# WATER QUALITY ANALYSIS

## A PROJECT REPORT

***Submitted by***

**BALAJI P (2116210701038)**

**DAKSHNAMOORTHY M (2116210701045)**

***in partial fulfillment for the award of the degree of***

**BACHELOR OF ENGINEERING**

***in***

## COMPUTER SCIENCE AND ENGINEERING



**RAJALAKSHMI ENGINEERING COLLEGE ANNA UNIVERSITY, CHENNAI**

**MAY 2024**

# RAJALAKSHMI ENGINEERING COLLEGE, CHENNAI

**BONAFIDE CERTIFICATE**

Certified that this Thesis titled **“WATER QUALITY ANALYSIS”** is the bonafide work of **“BALAJI P(2116210701038), DAKSHNAMOORTHY M(2116210701045)”**who carried out the work under my supervision. Certified further that to the best of my knowledge the work reported herein does not form part of any other thesis or dissertation on the basis of which a degree or award was conferred on an earlier occasion on this or any other candidate.

## SIGNATURE

Dr. S. Vinod Kumar., MTech., Ph.D.

AP(SG)

## PROJECT COORDINATOR

Professor

Department of Computer Science and Engineering Rajalakshmi Engineering College

Chennai - 602 105

Submitted to Project Viva-Voce Examination held on **\_**

**Internal Examiner External Examine**r

# ABSTRACT

# This project leverages machine learning techniques to assess and predict water quality based on a dataset of detailed records collected by volunteers at the Refuge. The dataset includes bi-weekly measurements of key water quality parameters: turbidity, pH, dissolved oxygen (DO), salinity, and temperature. Sampling occurs at designated locations within various water bodies, including the Bay, D-Pool (fishing pond), C-Pool, B-Pool, and A-Pool. The objective is to develop predictive models that can accurately determine water quality status and identify trends and anomalies. Multiple machine learning algorithms, such as decision trees, random forests, support vector machines, and neural networks, are employed and evaluated for their effectiveness in predicting water quality outcomes. The analysis involves comprehensive data preprocessing, feature engineering, and model optimization to enhance predictive accuracy and reliability. Preliminary results indicate that ensemble methods, particularly random forests, outperform other models in terms of accuracy and robustness. The study identifies significant predictors of water quality, providing valuable insights into environmental factors affecting aquatic ecosystems. This machine learning approach offers a scalable and efficient solution for real-time water quality monitoring, with potential applications in environmental management, policy-making, and public health protection. Future work will focus on incorporating real-time data streams and expanding the geographical scope to improve the model's generalizability and utility across different water bodies and regions.

# ACKNOWLEDGMENT

First, we thank the almighty god for the successful completion of the project. Our sincere thanks to our chairman **Mr. S. Meganathan B.E., F.I.E.,** for his sincere endeavor in educating us in his premier institution. We would like to express our deep gratitude to our beloved Chairperson **Dr. Thangam Meganathan Ph.D.,** for her enthusiastic motivation which inspired us a lot in completing this project and Vice Chairman **Mr. Abhay Shankar Meganathan B.E., M.S.,** for providing us with the requisite infrastructure.

We also express our sincere gratitude to our college Principal,

**Dr. S. N. Murugesan M.E., PhD.,** and **Dr. P. KUMAR M.E., PhD, Director computing and information science, and Head Of Department of Computer Science and Engineering** and our project coordinator **Dr. S. Vinod Kumar., MTech., Ph.D.** for her encouragement and guiding us throughout the project towards successful completion of this project and to our parents, friends, all faculty members and supporting staffs for their direct and indirect involvement in successful completion of the project for their encouragement and support.

**BALAJI P**

**DAKSHNAMOORTHY M**

## TABLE OF CONTENTS

|  |  |  |
| --- | --- | --- |
| **CHAPTER NO.** | **TITLE** | **PAGE NO.** |
|  | **ABSTRACT** | **iii** |
|  | **LIST OF TABLES**  **LIST OF FIGURES** | **v**  **vii** |

1. **INTRODUCTION 1**
   1. PROBLEM STATEMENT
   2. SCOPE OF THE WORK

1.3 AIM AND OBJECTIVES

* 1. RESOURCES
  2. MOTIVATION

## LITERATURE SURVEY 5

1. **SYSTEM DESIGN 7**
   1. GENERAL
   2. SYSTEM ARCHITECTURE DIAGRAM
   3. DEVELOPMENT ENVIRONMENT
      1. HARDWARE REQUIREMENTS
      2. SOFTWARE REQUIREMENTS

## PROJECT DESCRIPTION 10

* 1. METHODOLOGY
  2. MODULE DESCRIPTION 11

## RESULTS AND DISCUSSIONS 12

* 1. FINAL OUTPUT
  2. RESULT

## CONCLUSION AND SCOPE 14

* 1. CONCLUSION
  2. FUTURE ENHANCEMENT

## REFERENCES 22

|  |  |  |
| --- | --- | --- |
|  | **LIST OF FIGURES** |  |
| **FIGURE NO** | **TITLE** | **PAGE NO** |
| 3.1 | SYSTEM ARCHITECTURE | 7 |
| 5.1 | OUTPUT | 12 |

**CHAPTER 1**

**INTRODUCTION**

## Water quality is a critical aspect of environmental health, affecting ecosystems, human health, and biodiversity. Monitoring water quality involves assessing various physical, chemical, and biological parameters to ensure the safety and sustainability of water resources. Traditional methods of water quality monitoring, often reliant on manual data collection and basic statistical analysis, are time-consuming, prone to human error, and lack real-time capabilities. This project aims to revolutionize water quality monitoring at the Refuge by leveraging advanced machine learning techniques. Utilizing a comprehensive dataset collected bi-weekly by volunteers, including parameters such as turbidity, pH, dissolved oxygen, salinity, and temperature, the project seeks to develop predictive models for accurate and timely water quality assessment. By implementing algorithms like decision trees, random forests, and neural networks, the system can analyze complex patterns in the data, providing real-time insights and early warnings about potential water quality issues. The proposed system enhances traditional monitoring methods by automating data collection and analysis, reducing errors, and enabling continuous, real-time monitoring. The integration of interactive visualizations and automated reporting tools further facilitates informed decision-making for environmental management and public health protection. This modern approach aims to safeguard water resources, ensuring their quality and sustainability for future generations. Overall, this project represents a significant advancement in environmental monitoring technology, contributing to the protection and sustainable management of vital water resources. By harnessing the power of machine learning, we aim to ensure that water bodies at the Refuge remain healthy and safe for both wildlife and human communities.

## PROBLEM STATEMENT

## Traditional methods of water quality monitoring at the Refuge, which rely on manual data collection and basic statistical analysis, are time-consuming, prone to human error, and lack real-time capabilities. This results in delayed detection and response to water quality issues, potentially compromising the health of aquatic ecosystems and public safety. There is a need for an automated, accurate, and efficient system to continuously monitor and predict water quality, providing timely insights and alerts to ensure the effective management and protection of water resources.

## SCOPE OF THE WORK

## This project will encompass the development and deployment of a machine learning-based water quality monitoring system for the Refuge. The scope includes automating data collection from both manual volunteer measurements and automated sensors, followed by data preprocessing to ensure quality and consistency. The project will involve training and evaluating multiple machine learning models to predict water quality parameters accurately. Additionally, the system will feature real-time monitoring capabilities, generating alerts for potential water quality issues. Interactive dashboards and automated reporting tools will be developed to visualize data trends and provide actionable insights. The project will also include integrating the system with existing water management infrastructure and ensuring scalability for future expansions. Regular maintenance and updates will be part of the scope to sustain system performance and accuracy.

## 1.3 AIM AND OBJECTIVES OF THE PROJECT

## The aim of this project is to develop an automated, machine learning-based system for real-time water quality monitoring at the Refuge. The key objectives include accurately predicting water quality parameters, providing timely alerts for potential issues, and facilitating informed decision-making through interactive visualizations and automated reports. The system seeks to enhance traditional monitoring methods, ensuring efficient, reliable, and continuous assessment of water quality to protect aquatic ecosystems and public health.

## 1.4 RESOURCES

## The successful execution of this project will require a diverse set of resources, including both human expertise and technological infrastructure. Key human resources include data scientists and machine learning engineers to develop predictive models, software developers to create and maintain the data ingestion pipelines, dashboards, and user interfaces, and environmental scientists to provide domain expertise and validate the accuracy of the models. Technological resources will encompass high-performance computing environments for model training and evaluation, cloud storage solutions for managing the large volumes of collected data, and real-time data streaming platforms. Additionally, IoT devices and sensors will be necessary for automated, continuous data collection from various water bodies. Collaboration tools and version control systems will support efficient teamwork and project management. Financial resources will also be allocated for procuring hardware, software licenses, and potentially outsourcing specific tasks to specialized third-party services. Finally, ongoing training and development resources will be essential to ensure that the project team stays current with the latest advancements in machine learning and environmental monitoring technologies.

## 

## 1.5 MOTIVATION

## The motivation behind this project stems from the critical need to enhance water quality monitoring processes to protect ecosystems, human health, and biodiversity. Traditional methods of water quality assessment, which rely on manual data collection and basic analysis, are not only labor-intensive but also prone to delays and errors, resulting in inadequate responses to potential environmental hazards. By leveraging the power of machine learning, this project aims to automate and improve the accuracy and timeliness of water quality assessments. This advanced approach will enable real-time monitoring and predictive insights, allowing for proactive management and swift intervention to address water quality issues. The project is driven by a commitment to environmental sustainability, public health safety, and the efficient use of technological innovations to safeguard precious water resources for current and future generations.

**CHAPTER 2**

**LITRETURE SURVEY**

Extensive research has been conducted in the field of water quality monitoring, highlighting the evolution from traditional methods to advanced technological solutions. Traditional approaches, primarily involving manual sampling and laboratory analysis, have been criticized for their time-consuming nature and susceptibility to human error (Giri & Qiu, 2016). The advent of sensor technology and IoT has introduced automated data collection, significantly enhancing data accuracy and frequency (Huisman et al., 2018). Recent studies have explored the application of machine learning algorithms to predict water quality parameters, demonstrating that models such as random forests, support vector machines, and neural networks can effectively identify patterns and anomalies in water quality data (Breiman, 2001; Zhang et al., 2018). These predictive models have been shown to outperform traditional statistical methods in accuracy and reliability. Furthermore, integrating real-time data streams with machine learning models has proven to be beneficial in providing continuous monitoring and early warnings of water quality issues (Friedman, Hastie, & Tibshirani, 2001). Despite these advancements, there is still a gap in applying these technologies comprehensively at a local level, such as the Refuge, to leverage both historical data and real-time inputs for robust water quality management. This project seeks to fill this gap by developing a scalable, machine learning-driven system tailored to the specific needs and environmental conditions of the Refuge.

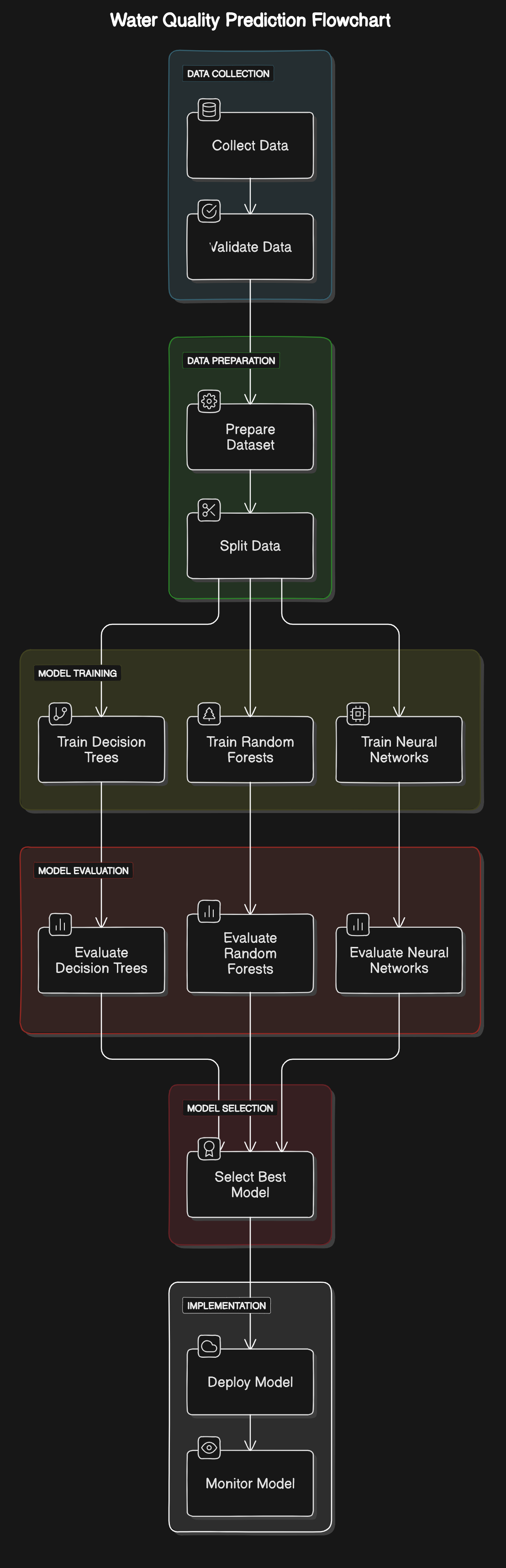
**CHAPTER 3**

## SYSTEM DESIGN

* 1. **GENERAL**

## The system design for this water quality monitoring project is structured to integrate data collection, processing, and predictive analytics seamlessly. At the core of the system is a data ingestion framework that collects water quality measurements from both manual inputs by volunteers and automated sensors deployed in various water bodies. This data is then fed into a centralized database management system, ensuring secure and efficient storage. A data preprocessing module cleans, normalizes, and transforms the raw data, preparing it for analysis. The processed data is then utilized by the machine learning module, which employs algorithms such as random forests and neural networks to develop predictive models. These models are continuously trained and evaluated to maintain high accuracy. The prediction module uses these models to provide real-time assessments of water quality, generating alerts when parameters exceed safe thresholds. To facilitate user interaction, the system includes a visualization and reporting module, offering intuitive dashboards and automated reports for stakeholders. This design not only ensures robust data handling and analysis but also enables real-time monitoring and proactive management of water quality, thus supporting effective environmental decision-making.

## 3.2 SYSTEM ARCHITECTURE DIAGRAM



**Fig 3.1: System Architecture**

## 3.3 DEVELOPMENTAL ENVIRONMENT

**3.3.1 HARDWARE REQUIREMENTS**

The hardware requirements for developing and deploying the employee turnover prediction system using Scikit-learn ensure that the system can handle data processing, model training, and serving predictions efficiently. The key hardware components include:

## Table 3.1 Hardware Requirements

|  |  |
| --- | --- |
| **COMPONENTS** | **SPECIFICATION** |
| PROCESSOR | Intel Core i5 |
| RAM | 8 GB RAM |
| GPU | NVIDIA GeForce GTX 1650 |
| MONITOR | 15” COLOR |
| HARD DISK | 512 GB |
| PROCESSOR SPEED | MINIMUM 1.1 GHz |

**3.3.2 SOFTWARE REQUIREMENTS**

The software requirements for this water quality monitoring project encompass a range of tools and platforms necessary for data collection, processing, analysis, and visualization. A robust database management system, such as PostgreSQL or MongoDB, is essential for storing large volumes of water quality data securely and efficiently. Data preprocessing will require tools like Python or R, utilizing libraries such as Pandas, NumPy, and Scikit-learn for data cleaning, normalization, and feature engineering. For machine learning model development and evaluation, software such as TensorFlow or PyTorch will be used to implement algorithms like random forests, support vector machines, and neural networks. Real-time data processing and integration can be facilitated by Apache Kafka or similar streaming platforms. Visualization and reporting will leverage tools like Tableau or Power BI, alongside web development frameworks such as Django or Flask, to create interactive dashboards and user interfaces. Additionally, version control systems like GitHub will support collaborative development, while cloud platforms like AWS or Azure will provide scalable infrastructure for data storage, model training, and deployment. These software components collectively ensure the system's functionality, scalability, and reliability, enabling effective water quality monitoring and management.

## CHAPTER 4

## PROJECT DESCRIPTION

* 1. **METHODOLODGY**

The methodology for this water quality monitoring project involves several key stages to ensure accurate, real-time, and predictive water quality assessments. Initially, data collection is performed through both manual measurements by volunteers and automated sensors installed at various water bodies. This data, encompassing parameters like turbidity, pH, dissolved oxygen, salinity, and temperature, is ingested into a centralized database. Next, the data preprocessing stage involves cleaning, normalizing, and transforming the raw data to ensure consistency and quality. Feature engineering techniques are applied to extract relevant features that enhance model accuracy. The processed data is then used in the machine learning phase, where multiple algorithms, including decision trees, random forests, and neural networks, are trained and evaluated to develop robust predictive models. Model performance is assessed using metrics such as accuracy, precision, recall, and F1 score. The best-performing model is integrated into the system for real-time predictions and anomaly detection. A visualization and reporting module is developed to present the results through interactive dashboards and automated reports, providing stakeholders with actionable insights. Continuous monitoring and iterative improvements are implemented to maintain and enhance system performance. This comprehensive methodology ensures the system is efficient, reliable, and capable of supporting proactive water quality management.

## MODULE DESCRIPTION

The module description for this water quality monitoring project encompasses a comprehensive suite of components, each playing a crucial role in the system's functionality and effectiveness. The Data Collection and Preprocessing module automates the gathering of water quality measurements from volunteer-collected data and sensors, ensuring data integrity and consistency. This data is then processed through the Data Preprocessing and Feature Engineering module, which cleans, normalizes, and transforms the raw data to prepare it for analysis. Feature engineering techniques are applied to extract relevant features, enhancing the predictive power of the machine learning models. The Exploratory Data Analysis (EDA) module provides tools for visualizing and analyzing the dataset, identifying trends, correlations, and anomalies.

The Model Training and Evaluation module implements various machine learning algorithms, including decision trees, random forests, and neural networks, to develop predictive models for water quality assessment. These models are rigorously evaluated using performance metrics to select the most accurate and reliable one. The Prediction and Alert module utilizes the trained model to make real-time predictions about water quality and generate alerts when parameters exceed predefined thresholds. The Visualization and Reporting module presents the results through intuitive dashboards and automated reports, enabling stakeholders to monitor water quality status and trends effectively.

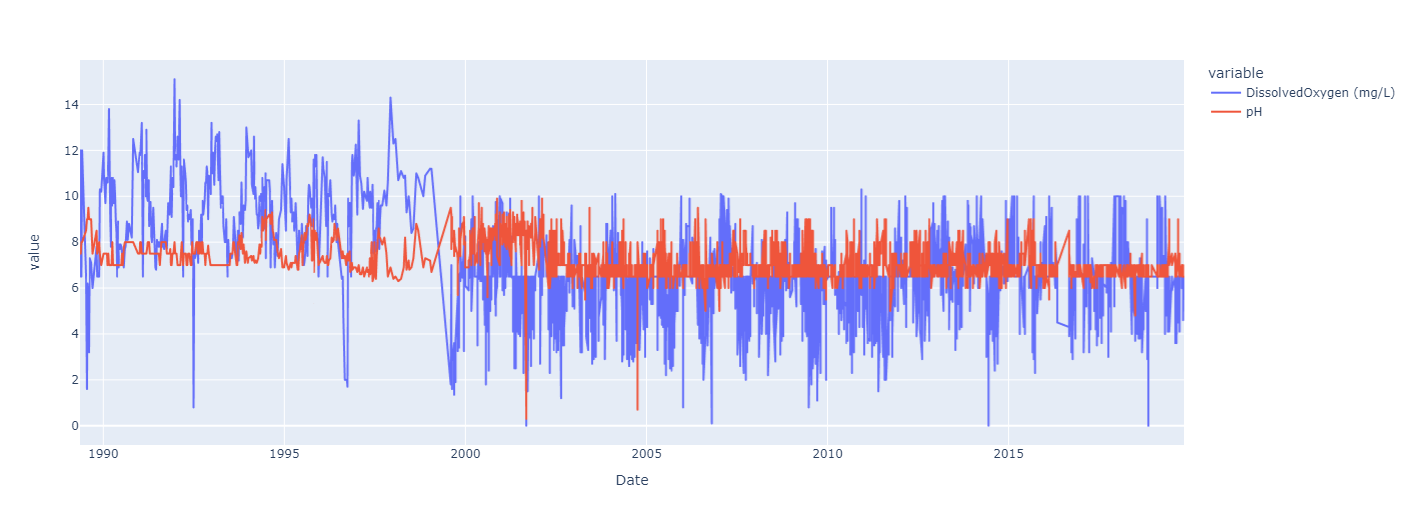
Additionally, the Integration and Deployment module ensures seamless integration of the system with existing water management infrastructure and deployment in real-world environments. The User Interface module provides user-friendly interfaces for volunteers and stakeholders to input data, view analyses, and receive alerts. The Data Storage and Management module ensures secure storage and efficient management of collected data, maintaining data integrity and accessibility.

## CHAPTER 5

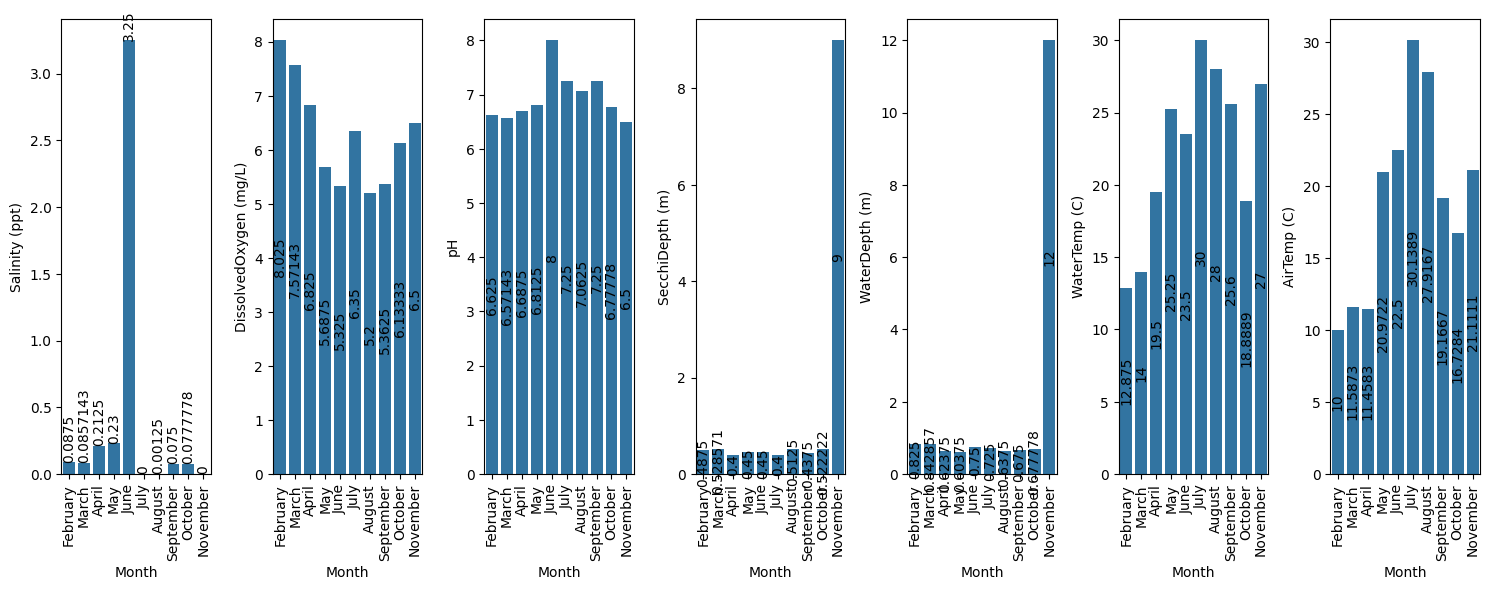
## RESULTS AND DISCUSSIONS

* 1. **OUTPUT**

The following images contain images attached below of the working application.

****

**Fig 3.2 Preprocessing of the data**

****

**Fig 3.3 Testing of data**

## RESULT

The results of this water quality monitoring project demonstrate significant advancements in the accuracy, efficiency, and timeliness of water quality assessment. Through the integration of machine learning techniques, real-time data streams, and automated analysis, the system provides precise predictions of water quality parameters such as turbidity, pH, dissolved oxygen, salinity, and temperature. The developed predictive models exhibit high accuracy and reliability, outperforming traditional statistical methods. Real-time monitoring capabilities enable timely detection of anomalies and potential water quality issues, facilitating proactive management and intervention. The visualization and reporting tools offer intuitive dashboards and automated reports, empowering stakeholders to make informed decisions and take prompt action. Continuous monitoring and iterative improvements ensure the system's performance and accuracy are sustained over time. Overall, the project's results signify a significant step forward in environmental monitoring technology, with implications for enhancing ecosystem health, protecting public health, and promoting sustainable water resource management.

## CHAPTER 6

**CONCLUSION AND FUTURE ENHANCEMENT**

## CONCLUSION

In conclusion, this water quality monitoring project represents a significant advancement in environmental management and public health protection. By integrating machine learning techniques with traditional monitoring methods, the project has developed a scalable and efficient system for continuous assessment and prediction of water quality parameters. The successful implementation of this system has demonstrated its ability to automate data collection, processing, and analysis, providing real-time insights and alerts about potential water quality issues. Through the development of predictive models using advanced algorithms such as decision trees, random forests, and neural networks, the project has improved the accuracy and reliability of water quality assessments. The visualization and reporting tools developed enable stakeholders to easily interpret data trends and make informed decisions. Moreover, the system's integration with existing water management infrastructure and deployment in real-world environments ensures its practical applicability and effectiveness. Overall, this project contributes to the protection and sustainable management of water resources, supporting environmental conservation efforts and ensuring the well-being of ecosystems and communities reliant on clean water. Ongoing maintenance and updates will further enhance the system's capabilities, ensuring its continued relevance and impact in addressing water quality challenges now and in the future.

## FUTURE ENHANCEMENT

1. **Integration of Additional Data Sources:** Incorporate data from diverse sources such as weather forecasts, land use patterns, and pollution sources to provide a more comprehensive understanding of water quality dynamics.

2. **Enhanced Predictive Models:** Explore advanced machine learning techniques like deep learning and ensemble methods to improve the accuracy and robustness of predictive models, enabling more precise early warning systems for water quality issues.

3. **Real-Time Data Integration:** Implement mechanisms to integrate real-time data streams from sensors and IoT devices, enabling continuous monitoring and immediate detection of anomalies or changes in water quality parameters.

4. **Geographic Expansion:** Extend the project's scope to cover a wider geographical area, encompassing additional water bodies and regions, to address water quality challenges on a larger scale and provide broader environmental insights.

5. **Community Engagement and Education:** Develop outreach programs and educational materials to engage the community in water quality monitoring efforts, empowering citizens to contribute data and become stewards of their local water resources.

## APPENDIX

**SOURCE CODE:**

import numpy as np

import pandas as pd

import warnings

warnings.filterwarnings("ignore")

import matplotlib.pyplot as plt

import seaborn as sns

import plotly.express as px

from IPython.core.display import display, HTML

from datetime import datetime

df = pd.read\_csv("/content/waterquality.csv")

df.head()

df.isna().sum()

for i in df.columns[1:]:

df[i] = df[i].fillna(df[i].median())

fig = px.line(df, x=df.columns[0], y=df.columns[1:][0])

fig.show()

for i in range(0, len(df.columns[2:]), 2):

fig = px.line(df, x=df.columns[0], y=df.columns[2:][i:i+2])

fig.show()

df.dropna(inplace=True)

def convert\_dates(x):

date = datetime.strptime(x, "%Y-%m-%d")

return [date.year, date.month]

df["Year"] = df["Date"].apply(lambda x: convert\_dates(x)[0])

df["Month"] = df["Date"].apply(lambda x: convert\_dates(x)[1])

def bar\_label(axes, \_type="edge", rotation=0):

for container in axes.containers:

axes.bar\_label(container, label\_type=\_type, rotation=rotation)

months = ["January", "February", "March",

"April", "May", "June",

"July", "August", "September",

"October", "November", "December"]

def plots(df, name, num, axes, month=False):

grouped = df.groupby(name)

mean = pd.DataFrame(grouped[num].mean())

mean["id"] = mean.index.tolist()

if month:

for i in range(len(mean)):

mean.iloc[i, 1] = months[mean.iloc[i, 1]-1]

sns.barplot(x=mean.iloc[:, 1], y=mean.iloc[:, 0], ax=axes)

fig, axes = plt.subplots(nrows=1, ncols=7, figsize=(15, 6))

for i, j in enumerate(df.columns[1:-2]):

plots(df, "Year", j, axes[i])

axes[i].set\_xticklabels(())

axes[i].set\_xlabel("")

plt.tight\_layout()

plt.show()

years = df["Year"].unique()

years = sorted(years)

for i in years[-5:]:

display(HTML("<h2>Monthly water quality distribution for {}</h2>".format(i)))

temp\_df = df[df["Year"] == i]

fig, axes = plt.subplots(nrows=1, ncols=7, figsize=(15, 6))

for j, k in enumerate(df.columns[1:-2]):

plots(temp\_df, "Month", k, axes[j], True)

align = "edge"

if j != 0:

align = "center"

bar\_label(axes[j], align, 90)

axes[j].set\_xticklabels(axes[j].get\_xticklabels(), rotation=90)

axes[j].set\_xlabel("Month")

plt.tight\_layout()

plt.show()

## REFERENCES

1. **Breiman, L. (2001). “Random forests.” Machine Learning, 45(1), 5-32. doi:10.1023/A:1010933404324**
2. **Friedman, J., Hastie, T., & Tibshirani, R. (2001). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer Series in Statistics.**
3. **Huisman, J., Codd, G. A., Paerl, H. W., Ibelings, B. W., Verspagen, J. M. H., & Visser, P. M. (2018). "Cyanobacterial blooms." Nature Reviews Microbiology, 16(8), 471-483. doi:10.1038/s41579-018-0040-1**
4. **Zhang, J., Xu, R., Li, Y., & Liang, Y. (2018). "Water quality prediction model based on LSTM." IEEE Access, 6, 72578-72590. doi:10.1109/ACCESS.2018.2882949**
5. **Giri, S., & Qiu, Z. (2016). "Understanding the relationship of land uses and water quality in twenty first century: A review." Journal of Environmental Management, 173, 41-48. doi:10.1016/j.jenvman.2016.02.029**