.columns)
```

```
Index(['ph', 'Hardness', 'Solids', 'Chloramines', 'Sulfate', 'Conductivity',
       'Organic_carbon', 'Trihalomethanes', 'Turbidity', 'Potability'],
      dtype='object')
```

In [7]:

```
df.describe()
```

Out[7]:

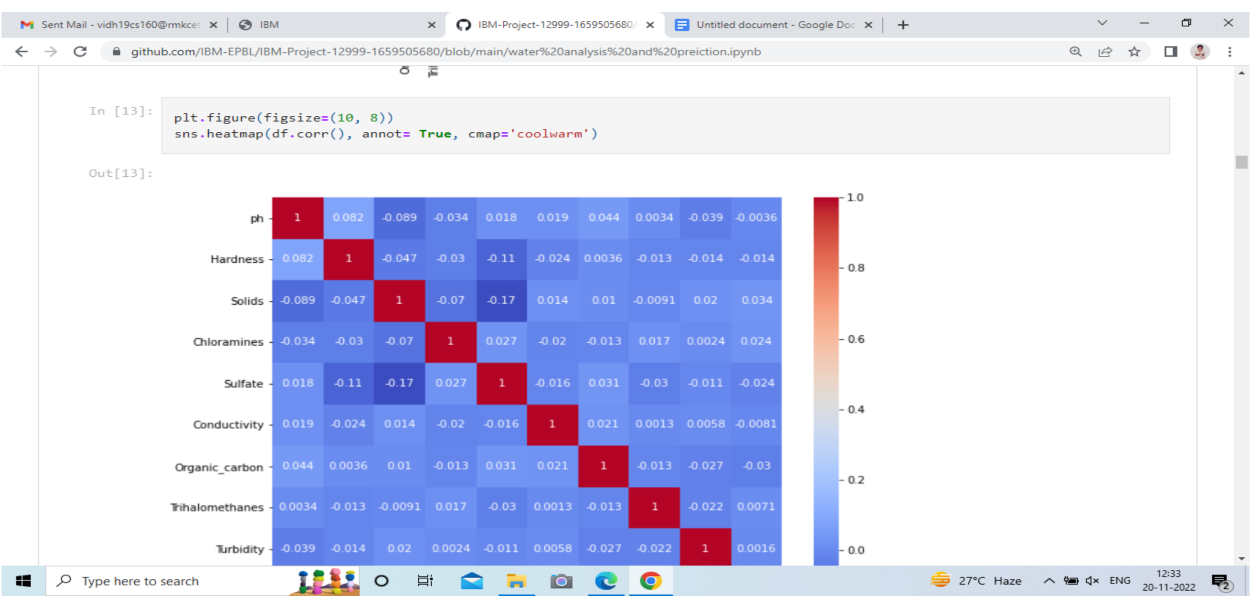
	ph	Hardness	Solids	Chloramines	Sulfate	Conductivity	Organic_carbon	Trihalomethanes	Turbidity	Potability
count	2785.000000	3276.000000	3276.000000	3276.000000	2495.000000	3276.000000	3276.000000	3114.000000	3276.000000	3276.000000
mean	7.080795	196.369496	22014.092526	7.122277	333.775777	426.205111	14.284970	66.396293	3.966786	0.390110
std	1.594320	32.879761	8768.570828	1.583085	41.416840	80.824064	3.308162	16.175008	0.780382	0.487849
min	0.000000	47.432000	320.942611	0.352000	129.000000	181.483754	2.200000	0.738000	1.450000	0.000000
25%	6.093092	176.850538	15666.690297	6.127421	307.699498	365.734414	12.065801	55.844536	3.439711	0.000000
50%	7.036752	196.967627	20927.833607	7.130299	333.073546	421.884968	14.218338	66.622485	3.955028	0.000000
75%	8.062066	216.667456	27332.762127	8.114887	359.950170	481.792304	16.557652	77.337473	4.500320	1.000000
max	14.000000	323.124000	61227.196008	13.127000	481.030642	753.342620	28.300000	124.000000	6.739000	1.000000

In [8]:

```
df.info()
```

```
RangeIndex: 3276 entries, 0 to 3275
Data columns (total 10 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   ph               2785 non-null   float64
 1   Hardness         3276 non-null   float64
 2   Solids           3276 non-null   float64
```


github.com/IBM-EPBL/IBM-Project-12999-1659505680/blob/main/water%20analysis%20and%20preiction.ipynb
In [12]:
sns.heatmap(df.isnull())
Out[12]:
ph
Hardness
Solids
Chloramines
Sulfate
Conductivity
Organic_carbon
Trihalomethanes
Turbidity
Potability
In [13]:
plt.figure(figsize=(10, 8))
sns.heatmap(df.corr(), annot= True, cmap='coolwarm')
Out[13]:


github.com/IBM-EPBL/IBM-Project-12999-1659505680/blob/main/water%20analysis%20and%20preiction.ipynb
In [13]:
plt.figure(figsize=(10, 8))
sns.heatmap(df.corr(), annot= True, cmap='coolwarm')
Out[13]:
ph 1 0.082 -0.089 -0.034 0.018 0.019 0.044 0.0034 -0.039 -0.0036
Hardness 0.082 1 -0.047 -0.03 -0.11 -0.024 0.0036 -0.013 -0.014 -0.014
Solids -0.089 -0.047 1 -0.07 -0.17 0.014 0.01 -0.0091 0.02 0.034
Chloramines -0.034 -0.03 -0.07 1 0.027 -0.02 -0.013 0.017 0.0024 0.024
Sulfate 0.018 -0.11 -0.17 0.027 1 -0.016 0.031 -0.03 -0.011 -0.024
Conductivity 0.019 -0.024 0.014 -0.02 -0.016 1 0.021 0.0013 0.0058 -0.0081
Organic_carbon 0.044 0.0036 0.01 -0.013 0.031 0.021 1 -0.013 -0.027 -0.03
Trihalomethanes 0.0034 -0.013 -0.0091 0.017 -0.03 0.0013 -0.013 1 -0.022 0.0071
Turbidity -0.039 -0.014 0.02 0.0024 -0.011 0.0058 -0.027 -0.022 1 0.0016

Sent Mail - vidh19cs160@rmkce | IBM | IBM-Project-12999-1659505680/ | Untitled document - Google Doc

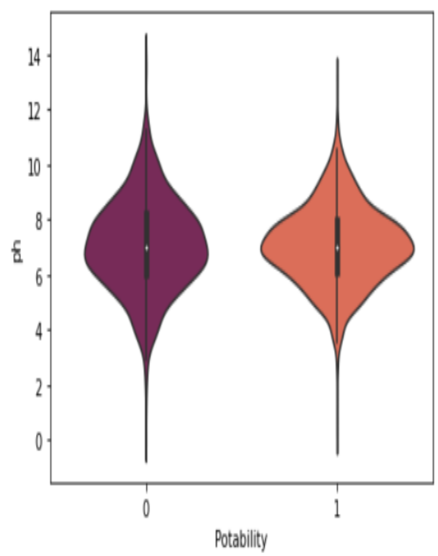
github.com/IBM-EPBL/IBM-Project-12999-1659505680/blob/main/water%20analysis%20and%20preiction.ipynb

```
1    1278
Name: Potability, dtype: int64
```

In [17]:

```
sns.violinplot(x='Potability', y='ph', data=df, palette='rocket')
```

Out[17]:



In [18]:

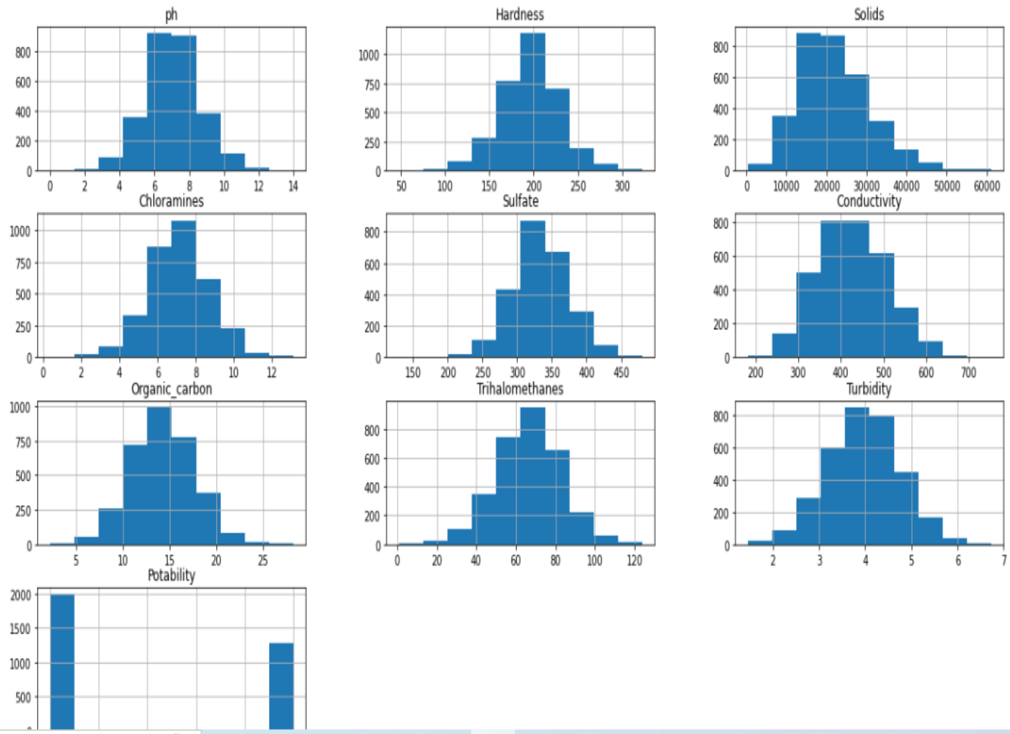
```
# Visualizing dataset and also checking for outliers

fig, ax = plt.subplots(ncols = 5, nrows = 2, figsize = (20, 10))
index = 0
ax = ax.flatten()

for col, value in df.items():
```

Type here to search

27°C Haze ENG 12:34 20-11-2022

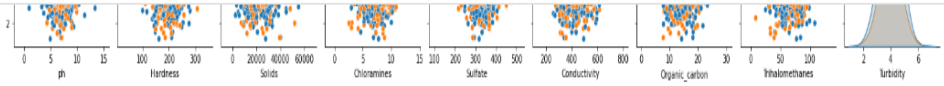
In [19]:

```
plt.rcParams['figure.figsize'] = [20,10]
df.hist()
plt.show()
```



Sent Mail - vidh19cs160@rmkcet | IBM | IBM-Project-12999-1659505680/ | Untitled document - Google Doc

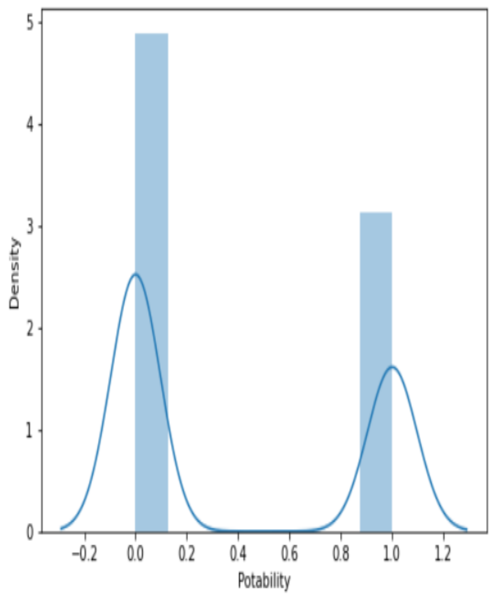
github.com/IBM-EPBL/IBM-Project-12999-1659505680/blob/main/water%20analysis%20and%20preiction.ipynb



In [21]:

```
plt.rcParams['figure.figsize'] = [7,5]
sns.distplot(df['Potability'])
```

Out[21]:



In [22]:

```
df.hist(column='ph', by='Potability')
```

Type here to search

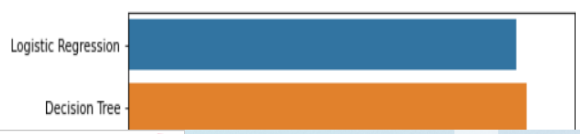
27°C Haze ENG 12:35 20-11-2022

In [96]:

```python
models = pd.DataFrame({
    'Model':['Logistic Regression', 'Decision Tree', 'Random Forest', 'XGBoost', 'KNeighbours', 'SVM', 'AdaBoost'],
    'Accuracy_score' :[lg, dt, rf, xgb, kn, sv, ada]
})
models
sns.barplot(x='Accuracy_score', y='Model', data=models)

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

Out[96]:

	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



1	Decision Tree	0.645102
6	AdaBoost	0.634011
0	Logistic Regression	0.628466
2	Random Forest	0.628466