

PROFESSIONAL TRAINING REPORT - II

entitled

PREDICTING AND ANALYSING URBAN WATER QUALITY USING MACHINE LEARNING

Submitted in partial fulfillment of the requirements for the award of
Bachelor of Engineering degree in Computer Science and Engineering with
specialization in Cyber Security

by

GARAPATI RAM CHOWDARY [41614025]

GUNTAKA VENKATA SAI CHENNA REDDY [41614029]

ARIGONDA HARI KUMAR [41614007]



**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
SCHOOL OF COMPUTING**

SATHYABAMA

**INSTITUTE OF SCIENCE AND TECHNOLOGY
(DEEMED TO BE UNIVERSITY)**

CATEGORY -1 UNIVERSITY BY UGC

**Accredited with Grade "A++" by NAAC | 12B Status by UGC | Approved by AICTE
JEPPIAAR NAGAR, RAJIV GANDHI SALAI, CHENNAI - 600 119**

MAY 2024



SATHYABAMA

**INSTITUTE OF SCIENCE AND TECHNOLOGY
(DEEMED TO BE UNIVERSITY)**

CATEGORY-I UNIVERSITY BY UGC

Accredited "A++" Grade by NAAC | 12B Status by UGC | Approved by AICTE

www.sathyabama.ac.in

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

BONAFIDE CERTIFICATE

This is to certify that this Professional Training Report is the bonafide work of **Mr. GARAPATI RAM CHOWDARY (41614025)** who carried out the project entitled **"PREDICTING AND ANALYSING URBAN WATER QUALITY USING MACHINE LEARNING"** under my supervision from January 2024 to May 2024.

Internal Guide

Dr. B.U. ANU BARATHI, M.E., Ph.D.,

Head of the Department

Dr. S. VIGNESHWARI, M.E., Ph.D.,

Submitted for Viva voce Examination held on _____

Internal Examiner

External Examiner

DECLARATION

I, **GARAPATI RAM CHOWDARY (41614025)**, hereby declare that the Professional Training Report-II entitled **“PREDICTING AND ANALYSING URBAN WATER QUALITY USING MACHINE LEARNING”** done by me under the guidance of **Dr. B. U. ANU BARATHI, M.E., Ph.D.**, is submitted in partial fulfilment of the requirements for the award of Bachelor of Engineering degree in Computer Science and Engineering with specialization in Cyber Security.

DATE:

PLACE:

SIGNATURE OF THE CANDIDATE

ACKNOWLEDGEMENT

I am pleased to acknowledge my sincere thanks to **Board of Management of SATHYABAMA** for their kind encouragement in doing this project and for completing it successfully. I am grateful to them.

I convey my thanks to **Dr. T.Sasikala M.E., Ph.D., Dean, School of Computing, Dr. S.Vigneshwari M.E., Ph.D., Head of the Department of Computer Science and Engineering** for providing me necessary support during the progressive reviews.

I would like to express my sincere and deep sense of gratitude to my Internal Guide **Dr. B.U. Anu Barathi, M.E., Ph.D.,** for her valuable guidance, suggestions and constant encouragement which paved way for the successful completion of my phase-2 Professional Training.

I wish to express my thanks to all Teaching and Non-teaching staff members of the **Department of Computer Science and Engineering** who were helpful in many ways for the completion of the project.

COURSE CERTIFICATE



ABSTRACT

The deteriorating quality of natural water resources like lakes, streams and estuaries, is one of the direst and most worrisome issues faced by humanity. The effects of unclean water are far-reaching, impacting every aspect of life. Therefore, management of water resources is very crucial in order to optimize the quality of water. The effects of water contamination can be tackled efficiently if data is analysed and water quality is predicted beforehand. This issue has been addressed in many previous researches however, more work needs to be done in terms of effectiveness, reliability, accuracy as well as usability of the current water quality management methodologies. The goal of this study is to develop a water quality prediction model with the help of water quality factors using Artificial Neural Network (ANN) and time-series analysis. This research uses the water quality historical data of the year of 2014, with 6-minutes time interval. Data is obtained from the United States Geological Survey (USGS) online resource called National Water Information System (NWIS). For this paper, the data includes the measurements of 4 parameters which affect and influence water quality. For the purpose of evaluating the performance of model, the performance evaluation measures used are Mean Squared Error (MSE), Root Mean-Squared Error (RMSE) and Regression Analysis. Previous works about Water Quality prediction have also been analysed and future improvements have been proposed in this paper.

TABLE OF CONTENTS

CHAPTER NO.	TITLE		PAGE NO.
	ABSTRACT		vi
	LIST OF FIGURES		viii
1	INTRODUCTION 1.1 Overview		1
2	LITERATURE SURVEY 2.1 survey		2
3	REQUIREMENTS ANALYSIS		11
	3.1	Objective	11
	3.2	3.2.1 Hardware Requirements 3.2.2 Software Requirements	12
4	DESIGN DESCRIPTION OF PROPOSED PRODUCT		13
	4.1	4.1.1 Ideation Map/Architecture Diagram 4.1.2 Various stages 4.1.3 Internal or Component design structure 4.1.4 working principles	17 17 19 19
	4.2	4.2.1 Novelty of the Project	23
5	RESULTS AND DISCUSSION		25
6	SUMMARY AND CONCLUSIONS		29
	REFERENCES		30
	APPENDIX A. Research Paper B. Source Code		31 41

LIST OF FIGURES

Figure No.	Figure Name	Page No.
3.1	Random Forest Classifier Example	11
4.1.4	Decision Tree example	20
4.1.4	Random Forest Classifier	22
5.1.1	Heat Map	25
5.1.2	Box Plot	25
5.1.3	3D Scatter Plot	26
5.1.4	Scatter Plot	26
5.1.5	Count Plot	26
5.1.6	Spitter Of Data Set	27
5.1.7	Random Forest Regression	27
5.1.8	Dist Plot	28
5.1.9	Accuracy And Saving The Model	28

CHAPTER 1

INTRODUCTION

1.1 OVERVIEW

The deteriorating quality of natural water resources like lakes, streams and estuaries, is one of the direst and most worrisome issues faced by humanity. The effects of unclean water are far-reaching, impacting every aspect of life. Therefore, management of water resources is very crucial in order to optimize the quality of water. The effects of water contamination can be tackled efficiently if data is analyzed and water quality is predicted beforehand. This issue has been addressed in many previous researches, however, more work needs to be done in terms of effectiveness, reliability, accuracy as well as usability of the current water quality management methodologies. The goal of this study is to develop a water quality prediction model with the help of water quality factors using Artificial Neural Network (ANN) and time-series analysis. This research uses the water quality historical data of the year of 2014, with 6-minutes time interval. Data is obtained from the United States Geological Survey (USGS) online resource called National Water Information System (NWIS). For this paper, the data includes the measurements of 4 parameters which affect and influence water quality. For the purpose of evaluating the performance of model, the performance evaluation measures used are MeanSquared Error (MSE), Root Mean-Squared Error (RMSE) and Regression Analysis. Previous works about Water Quality prediction have also been analyzed and future improvements have been proposed in this paper.

CHAPTER 2

LITERATURE REVIEW

2.1 SURVEY

1. Ping Liu Jin Wang Arun Kumar Sangaiah Yang Xie Xinchun Yin, 2019. **Analysis and Prediction of Water Quality Using LSTM Deep Neural Networks in IoT Environment, Sustainability, MDPI, Open Access Journal, vol. 11(7), pages 1- 14, April.**

Advantages:

1. **Temporal Modeling:** LSTM networks excel in capturing temporal dependencies in data, making them suitable for time-series analysis such as water quality prediction.
2. **Handling Long-Term Dependencies:** LSTMs are designed to overcome the vanishing gradient problem, allowing them to retain information over long sequences, which is crucial for accurately predicting water quality fluctuations.
3. **Feature Extraction:** LSTMs can automatically extract relevant features from raw sensor data, reducing the need for manual feature engineering.
4. **Flexibility:** LSTMs can handle various types of data, including multivariate time series, making them adaptable to different IoT environments with diverse sensor inputs.

Disadvantages:

- 1.**Complexity:** Training LSTM networks requires substantial computational resources and time due to their complex architecture and large number of parameters.
- 2.**Data Dependency:** LSTMs heavily rely on the availability of historical data for accurate prediction, which might be limited or noisy in certain IoT environments.
- 3.**Overfitting:** LSTM models are prone to overfitting, especially when trained on small datasets or noisy data, leading to poor generalization performance.

4.Interpretability: The black-box nature of LSTM networks makes it challenging to interpret the underlying mechanisms driving water quality predictions, which can be crucial for stakeholders seeking insights into the system.

These points highlight both the strengths and limitations of using LSTM deep neural networks for water quality analysis and prediction in IoT environments.

2. Manish Kumar Jha ; Rajni Kumari Sah ; M. S. Rashmitha ; Rupam Sinha ; B. Sujatha ; K. V. Suma Smart Water Monitoring System for Real-Time Water Quality and Usage Monitoring||,IEEE Conference paper, July 2018.

Advantages:

1.Real-Time Monitoring: These systems offer real-time data on water quality and usage, allowing for immediate response to any changes or issues, leading to improved safety and efficiency.

2. Early Detection of Contaminants: With real-time monitoring, these systems can quickly detect changes in water quality, such as contamination, allowing for prompt action to prevent public health risks.

3. Resource Management: Smart water monitoring systems can help manage water resources efficiently by tracking usage patterns, reducing waste, and promoting sustainability.

4. Automation and Remote Access: The systems often include automated functions and remote monitoring capabilities, allowing users to monitor water conditions from a distance, reducing the need for manual checks.

5.Data Analytics and Trends: By collecting data over time, these systems can provide valuable insights into trends, helping to make informed decisions about water management and conservation efforts.

6.Compliance and Regulation: Real-time monitoring can assist in ensuring compliance with environmental regulations, providing a record of water quality and usage to meet legal requirements.

Disadvantages:

- 1. Cost and Complexity:** Implementing a smart water monitoring system can be costly due to the technology, sensors, and infrastructure needed. The system's complexity can also require specialized skills for setup and maintenance.
- 2. Reliability and Accuracy:** The accuracy of sensors and data reliability can be a concern, particularly in harsh environmental conditions or with complex water systems. Calibration and maintenance are crucial.
- 3. Data Security and Privacy:** As these systems are connected to networks, there is a risk of data breaches or unauthorized access. Ensuring data security and privacy is a significant challenge.
- 4. Limited Accessibility:** In regions with limited access to technology or infrastructure, implementing smart water monitoring systems can be challenging, potentially limiting their benefits to certain areas.
- 5. Technical Support and Maintenance:** The need for ongoing technical support and maintenance can be a disadvantage, especially in remote or resource-limited locations where expertise may not be readily available.
- 6. Integration Challenges:** Integrating with existing systems or infrastructure may require significant effort and adaptation, potentially leading to additional costs and complexities.

- 3. K. Gopavanitha and S. Nagaraju A low cost system for real time water quality monitoring and controlling using IoT||, IEEE Conference paper, August 2017.**

Advantages:

- 1. Affordability:** Low-cost systems are more accessible, making them suitable for deployment in resource-constrained areas where budget is a concern.
- 2. Scalability:** Due to their low cost, these systems can be easily scaled up for widespread deployment across multiple locations, enhancing monitoring coverage.
- 3. Real-Time Monitoring:** IoT-enabled systems provide real-time data on water quality parameters, allowing for prompt detection and response to any deviations from desired levels.

4. **Remote Accessibility:** Users can remotely access and monitor water quality data through IoT platforms, facilitating timely decision-making and intervention.

Disadvantages:

1. **Accuracy and Reliability:** Low-cost sensors may not offer the same level of accuracy and reliability as more expensive counterparts, potentially leading to less precise measurements and monitoring.
2. **Limited Functionality:** Low-cost systems may have limited features and capabilities compared to higher-end solutions, restricting the range of parameters that can be monitored and controlled.
3. **Maintenance and Calibration:** These systems may require more frequent maintenance and calibration to ensure accurate and reliable operation, which can incur additional costs and effort.
4. **Data Security Concerns:** IoT devices are susceptible to cybersecurity threats, and low-cost systems may lack robust security measures, posing risks to the integrity and confidentiality of water quality data.

These points highlight the trade-offs associated with using a low-cost system for real-time water quality monitoring and controlling using IoT, emphasizing the balance between affordability and performance.

4. Vaishnavi V. Daigavane, Dr. M.A Gaikwad. ||Water Quality Monitoring System Based on IOT|| 2017 Advances in Wireless and Mobile Communications, Nov 2017 ISSN 0973-6972

The paper titled "Water Quality Monitoring System Based on IoT" by Vaishnavi V. Daigavane and Dr. M.A. Gaikwad describes an Internet of Things (IoT)-based approach to monitor water quality. Let's discuss the potential advantages and disadvantages of this kind of system.

Advantages:

1.Remote Monitoring: IoT-based systems enable remote monitoring of water quality, allowing users to track conditions from anywhere with internet access. This capability is particularly useful for distributed or hard-to-reach water sources.

2.Real-Time Data: These systems provide real-time updates on water quality, allowing for quick responses to any detected issues, enhancing safety and reducing risks of contamination.

3. Automation and Alerts: IoT-based systems can be automated to send alerts or notifications when certain parameters exceed predefined thresholds, facilitating proactive management.

4.Scalability: IoT systems are inherently scalable, allowing for easy expansion or integration with additional sensors or monitoring points as needs grow.

5.Data Analytics and Visualization: These systems often come with data analytics capabilities, allowing users to analyze trends and visualize data, leading to better decision-making and resource management.

6.Reduced Manual Labor: With automated data collection, these systems reduce the need for manual sampling and analysis, saving time and reducing human error.

Disadvantages:

1.Security and Privacy Risks: IoT-based systems are connected to the internet, making them vulnerable to cyber-attacks and data breaches. Ensuring security and privacy is a significant concern.

2.Infrastructure Requirements: Implementing IoT-based systems requires a reliable internet connection and supporting infrastructure, which may not be readily available in all regions or rural areas.

3.Technical Complexity: The use of IoT technology requires specialized knowledge for setup, configuration, and maintenance, which can be a barrier in areas with limited technical expertise.

4.Cost: While IoT systems can be scalable, the initial setup costs, including sensors, connectivity, and data management platforms, can be high.

5.Reliability and Maintenance: IoT sensors can be sensitive to environmental conditions, requiring regular maintenance and calibration to ensure data accuracy and reliability.

6.Integration Challenges: Integrating with existing systems or legacy infrastructure can be complex and may require additional resources or customization.

In summary, IoT-based water quality monitoring systems offer significant benefits in terms of real-time monitoring, scalability, and automation. However, they also present challenges related to security, technical complexity, and infrastructure requirements.

5. Manish Kumar Jha ; Rajni Kumari Sah ; M. S. Rashmitha ; Rupam Sinha ; B. Sujatha ; K. V. Suma Smart Water Monitoring System for Real-Time Water Quality and Usage Monitoring||,IEEE Conference paper,july 2018.

The IEEE conference paper titled "Smart Water Monitoring System for Real-Time Water Quality and Usage Monitoring" by Manish Kumar Jha, Rajni Kumari Sah, M. S. Rashmitha, Rupam Sinha, B. Sujatha, and K. V. Suma, likely discusses a system designed for real-time monitoring of water quality and water usage. Without direct access to the specific content of the paper, I can provide general advantages and disadvantages associated with smart water monitoring systems that focus on real-time water quality and usage monitoring.

Advantages:

1. Real-Time Monitoring: The system provides instant data on water quality and usage, enabling immediate response to any anomalies or contamination.

2. Enhanced Safety and Quality Control: Early detection of water contamination or quality degradation helps to prevent health risks and ensures compliance with safety standards.

3.Efficiency in Water Resource Management: By monitoring water usage in real-time, these systems can help identify wasteful practices, optimize water distribution, and promote conservation.

4.Automation and Remote Access:The system can automate data collection and processing, providing remote access to information, which reduces the need for manual checks and physical presence.

5.Data Analytics and Decision-Making: The collected data can be analyzed to identify trends, aiding in informed decision-making regarding water quality management and resource allocation.

6.Regulatory Compliance: Real-time monitoring systems can facilitate compliance with environmental and safety regulations by maintaining accurate records of water quality and usage.

Disadvantages:

1.Cost and Complexity: Implementing such a system can be costly, requiring investment in sensors, infrastructure, and data processing technology.

2.Technical Challenges: The complexity of installing and maintaining smart water monitoring systems may require specialized expertise and ongoing technical support.

3.Data Security and Privacy: Since these systems often use wireless communication and cloud-based storage, they are vulnerable to cybersecurity risks, requiring robust security measures.

4.Reliability and Maintenance: Environmental factors can affect sensor reliability, necessitating regular calibration and maintenance to ensure accurate data.

5.Limited Accessibility: In regions with limited infrastructure or connectivity, implementation of such systems may be challenging or impractical.

6.Integration with Legacy Systems: Integrating smart water monitoring systems with existing or older infrastructure may require additional effort and resources, presenting a potential barrier to adoption.

These advantages and disadvantages cover the general aspects of smart water monitoring systems for real-time quality and usage monitoring. Specific points in the paper might vary based on the scope of the study, the technology used, and the context of implementation.

6. Najah Ahmed, A.; Binti Othman, F.; Abdulmohsin Afan, H.; Khaleel Ibrahim, R.; Ming Fai, C.; Shabbir Hossain, M.; Ehteram, M.; Elshafie, A. Machine learning methods for better water quality prediction. J. Hydrol. 2019, 578, 124084

The paper "Machine learning methods for better water quality prediction" by Najah Ahmed, A., et al., published in the Journal of Hydrology in 2019, discusses the use of machine learning (ML) techniques to improve predictions related to water quality. Although I don't have access to the specific content of the paper, I can outline the general advantages and disadvantages of applying machine learning to water quality prediction based on industry trends and common practices.

Advantages:

1.Improved Prediction Accuracy: Machine learning algorithms can analyze large datasets and identify complex patterns, leading to more accurate water quality predictions compared to traditional methods.

2.Data-Driven Insights: ML can derive insights from data that might not be readily apparent, helping to understand factors influencing water quality and detect correlations.

3.Automation and Efficiency: Once trained, ML models can automate the process of predicting water quality, reducing manual labor and enhancing efficiency in monitoring and management.

4.Adaptability and Scalability: ML models can adapt to changing conditions and scale with additional data, providing flexibility as monitoring requirements grow.

5.Early Detection of Anomalies: ML algorithms can detect anomalies or trends that could indicate potential contamination or other issues, allowing for early intervention.

6.Reduced Human Error: Automation through ML reduces the chances of human error in analyzing data and making predictions.

Disadvantages:

1.Data Quality and Availability: Machine learning models require large amounts of high-quality data for training. Inconsistent or incomplete data can negatively impact prediction accuracy.

2. Technical Complexity: Developing, training, and maintaining ML models can be complex and may require specialized expertise, potentially increasing implementation costs.

3.Overfitting and Generalization: ML models can overfit the training data, leading to poor generalization when applied to new data. This requires careful model design and validation.

4.Black Box Nature: Many ML methods, especially deep learning, operate as "black boxes," making it difficult to understand the rationale behind specific predictions. This can complicate regulatory compliance and decision-making.

5.Computational Resources: Training and running ML models can require significant computational resources, which might not be readily available in all environments.

6.Ethical and Security Concerns: The use of ML for water quality prediction involves handling sensitive data, raising concerns about data privacy and security. There is also a need to ensure ethical use of ML in public health and environmental monitoring.

These advantages and disadvantages highlight the potential benefits and challenges of applying machine learning to water quality prediction. The specifics can vary based on the chosen ML techniques, the context of application, and the quality of available data.

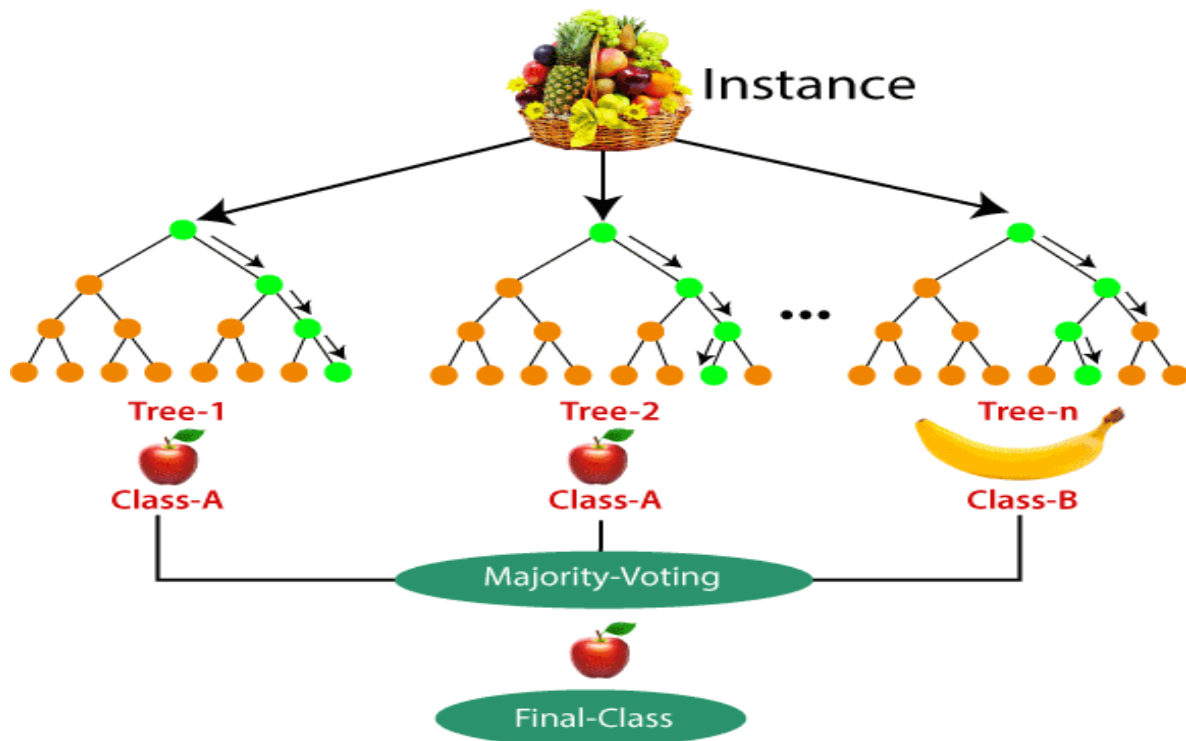
CHAPTER 3

REQUIREMENTS ANALYSIS

3.1 OBJECTIVE OF THE PROJECT

Let's take an example of a training dataset consisting of various fruits such as bananas, apples, pineapples, and mangoes. The random forest classifier divides this dataset into subsets. These subsets are given to every decision tree in the random forest system. Each decision tree produces its specific output. For example, the prediction for trees 1 and 2 is apple.

Another decision tree (n) has predicted banana as the outcome. The random forest classifier collects the majority voting to provide the final prediction. The majority of the decision trees have chosen apple as their prediction. This makes the classifier choose apple as the final prediction.



Figno 3.1 : Random Forest Classifier Example

3.2.1 HARDWARE REQUIREMENTS

Functional requirements for a secure cloud storage service are straight forward:

- a. The service should be able to store the user's data;
- b. The data should be accessible through any devices connected to the Internet;
- c. The service should be capable to synchronize the user's data between multiple devices (notebooks, smart phones, etc.);
- d. The service should preserve all historical changes (versioning);
- e. Data should be shareable with other users;
- f. The service should support SSO; and
- g. The service should be interoperable with other cloud storage services, enabling data migration from one CSP to another.
- h. Operating system: Windows
- i. Coding language: Python 3.7

3.2.2 SOFTWARE REQUIREMENTS

- a) **Processor** – Intel core i5
- b) **Speed** – 1.19 GHz
- c) **RAM** - 512 MB (min)
- d) **Hard Disk** - 20 GB
- e) **Key Board** - Standard Keyboard
- f) **Monitor** – 15 VGA Colour

CHAPTER 4

DESIGN DESCRIPTION OF PROPOSED PROJECT

4.1 PROPOSED METHODOLOGY

As mentioned above, Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model.

As the name suggests, "Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset." Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output.

ADVANTAGES OF PROPOSED SYSTEM

1. It overcomes the problem of overfitting by averaging or combining the results of different decision trees.
2. Random forests work well for a large range of data items than a single decision tree does.
3. Random forest has less variance than single decision tree.
4. Random forests are very flexible and possess very high accuracy.
5. Scaling of data does not require in random forest algorithm. It maintains accuracy even after providing data without scaling.
6. Random Forest algorithms maintains good accuracy even a large proportion of the data is missing.

MACHINE LEARNING

Before we take a look at the details of various machine learning methods, let's start by looking at what machine learning is, and what it isn't. Machine learning is often categorized as a subfield of artificial intelligence, but I find that categorization can often be misleading at first brush. The study of machine learning certainly arose from research in this context, but in the data science application of machine learning methods, it's more helpful to think of machine learning as a means of building models of data.

CHALLENGES IN MACHINE LEARNING

While Machine Learning is rapidly evolving, making significant strides with cyber security and autonomous cars, this segment of AI as whole still has a long way to go.

The reason behind is that ML has not been able to overcome number of challenges. The challenges that ML is facing currently are :

- **Quality of data**

Having good-quality data for ML algorithms is one of the biggest challenges. Use of low-quality data leads to the problems related to data preprocessing and feature extraction.

- **Time-Consuming task**

Another challenge faced by ML models is the consumption of time especially for data acquisition, feature extraction and retrieval.

- **Lack of specialist persons**

As ML technology is still in its infancy stage, availability of expert resources is a tough job.

- **No clear objective for formulating business problems**

Having no clear objective and well-defined goal for business problems is another key challenge for ML because this technology is not that mature yet.

- **Issue of over fitting & under fitting**

If the model is over fitting or under fitting, it cannot be represented well for the problem.

- **Curse of dimensionality**

Another challenge ML model faces is too many features of data points. This can be a real hindrance.

- **Difficulty in deployment**

Complexity of the ML model makes it quite difficult to be deployed in real life.

MODULES USED IN PROJECT

1. NUMPY

Numpy is a general-purpose array-processing package. It provides a high performance multidimensional array object, and tools for working with these arrays. It is the fundamental package for scientific computing with Python. It contains various features including these important ones:

Besides its obvious scientific uses, Numpy can also be used as an efficient multi-dimensional container of generic data. Arbitrary data-types can be defined using Numpy which allows Numpy to seamlessly and speedily integrate with a wide variety of databases.

1. PANDAS

Pandas is an open-source Python Library providing high-performance data manipulation and analysis tool using its powerful data structures. Python was majorly used for data munging And preparation. It had very little contribution towards data analysis. Pandas solved this problem. Using Pandas, we can accomplish five typical steps in the processing and analysis of data, regardless of the origin of data load, prepare, manipulate, model, and analyze. Python with Pandas is used in a wide range of fields including academic and commercial domains including finance, economics, Statistics, analytics, etc .

2. MATPLOTLIB

Matplotlib is a Python 2D plotting library which produces publication quality figures in a variety of hardcopy formats and interactive environments across platforms. Matplotlib can be used in Python scripts, the Python and IPython shells, the Jupyter Notebook, web application servers, and four graphical user interface toolkits. Matplotlib tries to make easy things easy and hard things possible. You can generate plots, histograms, power spectra, bar charts, error charts, scatter plots, etc., with just a few lines of code. For examples, see the sample plots and thumbnail gallery. For simple plotting the pyplot module provides a MATLAB-like interface, particularly when combined with IPython. For the power user, you have full control of line styles, font properties, axes properties, etc, via an object oriented interface or via a set of functions familiar to MATLAB users.

3. SCIKIT – LEARN

Scikit-learn provides a range of supervised and unsupervised learning algorithms via a consistent interface in Python. It is licensed under a permissive simplified BSD license and is distributed under many Linux distributions, encouraging academic and commercial use.

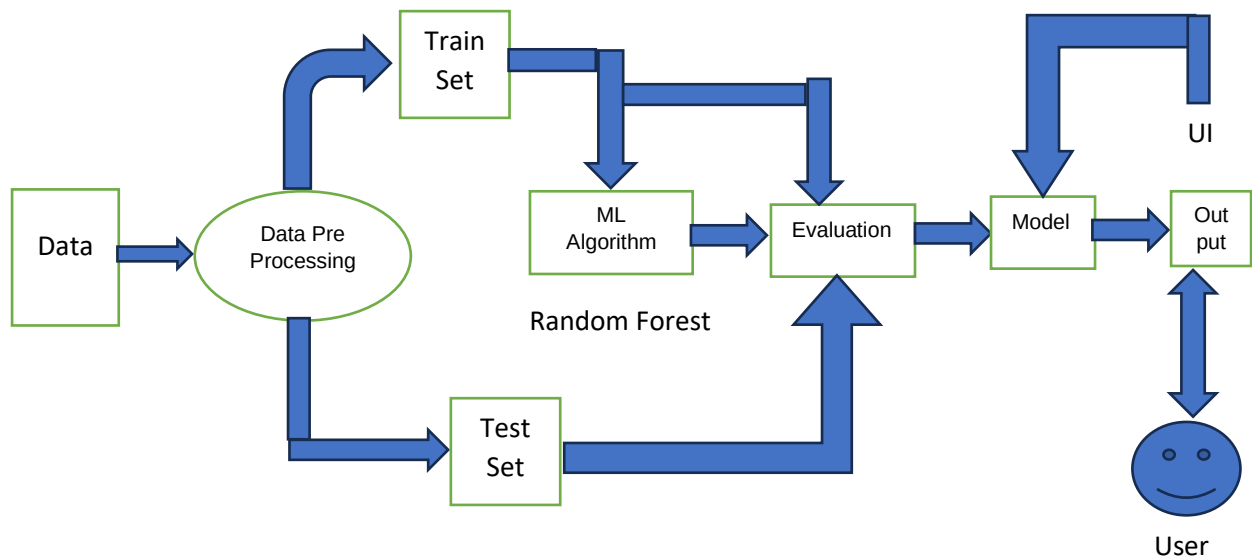
4. FLASK APPLICATION

Flask is a small and lightweight **Python web framework** that provides useful tools and features that make creating web applications in Python easier. It gives developers flexibility and is a more accessible framework for new developers since you can build a web application quickly using only a single Python file.

How to run a flask application on a local development server?

Finally, the run method of the Flask class is used to run the flask application on the local development server. The syntax is given below. The default hostname is 127.0.0.1, i.e. localhost. The port number to which the server is listening to. The default port number is 5000.

4.1.1 System Architecture



4.1.2 Various Stages

Urban water quality prediction involves several stages, each crucial for accurately forecasting the quality of water in urban areas. Here are the key stages involved:

Data Collection and Preprocessing: The first stage involves collecting relevant data on various water quality parameters such as pH, dissolved oxygen, turbidity, nutrient levels, and pollutant concentrations. This data can be collected from various sources including monitoring stations, sensors, satellite imagery, and citizen science initiatives. Once collected, the data needs to be preprocessed to remove noise, correct errors, and ensure consistency.

Feature Selection and Extraction: In this stage, relevant features or variables that influence water quality are selected from the collected data. This may involve statistical analysis, correlation studies, or domain knowledge. Additionally, feature extraction techniques may be employed to derive new variables or indicators from the raw data, enhancing the predictive power of the models.

Model Development: The core stage involves developing predictive models that relate the selected features to water quality parameters. Various modeling techniques

can be employed, including statistical regression models, machine learning algorithms (such as decision trees, random forests, support vector machines, and neural networks), and physical process-based models (such as hydrological models and water quality models). The choice of model depends on the complexity of the system, available data, and prediction requirements.

Model Calibration and Validation: Once the models are developed, they need to be calibrated and validated using historical data. Calibration involves adjusting model parameters to minimize the difference between observed and predicted values, while validation assesses the performance of the model using independent data. This stage ensures that the models accurately capture the relationships between predictors and water quality parameters.

Prediction and Forecasting: After calibration and validation, the models are used to make predictions and forecasts of water quality parameters for future time periods. These predictions can range from short-term forecasts (e.g., hours to days) to long-term projections (e.g., months to years). The predictions can be spatially explicit, providing information on water quality at specific locations within the urban area.

Uncertainty Assessment: Assessing the uncertainty associated with water quality predictions is essential for understanding the reliability and robustness of the forecasts. Uncertainty can arise from various sources such as data errors, model parameter uncertainty, and variability in environmental conditions. Techniques such as sensitivity analysis, Monte Carlo simulations, and ensemble modeling can be employed to quantify and communicate uncertainty in the predictions.

Decision Support and Communication: The final stage involves utilizing the predicted water quality information to support decision-making and communicate findings to stakeholders. This may involve developing decision support tools, providing actionable recommendations for water quality management, and disseminating information to policymakers, water resource managers, and the general public.

4.1.3 Internal or Component design structure



The screenshot shows a web application titled "Urban Water Quality Prediction" with a blue water drop icon. The interface features a central dark grey rounded rectangle containing a vertical stack of input fields and a "Predict" button. The input fields contain the following values from top to bottom: 2014, 6.7, 6.9, 203, 2, 0.2, and 8391. The "Predict" button is a blue-to-purple gradient. Below the input fields, a green box displays the output: "Good, The predicted value is [80.38]".

4.1.4 working principles

ALGORITHMS USED IN THIS PROJECT

RANDOM FOREST CLASSIFIER ALGORITHM

Random forest is a supervised learning algorithm which is used for both classification as well as regression. But however, it is mainly used for classification problems. As we know that a forest is made up of trees and more trees means more robust forest. Similarly, random forest algorithm creates decision trees on data samples and then gets the prediction from each of them and finally selects the best solution by means of voting. It is an ensemble method which is better than a single decision tree because it reduces the over-fitting by averaging the result.

Working Of Random Forest Algorithm:

Entropy is a metric for calculating uncertainty. Information gain is a measure of how uncertainty in the target variable is reduced, given a set of independent variables. The information gain concept involves using independent variables (features) to gain information about a target variable (class). The entropy of the target variable (Y) and the conditional entropy of Y (given X) are used to estimate the information gain. In this case, the conditional entropy is subtracted from the entropy of Y.

Information gain is used in the training of decision trees. It helps in reducing uncertainty in these trees. A high information gain means that a high degree of uncertainty (information entropy) has been removed. Entropy and information gain are important in splitting branches, which is an important activity in the construction of decision trees.

Let's take a simple example of how a decision tree works. Suppose we want to predict if a customer will purchase a mobile phone or not. The features of the phone form the basis of his decision. This analysis can be presented in a decision tree diagram.

The root node and decision nodes of the decision represent the features of the phone mentioned above. The leaf node represents the final output, either buying or not buying. The main features that determine the choice include the price, internal storage, and Random Access Memory (RAM). The decision tree will appear as follows.

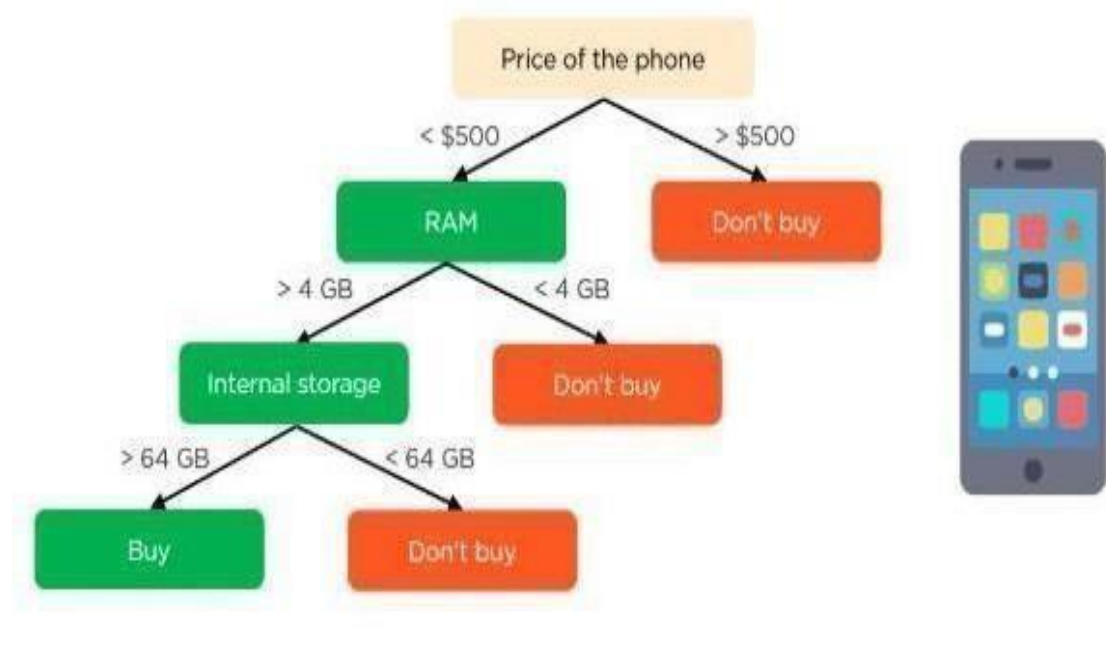


Fig 4.1.4 Decision Tree Example

Applying Decision Trees In Random Forest:

The main difference between the decision tree algorithm and the random forest algorithm is that establishing root nodes and segregating nodes is done randomly in the latter. The random forest employs the bagging method to generate the required prediction.

Bagging involves using different samples of data (training data) rather than just one sample. A training dataset comprises observations and features that are used for making predictions. The decision trees produce different outputs, depending on the training data fed to the random forest algorithm. These outputs will be ranked, and the highest will be selected as the final output.

Our first example can still be used to explain how random forests work. Instead of having a single decision tree, the random forest will have many decision trees. Let's assume we have only four decision trees. In this case, the training data comprising the phone's observations and features will be divided into four root nodes.

The root nodes could represent four features that could influence the customer's choice (price, internal storage, camera, and RAM). The random forest will split the nodes by selecting features randomly. The final prediction will be selected based on the outcome of the four trees. The outcome chosen by most decision trees will be the final choice. If three trees predict buying, and one tree predicts not buying, then the final prediction will be buying. In this case, it's predicted that the customer will buy the phone.

Classification in random forests:

Classification in random forests employs an ensemble methodology to attain the outcome. The training data is fed to train various decision trees. This dataset consists of observations and features that will be selected randomly during the splitting of nodes.

A rain forest system relies on various decision trees. Every decision tree consists of decision nodes, leaf nodes, and a root node. The leaf node of each tree is the final output produced by that specific decision tree. The selection of the final output follows the majority-voting system. In this case, the output chosen by the majority of the

decision trees becomes the final output of the rain forest system. The diagram below shows a simple random forest classifier.

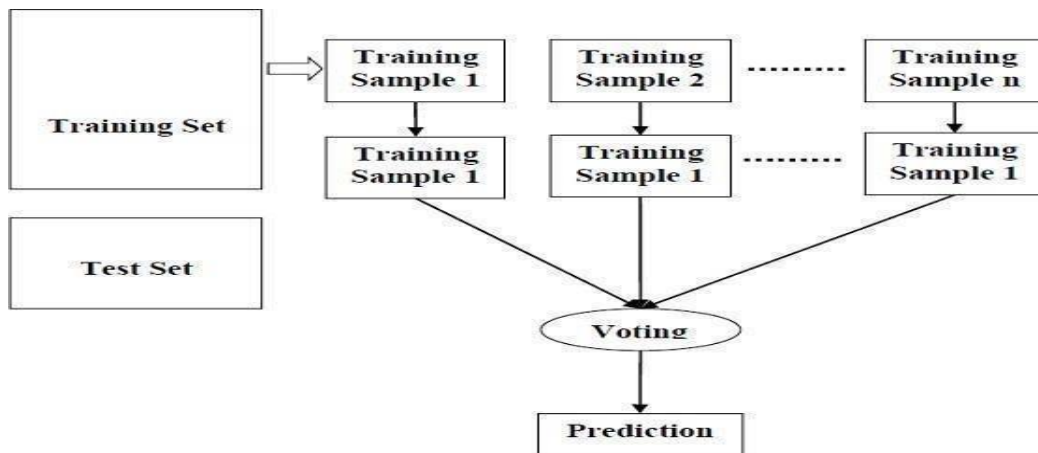


Fig 4.1.4 Random Forest Classifier

4.2 FEATURES

- In future works, we propose integrating the findings of this research in a large-scale IoT-based online monitoring system using only the sensors of the required parameters. The tested algorithms would predict the water quality immediately based on the real-time data fed from the IoT system. The proposed IoT system would employ the parameter sensors of pH, turbidity, temperature.
- It would identify poor quality water before it is released for consumption and alert concerned authorities. It will hopefully result in curtailment of people consuming poor quality water and consequently de-escalate harrowing diseases like typhoid and diarrhea. In this regard, the application of a prescriptive analysis from the expected values would lead to future facilities to support decision and policy makers.

4.2.1 Novelty of the proposal

Water quality prediction involves using various techniques and models to forecast the quality of water in a particular area over a certain period. The novelty in water quality prediction lies in several aspects:

1.Advanced Modeling Techniques: Traditional methods for water quality prediction often relied on simplistic statistical models. However, with advancements in technology and computational power, more sophisticated techniques such as machine learning, artificial intelligence, and computational fluid dynamics are being employed. These methods can handle complex interactions between different water quality parameters and environmental factors, resulting in more accurate predictions.

2.Integration of Big Data: With the proliferation of sensors, satellites, and other monitoring devices, vast amounts of data on water quality are being collected. Novel approaches involve integrating this big data into predictive models, allowing for real-time or near real-time predictions. This includes data from sources such as remote sensing, IoT devices, citizen science initiatives, and government monitoring programs.

3.Spatial and Temporal Resolution: Traditional water quality models often provided predictions at coarse spatial and temporal resolutions. However, newer approaches aim to increase the resolution of predictions, providing detailed insights into water quality variations at smaller spatial scales (e.g., individual water bodies, watersheds) and finer temporal scales (e.g., hourly or daily predictions).

4.Uncertainty Quantification: Assessing the uncertainty associated with water quality predictions is crucial for informed decision-making. Novel approaches incorporate uncertainty quantification techniques, such as probabilistic modeling, ensemble methods, and sensitivity analysis, to provide confidence intervals and assess the reliability of predictions.

5.Integration of Multi-Source Data: Water quality is influenced by a wide range of factors, including land use, weather patterns, hydrological processes, and human activities. Novel prediction models integrate multi-source data from diverse disciplines.

6.Adaptive and Dynamic Models: Environmental conditions and human activities can change rapidly, leading to shifts in water quality dynamics. Novel prediction models incorporate adaptive and dynamic components that can adjust in real-time to changing conditions, ensuring that predictions remain relevant and accurate under varying circumstances.

Overall, the novelty in water quality prediction lies in the integration of advanced modeling techniques, big data analytics, increased spatial and temporal resolution, uncertainty quantification, multi-source data integration, and adaptive modeling approaches, leading to more accurate and actionable predictions for water quality management and decision-making.

CHAPTER 5

RESULT AND DISCUSSION

5.1 SCREENSHOTS



Fig 5.1.1 Heat Map

```
In [35]: #Data Visualization
data1.boxplot(column='year')
```

```
Out[35]: <AxesSubplot:>
```

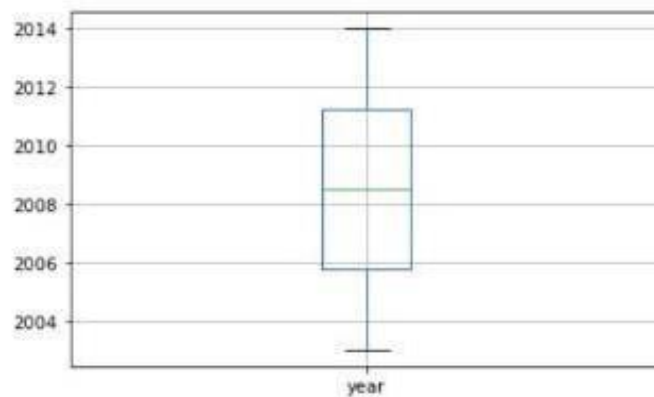


Fig 5.1.2 Box Plot

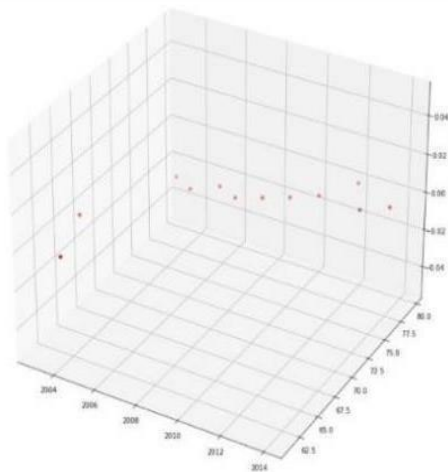


Fig 5.1.3

```
#visualizing the filtered data

year=data1['year'].values
AQI=data1['wqi'].values
data1['wqi']=pd.to_numeric(data1['wqi'])
data1['year']=pd.to_numeric(data1['year'])

import matplotlib.pyplot as plt
plt.rcParams['figure.figsize'] = (20.0, 10.0)
from mpl_toolkits.mplot3d import Axes3D
fig = plt.figure()
ax = Axes3D(fig)
ax.scatter(year,AQI, color='red')
plt.show()
data1
```

3D Scatter Plot

```
In [37]: #scatter plot of data points
cols = ['year']
y = data1['wqi']
x=data1[cols]

plt.scatter(x,y)
plt.show()
```

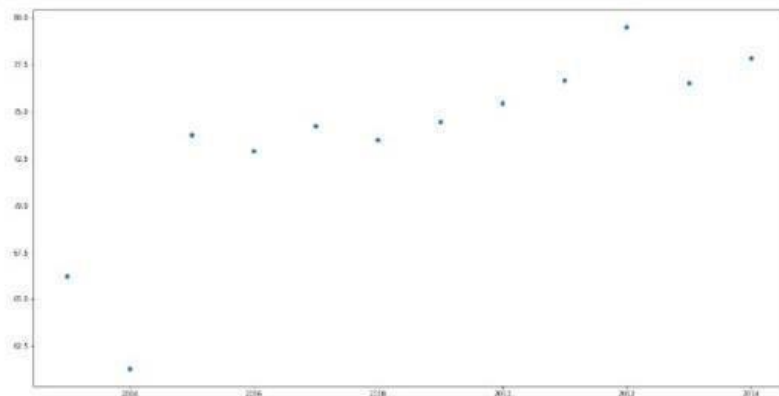


Fig 5.1.4 Scatter Plot

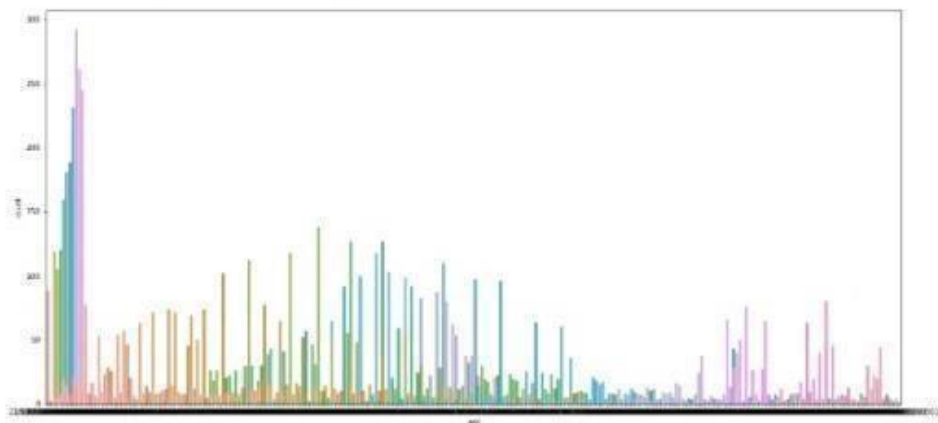


Fig 5.1.5 Count Plot

Splitting the data into training and testing set

```
In [46]: 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_stat

In [47]: x_train.shape
Out[47]: (1592, 7)

In [48]: x_test.shape
Out[48]: (399, 7)

In [49]: y_train.shape
Out[49]: (1592, 1)

In [50]: y_test.shape
Out[50]: (399, 1)
```

Fig 5.1.6 Splitting of Data Set

Random Forest Regressor

A random forest is a meta estimator that fits a number of classifying decision trees on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting.

`n_estimators` = number of trees in the forest

```
In [51]: from sklearn.ensemble import RandomForestRegressor
reg_rf = RandomForestRegressor(n_estimators = 10 ,random_state = 0)
reg_rf.fit(x_train, y_train)
y_pred = reg_rf.predict(x_test)
```

Fig 5.1.7 Random Forest Regressor

```
In [58]: sns.distplot(y_test-y_pred)
plt.show()
```

C:\Users\DELL\anaconda3\lib\site-packages\seaborn\distributions.py:2619: FutureWarning: `distplot` is a deprecated function and will be removed in a future version. Please adapt your code to use either `displot` (a figure-level function with similar flexibility) or `histplot` (an axes-level function for histograms).

warnings.warn(msg, FutureWarning)

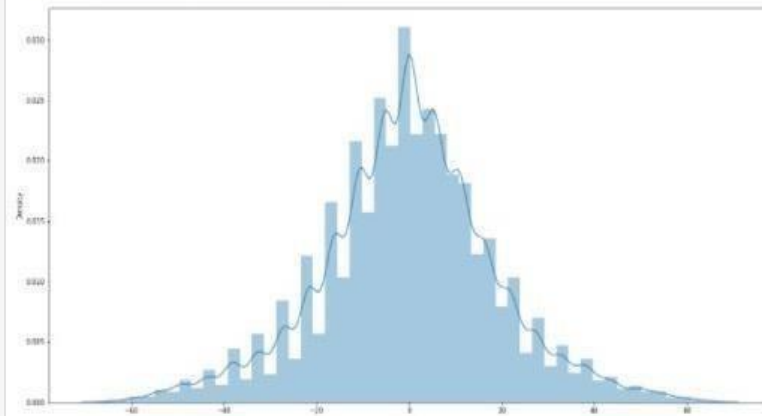


Fig 5.1.8 Dist Plot

```
In [58]: metrics.r2_score(y_test, y_pred)
```

```
Out[58]: 0.9669439427778931
```

```
In [ ]:
```

Save the model to reuse it again

```
In [59]: import pickle
pickle.dump(reg_rf,open('wqi.pkl','wb'))
model = pickle.load(open('wqi.pkl','rb'))
```

```
pickle.dump(sc,open('sc.pkl','wb'))
model = pickle.load(open('sc.pkl','rb'))
```

Fig 5.1.9 Accuracy and Saving the model

CHAPTER 5

CONCLUSION

Monitoring the water quality is a vital necessity as it provides specific data on the nature of the water. Different types of physical, material, and organic data are combined to determine the list of different attributes. Diversity indices for water quality monitoring can be used for resource allocation, location ranking, standard practice, trend analysis, public information and research. The present study describes the dominant physiochemical variables and their relationship to phytoplankton density and community and a pollution gradient. The proposed operation has developed a performance-based intelligence algorithm, the NARNET and LSTM models used to predict WQI. In addition, it classifies WQI data using machine learning algorithms such as SVM, KNN, and Naive Bayes. The proposed model is tested and evaluated by other mathematical parameters. Through WQI predictions, the results show that the performance of the NARNET model is slightly better than that of the LSTM model, based on the R-values obtained. However, the SVM algorithm achieved the highest WQC prediction accuracy compared to the KNN and Naive Bayes algorithms. After assessing the robustness and efficiency of the proposed WQI prediction model, improved models will be used in future work to predict the water quality of different water types in Saudi Arabia. Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset." Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output.

This paper tried to investigate and evaluate the teaching assistant staff to adopt the prediction process based on the data size, time process, accuracy, estimated error factor. The quality of teaching assistant staff was emphasized using artificial neural network algorithms in this study, the experiments conducted based on UCI teaching assistant dataset. The results of the evaluation obtained using with different size in training and testing phases. The deep examinations highlighted that the group of 90% achieved better results in the prediction accuracy, estimated time, and error.

REFERENCES

1. Ping Liu Jin Wang Arun Kumar Sangaiah Yang Xie Xinchun Yin, 2019. Analysis and Prediction of Water Quality Using LSTM Deep Neural Networks in IoT Environment, Sustainability, MDPI, Open Access Journal, vol. 11(7), pages 1- 14, April.
2. K. Gopavanitha and S. Nagaraju A low cost system for real time water quality monitoring and controlling using IoT||, IEEE Conference paper, August 2017.
3. Manish Kumar Jha ; Rajni Kumari Sah ; M. S. Rashmitha ; Rupam Sinha ; B. Sujatha ; K. V. Suma Smart Water Monitoring System for Real-Time Water Quality and Usage Monitoring||, IEEE Conference paper, July 2018
4. Asma Al Khaili, Aisha Al Mamari, Hoda Amer and Walid Ibrahim An Affordable System for Remotely Monitoring Water Quality in Residential Water Tanks||, IEEE paper, November 2018
5. Vaishnavi V. Daigavane, Dr. M. A. Gaikwad. ||Water Quality Monitoring System Based on IOT|| 2017 Advances in Wireless and Mobile Communications, Nov 2017 ISSN 0973-6972
6. Najah Ahmed, A.; Binti Othman, F.; Abdulmohsin Afan, H.; Khaleel Ibrahim, R.; Ming Fai, C.; Shabbir Hossain, M.; Ehteram, M.; Elshafie, A. Machine learning methods for better water quality prediction. J. Hydrol. 2019, 578, 124084.
7. Yang, G.; Moyer, D.L. Estimation of nonlinear water-quality trends in high-frequency monitoring data. Sci. Total Environ. 2020, 715, 136686.
8. Grayman, W.M.; Day, H.J.; Luken, R. Regional water quality management for the Dong Nai River Basin, Vietnam. Water Sci. Technol. 2003, 48, 17–23.