UTILISING ML TO ANALYSE AND FORECAST INDIAN WATER QUALITY

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

Bachelor of Technology
in
Computer Science & Engineering

By

J. SHIVA SAI	(19UECS0389)	**(VTU11524)**
J. MANIKANTA	(19UECS0383)	**(VTU11527)**
P. SWAMY SATISH	(19UECS0776)	**(VTU12766)**

Under the guidance of
Dr. A. Suresh, M.E, 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

UTILISING ML TO ANALYSE AND FORECAST INDIAN WATER QUALITY

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

Bachelor of Technology
in
Computer Science & Engineering

By

J. SHIVA SAI	(19UECS0389)	**(VTU11524)**
J. MANIKANTA	(19UECS0383)	**(VTU11527)**
P. SWAMY SATISH	(19UECS0776)	**(VTU12766)**

Under the guidance of
Dr. A. SURESH, M.E, 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 "UTILISING ML TO ANALYSE AND FORECAST INDIAN WATER QUALITY" by J. SHIVA SAI (19UECS0389), J. MANIKANTA (19UECS0383), P. SWAMY SATISH (19UECS0776) has been carried out under my supervision and that this work has not been submitted elsewhere for a degree.

Signature of Supervisor
Dr. A. Suresh
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
Dr. M. S. Murali Dhar
Associate Professor
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.

J. SHIVA SAI

Date: / /

J. MANIKANTA

Date: / /

P. SWAMY SATISH

Date: / /

APPROVAL SHEET

This project report entitled “UTILISING ML TO ANALYSE AND FORECAST INDIAN WATER QUALITY” by J. SHIVA SAI (19UECS0389), J. MANIKANTA (19UECS0383), P. SWAMY SATISH (19UECS0776) is approved for the degree of B.Tech in Computer Science & Engineering.

Examiners **Supervisor**

Dr. A. Suresh, M.E, 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. A. SURESH, M.E, Ph.D.,** for his cordial support, valuable information and guidance, he 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.

J. SHIVA SAI **(19UECS0389)**
J. MANIKANTA **(19UECS0383)**
P. SWAMY SATISH **(19UECS0776)**

ABSTRACT

Water quality analysis is essential for ensuring the safety and cleanliness of water resources. Traditional methods of water quality analysis are often time-consuming and expensive. The emergence of Machine Learning (ML) as a tool for water quality analysis has the potential to automate the process and improve accuracy. ML algorithms can be trained to identify patterns and relationships in large datasets of water quality parameters, enabling efficient decision-making for water management. This technology can help to identify water quality issues early, reduce costs associated with manual sampling and testing, and ensure the safety and sustainability of water resources. The various ways in which machine learning is being used in water quality analysis and its potential to revolutionize the field.Thus, the quality of water is very important in both environmental and economic aspects. Thus, water quality analysis is essential for using it in any purpose. After years of research, water quality analysis is now consists of some standard protocols. There are guidelines for sampling, preservation and analysis of the samples. Here the standard chain of action is discussed briefly so that it may be useful to the analysts and researchers. The accuracy of our project is 56.2% .

Keywords : Machine learning(ML), Parameters, Revoluionize, Preservation, Accuracy

LIST OF FIGURES

LIST OF ACRONYMS AND ABBREVIATIONS

GB	Giga Byte
IDE	Integrated Development Environment
LSTM	Long Short Term Memory
ML	Machine Learning
PH	Potential of Hydrogen
RAM	Random Access Memory
SMS	Short Message Service
UI	User Interface
UML	Unified Modeling Language

TABLE OF CONTENTS

Chapter 1

INTRODUCTION

1.1 Introduction

Water quality analysis is the process of measuring and monitoring the physical, chemical, and biological properties of water. It is essential to ensure the safety and cleanliness of water for human consumption and various other purposes. Traditional methods for water quality analysis involve manual collection of water samples and laboratory testing, which can be time-consuming and expensive.Machine learning (ML) has emerged as a promising tool for water quality analysis as it can help to automate the process and improve accuracy. ML algorithms can be trained to identify patterns and relationships in large datasets of water quality parameters and predict water quality conditions in real-time. This technology can help to identify water quality issues early, reduce costs associated with manual sampling and testing, and enable efficient decision-making for water management. In this way, machine learning can help to ensure the safety and sustainability of water resources for future generations.ML algorithms can help to identify patterns in historical water quality data and predict future changes in water quality, providing valuable insights for policymakers, water managers, and the general public. By utilizing ML to analyse and forecast Indian water quality, it may be possible to develop more effective strategies for water management and protect the health and well being of indian communities.

1.2 Aim of the Project

The aim of this project is to develop a machine learning based system for water quality analysis. The system will utilize machine learning algorithms to analyze large datasets of water quality parameters and provide accurate and real time predictions of water quality conditions.

1.3 Project Domain

The Project is based on machine learning which is sub field of deep learning.

we predict the water quality. This project would involve analyzing large amounts of water quality data from various sources in India, such as government agencies, NGOs, and research institutions, using machine learning algorithms.

1.4 Scope of the Project

Evaluation of the system's performance and comparison with traditional methods of water quality analysis. Development and testing of machine learning algorithms for water quality prediction.

Chapter 2

LITERATURE REVIEW

Juntao Lui et al., [2020] In the construction of prediction model, the deep Bi-S-SRU network used in the experiment is superior to most other neural networks in terms of prediction accuracy. The experimental results also show that the Bi-S-SRU-based prediction method is only slightly higher in time complexity than the traditional RNN-based or LSTM-based prediction method. In the actual prediction, the average prediction time taken was 12.5ms, and the prediction accuracy can reach 94.42 percent.[3]

Nur Aqulah Paskhal Rostam et al., [2021] Based on discussion and analysis, it was observed that LSTM with the right features outperformed the other methods and grasped the temporal behaviour and tackled the dynamic issues. Besides, even though during this study the MF was excluded, and more CF and PF were included, this study outperformed the other studies. a complete framework that discusses in detail both IoT and predictive modelling that consists of the main phases such as data acquisition, data management and lastly, predictive modelling.[6]

Dhruti Dheda et al., [2021] The proposed LSTM based ensemble scheme improved the tolerance (mitigated the discrepancies of the individual LSTM models) of the hybrid GA-optimised LSTM water quality prediction models, for different water quality datasets taken from different sites and different times. Future studies can identify the optimum number of LSTM based models required to make the most tolerant ensemble model for water quality prediction and possibly in other areas such as energy, finance, geology, and many more.[7]

Ali Omran Al-Sulttani et al., [2021] This study was proposed five relatively new explored ML models for BOD of surface WQ prediction. These models were considered in this work as a robust approach towards the prediction of WQ parameters rather than relying on laboratory analysis. Various categories of water parameters, including physical, chemical, and biological parameters were used for the develop-

ment of the proposed models as the input attributes. Future studies are aimed at the prediction of other WQ parameters, as well as the inclusion of more input attributes, such as climatological or hydrological factors.[5]

K. P. Rasheed Abdul Haq And V. P. Harigovindan., [2022] The water quality parameters data was collected from aquaculture ponds located in Kollam, Kerala, under ADAK. Another dataset used was the MAC dataset which was collected from the marine aquaculture base in Xincun Town, LingShui County, Hainan Province, China. The hybrid models have a similar performance compared with the attention-based models. Still, they outperform the attention based models in computation time, offering a realistic solution for predicting water quality parameters in smart aquaculture.[8]

Chapter 3

PROJECT DESCRIPTION

3.1 Existing System

In Existing System we have used Linear regression, Linear regression is a statistical modeling technique that seeks to model the relationship between a dependent variable and one or more independent variables. In the context of water quality analysis, linear regression can be used to identify the relationship between water quality parameters (such as pH, temperature, or dissolved oxygen) and other factors such as time of day, season, or weather conditions. Linear regression can help to identify which factors are most important in determining water quality and can be used to make predictions about future water quality based on historical data. In terms of their strengths and weaknesses, linear regression is a simple and widely used technique that can be effective when the relationship between the dependent and independent variables is linear.

3.2 Proposed System

In proposed system we have used decision trees are a type of machine learning algorithm that can be used for both classification and regression analysis. Decision trees are constructed by dividing the data into smaller subsets based on a set of rules or criteria, with the goal of creating a tree like structure that can be used to make predictions or classify new data. In the context of water quality analysis, decision trees can be used to identify the most important parameters or factors in determining water quality, and can be used to make predictions or classifications based on historical data. It capture more complex relationships and interactions between variables, but may be prone to overfitting or creating overly complex models that do not generalize well to new data.

3.3 Feasibility Study

A feasibility study for the "Utilizing ML to analyse and forecast Indian water quality" project would involve an analysis of the technical, economic, and social.

3.3.1 Economic Feasibility

The cost of implementing the project should be assessed to determine whether it is economically feasible. This would include costs associated with acquiring and maintaining the necessary hardware and software, as well as the cost of hiring experts in data analysis and ML. Additionally, the potential benefits of the project should be considered, such as the ability to identify and address water quality issues before they become major problems.

3.3.2 Technical Feasibility

ML can be used to analyze large datasets and make predictions based on patterns identified in the data. To make this project technically feasible, we would need access to large amounts of reliable data on water quality in India. This data should include information on key parameters such as pH, temperature, dissolved oxygen, and conductivity, as well as information on weather patterns, rainfall, and human activities that may impact water quality.

3.3.3 Social Feasibility

Water is a vital resource, and water quality is a critical concern for public health, agriculture, and the environment. The project should be assessed for its ability to address these concerns and improve water quality in india. The potential benefits of the project, such as identifying and addressing water quality issues before they become major problems, should be communicated effectively to key stakeholders.

3.4 System Specification

3.4.1 Hardware Specification

• RAM : 8.00 GB(7.78 GB usable).
• Hard Disk space : 8GB

• Processor : Intel Core i5

3.4.2 Software Specification

• Operating System : Windows 7,8,10 (or) Mac (or) Linux.
• Coding Language : Python - version(3.5).
• Platform Requirement: Jupyter notebook.

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.

Standard Used: ISO/IEC 27001

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.

Standard Used: ISO/IEC 27001

Chapter 4

METHODOLOGY

4.1 Architecture Diagram for Water Quality Prediction

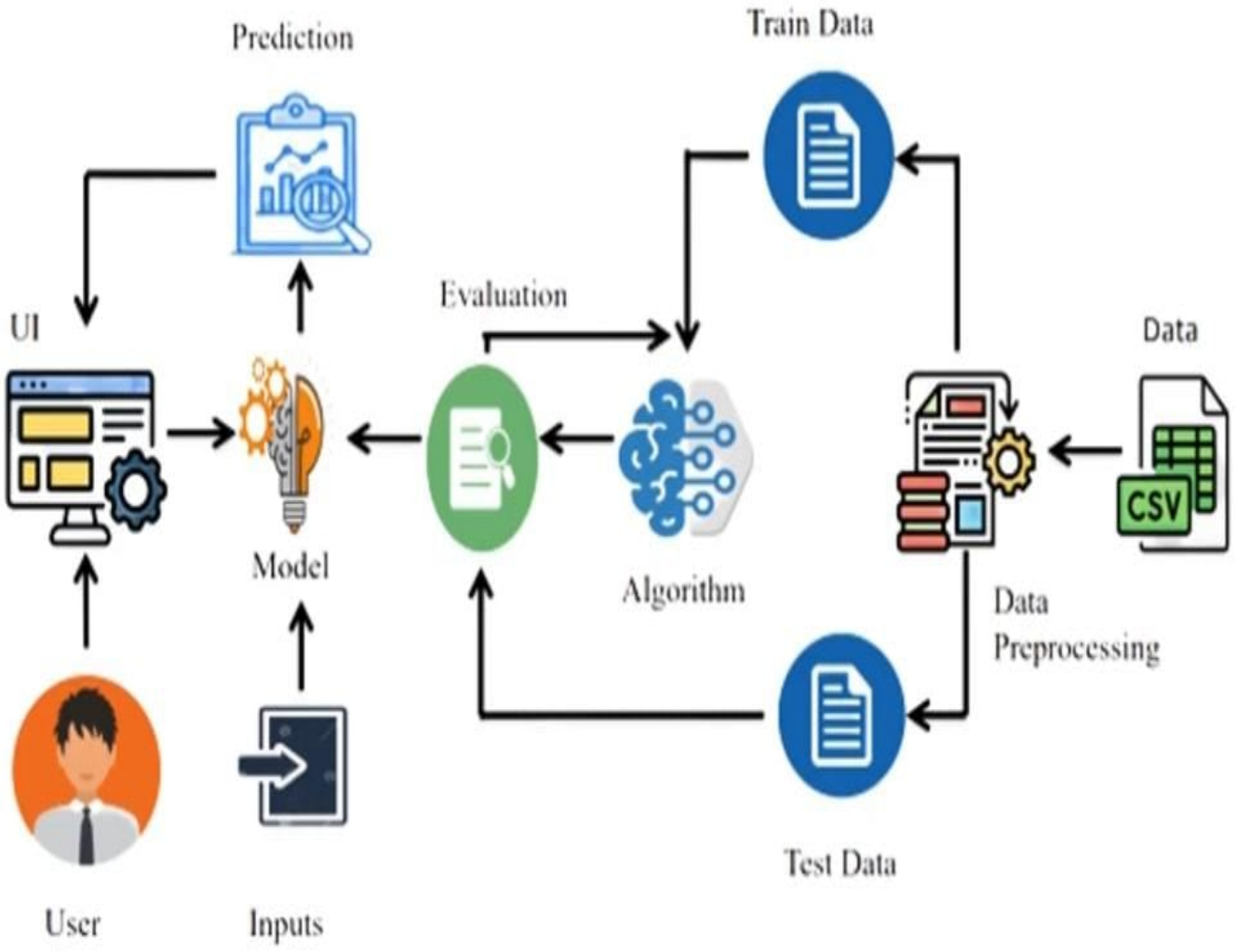


Figure 4.1: **Architecture Diagram for Water Quality Prediction**

The above Figure 4.1 is a Architecture digram for water quality analysis, we train the dataset and preprocess the data then training the data and testing the data, we use ML algorithm and evaluate then the model shows the predictions and it reaches user interface.

4.2 Design Phase

4.2.1 Data Flow Diagram

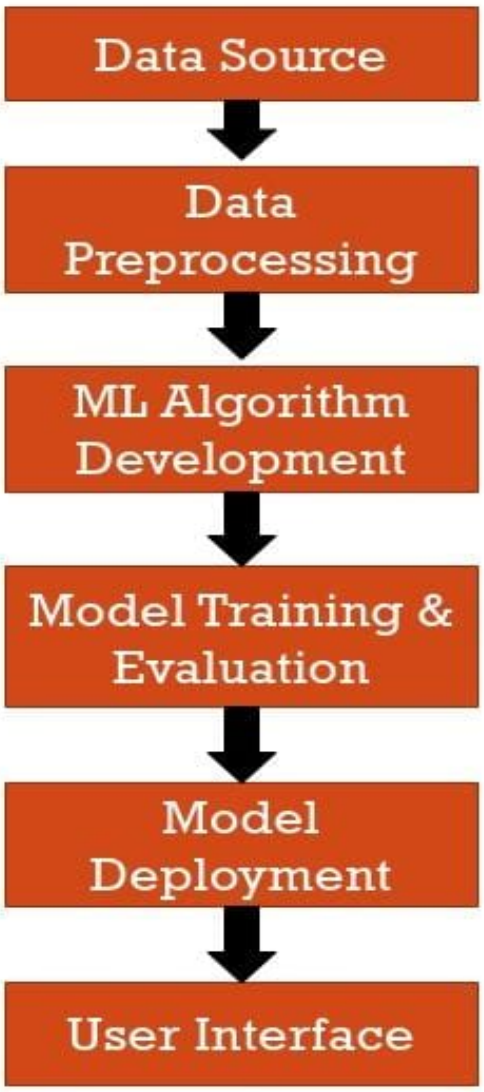


Figure 4.2: **Data Flow Diagram**

The Figure 4.2 shows a data flow diagrm, the data source refers to the various sources of water quality data in india. This raw data is then preprocessed to clean and prepare it for analysis. Next, feature extraction techniques are applied to convert the data into a set of feature vectors that can be used by the ML model. The ML model is then trained on this data, using various algorithms and techniques to learn patterns and relationships in the data. Once the model is trained, it is tested to ensure that it is accurate and can make reliable predictions. Finally, the model is used to forecast future water quality in India, based on the data it has learned from the past.

4.2.2 Usecase Diagram

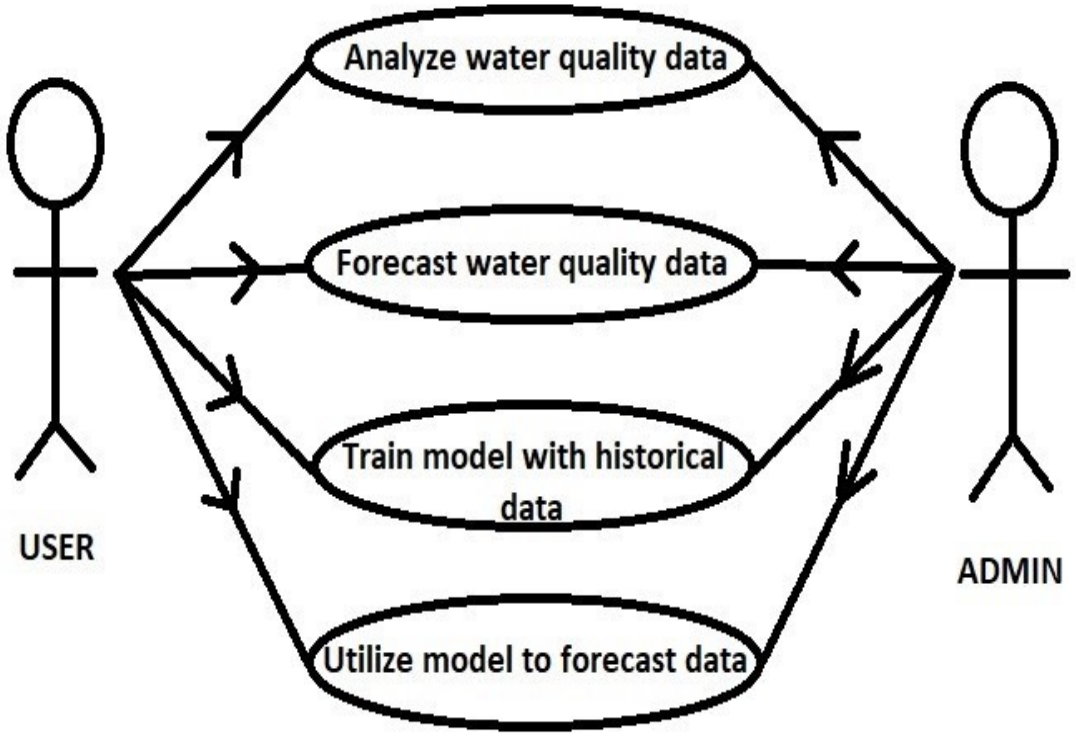


Figure 4.3: **Usecase Diagram**

The Figure 4.3 shows usecase diagram, it summarizes some of the relationships between use cases, actors, and systems. It does not show the order in which steps are performed to achieve the goals of each use case. To check water quality analyze water data and gather water quality data, To forecast water quality train the machine learning models. The user views the predicted water quality results. The system sends notifications to the user regarding the predicted water quality and any alerts. The use case diagram provides a highl evel overview of the main functionalities and interactions of the water quality prediction system, which can be useful for system design and communication.

4.2.3 Sequence Diagram

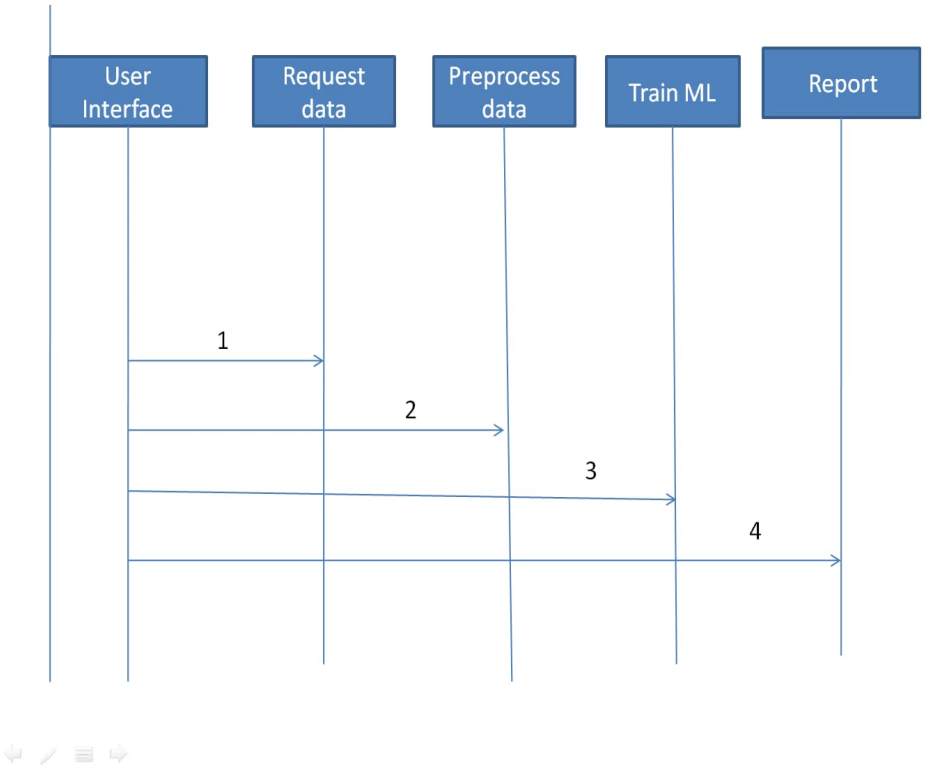


Figure 4.4: **Sequence Diagram**

The Figure 4.4 shows the sequence diagram, the user interface (UI) where users can interact with the system to perform various tasks, such as requesting water quality data analysis or viewing reports. The UI can be designed to be user friendly and intuitive, providing the user with the necessary information and options to make informed decision. Users can request water quality data to be analyzed through the UI. The request can include parameters such as the location, type of water source, and time frame. Once the request is received, the data collection component can collect the relevant water quality data from various sources, such as sensors or manual measurements. The collected data can then be preprocessed, including cleaning, filtering, and transforming, to prepare it for analyse. Once the ML model is trained, it can be further optimized and fine tuned to improve its accuracy and performance. This can involve adjusting the model parameters, changing the algorithms used, or adding more data to the training set. Once the ML model is ready, it can be used to generate reports or predictions about future water quality. The reports can be presented to the user through the UI or other channels, such as email or SMS.

4.2.4 Class Diagram

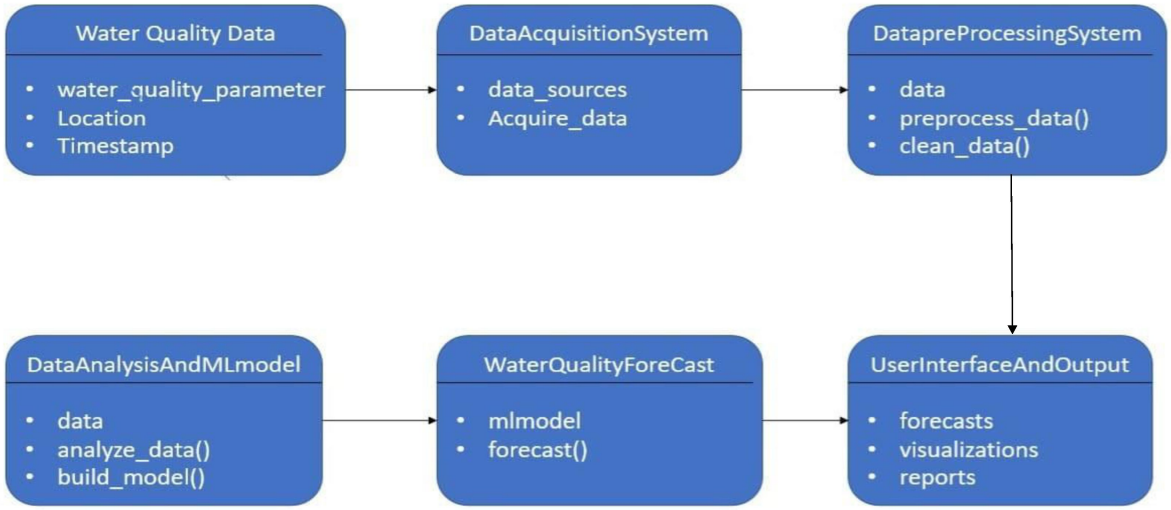


Figure 4.5: **Class Diagram**

The Figure 4.5 shows the class diagram, A class diagram is an illustration of the relationships and source code dependencies among classes in the Unified Modeling Language (UML). In this context, a class defines the methods and variables in an object, which is a specific entity in a program or the unit of code representing that entity.

4.3 Algorithm & Pseudo Code

4.3.1 Decision Tree Algorithm

In “utilizing ml to analyze and forecast Indian water quality prediction”, the decision tree algorithm can be used as a machine learning technique to develop a predictive model for water quality. The decision tree algorithm is a supervised learning algorithm that can be used for both classification and regression tasks. In the context of water quality prediction, the decision tree algorithm can be used to predict the values of water quality parameters such as pH, temperature, dissolved oxygen, or turbidity. The decision tree algorithm works by recursively splitting the data into

subsets based on the selected features, such as pH or temperature. Each split creates a decision node, which is associated with a test on the selected feature. The algorithm continues to split the data into smaller subsets until the subsets contain only one class or meet a predefined stopping criterion.

4.3.2 Pseudo Code

BEGIN
df = pd.read csv(”waterpotability.csv”)
df shape()
df.isnull().sum()
df.info()
df.describe()
df.fillna(df.mean(), inplace=True)
df.isnull().sum()
df.hist(figsize=(14,14)):
plt.figure(figsize=(13,8))
sns.heatmap(df.corr(),annot=True,cmap=’terrain’
return
END

4.4 Module Description

4.4.1 Collection of Data

We have collected data from the UCI repository which is main domain to collect the data sets for the Machine Learning which has 10 different attributes related to our project.

4.4.2 Splitting Dataset

Dividing the dataset into two sets should be done precisely. The dataset can be divided into the ratio of 80% train set, 20% test set or 70% train set, 30% test set, or any other way. The division of the dataset also affects the accuracy of the training model. A slicing operation can be performed to separate the dataset. Wehavve take care while splitting the dataset, assure that the test set must hold an equivalent features as the train set and also the datasets must be statistically meaningful.

4.4.3 Pre-Processing

Data Cleaning:

The process of preparing data for analysis by removing or modifying data that is incorrect or irrelevent and improperly formatted.

Handling missing values:

Using attribute mean the missing values are replaced by the mean of all attribute values.

Handling outliers:

Outliers are defined as samples that are significantly differ from the remaing data. It refers to the conversion of continuous attributes discretized or nominal attributes. Equal areas are used to discretize the continuous attributes in the dataset.

4.5 Steps to execute/run/implement the project

4.5.1 Install

• Install the required software and libraries, such as
Python
NumPy
Pandas.

4.5.2 Process

• Import all the required libraries which are used to train the model or visualise the data. Then load the data set using a Pandas's function read csv() and display the top five rows of the data set.
• Then finally handle the missing values. fill the missing values in our feachers using a mean value of each feature which means fill the mean value to handle missing data. Then again check that there are null values present or not.
• Now visualize the pH value using a distplot function to check that it contains a normal distribution or not. So, you can see that it is a normal distribution.

4.5.3 Code Execution

• Run the code using run command in jupyter notebook.

Chapter 5

IMPLEMENTATION AND TESTING

5.1 Input and Output

The type of input data required for water quality analysis depends on the specific objectives of the project. However, some common types of data that may be used in water quality analysis.

5.1.1 Input Design

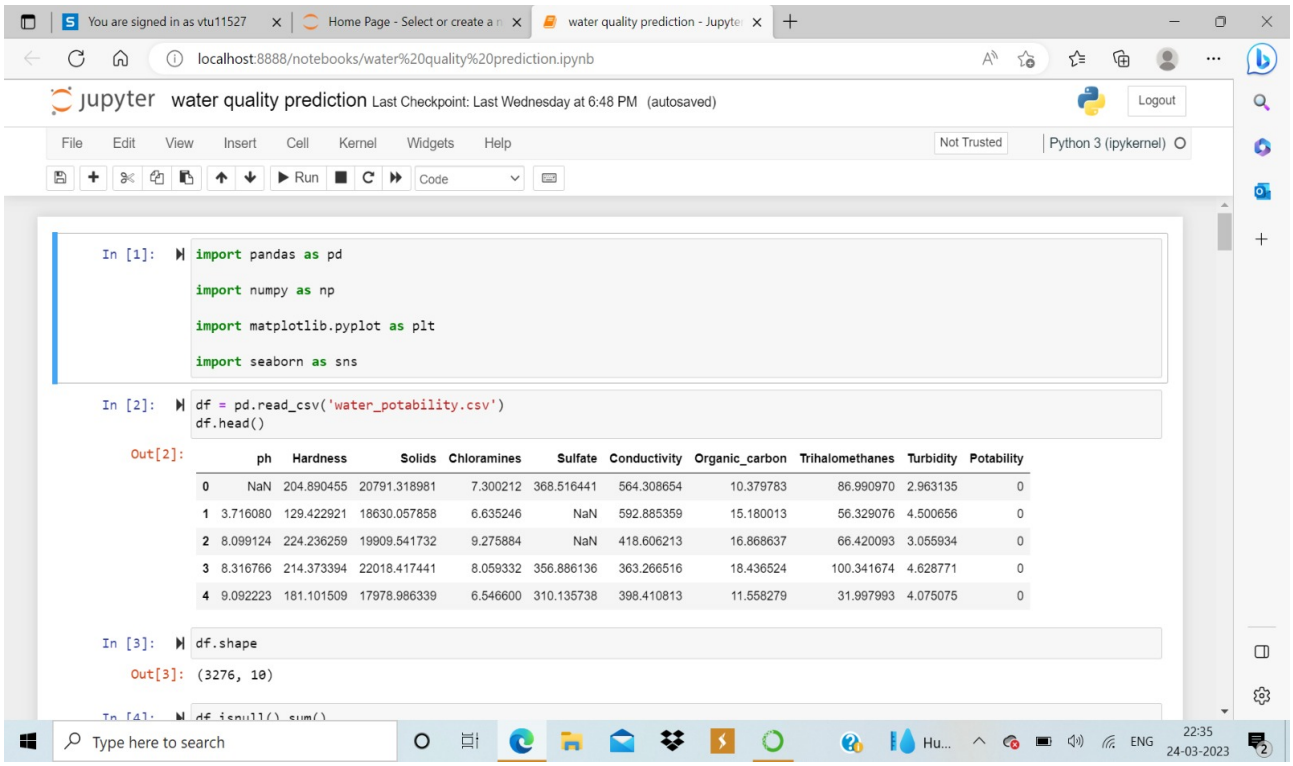


Figure 5.1: **Water Quality Prediction**

The Figure 5.1 shows water quality prediction, it imports all the required libraries which are used to train the model or visualise the data. Then load the data set using a pandas's function read csv() and display the top five rows of the data set.

5.1.2 Output Design

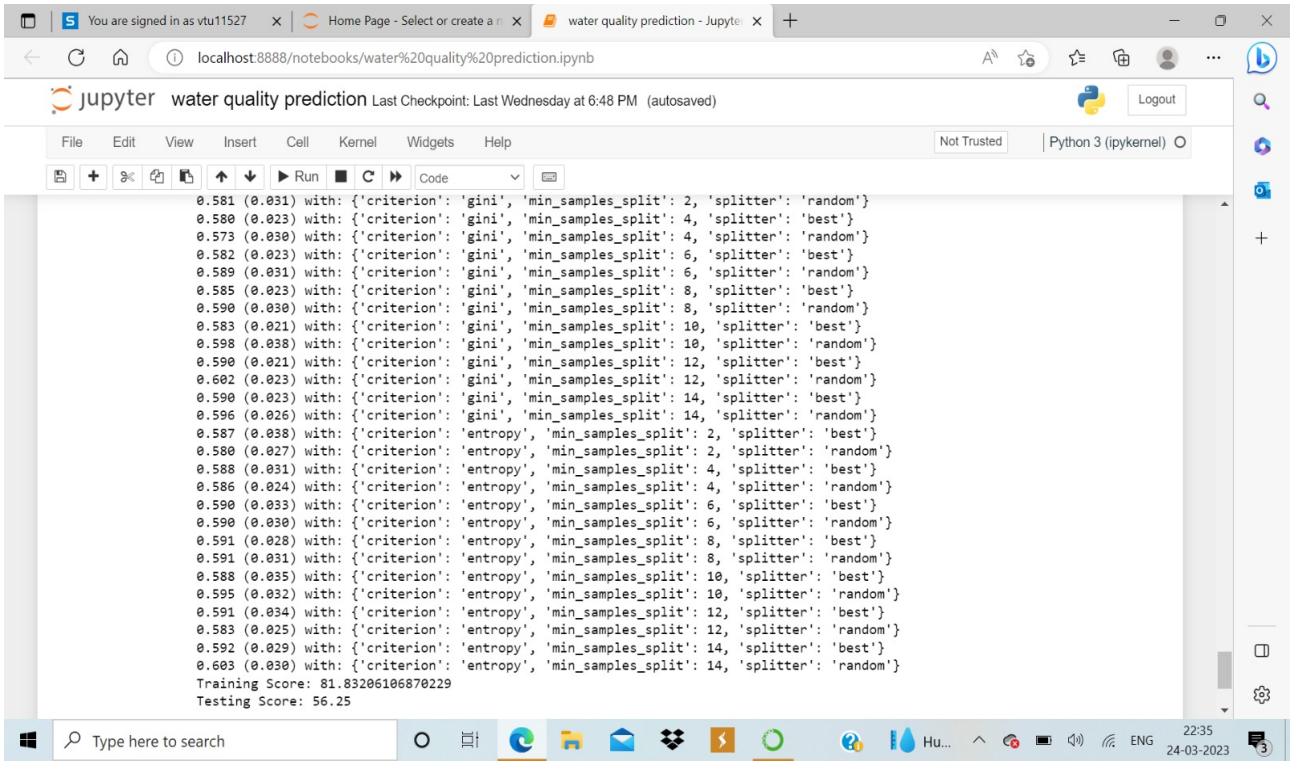


Figure 5.2: **Training and Testing score**

The Figure 5.2 shows traning and testing score of the water quality reports, water quality index, maps, results and that shows the real quality of indian water, and shows the how much percent water is safe to drink.

5.2 Testing

Testing is an essential part of any software project, including projects that utilize Machine earning (ML) algorithms to analyze and forecast Indian water quality. The testing process in such process that involves verifying the ML models that are accurate, reliable, and robust enough to provide meaningful insights into the water quality in India. Identify the objectives and goals of the project, such as identifying the parameters that affect water quality and forecasting future trends. Collect and prepare the data for training and testing the ML models. This involves cleaning, filtering, and transforming the data to ensure that it is of high quality and suitable for use in the ML algorithms. Choose the appropriate ML algorithms for the project based on the goals and objectives. This could include supervised or unsupervised learning algorithms, such as regression, decision trees, or neural networks. Train

the ML models using the prepared data and evaluate their accuracy and performance using a validation dataset.

5.3 Types of Testing

5.3.1 Unit Testing

Unit testing is a software testing technique where individual units or components of a software application are tested in isolation from the rest of the system to ensure that each unit works as expected. By performing unit testing, we can ensure that each component of our water quality analysis software application works as expected and that any issues or bugs are identified and fixed early in the development process, reducing the risk of costly and time-consuming rework later on.

5.3.2 Integration Testing

Integration testing is a type of software testing that verifies the proper functioning of multiple modules or components of an application when integrated together. By performing integration testing, we can ensure that the various modules or components of our Indian water quality analysis software application work together seamlessly and produce accurate results. This can help to minimize the risk of errors or inaccuracies in the analysis, which is essential for maintaining the quality of water in India.

5.3.3 System Testing

System testing is a type of software testing that evaluates the complete system or software application as a whole. By performing system testing,we can ensure that the water quality analysis software application meets the functional and non functional requirements of the project and performs accurately and reliably in real world scenarios. This can help to ensure that the quality of water in a given region is maintained and protected from potential hazards.

5.3.4 Test Result

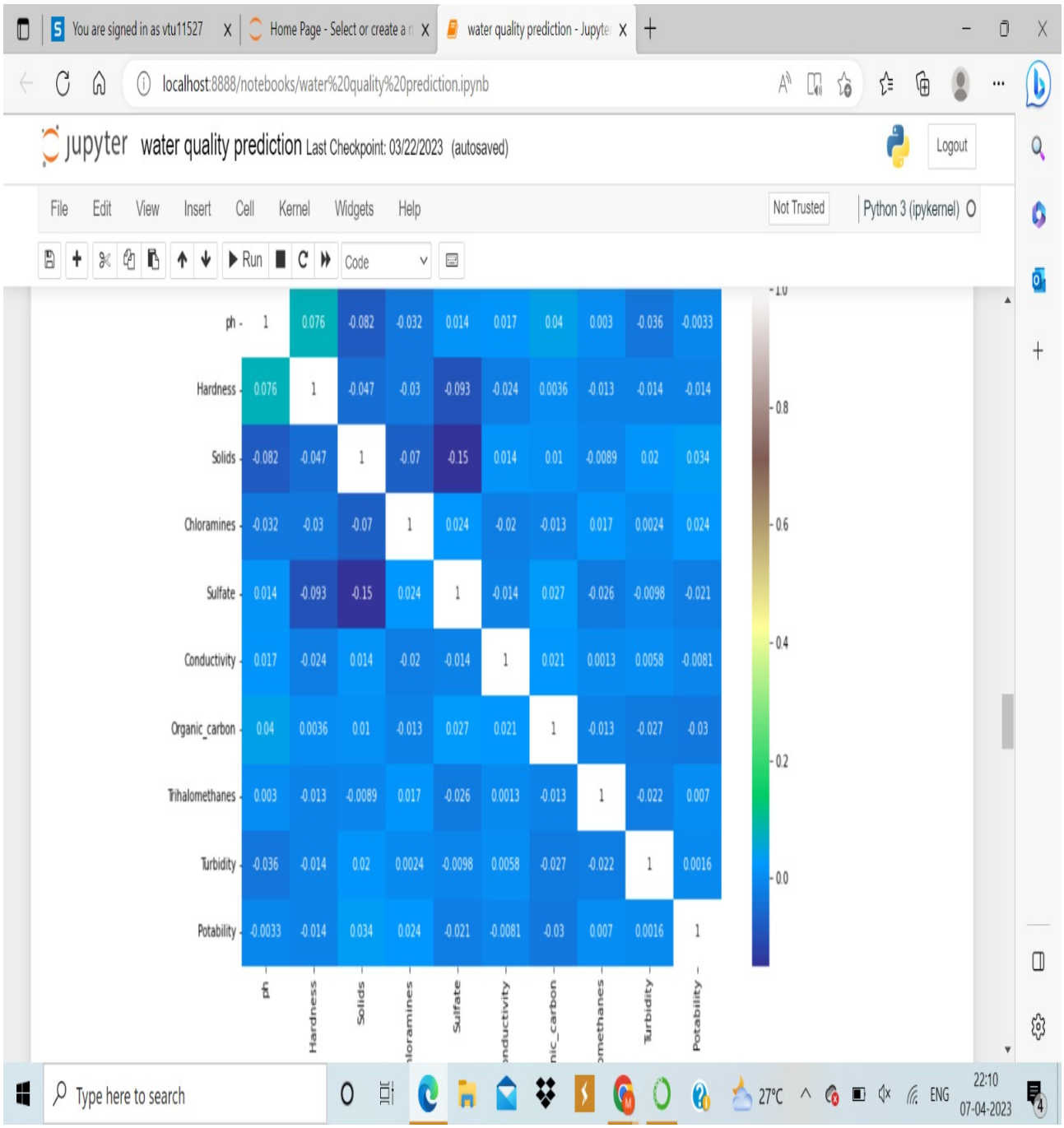


Figure 5.3: **Heat Map**

Chapter 6

RESULTS AND DISCUSSIONS

6.1 Efficiency of the Proposed System

The proposed system is based on the Decision Tree Algorithm.Accuracy of proposed system is done by using decision tree gives the output. Decision Tree gives more accuracy than Linear Regression. Decision trees are a type of machine learning algorithm that can be used for classification and regression analysis. In the context of the "utilizing ML to analyze and forecast Indian water quality prediction" project, decision trees can be used to identify the factors that are most important in predicting water quality.

For example, a decision tree can be constructed using historical data on water quality, which can include parameters such as pH, temperature, dissolved oxygen, and levels of various pollutants. The decision tree algorithm would then analyze the data and identify which parameters are most important in determining water quality. This information can be used to develop a predictive model that can be used to forecast future water quality. The decision tree algorithm can also be used for classification analysis, which can help identify whether a water sample is contaminated or not. The algorithm can be trained on historical data on contaminated water samples and non-contaminated water samples, and then used to classify new water samples based on their parameters.

6.2 Comparison of Existing and Proposed System

Linear Regression:(Existing system)

Linear regression is a statistical modeling technique that seeks to model the relationship between a dependent variable and one or more independent variables. In the context of water quality analysis, linear regression can be used to identify the relationship between water quality parameters (such as pH, temperature, or dissolved oxygen) and other factors such as time of day, season, or weather conditions. Linear

regression can help to identify which factors are most important in determining water quality and can be used to make predictions about future water quality based on historical data. In terms of their strengths and weaknesses, linear regression is a simple and widely used technique that can be effective when the relationship between the dependent and independent variables is linear.

Decision Tree Algorithm:(Proposed system)

Decision trees are a type of machine learning algorithm that can be used for both classification and regression analysis. Decision trees are constructed by dividing the data into smaller subsets based on a set of rules or criteria, with the goal of creating a tree-like structure that can be used to make predictions or classify new data. In the context of water quality analysis, decision trees can be used to identify the most important parameters or factors in determining water quality, and can be used to make predictions or classifications based on historical data.It capture more complex relationships and interactions between variables, but may be prone to overfitting or creating overly complex models that do not generalize well to new data.

6.3 Sample Code

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
df = pd.read_csv('water_potability.csv')
df.head()
df.shape
df.isnull().sum
df.info()
df.describe()
df.fillna(df.mean(), inplace=True)
df.isnull().sum()
df.Potability.value_counts()
sns.countplot(df['Potability'])
plt.show()
sns.distplot(df['ph'])
plt.show()
df.hist(figsize=(14,14))
plt.show()
plt.figure(figsize=(13,8))
sns.heatmap(df.corr(),annot=True,cmap='terrain')
plt.show()
df.boxplot(figsize=(14,7))
X = df.drop('Potability',axis=1)
```

```
25 Y= df['Potability']
```

Output

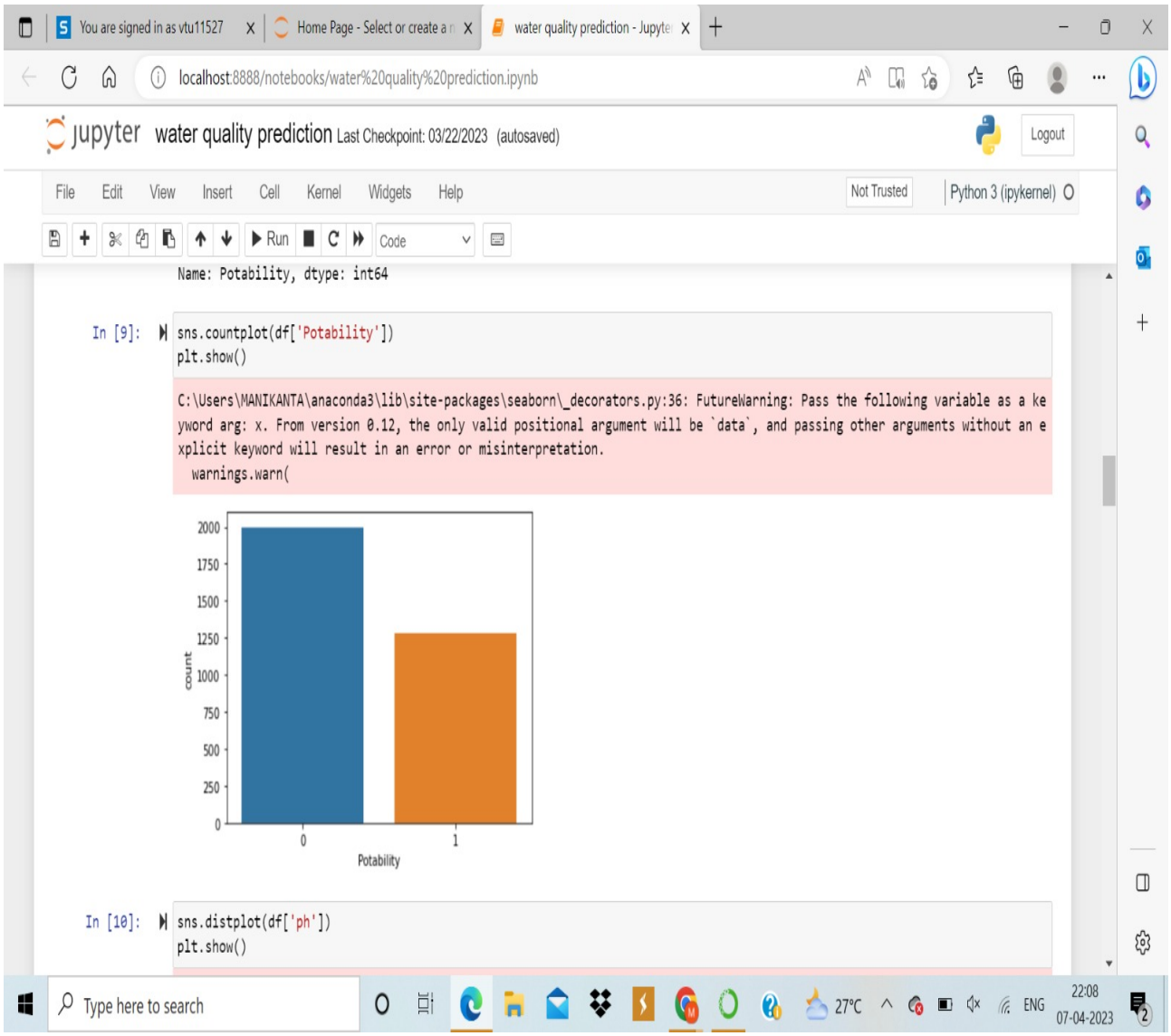


Figure 6.1: **Potability**

In figure 6.1 shows the Potability, Check the value counts of our target feature Potability. Then visualize the portability using a countplot function of seaborn. Now visualize the pH value using a distplot function to check that it contains a normal distribution or not. So, you can see that it is a normal distribution.

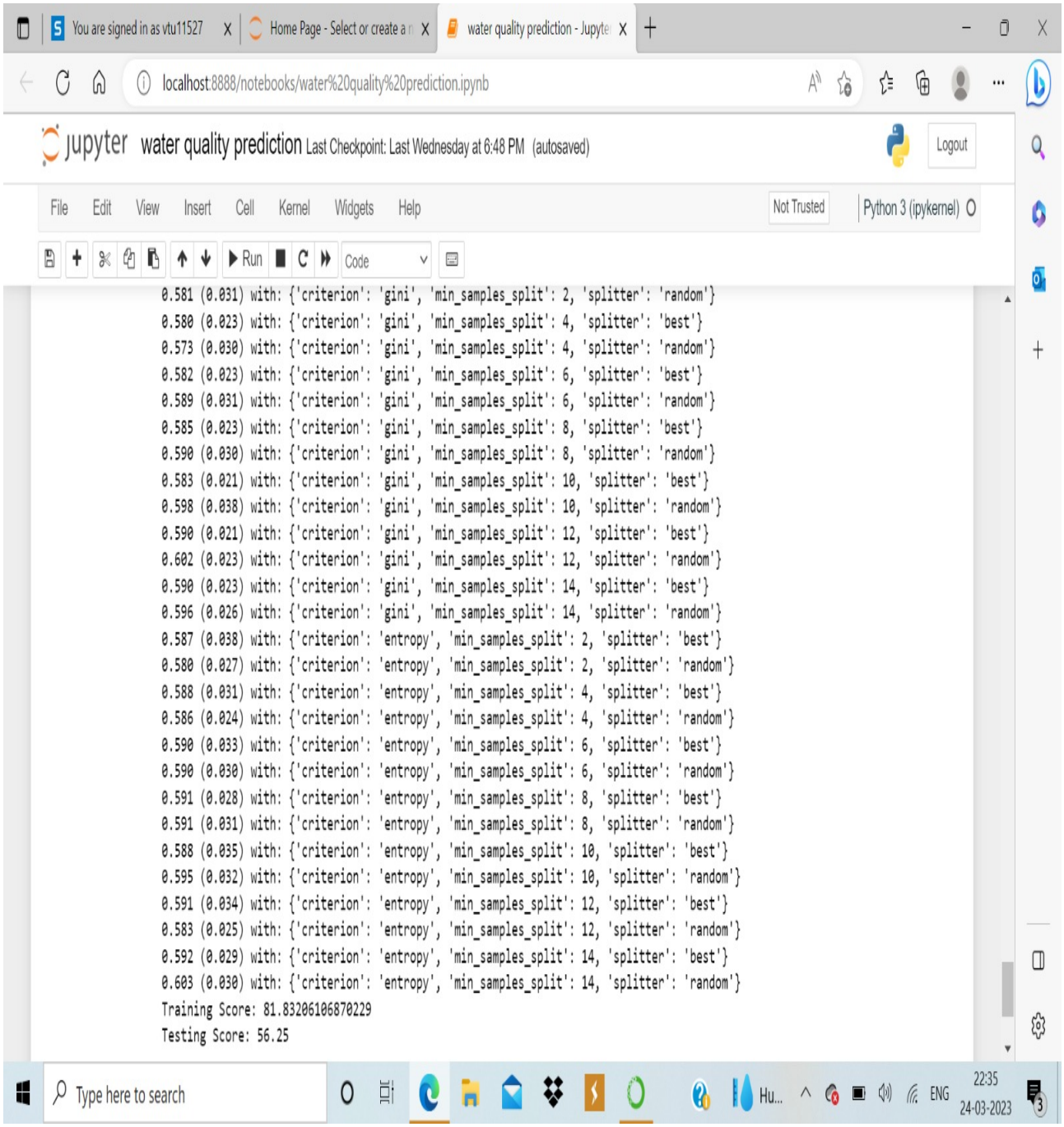


Figure 6.2: **Training And Testing Score**

In figure 6.2 shows the training and testing score, Now see how much time the model has trained with different parameters. Also check the training and testing accuracy as you can see that the training accuracy is 81.8 % and testing accuracy is 56.2 % which is okay.

Chapter 7

CONCLUSION AND FUTURE ENHANCEMENTS

7.1 Conclusion

The project of utilizing machine learning to analyze and forecast Indian water quality is a valuable endeavor that can provide critical information for policymakers, researchers, and the public. The entity relationship diagram presented here shows the various entities and relationships involved in the project, including the water samples collected, the tests performed on them, the machine learning models used for analysis and forecasting, and the geographic locations of the samples. Overall, this project represents a valuable application of machine learning and data analysis techniques to address an important issue in public health and environmental management.

7.2 Future Enhancements

In order to improve the accuracy of the water quality predictions, additional data sources could be incorporated into the ML model. For example, information on weather patterns, river flow rates, and agricultural practices could be included to provide a more comprehensive picture of the factors that impact water quality.Currently, many water quality prediction models rely on simple regression algorithms. However, more advanced ML algorithms, such as deep learning or ensemble methods, may be able to provide more accurate predictions. These algorithms can also help to identify complex patterns in the data that may not be evident using simpler methods. By continuing to improve and refine these methods, it may be possible to better protect the health and well being of indian communities that rely on clean water.

Chapter 8

INDUSTRY DETAILS

8.1 Industry name

BOSTON IT SOLUTIONS INDIA PVT LTD

8.1.1 Duration of Internship (From Date - To Date)

From 18-02-2023 To 20-05-2023

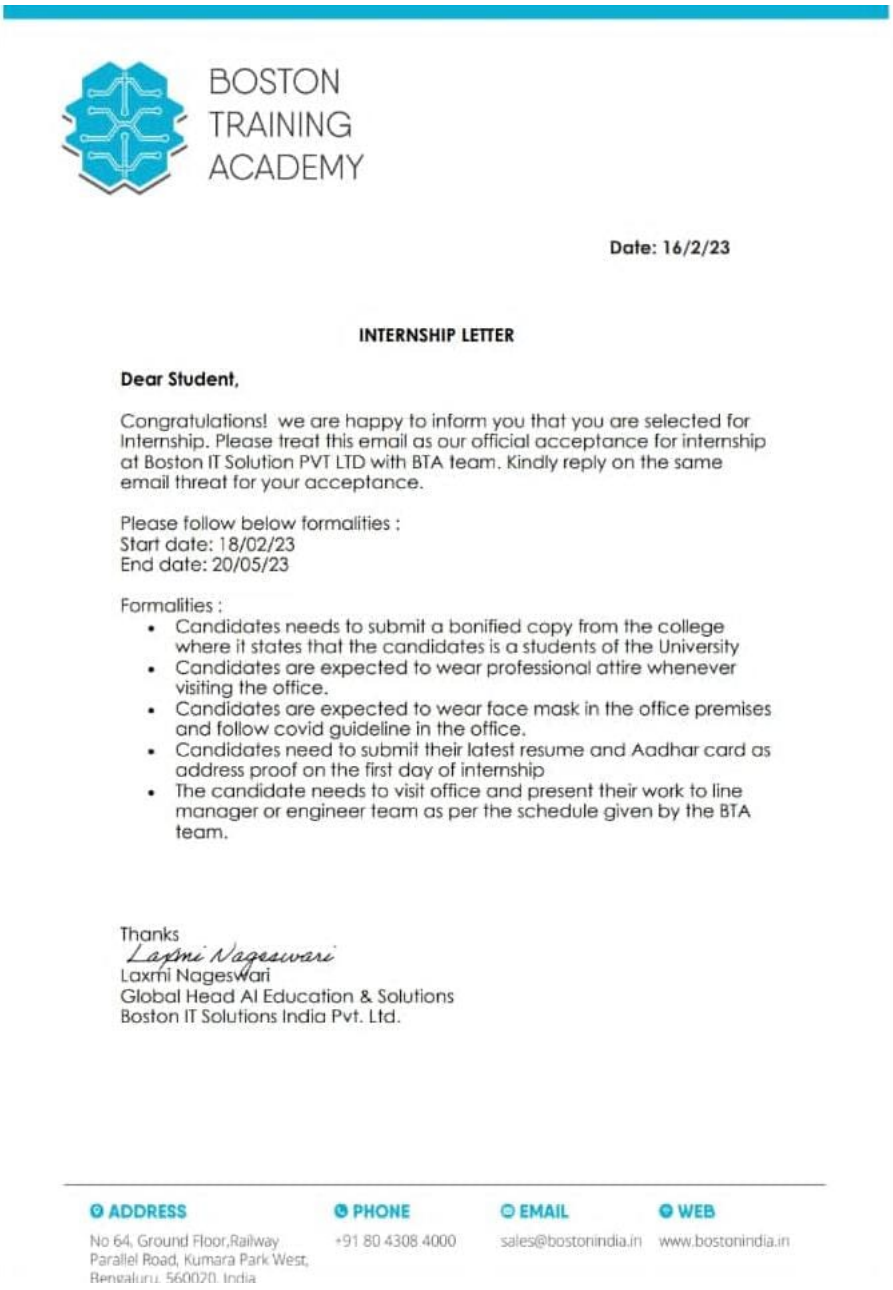
8.1.2 Duration of Internship in months

3 months

8.1.3 Industry Address

No 64,Ground Floor,Railway Parallel Road,Kumara Park West,Bangaluru,560020,India.

8.2 Internship offer letter



BOSTON
TRAINING
ACADEMY

Date: 16/2/23

INTERNSHIP LETTER

Dear Student,

Congratulations! we are happy to inform you that you are selected for Internship. Please treat this email as our official acceptance for internship at Boston IT Solution PVT LTD with BTA team. Kindly reply on the same email threat for your acceptance.

Please follow below formalities :
Start date: 18/02/23
End date: 20/05/23

Formalities :

- Candidates needs to submit a bonified copy from the college where it states that the candidates is a students of the University
- Candidates are expected to wear professional attire whenever visiting the office.
- Candidates are expected to wear face mask in the office premises and follow covid guideline in the office.
- Candidates need to submit their latest resume and Aadhar card as address proof on the first day of internship
- The candidate needs to visit office and present their work to line manager or engineer team as per the schedule given by the BTA team.

Thanks
Laxmi Nageswari
Laxmi Nageswari
Global Head AI Education & Solutions
Boston IT Solutions India Pvt. Ltd.

ADDRESS
No 64, Ground Floor,Railway Parallel Road, Kumara Park West, Bengaluru, 560020, India

PHONE
+91 80 4308 4000

EMAIL
sales@bostonindia.in

WEB
www.bostonindia.in

Chapter 9

PLAGIARISM REPORT

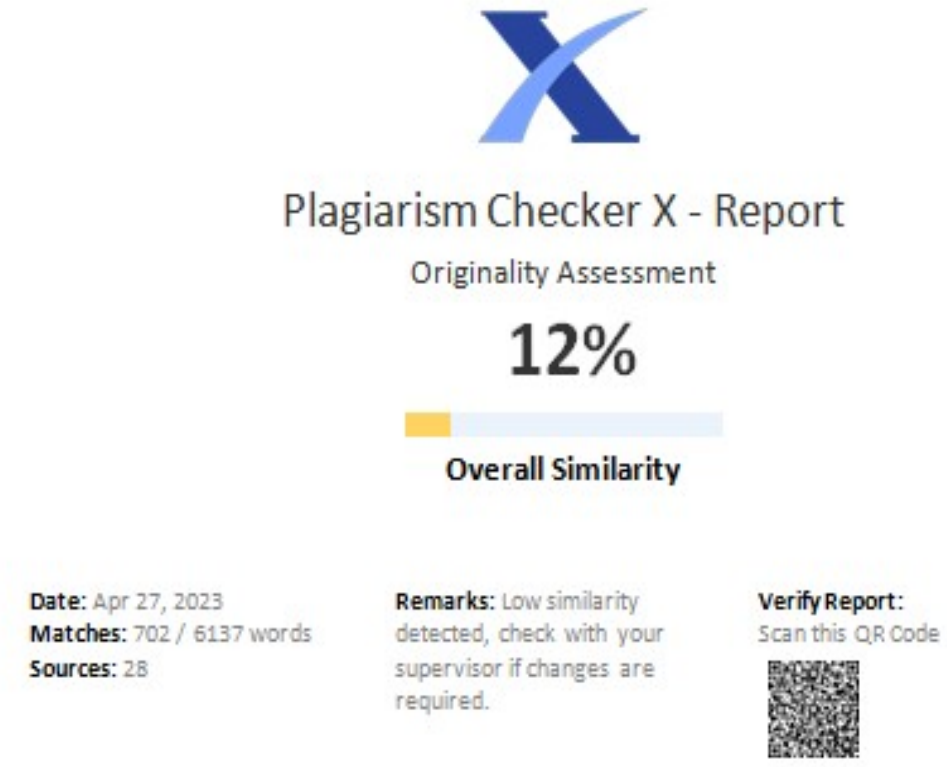


Figure 9.1: **Plagiarism Report**

Chapter 10

SOURCE CODE & POSTER PRESENTATION

10.1 Source Code

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
df = pd.read_csv('water_potability.csv')
df.head()
df.shape
df.isnull().sum
df.info()
df.describe()
df.fillna(df.mean(), inplace=True)
df.isnull().sum()
df.Potability.value_counts()
sns.countplot(df['Potability'])
plt.show()
sns.distplot(df['ph'])
plt.show()
df.hist(figsize=(14,14))
plt.show()
plt.figure(figsize=(13,8))
sns.heatmap(df.corr(),annot=True,cmap='terrain')
plt.show()
df.boxplot(figsize=(14,7))
X = df.drop('Potability',axis=1)
Y= df['Potability']
from sklearn.model_selection import train_test_split
X_train, X_test, Y_train, Y_test = train_test_split(X,Y, test_size= 0.2, random_state=1,shuffle=True
    )
from sklearn.tree import DecisionTreeClassifier
from sklearn.metrics import accuracy_score,confusion_matrix,classification_report
dt=DecisionTreeClassifier(criterion= 'gini', min_samples_split= 10, splitter= 'best')
dt.fit(X_train, Y_train)
prediction=dt.predict(X_test)
print(f"Accuracy Score = {accuracy_score (Y_test,prediction)*100}")
print (f"Confusion Matrix = {confusion_matrix(Y_test,prediction)}")
```

```
print (f"classification Report =\n {classification_report(Y_test, prediction)}")
res = dt.predict([[5.735724, 158.318741,25363.016594,7.728601,377.543291,568.304671
,13.626624,75.952337,4.732954]])[0]
res
from sklearn.model_selection import RepeatedStratifiedKFold
from sklearn.model_selection import GridSearchCV
from sklearn.model_selection import RepeatedStratifiedKFold
from sklearn.model_selection import GridSearchCV
model = DecisionTreeClassifier()
criterion = ["gini", "entropy"]
splitter = ["best", "random"]
min_samples_split = [2,4,6,8,10,12,14]
grid = dict(splitter=splitter, criterion=criterion, min_samples_split=min_samples_split)
cv = RepeatedStratifiedKFold(n_splits=10, n_repeats=3, random_state=1)
grid_search_dt = GridSearchCV(estimator=model, param_grid=grid, n_jobs=-1, cv=cv,scoring='accuracy',
    error_score=0)
grid_search_dt.fit(X_train,Y_train)
print(f"Best: {grid_search_dt.best_score_:.3f} using {grid_search_dt.best_params_}")

grid_search_dt.cv_results_['mean_test_score']
stds=grid_search_dt.cv_results_['std_test_score']
params = grid_search_dt.cv_results_['params']
for mean, stdev, param in zip(means, stds, params):
print(f"{mean:.3f} ({stdev:.3f}) with: {param}")
print("Training Score:",grid_search_dt.score(X_train, Y_train)*100)
print("Testing Score:", grid_search_dt.score(X_test, Y_test)*100)
```

10.2 Poster Presentation

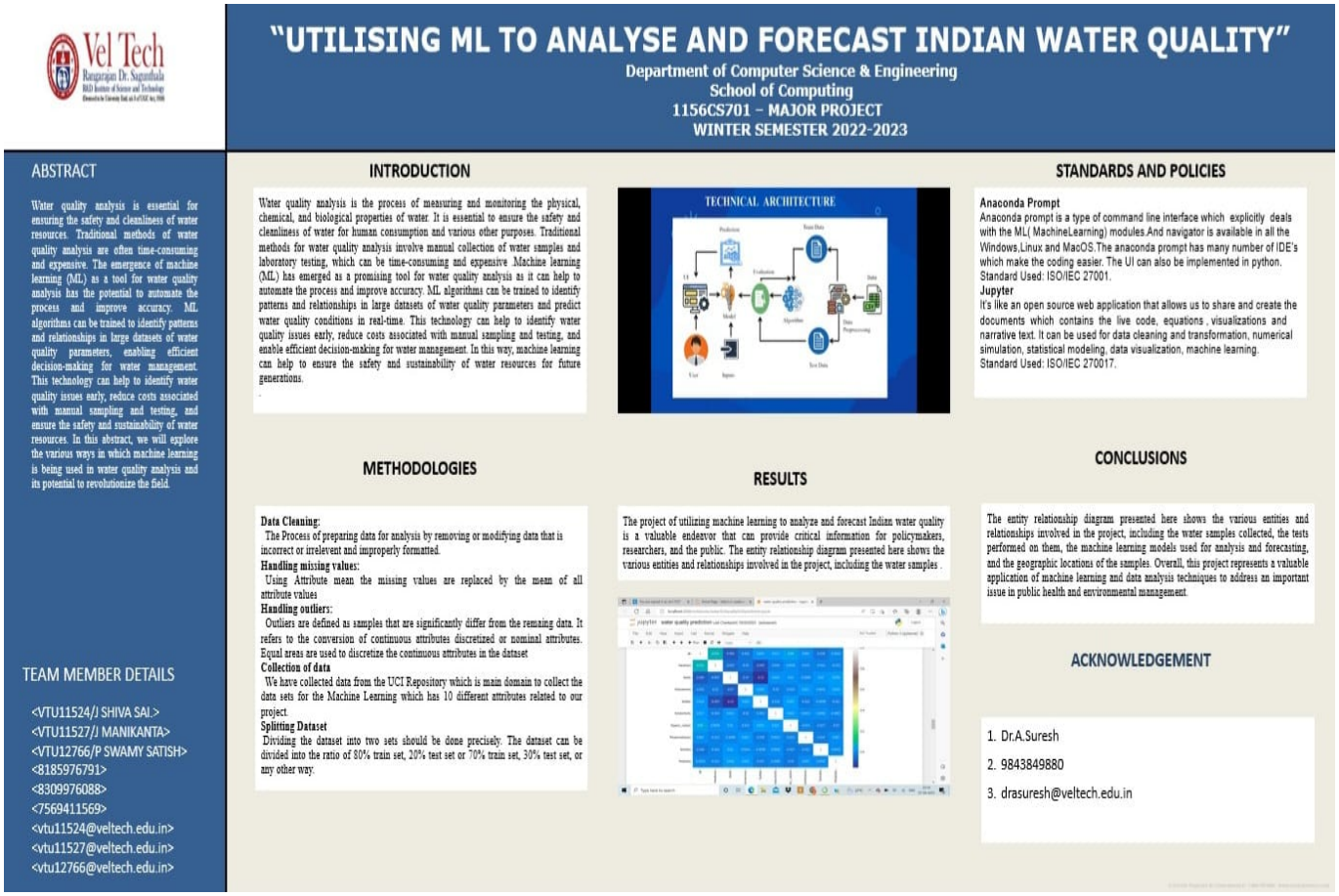


Figure 10.1: **Poster**

References

[1] Yinguo Qiu., "A Novel Spatiotemporal Data Model for River Water Quality Visualization and Analysis "IEEE Access Volume 7, 2019: 155455 – 155461.

[2] Liang Kuang., "An Enhanced Extreme Machine learning for Dissolved Oxygen Prediction in Wireless Sensor Networks" IEEE Access Volume 8, 2020 : 198730-198739.

[3] Juntao Liu., "Accurate Prediction Scheme of Water Quality in Smart Mariculture With Deep Bi-S-SRU Learning Network" IEEE Access Volume 8,2020: 24784-24798.

[4] Hadi Mohammed, and Razak Seidu "Quality Risk Analysis for Sustainable Smart Water Supply Using Data Perception" Volume 5,2020: 377-388.

[5] Ali Omran Ai-Sulttani., "Proposition of New Ensemble Data Intelligence Models for Surface Water Quality Prediction" Volume 9, 2021 : 108527 – 108541 .

[6] Nur Aqilah Paskhal Rostam., "A Complete Proposed Framework for Coastal Water Quality Monitoring System With Algae Predictive Model" IEEE Access Volume 9, 2021:108249 -108265.

[7] Dhruti Dheda et al., "Long ShortTerm Memory Water Quality Predictive Model Discrepancy Mitigation Through Genetic Algorithm Optimisation and Ensemble Modeling" Volume 10, 2021 : 24638- 24658.

[8] K. P. Rasheed Abdul Haq and V. P. Harigovindan., "Water Quality Prediction for Smart Aquaculture Using Hybrid Deep Learning Models" Volume 10, 2022 : 60078- 60098.

[9] L. Li, P. Jiang, H. Guang, L. Dong, G. Wu, and H. Wu., "Water quality prediction based on recurrent neural network and improved evidence theory: a case study of Qiantang River, China," vol. 26, no. 19, pp. 19879-19896, Mar. 2022

[10] A. Sankar and P. Panday, "River Water Quality Modelling Using Artificial Neural Network Technique," Aquatic Procedia, vol. 4, pp. 1070-1077, 2022.