Water Quality Analysis and Prediction

SUBMITTED IN PARTIAL FULFILLMENT REQUIREMENT FOR THE AWARD OF DEGREE OF

BACHELOR OF TECHNOLOGY

SUBMITTED BY:

Nihar (220381)

Divyanshu (220375)

Piyush Garg (220491)

Nirbhay (220412)

UNDER THE SUPERVISION OF:

Dr. Hirdesh Pharasi

SCHOOL OF ENGINEERING AND TECHNOLOGY

BML MUNJAL UNIVERSITY,

GURUGRAM, HARYANA

APRIL 2024

Candidate's Declaration

We hereby certify that the work on the project entitled, "Project Name –**Water Quality Analysis and Prediction**", in partial fulfillment of requirements for the award of a Degree of Bachelor of Technology in the School of Engineering and Technology at BML Munjal University, having University Roll No. 220381, 220375, 220412, 220491, is an authentic record of our work carried out during a period from March 2024 to May 2024 under the supervision of DR. HIRDESH PHARASI.

Nihar (220381)

Divyanshu (220375)

Piyush Garg (220491)

Nirbhay (220412)

Supervisor’s Declaration

This is to certify that the above statement made by the candidate is correct to the best of my knowledge.

Faculty Supervisor Name: Dr. Hirdesh Pharasi

Signature:

Acknowledgment

I am highly grateful to DR. HIRDESH PHARASI, ASSISTANT PROFESSOR, BML Munjal University, Gurugram, for supervising the Machine Learning project from March to May 2024.

DR. HIRDESH PHARASI has provided great help in carrying out my work and is acknowledged with reverential thanks. Without wise counsel and able guidance, completing the project in this manner would have been impossible.

I would like to express thanks profusely to thank DR. HIRDESH PHARASI, for stimulating me from time to time. I would also like to thank the entire team at BML Munjal University. I would also thank my friends who devoted their valuable time and helped me in all possible ways toward successful completion.

Table of Contents

Abstract

The process of attempting to analyze and predict changes in water quality parameters is known as water quality analysis and prediction. Accurate analysis regarding water quality could lead to significant improvements in managing water resources and mitigating potential hazards. In recent years, the application of sentiment analysis techniques to water quality prediction has gained traction. This report presents an analysis of water quality predictions using machine learning models based on physicochemical parameters. Data spanning from 2017 to 2021 from various monitoring stations across different states were collected and analyzed. Four key parameters - Dissolved Oxygen, pH, Biochemical Oxygen Demand (BOD), and Nitrate - were used as features to predict water quality classes. The Support Vector Classifier (SVC) machine learning model was trained and applied to the datasets to analyze and predict the quality of water at different monitoring locations. The results revealed insights into the distribution of predicted water quality classes, highlighting areas of concern for water pollution and environmental protection. The findings underscore the importance of continued monitoring and assessment of water quality for public health and environmental sustainability.

Chapter 1: INTRODUCTION

Water quality prediction is of paramount importance for ensuring the safety and sustainability of our natural resources. With data collected spanning from 2017 to 2021 from various monitoring stations across different states, this project endeavors to analyze and predict water quality levels. Four critical parameters - Dissolved Oxygen, pH, Biochemical Oxygen Demand (BOD), and Nitrate - serve as the foundational features for this predictive analysis.

Stakeholders in water management must have access to accurate predictions to safeguard public health and environmental integrity. Water quality analysts aim to ascertain the highest standards in water quality for the well-being of communities and ecosystems.

Utilizing the Support Vector Classifier (SVC) machine learning model, this study applies advanced computational techniques to the datasets, aiming to forecast water quality across different monitoring locations. By discerning patterns and trends in water quality data, this research contributes valuable insights into the distribution of predicted water quality classes.

The implications of this study extend far beyond data analysis; they underscore the urgent need for proactive measures in combating water pollution and advocating for environmental protection. By identifying areas of concern and potential risks, stakeholders can prioritize resource allocation and intervention strategies to mitigate environmental degradation and ensure sustainable water management practices.

In the subsequent sections of this report, we delve into the methodology employed, the results obtained, and the implications of our findings. Through rigorous analysis and interpretation, this study seeks to inform policy-makers, environmentalists, and the general public about the critical importance of continued monitoring and assessment of water quality for the collective well-being of society and the environment.

Chapter 2: LITERATURE REVIEW

2.1 Introduction

Water quality prediction is increasingly becoming a critical area of research due to growing concerns about environmental sustainability and public health. Traditional methods of water quality analysis, while effective, can be enhanced significantly with the integration of modern technologies like machine learning (ML). This literature review explores various studies that have applied machine learning techniques to predict water quality, focusing on the use of physicochemical parameters and the application of Support Vector Machines (SVM).

2.2 Background and Importance of Water Quality Prediction

Water quality prediction serves a pivotal role in environmental science as it helps in the proactive management of water resources, ensuring safe drinking water, and maintaining ecological balance. Studies in this field typically use parameters such as Dissolved Oxygen (DO), pH, Biochemical Oxygen Demand (BOD), and Nitrate levels, which are indicative of the overall health of water bodies .The accurate prediction of water quality assists in timely decision-making and policy formulation.

2.3 Machine Learning in Water Quality Prediction

Machine learning offers robust methodologies for predicting complex nonlinear relationships inherent in environmental data. In recent years, its application in predicting water quality has been documented extensively:

Project Name	Observationion	Result
Water Quality using Machine Learning and Fuzzy Techniques	water quality of the Ganga, a fuzzy knowledge-based method has been created . Four parameters are used as decision parameters for the prediction: dissolved oxygen (DO), biochemical oxygen demand (BOD), pH, and total coli-form.	The experimental study shows that SVM model performs better when applied to highly imbalanced dataset.
Machine Learning Approach for Predicting the Quality of Water	The model shows high accuracy through simulation testing in estimating the quality of water .In this they conducted th research study of two lakes	A tree based decision making model has been implemented to decide the level of chorophyll found in common water samples.
Water quality analysis and Prediction (project we have made)	In this project we have used several machine learning models on different parameters like dissolved oxygen(DO), pH, BOD and nitrate content.	We found out that SVC performs best on the parameters we have specified.

2.4 Comparison Study

The comparison of three water quality prediction projects highlights different methods using machine learning and fuzzy techniques. The first project predicts Ganga river's water quality using a fuzzy knowledge-based approach and SVM models. The second project focuses on two lakes, using simulation testing and a tree-based model to assess chlorophyll levels. Our project, the third one, employs various machine learning models on parameters like dissolved oxygen and pH, with the SVC model performing best. These projects collectively show how machine learning can effectively predict water quality, each with its unique approach tailored to specific needs.

Chapter 3: DATASET DESCRIPTION

3.1. Dataset

The dataset utilized has been sourced from the official website of the Government of India, specifically the "Central Pollution Control Board." This dataset encompasses six significant rivers: Ganga, Beas, Brahmaputra, Godavari, Krishna, and Satluj. Spanning a period of five years, from 2017 to 2021, this comprehensive dataset serves as a valuable resource for analyzing and understanding the water quality trends and dynamics within these vital water bodies.

Year	Number of Rows	Number of Columns
2017	202	9
2018	209	9
2019	249	9
2020	237	9
2021	263	9

- **station_code**: Unique numeric code assigned to each monitoring station location.
- **location**: The specific name of the location where the monitoring station is situated. Locations are along major rivers like Ganga, Brahmaputra, Godavari, Krishna, Beas, and Satluj.
- **state**: State in India where the monitoring location is situated.
- **DissolvedOxygen**: Measure of the concentration of dissolved oxygen gas in the water body (in mg/L).
- **pH**: Measures the acidity or basicity of the water on a scale of 0-14.
- **BOD**: Biochemical Oxygen Demand is a measure of the amount of dissolved oxygen required/consumed by microorganisms to break down organic matter present in the water. Higher values indicate more organic pollution load.
- **Nitrate**: Measures the concentration of nitrate in the water (in mg/L). Nitrates can come from agricultural runoff, sewage, etc.
- **Latitude**: The latitude coordinate of the monitoring location.
- **Longitude**: The longitude coordinate of the monitoring location.
- **year**: Year in which the measurements were taken.

In addition to the primary water quality monitoring data, an auxiliary dataset has been furnished, consisting of 1,755 rows of observations and 5 columns of variables. The inclusion of this secondary dataset serves the purpose of enriching the training process for the development of predictive models pertinent to water quality assessment. By integrating this supplementary data, we aim to enhance the depth and accuracy of our analytical endeavors in understanding and managing water quality dynamics.

3.2. Exploratory Data Analysis and Visualizations

Exploratory data analysis (EDA) plays a crucial role in understanding the characteristics of the dataset and gaining insights that can inform the feature engineering and modeling processes. In this section, we present a detailed analysis of the dataset, including visualizations and potential feature transformations.

Based on the EDA, several key insights were obtained:

	DissolvedOxygen	pH	BOD	Nitrate
count	1754.000000	1754.000000	1754.000000	1754.000000
mean	6.499177	7.631383	3.086884	1.905298
std	2.073213	0.562628	6.038215	7.150219
min	0.000000	5.350000	0.000000	0.000000
25%	5.183945	7.400000	1.100000	0.586036
50%	6.550000	7.738659	2.099524	1.058674
75%	8.200000	7.995308	3.250000	2.524427
max	10.900000	8.974343	84.000000	290.100000

Figure 1: Statistics of Water Quality Parameters

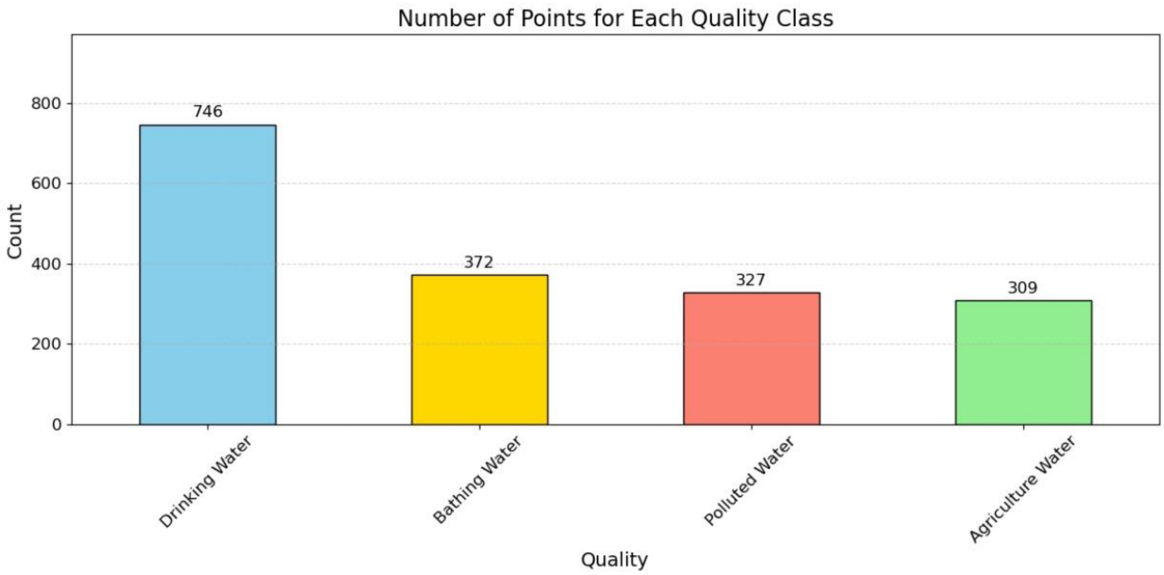


Figure 2: Data points of auxiliary dataset

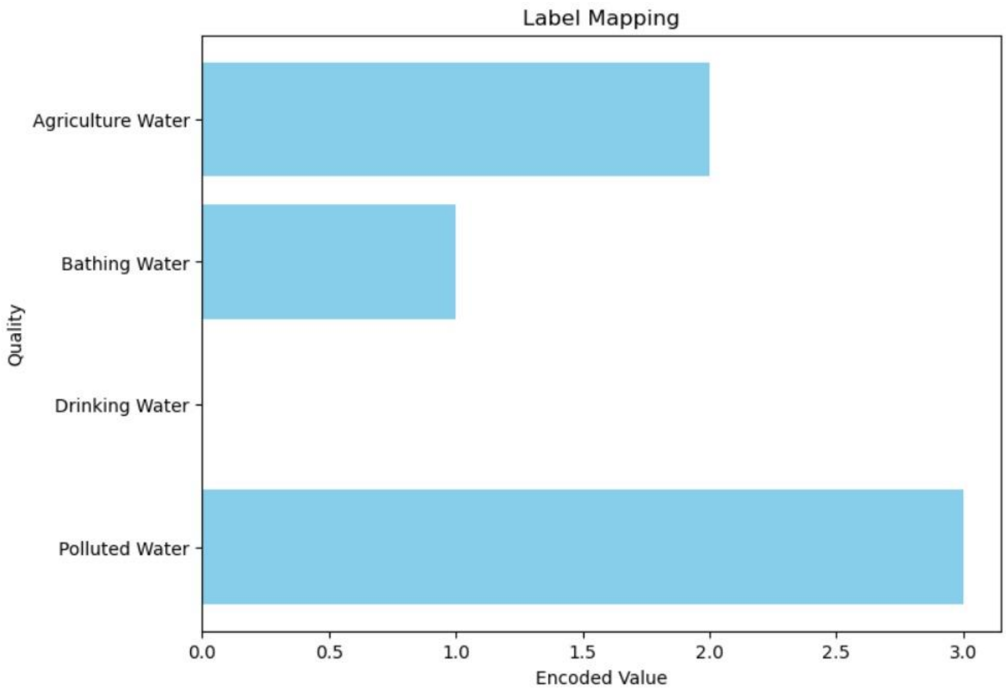


Figure 3: Encoded Values

Chapter 4: METHODOLOGY

1. **Data Loading and Exploration**:
   - The code loads water quality data from an Excel file (`water_quality.xlsx`) using pandas.
   - It performs basic data exploration, such as displaying a sample of the data, checking the shape (number of rows and columns), obtaining information about the data types, and calculating descriptive statistics.
   - It also checks for duplicate entries and counts the number of instances for each class in the 'Quality' column.
2. **Data Visualization**:
   - The code creates a bar plot to visualize the distribution of water quality classes ('Quality') using matplotlib and seaborn libraries.
   - It also generates a heatmap to show the correlation between different water quality parameters.
3. **Data Preprocessing**:
   - Manual label encoding is performed on the 'Quality' column, mapping string labels to numerical values.
   - The code visualizes the label mapping for better understanding.
4. **Feature Selection and Split**:
   - The features 'DissolvedOxygen', 'pH', 'BOD', and 'Nitrate' are selected as predictors (X), while 'Quality' is chosen as the target variable (y).
   - The data is split into training and testing sets using `train_test_split` from scikit-learn.
5. **Feature Scaling**:
   - The predictor variables are scaled using `StandardScaler` from scikit-learn to ensure consistent scales across features.
6. **Model Building and Evaluation**:
   - The code defines a function `model_building` to fit various machine learning models, make predictions, and calculate evaluation metrics like accuracy score and classification report.
   - Several models are trained and evaluated, including Logistic Regression, Random Forest Classifier, Decision Tree Classifier, and Support Vector Classifier (SVC).
   - The models' performance is evaluated using accuracy scores and classification reports.
   - Confusion matrices are generated for each model and visualized using heatmaps.

- ## Random Forest

Applications for classification and regression commonly use supervised machine learning techniques like random forest. When doing regression on different data, it builds decision trees and utilises their average for categorization and majority vote for voting. There are two models: the Random Forest Classifier, an ensemble technique in which the model is built from several little decision trees, or estimators, each of which generates a separate set of predictions. A more precise forecast is made by combining all of the estimators' estimations rather than just one tree. Using supervised learning, the Random Forest Regressor employs ensemble learning techniques for regression. The ensemble prediction of a machine learning algorithm is created by merging the predictions of many machine learning algorithms. By combining predictions from various algorithms, it generates predictions that are more accurate than those from a single model.

- **Logistic Regression**

Supervised machine learning techniques commonly utilize logistic regression for classification tasks. It models the probability of a binary outcome by fitting data to a logistic curve. Logistic regression operates by estimating coefficients for each feature and applying them to a logistic function to predict the probability of an event. Unlike random forest, which utilizes an ensemble of decision trees, logistic regression is a single-model approach. It calculates the probability of a binary outcome using a linear combination of features and applies a logistic function to map this result to a range between 0 and 1, representing the probability of belonging to a particular class.

- **Decision Tree Classifier**

Classification and regression tasks frequently employ supervised machine learning methods like decision trees. Decision Tree Classifier, akin to Random Forest, operates by constructing decision trees on diverse data sets. Each tree provides its own predictions, which are then combined to produce a more accurate forecast. Similarly, Decision Tree Regressor utilizes ensemble learning techniques to generate regression predictions. By amalgamating predictions from multiple decision trees, it achieves greater accuracy compared to individual models.

- **Support Vector Classifier (SVC)**

Supervised machine learning commonly employs support vector classifier (SVC) for classification tasks. SVC constructs decision boundaries based on training data points to classify new instances. It separates classes by finding the hyperplane that maximizes the margin between them. The SVC model utilizes a subset of training data points, known as support vectors, to define the decision boundary. It aims to maximize the margin while minimizing classification errors. By using supervised learning, SVC ensures accurate classification by finding the optimal hyperplane that best separates different classes in the feature space.

7. **Model Selection and Testing**:
    - The SVC model is selected, and its performance is further evaluated by making predictions on separate datasets for the years 2017, 2018, 2019, 2020, and 2021.
    - The predictions are added to each dataset as a new column ('Predicted_Quality'), and the distribution of predicted water quality classes is analyzed.
8. **Visualization on Maps**:
    - For each year (2017-2021), the code creates an interactive map of India using the folium library.
    - The monitoring locations are plotted on the map as markers, with colors representing different predicted water quality classes.
    - Marker clusters are used to group nearby markers for better visualization.
9. **COVID-19 Period Analysis**:
    - The code filters the data for the years 2019, 2020, and 2021, representing the COVID-19 period.
    - It creates a count plot using seaborn to visualize the distribution of water quality classes during this period, grouped by year.

The methodology covers data loading, exploration, visualization, pre-processing, feature selection, model building, evaluation, prediction, and spatial visualization on maps. Additionally, it includes a specific analysis of the water quality distribution during the COVID-19 period (2019-2021).

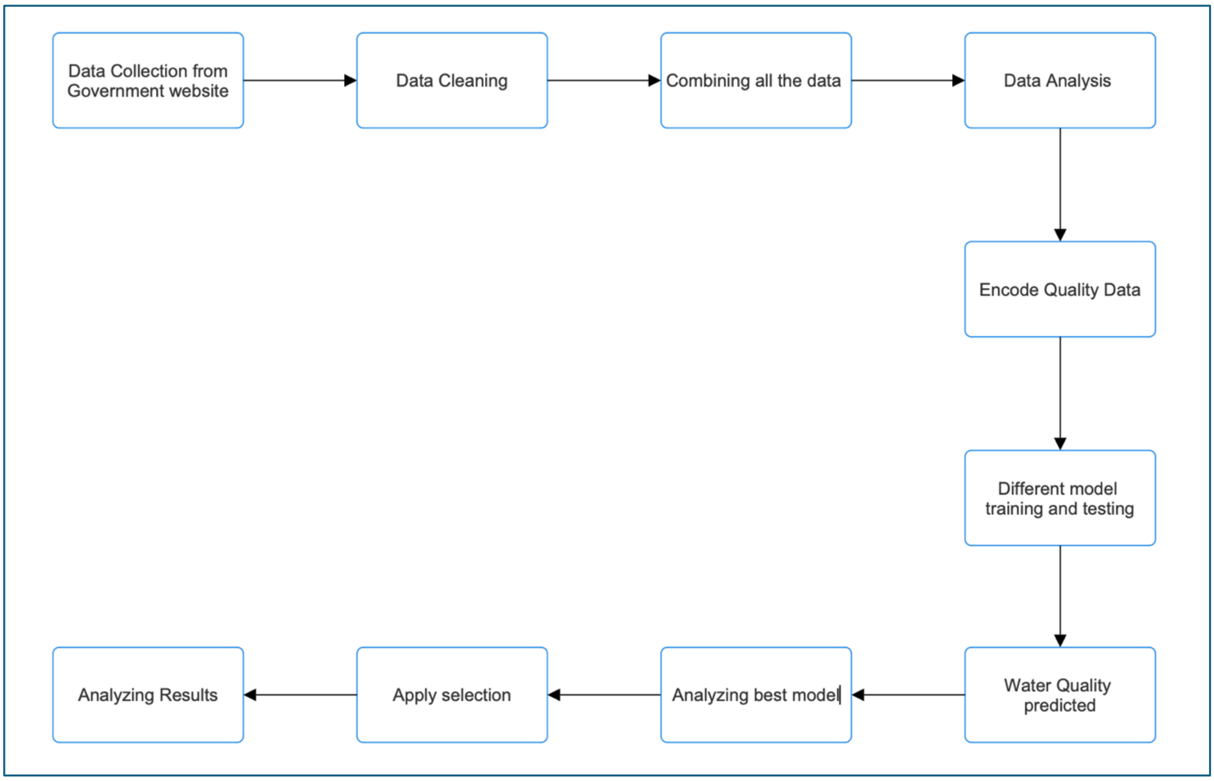
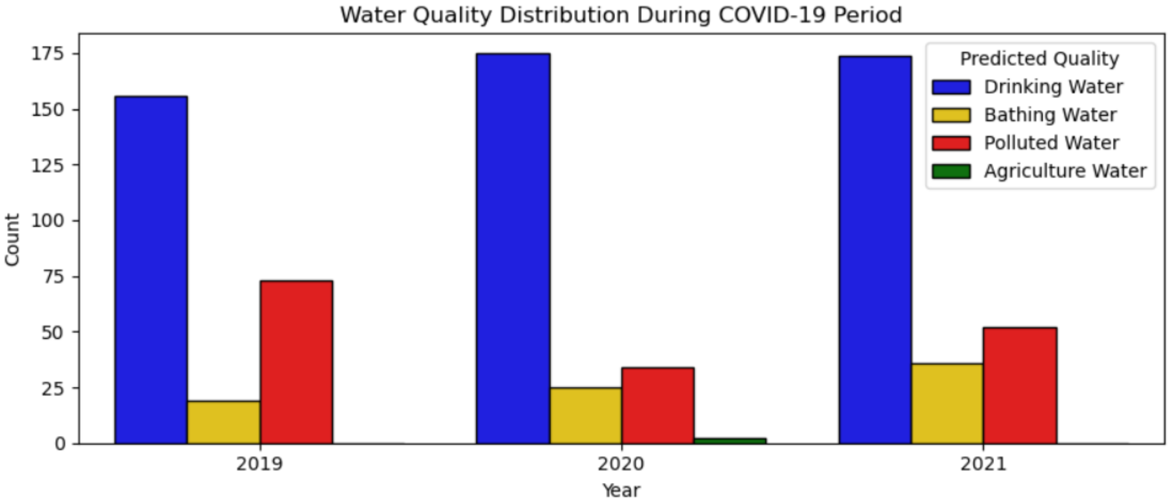
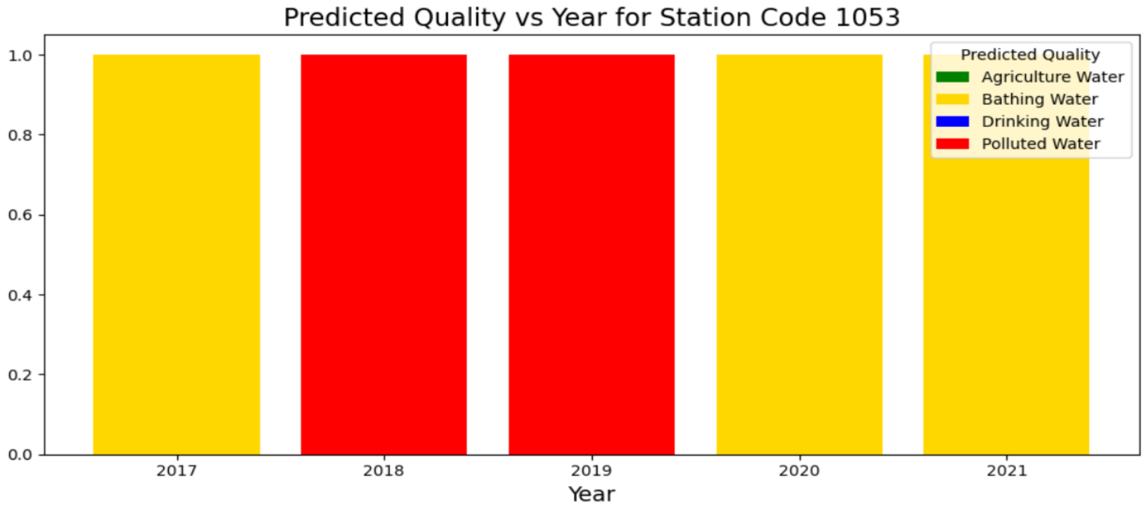


Figure 4:Flowchart showcasing methodology implemented

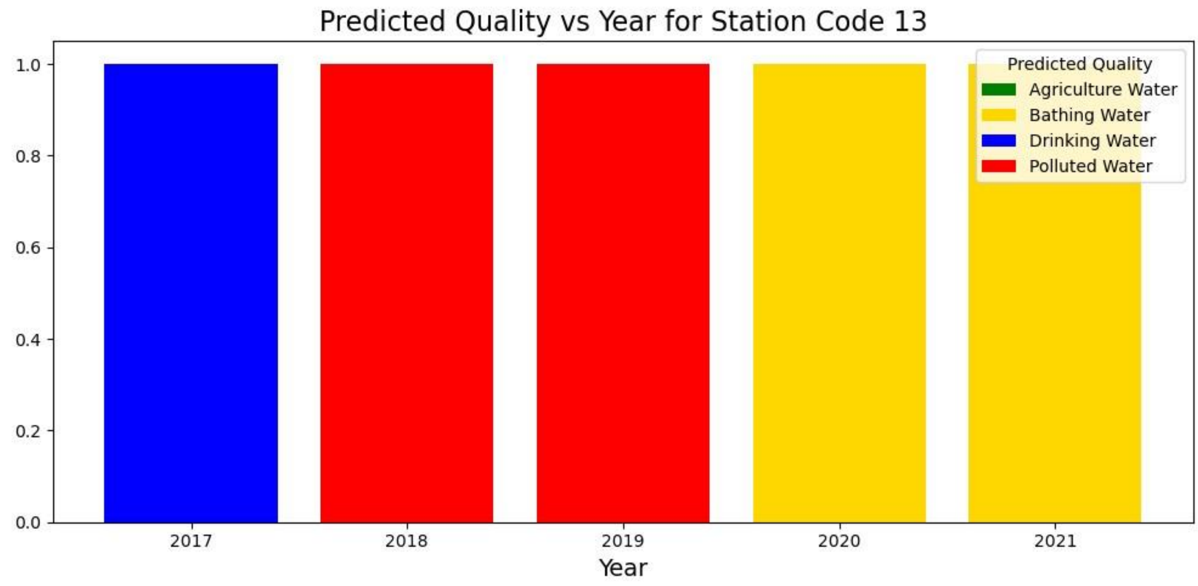
Chapter 5: Analysis of Fluctuation of Water Quality



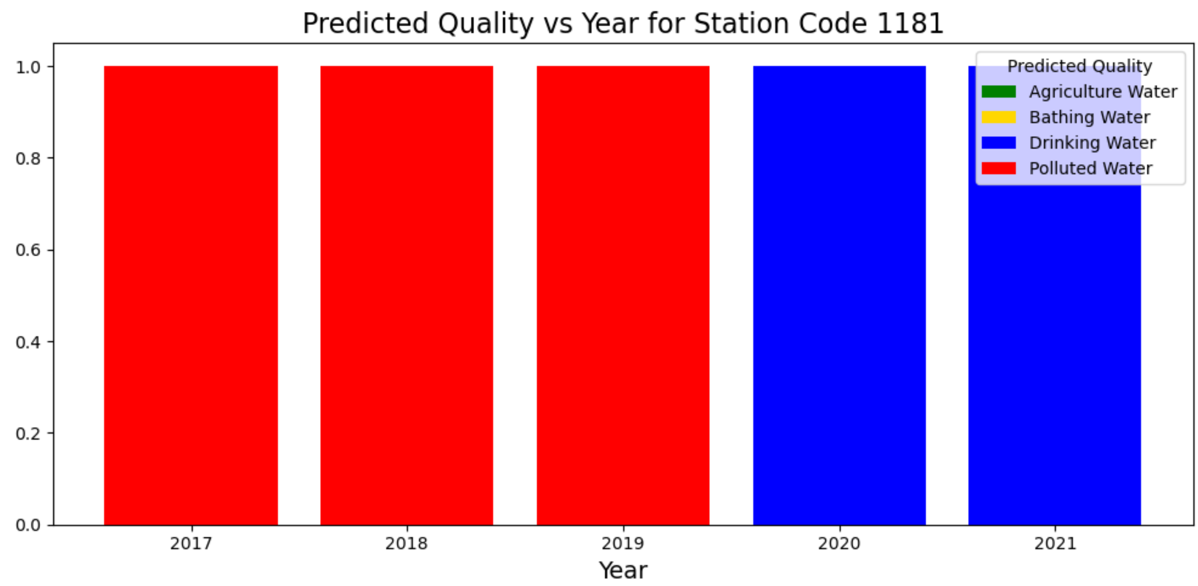
The fluctuation in water quality observed over the years can be attributed to the varying levels of human activities influenced by COVID-19 lockdown measures. The absence of lockdown in 2019 led to increased pollution, while the lockdown in 2020 resulted in improved water quality. The relaxation of lockdown measures in 2021 potentially contributed to the deterioration of water quality once again.



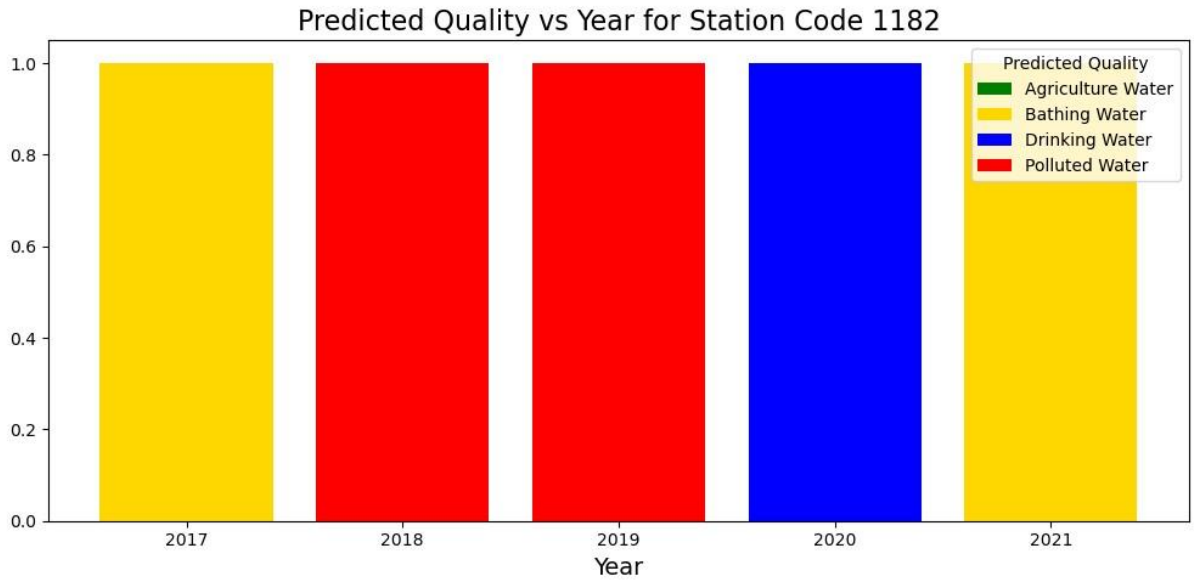
The fluctuation in water quality at station 1053, GANGA AT DAKSHMINESHWAR, KOLKATA, WEST BENGAL, observed as suitable for bathing and domestic use in 2017, deteriorating to polluted levels in 2018 and 2019, before returning to suitability for bathing in 2020 and 2021, could be influenced by various factors. Increased industrial activities near the water source may have contributed to the pollution observed in 2018 and 2019. The discharge of untreated or partially treated industrial effluents containing pollutants such as heavy metals, chemicals, and organic compounds can significantly degrade water quality. Additionally, rapid urbanization and associated infrastructure development can lead to increased runoff containing pollutants like oils, chemicals, and debris from roads and construction sites, which might have further contributed to the decline in water quality.



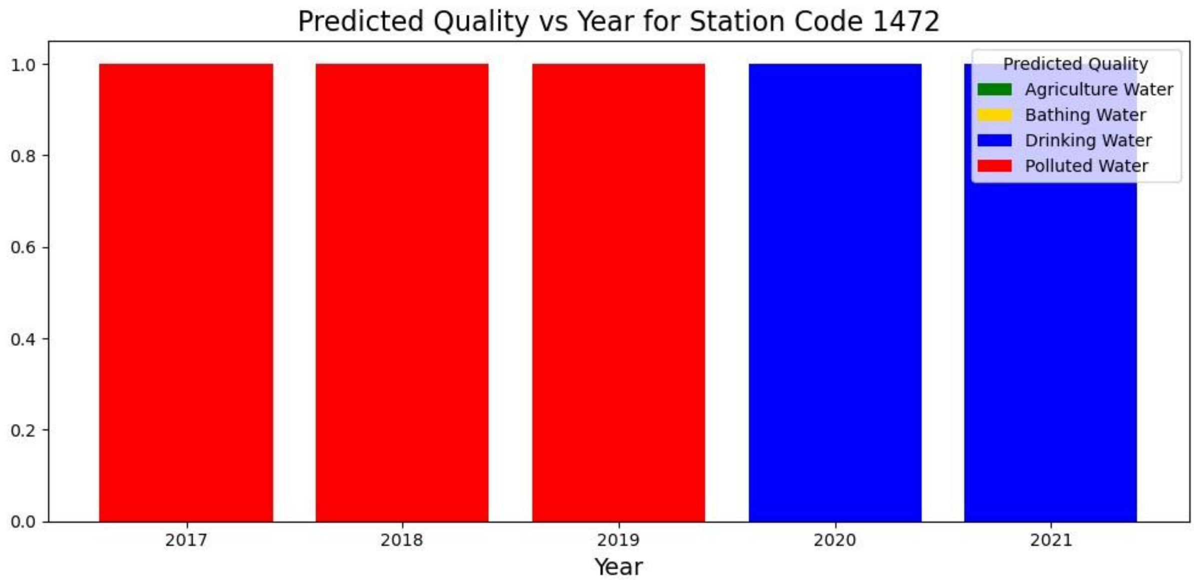
The variability in water quality at station 13 transitioning from polluted to suitable for bathing, then further improving to drinking water quality in subsequent years before declining again, is likely influenced by several factors, with the presence of nearby coal mines being a significant contributor. The proximity of these mines suggests a potential for contamination of water bodies due to runoff carrying heavy metals, sedimentation, and leaching of harmful chemicals used in mining operations. The initial pollution observed may be attributed to increased mining activities or inadequate waste disposal practices, resulting in heightened levels of pollutants in the water.



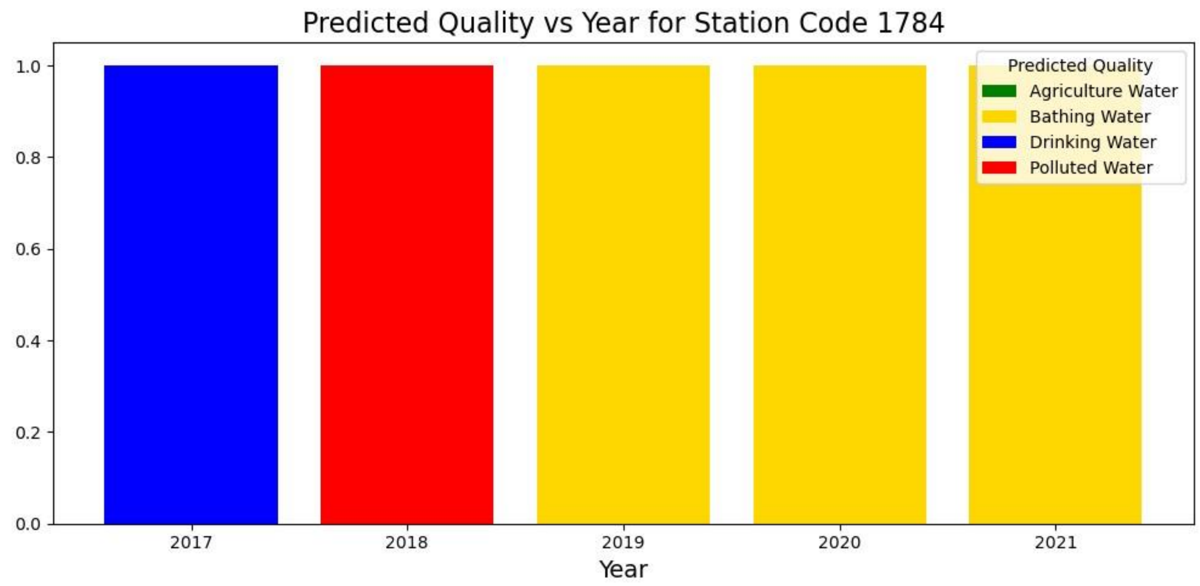
Analyzing the water quality predictions for station code 1181 spanning from 2017 to 2020 reveals a consistent pattern of water being predicted as polluted from 2017 to 2019, followed by an improvement to drinking water quality in 2020. Industrial Pollution emerges as a prominent factor influencing these fluctuations in water quality. The presence of industries near the water source likely played a pivotal role in exacerbating pollution levels during 2017, 2018, and 2019. Industrial effluents, laden with pollutants such as heavy metals, chemicals, and organic matter, might have seeped into the water, consequently leading to the prediction of polluted water quality during those years.



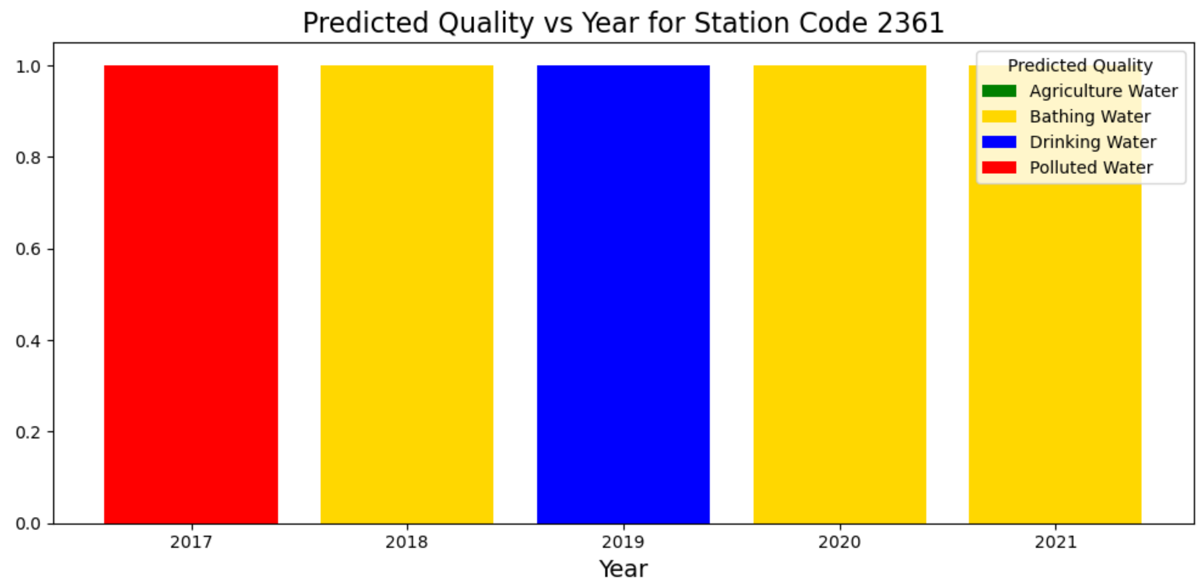
Analyzing the water quality predictions at station code 1182, situated at the Krishna river upstream of Ugarkhurd, several factors may have influenced the variations in water quality predictions over the years. One significant factor to consider is Agricultural Runoff. If there are agricultural activities upstream of the station, runoff containing fertilizers, pesticides, and other agricultural chemicals could flow into the river, impacting its water quality. This could contribute to the prediction of polluted water quality, particularly during years with heavy rainfall events, as runoff carrying agricultural pollutants is more likely to enter the river during such periods.



Analyzing the water quality predictions for station code 1472, located at the Ganga river in Serampur, from 2017 to 2020, reveals a consistent trend of water being predicted as polluted from 2017 to 2019, followed by an improvement to drinking water quality in 2020. A major contributing factor to these fluctuations in water quality is Industrial Pollution. The presence of industries near the water source could have significantly impacted pollution levels in 2017, 2018, and 2019. Industrial effluents containing pollutants such as heavy metals, chemicals, and organic matter might have contaminated the water, leading to predictions of polluted water quality during those years.



Analyzing the water quality trends at station code 1784 Krishna Thangadi reveals fluctuations over the years, with water initially being suitable for drinking, then deteriorating to polluted levels in 2018, followed by an improvement to being suitable for bathing and daily purposes in 2019, 2020, and 2021. A major factor that contributed to the fluctuation in water quality is Agricultural Activities. Initially, the water quality might have been good due to minimal agricultural pollution. However, increased industrial activities or intensified agricultural practices in the vicinity could have led to the deterioration of water quality in 2018. Discharge of industrial effluents, agricultural runoff containing pesticides and fertilizers, and improper waste disposal could have contributed to pollution levels.



The variability in water quality at station 2361 (GODAVARI AT MANCHERIAL, NEAR RLY BDG B/C OF RALLAVAGU), transitioning from polluted to suitable for bathing, then further improving to drinking water quality in subsequent years before declining again, is likely influenced by several factors, with the presence of nearby coal mines being a significant contributor. The proximity of these mines suggests a potential for contamination of water bodies due to runoff carrying heavy metals, sedimentation, and leaching of harmful chemicals used in mining operations. The initial pollution observed may be attributed to increased mining activities or inadequate waste disposal practices, resulting in heightened levels of pollutants in the water.

Chapter 6: Results

6.1. Model/Algorithm Performance

We conducted a comprehensive analysis of water quality prediction using several classification models: Logistic Regression, Random Forest Classifier, Decision Tree Classifier, and Support Vector Classifier (SVC). Each model was evaluated based on key performance metrics including accuracy, precision, recall, and F1-score.
Here's a detailed comparison of the results obtained from each model:
Table 6.1: Scores obtained for each of these models

Model	Accuracy	Precision	Recall	F1-score
Logistic Regression	90%	0.91	0.90	0.90
Random Forest Classifier	91%	0.93	0.91	0.91
Decision Tree Classifier	82%	0.88	0.82	0.81
Support Vector Classifier	94%	0.95	0.94	0.95

6.2. Comparative Analysis

Among these models, the Support Vector Classifier (SVC) demonstrated the highest accuracy of 94% and achieved the highest precision, recall, and F1-score. Therefore, we selected the SVC model as our final predictor for water quality.

Reasons for Fluctuation in Water Quality:

The analysis of water quality fluctuations revealed several contributing factors:

1. **COVID-19 Pandemic:** During the COVID-19 pandemic, reduced industrial activities and human movements might have positively impacted water quality due to decreased pollution levels.
2. **Industrial Pollution:** Industrial effluents containing pollutants such as heavy metals, chemicals, and organic matter can contaminate water sources, leading to predictions of polluted water quality.
3. **Agricultural Waste:** Runoff from agricultural fields carrying fertilizers, pesticides, and other chemicals can enter water bodies, affecting water quality.
4. **Natural Factors:** Environmental factors such as rainfall, temperature variations, and seasonal changes can also influence water quality fluctuations.

The study highlights the importance of employing machine learning models for predicting water quality and understanding the underlying factors contributing to its fluctuations. By leveraging advanced analytics, policymakers and environmental agencies can make informed decisions to mitigate pollution and ensure sustainable management of water resources.

Chapter 7: Conclusion and Future Scope

7.1. Conclusion

This study has successfully demonstrated the application of machine learning techniques, specifically the Support Vector Classifier (SVC) model, for accurate prediction of water quality using physicochemical parameters. The analysis of water quality data from various monitoring stations across six major Indian rivers from 2017 to 2021 has provided valuable insights into fluctuations in water quality over time.

The SVC model exhibited superior performance, achieving an accuracy of 94% and outperforming other models such as Logistic Regression, Random Forest Classifier, and Decision Tree Classifier. The high precision, recall, and F1-score obtained further validate the effectiveness of the SVC model in predicting water quality classes.

Through this study, several factors contributing to fluctuations in water quality were identified, including the COVID-19 pandemic, industrial pollution, agricultural waste, and natural factors. The analysis highlighted the positive impact of reduced human activities during the COVID-19 lockdown on water quality, as well as the detrimental effects of industrial effluents, agricultural runoff, and environmental factors on water pollution levels.

The findings of this research underscore the importance of continuous monitoring and assessment of water quality, as well as the implementation of effective strategies to mitigate pollution and ensure sustainable water resource management. The application of machine learning models, such as the SVC, can provide valuable insights and support informed decision-making processes for policymakers and environmental agencies.

7.2. Future Scope

While this study has made significant strides in water quality prediction, there are several avenues for future research and improvements:

1. **Incorporation of Additional Parameters**: Expanding the dataset to include more physicochemical parameters, such as heavy metal concentrations, total dissolved solids, and microbial indicators, could enhance the predictive capabilities of the machine learning models.
2. **Integration of Remote Sensing Data**: Combining water quality data with remote sensing imagery and other geospatial data could provide additional contextual information and improve the understanding of the factors influencing water quality fluctuations.
3. **Real-time Monitoring and Prediction**: Developing a real-time monitoring and prediction system by integrating machine learning models with IoT (Internet of Things) devices and sensor networks could enable continuous monitoring and timely interventions to address water quality issues.
4. **Advanced Machine Learning Techniques**: Exploring the application of more advanced machine learning techniques, such as deep learning and ensemble methods, could potentially improve the accuracy and robustness of water quality predictions.
5. **Socioeconomic Impact Analysis**: Conducting socioeconomic impact analyses to assess the effects of water quality fluctuations on public health, agriculture, and other sectors could provide valuable insights for policymakers and stakeholders.

6. **Collaborative Research and Data Sharing**: Fostering collaborative research efforts and promoting data sharing among researchers, institutions, and government agencies could facilitate the development of more comprehensive and generalizable water quality prediction models.

By addressing these prospects, researchers and policymakers can further enhance our understanding of water quality dynamics, enabling more effective decision-making processes and contributing to the sustainable management of water resources for the benefit of both human populations and ecosystems.

References

- https://www.researchgate.net/publication/374163634_Water_Quality_Assessment_Through_Predictive_Machine_Learning
- https://www.researchgate.net/publication/371860089_Assessment_of_Water_Quality_using_Machine_Learning_and_Fuzzy_Techniques
- https://www.researchgate.net/publication/340686027_Machine_Learning_Approach_for_Predicting_the_Quality_of_Water
- https://cpcb.nic.in/index.php
- Reza, M. M., & Hasan, M. K. (2021). Machine Learning Applications for Water Quality Analysis and Prediction. CRC Press. https://doi.org/10.1201/9781003156284
- Sadiku, M. N. O., Musa, S. M., & Akpez, O. P. (2018). Environmental Modeling and Monitoring for Human Health. CRC Press. https://doi.org/10.1201/9781315152899