8-2023

COMPARING THE EFFECTIVENESS OF DIFFERENT BOOSTING ALGORITHMS FOR GROUND WATER QUALITY IN TELANGANA REGION

Divy Jot

COMPARING THE EFFECTIVENESS OF DIFFERENT BOOSTING ALGORITHMS FOR WATER QUALITY OF GROUND WATER IN TELANGANA REGION

A Project

Presented to the

Faculty of

California State University,

San Bernardino

In Partial Fulfillment

of the Requirements for the Degree

Master of Science

in

Information Systems and Technology

---

by

Divy Jot

(August 2023)

COMPARING THE EFFECTIVENESS OF DIFFERENT BOOSTING ALGORITHMS FOR WATER QUALITY OF GROUND WATER IN TELANGANA REGION

A Project

Presented to the

Faculty of

California State University,

San Bernardino

---

by

Divy Jot

(August 2023)

Approved by:

Dr William Butler, Committee Chair

Dr Conrad Shayo, Chair of Department of Information and Decision Sciences

ABSTRACT

This culminating experience research project explores the parameters needed to predict the water quality levels for use in different climatic conditions pre and post monsoon from 2018 to 2020 in Telangana State, India. A study was conducted on the water quality analysis by using linear regression with water quality Index in Telangana region. However, in this study we are replicating the water quality analysis by using stack model and machine learning algorithms such as Light Gradient Boosting Machine, Random Forest, and Artificial Neural Network. The research Questions are: Q1. What are the sources of the significant parameters that impact groundwater quality in a location? Q2. Will the use of the stacked model analysis approach produce different results when applied to the Telangana dataset? Q3. How does the size and nature of a dataset impact the effectiveness of ensemble techniques, such as stacking, for addressing class imbalance in groundwater quality prediction models? The findings are: 1. Sodium and Magnesium parameters values have been calculated for Sodium Adsorption Ratio (SAR) for the ground water samples. Based on these parameter electrical conductivity EC and SAR values, Salinity hazard values calculated and converted into different classes like Low C1 (EC<250), Medium C2 (EC 250 – 750), High C3 (750-2250), Very High C4 (>2250). Sodium Hazard Classes Low S1 (SAR < 10), Medium S2 (SAR 10 – 18), High S3 (SAR

18-26), Very High S4 (SAR > 26). In comparing of 2018, 2019 and 2020 dataset of water quality analysis, increased in ranges Sodium (5.07 to 748), Calcium (1.2 to 640.0), Magnesium (4.86 to 457.02), Electrical Conductivity (102 to 9499). 2. Stacked models achieved the best performance with use of different classifiers in terms of accuracy (the individual models of Random forecast 97%, Light GBM 97% and calculation of two predicted probability values passes through ANN which model accuracy increasers to 98%) to predict the water quality by collecting the data from different regions and climatic conditions based on the suitability of water salinity and sodium content. 3. To manage imbalanced data and increase prediction accuracy by calculating the model performance by using classification report of random forest, LGBM and ANN these are the values which are varying in performance F1 Score. For Class Marginal (RF-0.63), (LGBM-0.67), ANN increased to performance to (0.76). Class Poor (RF-0.95), (LGBM-0.95), ANN increased to performance to (0.96), Class Very Poor (RF-0.77), (LGBM-0.77), ANN increased to performance to (0.86). For classes Excellent and good F1 Score for 3 models are1 and for Permissible three models got 0.99. The conclusions are: 1. This Research provides helpful information to understand and handle the potential risks of salinity and sodium in the researched region by classifying the salinity hazard levels into four classes (C1 to C4 and S1 to S4) based on electrical conductivity (EC) and SAR values. 2. To conclude, our research demonstrates that stacked models, employing different classifiers, have proven to be highly effective in predicting water quality with remarkable accuracy.

When we utilized the predicted probability values by passing them through the Artificial Neural Network (ANN), the accuracy further improved to an impressive 98%. 3. The stacked model technique, which combines random forest, light GBM, and ANN, seems to be an effective means of dealing with imbalanced data and enhancing prediction accuracy. The significant improvement in F1 Scores for a few classes, especially when using ANN, demonstrates how effectively this ensemble approach handles challenging classification problems.

Furthermore, emerging areas for future research that emerged from this study include the opportunity for training and testing using our model with a larger dataset and modifying different hypermeters for further improvement.

DEDICATION

To my beloved Mom and Dad

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER ONE

INTRODUCTION

Water is a fundamental resource for human survival and the health of ecosystems, playing a critical role in maintaining the well-being of humans, animals, and ecosystems. However, human activities can lead to contamination and deterioration of water quality, particularly in water-scarce areas where groundwater serves as a crucial source of drinking water and irrigation (Azrour et al., 2021). A study was conducted on the water quality analysis by using linear regression with WQI water quality Index in Telangana region (Saikrishna et al., 2020). However, in this study we are replicating the water quality analysis by using stack model and machine learning algorithms such as Light Gradient Boosting Machine, Random Forest, and Artificial Neuro Network. To implement the accurate and reliable water quality prediction models are necessary to assess environmental and health risks, as well as to monitor and control water quality (Kumar et al., 2017). Nevertheless, developing such models is challenging due to the complex relationships between water quality parameters and the multitude of influencing factors.

These Factors can by influenced by numerous interconnected variables such as temperature, dissolved oxygen levels, pH, turbidity, nutrient concentrations, and the presence of pollutants. These parameters do not act in isolation but interact with one another, creating intricate and dynamic

relationships that are often difficult to untangle. Moreover, external factors like land use, climate patterns, human activities, and natural processes further complicate the picture, as they can exert significant influence on water quality. To build accurate and reliable models, scientists and researchers must account for these complex relationships, considering both direct and indirect influences on water quality parameters. It requires comprehensive data collection, sophisticated analytical techniques, and advanced modelling approaches to unravel these intricate relationships and develop effective tools for water quality assessment and management.

This study finds that Telangana region sheds light on the major influence of human activities on groundwater quality. For instance, a study by (Subba et al., 2021) in the Yellareddygudem watershed highlighted the impact of human activities on groundwater quality, by using Geographical information system (GIS) to delineate the seasonal and spatial variations for vulnerable zones related to the drinking groundwater quality index (DGQI) and irrigation groundwater quality index (IGQI). Furthermore, Studies have also emphasized on evaluating groundwater quality in specific areas of Telangana. (Durgasrilakshmi et al., 2021) assessed the groundwater quality in the Uppal Kalan area and found severe pollution, with non-compliance to drinking water standards. Experts recommended regular monitoring, appropriate treatment, and further research to investigate pollution-causes and its impact on human health. Similarly, (Suresh et al., 2017) examined water contamination in Kandlakoya

Village caused by industrial operations and domestic waste, emphasizing the negative effects on human health.

To Examine the water contamination in particular region, Machine learning algorithms have proven to be valuable tools in predicting water quality and have been successfully utilized in various environmental monitoring applications (Wang et al., 2018). This study focused on machine learning tools, mainly MLP Algorithms demonstrated their effectiveness in enhancing the accuracy and reliability of predictive models. Their ability to handle large and complex datasets enables them to capture the intricate relationships between water quality parameters and influencing factors. By analysing historical data, machine learning models can identify patterns and correlations that may not be easily discernible through traditional statistical methods. Moreover, these algorithms can adapt and learn from new data, continuously improving their predictive capabilities (El Baba et al., 2020).

Problem Statement

The main objective of this culminating experience project is to provide a comprehensive overview of groundwater quality assessment in Telangana by synthesizing the findings of various case studies conducted in the region. It is facing severe contamination due to human activities such as industrialization, urbanization practices. This study major focuses on the Imbalance in ground water quality predictions using various type of stacked models, which would be

helpful to predict the ground water quality accuracy and reliability. Another aspect to work on this study combination of various parameters which are impacted on ground water quality in Telangana Region. Another considerable factor of this research work on Stacked model-based machine learning algorithms which predicts the drastic changes in climatic conditions and other parameters in water quality. The outcomes of this study will contribute to a better tool for analysing groundwater quality and its implications for human health and ecosystems, facilitating effective water resource management and conservation efforts in Telangana and beyond.

Research Questions

This project will seek to answer the following questions:

1. What are the sources of the significant parameters that impact groundwater quality in the location? (ALMuhisen et al., 2019).
2. Will the use of the stacked model analysis approach produce different results when applied to the Telangana dataset?
3. How does the size and nature of the dataset impact the effectiveness of ensemble techniques, such as stacking, for addressing class imbalance in groundwater quality prediction models? (Malek et al., 2022)

Organization of Project

This culminating experience project is organized as follows*:* Chapter 1 provided the introduction, Problem statement, Research questions, and motivation of the study. Chapter 2 covers the Literature review, Chapter 3 provide the Methods used to answer the research questions, Chapter 4 covers Data collection, Analysis, and findings; and Chapter 5 provides the discussion of the Findings, Conclusions and Recommendations for Future Research.

CHAPTER TWO

LITERATURE REVIEW

Q1. What are the sources of the significant parameters that impact groundwater quality in the location?

Water quality in location state is required for food security, cattle feeding, industrial production, and environmental conservation. However, the scarcity of water is becoming a growing concern due to both natural and human causes. Water quality decreasing caused by the accumulation of industrial waste, solid waste, pollutions, contaminated materials in water systems like canals, rivers, ponds, and lakes (Guleria et al., 2018).

In regions with low rainfall and limited surface water resources, groundwater plays a crucial role in sustaining communities and supporting agricultural needs. (Singh et al., 2019) By conducted a study in a hyper-arid location to analyse the quality of groundwater and by employing multivariate statistical analysis, the researchers collected and examined total 43 groundwater samples. They utilized techniques such as principal component analysis (PCA) and hierarchical cluster analysis (HCA) to understand the complex interrelationships among various physicochemical variables that affect groundwater quality in the study area.

Furthermore, the statistical techniques revealed that the variables in the groundwater samples were highly associated with one another. The study

discovered that cluster analysis was a good approach for determining the spatial similarity between the contributing variables (Reghunath et al., 2002). According to (Singh et al., 2019) conducted a study that used multivariate statistical analysis to analyse groundwater quality in a hyper-arid location. The findings of the study have significant implications for groundwater management and protection in hyper-arid locations, where the sustainable use and conservation of groundwater resources are particularly crucial (Huang et al., 2019).

Q2. Will the use of the stacked model analysis approach produce different results when applied to the Telangana dataset?

A significant study conducted by (Deepika et al., 2020) emphasizes the importance of groundwater quality parameters in evaluating the suitability of groundwater for various applications such as domestic, industrial, agricultural, and environmental purposes. These Parameters such as pH, electrical conductivity, total dissolved solids, chlorides, total hardness, nitrates, sulphates, sodium, calcium, and magnesium are commonly used to assess groundwater quality. The APHA method was used to determine the physicochemical parameters of the water samples, and the Water Quality Index (WQI) was used to express the quality of the groundwater. The findings indicated that two sites had poor groundwater quality, primarily due to high levels of nitrate contamination and extreme water hardness. The region's groundwater is mismatched for both human usage and present-day use in case it isn't managed appropriately.

According to Deepika et al., (2020) the Water Quality Index is a thorough measure that offers a solitary mathematical benefit to address the general nature of a specific water source. It contains various considerable parameters PH, Oxygen, Dissolved Solids. The availability of contaminants, this record and it imply as a water quality quantitative performance indicator. It will track the water quality changes over the time durations.

A study by Adimalla et al., (2021) conducted water quality predictions based on multiple approaches, where researcher chooses two best approaches such as Entropy weighted water Index and Pollution Index groundwater. This study shows up that 60% of the groundwater samples has met the level of satisfaction for drinking wate and the remaining 40% of the samples were in the bad condition for human consumption. Total 63 % water samples are suitable for the human consumption as predicted by the pollution index groundwater. This study further identified that the groundwater in the region exhibited alkaline characteristics, with sodium (Na+) being the most abundant cation and chloride (Cl) being the least abundant.

These studies shed light on the significance of evaluating groundwater quality using various parameters and indices. The findings highlight the need for proper water treatment to ensure safe drinking water and sustainable use of groundwater resources in the examined regions. The stacked model approach has not been used to analyze water quality in the Telangana dataset.

Q3. How does the size and nature of the dataset impact the effectiveness of ensemble techniques, such as stacking, for addressing class imbalance in groundwater quality prediction models?

Several studies have focused on water quality prediction models and the calculation of the water quality index (WQI) due to the increasing global challenge of water quality deterioration. According to Khan et al., (2021) researcher employed that a principal component regression approach to predict water quality. They used the weighted arithmetic index to calculate the WQI, considering various physicochemical properties of water samples. The dataset underwent principal component analysis (PCA) to identify the most dominant water quality index features, which were then used to forecast the quality Index. Also, they have applied Multiple regression algorithms to the principal component Analysis output for accurate water quality index prediction. They projected WQI was further classified using the Gradient Boosting Classifier, yielding a 95% prediction accuracy for the principal component regression method and a 100% classification accuracy for the Gradient Boosting Classifier, demonstrating their effectiveness compared to state-of-the-art models.

Elbeltagi et al., (2021) researcher concluded that, when dealing with multiple variables, conventional methods for calculating the WQI may be time-consuming and error prone. To address this, the researchers applied four standalone approaches (AR, M5P, RSS, and SVM) using a variable elimination technique to predict the WQI. Among these approaches, the additive regression

(AR) technique performed the best, exhibiting high prediction precision with a small number of input parameters. Based on the performance metrics (R2 = 0.9993, MAE = 0.5243, RMSE = 0.06356%, RAE = 3.8449, and RRSE = 3.9925%), the researchers considered AR as a reliable and accurate approach for predicting WQI in the research area. An Additive Regression performed well for water quality observation and production process. Groundwater is a major source of income in particular areas, people are mostly depending on it for the agriculture.

The literature review reveals that, water plays essential roles for all living species and the facing a major problem of water insufficiency generated by natural and human factors. To examine, government should implement essential measures to protect ground water resources. In various regions groundwater is actual source of income which is characterized by less rain fall and limited water availability. A study conducted that in various locations cluster analysis was an effective method for predicting the contributing variables in the ground water. The Water Quality Index is an extensive used to measure various water quality parameters. Several methodologies have been employed to measure groundwater quality, such as the entropy weighted water quality index (EWQI), groundwater pollution index (PIG), and traditional as well as standalone approaches for predicting WQI. The various methodologies used to measure groundwater quality, including the entropy weighted water quality index (EWQI), groundwater pollution index (PIG), and the traditional and standalone

approaches for predicting WQI. The prediction model is evaluated based on performance and error calculation by using MAE, MSE, RMSE and R Square. (Ahmad et al., 2022) The findings from such evaluations contribute to the advancement of water quality monitoring practices and support informed decision-making processes regarding water resource management.

CHAPTER THREE

RESEARCH METHODS

As noted in Chapter 1, this project will seek to answer the following questions:

1. What are the sources of the significant parameters that impact groundwater quality in the location?
2. Will the use of the stacked model analysis approach produce different results when applied to the Telangana dataset?
3. How does the size and nature of the dataset impact the effectiveness of ensemble techniques, such as stacking, for addressing class imbalance in groundwater quality prediction models?

Question 1: What are the sources of the significant parameters that impact groundwater quality in the location?

To gain insights into groundwater quality in Telangana, it's necessary to select a study area that encompasses diverse land uses, geological formations, and potential sources of pollution. The choice of the study area should take into consideration the availability of groundwater data and the feasibility of conducting field sampling.

The data collection stage will include the collection of groundwater tests from numerous areas inside the region. Wells or boreholes will serve as the sources of these samples, and their geographical coordinates will be recorded using GPS technology. These tests will at that point experience a comprehensive investigation to decide different water quality parameters, counting pH esteem,

add up to broken down solids (TDS), nitrates, phosphates, overwhelming metals, and other pertinent pointers. Additionally, data regarding land use patterns, climate conditions, and geological characteristics will be gathered from reliable sources.

There are some noticeable factors which are influencing groundwater quality, there are data analysis will be conducted and statistical analysis techniques, will be implemented to investigate the corelation between groundwater quality and their parameters. By finding various parameters, this analysis will contribute to the identify pollution-based sources, give the chance to get more clarity of their impact in Water Quality.

Question 2: Will the use of the stacked model analysis approach produce different results when applied to the Telangana dataset?

To answer this question, we considered two different models Random Forest Classifier and of Gradient Boosting Machines algorithm including their advantages, and disadvantages on water quality predictions. To research on this question Data Preparation played a vast role. To get this comprehensive dataset will be containing various parameters related to water quality, which will include water quality indices as a Water Quality Index. To predict groundwater quality of the region, a stacked model methodology will be use and it involves individual base models using Gradient Boosting and Random Forest learning algorithms. An ensemble method would be generated by adding the predictions from base

models and this prediction will serve as features for the ground water quality by using Artificial Neural Network.

The performance of the stacked ensemble model will be conducted by various metrics including Mean Absolute Error, Mean Squared Error, Root Mean Squared Errors. This evaluation of ensemble model metrics will provide the effectiveness of the stacked model approach in predicting groundwater quality with the combination of parameters.

The Random Forest classifier is problem solving machine learning algorithm which perform effective calculations and develop multiple decision trees for their combine outputs (Kumar et al., 2017). This algorithm is adaptable because it can handle both numerical and categorical data and is well-suited for datasets with many features (Blanchet et al., 2021). However, it may not be appropriate for small datasets with few features; the dataset's size and complexity should be taken into consideration (Hemant et al., 2017). The Key parameters to consider for the Random Forest algorithm include the number of trees, maximum tree depth, and the number of features at each split (Boulesteix et al., 2022). These numbers of trees should be carefully selected to capture dataset complexity without overfitting, using approaches like cross-validation and heuristic rules (Louppe et al., 2014).

The maximum tree depth also impacts model performance, where a balance between capturing relationships and avoiding overfitting should be achieved (Probst et al., 2019). The Random Forest algorithm has advantages in

handling complex relationships and robustness to noisy data, making it suitable for various applications (Ahmed et al., 2015). However, it can be computationally expensive with many trees and prone to overfitting in the presence of noisy or high-dimensional data (Salles et al., 2015). When implementing the Random Forest algorithms, Python libraries and programming languages are used and evaluating these algorithm's performance with metrics such as accuracy, precision, recall, and F1 score is essential, and optimizing its performance involves adjusting hyperparameters like the number of trees, maximum tree depth, and the number of features considered at each split.

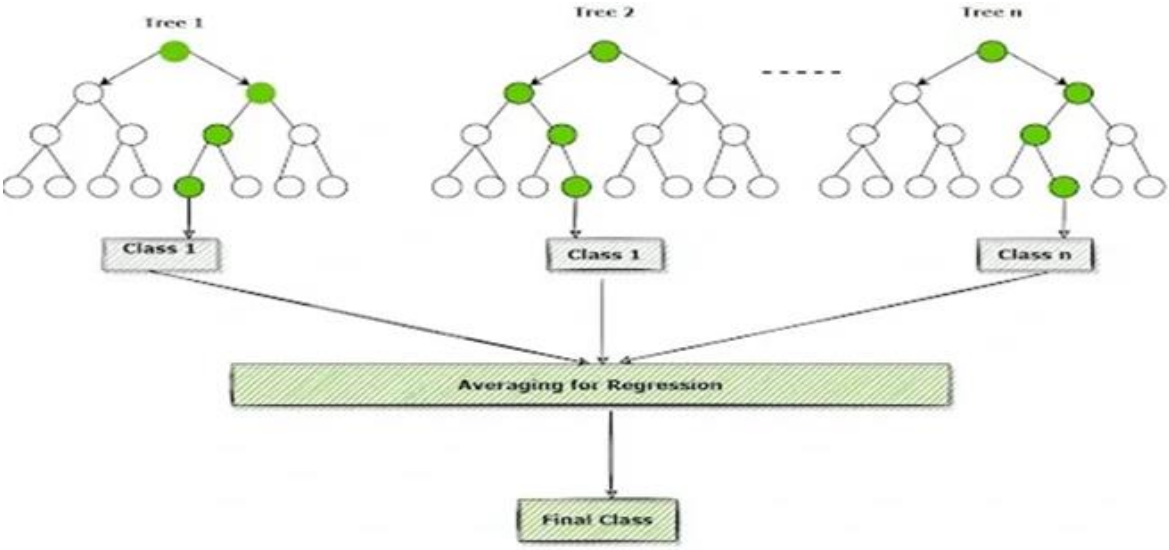


Figure 1: Random Forest Model

Gradient Boosting Machine (GBM) is a widely used and effective ensemble learning algorithm in machine learning applications (Aggarwal et al., 2018). It combines multiple weak learners to create a strong predictive model, making it suitable for complex and large datasets. GBM is versatile and can handle both numerical and categorical data, as well as noisy and missing data (Jordanov et al., 2017). It excels in situations that demand high accuracy and can capture complex relationships between features. Moreover, GBM finds applications in various tasks such as classification and regression (Kawaguchi et al., 2021).

When utilizing GBM, important parameters to consider include the learning rate, number of trees, maximum tree depth, and subsampling fraction (Gong et al., 2019). These parameters can be adjusted to optimize the model's performance. The subsampling fraction determines the proportion of data used to train each tree in the ensemble (Natekin et al., 2013). Selecting a fitting subsampling division makes contrast expect overfitting and update the algorithm's execution. Experimentation and cross-validation are typically employed to find the optimal subsampling fraction value.

GBM offers several advantages as an ensemble learning algorithm. It effectively handles complex and large datasets, is robust against noise and missing data, and can be tuned for optimal performance (Feng et al., 2020). It’s widely applied in various domains due to its ability to achieve high accuracy and versatility in classification and regression tasks. However, it is important to note

that GBM can be computationally expensive and may require significant computing resources for training large models. Additionally, without proper tuning or regularization, GBM is susceptible to overfitting (Park et al., 2018).

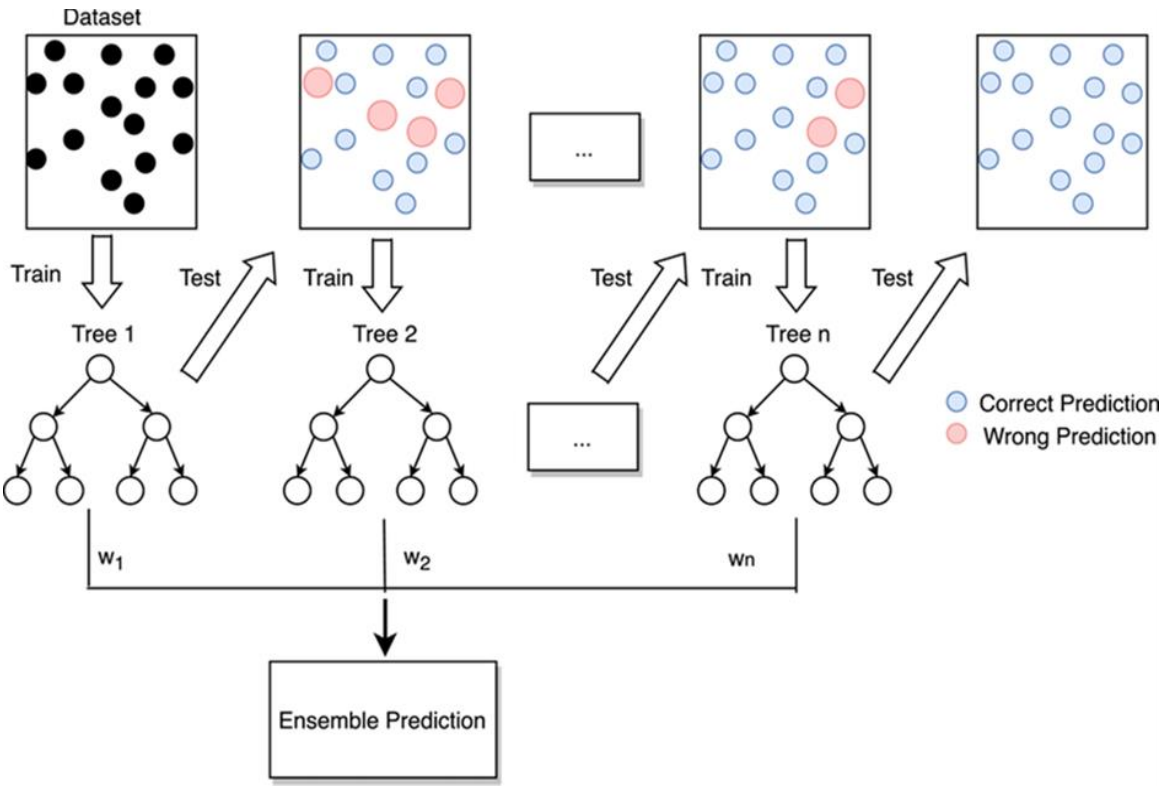


Figure 2: Gradient Boosting Machine

Multi-layer Perceptron (MLP) classifiers are widely used in machine learning as a type of artificial neural network that consists of multiple layers of interconnected nodes or "neurons". Multi-layer Perceptron classifiers are known for their remembrance complex designs and cope up with miscellaneous tasks

and the key parameters to consider when working with MLP classifiers include the number of hidden layers, the number of neurons in each layer, the choice of activation function, and the learning rate. These parameters affect the model's capacity to capture and represent data designs. MLP classifiers offer advantages such as their ability to handle non-linear relationships, flexibility in modelling complex decision boundaries, and capability to process both numerical and categorical data. However, they can be computationally requesting, particularly for huge and high dimensional datasets, and may require cautious regularization to anticipate overfitting. While Implementing MLP classifiers involves using libraries and programming languages such as Python with popular frameworks like scikit-learn or TensorFlow. Evaluating the performance of MLP classifiers involves using appropriate metrics such as accuracy, precision, recall, and F1 score on a separate test dataset, and tuning hyperparameters to optimize the model's performance.

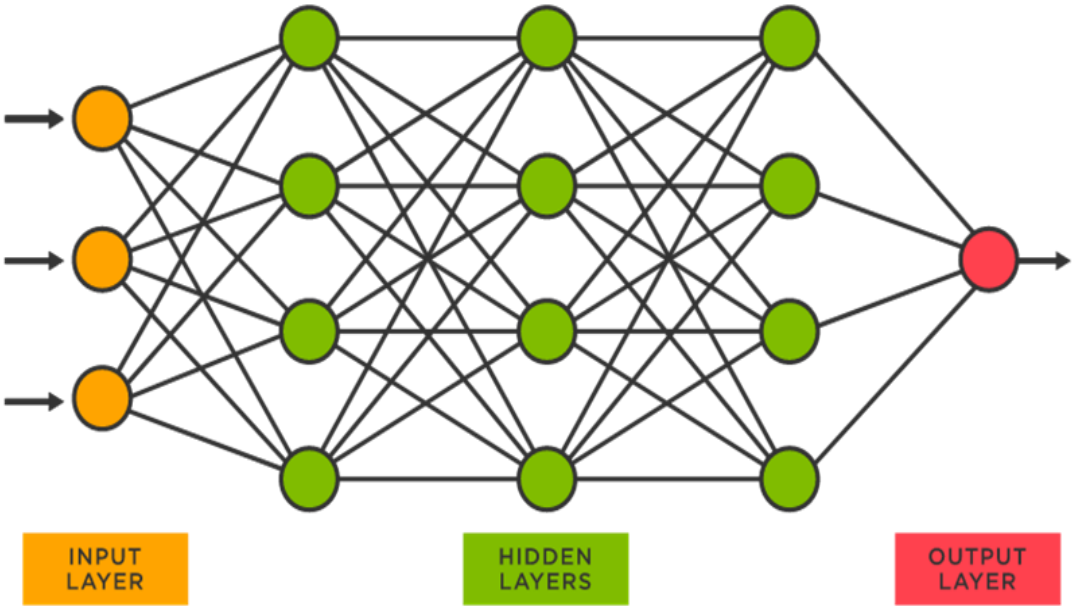


Figure 3: Artificial Neural Networks

Question 3: How does the size and nature of the dataset impact the effectiveness of ensemble techniques, such as stacking, for addressing class imbalance in groundwater quality prediction models?

Dataset creation is an essential step in addressing Question 3 of this research study. Various variations of the dataset will be created to explore different scenarios, encompassing different dataset sizes and levels of class imbalance. The dataset size manipulation will involve randomly selecting subsets of different sizes from the original dataset. Additionally, class imbalance will be introduced by either oversampling the minority class or under sampling the majority class.

As same approach described in Question 2, stacked models using ensemble techniques, such as stacking, will be developed for each variation of the dataset. The base models will be trained on the modified datasets, considering the specific dataset size and class distribution characteristics. The ensemble model will be constructed by combining the predictions from the base models.

The performance of the stacked ensemble models will be thoroughly evaluated for each dataset variation. Evaluation metrics such as accuracy, precision, recall, F1-score, and the area under the receiver operating characteristic curve (AUC-ROC) will be employed. These metrics will provide a comprehensive assessment of the model's performance, considering the dataset variations, and will aid in understanding the effectiveness of the stacked model approach in different scenarios.

CHAPTER FOUR

FINDINGS AND ANALYSIS

This chapter describes the procedures and methods used for data collection and analysis for this culminating experience project. The Telangana Ground Water Department Dataset is available from 2018 to 2020. For this project's purposes, it is important to examine Pre-Monsoon, and post-Monsoon Water Quality details in Telangana. Data Contains the below columns: | district | mandal | village | temp_id | long_gis | lat_gis | gwl | season | pH | E.C | TDS | CO3 | HCO3 | Cl | F | NO3 | SO4 | Na | K | Ca | Mg | T.H | SAR | Classification | RSC meq / L | Classification. Some of the attributes have been deleted. This dataset consists of observations of 2018 to 2020 and contains 16 variables. The software used to collect the data and perform the analysis was Microsoft Excel. The computer program utilized to gather the information and perform the examination was Microsoft Excel. The variables of interest for which data was collected were attribute names for each chemical with their range value, from 2018 to 2020. These attributes name contains descriptions and range of value like, SO4 sulphates, Na Sodium, K potassium and Ca Calcium. The remainder of this chapter covers the data collection, analysis, and findings.

Table 1: Ground Water Department Dataset

S.NO	Attribute name	Description	Range of value
1	SO4	Sulphates	1 to 860
2	Na	Sodium	5.07 to 748.1
3	K	Potassium	0.07 to 354.6
4	Ca	Calcium	1.2 to 640.0
5	Mg	Magnesium	4.86 to 457.02
6	TH	Total hardness	39.99 to 3479.22
7	HCO3	Bicarbonate	10.13 to 1070
8	Cl	Chloride	10 to 2480
9	F	Fluoride	0.04 to 7.7
10	NO3	Nitrate	0.02 to 1028
11	pH	Potential of Hydrogen	6.11 to 10.59
12	EC	Electrical conductivity	102 to 9499
13	TDS	Total dissolved solids	65.28 to 6079.36
14	SAR	Sodium adsorption ratio	0.181178 to 31.435063
15	RSC	Residual Free Chlorine	-59.584539 to 18.200822
16	Classification		

Data Information

Telangana Ground Water Department dataset provides valuable information about pre-monsoon water quality in Telangana, allowing for the assessment of the suitability of water for various purposes and the identification of potential health hazards (Delahaut et al., 2021). In this dataset attribute name has chemical name with their descriptions, which contains, SO4 provides information about the concentration of sulphate in the water. Elevated levels of sulphate in water can cause gastrointestinal issues and other health problems (Iyuke et al., 2012). Also, Sodium (Na) in water can cause high blood (Khan et al., 2016). Potassium in water can cause health problems (Walia et al., 2007). Calcium is a crucial mineral for bone (Eby et al., 2006). Magnesium is an important mineral for various body functions, but elevated levels of magnesium in water can cause health problems (Eby et al., 2006). Bicarbonate (HCO3) is an important buffer in the body, but high levels of bicarbonate in water can cause health problems (Casey et al., 2008). Chloride (Cl) causes gastrointestinal health problems in the human body (Shahnaz et al., 2013). Fluoride (F) is plays major role in human dental health (Abouleish et al., 2016). Nitrate (NO3) high concentration can impact the young children's health drastically (Gruener et al., 2013).

Programming and Libraries

Python programming language which is useful for building a various mobile applications and websites, scientific computing and exploring data. (Hao & Ho, 2019).

Data Set Analysis

The data set contains the various parameter samples with their ranges. PH Potential of Hydrogen, which can rely in the range 6.11 to 10.59. The electrical conductivity (E.C.) examine the electric current available in the water, which can range from 102 to 9499. The TDS column provides information on the total dissolved solids in the water sample, which can range from 65.28 to 6079.36. SAR (Sodium Adsorption Ratio). The selected columns were SO4, Na, K, Ca, Mg, TH, HCO3, Cl, F, NO3, Ph, EC, TDS, SAR, RSC, meg/L, and Classification. The range of Sodium Adsorption Ratio can vary between 0.181178 to 31.435063. Elevated levels of Sodium Adsorption Ratio can cause soil degradation and can reduce the productivity of crops (Quillérou et al., 2014). RSC provides information about the residual sodium carbonate in the water. Elevated level of Residual Sodium Carbonate can cause soil degradation and reduce crop yields (Singh et al., 2022). The classification column provides information about water classification based on its suitability for irrigation. The water quality classification can be suitable, marginally suitable, or unsuitable for irrigation (Zahraw et al., 2017).

This research used the Anaconda platform and Jupyter notebooks. It is an integrated development environment used primarily by Python programmers for a wide range of essential tools for building machine learning models, Firstly, we have installed the panda's library to read the dataset and then load six different CSV files into separate data frames. Each CSV file corresponds to pre-monsoon and post-monsoon water quality data from the years 2018, 2019, and 2020, and each pre-monsoon and post-monsoon dataset is loaded into a separate data frame. The classification SO4, Na, K, Ca, Mg, TH, HCO3, Cl, F, NO3, pH, EC, TDS, SAR, RSC meq/L were the selected columns. The Classification column represents the classification of the water sample based on its quality. The suffixes like -2 and +2 were used to represent the charge of the ions in the column names. Once the CSV files have been loaded into their respective data frames, the columns are being renamed to abbreviations for specific chemical elements and water quality parameters. We are concatenating six data frames into one single data frame and the resulting data frame will have all the rows of the original data frames stacked on top of each other, with the same column headers. The original code values in the 'Classification' column with their respective quality mapped with the levels.

The classification codes C2S1, C3S1, and C4S1 are used to indicate the suitability of water for irrigation purposes based on its salinity and sodium content. C2S1 indicates water that is safe for irrigation purposes, C3S1 predict that water with high level of salinity and low sodium content, that can be used for

all types of soil with minimal risk of exchangeable sodium and classification codes C4S1 indicates water with high level of salinity and low level of sodium content, which is appropriate for plants with good salt tolerance but unsuitable for irrigation in soils with restricted drainage. We have split the data into training and testing sets using the train_test_split () function from sklearn.model_selection module. This allows us to train our machine learning model on a subset of the data and evaluate its performance on a separate subset. Based on the target variable, we chose the Random Forest and LightGBM machine learning calculations from the classification. We are importing the Random Forest Classifier from sklearn. ensemble module, importing the LightGBM library and initializing a LightGBM classifier, which is an algorithm used for classification problems.

Results

In this section, we are going to compare the findings that we achieved from various machine learning algorithms such as Light Gradient Boosting Machine, Random Forest, stack models using Artificial Neural Network by building classification codes and report as mentioned above and compared those results with the results from the previous Telangana Studies. The classification report is a function in scikit-learn library that provides report on the performance of a classification model and includes important metrics such as precision, recall, and F1 score. This report analysed the classification performance and the room

of improvement. This Classification performs consistent method of evaluating water quality for irrigations. This classification model has the set of rules and principles to implement across the regions.

Finally, we are implementing Random Forest Classifier model with training set using the fit () method and this will allow the model to recall the patterns in the data and make predictions accordingly. Random forecast Classifier has capacity to handle complex non liner connection between input factors and target variables.

Table 2: Random Forest Classifier

Algorithm	Additional Parameters Information	Accuracy %
Random Forest	n_estimators=100, criterion='gini', max_features='sqrt	97%

Now, we are implementing the library such as LightGBM, to initializing the classifiers, and fitting the classifier on the training data X and Y using the Fit method. Finally, we are using predict method of Light Classifier to get the predictions on Test X and adding them in variable Prediction LGBM.By Implementing the predict method will generate the accurate performance of the

LightGBM classifier on the Test X.The LightGBM has major advantage to customize the training process.

Table 3: Accuracy

Algorithm	Additional Parameters information	Accuracy %
Light GBM	max_depth= -1, learning_rate: 0.1	97%

The classification report provides for each class in the target variable and provides an overall accuracy. This report can be used to evaluate the performance of the classification model and identify areas for improvement.

Table 4: Random Forest for Evaluation

Each Class	Precision	Recall	F1 Score	Support
Excellent	1.00	1.00	1.00	1
Good	1.00	1.00	1.00	92
Marginal	0.83	0.42	0.56	12
Permissible	0.97	1.00	0.98	289
Poor	0.95	0.95	0.95	37
Very Poor	0.83	0.71	0.77	7

Table 5: Light GBM

Each Class	Precision	Recall	F1 Score	Support
Excellent	1.00	1.00	1.00	1
Good	1.00	1.00	1.00	92
Marginal	0.78	0.58	0.67	12
Permissible	0.98	0.99	0.99	289
Poor	0.95	0.95	0.95	37
Very Poor	0.83	0.71	0.77	7

We followed the stacked model methodology by concatenating predictions probability of the random forest model and Light GBM model applied stack model predictions by passing by values as input parameters into an Artificial Neural Network (ANN) multiclass classifier to improve the accuracy of the final prediction.

Table 6: Accuracy

Algorithm	Additional Parameters information	Accuracy %
ANN	num_hidden_layers: 1-3 hidden_dim: 32-256	98%

The ANN used three hidden layers with 100, 50, and 25 nodes with the activation function used in the hidden layers is ReLU. The performance of the final prediction is then evaluated using several metrics such as accuracy, precision, recall, confusion matrix, and classification report, these metrics are calculated provided by the 'sklearn.metrics' module used by Artificial Neuro Network.

Table 7: Artificial Neural Networks

Each Class	Precision	Recall	F1 Score	Support
Excellent	1.00	1.00	1.00	1
Good	1.00	1.00	1.00	92
Marginal	0.73	0.67	0.70	12
Permissible	0.98	0.99	0.99	289
Poor	0.97	0.95	0.96	37
Very Poor	0.86	0.86	0.86	7

This chapter has presented the analysis process that provides insights into the three research questions relating to the Telangana Ground Water Department dataset provides valuable information about water quality in Telangana, allowing for the assessment of the suitability for various purposes and the identification of potential health problems. Chapter 5 next, provides a discussion, conclusion, and recommendations for further research

CHAPTER FIVE

DISCUSSION

This last chapter will discuss the project findings, and provide a conclusion, and areas for further study for each of the three questions.

Question 1 focuses on the various parameter's values for ground water quality like locations, pre and post monsoon of ground water quality information with different years. The results from question 1 found that groundwater quality data from various locations and time periods were analysed, and it was found that pre-monsoon water quality has higher levels of Total hardness (TH) which has range of 39.99 to 3479.22 and total dissolved solids in the water sample, which can range from 65.28 to 6079.36. SAR (Sodium Adsorption Ratio) and Nitrate (NO3) concentration can vary between 0.02 to 1028. The study also utilized Random Forest, Light Gradient Boosting Machine and Multi-layer Perceptron machine learning algorithms to predict groundwater quality based on water salinity and sodium content suitability.

Question 2 focuses on the Machine Learning algorithms are trained based on the historical data by learning the patterns of the features. The results of question 2 found that Stacked model achieved 98% accuracy by assembling the random forest and LGBM with MLP Classifier. The stacked models achieve the best performance with use of different classifiers in terms of accuracy (Random forecast 97%, Light GBM 97%, ANN 98%) to predict the water quality by

collecting the data from different regions and climatic conditions based on the suitability of water salinity and sodium content. A stacked model was created by combining the random forest and LGBM algorithms with MLP Classifier.

Question 3 examined on the classes such as excellent, good, permissible, poor, and very poor, imbalanced score has been increased when we applied in the stacked models' predictions. These imbalanced data increased in the stacked model methodology by concatenating predictions probability of the random forest model and Light GBM model applied stack model predictions by passing by values as input parameters into an Artificial Neural Network (ANN) multiclass classifier to improve the accuracy of the final prediction. These Final Predictions are by passing values of 1 Random Forest, LGBM and ANN increase score of 1 that is excellent prediction. Again, passing support 92 these three stacked models increase of 1 that is good. By passing support 289 RF, LGBM and ANN scored 0.99 which is permissible. Passing support 37 RF, LGBM and ANN scored 0.95, 0.95, 0.96 which is poor. Finally, to know the imbalanced of class in water quality Passing support 7 predicted that RF, LGBM scored 0. 77 and ANN scored 0.86 which are very poor for the predictions.

Conclusion

Overall, this study highlights the importance of using advanced data analysis techniques to inform better management and protection of groundwater resources. This study suggest area to be researched to Optimize

hyperparameters for each base model (random forest, light GBM, and ANN) and for the overall stacked model. Hyperparameter tuning can significantly affect the model's performance, so it's essential to find the best combination for stacked models.

Areas of Further Studies

For Future studies could explore the relationship between climate and weather patterns and groundwater quality in more detail. Additionally, it may be useful to examine the performance of other machine learning algorithms on predicting groundwater quality parameters. Further research may perform an in-depth analysis of the factors that focus on the need to improve the accuracy of the machine learning models to predict the ground water quality based on the various soils and climatic conditions. These studies could explore additional algorithms to stacked methods for improving the accuracy of the models.

APPENDIX A

CODES

Importing Necessary Libraries

In [1]:
```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
```

Loading 2018,2019,2020 pre and post Monsoon

In [2]:
```
df_pr_2018 = pd.read_csv("ground_water_quality_2018_pre.csv")
df_pr_2019 = pd.read_csv("ground_water_quality_2019_pre.csv")
df_pr_2020 = pd.read_csv("ground_water_quality_2020_pre .csv")
df_ps_2018 = pd.read_csv("ground_water_quality_2018_post.csv")
df_ps_2019 = pd.read_csv("ground_water_quality_2019_post.csv")
df_ps_2020 = pd.read_csv("ground_water_quality_2020_post.csv")
```

In [3]:
```
df_pr_2018.columns
```

Out[3]:
```
Index(['sno', 'district', 'mandal', 'village', 'temp_id', 'long_gis',
       'lat_gis', 'gwl', 'season', 'pH', 'E.C', 'TDS', 'CO3', 'HCO3', 'Cl',
       'F', 'NO3 ', 'SO4', 'Na', 'K', 'Ca', 'Mg', 'T.H', 'SAR',
       'Classification', 'RSC  meq  / L', 'Classification.1'],
      dtype='object')
```

Selecting necessary Columns

In [4]:
```
df_2019_ps = df_ps_2019[['SO4-2','Na+','K+','Ca+2','Mg+2','T.H','HCO_ - ','Cl -','F -','NO3- ','pH', 'EC', 'TDS','SAR','RSC  meq
df_2020_ps = df_ps_2020[['SO4','Na','K','Ca','Mg','T.H','HCO3','Cl','F','NO3 ','pH', 'E.C', 'TDS','SAR','RSC  meq  / L','Classif
df_2018_ps = df_ps_2018[['SO4','Na','K','Ca','Mg','T.H','HCO3','Cl','F','NO3 ','pH', 'E.C', 'TDS','SAR','RSC  meq  / L','Classif
df_2020_pr = df_pr_2020[['SO4','Na','K','Ca','Mg','T.H','HCO3','Cl','F','NO3 ','pH', 'E.C', 'TDS','SAR','RSC  meq  / L','Classif
df_2019_pr = df_pr_2019[['SO4','Na','K','Ca','Mg','T.H','HCO3','Cl','F','NO3 ','pH', 'E.C', 'TDS','SAR','RSC  meq  / L','Classif
df_2018_pr = df_pr_2018[['SO4','Na','K','Ca','Mg','T.H','HCO3','Cl','F','NO3 ','pH', 'E.C', 'TDS','SAR','RSC  meq  / L','Classif
```

Dropping Null values

In [5]:
```
df_2019_pr = df_2019_pr.dropna()
df_2018_pr = df_2018_pr.dropna()
```

Changing the Column Names

In [6]:
```
df_2019_ps.columns.values[0:16] =['SO4','Na', 'K', 'Ca', 'Mg', 'TH', 'HCO3', 'Cl', 'F', 'NO3', 'pH', 'EC', 'TDS','SAR', 'RSC', '
df_2020_ps.columns.values[0:16] =['SO4','Na', 'K', 'Ca', 'Mg', 'TH', 'HCO3', 'Cl', 'F', 'NO3', 'pH', 'EC', 'TDS','SAR', 'RSC', '
df_2018_ps.columns.values[0:16] =['SO4','Na', 'K', 'Ca', 'Mg', 'TH', 'HCO3', 'Cl', 'F', 'NO3', 'pH', 'EC', 'TDS','SAR', 'RSC', '
df_2020_pr.columns.values[0:16] =['SO4','Na', 'K', 'Ca', 'Mg', 'TH', 'HCO3', 'Cl', 'F', 'NO3', 'pH', 'EC', 'TDS','SAR', 'RSC', '
df_2019_pr.columns.values[0:16] =['SO4','Na', 'K', 'Ca', 'Mg', 'TH', 'HCO3', 'Cl', 'F', 'NO3', 'pH', 'EC', 'TDS','SAR', 'RSC', '
df_2018_pr.columns.values[0:16] =['SO4','Na', 'K', 'Ca', 'Mg', 'TH', 'HCO3', 'Cl', 'F', 'NO3', 'pH', 'EC', 'TDS','SAR', 'RSC', '
```

Concatinating all Data Frames

In [7]:
```
frames = [df_2019_ps, df_2020_ps, df_2018_ps,df_2020_pr,df_2019_pr,df_2018_pr]
df = pd.concat(frames)
df.head(3)
```

Out[7]:

	SO4	Na	K	Ca	Mg	TH	HCO3	Cl	F	NO3	pH	EC	TDS	SAR	RSC	Classification
0	377.0	273.0	113.0	80.0	82.654	539.860197	320.0	340.0	0.64	66.435000	8.32	2355.0	1507.20	5.106509	-3.797204	C4S2
1	52.0	89.0	14.0	48.0	4.862	139.991776	180.0	40.0	2.21	38.049136	8.3	565.0	361.60	3.196982	1.200164	C2S1
2	43.0	39.0	10.0	40.0	24.310	199.958882	190.0	20.0	0.55	41.270227	8.12	766.0	490.24	1.199130	-0.199178	C3S1

Shape of the Data Set

In [8]:
```
df.shape
```

Out[8]:
```
(2197, 16)
```

Displaying Classification Column Names

In [9]:
```
df['Classification'].value_counts()
```

Out[9]:
```
C3S1               1404
C2S1                477
C4S1                138
C4S2                 72
C3S2                 44
C4S4                 16
C4S3                 13
C1S1                 12
C3S3                 10
0.G                   4
OG                    3
C3S4                  2
C2S2                  1
BELOW THE GRAPH       1
Name: Classification, dtype: int64
```

Displaying Classification column dtype

In [10]:
```
df['Classification'].dtype
```

Out[10]:
```
dtype('O')
```

Converting different categories by using mapping

In [11]:
```
dict = {'C1S1':'Excellent',
        'C1S2':'Good','C2S1':'Good','C2S2':'Good',
       'C1S3':'Permissible','C3S1':'Permissible',
       'C2S3':'Marginal','C3S2':'Marginal','C3S3':'Marginal',
       'C1S4':'Poor','C2S4':'Poor','C3S4':'Poor','C4S1':'Poor','C4S2':'Poor',
       'C4S3':'Very Poor','C4S4':'Very Poor'}
```

In [12]:
```
df['Classification'] = df['Classification'].map(dict)
```

Checking value counts in the classification column

In [13]:
```
df['Classification'].value_counts()
```

Out[13]:
```
Permissible    1404
Good            478
Poor            212
Marginal         54
Very Poor        29
Excellent        12
Name: Classification, dtype: int64
```

Checking Null values

In [14]:
```
df.isnull().sum().sum()
```

Out[14]:
```
8
```

Shape of the Data Set

In [8]:
```
df.shape
```

Out[8]:
```
(2197, 16)
```

Displaying Classification Column Names

In [9]:
```
df['Classification'].value_counts()
```

Out[9]:
```
C3S1               1404
C2S1                477
C4S1                138
C4S2                 72
C3S2                 44
C4S4                 16
C4S3                 13
C1S1                 12
C3S3                 10
0.G                   4
OG                    3
C3S4                  2
C2S2                  1
BELOW THE GRAPH       1
Name: Classification, dtype: int64
```

Displaying Classification column dtype

In [10]:
```
df['Classification'].dtype
```

Out[10]:
```
dtype('O')
```

Converting different categories by using mapping

In [11]:
```
dict = {'C1S1':'Excellent',
        'C1S2':'Good','C2S1':'Good','C2S2':'Good',
       'C1S3':'Permissible','C3S1':'Permissible',
       'C2S3':'Marginal','C3S2':'Marginal','C3S3':'Marginal',
       'C1S4':'Poor','C2S4':'Poor','C3S4':'Poor','C4S1':'Poor','C4S2':'Poor',
       'C4S3':'Very Poor','C4S4':'Very Poor'}
```

In [12]:
```
df['Classification'] = df['Classification'].map(dict)
```

Checking value counts in the classification column

In [13]:
```
df['Classification'].value_counts()
```

Out[13]:
```
Permissible    1404
Good            478
Poor            212
Marginal         54
Very Poor        29
Excellent        12
Name: Classification, dtype: int64
```

Checking Null values

In [14]:
```
df.isnull().sum().sum()
```

Out[14]:
```
8
```

Dropping Null values rows

In [15]:
```
df = df.dropna()
```

Checking null values

In [16]:
```
df.isnull().sum().sum()
```

Out[16]:
```
0
```

Checking datatypes

In [17]:
```
df.dtypes
```

Out[17]:
```
SO4               float64
Na                float64
K                 float64
Ca                float64
Mg                float64
TH                float64
HCO3              float64
Cl                float64
F                 float64
NO3               float64
pH                 object
EC                float64
TDS               float64
SAR               float64
RSC               float64
Classification     object
dtype: object
```

Replacing special chars and datatype of the ph column

In [18]:
```
df['pH']=df['pH'].replace('8..05','8.05')
df['pH']=df['pH'].astype('float')
df['pH'].dtypes
```

Out[18]:
```
dtype('float64')
```

Displaying the Quantitative columns statistics

In [19]:
```
df.describe()
```

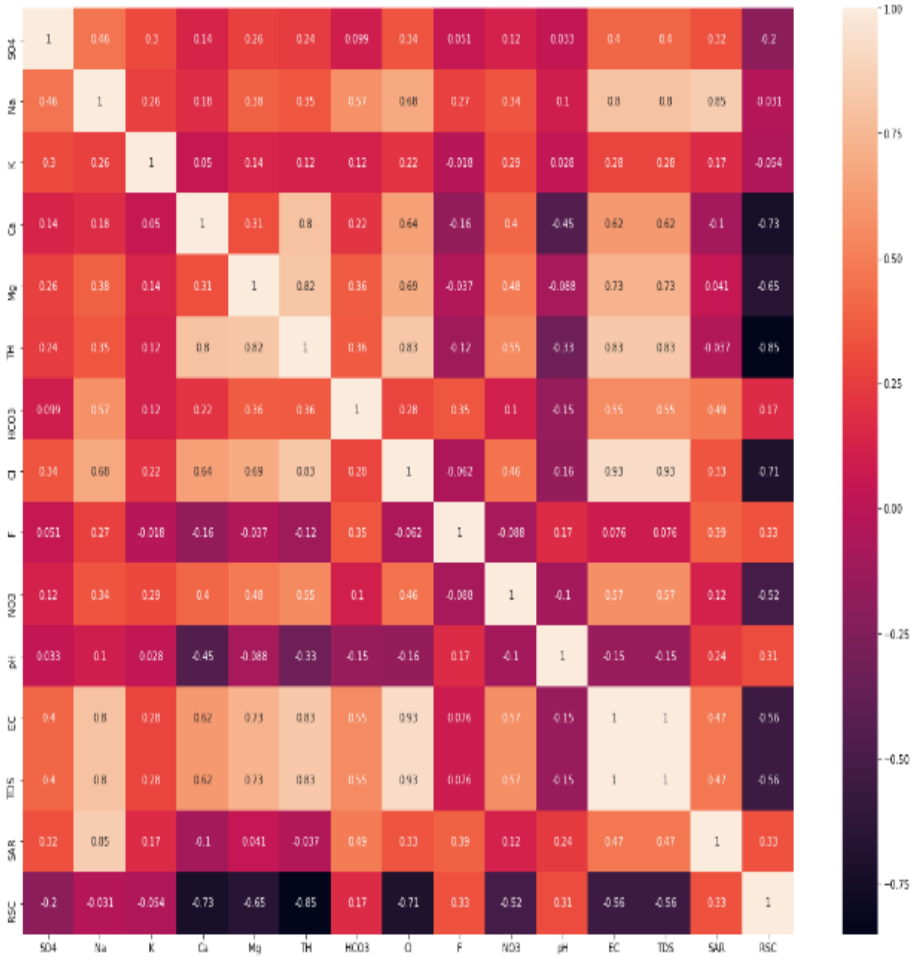
Out[19]:

	SO4	Na	K	Ca	Mg	TH	HCO3	Cl	F	NO3	pH	
count	2189.000000	2189.000000	2189.000000	2189.000000	2189.000000	2189.000000	2189.000000	2189.000000	2189.000000	2189.000000	2189.000000	2189.0
mean	40.174043	122.376853	7.962316	75.775788	50.857176	398.415941	279.067732	192.032692	1.124188	73.301036	7.948968	1314.9
std	53.152437	103.917804	20.270677	58.778972	36.826592	242.038753	129.842964	185.105649	0.780462	93.953025	0.451728	778.8
min	1.000000	5.076154	0.070000	1.200000	4.862000	39.991776	10.131754	10.000000	0.040000	0.026574	6.110000	102.0
25%	14.000000	56.940840	2.000000	40.000000	24.310000	239.942434	186.186772	70.000000	0.620000	17.716000	7.660000	789.0
50%	23.000000	94.000000	3.220000	64.000000	43.758000	339.958882	264.196846	140.000000	0.940000	40.867591	7.980000	1149.0
75%	41.000000	154.000000	6.000000	96.000000	68.068000	499.893092	353.209021	260.000000	1.420000	94.619545	8.250000	1625.0
max	860.000000	748.100000	354.600000	640.000000	457.026000	3479.226974	1070.000000	2480.000000	7.700000	1026.000000	10.590000	9499.0

Finding the correlation

In [20]:
```
plt.figure(figsize=(20,15))
sns.heatmap(df.corr(),annot=True)
plt.show()
```

```
<ipython-input-20-8a86e50fe70c>:2: FutureWarning: The default value of numeric_only in DataFrame.corr is deprecated. In a futur
e version, it will default to False. Select only valid columns or specify the value of numeric_only to silence this warning.
  sns.heatmap(df.corr(),annot=True)
```



Converting data into independent and dependent variables

In [21]:
```
X = df.drop(['Classification'],axis=1)
y = df['Classification']
```

Splitting Data into Train 80% and Test 20%

In [22]:
```
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.2, random_state = 42)
```

Impoting Evaluation Matrix libraries

In [23]:
```
from sklearn.metrics import confusion_matrix
from sklearn.metrics import ConfusionMatrixDisplay
from sklearn.metrics import classification_report
```

Algorithms

1. Random Forest

In [24]:
```
from sklearn.ensemble import RandomForestClassifier
```

In [25]:
```
model = RandomForestClassifier()
model.fit(X_train,y_train)
```

Out[25]:
```
RandomForestClassifier()
```

In a Jupyter environment, please rerun this cell to show the HTML representation or trust the notebook.
On GitHub, the HTML representation is unable to render, please try loading this page with nbviewer.org.

In [26]:
```
y_pred = model.predict(X_test)
```

In [27]:
```
y_test.value_counts()
```

Out[27]:
```
Permissible    289
Good            92
Poor            37
Marginal        12
Very Poor        7
Excellent        1
Name: Classification, dtype: int64
```

In [28]:
```
pd.Series(y_pred).value_counts()
```

Out[28]:
```
Permissible    295
Good            92
Poor            37
Marginal         7
Very Poor        6
Excellent        1
dtype: int64
```

Classification Model

In [29]:
```
print("\033[1m Classification Report: \033[0m \n")
print(classification_report(y_test, y_pred))
```

```
 Classification Report:

              precision    recall  f1-score   support

   Excellent       1.00      1.00      1.00         1
        Good       1.00      1.00      1.00        92
    Marginal       0.86      0.50      0.63        12
Permissible       0.98      1.00      0.99       289
        Poor       0.95      0.95      0.95        37
   Very Poor       0.83      0.71      0.77         7

    accuracy                           0.97       438
   macro avg       0.94      0.86      0.89       438
weighted avg       0.97      0.97      0.97       438
```

2. LGBM

In [30]:
```
import lightgbm as lgb
lgbm_Classifier = lgb.LGBMClassifier()
lgbm_Classifier.fit(X_train,y_train)

# Make final predictions
pred_lgbm = lgbm_Classifier.predict(X_test)
```

```
C:\Users\divesh\anaconda3\lib\site-packages\dask\dataframe\utils.py:369: FutureWarning: pandas.Int64Index is deprecated and wil
l be removed from pandas in a future version. Use pandas.Index with the appropriate dtype instead.
  _numeric_index_types = (pd.Int64Index, pd.Float64Index, pd.UInt64Index)
C:\Users\divesh\anaconda3\lib\site-packages\dask\dataframe\utils.py:369: FutureWarning: pandas.Float64Index is deprecated and w
ill be removed from pandas in a future version. Use pandas.Index with the appropriate dtype instead.
  _numeric_index_types = (pd.Int64Index, pd.Float64Index, pd.UInt64Index)
C:\Users\divesh\anaconda3\lib\site-packages\dask\dataframe\utils.py:369: FutureWarning: pandas.UInt64Index is deprecated and wi
ll be removed from pandas in a future version. Use pandas.Index with the appropriate dtype instead.
  _numeric_index_types = (pd.Int64Index, pd.Float64Index, pd.UInt64Index)
```

In [31]:
```
y_test.value_counts()
```

Out[31]:
```
Permissible    289
Good            92
Poor            37
Marginal        12
Very Poor        7
Excellent        1
Name: Classification, dtype: int64
```

In [32]:
```
pd.Series(pred_lgbm).value_counts()
```

Out[32]:
```
Permissible    293
Good            92
Poor            37
Marginal         9
Very Poor        6
Excellent        1
dtype: int64
```

Classification Model

In [33]:
```
print("\033[1m Classification Report: \033[0m \n")
print(classification_report(y_test, pred_lgbm))
```

Classification Report:

	precision	recall	f1-score	support
Excellent	1.00	1.00	1.00	1
Good	1.00	1.00	1.00	92
Marginal	0.78	0.58	0.67	12
Permissible	0.98	0.99	0.99	289
Poor	0.95	0.95	0.95	37
Very Poor	0.83	0.71	0.77	7
accuracy			0.97	438
macro avg	0.92	0.87	0.89	438
weighted avg	0.97	0.97	0.97	438

3. MLPClassifier

In [34]:
```
from sklearn.neural_network import MLPClassifier
from sklearn.metrics import classification_report, accuracy_score, precision_score, recall_score, f1_score, confusion_matrix
xg_pred = model.predict_proba(X_test)
lgbm_pred = lgbm_Classifier.predict_proba(X_test)
```

Stack predictions using ANN

In [35]:
```
stacked_pred = np.concatenate((xg_pred, lgbm_pred), axis=1)
ann = MLPClassifier(hidden_layer_sizes=(100, 50, 25), activation='relu', solver='adam', max_iter=1000)
ann.fit(stacked_pred, y_test)
final_pred = ann.predict(stacked_pred)
```

Evaluate performance

In [36]:
```
accuracy = accuracy_score(y_test, final_pred)
precision = precision_score(y_test, final_pred, average='weighted')
recall = recall_score(y_test, final_pred, average='weighted')
f1 = f1_score(y_test, final_pred, average='weighted')
conf_matrix = confusion_matrix(y_test, final_pred)
class_report = classification_report(y_test, final_pred)
```

Print evaluation metrics

In [38]:
```
print('Accuracy: ', accuracy)
print('Precision: ', precision)
print('Recall: ', recall)
print('F1-Score: ', f1)
print('Confusion Matrix:\n', conf_matrix)
print('Classification Report:\n', class_report)
```

```
Accuracy:  0.9817351598173516
Precision:  0.981066548572042
Recall:  0.9817351598173516
F1-Score:  0.9809199263798006
Confusion Matrix:
 [[  1   0   0   0   0   0]
 [  0  92   0   0   0   0]
 [  0   0   8   4   0   0]
 [  0   0   1 288   0   0]
 [  0   0   0   1  35   1]
 [  0   0   0   0   1   6]]
Classification Report:
              precision    recall  f1-score   support

   Excellent       1.00      1.00      1.00         1
        Good       1.00      1.00      1.00        92
    Marginal       0.89      0.67      0.76        12
 Permissible       0.98      1.00      0.99       289
        Poor       0.97      0.95      0.96        37
   Very Poor       0.86      0.86      0.86         7

    accuracy                           0.98       438
   macro avg       0.95      0.91      0.93       438
weighted avg       0.98      0.98      0.98       438
```

REFERENCES

Ahmad, A. (2021). Compressive Strength prediction of fly ash-based geoplymer concerte via machine leaning techniques. Retrieved from https://doi.org/10.1016/j.cscm.2021.e00840

Azrour, M., Mabrouki, J., Fattah, G., Guezzaz, A., & Aziz, F. (2021). Machine learning algorithms for efficient water quality prediction. *Modeling Earth Systems and Environment*, *8*(2), 2793–2801. https://doi.org/10.1007/s40808-021-01266-6

Adimalla, N. Application of the Entropy Weighted Water Quality Index (EWQI) and the Pollution Index of Groundwater (PIG) to Assess Groundwater Quality for Drinking Purposes: A Case Study in a Rural Area of Telangana State, India. *Arch Environ Contam Toxicol* **80**, 31–40 (2021). https://doi.org/10.1007/s00244-020-00800-4

Adimalla, N. (2021). Application of the Entropy Weighted Water Quality Index (EWQI) and the Pollution Index of Groundwater (PIG) to assess groundwater quality for drinking purposes: a case study in a rural area of Telangana State, India. *Archives of Environmental Contamination and Toxicology*, *80*(1), 31–40. https://doi.org/10.1007/s00244-020-00800-4

Anne-Laure Boulesteix, Sike Janitza, Jochen Kruppa, Inke R. Kong “Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics” (2012). https://doi.org/10.1002/widm.1072

Alexey Natekin,“Gradient boosting machines, a tutorial, 2013 https://doi.org/10.3389/fnbot.2013.00021

Boulesteix, A., Janitza, S., Kruppa, J., & König, I. R. (2012). Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics. *Random Forest Methodology*, *2*(6), 493–507. https://doi.org/10.1002/widm.1072

Bhakar, P., Singh, A.P. Groundwater Quality Assessment in a Hyper-arid Region of Rajasthan, India. Nat Resource Res 28, 505–522 (2019).https://doi.org/10.1007/s11053-018-9405-4

Barzegar, R., Asghari Moghaddam, A., Adamowski, J. et al. multi-step water quality forecasting using a boosting ensemble multi-wavelet extreme learning machine model. Stoch Environ

Res Risk Assess 32, 799–813 (2018). https://doi.org/10.1007/s00477-017-1394-z

Chen, Y., & Wang, X. (2020). A review of machine learning algorithms for water quality prediction. Water, 12(7), 2022. https://doi.org/10.3390/w12072022

Chen, Y., & Wang, X. (2018). Machine learning algorithms for water quality prediction: A comparative study. Environmental Monitoring and Assessment, 190(9), 420 https://doi.org/10.1787/19934351

Charlotte Van Ooijen, Barbara Ubaldi 2019, Enabling the strategic use of data for productive, inclusive, and trustworthy governance. https://doi.org/10.1007/s10661-018-6563-z

Chaudhary Random Forest Algorithm - How It Works and Why It Is So Effective. https://www.turing.com/kb/random-forest-algorithm

Durgasrilakshmi Hari, Y Rudraksh Goud and Dodla Meghana, Spatial Analysis and Mapping of Groundwater Quality in Uppal Kalan, Hyderabad,2020 https://iopscience.iop.org/article/10.1088/1757-899X/1070/1/012038/meta

Di Nunno, F., Zhu, S., Ptak, M., Sojka, M., & Granata, F. (2023). A stacked machine learning model for multi-step ahead prediction of lake surface water temperature. *Science of the Total Environment*, *890*, 164323. https://doi.org/10.1016/j.scitotenv.2023.164323

Di Nunno, F., Zhu, S., Ptak, M., Sojka, M., & Granata, F. (2023). A stacked machine learning model for multi-step ahead prediction of lake surface water temperature. *Science of the Total Environment*, *890*, 164323. https://doi.org/10.1016/j.scitotenv.2023.164323

Duvva, L.K., Panga, K.K., Dhakate, R. et al. Health risk assessment of nitrate and Fluoride toxicity in groundwater contamination in the semi-arid area of Medchal, South India. Appl Water Sci 12, 11 (2022) https://doi.org/10.1007/s13201-021-01557-4

Dritsas, E. (n.d.). Efficient Data-Driven Machine Learning Models for Water Quality Prediction. MDPI. https://www.mdpi.com/2079-3197/11/2/16

Deepika, B.V., Ramakrishnaiah, C.R. & Naganna, S.R. Spatial variability of ground water quality: a case study of Udupi district, Karnataka State, India. *J Earth Syst Sci* **129**, 221 (2020).

https://doi.org/10.1007/s12040-020-01471-4

Emmanuelle CordatJoseph R. Casey Emmanuella Cordate, Joseph R. Casey Bicarbonate transport in cell physiology and disease, (2008). https://doi.org/10.1042/BJ20081634

Elbeltagi, A., Pande, C.B., Kouadri, S. *et al.* Applications of various data-driven models for the prediction of groundwater quality index in the Akot basin,Maharashtra, India. *Environ Sci Pollute Res* **29**, 17591–17605 (2022). https://doi.org/10.1007/s11356-021-17064-7Fei Tang,

Hemant Ishwaran Hemant Ishwaran Random Forest missing data algorithms, (2017) https://doi.org/10.1002/sam.11348

Geoffrey.S.Simate, Sunny E.Lyuke, Sehliselo Ndlovu, Human health effects of residual carbon nanotubes and traditional water treatment chemicals indrinking water, (2011). https://doi.org/10.1016/j.envint.2011.09.006

George A. Eby, Karen L. Eby, Rapid recovery from major depression using magnesium treatment, (2006)https://doi.org/10.1016/j.mehy.2006.01.047

Gudur, U., & Kumar, V. (2017). An analysis of water quality data using machine learning algorithms. Journal of Water and Climate Change, 8(1), 87-100. https://doi.org/10.2166/wcc.2017.005

Ghosal, S., & Bhattacharya, P. (2015). Machine learning approaches for water quality prediction. Journal of Hydrology, 523, 452-462. https://doi.org/10.1016/j.jhydrol.2015.03.071

Gilles Louppe “Understanding Random Forests: From Theory to Practice,” 2014 https://doi.org/10.48550/arXiv.1407.7502

Hongren Gong, Iren Sun, Baoshan Huang “Gradient Boosted Models for Enhancing Fatigue Cracking Prediction in Mechanistic-Empirical Pavement Design Guide” 2019 https://doi.org/10.1061/JPEODX.0000121

Hao, J., & Ho, T. K. (2019). Machine learning made easy: A review of scikit-learn package in python programming language. Journal of Educational and Behavioral Statistics, 44(3), 348–361. https://doi.org/10.3102/1076998619832248

Indrani Mukherjee, Umesh Kumar Singh, Sankar Chakma,

Evaluation of groundwater quality for irrigation water supply using

multi-criteria decision-making techniques and GIS in an agroeconomic tract of Lower Ganga basin, India, (2022) https://doi.org/10.1016/j.jenvman.2022.114691

Ivan Jordanov, Nedyalko Petrov. Alessio Petrozziello “Classifiers Accuracy Improvement Based on Missing Data Imputation”, (2018)https://doi.org/10.1515/jaiscr-2018-0002

Ji Feng, Yi-Xuan Xu, Yuan Jiang, Zhi-Hua Zhou “Soft Gradient Boosting Machine”2020https://doi.org/10.48550/arXiv.2006.04059

Jiabao Yan, Shaofeng Jia, Aifeng Lv,“Water Resources Assessment of China's Transboundary River Basins Using a Machine Learning Approach”2018 https://doi.org/10.1029/2018WR023044

Khan, S. I., Islam, N., Uddin, J., Islam, S., & Nasir, M. K. (2021, June 3). Water quality prediction and classification based on principal component regression and gradient boosting classifier approach, Journal of King Saud University - Computer and Information Sciences. https://doi.org/10.1016/j.jksuci.2021.06.003

Liu, Y., Wang, Y., & Hu, S. (2017). Boosting algorithms for water quality prediction: A comparative study. Water Research, 114, 177-185.https://doi.org/10.1016/j.watres.2017.03.064

Lionel Blanchet, Raffaele Vitale, Robert van,Vorstenbosch, George Stavropoulos, John Pender,Daisy Jonkers, Frederik-Jan van Schooten, Agnieszka Smolinska, “Constructing bi-plots for random forest” (2021).https://doi.org/10.1016/j.aca.2020.06.043

Luv Aggarwal, Manshubh Singh Rihal, Swati Aggarwal “EEG Based Participant Independent Emotion Classification using Gradient Boosting Machines” (2018). https://ieeexplore.ieee.org/abstract/document/8692106

Malek, N. H. A., Wan Yaacob, W. F., Md Nasir, S. A., & Shaadan, N. (2022). Prediction of Water Quality Classification of the Kelantan River Basin, Malaysia, Using Machine Learning Techniques. *Water*, *14*(7), 1067. https://doi.org/10.3390/w14071067

Martin Siegert, Richard B.Alley 2020, Twenty-first century sea-level rise could exceed IPCC projections for strong-warming futures https://doi.org/10.1016/j.oneear.2020.11.002

Mengyuan Zhu, Jiawei Wang 2022, A review of the application of machine learning in water quality evaluation,https://doi.org/10.1016/j.eehl.2022.06.001

Mohamed Yehia Z. Abouleish, Evaluation of fluoride levels in bottled water and their contribution to health and teeth problems in the United Arab Emirates, (2016) https://doi.org/10.1016/j.sdentj.2016.08.002

M. Qadir, E. Quillérou, V. Nangia, G. Murtaza, M. Singh, R.J. Thomas, P. Drechsel, A.D. Noble, Economics of salt-induced land degradation and restoration, (2014) https://doi.org/10.1111/1477-8947.12054

Maxime Gavage, Philippe Delahaut, Nathalie Gillard Suitability of High-Resolution Mass Spectrometry for Routine Analysis of Small Molecules in Food, Feed and Water for Safety and Authenticity Purposes: A Review, (2021)https://doi.org/10.3390/foods10030601

M. Vadiati, A. Asghari,Moghaddam, M. Nakhaei, J. Adamowski, A.H. Akbarzadeh, A fuzzy logic based decision-making approach for identification of groundwater quality based on groundwater quality indices (2016)https://doi.org/10.1016/j.jenvman.2016.09.082

Mogaraju, J. K. (2023, January 15). Application of machine learning algorithms in the investigation of groundwater quality parameters over YSR district, India. https://dergipark.org.tr/en/pub/tuje/issue/69252/1032314

Morin-Crini, N., Lichtfouse, E., Liu, G. et al. Worldwide cases of water pollution by emerging contaminants: a review. Environ Chem Lett 20, 2311–2338 (2022). https://doi.org/10.1007/s10311-022-01447-4

N. Subba Rao, A. Dinakar 2020, Seasonal and Spatial Variation of Groundwater Quality Vulnerable Zones of Yellareddygudem Watershed, Nalgonda District, Telangana State, India h

N.Subba Rao, A Dinakar, B Karnua Kumari, D Karunanidhi, T Kamakesh, 2022, Seasonal and Spatial Variation of Groundwater Quality Vulnerable Zones of Yellareddygudem Watershed, Nalgonda District, Telangana State Indiahttps://pubmed.ncbi.nlm.nih.gov/33236187/

Natekin, A., & Knoll, A. (2013b). Gradient boosting machines, a tutorial. Frontiers in Neurorobotics, 7. https://doi.org/10.3389/fnbot.2013.00021

Nikita Saklani, Ashli Khurana 2020, Global Warming:

Effect on Living Organisms,Causes and its Solutions.https://ssrn.com/abstract=3517122

Narsimha Adimalla, Application of the Entropy Weighted Water Quality Index (EWQI) and the Pollution Index of Groundwater (PIG) to Assess Groundwater Quality for Drinking Purposes: A Case Study in a Rural Area of Telangana State, India,(2021)https://doi.org/10.1007/s00244-020-00800-4

Nida Nasir, Afreen Kansal, Omar Alshaltone, Feras Barneih, Mustafa Sameer, Abdallah Shanableh, Ahmed Al-Shamma'a, Water quality classification using machine learning algorithms, Journal of Water Process Engineering, Volume 48, 2022, 102920, ISSN 2214-7144, https://doi.org/10.1016/j.jwpe.2022.102920

Pauline F.D. Scheelbeek, Aneire E. Khan,Sontosh Mojumder,Paul Elliott, Paolo Vineis, Drinking Water Sodium and Elevated Blood Pressure of Healthy Pregnant Women in Salinity-Affected Coastal Areas, (2006)https://doi.org/10.1161/HYPERTENSIONAHA.116.07743

Phillipp Probst, Marvin N. Wright, Anne-Laure Boulesteix, “Hyperparameters and tuning strategies for random forest,” 2019 https://doi.org/10.1002/widm.1301

Rana Ammar Aslam a, Sangam Shresth 2018, Groundwater vulnerability to climate change: A review of the assessment methodologyhttps://doi.org/10.1016/j.scitotenv.2017.08.237

Rakesh Kumar Mahajan, T.P.S Walia, B.S. Lark, Analysis of physical and chemical parameters of bottled drinking water, (2017)https://doi.org/10.1080/09603120500538184

Ruibin Zhang, Xin Qian, Huiming Li, Xingcheng Yuan, Rui Ye, Selection of optimal river water quality improvement programs using QUAL2K: A case study of Taihu Lake Basin, China, (2012) https://doi.org/10.1016/j.scitotenv.2012.05.063.

Risa K Kawaguchi, Masamichi Takahashi, Mototaka Miyake “Assessing Versatile Machine Learning Models for Glioma Radiogenomic Studies across Hospitals” 2021https://doi.org/10.3390/cancers13143611

S. Kayalvizhi, K. Ferents Koni Jiavana 2023, Prediction of ground water quality in western regions of Tamil Nadu using deep encoderhttps://doi.org/10.1016/j.uclim.2023.101458

Sardar Khan, Maria Shahnaz, Noor Jehan, Drinking water quality and human health risk in Charsadda district, Pakistan, (2016)https://doi.org/10.1016/j.jclepro.2012.02.016

Sharma, S., Bhattacharya, A. Drinking water contamination and treatment techniques. Appl Water Sci 7, 1043–1067 (2017). https://doi.org/10.1007/s13201-016-0455-7

Sao, V. (n.d.). Assessing drinking water quality and health risks of contaminants in the coastal areas of Cambodia | Journal of Water and Health | IWA Publishing. https://iwaponline.com/jwh/article/doi/10.2166/wh.2023.215/93119/Assessing-drinking-water-quality-and-health-risks

Travassos, X. L. (n.d.). 2020 *Artificial Neural Networks and machine learning techniques applied to ground penetrating radar: A Review*. Applied Computing Informatics. https://www.emerald.com/insight/content/doi/10.1016/j.aci.2018.10.001/full/html

Timo Vihma, James Screen 2015, The atmospheric role in the Arctic water cycle: A review on processes, past and future changes, and their impacts,https://doi.org/10.1002/2015JG003132

Vasudevan, P., & Anbalagan, S. (2018). Boosting algorithms for water quality prediction: A comparative study. Journal of Cleaner Production, 185, 861-874. https://doi.org/10.1016/j.jclepro.2018.05.100

Wu, Q., & Li, Y. (2018). Boosting algorithms for water quality prediction: A comparative study. Environmental Science and Pollution Research, 25(26), 26406-26417. https://doi.org/10.1007/s11356-018-2469-y

Yubin Park, Joyce C. Ho “An Overfitting-robust TreeBoost with Out-of-Bag Sample Regularization Techniques,” 2018https://doi.org/10.48550/arXiv.1807.08383

Zhao, Y., & Li, Y. (2019). A comparison of boosting algorithms for water quality prediction. Water, 11(11), 2272. https://doi.org/10.3390/w11112272

Zhang, Tao. (n.d.). *Improving convection trigger functions in deep convective ...* Improving Convection Trigger Functions in Deep Convective Parameterization Schemes Using Machine Learning. https://agupubs.onlinelibrary.wiley.com/doi/full/10.1029/2020MS002365