PREDICTIVE MODEL FOR ASSESSMENT OF CONSUMABLE WATER

Major Project Report

Submitted in partial fulfillment of the requirement of University of Mumbai

For the Degree of

(Computer Engineering)

By

1) Vaishnavi Nikam	**ID No: TU3F2021036**
2) Ruchika Naik	**ID No: TU3F2021162**
3) Manasi Chavan	**ID No: TU3F2021013**
4) Swati Utekar	**ID No: TU3F2021017**

Under the Guidance of

Dr.Rohini Patil

Department of Computer Engineering

TERNA ENGINEERING COLLEGE

Plot no.12, Sector-22, Opp. Nerul Railway station,
Phase-11, Nerul (w), Navi Mumbai 400706

UNIVERSITY OF MUMBAI

TERNA ENGINEERING COLLEGE, NERUL,

NAVI MUMBAI

Department of Computer Engineering

Academic Year 2023-2024

CERTIFICATE

This is to certify that the major project entitled “Predictive Model for Assessment of Consumable Water” is a bonafide work of

1) Vaishnavi Nikam	**ID No: TU3F2021036**
2) Ruchika Naik	**ID No: TU3F2021162**
3) Manasi Chavan	**ID No: TU3F2021013**
4) Swati Utekar	**ID No: TU3F2021017**

submitted to the University of Mumbai in partial fulfillment of the requirement for the award of the Bachelor of Engineering (Computer Engineering).

Guide **Head of Department** **Principal**

Project Report Approval

This Major Project Report – entitled “**Predictive Model for Assessment of Consumable Water**” by following students is approved for the degree of ***B.E. in "Computer Engineering"***.

Submitted by:

1) Vaishnavi Nikam	**ID No: TU3F2021036**
2) Ruchika Naik	**ID No: TU3F2021162**
3) Manasi Chavan	**ID No: TU3F2021013**
4) Swati Utekar	**ID No: TU3F2021017**

Examiners Name & Signature:

1.---

2.---

Date: --------------------------------

Place: --------------------------------

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.

Vaishnavi Nikam	**ID No:TU3F2021036**	--------------------------
Ruchika Naik	**ID No: TU3F2021162**	--------------------------
Manasi Chavan	**ID No: TU3F2021013**	--------------------------
Swati Utekar	**ID No: TU3F2021017**	--------------------------

Date: ______________________

Place: ______________________

Acknowledgement

We would like to express our sincere gratitude towards our guide **Dr.Rohini Patil**, Project Coordinators **Prof. Pramila Mate** for their help, guidance and encouragement, they provided during the project development. This work would have not been possible without their valuable time, patience and motivation. We thank them for making our stint thoroughly pleasant and enriching. It was great learning and an honor being their student.

We are deeply thankful to **Prof. Kishor Sakhure (H.O.D Computer Department)** and entire team in the Computer Department. They supported us with scientific guidance, advice and encouragement, they were always helpful and enthusiastic and this inspired us in our work.

We take the privilege to express our sincere thanks to **Dr. L. K. Ragha** our Principal for providing the encouragement and much support throughout our work.

Vaishnavi Nikam	**ID No:TU3F2021036**	--------------------------
Ruchika Naik	**ID No: TU3F2021162**	--------------------------
Manasi Chavan	**ID No: TU3F2021013**	--------------------------
Swati Utekar	**ID No: TU3F2021017**	--------------------------

Date: ______________________

Place: ______________________

Index

TABLE OF CONTENTS

Abstract

Water is one of the vital resources for sustaining life and supporting varied ecosystems. Although more than two-third of the earth's surface is covered by water, only a small portion is available for diverse activities by humans and other animals. The degradation of water quality due to pollutants, contaminants, and changes in natural conditions has far-reaching consequences, thus preservation of water quality has become crucial for public health and environmental sustainability. When it comes to preventing water pollution and building a network of automatic water quality monitoring stations, water quality prediction is a hotspot for research in the ecological environment discipline. In this paper, we have examined development and research trends in the field of water quality prediction by analyzing various prediction models that utilize machine learning techniques to estimate water quality. Finally, based on the study context our prediction model is proposed.

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Introduction:

Water quality is a critical aspect of environmental health, impacting both human well-being and ecosystem sustainability.Ensuring the availability of clean and drinkable water is essential for sustaining life on Earth. With the continuous evolution and expansion of human activities, the concern for maintaining the quality of our water resources has become increasingly critical. Monitoring and predicting water quality is crucial for ensuring safe drinking water, maintaining aquatic ecosystems, and managing industrial processes.While traditional methods of assessing water quality hold value, they often struggle to keep up with the swiftly changing environmental conditions and the emergence of new pollutants.In recent times, the incorporation of machine learning techniques into environmental monitoring has emerged as a powerful tool in this domain, offering the potential to enhance the accuracy and efficiency of water quality assessments.This innovative approach harnesses the capabilities of artificial intelligence to analyze extensive datasets, identify intricate patterns, and make remarkably accurate predictions regarding water quality parameters.

ML algorithms, when trained on large datasets of water quality parameters and corresponding environmental conditions, can learn complex relationships and patterns that may not be readily apparent to human observers. These models can then be used to make accurate predictions about future water quality based on input data.

In this context, this study aims to explore the application of ML techniques for water quality prediction. By leveraging advanced algorithms and data-driven approaches, we seek to improve our understanding of the factors influencing water quality and develop robust models for forecasting.

1.2 Scope of the project:

The proposed Machine Learning (ML) model for water quality prediction aims to accomplish several key tasks. Firstly, it will integrate diverse data sources encompassing parameters like pH levels, dissolved oxygen, turbidity, temperature, and nutrient concentrations, while also handling preprocessing tasks such as managing missing values, normalizing features, and removing outliers. Additionally, the model will engage in feature selection and engineering, identifying the most pertinent variables that contribute to accurate water quality predictions and potentially creating new features for enhanced predictive performance. Following this, the ML model will undergo training on historical water quality data, learning intricate relationships between environmental factors and water quality parameters through parameter adjustments to minimize prediction errors. Once trained, the model will be proficient in predicting specific water quality parameters, such as pollutant concentrations, pH levels, and dissolved oxygen content. Furthermore, it may be seamlessly integrated into a system for real-time water quality monitoring, furnishing continuous predictions and alerts to facilitate proactive responses to changes in water conditions.

1.3 Organization of The Report

Chapter 1 contains a brief introduction of our project with the aim and scope of the project.

Chapter 2 contains a brief idea of our Problem Statement and Objective.

Chapter 3 contains a Literature Survey of our project.In this chapter, we have studied and reviewed the previous work done on the topics related to our project. We have included different papers published by their respective authors.

Chapter 4 Software Standards this chapter contains details of the Proposed system, Software Requirements, Hardware Requirements.

Chapter 5 Design Contains the Use case Diagram and the sequence diagram of our project.

Chapter 6 Methodology deals with the Overview of the proposed system.

Chapter 7 is the conclusion of the project. We have also discussed the Future work of our project.

Lastly, it has a list
of references.

Chapter 2
Problem Statement

2.1 Problem statement

The degradation of water quality in various water bodies globally, driven by human activities, has led to an urgent requirement for accurate and dependable water quality prediction models. As a result of factors like industrial pollution, agricultural runoff, and urban development, natural water systems are constantly under pressure, emphasizing the crucial need for the ability to anticipate changes in water quality parameters. These parameters encompass a wide range of physical, chemical, and biological characteristics, including temperature, pH levels, turbidity, dissolved oxygen levels, and the presence of contaminants. This endeavor will involve the collection and integration of diverse datasets. Through rigorous data preprocessing, feature engineering, and model development, the project aims to provide accurate and interpretable models that can forecast water quality trends. The successful implementation of such models will not only enhance our understanding of water quality dynamics but also equip stakeholders and decision-makers with valuable tools for proactively managing and safeguarding our precious water resources, ensuring both environmental sustainability and public health.

2.2 Objectives

2.2.1 Health Protection: Keeping drinking water safe is crucial. Predicting issues like harmful substances helps authorities act to keep people healthy.

2.2.2 Sustainable Water: Predicting water quality helps manage lakes, rivers, etc. This helps balance human needs with nature's needs, ensuring long-term water use.

2.2.3 Nature Protection: Predicting water quality protects aquatic life. It prevents harm and maintains diverse ecosystems in water bodies.

2.2.4 Helping Industries: Industries like farming, fishing, and tourism need water. Prediction guides them to use water wisely, boosting the economy.

Chapter 3

Literature Review

Author(Year)	Title of Paper	Findings	Research Gap
Yong Ye et.al (2023) [1]	Predictive Simulation Study on the Effect of Small and Medium River Basin Outfall Treatment Measures on Water Quality Improvement	Hydrodynamic simulation on the total nitrogen (TN) concentration's movement was implemented and the time of the nitrogen concentration to reach the standard was predicted.	In this paper, the water environment model only considered a single water quality indicator of TN, and other pollution factors were not included in the model. Therefore, later studies can extend and improve the model with additional data to make the water environment model more comprehensive.
Dr. D. Brindha, et.al(2023) [2]	Water Quality Analysis and Prediction using Machine Learning	The goal of this study is to estimate water quality by acquiring several parameters such as temperature, conductivity, pH, dissolved oxygen (% sat), nitrates, and fecal and total coliforms and using the machine learning method Random Forest regression, Decision Tree. Dataset has 1992 samples.	The parameters can be reduced and can make it available for all users. More functionalities and stylings can be added to the web interface, so that the web application can be more interactive. the suggestions can be elaborated with what minerals can be added to the water.
Rui Tan et.al (2023) [3]	A data-driven model for water quality prediction in Tai Lake, China, using secondary modal	The proposed model based on WOA-CVMD-CBILSTM-AT-F derives a water temperature regulation mechanism to control the DO content and thus contribute to	The proposed water temperature regulation mechanism is not yet mature enough, and it is currently only a stepwise increase or

	decomposition with multidimensional external features	strengthening the protection of water resources and the management of fishery production. WOA-CVMD-CBiLSTM-AT-F -WOA-CVMD-CBiLSTM-AT-F model BP has 0.9954 highest accuracy for training set and 0.9947 accuracy for testing set	decrease in water temperature. If corresponding research theories in the field are proposed later, the regulatory effect of DO can be further improved.
Jinal Patel et.al.(2022) [4]	A Machine Learning-Based Water Potability Prediction Model by Using Synthetic Minority Oversampling Technique and Explainable AI	This work shows the comparative analysis of different machine learning approaches like Support Vector Machine (SVM), Decision Tree (DT), Random Forest, Gradient Boost, and Ada Boost, used for the water quality classification. The model is trained on the Water Quality Index dataset available on Kaggle. Gradient Boost has the accuracy of 0.76 Dataset available on kaggle but requires authority to use	In this paper multiple variables have been chosen from well known datasets that include variables such as pH, hardness, solids, turbidity. Only XGB and RF had the best performance. The performance of other algorithms like SVC, ADA, decision tree can be improved by carrying out more detailed study.
Mohamed Torky et.al (2023) [5]	Recognizing Safe Drinking Water and Predicting Water Quality Index using Machine Learning Framework	The framework uses nine classification models, including XGBoost, LightGB, Decision Tree, Extra Tree, MLP, Gradient Boosting, SVM, ANN, and Random Forest. Results show good classification with average accuracy of 94.7%. However, Light	More novel deep learning models can be developed for predicting water quality which is suitable for human drinking, irrigation of plants and other industrial or environmental purposes.

		GBM outperformed other classifiers in testing accuracy of 0.97%.	
Sanaa Kaddoura (2022) [6]	Evaluation of Machine Learning Algorithm on Drinking Water Quality for Better Sustainability	In this paper, a machine learning classifier model is built to predict the quality of water using a real dataset.These are some algorithm used: 1.Random Forest 2.Gradient Boosted Trees 3.Logistic Regression 4.Logistic regression algorithm details 5.XGBoost 6.LASSO LARS	The proposed approach will be modified to enhance the performance of these algorithms. Hyperparameter tuning can be performed on each algorithm to find the best model setup to obtain the most optimized result.
Liming Sheng et.al (2020) [7]	Water quality prediction method based on preferred classification	The water quality prediction method based on preferred classification combines the advantages of three different prediction models. These are some of the models used: 1.BPNN 2.SVR 3.LSTM	BPNN and LSTM require a large amount of data for training; the proposed method in this paper is limited by the amount of historical water quality data. The proposed method will be improved on this point in the future.
Nur Afyfah Suwadi et.al (2022) [8]	An Optimized Approach for Predicting Water Quality Features Based on Machine Learning	This study aimed to improve prediction efficiency by selecting the best feature subset from the fewest number of available features. As a result, the reduced number of features reduces the time required to predict the	The study's limitations include its specificity to a region, requiring consideration for extending results to other regions and water quality parameters. They are aiming to perform future research which will use

		class. SMOTE sampling methods were applied to the raw dataset to ensure the accuracy of the training data	augmented reality for better monitoring and interaction techniques.
Abirami K et.al (2023) [9]	Water Quality Analysis and Prediction using Machine Learning	They conducted a comparative analysis of various machine learning classification algorithms to predict Water Quality Classification (WQC) based on the Water Quality Index (WQI). This research utilizes a Kaggle dataset to conduct an analysis of water quality. It encompasses data of Water samples that are taken from two locations in Chennai. Among the tested models, the Random Forest Classifier algorithm demonstrated superior performance, achieving an impressive accuracy score of 91.97% in efficiently predicting WQC.	The system can be improved by designing it in such way that it can notify the concerned authorities using GSM module in case of deteriorating water quality. This could be useful particularly in the agriculture and other industry sectors where they can assess the quality of water before releasing it for utilization.
Rongli Gai, Jiahui Yang (2021) [10]	Summary of water quality prediction models based on machine learning	The water quality prediction method based on machine learning is mainly introduced, focusing on time series prediction method, regression analysis method, neural network method and combination prediction method.	The application of water quality prediction based on machine learning is still immature, the existing work results show that the prediction method of machine learning has been greatly improved compared with the

		According to the research history and present situation of water quality prediction model, the development trend of water quality prediction model is prospected.	traditional prediction method, but there is still a lot of room for improvement. It can be expected that the prediction of water quality by machine learning deep learning method will become a research hotspot in this field in the future.
S. Babu et.al (2023) [11]	Water Quality Prediction using Neural Networks	They used the Long Short Term Memory(LSTM) algorithm in Neural Network and the Decision tree and Naive Bayes classifiers for classification for Water Quality Index (WQI).The primary intention of the project is to design and create a method for measuring water quality utilising MLP and a feed-forward neural network. LSTM, ANN and ARIMA are the three methods applied	The proposed approach will further do classification of water quality by improving its accuracy. The study proposes a new LSTM NN model to predict the pH value and turbidity, which are measures of water quality. They aim to determine the adaptability of water for aquatic species' habitats and classify whether it is polluted, palatable, potable, or infected water.
Theyazn H. H Aldhyani et.al (2020) [12]	Water Quality Prediction Using Artificial Intelligence Algorithms	They introduced a methodology for predicting and classifying water quality using machine learning techniques. The study utilises various water metrics such as pH, dissolved oxygen, suspended solids,	The proposed approach will further apply the developed models to real-world scenarios and explore new approaches for analysing and predicting water quality.

		electrical conductivity, turbidity, chloride, chemical oxygen demand, total dissolved solids, and alkalinity. The dataset used for study consisted of 108 specimens. The collected samples were obtained from the Environment Department (DOE) and the Environment and Forest Ministry, Bangladesh	
Md. Saikat Islam Khan et.al (2021) [13]	Water quality prediction and classification based on principal component regression and gradient boosting classifier approach	The paper proposes a method for predicting water quality using machine learning algorithms. The authors used nine water quality parameters, including pH, DO, SS, EC, Turbidity, Chloride, COD, TDS, and Alkalinity, and applied the principal component regression (PCR) method. The results showed that the PCR model achieved a prediction accuracy of 95%.	Some limitations mentioned in this paper are for collecting more training samples. These will be overcome using proper tuning of the PCR model and using deep neural networks.
Jitha P Nair, Vijaya M S (2021) [14]	Predictive Models for River Water Quality using Machine Learning and Big Data Techniques - A	They applied various time series analysis, machine learning and deep learning techniques involved in identifying and evaluating the quality of the water.	The further research on water quality prediction can be carried out by implementing hybrid techniques using deep learning, GIS and remote sensing

	Survey	Autoregressive (AR) and Autoregressive Moving Average (ARMA) were used for the time series Analysis of WQP. Whereas, Linear Regression and Logistic Regression Artificial Neural Network were used for the Machine Learning algorithms.	algorithms on time series meteorological data, geographical data, and remote sensing images.
Shubham Palkar et.al(2021) [15]	WQ-Net: A Deep Neural Network Model For Water Quality Prediction	They developed WQ-Net, a Deep Neural Network model to estimate the Water Quality Index and Water Quality Class. It contains 1744 samples from different Indian states. The main goal of study was to develop a model that will predict WQI and WQC with fewer parameters.	This research lacks in a real-time monitoring system using. By applying it the model would immediately predict the water quality based on the monitoring system's real-time data. An android application can be developed to check the WQI value from any remote area.

Table 3.1 Literature Survey

Chapter 4

Software Requirements Specification

4.1.External Interface Requirements

4.1.1 User Interfaces
The user interface for system shall be compatible to any type of web browser such as Mozilla ,Firefox, Google Chrome, and Internet Explorer.

4.1.2 Software Interfaces
Google Colab - Python 3.9.2
Operating System - Windows(64-bit) macOS 10.12.6

4.1.3 Hardware Interfaces
Processor -2.10GHz or faster
RAM - Minimum 2GB
Memory - Minimum 200MB

4.2 Functional Requirements

4.2.1 Give input
User Should fill the input filled which are required for prediction of water quality.

4.2.2 Submit input
User should submit after filling all parameters.

4.3 Performance Requirements

4.3.1 Response Time
Acceptable response times for various system operations, such as data input, model training, and prediction generation.

4.3.2 Scalability
System should scale with increased data volume or user load.

4.3.3 Prediction Throughput
Predictions can be processed quickly, especially for real-time monitoring.

4.3.4 Security Performance
Performance requirements for security measures, such as encryption and access control.

4.3.5 Data Processing Speed
Data should be processed, including data cleaning, feature extraction, and preprocessing steps. Optimize data processing algorithms for efficiency.

4.4.Security requirements

4.4.1 Data Encryption
Ensure that all sensitive data, including historical water quality data and model parameters, are encrypted both in transit and at rest to prevent unauthorized access.

4.4.2 Access Control

Implement strict access controls to restrict access to the prediction system based on user roles and responsibilities. Only authorized personnel should be able to interact with and modify the system.

4.4.3 Authentication and Authorization Require strong authentication mechanisms, such as multi-factor authentication, to verify the identity of users. Authorization mechanisms should be in place to define what actions users are allowed to perform within the system.

4.4.4 Data Integrity

Implement measures to ensure the integrity of data, such as checksums and data validation, to detect and prevent data tampering or corruption.

Chapter 5

Design

5.1 Data Flow Diagram (DFD)

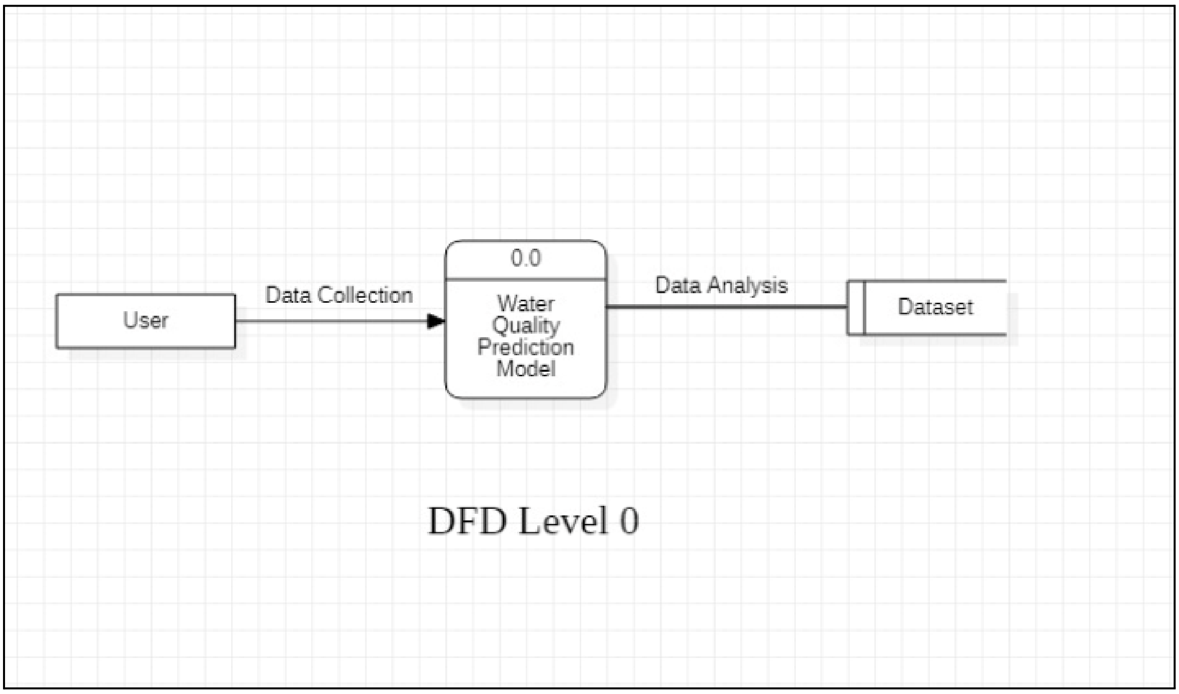


Fig 5.1.1 DFD Level-0

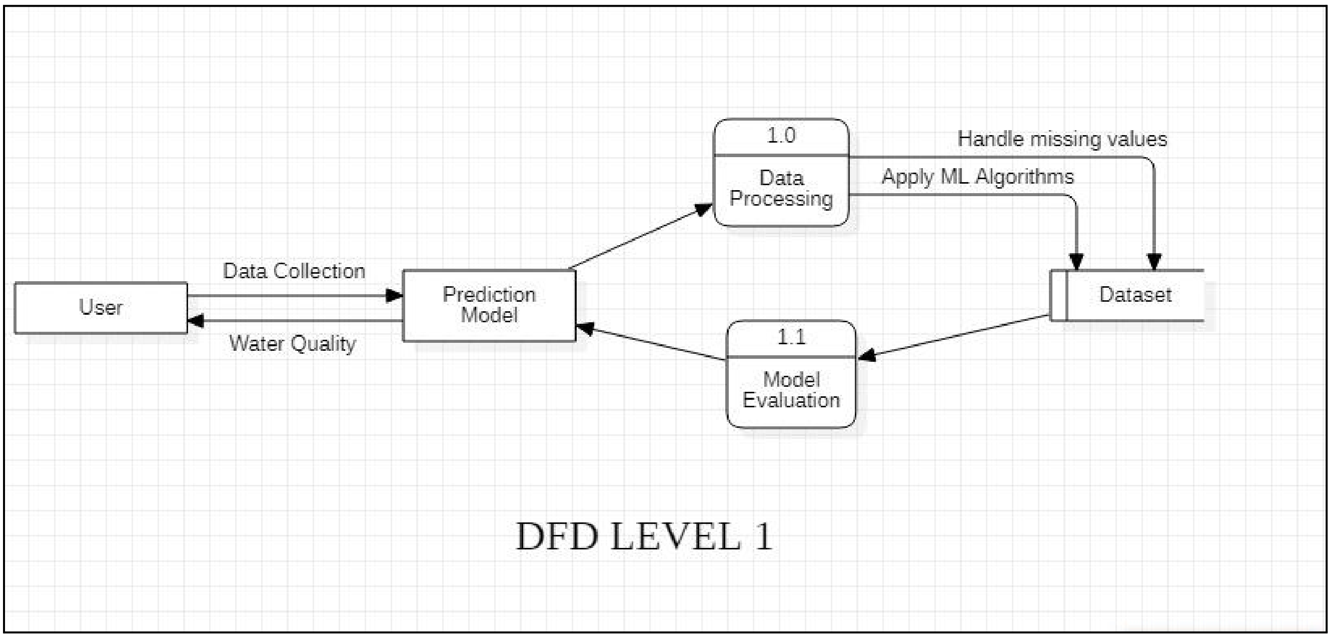


Fig 5.1.2 DFD Level-1

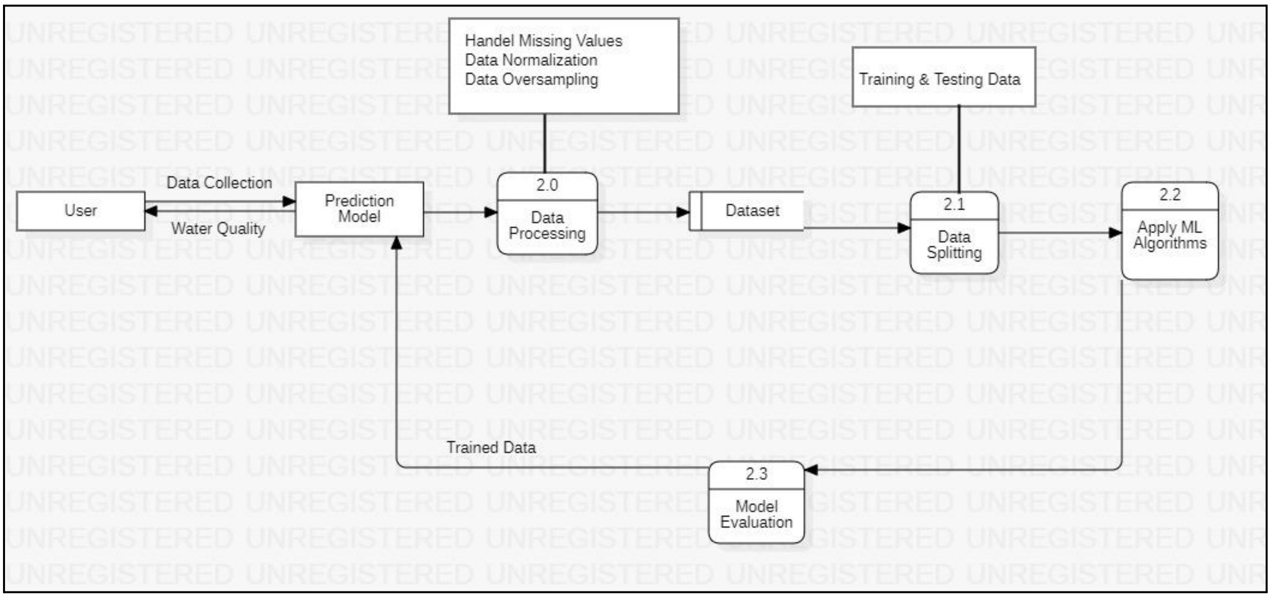


Fig 5.1.3 DFD Level-2

5.2 Use Case Diagram

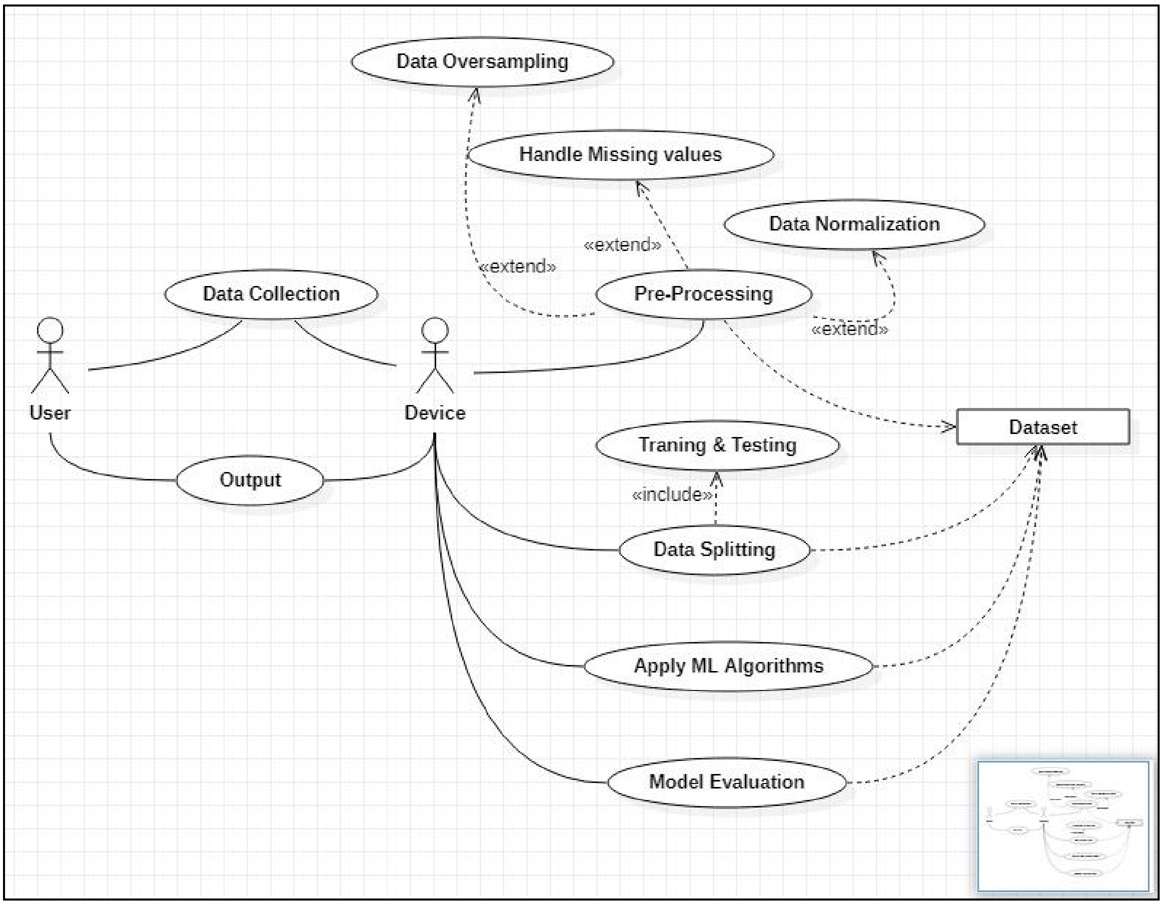


Fig 5.2 Use Case

The above figure shows the user and the actions performed by the user in the system.

5.3 Sequence Diagram

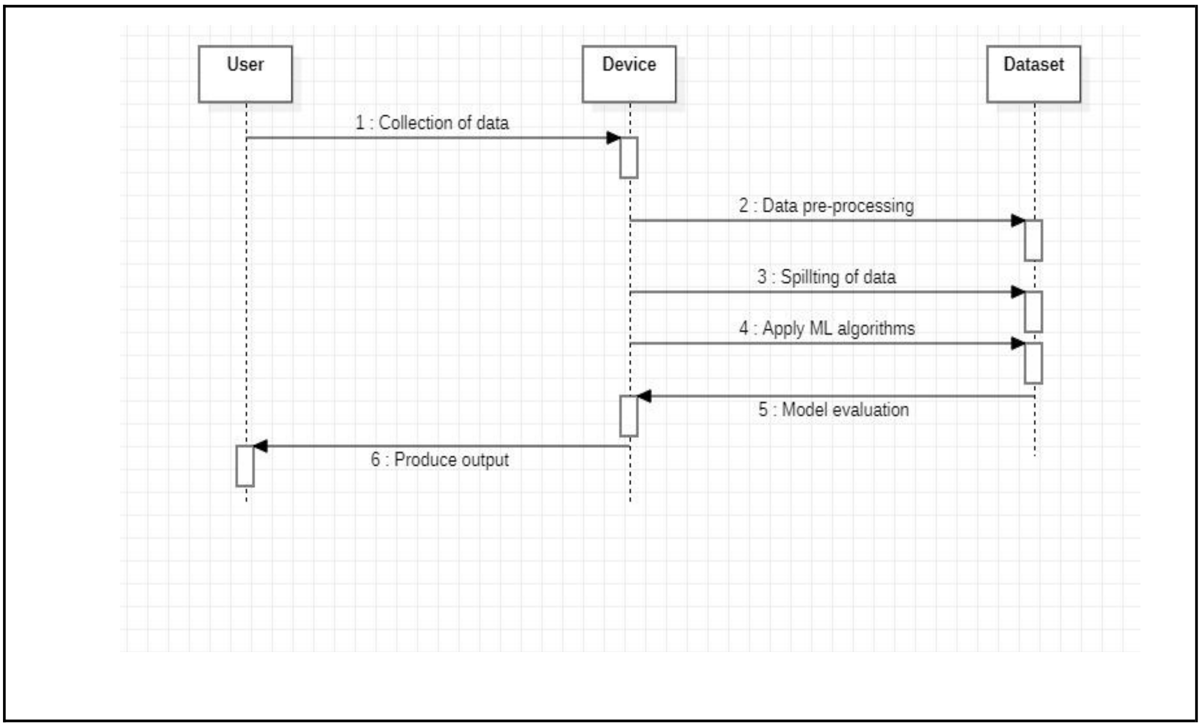
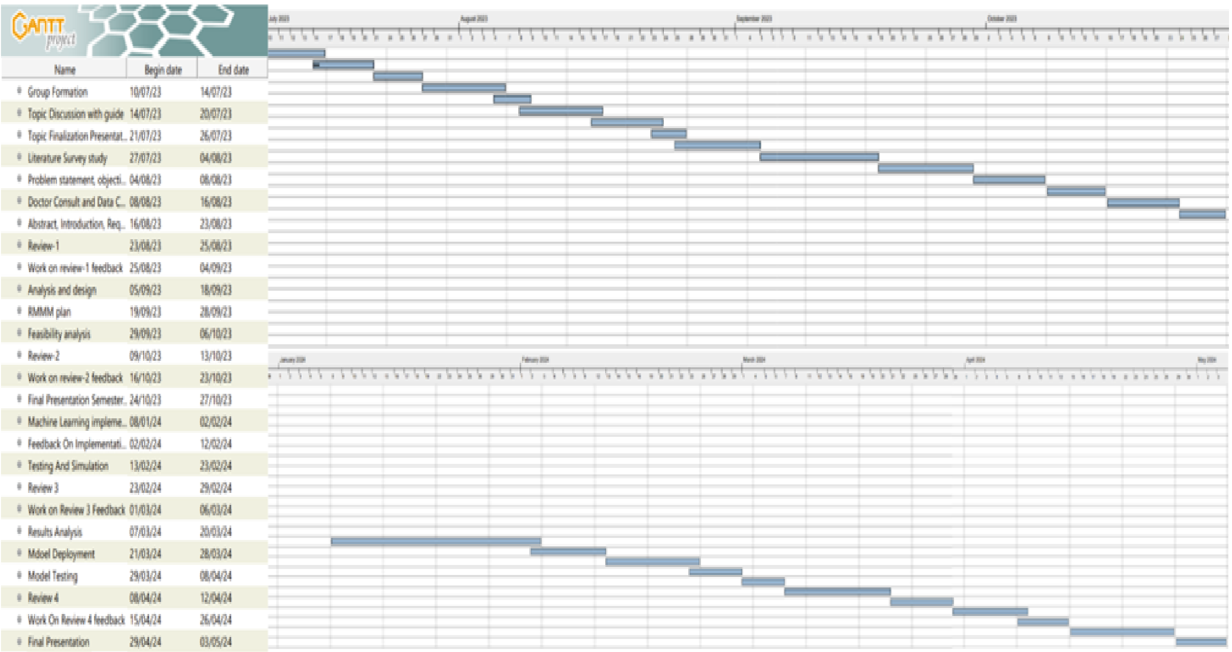


Fig 5.3 Sequence Diagram

The above figure shows the sequence of the system.

5.4 Gantt Chart



Name	Begin date	End date
Group Formation	10/07/23	14/07/23
Topic Discussion with guide	14/07/23	20/07/23
Topic Finalization Presentat...	21/07/23	26/07/23
Literature Survey study	27/07/23	04/08/23
Problem statement, objecti...	04/08/23	08/08/23
Doctor Consult and Data C...	08/08/23	16/08/23
Abstract, Introduction, Req...	16/08/23	23/08/23
Review-1	23/08/23	25/08/23
Work on review-1 feedback	25/08/23	04/09/23
Analysis and design	05/09/23	18/09/23
RMMM plan	19/09/23	28/09/23
Feasibility analysis	29/09/23	06/10/23
Review-2	09/10/23	13/10/23
Work on review-2 feedback	16/10/23	23/10/23
Final Presentation Semester...	24/10/23	27/10/23
Machine Learning impleme...	08/01/24	02/02/24
Feedback On Implementati...	02/02/24	12/02/24
Testing And Simulation	13/02/24	23/02/24
Review 3	23/02/24	29/02/24
Work on Review 3 Feedback	01/03/24	06/03/24
Results Analysis	07/03/24	20/03/24
Mdoel Deployment	21/03/24	28/03/24
Model Testing	29/03/24	08/04/24
Review 4	08/04/24	12/04/24
Work On Review 4 feedback	15/04/24	26/04/24
Final Presentation	29/04/24	03/05/24

Fig 5.4 Gantt Chart

The above figure shows the sequence of the tasks performed during the project.

Chapter 6

Methodology

6.1 Methodology

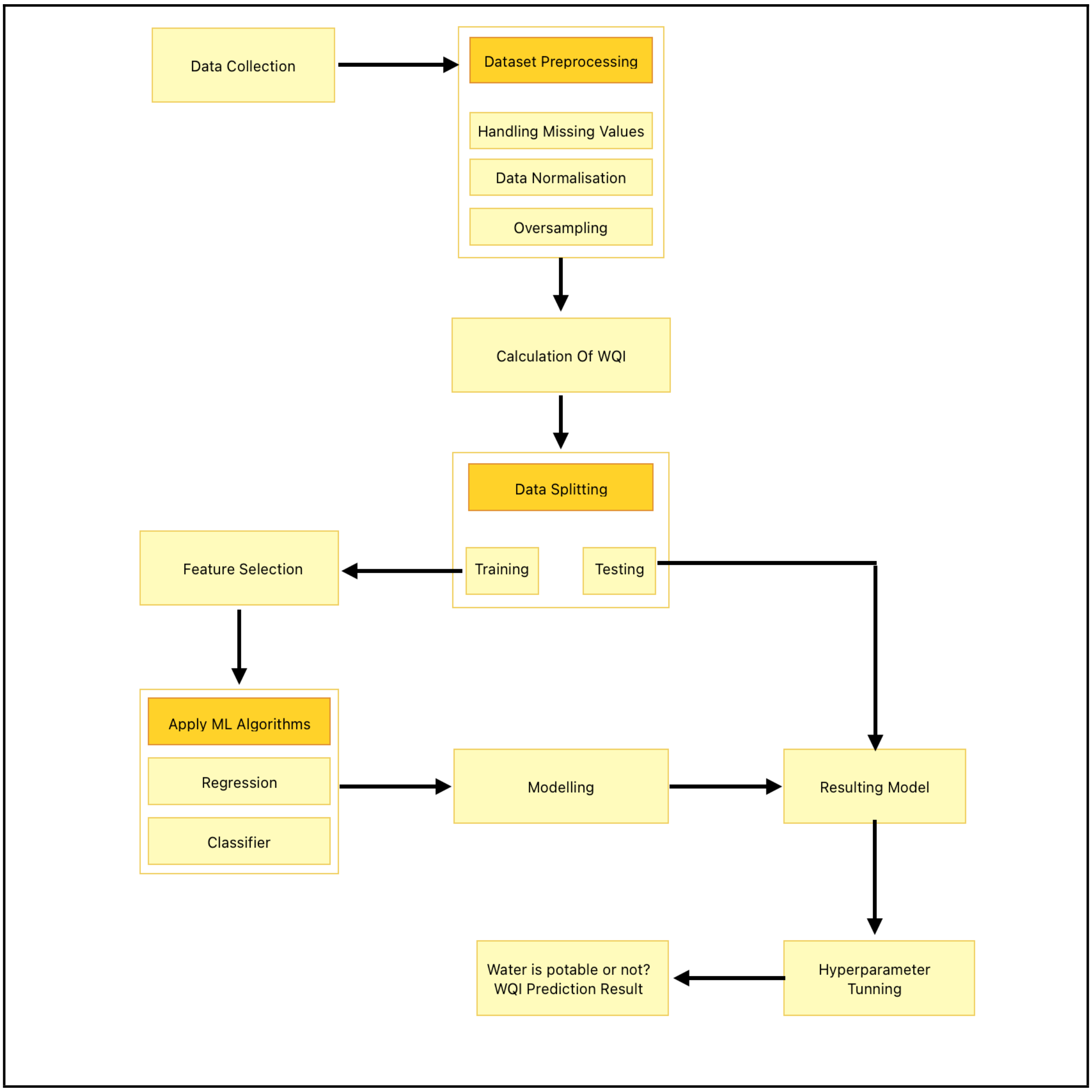
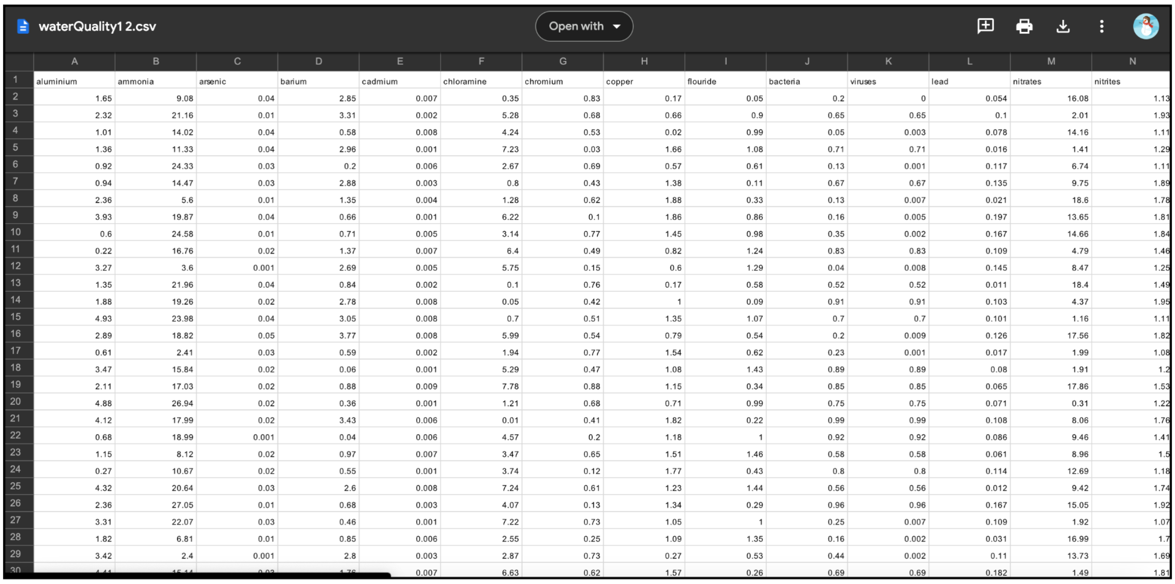


Fig 6.1 Overview of the proposed system

6.2 Data Collection: This study utilized the Water Quality Dataset available on Kaggle as its primary source of data. It has 7999 records and 20 features.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	aluminium	ammonia	arsenic	barium	cadmium	chloramine	chromium	copper	flouride	bacteria	viruses	lead	nitrates	nitrites
2	1.65	9.08	0.04	2.85	0.007	0.35	0.83	0.17	0.05	0.2	0	0.054	16.08	1.13
3	2.32	21.16	0.01	3.31	0.002	5.28	0.68	0.66	0.9	0.65	0.65	0.1	2.01	1.93
4	1.01	14.02	0.04	0.58	0.008	4.24	0.53	0.02	0.99	0.05	0.003	0.078	14.16	1.11
5	1.36	11.33	0.04	2.96	0.001	7.23	0.03	1.66	1.08	0.71	0.71	0.016	1.41	1.29
6	0.92	24.33	0.03	0.2	0.006	2.67	0.69	0.57	0.61	0.13	0.001	0.117	6.74	1.11
7	0.94	14.47	0.03	2.88	0.003	0.8	0.43	1.38	0.11	0.67	0.67	0.135	9.75	1.89
8	2.36	5.6	0.01	1.35	0.004	1.28	0.62	1.88	0.33	0.13	0.007	0.021	18.6	1.78
9	3.93	19.87	0.04	0.66	0.001	6.22	0.1	1.86	0.86	0.16	0.005	0.197	13.65	1.81
10	0.6	24.58	0.01	0.71	0.005	3.14	0.77	1.45	0.98	0.35	0.002	0.167	14.66	1.84
11	0.22	16.76	0.02	1.37	0.007	6.4	0.49	0.82	1.24	0.83	0.83	0.109	4.79	1.46
12	3.27	3.6	0.001	2.69	0.005	5.75	0.15	0.6	1.29	0.04	0.008	0.145	8.47	1.25
13	1.35	21.96	0.04	0.84	0.002	0.1	0.76	0.17	0.58	0.52	0.52	0.011	18.4	1.49
14	1.88	19.26	0.02	2.78	0.008	0.05	0.42	1	0.09	0.91	0.91	0.103	4.37	1.95
15	4.93	23.98	0.04	3.05	0.008	0.7	0.51	1.35	1.07	0.7	0.7	0.101	1.16	1.11
16	2.89	18.82	0.05	3.77	0.008	5.99	0.54	0.79	0.54	0.2	0.009	0.126	17.56	1.82
17	0.61	2.41	0.03	0.59	0.002	1.94	0.77	1.54	0.62	0.23	0.001	0.017	1.99	1.08
18	3.47	15.84	0.02	0.06	0.001	5.29	0.47	1.08	1.43	0.89	0.89	0.08	1.91	1.2
19	2.11	17.03	0.02	0.88	0.009	7.78	0.88	1.15	0.34	0.85	0.85	0.065	17.86	1.53
20	4.88	26.94	0.02	0.36	0.001	1.21	0.68	0.71	0.99	0.75	0.75	0.071	0.31	1.22
21	4.12	17.99	0.02	3.43	0.006	0.01	0.41	1.82	0.22	0.99	0.99	0.108	8.06	1.76
22	0.68	18.99	0.001	0.04	0.006	4.57	0.2	1.18	1	0.92	0.92	0.086	9.46	1.41
23	1.15	8.12	0.02	0.97	0.007	3.47	0.65	1.51	1.46	0.58	0.58	0.061	8.96	1.5
24	0.27	10.67	0.02	0.55	0.001	3.74	0.12	1.77	0.43	0.8	0.8	0.114	12.69	1.18
25	4.32	20.64	0.03	2.6	0.008	7.24	0.61	1.23	1.44	0.56	0.56	0.012	9.42	1.74
26	2.36	27.05	0.01	0.68	0.003	4.07	0.13	1.34	0.29	0.96	0.96	0.167	15.05	1.92
27	3.31	22.07	0.03	0.46	0.001	7.22	0.73	1.05	1	0.25	0.007	0.109	1.92	1.07
28	1.82	6.81	0.01	0.85	0.006	2.55	0.25	1.09	1.35	0.16	0.002	0.031	16.99	1.7
29	3.42	2.4	0.001	2.8	0.003	2.87	0.73	0.27	0.53	0.44	0.002	0.11	13.73	1.69
30	4.41	15.14	0.03	1.76	0.007	6.63	0.62	1.57	0.26	0.69	0.69	0.182	1.49	1.81

Fig 6.2 Screenshot of Dataset

6.3 Data Preprocessing:

6.3.1 Handling Missing Values: one of the most commonly used approaches for addressing missing values in numeric columns is to calculate the mean. However, it's important to note that the mean may not be appropriate when dealing with outliers; in such cases, addressing outliers should be the initial step before applying the mean imputation method

6.3.2 Data Normalization: The z-score is a popular method of normalization that indicates the number of standard deviations. It is best if it is between -3 and +3. It converts all the values with different scales to the default scale by normalizing the dataset.

6.3.3 Oversampling using SMOTE: SMOTE, which stands for Synthetic Minority Over-sampling Technique, is a common method in machine learning and data analysis. Its purpose is to tackle the problem of class imbalance in datasets. It achieves this by creating artificial data points for the minority class (the less common category) to balance the class distribution. This approach helps prevent the model from favoring the majority class and can enhance the performance of algorithms, particularly in classification tasks.

6.4 WQI Calculation: Selected Parameters are used to calculate the WQI in the traditional way.

6.5 Data Splitting: In order to train the model, the data must be split, tested with a subset of the data, and computed with accuracy measures to determine the model's performance in the final stage before applying the machine learning model. Training data and test data were created from the dataset.

6.6 Feature Selection: We conducted a correlation analysis to identify potential relationships among all the features, aiming to discover dependent features using readily available variables.

6.7 ML Algorithms used:

6.7.1 Logistic Regression

6.7.2 Support Vector Machine Classifier

6.7.3 Decision Tree Classifier

6.7.4 Random Forest Classifier

6.7.5 Gradient Boost Classifier

6.8 Modeling: The Performance of model is analyzed on the basis of Accuracy, Precision, Recall, F1 score. Choose suitable machine learning algorithms for classification.

6.9 Hyperparameter Tuning: will be performed in order to improve performance of overall model

6.10 Cross Validation: used cross validation to evaluate the final model.

6.11 Comparative Analysis: Comparing the various results and filtering the best one out according to their accuracy.

6.12 Result: For given input, the respective result is predicted.

Chapter 7

Conclusion and Future Scope

7.1 Conclusion

This machine learning project focused on predicting water quality has demonstrated promising results. By leveraging advanced algorithms and a comprehensive dataset, we were able to develop a reliable model for estimating water quality parameters. The model's performance was assessed through rigorous testing and validation, showcasing its ability to provide accurate predictions.

This project holds significant potential for practical applications in monitoring and managing water resources. It can serve as a valuable tool for environmental agencies, ensuring the safety and quality of water sources for communities. Overall, this endeavor highlights the immense potential of machine learning in addressing critical environmental concerns, and underscores the importance of continued research in this field to enhance our ability to safeguard our natural resources.

7.2 Future Scope

There are several key areas for advancement in the field of water quality prediction using machine learning. Firstly, expanding the scope of data collection by incorporating additional variables could significantly enhance the model's predictive capabilities. Furthermore, delving into more sophisticated feature engineering techniques and exploring advanced algorithms.

Further research and refinement of the model could lead to even more accurate and robust predictions.

References

[1] Ye, Y.; Zhang, J.; Liu, H.; Zhu, W. Predictive Simulation Study on the Effect of Small and Medium River Basin Outfall Treatment Measures on Water Quality Improvement. Water 2023, 15, 2359. https://doi.org/10.3390/w15132359

[2] D. Brindha, V. Puli, B. K. S. NVSS, V. S. Mittakandala and G. D. Nanneboina, "Water Quality Analysis and Prediction using Machine Learning," 2023 7th International Conference on Computing Methodologies and Communication (ICCMC), Erode, India, 2023, pp. 175-180, doi: 10.1109/ICCMC56507.2023.10083776.

[3] Rui Tan, Zhaocai Wang, Tunhua Wu, Junhao Wu,A data-driven model for water quality prediction in Tai Lake, China, using secondary modal decomposition with multidimensional external features,Journal of Hydrology: Regional Studies,Volume 47,2023,101435,ISSN 2214-5818,https://doi.org/10.1016/j.ejrh.2023.101435.

[4] Kumar, Vijay Patel, Jinal Amipara, Charmi Ahanger, Tariq Ahamed Ladhva, Komal Gupta, Rajeev Kumar Alsaab, Hashem O. Althobaiti, Yusuf S. “A Machine Learning-Based Water Potability Prediction Model by Using Synthetic Minority Oversampling Technique and Explainable AI”.A Machine Learning-Based Water Potability Prediction Model by Using Synthetic Minority Oversampling Technique and Explainable AI. Volume 2022. 2022 https://doi.org/10.1155/2022/9283293.

[5] Mohamed Torky, Ali Bakhiet, Mohamed Bakrey, Ahmed Adel Ismail, Ahmed I. B. EL Seddawy. "Recognizing Safe Drinking Water and PredictingWater Quality Index using Machine Learning Framework".(IJACSA) International Journal of Advanced Computer Science and Applications,Vol. 14, No. 1, 2023.

[6] Kaddoura, S. Evaluation of Machine Learning Algorithm on Drinking Water Quality for Better Sustainability. Sustainability 2022, 14, 11478. https://doi.org/10.3390/su141811478

[7] L. Sheng, Z. Jian, P. Yifan, & L. Liu, "Water quality prediction method based on preferred classification", IET Cyber-Physical Systems: Theory &Amp; Applications, vol. 5, no. 2, p. 176-180, 2020. https://doi.org/10.1049/iet-cps.2019.0062.

[8] Nur Afyfah Suwadi, Morched Derbali, Nor Samsiah Sani, Meng Chun Lam, Haslina Arshad, Imran Khan, Ki-Il Kim, "An Optimized Approach for Predicting Water Quality

Features Based on Machine Learning", Wireless Communications and Mobile Computing, vol. 2022, Article ID 3397972, 20 pages, 2022. https://doi.org/10.1155/2022/3397972

[9] K. Abirami, P. C. Radhakrishna and M. A. Venkatesan, "Water Quality Analysis and Prediction using Machine Learning," 2023 IEEE 12th International Conference on Communication Systems and Network Technologies (CSNT), Bhopal, India, 2023, pp. 241-245, doi: 10.1109/CSNT57126.2023.10134661.

[10] R. Gai and J. Yang, "Summary of Water Quality Prediction Models Based on Machine Learning," 2021 IEEE 23rd Int Conf on High Performance Computing & Communications; 7th Int Conf on Data Science & Systems; 19th Int Conf on Smart City; 7th Int Conf on Dependability in Sensor, Cloud & Big Data Systems & Application (HPCC/DSS/SmartCity/DependSys), Haikou, Hainan, China, 2021, pp.2338-2343, doi: 10.1109/HPCC-DSS-SmartCity-DependSys53884.2021.00353.

[11] S. Babu, B. B. Nagaleela, C. G. Karthik and L. N. Yepuri, "Water Quality Prediction using Neural Networks," 2023 International Conference on Artificial Intelligence and Knowledge Discovery in Concurrent Engineering (ICECONF), Chennai, India, 2023, pp. 1-6, doi:10.1109/ICECONF57129.2023.10084120.

[12] Algalil, Fahd Abd, Aldhyani, Theyazn H. H,Al-Yaari, Mohammed,Alkahtani, Hasan,Maashi, Mashael."Water Quality Prediction Using Artificial Intelligence Algorithms".Applied Bionics andBiomechanics.Volume- 2020. 2020 https://doi.org/10.1155/2020/6659314.

[13] Md. Saikat Islam Khan, Nazrul Islam, Jia Uddin, Sifatul Islam, Mostofa Kamal Nasir.“Water quality prediction and classification based on principal component regression and gradient boosting classifier approach”.Journal of King Saud University - Computer and Information Sciences.Volume 34 PartA,2022. https://doi.org/10.1016/j.jksuci.2021.06.003.

[14] J. P. Nair and M. S. Vijaya, "Predictive Models for River Water Quality using Machine Learning and Big Data Techniques - A Survey," 2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS), Coimbatore, India, 2021, pp. 1747-1753, doi: 10.1109/ICAIS50930.2021.9395832.

[15] S. Palkar, S. Usgaonkar and S. Ansari, "Wq-Net: A Deep Neural Network Model For Water Quality Prediction," OCEANS 2022 - Chennai, Chennai, India, 2022, pp. 1-6, doi: 10.1109/OCEANSChennai45887.2022.9775235.