DATA ENGINEERING CONCEPTS

Mini Project Report

On

WATER QUALITY ANALYSIS

ON

RIVER YAMUNA AND RIVER INDRAYANI

Submitted to: **Prof. Varsha Powar**

Submitted by:

Drishti Chauhan PD38 1032211866

Shivanshi Singh PD 40 1032211883

Akshat Lahariya PD41 1032211891

School of Computer Engineering and Technology, MIT World Peace University Pune

2023-2024

ABSTRACT

This project is centred around a detailed examination of the water quality in two significant rivers, namely River Yamuna and River Indrayani. It commenced by assimilating and meticulously preprocessing two distinct datasets. The primary objective was to ensure the reliability and accuracy of the data, laying a solid foundation for subsequent analysis. The datasets encapsulated diverse parameters related to water quality, laying the groundwork for comprehensive evaluation.

Upon completing the preprocessing phase, an array of sophisticated machine learning models was deployed. These models encompassed Decision Tree, Logistic Regression, Gradient Boosting, Random Forest, Keans Clustering, SVM, and K-Nearest Neighbours. Their implementation was aimed at extracting meaningful insights and predictive patterns from the datasets. The application of varied models facilitated a multifaceted analysis, empowering the project to uncover intricate relationships between different water quality parameters. This approach enabled a nuanced understanding of the characteristics and health of the rivers' water.

The culmination of this project represents a pivotal stride in comprehending the dynamics of River Yamuna and River Indrayani's water quality. Through extensive analysis and model utilization, the project unearthed correlations, trends, and predictive indicators crucial for effective water resource management. The insights derived from this endeavour hold profound significance for environmental conservation efforts and informed decision-making regarding these rivers' preservation. This project serves as a valuable resource for policymakers, environmentalists, and researchers, offering comprehensive data-driven insights essential for devising strategies aimed at safeguarding these vital water bodies.

TABLE OF CONTENTS

INTRODUCTION

The exploration of water quality is a fundamental pursuit crucial to environmental sustainability. Within this realm, our project embarks on a comprehensive analysis centred around vital parameters sourced from a dataset encompassing sampling seasons, sample numbers, and a diverse array of critical metrics. Through meticulous examination of parameters such as air and water temperature, pH levels, electrical conductivity, turbidity, dissolved oxygen, biological oxygen demand, chemical oxygen demand, and various elemental compositions, our objective stands resolute: to unravel the intricate tapestry of water quality dynamics.

Amidst the intricate interplay of these parameters, our mini-project takes shape as an endeavour to decipher the health and characteristics of water bodies. Through rigorous analysis and interpretation of this dataset, we endeavour to uncover correlations, trends, and predictive patterns that illuminate the condition and potential implications on the environment. This report stands as a testament to our commitment to understanding, assessing, and contributing insights pivotal for informed decision-making and proactive measures in the realm of water resource management and environmental conservation.

MOTIVATION

This study is motivated by the increasing concern and urgency to address the declining water quality of the Indrayani and Yamuna rivers. Water is a vital resource for the environment, for communities, and for industries. It is essential to understand and address any issues affecting water quality in order to ensure the well-being of the environment and public health. The reasons for this analysis are as follows:

Environmental Impact: The health of rivers plays an important role in the overall environmental health of a region. The Indrayani and the Yamuna River are both essential for the preservation of local ecosystems. A decrease in water quality can have serious consequences for flora and fauna, as well as for public health.

Industrial and Agricultural Impact: Industries and agriculture rely heavily on river water for their operations, and any high levels of pollutants can have a negative impact on crops, livestock and industrial processes.

Regulatory Compliance: Governments and regulatory bodies set standards for water quality to ensure environmental sustainability and public health. The study is motivated by the need to assess compliance with these standards and identify areas where corrective measures may be necessary.

Long-term Sustainability: Sustainable management of water resources is vital for future generations. By comprehensively analysing the water quality parameters, the study aims to contribute insights that support long-term sustainability goals and encourage responsible water resource management practices.

PROBLEM DEFINITION

The focal point of this study delves into the alarming decline of water quality within the Indrayani and Yamuna rivers, primarily attributed to an intricate web of contributors. Industrial discharges, the ramifications of rapid urbanization, agricultural runoff, and inadequate wastewater treatment collectively compound the escalating pollution levels in these crucial water bodies. The cumulative impact of these factors poses significant threats to both environmental integrity and public health.

The burgeoning pace of urbanization and industrial expansion acts as catalysts, elevating the influx of pollutants into these rivers. Agricultural runoff, coupled with untreated household effluents, further exacerbates the pollution burden, intensifying its adverse effects. The inadequacy of wastewater treatment facilities exacerbates this onslaught, significantly compromising the quality of these river systems. The resultant imbalance not only imperils aquatic ecosystems but also escalates risks to public health.

Given the intricate interplay of these multifaceted factors, a comprehensive inquiry becomes imperative. Pinpointing the specific sources and meticulously analysing their individual contributions form the crux of this investigation. Our pursuit lies in unveiling the nuanced nuances of each contributor, discerning their unique impact on water quality. Such granular insights will pave the way for tailored strategies aimed at mitigating the overarching deterioration of these vital river systems' water quality.

OBJECTIVES

Temporal Trend Analysis: Find patterns, seasonality and potential trends in water quality parameters over different months and years.

Spatial Variation Assessment: Analyse spatial variation of water quality parameters to identify areas with consistently low or rising water quality.

Comparison with Regulatory Standards: Compare observed water quality values to regulatory standards and guidelines for compliance and identify locations where water quality is below acceptable levels.

Seasonal Variations and Influences: Analyse seasonal variations and influences on water quality by examining how factors such as temperature and precipitation may affect changes in pH, COD, BOD, and DO.

Correlation analysis: Find correlations between water quality parameters that provide insight into possible causal relationships that help to understand the intricate dynamics that affect water quality

Public health risk assessment: Assess the public health risks associated with the water quality parameters observed, with particular attention to faecal coliform levels as a risk indicator for waterborne disease

Mitigation recommendations: Based on the results of the correlation analysis, suggest specific mitigation measures and interventions to improve water quality in the problem areas identified, taking into account both point as well as non-point sources of pollution

Long-term monitoring recommendations: Develop a plan for continuous monitoring and evaluation to monitor the success of interventions and adjust strategies as needed to continuously improve water quality.

TOOLS USED

- R or Python: for data preprocessing, statistical analysis, and visualization.
- Pandas, NumPy, SciPy: for data manipulation and analysis.
- Matplotlib, Seaborn: for data visualization.
- Scikit-learn: for implementing machine learning models like Decision Trees, Logistic Regression, Gradient Boosting, Random Forest, SVM, KMeans, and KNeighborsRegressor.
- TensorFlow or PyTorch: for more complex machine learning or neural network models.
- Power BI: for interactive data visualization and dashboard creation.
- Microsoft Word: for formatting and compiling the project report.
- Microsoft Excel: Excel allows easy visualization of CSV files, enabling quick inspection of data rows, columns, and basic statistics.

DATASET DESCRIPTION

The provided dataset comprises water quality measurements for the Indrayani River and Yamuna River across different locations and over several months from January 2020 to July 2023. The dataset contains the following key parameters: pH, COD (Chemical Oxygen Demand) in mg/l, BOD (Biochemical Oxygen Demand) in mg/l, DO (Dissolved Oxygen) in mg/l, and Faecal Coliform levels in MPN/100ml.

- **Temporal Scope:** The dataset covers a period of several months, starting from January 2020 to July 2023, allowing for a comprehensive examination of seasonal variations and trends in water quality.
- **Spatial Scope:** Measurements are recorded at nine different locations along the rivers, including Palla, Surghat, Palton Pool, Kudesia Ghat, ITO Bridge, Nizamuddin Bridge, Agra Canal, Jaitpur, and D/S Okhla Barrage. This spatial granularity enables a detailed analysis of water quality variations across diverse environments.
- **Parameters:**
    - **pH:** The measure of acidity or alkalinity, influencing the overall chemical and biological processes in the water.
    - **COD (Chemical Oxygen Demand):** An indicator of the amount of oxygen required to chemically oxidize organic matter, often reflective of pollution levels.
    - **BOD (Biochemical Oxygen Demand):** The amount of dissolved oxygen needed by microorganisms to break down organic material in water, indicating the level of organic pollution.
    - **DO (Dissolved Oxygen):** The concentration of oxygen dissolved in water, crucial for the survival of aquatic organisms.
    - **Faecal Coliform:** A measure of bacterial contamination, specifically indicative of faecal matter, and a key parameter for assessing water's suitability for human use.
- **Units:** Measurements are provided in standard units - pH is dimensionless, COD, BOD, and DO are measured in mg/l, and Faecal Coliform levels are presented in MPN (Most Probable Number) per 100ml of water.
- **Data Granularity:** Recorded on a monthly basis for each location, the dataset provides a detailed temporal resolution for understanding variations and trends.
- **Quality Indicators:** The dataset includes instances where certain parameters are marked as not measured or missing, which should be considered during data analysis and interpretation.

DATA PRE-PROCESSING

1. **Handling Missing Values:**
    - Identify and address missing values in each parameter (pH, COD, BOD, DO, Faecal Coliform) at different locations and time points. Options include imputation techniques such as mean, median, or forward/backward filling, considering the potential impact on the analysis.

2. **Consistent Date-Time Format:**
    - Ensure uniformity in the date-time format to facilitate chronological analysis. Convert date and month entries to a standard format across the dataset.

3. **Unit Standardization:**
    - Verify that all measurements are in consistent units (mg/l for pH, COD, BOD, DO), facilitating accurate analysis and interpretation. Address any discrepancies in units if present.

4. **Outlier Detection and Handling:**
    - Identify outliers in each parameter at different locations. Evaluate whether outliers are indicative of measurement errors or valid data points. If necessary, consider techniques like transformation to manage outliers. Zscore analysis has been done.

DATA MINING TASKS PERFORMED

1. **Identification of Most Polluted Season:**
    - Utilized various data mining algorithms such as decision trees, logistic regression, gradient boosting, random forest, SVM, KMeans, and KNeighborsRegressor.
    - Leveraged these models to analyze and identify the season(s) with the highest pollution levels in the Indrayani and Yamuna rivers.

2. **Analysis of Key Factors Affecting pH Levels:**
    - Employed data mining techniques to determine the most influential factors impacting pH levels in the rivers.
    - Identified and assessed the significance of variables like industrial discharges, urbanization effects, agricultural runoff, or other pollutants on pH levels.

3. **Determining Most Influential Factors for Specified Pollutants:**
    - Analysed the dataset to uncover the primary contributors or factors significantly affecting specific pollutants within the rivers.
    - Investigated and quantified the impact of various contributors such as industrial discharge, urbanization, or agricultural runoff on the concentration of specified pollutants.

4. **Identification of Least Present Element:**
    - Utilized data mining approaches to identify and quantify the element that appeared least frequently or had the lowest presence in the water quality dataset.
    - Explored the dataset to determine the least prevalent element and potentially understand its implications in the context of water quality.

ALGORITHM

1. Logistic Regression for pH Classification:

A. Load the CSV and store in a DataFrame.
B. Define a function to classify pH levels.
C. Create a 'PhLevel' column based on the classification of pH.
D. Create features and target variables.
E. Handle missing values.
F. Split the data into training and testing sets.
G. Train the Logistic Regression model.
H. Make predictions using the test set.
I. Evaluate the model:
    a. Calculate accuracy using accuracy_score().
    b. Generate a classification report using classification_report().
    c. Display confusion matrix using confusion_matrix().
J. Display and store evaluation metrics in the DataFrame.

2. Random Forest for pH Classification:

A. Load the CSV and store in a DataFrame.
B. Define a function to classify pH levels.
C. Create a 'PhLevel' column based on the classification of pH.
D. Create features and target variables.
E. Handle missing values.
F. Split the data into training and testing sets.
G. Train the Random Forest model.
H. Make predictions using the test set.
I. Evaluate the model:
    a. Calculate accuracy using accuracy_score().
    b. Generate a classification report using classification_report().
    c. Display confusion matrix using confusion_matrix().
J. Display and store evaluation metrics in the DataFrame.

3. Random Forest Regressor for Predicting Total Pollution:

A. Load the CSV and store in a DataFrame.
B. Create a new column for the sum of COD and BOD.
C. Define features and target variable for Total Pollution prediction.
D. Handle missing values.
E. Split the data into training and testing sets.
F. Train the RandomForestRegressor model.
G. Predict total pollution for each season.
H. Identify the most polluted season.
I. Display and store relevant information in the DataFrame.

4. Gradient Boosting Regressor for pH Prediction:

1. Load the CSV and store in a DataFrame.
2. Define features and target variable for pH prediction.
3. Handle missing values.
4. Split the data into training and testing sets.
5. Train the GradientBoostingRegressor model.
6. Predict pH levels using the test set.
7. Calculate feature importance for pH prediction.
8. Identify the most important feature affecting pH levels.
9. Store feature importance details in the DataFrame.

5. K-Nearest Neighbors for Water Quality Classification:

1. Load the CSV and store in a DataFrame.
2. Create a target variable to categorize water quality.
3. Define features and target variable for classification.
4. Handle missing values.
5. Split the data into training and testing sets.
6. Train the KNeighborsClassifier model.
7. Predict water quality for each row in the dataset.
8. Evaluate the model's accuracy.
9. Display and store evaluation metrics in the DataFrame.

OUTPUT

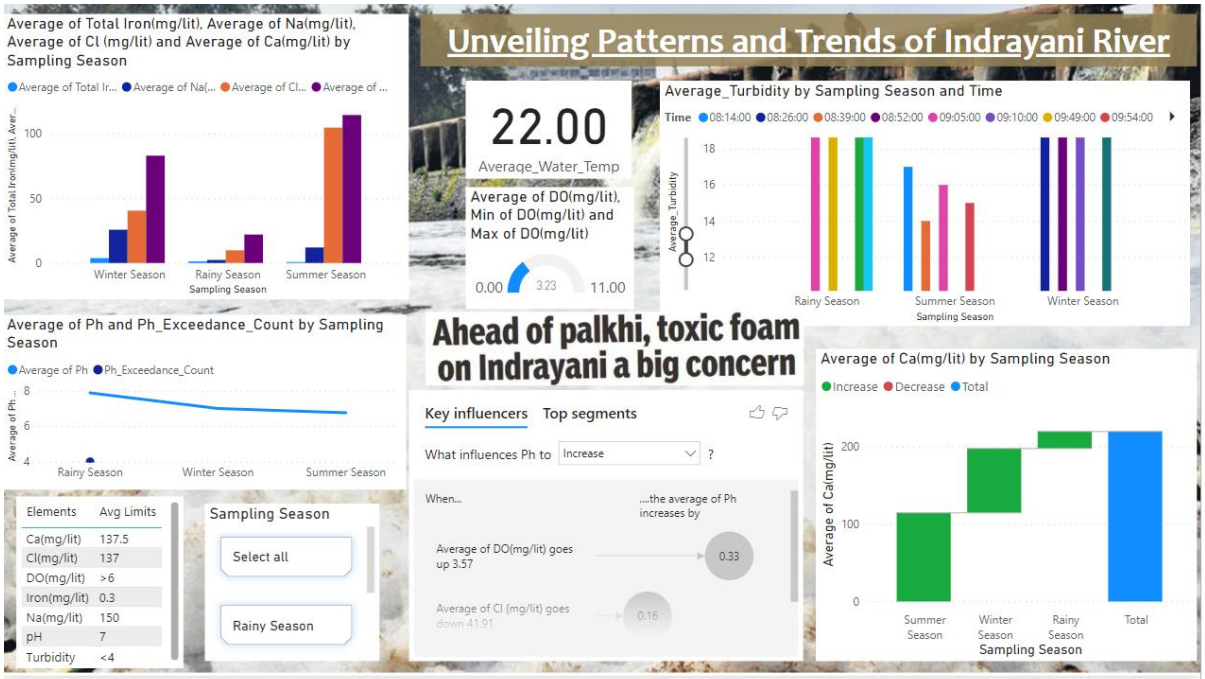
```
Enter the pollutant (Turbidity, Cl, BOD): Cl
Mean Squared Error for Cl: 58.53885533333312
Mean Absolute Error for Cl: 7.523333333333319
Root Mean Squared Error for Cl: 7.651068901358366

Feature Importance for Cl:
                Feature  Importance
9           PO(mg/lit)    0.211148
10  Total Iron(mg/lit)    0.185669
5           Ca(mg/lit)    0.171637
11      Silica(mg/lit)    0.150910
0          Air temp(°)    0.070726
2           EC (pS/cm)    0.046482
3           DO(mg/lit)    0.036310
4           Na(mg/lit)    0.027494
7           SO(mg/lit)    0.025972
6         HCO3(mg/lit)    0.023334
1       Water temp(*C)    0.018653
8          NO3(mg/lit)    0.018596
12           TS(mg/lit    0.013070

The most important feature for Cl is: PO(mg/lit)
```

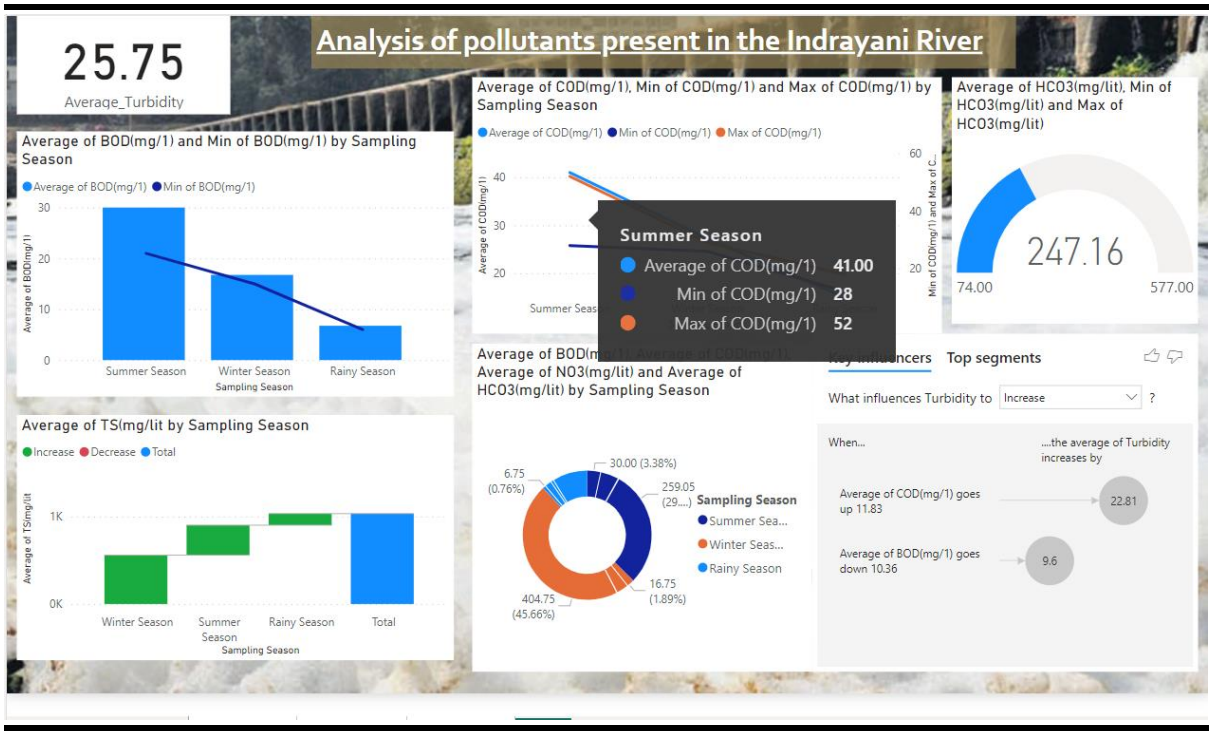
```
Mean Squared Error: 770.5424666666668
The most polluted season (predicted) is: Summer Season

All values for Summer Season season:
   Sampling Season  SampleNo  Air temp(°)  Water temp(*C)   Ph  EC (pS/cm)  \
8    Summer Season         1           25              26  7.0         352
9    Summer Season         2           26              22  6.0         551
10   Summer Season         3           28              22  7.0         565
11   Summer Season         4           30              24  7.0         624

    Turbidity   DO(mg/lit)  BOD(mg/1)  COD(mg/1)  ...  NO3(mg/lit)  \
8          17         0.8         26         40  ...         0.93
9          14         0.0         34         44  ...         1.00
10         16         2.0         39         52  ...         1.00
11         15         2.0         21         28  ...         1.60

    PO(mg/lit)  Total Iron(mg/lit)  Silica(mg/lit)  TS(mg/lit  Hour  Minute  \
8       0.3600              0.290          0.8700        230     8      14
9       0.6200              0.270          0.9700        360     8      39
10      0.7920              0.281          1.0000        370     9       5
11      0.9741              0.273          0.9088        410     9      54

    WaterQuality  PhLevel  Total_Pollution
8            Bad  Neutral               66
9            Bad   Acidic               78
10           Bad  Neutral               91
11           Bad  Neutral               49
```

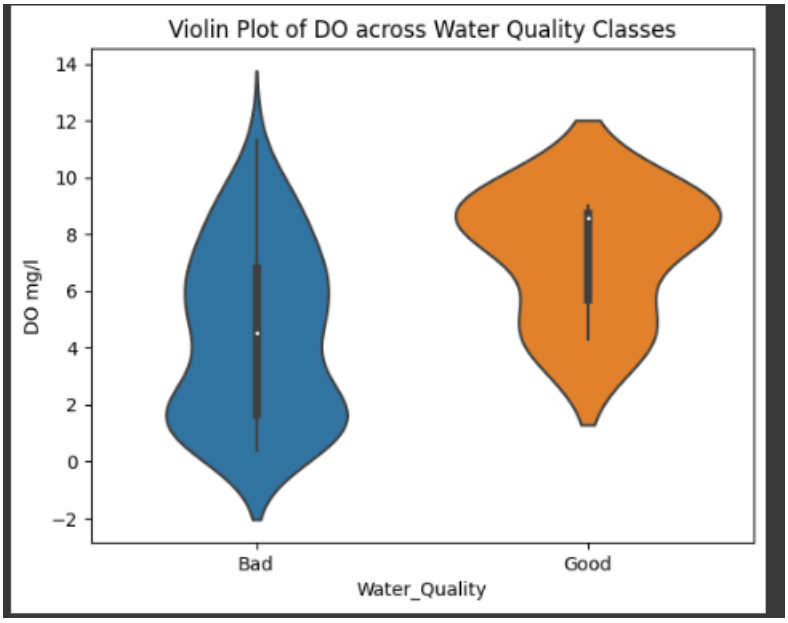
```
    For the month of May 2020, at location Palla:
      Actual BOD mg/l: 2.80, Predicted BOD mg/l: 2.81
      Actual COD mg/l: 12.00, Predicted COD mg/l: 11.70
    For the month of May 2020, at location Surghat:
      Actual BOD mg/l: 3.80, Predicted BOD mg/l: 3.83
      Actual COD mg/l: 16.00, Predicted COD mg/l: 16.28
    For the month of May 2020, at location Palton Pool:
      Actual BOD mg/l: 33.00, Predicted BOD mg/l: 32.97
      Actual COD mg/l: 116.00, Predicted COD mg/l: 114.76
    For the month of May 2020, at location Kudesia Ghat:
      Actual BOD mg/l: 25.00, Predicted BOD mg/l: 24.91
      Actual COD mg/l: 60.00, Predicted COD mg/l: 59.88
    For the month of May 2020, at location ITO Bridge:
      Actual BOD mg/l: 22.00, Predicted BOD mg/l: 21.98
      Actual COD mg/l: 32.00, Predicted COD mg/l: 31.86
    For the month of May 2020, at location Nizamuddin Bridge:
      Actual BOD mg/l: 16.00, Predicted BOD mg/l: 16.00
      Actual COD mg/l: 42.00, Predicted COD mg/l: 42.20
    For the month of May 2020, at location Agra Canal:
      Actual BOD mg/l: 16.00, Predicted BOD mg/l: 16.01
      Actual COD mg/l: 42.00, Predicted COD mg/l: 42.00
    For the month of May 2020, at location Jaitpur:
      Actual BOD mg/l: 17.00, Predicted BOD mg/l: 16.85
      Actual COD mg/l: 48.00, Predicted COD mg/l: 47.98
    For the month of May 2020, at location D/S okhla Barrage:
      Actual BOD mg/l: 23.00, Predicted BOD mg/l: 22.89
      Actual COD mg/l: 76.00, Predicted COD mg/l: 75.96
```

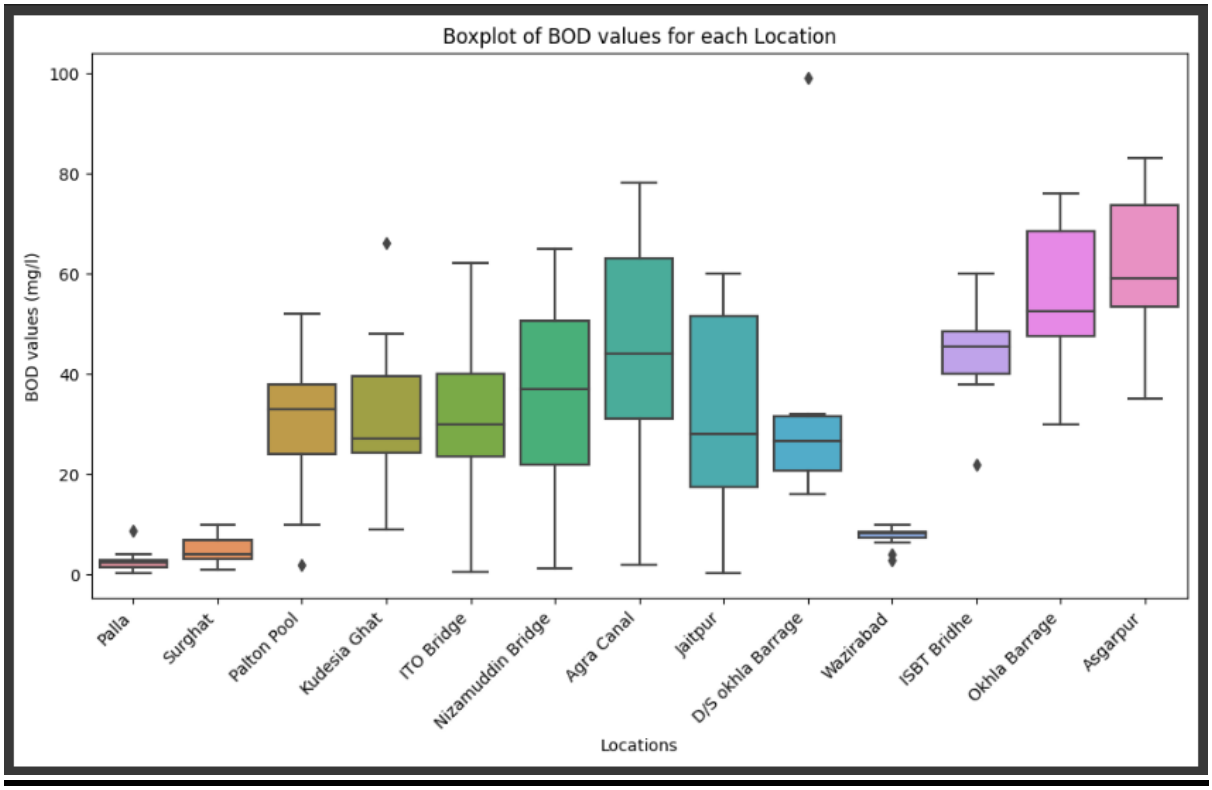
```
Enter the prediction year: 2020
Choose a model (L for Logistic Regression, D for Decision Tree, G for Gradient Boosting): D
Model: DecisionTreeClassifier
Accuracy: 1.00
F1 Score: 1.00
Confusion Matrix:
[[ 4  0]
 [ 0 41]]
Classification Report:
              precision    recall  f1-score   support

      Acidic       1.00      1.00      1.00         4
    Alkaline       1.00      1.00      1.00        41

    accuracy                           1.00        45
   macro avg       1.00      1.00      1.00        45
weighted avg       1.00      1.00      1.00        45

Row 9: Actual pH Category - Alkaline, Predicted pH Category - Alkaline
Row 184: Actual pH Category - Alkaline, Predicted pH Category - Alkaline
Row 120: Actual pH Category - Alkaline, Predicted pH Category - Alkaline
Row 207: Actual pH Category - Alkaline, Predicted pH Category - Alkaline
Row 148: Actual pH Category - Alkaline, Predicted pH Category - Alkaline
Row 214: Actual pH Category - Alkaline, Predicted pH Category - Alkaline
Row 182: Actual pH Category - Alkaline, Predicted pH Category - Alkaline
Row 86: Actual pH Category - Acidic, Predicted pH Category - Acidic
Row 178: Actual pH Category - Alkaline, Predicted pH Category - Alkaline
```

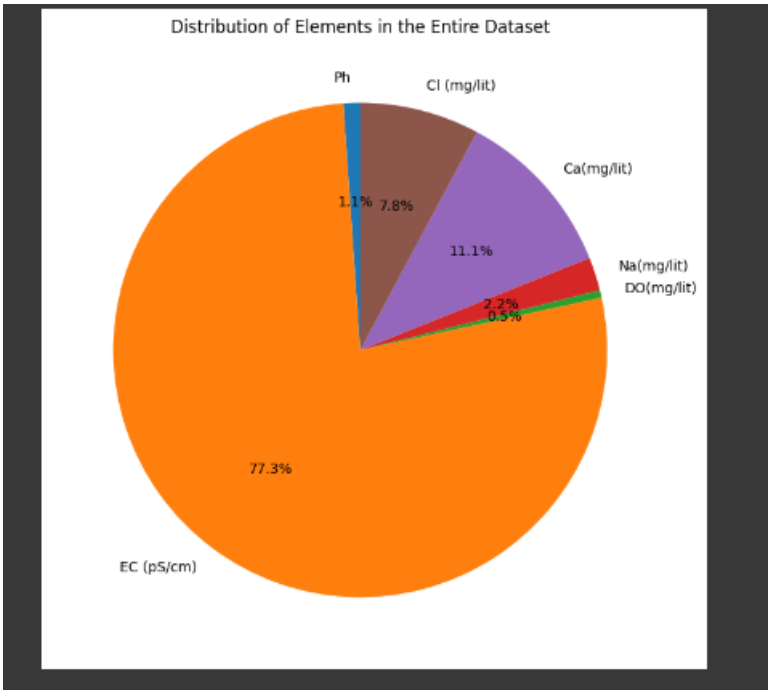
VISUALISATION SCREENSHOTS

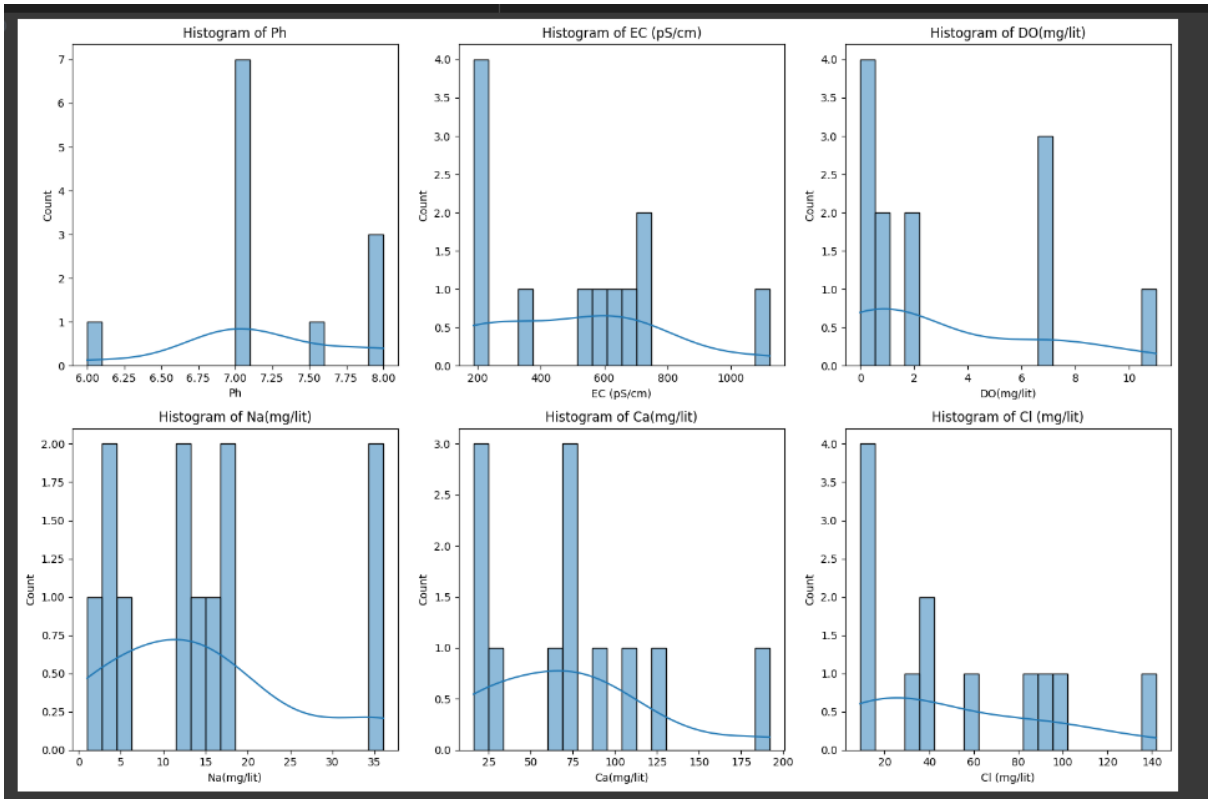





Violin Plot of DO across Water Quality Classes
DO mg/l
14
12
10
8
6
4
2
0
−2
Bad
Good
Water_Quality


Boxplot of BOD values for each Location
BOD values (mg/l)
100
80
60
40
20
0
Palla
Surghat
Palton Pool
Kudesia Ghat
ITO Bridge
Nizamuddin Bridge
Agra Canal
Jaitpur
D/S okhla Barrage
Wazirabad
ISBT Bridhe
Okhla Barrage
Asgarpur
Locations


Distribution of Elements in the Entire Dataset
Ph
Cl (mg/lit)
Ca(mg/lit)
Na(mg/lit)
DO(mg/lit)
EC (pS/cm)
1.1%
7.8%
11.1%
2.2%
0.5%
77.3%


Histogram of Ph
Histogram of EC (pS/cm)
Histogram of DO(mg/lit)
Histogram of Na(mg/lit)
Histogram of Ca(mg/lit)
Histogram of Cl (mg/lit)
Count
Ph
EC (pS/cm)
DO(mg/lit)
Na(mg/lit)
Ca(mg/lit)
Cl (mg/lit)

CONCLUSION

To sum up, our analysis of Indrayani River and Yamuna River water quality dataset provides important insights into time and spatial variability, compliance with regulations, and potential consequences on ecosystems and human health. Identifying hotspots with high Faecal coliform levels highlights immediate need for intervention. Correlation analysis provides a nuanced view of parameter relationships that guide targeted mitigation efforts. The use of machine learning exploration hints at the potential of predictive modelling in water quality management in the future. Proactive recommendations for continuous monitoring and mitigation highlight the study's pro-active approach towards sustainability. As water quality dynamics continue to evolve, compliance with regulatory standards and collaboration are critical for maintaining these critical water resources and safeguarding the health of communities and ecosystems around them. This study provides a basis for informed decision making and highlights the importance of ongoing environmental stewardship.

REFERENCES

- Dubey, R. S. (2016). Impact of urban runoff on the water quality of the Yamuna River in Delhi stretch. *EIACP Programme Centre on Hygiene, Sanitation, Sewage Treatment Systems and Technology*, 2023-11-28.

- H. V. Le and B. T. Pham, "Analysis of Water Pollution Using Different Physicochemical Parameters: A Study of Yamuna River," *Front. Environ. Sci.*, vol. 8, p. 581591, Nov. 2020.

- S. Sharma et al., "Managing water quality of River Yamuna in NCR Delhi," *ScienceDirect.com*.