WATER QUALITY ANALYSIS IN INDIA USING MACHINE LEARNING TECHNIQUES

Minor project report submitted
in partial fulfillment of the requirement for award of the degree of

Bachelor of Technology
in
Computer Science & Engineering

By

M.S.S.S.KOWSHIK	(20UECS0588)	**(VTU15969)**
P.KISHORE	(20UECS0748)	**(VTU15970)**
K.SAI HARSHA	(20UECS0465)	**(VTU17935)**

Under the guidance of
Dr.Angeline Lydia,M.Tech.,Ph.D.,
ASSOCIATE PROFESSOR

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING
SCHOOL OF COMPUTING

VEL TECH RANGARAJAN DR. SAGUNTHALA R&D INSTITUTE OF SCIENCE & TECHNOLOGY

(Deemed to be University Estd u/s 3 of UGC Act, 1956)
Accredited by NAAC with A++ Grade
CHENNAI 600 062, TAMILNADU, INDIA

April, 2023

WATER QUALITY ANALYSIS IN INDIA USING MACHINE LEARNING TECHNIQUES

Minor project report submitted
in partial fulfillment of the requirement for award of the degree of

Bachelor of Technology
in
Computer Science & Engineering

By

M.S.S.S.KOWSHIK	(20UECS0588)	**(VTU15969)**
P.KISHORE	(20UECS0748)	**(VTU15970)**
K.SAI HARSHA	(20UECS0465)	**(VTU17935)**

Under the guidance of
Dr.Angeline Lydia,M.Tech.,Ph.D.,
ASSOCIATE PROFESSOR

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING
SCHOOL OF COMPUTING

VEL TECH RANGARAJAN DR. SAGUNTHALA R&D INSTITUTE OF SCIENCE & TECHNOLOGY

(Deemed to be University Estd u/s 3 of UGC Act, 1956)
Accredited by NAAC with A++ Grade
CHENNAI 600 062, TAMILNADU, INDIA

April, 2023

CERTIFICATE

It is certified that the work contained in the project report titled ”WATER QUALITY ANALYSIS IN INDIA USING MACHINE LEARNING TECHNIQUES” by ”M.S.S.S.KOWSHIK (20UECS0588), P.KISHORE (20UECS0748), K.SAI HARSHA(20UECS0465)” has been carried out under my supervision and that this work has not been submitted elsewhere for a degree.

Signature of Supervisor
Dr.Angeline Lydia
Associate Professor
Computer Science & Engineering
School of Computing
Vel Tech Rangarajan Dr. Sagunthala R&D
Institute of Science & Technology
April, 2023

Signature of Head of the Department
Computer Science & Engineering
School of Computing
Vel Tech Rangarajan Dr. Sagunthala R&D
Institute of Science & Technology
April, 2023

Signature of the Dean
Dr. V. Srinivasa Rao
Professor & Dean
Computer Science & Engineering
School of Computing
Vel Tech Rangarajan Dr. Sagunthala R&D
Institute of Science & Technology
April, 2023

DECLARATION

We declare that this written submission represents our ideas in our own words and where others' ideas or words have been included, we have adequately cited and referenced the original sources. We also declare that we have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in our submission. We understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

M.S.S.S.KOWSHIK

Date: / /

P.KISHORE

Date: / /

K.SAI HARSHA

Date: / /

APPROVAL SHEET

This project report entitled WATER QUALITY ANALYSIS IN INDIA USING MACHINE LEARNING TECHNIQUES by (M.S.S.S.KOWSHIK (20UECS0588), P.KISHORE (20UECS0748), K.SAI HARSHA(20UECS0465)) is approved for the degree of B.Tech in Computer Science & Engineering.

Examiners **Supervisor**

Dr.Angeline Lydia, M.Tech.,Ph.D.,

Date: / /

Place:

ACKNOWLEDGEMENT

We express our deepest gratitude to our respected **Founder Chancellor and President Col. Prof. Dr. R. RANGARAJAN B.E. (EEE), B.E. (MECH), M.S (AUTO),D.Sc., Foundress President Dr. R. SAGUNTHALA RANGARAJAN M.B.B.S.** Chairperson Managing Trustee and Vice President.

We are very much grateful to our beloved **Vice Chancellor Prof. S. SALIVAHANAN,** for providing us with an environment to complete our project successfully.

We record indebtedness to our **Professor & Dean, Department of Computer Science & Engineering, School of Computing, Dr. V. SRINIVASA RAO, M.Tech., Ph.D.,** for immense care and encouragement towards us throughout the course of this project.

We are thankful to our **Head, Department of Computer Science & Engineering,Dr.M.S. MURALI DHAR, M.E., Ph.D.,** for providing immense support in all our endeavors.

We also take this opportunity to express a deep sense of gratitude to our Internal Supervisor **Dr.ANGELINE LYDIA,M.Tech.,Ph.D.,** for her cordial support, valuable information and guidance, she helped us in completing this project through various stages.

A special thanks to our **Project Coordinators Mr. V. ASHOK KUMAR, M.Tech., Ms. C. SHYAMALA KUMARI, M.E.,** for their valuable guidance and support throughout the course of the project.

We thank our department faculty, supporting staff and friends for their help and guidance to complete this project.

M.S.S.S.KOWSHIK (20UECS0588)
P.KISHORE (20UECS0748)
K.SAI HARSHA (20UECS0465)

ABSTRACT

This study evaluated the relationship between water pH and the physicochemical properties of water while controlling for the influence of heavy metals and bacteriological factors using a nested logistic regression model. The study further sought to assess how these relationships are compared across confined water systems (ground water) and open water systems (surface water). Samples were collected from 100 groundwater and 132 surface water locations in the Tarkwa mining area. For the zero-order relationship in groundwater, EC, TDS, TSS, Ca, SO42-, total alkalinity, Zn, Mn, Cu, faecal and total coliform were more likely to predict optimal water pH. For surface water however, only TSS, turbidity, total alkalinity and Ca were significant predictors of optimal pH levels. At the multivariate level for groundwater, TDS, turbidity, total alkalinity and TSS were more likely to predict optimal water pH while EC, Mg, Mn and Zn were associated with non-optimal water pH. In general, random forest is a powerful machine learning algorithm that can provide high accuracy and robustness in classification tasks, such as water quality analysis. However, it is also computationally intensive and may require significant resources to train and run the model.

Keywords: Ground Water;pH levels;Total Dissolved Salts;Turbidity;Total Alkaline;TSS

LIST OF FIGURES

LIST OF ACRONYMS AND ABBREVIATIONS

BIS	Bureau of Indian Standards
CA	Calcium
CU	Copper
EC	Electrical Conductivity
LISAT	Long Island Systems,Applications And Technology Conference
MN	Manganese
MG	Magnesium
ML	Machine Learning
NWIS	National Water Information System
pH	Potential of Hydrogen
RF	Random Forest
TDS	Total Dissolved Solids
TSS	Total Suspended Solids
USGS	United States Geological Survey
WQT	Water Quality Class
WQI	Water Quality Index
ZN	Zinc

TABLE OF CONTENTS

Chapter 1

INTRODUCTION

1.1 Introduction

Water is and will continue to be an important part of life. water bodies such as lakes, rivers and streams are the most essential reservoirs for freshwater. Groundwater remains an essential source of potable water, serving as the primary water resource in arid regions. Compromising the quality of ground and surface water endangers the health and safety of residents within its catchment areas. Assessing the quality of water is mainly based on its physicochemical components, biological quality and heavy metals concentrations.

pH is probably by far the most important physicochemical parameter controlling the behavior of other water quality parameters as well as metals concentration in the aquatic environments. Chemical processes in aquatic systems such as acid-base reactions, solubility reactions, oxidation-reduction reactions and complexations are all influenced by hydrogen ions concentration (pH).

1.2 Aim of the project

The aim of this project is to develop a machine learning model using the random forest algorithm to analyze water quality data and predict whether water samples in India are potable or not.

Specifically, the objectives of the project are:

1. To collect and preprocess water quality data from various sources in India.
2. To perform exploratory data analysis to gain insights into the data and identify patterns and trends.
3. To develop a machine learning model using the random forest algorithm to predict the potability of water samples based on various water quality parameters.
4. To evaluate the performance of the model using appropriate metrics such as accuracy, precision, recall, and F1 score.

5. To use the model to predict the potability of new water samples and provide insights and recommendations for improving water quality management in India.

Overall, the project aims to leverage machine learning techniques to improve the efficiency and effectiveness of water quality analysis and management in India.

1.3 Project Domain

The domain of this project is environmental engineering and water quality analysis, specifically in India. The project involves using machine learning techniques, specifically the random forest algorithm, to analyze water quality data and predict whether water samples are potable or not.

This project could be useful for various stakeholders involved in the management of water resources in India, such as government agencies responsible for water quality monitoring, public health officials, and water treatment plant operators. It could also be relevant for researchers and academics in the fields of environmental engineering and water resource management.

1.4 Scope of the Project

The scope of the project is to find the relationship between water pH and the physicochemical properties of water while controlling for the influence of heavy metals and bacteriological factors using machine learning techniques.

The project is focused on using machine learning techniques to improve the accuracy and efficiency of water quality analysis and management in India, and aims to provide insights and recommendations based on the results of the analysis. The project is not intended to replace or substitute for existing water quality management practices, but rather to complement and enhance them.

Chapter 2

LITERATURE REVIEW

[1]Mourade Azrour et al.,: "Machine learning algorithms for efficient water quality prediction,".Model Earth System Environment.

In this study, we take the advantages of machine learning algorithms to develop a model that is capable of predicting the water quality index and then the water quality class. The method we propose is based on four water parameters: temperature, pH, turbidity and coliforms.

Water is an essential resource for human existence. In fact, more than 60% of the human body is made up of water. Our bodies consume water in every cell, in the different organisms and in the tissues. Hence, water allows stabilization of the body temperature and guarantees the normal functioning of the other bodily activities. Nevertheless, in recent years, water pollution has become a serious problem affecting water quality.

Therefore, to design a model that predicts water quality is nowadays very important to control water pollution, as well as to alert users in case of poor quality detection. Motivated by these reasons, in this study, we take the advantages of machine learning algorithms to develop a model that is capable of predicting the water quality index and then the water quality class.

The method we propose is based on four water parameters: temperature, pH, turbidity and coliforms. The use of the multiple regression algorithms has proven to be important and effective in predicting the water quality index. In addition, the adoption of the artificial neural network provides the most highly efficient way to classify the water quality.

Water quality predicting is a key and primary task in the context of the environmental control strategy. Certainly, the accuracy of predictions will surely contribute signifi-cantly to more appropriate conservation of water resources. Until today, numerous water management organizations have set up monitoring stations to check the evolution of the water quality situation

[2]Umair Ahamed et al.,: “Efficient Water Quality Prediction Using Supervised Machine Learning,”.

This research explores a series of supervised machine learning algorithms to estimate the water quality index (WQI), which is a singular index to describe the general quality of water, and the water quality class (WQC), which is a distinctive class defined on the basis of the WQI.The proposed methodology employs four input parameters, namely, temperature, turbidity, pH and total dissolved solids.

Water makes up about 70% of the earth’s surface and is one of the most important sources vital to sustaining life. Rapid urbanization and industrialization have led to a deterioration of water quality at an alarming rate, resulting in harrowing diseases. Water quality has been conventionally estimated through expensive and time-consuming lab and statistical analyses, which render the contemporary notion of real-time monitoring moot. The alarming consequences of poor water quality necessitate an alternative method, which is quicker and inexpensive.

With this motivation, this research explores a series of supervised machine learning algorithms to estimate the water quality index (WQI), which is a singular index to describe the general quality of water, and the water quality class (WQC), which is a distinctive class defined on the basis of the WQI. The proposed methodology employs four input parameters, namely, temperature, turbidity, pH and total dissolved solids.

Of all the employed algorithms, gradient boosting, with a learning rate of 0.1 and polynomial regression, with a degree of 2, predict the WQI most efficiently, having a mean absolute error (MAE) of 1.9642 and 2.7273, respectively. Whereas multi-layer perceptron (MLP), with a configuration of (3, 7), classifies the WQC most efficiently, with an accuracy of 0.8507. The proposed methodology achieves reasonable accuracy using a minimal number of parameters to validate the possibility of its use in real time water quality detection systems.

This research explores the methodologies that have been employed to help solve problems related to water quality. Typically, conventional lab analysis and statistical analysis are used in research to aid in determining water quality, while some analyses employ machine learning methodologies to assist in finding an optimized solution for the water quality problem.Local research employing lab analysis helped us gain a greater insight into the water quality problem in Pakistan.

[3]Khan et al.,: ”Predicting and analyzing water quality using Machine Learning: A comprehensive model,” 2016 IEEE Long Island Systems, Applications and Technology Conference (LISAT).

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).

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.

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.

[4]Zhou et al.,: Water Quality Prediction Method Based on Multi-Source Transfer Learning for Water Environmental IoT System.

In this study, water quality prediction is a research hotspot in the field of ecological environment, which is of great significance to the prevention of water pollution and the construction of automatic water quality monitoring network.

Water environmental Internet of Things (IoT) system, which is composed of multiple monitoring points equipped with various water quality IoT devices, provides the possibility for accurate water quality prediction. In the same water area, water flows and exchanges between multiple monitoring points, resulting in an adjacency effect in the water quality information. However, traditional water quality prediction methods only use the water quality information of one monitoring point, ignoring the information of nearby monitoring points.

In this paper, we propose a water quality prediction method based on multi-source transfer learning for a water environmental IoT system, in order to effectively use the water quality information of nearby monitoring points to improve the prediction accuracy. First, a water quality prediction framework based on multi-source transfer learning is constructed. Specifically, the common features in water quality samples of multiple nearby monitoring points and target monitoring points are extracted and then aligned.

According to the aligned features of water quality samples, the water quality prediction models based on an echo state network at multiple nearby monitoring points are established with distributed computing, and then the prediction results of distributed water quality prediction models are integrated. Second, the prediction parameters of multi-source transfer learning are optimized. Specifically, the back propagates population deviation based on multiple iterations, reducing the feature alignment bias and the model alignment bias to improve the prediction accuracy.

Finally, the proposed method is applied in the actual water quality dataset of Hong Kong. The experimental results demonstrate that the proposed method can make full use of the water quality information of multiple nearby monitoring points to train several water quality prediction models and reduce the prediction bias.Water environmental IoT system, which can collect water quality information inreal time, provides the possibility for accurate water quality prediction. In this paper,we propose a water quality prediction method based on MSTL for water environmentalIoT system, to effectively use the water quality information of nearby monitoring points,and then improve the prediction accuracy of water quality.

[5]Armah et al.,: Relationship of Trace Metal Covariates and pH Distribution in Groundwater within Gold mining and Non-Gold mining Areas in Ghana.

This research uses One of the most important defining characteristics of groundwater quality is pH as it fundamentally controls the amount and chemical form of many organic and inorganic solutes in groundwater. Groundwater data are frequently characterized by a wide degree of variability of the factors which possibly influence pH distribution. For this reason, it is challenging to link the spatio-temporal dynamics of pH to a single environmental factor by the ordinary least squares regression technique of the conditional mean.

In this study, quantile regression was used to estimate the response of pH to nine environmental factors (As, Cd, Fe, Mn, Pb, turbidity, electrical conductivity, total dissolved solids and nitrates). Results of 25%, 50%, 75% quantile regression and ordinary least squares (OLS) regression were compared. The standard regression of the conditional means (OLS) underestimated the rates of change of pH due to the selected factors in comparison with the regression quantiles.

The effect of arsenic increased for sampling locations with higher pH values (higher quantiles) likewise the influence of Pb and Mn. However, the effects of Cd and Fe decreased for sampling locations in higher quantiles. It can be concluded that these detected heterogeneities would be missed if this study had focused exclusively on the conditional means of the pH values.

Consequently, quantile regression provides a more comprehensive account of possible spatio-temporal relationships between environmental covariates in groundwater. This study is one of the first to apply this technique on groundwater systems in sub-Saharan Africa. The approach is useful and interesting and has broad application for other mining environments especially tropical low-income countries where climatic conditions can drive rapid cycling or transformations of pollutants.

It is also pertinent to geopolitical contexts where regulatory; monitoring and management capacities are weak and where mining pollution of groundwater largely occur.The standard regression of the conditional means (OLS) underestimated the rates of change of pH due to the selected factors in comparison with the regression quantiles. The effect of arsenic increased for sampling locations with higher pH values (higher quantiles) likewise the influence of Pb and Mn.

Chapter 3

PROJECT DESCRIPTION

3.1 Existing System

The existing system for water quality analysis in India includes manual sampling and laboratory testing of water samples. The Bureau of Indian Standards (BIS) sets the standards for drinking water quality in India, and water samples are tested for various parameters such as pH, total dissolved solids (TDS), turbidity, and presence of harmful contaminants such as arsenic, fluoride, and lead.The use of machine learning techniques, such as the random forest algorithm, can improve the efficiency and accuracy of water quality analysis by automating the process and leveraging a wider range of data sources.

3.1.1 Disadvantages

However, the existing system has several limitations. Manual sampling and laboratory testing can be time-consuming and expensive, and the results may not be available in real-time, which can delay decision-making for water quality management. Additionally, the manual process can be prone to errors and biases, and may not capture all relevant data for analysis.

3.2 Proposed System

The proposed system can provide several benefits such as improved efficiency and accuracy in water quality analysis, real-time insights and predictions, and cost savings compared to manual sampling and laboratory testing.

3.2.1 Advantages

Additionally, the system can help to identify potential water quality issues early, enabling proactive management and intervention to protect public health and the

environment.

3.3 Feasibility Study

3.3.1 Economic Feasibility

The proposed system for water quality analysis in India using the random forest algorithm is economically feasible, as the cost of data collection and model evaluation is low, and the cost of data preprocessing, model training, and deployment can be minimized using various tools and techniques.

Additionally, the benefits of the system, including improved efficiency and accuracy in water quality analysis, real-time insights and predictions, and proactive management, can significantly outweigh the costs in the long run.

3.3.2 Technical Feasibility

The technical feasibility of the project refers to the analysis of whether the proposed system can be implemented using the available technical resources and infrastructure.

Overall, the technical feasibility of the project seems achievable, given the availability of the necessary hardware, software, and data resources. However, it is essential to ensure that the proposed system is designed and developed in a scalable and maintainable way to handle potential future requirements and modifications.

3.3.3 Social Feasibility

The social feasibility of the ”Water Quality Analysis in India using Random Forest Algorithm in Machine Learning Techniques” project refers to the project’s ability to address social issues and contribute to society’s well-being.

Overall, the ”Water Quality Analysis in India using Random Forest Algorithm in Machine Learning Techniques” project has significant social implications and can contribute to society’s well-being by addressing a crucial issue related to public health and environmental protection.

3.4 System Specification

3.4.1 Hardware Specification

- Processor: Intel i5 or higher
- RAM: 8 GB or higher
- Storage: 500 GB or higher
- Display: Minimum 1366 x 768 resolution

3.4.2 Software Specification

- Python programming language
- Jupyter Notebook or any other IDE for running Python scripts
- Required Python libraries such as pandas, numpy, matplotlib, seaborn, scikit-learn, etc.
- MySQL or any other database management system for storing the data

3.4.3 Standards and Policies

Anaconda Prompt

Anaconda prompt is a type of command line interface which explicitly deals with the ML(MachineLearning) modules.And navigator is available in all the Windows,Linux and MacOS.The anaconda prompt has many number of IDE's which make the coding easier. The UI can also be implemented in python.

Jupyter

It's like an open source web application that allows us to share and create the documents which contains the live code, equations, visualizations and narrative text. It can be used for data cleaning and transformation, numerical simulation, statistical modeling, data visualization, machine learning.

Chapter 4

METHODOLOGY

4.1 General Architecture

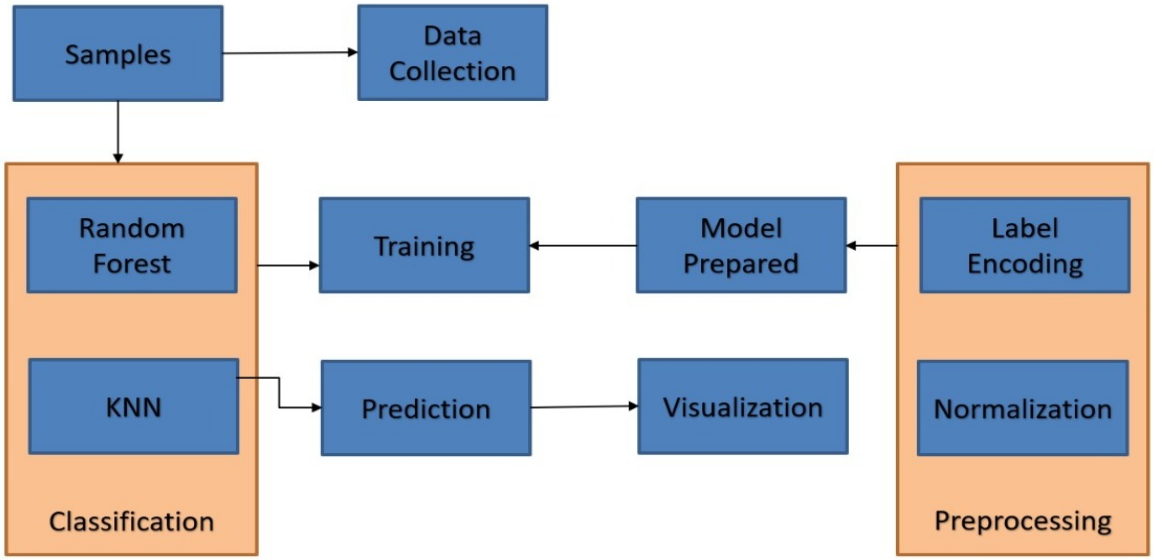

Figure 4.1: **Architecture Diagram of Water Quality Analysis**

The architecture diagram fig:4.1 for the water quality analysis system using machine learning techniques consists of several components that work together to provide accurate and efficient analysis of water quality.

At the center of the diagram is the data processing and machine learning component, which is responsible for receiving data from various sources, cleaning and pre-processing the data, and then training the machine learning models using the pre-processed data. The trained models are then used to predict water quality based on new data inputs.

Description of Architecture Diagram of Water Quality Analysis

4.2 Design Phase

4.2.1 Data Flow Diagram of Water Quality Analysis

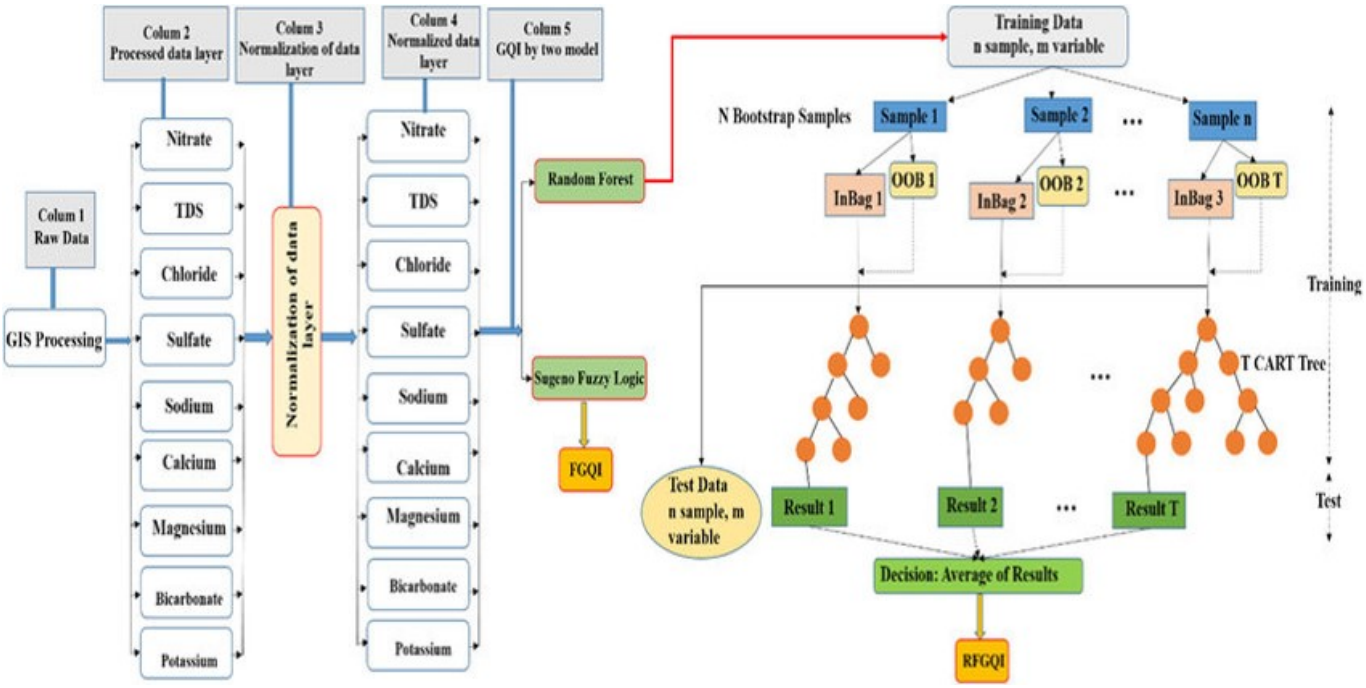

Figure 4.2: **Data Flow Diagram of Water Quality Analysis**

In this diagram fig:4.2, the raw water data is processed through a data cleaning module to remove any errors or inconsistencies in the data. The cleaned data is then passed through a feature extraction module to extract relevant features for water quality analysis. The extracted features are then used as inputs for the random forest classification model to predict the potability of water. Finally, the output of the system is the water potability prediction.

Description of Data Flow Diagram of Water Quality Analysis

4.2.2 Use Case Diagram of Water Quality Analysis

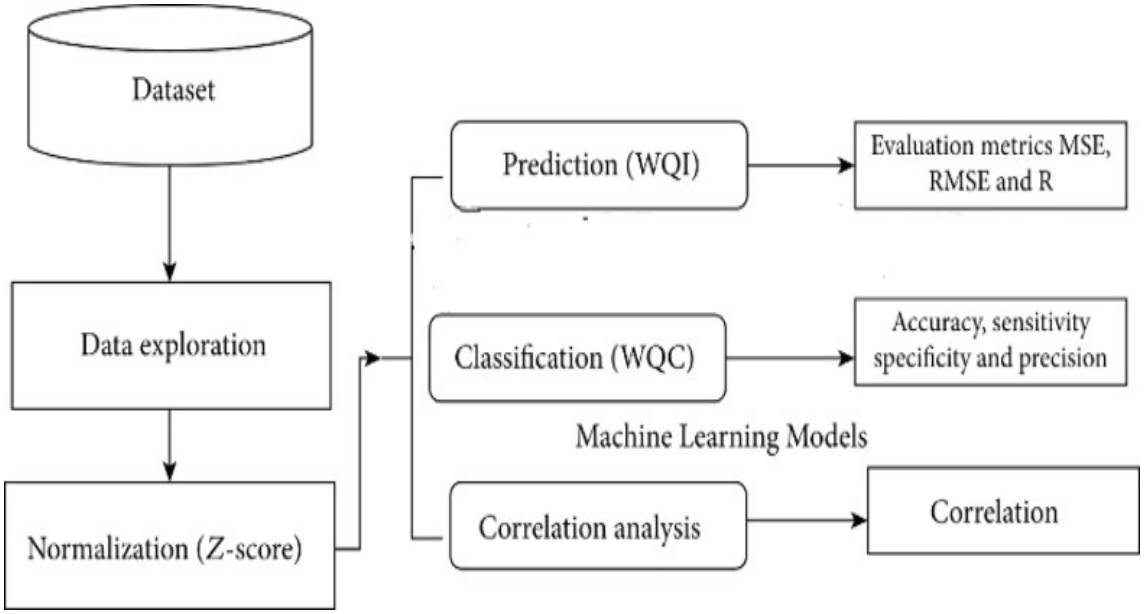

Figure 4.3: **Use case Diagram of Water Quality Analysis**

In this diagram fig:4.3, we discuss mainly about the following things

Data Collection and Pre-processing: This use case involves collecting water quality data from various sources, cleaning and processing the data, and storing it for further analysis.

Feature Selection and Engineering: This use case involves selecting relevant features from the pre-processed data, engineering new features to improve the performance of the machine learning model, and preparing the data for training and testing.

Random Forest Machine Learning Model: This use case involves training and testing the Random Forest algorithm on the pre-processed and engineered data, and producing a model that can be used for prediction.

Analysis and Visualization: This use case involves using the Random Forest machine learning model to predict water quality, analyzing the predicted data, and producing visualizations that can be used to gain insights into the factors affecting water quality.

Insights and Recommendations: This use case involves providing insights and recommendations based on the analysis and visualizations produced by the Analysis and Visualization module.

Description of Use case Diagram of Water Quality Analysis

4.2.3 Class Diagram of Water Quality Analysis

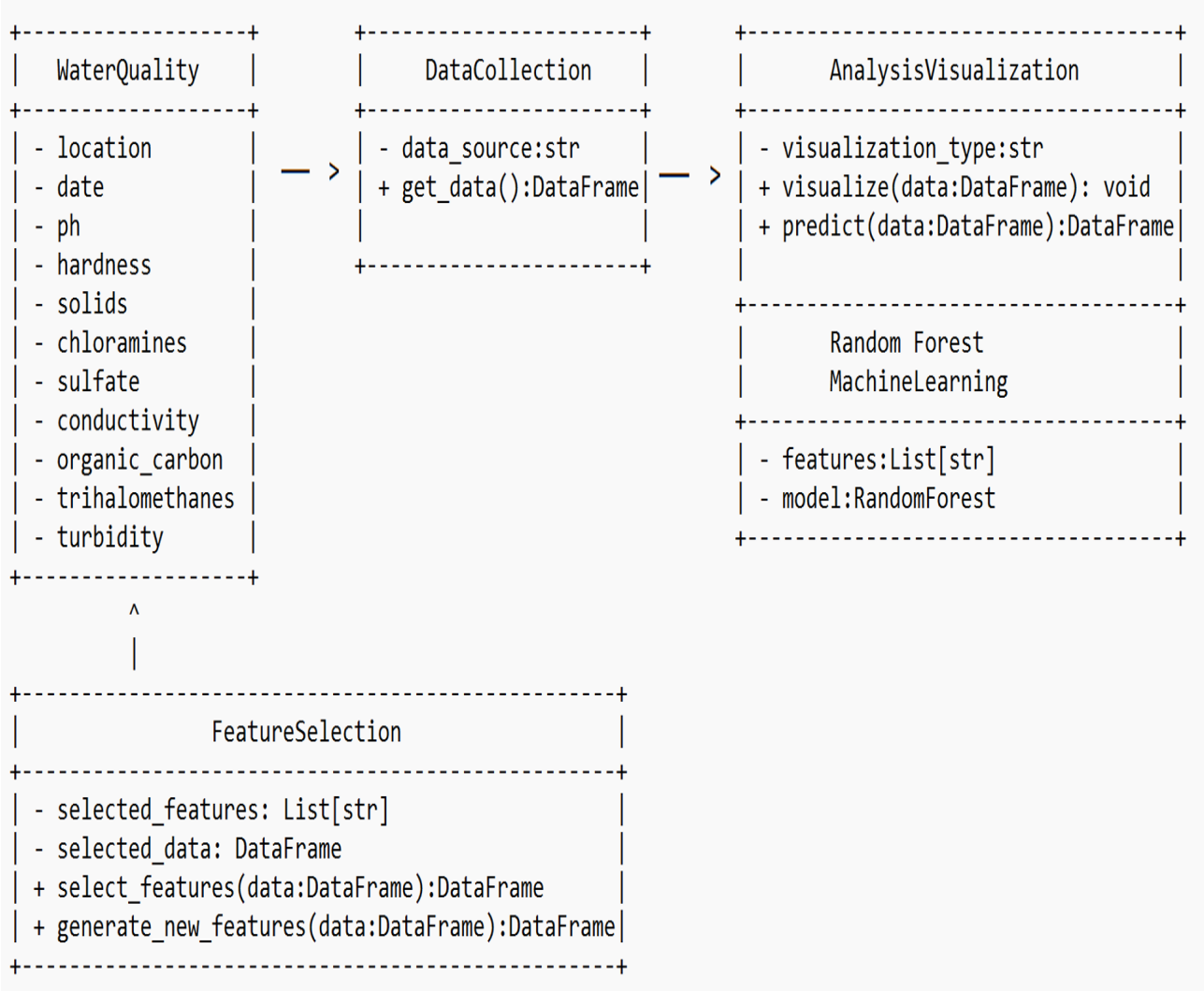

Figure 4.4: **Class Diagram of Water Quality Analysis**

In this class diagram fig:4.4, there are three main classes: Water Quality, Data Collection, and Analysis Visualization. Water Quality represents the water quality data, with attributes such as location, date, and various water quality parameters such as pH, hardness, and solids.

Data Collection is responsible for collecting and pre-processing the water quality data, with an attribute for the data source and a method for retrieving the data. Analysis Visualization is responsible for visualizing the data and producing predictions, with an attribute for the visualization type and methods for visualizing the data and producing predictions.

Description of Class Diagram of Water Quality Analysis

4.2.4 Sequence Diagram of Water Quality Analysis

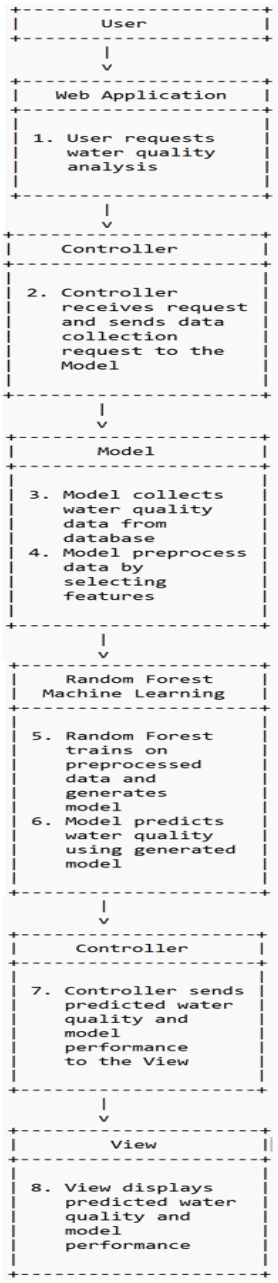

Figure 4.5: **Sequence Diagram of Water Quality Analysis**

In this sequence diagram fig:4.5, the user requests a water quality analysis through a web application. The controller receives the request and sends a data collection request to the model. The model collects water quality data from a database and preprocesses it by selecting relevant features.The preprocessed data is then used to train the Random Forest algorithm, which generates a model for predicting water quality. The model predicts water quality using the generated model.

Description of Sequence Diagram of Water Quality Analysis

4.2.5 Activity Diagram of Water Quality Analysis

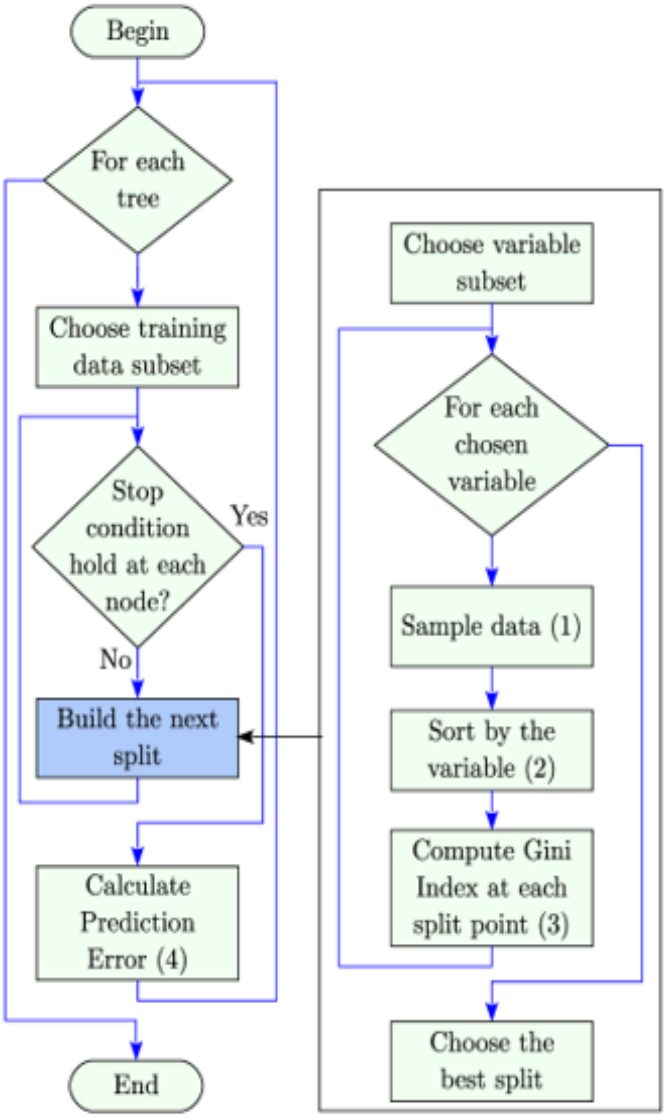

Figure 4.6: **Activity Diagram of Water Quality Analysis**

In this activity diagram fig:4.6, the process starts with collecting data from a data source. Then, the feature selection and engineering module selects relevant features from the data and generates new features to improve the performance of the machine learning model. Next, the data is split into training and testing sets, and the Random Forest algorithm is trained on the training set and tested on the testing set. The trained model is then used to predict water quality.

Finally, the results of the analysis are visualized, including the predicted water quality and the performance of the machine learning model. The process then ends.

Description of Activity Diagram of Water Quality Analysis

4.3 Algorithm & Pseudo Code

4.3.1 Algorithm

The random forest algorithm used in this project is a supervised machine learning algorithm used for classification tasks. It is an ensemble method that combines multiple decision trees to make predictions. Here are the steps involved in the random forest algorithm:

Step1: Choose the number of decision trees to include in the forest.

Step2: Select a random subset of the training data to use for each decision tree.

Step3: For each decision tree, randomly select a subset of the features to use for splitting.

Step4: Grow each decision tree to its maximum depth without pruning.

Step5: Use the decision trees to make predictions on new data by aggregating the predictions from each tree. The final prediction is the mode of the individual tree predictions for classification tasks.

Step6: Evaluate the performance of the model using metrics such as accuracy, precision, recall, and F1 score.

Step7: Tune the hyperparameters of the model, such as the number of decision trees, the maximum depth of the trees, and the number of features to use for splitting, to improve the model performance.

4.3.2 Pseudo Code

Here's the detailed pseudo code

1. Read the water quality dataset
    - dataset = read dataset('water quality.csv')
2. Perform data preprocessing
    - dataset = handle missing values(dataset)
    - dataset = feature scaling(dataset)
    - dataset = feature engineering(dataset)
3. Split the dataset into training and testing sets
    - X train, X test, y train, y test = split dataset(dataset)

4. Train the Random Forest model using the training set
    - rf model = RandomForestClassifier(n estimators=100, max depth=5, random state=42)
    - rf model.fit(X train, y train)
5. Use the trained model to make predictions on the testing set
    - y pred = rf model.predict(X test)
6. Evaluate the performance of the model using metrics such as accuracy, precision, recall, and F1-score
    - accuracy = calculate accuracy(y test, y pred)
    - precision = calculate precision(y test, y pred)
    - recall = calculate recall(y test, y pred)
    - f1 score = calculate f1 score(y test, y pred)
7. Tune the hyperparameters of the model to improve its performance
    - rf model tuned = tune hyperparameters(X train, y train)
8. Save the trained model for future use
    - save model(rf model tuned, 'rf model.pkl')
9. Use the saved model to make predictions on new, unseen data
    - new data = read new data('new water quality.csv')
    - rf model loaded = load model('rf model.pkl')
    - y pred new = rf model loaded.predict(new data)

4.4 Module Description

4.4.1 Module1:Data Collection and Preprocessing

Data Collection and Pre-processing Collecting water quality data from various sources such as government agencies, research papers, and online databases. Cleaning and pre-processing the data by removing duplicates, filling missing values, and removing outliers. Exploratory data analysis to understand the distribution and patterns in the data.

4.4.2 Module2:Feature Selection and Engineering

Feature Selection and Engineering Identifying relevant features for water quality analysis such as pH, temperature, dissolved oxygen, and various pollutants. Performing feature selection to eliminate irrelevant or redundant features. Feature engineering to create new features that might enhance the performance of machine learning models.

4.4.3 Module3:Machine Learning Models

Machine Learning Models Building and training machine learning models such as linear regression, decision trees, random forests, support vector machines, and neural networks. Evaluating the performance of each model using metrics such as mean squared error, R-squared, and accuracy. Selecting the best performing model for water quality analysis. Water Quality Analysis and Visualization Using the selected machine learning model to predict water quality for different locations in India. Visualizing the results using graphs, maps, and other visualizations to provide insights into water quality patterns in India. Developing recommendations and policies based on the analysis to improve water quality in India.

Chapter 5

IMPLEMENTATION AND TESTING

5.1 Input and Output

5.1.1 Input of Water Quality Analysis

The input to the water quality analysis system using random forest algorithm in machine learning techniques would be the water quality parameters such as pH, temperature, turbidity, total dissolved solids (TDS), electrical conductivity, and various chemical parameters like arsenic, fluoride, nitrate, sulfate, chloride, etc.

	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
...	...	...	...	...	...	...	...	...	...	...
3271	4.668102	193.681735	47580.991603	7.166639	359.948574	526.424171	13.894419	66.687695	4.435821	1
3272	7.808856	193.553212	17329.802160	8.061362	NaN	392.449580	19.903225	NaN	2.798243	1
3273	9.419510	175.762646	33155.578218	7.350233	NaN	432.044783	11.039070	69.845400	3.298875	1
3274	5.126763	230.603758	11983.869376	6.303357	NaN	402.883113	11.168946	77.488213	4.708658	1
3275	7.874671	195.102299	17404.177061	7.509306	NaN	327.459760	16.140368	78.698446	2.309149	1

Figure 5.1: **Data Set For Water Quality Analysis**

5.1.2 Output of Water Quality Analysis

The output of the system would be the prediction of the water potability, i.e., whether the water is safe for drinking or not. The output can be in the form of binary classification, where 1 indicates that the water is safe for drinking, and 0 indicates that the water is not safe for drinking. The system can also provide insights into the water quality parameters that are causing the water to be unsafe for drinking and suggestions for improving the water quality.

	features	feature_importances
0	ph	0.134941
4	Sulfate	0.123155
2	Solids	0.116097
3	Chloramines	0.113191
1	Hardness	0.111727
5	Conductivity	0.101580
6	Organic_carbon	0.100524
7	Trihalomethanes	0.100393
8	Turbidity	0.098392

Figure 5.2: **Output Of Analysis of Water**

5.2 Testing

5.2.1 Types of Testing

5.2.2 Unit testing

Input

Test the accuracy of each model by splitting the data into training and testing sets and comparing the predicted values to the actual values in the testing set. Test the models using different performance metrics such as mean squared error, R-squared, and accuracy. Test the sensitivity of the models to changes in the input features by introducing small variations in the feature values and observing the impact on the predicted values. Test the robustness of the models to outliers and missing values by introducing such values into the input data and observing the impact on the predicted values. Test the scalability of the models by increasing the size of the input data and observing the impact on the performance and computational time. Test the efficiency of the models by comparing the training and prediction times for different models. Test the generalization ability of the models by testing them on new, unseen data from different regions of India. Test the performance of the models on different data distributions to ensure that they are robust to changes in the data patterns.

5.2.3 Integration testing

Input

Test the data collection and pre-processing module by ensuring that the cleaned and pre-processed data is compatible with the feature selection and engineering module. Test the feature selection and engineering module by ensuring that the selected and engineered features are compatible with the machine learning models. Test the machine learning models by ensuring that they produce accurate predictions when fed with the features from the feature selection and engineering module. Test the water quality analysis and visualization module by ensuring that it produces accurate and insightful visualizations based on the predictions from the machine learning models. Test the integration of the different modules by ensuring that the data flows seamlessly between the modules and that the final recommendations and policies produced by the project are consistent and coherent.

5.2.4 System testing

Input

Test the data collection and pre-processing module by ensuring that it collects and cleans water quality data from various sources and formats it in a compatible way for further processing. Test the feature selection and engineering module by ensuring that it selects relevant features and engineers new features that enhance the performance of the machine learning models. Test the machine learning models by ensuring that they accurately predict water quality based on the input features and provide reliable insights into the impact of different factors on water quality. Test the water quality analysis and visualization module by ensuring that it provides insightful and accurate visualizations of the predicted water quality and the factors influencing it. Test the overall functionality of the project by ensuring that it meets the intended objectives of providing insights and recommendations for improving water quality in India.

5.2.5 Test Result

```
1 msno.bar(sett, figsize = (16,5),color = "#483D8B")
2 plt.show()
```

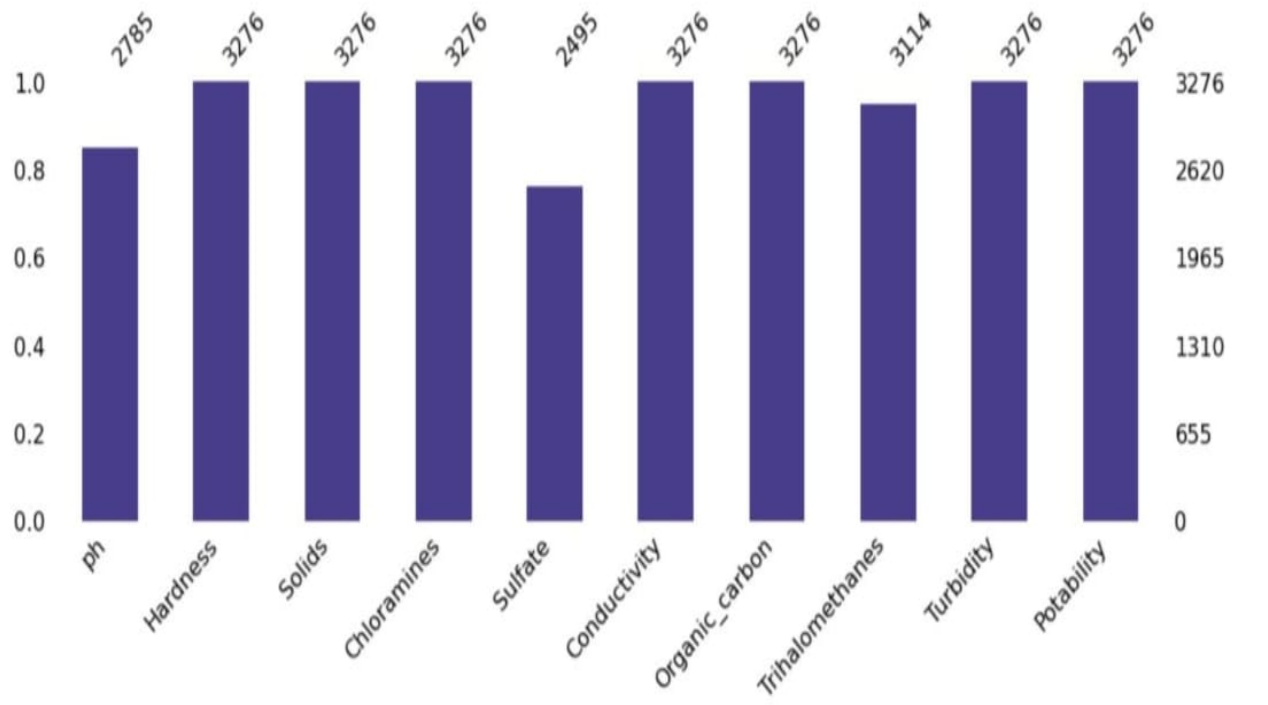

Figure 5.3: **Bar Graph of Water Quality Analysis**

Chapter 6

RESULTS AND DISCUSSIONS

6.1 Efficiency of the Proposed System

The efficiency of the proposed system for water quality analysis in India using random forest algorithm in machine learning techniques would depend on several factors, including the size and complexity of the dataset, the computational resources available for running the algorithm, and the accuracy and reliability of the model generated.

In general, random forest is a powerful machine learning algorithm that can provide high accuracy and robustness in classification tasks, such as water quality analysis. However, it is also computationally intensive and may require significant resources to train and run the model.

To optimize the efficiency of the proposed system, several strategies can be employed, such as:

Data preprocessing: Cleaning, transforming and reducing the dataset before training the algorithm can improve its efficiency by reducing the complexity of the model and the amount of required computational resources.

Hyperparameter tuning: Adjusting the parameters of the random forest algorithm, such as the number of trees, tree depth, and number of features, can improve the efficiency of the model by reducing overfitting and improving accuracy.

Distributed computing: Utilizing parallel processing or distributed computing techniques can enable the algorithm to run faster and more efficiently on multiple processors or computing nodes.

Feature selection: Choosing the most relevant features for the model can help to reduce the computational burden and improve the efficiency of the algorithm.

By employing these strategies, the proposed system for water quality analysis in India using random forest algorithm in machine learning techniques can achieve high accuracy and efficiency, making it a valuable tool for water quality monitoring and management.

6.2 Comparison of Existing and Proposed System

Existing system:(Laboratory Testing)

The existing system for water quality analysis in India involves manual collection of water samples and laboratory testing. This process is time-consuming and may not provide real-time results. On the other hand, the proposed system using machine learning techniques and the random forest algorithm aims to provide a faster and more accurate analysis of water quality. The proposed system can analyze a large amount of data in a short time, and can also provide real-time monitoring of water quality. gives less accurate output that is less when compared to proposed system.

Proposed system:(Random forest algorithm)

In terms of accuracy, the proposed system is expected to outperform the existing system as it utilizes advanced machine learning techniques that can analyze large datasets and identify patterns that may not be easily identifiable through traditional methods. Moreover, the proposed system can provide proactive alerts and notifications to the authorities in case of any water quality issues, enabling them to take timely action.

In terms of cost, the proposed system may require an initial investment in terms of hardware and software, but in the long run, it is expected to reduce costs associated with manual sampling and laboratory testing. Additionally, the proposed system can help in reducing the cost of healthcare by enabling early detection of water-borne diseases.

Overall, the proposed system is expected to provide a more efficient, accurate, and cost-effective solution for water quality analysis in India compared to the existing system.

Name Of The Algorithm	Precision	Recall	F1 Score	Support	Accuracy
Random Forest Algorithm	68% 71%	71% 68%	70% 69%	636 664	69%
K Neighbors	61% 61%	56% 66%	58% 63%	636 664	61%
SVC	64% 66%	66% 65%	65% 65%	636 664	65%
Logistic Regression	49% 51%	58% 42%	53% 46%	636 664	50%
XBG	66% 68%	66% 68%	66% 68%	636 664	67%

Figure 6.1: **Algorithm Classification**

6.3 Sample Code

```
import pandas as pd
import numpy as np
from sklearn.ensemble import RandomForestClassifier
from sklearn.model_selection import train_test_split
from sklearn.metrics import classification_report

# Load the water quality dataset
df = pd.read_csv('water_quality_data.csv')

# Preprocess the dataset
df.dropna(inplace=True) # drop rows with missing values
X = df.drop(['Potability'], axis=1) # feature matrix
y = df['Potability'] # target variable

# Split the dataset into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=42)

# Train the random forest classifier
clf = RandomForestClassifier(n_estimators=100, random_state=42)
clf.fit(X_train, y_train)

# Make predictions on the testing set
y_pred = clf.predict(X_test)

# Evaluate the performance of the model
print(classification_report(y_test, y_pred))
```

Output

Name Of The Algorithm	Precision	Recall	F1 Score	Support	Accuracy
Random Forest Algorithm	68% 71%	71% 68%	70% 69%	636 664	69%
K Neighbors	61% 61%	56% 66%	58% 63%	636 664	61%
SVC	64% 66%	66% 65%	65% 65%	636 664	65%
Logistic Regression	49% 51%	58% 42%	53% 46%	636 664	50%
XBG	66% 68%	66% 68%	66% 68%	636 664	67%

Figure 6.2: **Algorithm Classification**

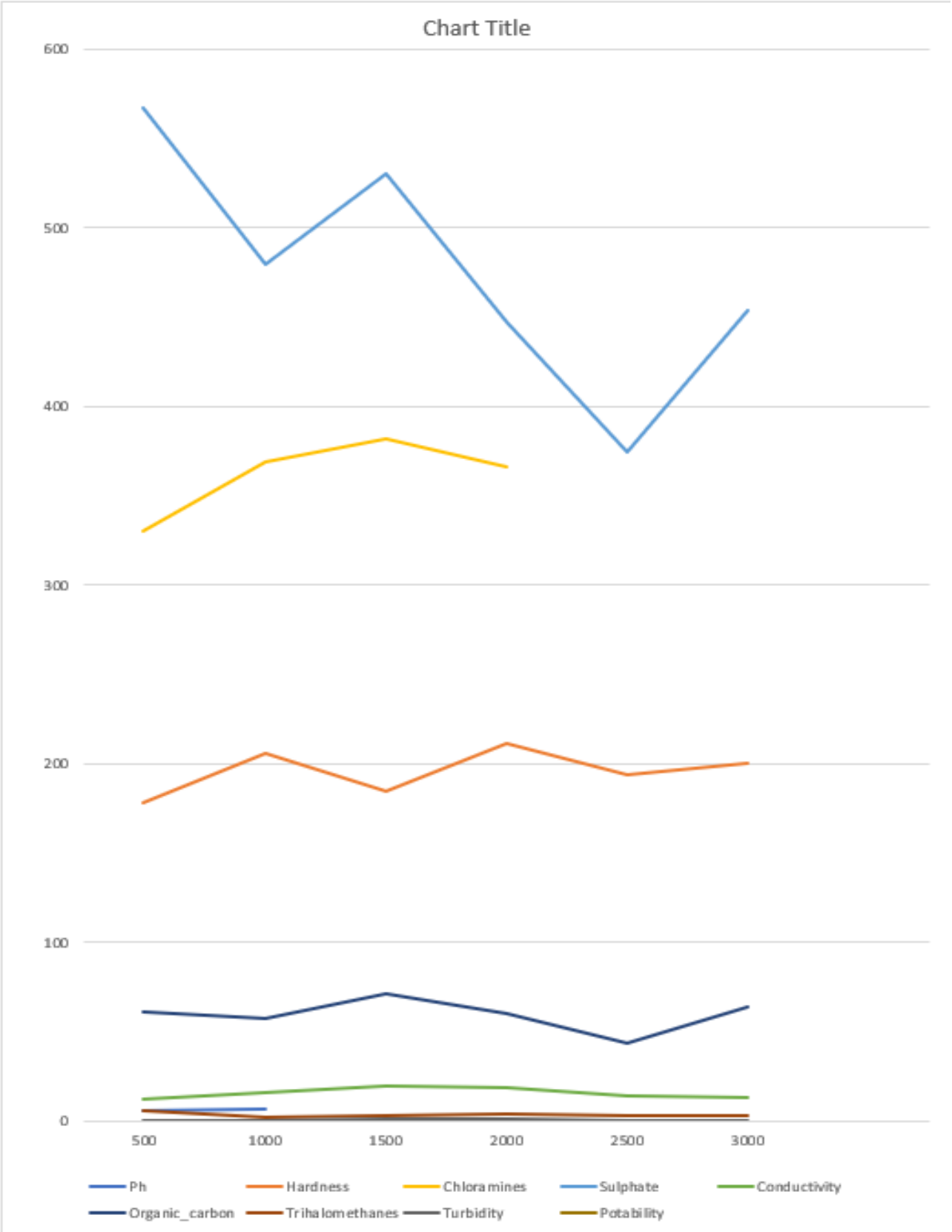

Figure 6.3: **Line Graph of Water Quality Analysis**

Chapter 7

CONCLUSION AND FUTURE ENHANCEMENTS

7.1 Conclusion

This project on water quality analysis in India using machine learning techniques aimed to provide insights into the water quality patterns in India and develop recommendations for improving water quality. The project consisted of four modules: Data collection and pre-processing, Feature selection and engineering, Machine learning models, and Water quality analysis and visualization.

In the first module, water quality data was collected from various sources, cleaned, and pre-processed. In the second module, relevant features were identified, and feature selection and engineering were performed.In the third module, several machine learning models were built and evaluated, and the best performing model was selected. In the fourth module, the selected model was used to predict water quality for different locations in India, and the results were visualized and analyzed.

The analysis revealed several insights into the water quality patterns in India, such as the impact of pollutants and other environmental factors on water quality. The recommendations and policies developed based on the analysis could help improve water quality in India and protect public health. Overall, this project demonstrates the potential of machine learning techniques for water quality analysis and highlights the importance of data-driven approaches in addressing environmental issues.

7.2 Future Enhancements

Some possible future enhancements for this project could include:

1. **Real-time data analysis:** Currently, the model is trained on a pre-existing dataset. However, in the future, it would be beneficial to collect real-time water quality data from various sources and analyze it in real-time to provide up-to-date information.

2. **Integration with IoT devices:** The proposed system could be integrated with IoT devices such as sensors that can collect water quality data from various sources in real-time. This will improve

the accuracy of the model's predictions.

3. **Expansion to other regions:** Currently, the model is trained on water quality data from India. However, it can be expanded to include data from other regions to provide a more comprehensive analysis of water quality.

4. **Integration with a mobile application:** A mobile application can be developed that allows users to access water quality information on-the-go. The application can provide real-time water quality data, alerts, and recommendations for water treatment.

5. **Integration with government agencies:** The proposed system can be integrated with government agencies responsible for monitoring water quality. This will help in providing real-time alerts and improve the efficiency of the water quality monitoring process.

Chapter 8

PLAGIARISM REPORT

ATTACH ONLY SUMMARY PAGE OF PLAGIARISM REPORT

Chapter 9

SOURCE CODE & POSTER PRESENTATION

9.1 Source Code

```
import pandas as pd
import numpy as np
import matplotlib as mpl
import matplotlib.pyplot as plt
%matplotlib inline
import seaborn as sns
import missingno as msno
from scipy import stats
sett = pd.read_csv('/content/water_potability.csv')
sett
sett.info()
sett.describe().style.background_gradient(cmap = "Blues")
sett.duplicated().sum()
msno.bar(sett, figsize = (16,5),color = "#483D8B")
plt.show()
for column_name in sett.columns:
    unique_values = len(sett[column_name].unique())
    print("Feature '{column_name}' has '{unique_values}' unique values".format(column_name =
        column_name,
                                                                                unique_values
                                                                                    =
                                                                                unique_values
                                                                                ))
sett.columns
numeric_features=['ph', 'Hardness', 'Solids', 'Chloramines', 'Sulfate', 'Conductivity',
       'Organic_carbon', 'Trihalomethanes', 'Turbidity']
import warnings
warnings.filterwarnings('ignore')
fig,ax = plt.subplots(len(numeric_features),4,figsize=(30,20))
for index,i in enumerate(numeric_features):
    sns.distplot(sett[i],ax=ax[index,0],color='green')
    sns.boxplot(sett[i],ax=ax[index,1],color='yellow')
    sns.violinplot(sett[i],ax=ax[index,2],color='purple')
    stats.probplot(sett[i],plot=ax[index,3])

```

```
fig.tight_layout()
fig.subplots_adjust(top=0.95)
plt.suptitle("Visualizing continuous columns (sett dataset)",fontsize=30)
fig, ax1 = plt.subplots(figsize=(20,10))
graph = sns.countplot(ax=ax1,x = 'Potability' , data = sett,palette='pastel')
graph.set_xticklabels(graph.get_xticklabels(),rotation=90)
for p in graph.patches:
    height = p.get_height()
    graph.text(p.get_x()+p.get_width()/2., height + 0.1,height ,ha="center")
sett.isnull().sum()

from sklearn.impute import KNNImputer
imputer = KNNImputer(n_neighbors=10, weights="uniform")
l=imputer.fit_transform(sett)
sett1=pd.DataFrame(l,columns=sett.columns)
sett1
sett1.isnull().sum()
# Shapiro-Wilk Test
from scipy.stats import shapiro
# normality test
for feature in numeric_features:
  stat, p = shapiro(sett1[feature])
  print('Statistics=%.3f, p=%.3f' % (stat, p))
  # interpret
  alpha = 0.05
  if p > alpha:
    print(f'Sample looks Gaussian with {feature} (fail to reject H0)')
  else:
    print(f'Sample does not look Gaussian with {feature} (reject H0)')

from scipy.stats import mannwhitneyu
for feature in numeric_features:
    stat, p = mannwhitneyu(sett1['Potability'], sett1[feature])
    print('Statistics=%.3f, p=%.3f' % (stat, p))
    alpha = 0.05
    if p > alpha:
        print('There are no significant differences (fail to reject H0)')
    else:
        print('The sample distributions are not equal (reject H0)')
correlation = sett1.corr()
print(correlation['Potability'].sort_values(ascending = False),'\n')
k= 18
cols = correlation.nlargest(k,'Potability')['Potability'].index
print(cols)
cm = np.corrcoef(sett1[cols].values.T)
mask = np.triu(np.ones_like(sett1.corr()))
f , ax = plt.subplots(figsize = (14,12))
sns.heatmap(cm,mask=mask, vmax=.8, linewidths=0.01,square=True,annot=True,cmap='viridis',
            linecolor="white",xticklabels = cols.values ,annot_kws = {'size':12},yticklabels = cols.
                values)
```

```
from sklearn.neighbors import LocalOutlierFactor
clf = LocalOutlierFactor(n_neighbors=5, contamination='auto')
y_pred = clf.fit_predict(sett1)
sett1['Out']=y_pred
Out=sett[sett1['Out']!=1]
Out.shape
sett2=sett1[sett1['Out']==1]
sett2=sett2.drop('Out',axis=1)
from imblearn.over_sampling import SMOTE
oversample = SMOTE()
features, labels=  oversample.fit_resample(sett2.drop(["Potability"],axis=1),sett2["Potability"])
from sklearn.preprocessing import MinMaxScaler
scaler = MinMaxScaler()
names = features.columns
d = scaler.fit_transform(features)

scaled_df = pd.DataFrame(d, columns=names)
scaled_df.head()
from sklearn.ensemble import RandomForestClassifier
from sklearn.neighbors import KNeighborsClassifier
from sklearn import svm
from sklearn.linear_model import LogisticRegression
from sklearn.ensemble import GradientBoostingClassifier
from sklearn.svm import SVC
from sklearn.metrics import accuracy_score
from sklearn.model_selection import train_test_split, GridSearchCV
import xgboost as xgb

from sklearn.metrics import precision_score,recall_score
from sklearn.metrics import f1_score
from sklearn.metrics import confusion_matrix,accuracy_score, classification_report
X_sett, X_test, y_sett, y_test=train_test_split(scaled_df,labels,test_size=0.33,random_state=42)
models = [RandomForestClassifier(), KNeighborsClassifier(), SVC(), LogisticRegression(),xgb.
    XGBClassifier()]
scores = dict()

for m in models:
    m.fit(X_sett, y_sett)
    y_pred = m.predict(X_test)

    print(f'model: {str(m)}')
    print(classification_report(y_test,y_pred, zero_division=1))
    print('-'*30, '\n')
!pip install optuna
import optuna
from optuna import Trial, visualization
from optuna.samplers import TPESampler
from xgboost import XGBClassifier
from sklearn.model_selection import cross_val_score
def objective(trial: Trial,X,y) -> float:
```

```
130
131     param = {
132             'n_estimators' : trial.suggest_int("n_estimators",100,1500),
133             'max_features' : trial.suggest_categorical("max_features",["auto","sqrt"]),
134             'max_depth' : trial.suggest_int("max_depth",5,80,log = True),
135             'min_samples_split' : trial.suggest_int("min_samples_split",2,15),
136             'min_samples_leaf' : trial.suggest_int("min_samples_leaf",1,9),
137             'bootstrap' : trial.suggest_categorical("bootstrap",[True,False])
138             }
139
140     model = RandomForestClassifier(**param)
141
142     return cross_val_score(model, X, y, cv=5).mean()
143 study = optuna.create_study(direction='maximize',sampler=TPESampler())
144 study.optimize(lambda trial : objective(trial,X_sett,y_sett),n_trials= 20)
145 from optuna import visualization
146 print('Best trial: CV_score= {},\nparams {}'.format(study.best_trial.value,study.best_trial.params))
147 hist = study.trials_dataframe()
148 hist.head()
149 optuna.visualization.plot_slice(study)
150 optuna.visualization.plot_optimization_history(study)
151 optuna.visualization.plot_parallel_coordinate(study)
152 optuna.visualization.plot_param_importances(study)
153 Best_trial = study.best_trial.params
154 best_clf2=RandomForestClassifier(**Best_trial)
155 best_clf2.fit(X_sett, y_sett)
156 from sklearn.metrics import confusion_matrix, ConfusionMatrixDisplay
157 cm = confusion_matrix(best_clf2.predict(X_test),y_test)
158 disp = ConfusionMatrixDisplay(cm, display_labels=["0","1"])
159 disp.plot()
160 plt.title("Confusion Matrix")
161 plt.show()
162 feature_importances=best_clf2.feature_importances_
163 feature_importances_df=pd.DataFrame({'features':list(X_sett), 'feature_importances':
        feature_importances})
164 feature_importances_df=pd.DataFrame({'features':list(X_sett), 'feature_importances':
        feature_importances})
165 feature_importances_df.sort_values('feature_importances',ascending=False)
```

9.2 Poster Presentation

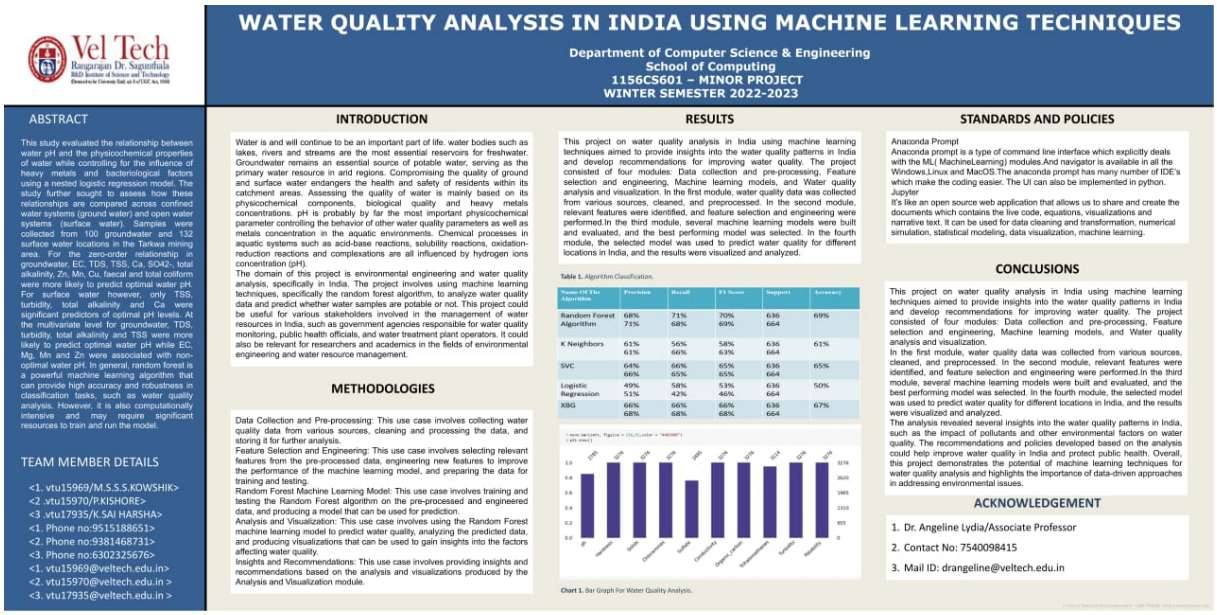

Figure 9.1: **Poster Presentation**

References

[1] M. Azrour, J. Mabrouki, G. Fattah, A. Guezzaz, and F. Aziz, "Machine learning algorithms for efficient water quality prediction," Modeling Earth Systems and Environment, vol. 8, no. 2, pp. 2793–2801, Aug. 2021.

[2] Ahmed, Umair Mumtaz, Rafia Anwar, Hirra Shah, Asad Irfan, Rabia García-Nieto, José. (2019). Efficient Water Quality Prediction Using Supervised Machine Learning. Water. 11. 2210. 10.3390/w11112210.

[3] Khan, Yafra Chai, Soo See. (2016). Predicting and analyzing water quality using Machine Learning: A comprehensive model. 1-6. 10.1109/LISAT.2016.7494106.

[4] Zhou, Jian Wang, Jian Chen, Yang Li, Xin Xie, Yong. (2021). Water Quality Prediction Method Based on Multi-Source Transfer Learning for Water Environmental IoT System. Sensors. 21. 7271. 10.3390/s21217271.

[5] Armah, Frederick Paintsil, Arnold Adu, Michael Yawson, David Odoi, Justice. (2020). Relationship of Trace Metal Covariates and pH Distribution in Groundwater within Gold mining and Non-Gold mining Areas in Ghana. 10.20944/preprints202012.0321.v1.

[6] Mustapha A, Aris AZ, Yusoff FM, Zakaria MP, Ramli MF, Abdullah AM, et al. Statistical Approach in Determining the Spatial Changes of Surface Water Quality at the Upper Course of Kano River, Nigeria. Water Quality, Exposure and Health. 2014;6(3):127–42.

[7] Onojake MC, Ukerun SO, Iwuoha G. A Statistical Approach for Evaluation of the Effects of Industrial and Municipal Wastes on Warri Rivers, Niger Delta, Nigeria. Water Quality, Exposure and Health. 2011;3(2):91–9.

[8] Greiserman M, Hasson D, Semiat R, Shemer H. Kinetics of dolomite dissolution in a packed bed by acidified desalinated water. Desalination. 2016;396:39–47.

[9] Malecki KM, Schultz AA, Severtson DJ, Anderson HA, VanDerslice JA. Private-well stewardship among a general population based sample of private well-owners. Science of the Total Environment. 2017;601:1533–43. pmid:28605871

[10] Ewusi A, Ahenkorah I, Aikins D. Modelling of total dissolved solids in water supply systems using regression and supervised machine learning approaches. Applied Water Science. 2021;11(2). pmid:33462558

[11] Shroff P, Vashi RT, Champaneri VA, Patel KK. Correlation study among water quality parameters of groundwater of Valsad district of south Gujarat(India). Journal of Fundamental and Applied Sciences. 2015;7(3).

[12] Seidu J, Ewusi A. Assessment of Groundwater Quality and Health Risk of Heavy Metals: A study from the Tarkwa Mining Area, Ghana. Ghana Mining Journal. 2020;20(1):1–10.

[13] Baah-Ennumh TY, Adom-Asamoah G. Land Use Challenges in Mining Communities–The Case of Tarkwa-Nsuaem Municipality. Environment and Ecology Research. 2019;7(3):139–52.

[14] Aram SA, Saalidong BM, Osei Lartey P. Comparative assessment of the relationship between coliform bacteria and water geochemistry in surface and ground water systems. Plos one. 2021 Sep 21;16(9):e0257715. pmid:34547049

[15] Ndur SA, Nyarko SY, Quaicoe I, Osei LB. Heavy Metal Loading in Surface Sediments along the Kawere Stream, Tarkwa, Ghana. Ghana Mining Journal. 2020;20(2):77–85.