Print ISSN:2645-8500
DOI: https://doi.org/10.3126/mefc.v8i1.60477

Water Quality Monitoring of River Ganga Using Non-Linear Data Analytics

B.D.K. Patro[1], Shivam Sharma[2], Abhishek Bajpai[3]

[1]Associate Professor, Department of CSE, Rajkiya Engineering College, Kannauj

[2]Department of CSE, Rajkiya Engineering College, Kannauj

[3]Assistant Professor, Department of CSE, Rajkiya Engineering College, Kannauj

Corresponding Author Email: bdkpatro@reck.ac.in

Abstract

The Ganga River is one of India's biggest and most significant rivers, and the health and welfare of millions of people depend on the purity of its water. The traditional linear models that have been used extensively to assess water quality have limitations in their ability to capture the intricate non-linear interactions between the water quality factors. On the other hand, non-linear data analytics are able to identify these linkages and can offer more precise and trustworthy estimates of water quality. In order to monitor the River Ganga's water quality, this study suggests a non-linear data analytics approach that entails gathering and studying a significant amount of water quality data. The results show that the proposed approach outperforms traditional linear models and can provide valuable insights into the water quality of the River Ganga.

Keywords: Non-Linear, analytics, Ganga, water, quality, river

Introduction

The Ganges River, originating from Gaumukh in the Himalayas and flowing through to the Ganges Delta in the Bay of Bengal, is not only India's most sacred river but also one of its most vital lifelines. It spans approximately 2510 kilo meters, passing through eleven states and providing water to over 40 percent of the nation's population. The Ganges valley hosts an extensive canal system, predominantly in Uttar Pradesh and Bihar, contributing significantly to agricultural productivity. This basin irrigates more than 140 million acres of land, which accounts for a substantial portion of India's GDP.

The Ganges River holds immense cultural and ecological significance. It supports a rich and diverse ecosystem, providing habitat to over 150 different animal and aquatic species. However, despite its spiritual and ecological importance, the Ganges has been grappling with severe water quality issues and is regrettably ranked among the world's most polluted rivers. In the realm of water quality monitoring and management, considerable efforts have been made in the context of the Ganges River. Various governmental and non-governmental organizations, along with international partnerships, have been actively involved in endeavors to restore and protect the river's water quality.Existing studies have shed light on the extent of pollution in the Ganges, pinpointing sources of contamination and

highlighting the urgency of addressing this crisis. Researchers have also conducted assessments of the river's health, measuring parameters such as pH, dissolved oxygen (DO), and pollutant concentrations (Bureau of Indian Standards, 2015). Additionally, ongoing initiatives, such as the Clean Ganga National Mission and the Second National Ganga River Basin Project, have been launched to mitigate pollution and promote sustainable water management in the Ganges basin (Rana & et al., 2022). However, despite these efforts, there remain gaps in the existing literature. First and foremost, while previous studies have identified the problems, there is still a need for comprehensive and real-time monitoring of critical water quality parameters to effectively manage and address pollution in the Ganges. Secondly, there is limited research that employs advanced non-linear data analytics techniques to gain deeper insights into the complex dynamics of water quality in the Ganges River. This study aims to bridge these gaps by utilizing cutting-edge sensor technology and non-linear data analytics to provide a more holistic understanding of water quality in the Ganges and offer actionable insights for its preservation. To achieve this, we have collected water samples from various points along the river and employed a range of sensors, including pH Sensors, ORP Sensors, DO Sensors, Water Sensors, Soil Moisture Sensors, Dust Sensors, and Temperature and Humidity Sensors.

Literature Survey

A research was conducted on quality of water in river Ganga on 3 sites in Rishikesh. In the study, it was observed that the turbidity value was above WHO’s permissible standards. This was attributed due to pollution from the organic matter, heavy rainfall and runoff. The water quality index (WQI) on some sites goes as high as 1714.6 in the monsoons i.e. July. While the highest 12.10 mg/L of dissolved oxygen was found to be in winter season (Jan 2007) with a minimum of 7.14 mg/L in monsoon during July (Chauhan, A., &etl2008). Another study was performed for measuring the quality of water using machine learning and IOT sensors. The model contained multiple sensors connected to a Node MCU to collect different parameters and analyse these parameters. In the study the standard values were taken to be as shown in the table 1 (Ashwini & et al., 2019).

Table 1

Standard values of water quality

S. No.	Parameters	Values
1.	pH	6.50 – 8.0
2.	Temperature	140 F
3.	Turbidity	5.0
4.	Dissolved Oxygen	5.0 mg/L
5.	Conductivity	0.05 S/cm

6.	Total Organic Carbon	55 mg/L
7.	Colour	Colourless

A study on water quality indices was conducted in Haridwar, Uttarakhand. Assessment on quality of water inGanga was made on the basis of 15 quality parameters for a period of 11 years. According to the River Ganga WQI, the research area's water quality was between good and medium. The river has a decent WQI according to the NSF, but according to the weighted Arithmetic approach, the river's water quality was subpar. Thus, it can be inferred that the 11-year study period saw a range in the water quality of the River Ganga from poor to good, which was consistent with other studies on the river's WQI. While bearing in mind the growing urbanisation and pollution loading of rivers, the proper actions should be taken to minimise future contamination loads from entering the river. The study found that the main causes of pollution were sewage, solid and liquid waste pollutants, or organic nature (Bhutiani, R. et al, 2016).

In another study of classification, the Chao Phraya River's numerous water parameters were identified and categorised in the research using river sensing processing actuators (rSPA), and a method for analysing water quality in Thailand's most important water resource was demonstrated. An experimental investigation on the conductivity, salinity, and total dissolved solid pollutants was conducted in the Chao Phraya River that had accumulated downstream (TDS). In particular for public consideration, the results have conveyed evaluations as comprehensible visuals. rSPA is regarded as an efficient method for implementing alert system in river's water quality studies (Veesommai, C. et al., 2015).

In another study in Bangladesh, they have proposed a Wireless Sensor Network (WSN). They showed the pH, temperature, turbidity, and ORP values that were as a result felt. The parameter values are continuously sensed, and the results are displayed in real-time on the LCD, PC, or mobile device. When the obtained value exceeds threshold value, "BAD" message is displayed in the interface (Salah, M. et al,2019). A system was proposed in a study where an early warning system was proposed for the ungauged river basins of china. The system was named mobile device early warning system (MEWSUB). The system provides a web services model and browser-based interface and makes use of risk response personnel and can use it any place, even if we have limited data (Wang et al., 2015). At Kumaon district of Uttarakhand, a correlation in water quality parameters and water quality index (WQI). They discovered that the majority of the results from the examined water samples fell within the Bureau of Indian Standards' (BIS) recommended ranges for physico-chemical parameters. However, microbial pollution is the primary factor that has impacted the water quality in this region. In 96 villages, the WQI index based on the physicochemical

and biological parameters was very low, but in 21 villages, it was excellent or good quality (Kothari et al, 2020).

This study uses real-time data collected from the Internet of Things (IoT) model and applies the Support Vector Machine (SVM) algorithm, which achieves an impressive accuracy of approximately 94%. When employing 7 distinct parameters, SVM demonstrates the best performance compared to other models. In addition to SVM, the study also explores the use of Decision Tree and Artificial Neural Network (ANN) models togain further insights and analysis from the data. (Jalal et al, 2019).

SWMS (Smart Water Monitoring System) has been designed to monitor water quality and usage in real time. A smart quality and quantity metre has been devised to check the purity of potable water and record the consumption of water. Parameters that are taken into account include pH, conductivity, turbidity, and temperature. For remote access to data, an online monitoring system is used. A notification system is also devised to notify the customer and the authorities. The Raspberry PiTM is used to determine water potability, and after water consumption is checked, a bill is generated using a three-slab system (Jha et al., 2018).

Integration of SCADA and IoT has been implemented in the Tiruneveli Corporation, Tamil Nadu, for monitoring the quality of water in real time. Parameters added to the system include turbidity, color, and temperature. The GSM module is used for the generation, collection, transfer, and storage of sensor data over the web browser. This system performs more efficiently and shows better results than existing systems that produce better results (Saravanan et al., 2018).

Another study describes that the River Ganges, a culturally and ecologically significant waterbody in India, faces serious water quality challenges. To assess and manage these challenges, researchers rely on a multidimensional approach, evaluating various physicochemical and microbiological parameters. These parameters include Dissolved Oxygen (DO), Biological Oxygen Demand (BOD), and Chemical Oxygen Demand (COD), along with temperature, pH, conductivity, and specific ions. However, the complexity of water quality data often necessitates multivariate statistical techniques like Principal Component Analysis (PCA) to streamline analysis and identify crucial variables. PCA has found successful application in various water quality studies, such as groundwater analysis, sediment contamination assessment, and nutrient gradient investigations. This study aims to apply PCA to the River Ganges, revealing the primary factors influencing its water quality. In summary, understanding and managing water quality in the River Ganges are essential due to its cultural and ecological significance, with PCA offering a valuable tool for data interpretation (Mishra et al., 2010).

A study on the Ganga River and its tributaries in the upstream regions of India revealed significant increases in silicate values at multiple locations (R1, R2, and R5), indicating a rise in mean silicate concentrations over the years but a decrease in higher

values. Additionally, a noteworthy increasing trend was observed for SAR, K+, and Na+ at most sites, suggesting an overall deterioration in water quality. The study also highlighted a significant upward trend in sulfate (SO2-4) concentration across all sites, except for a decreasing trend in upper values at R2. Furthermore, the analysis revealed varying trends in other water quality parameters, including magnesium (Mg+2), total alkalinity (TA), pH, calcium hardness (CH), and total hardness (TH), at different locations. Notably, the study emphasized the importance of addressing the direct input of water runoff from rock weathering areas as a potential threat to water quality. Overall, this research provides critical insights into the changing water quality dynamics in the Ganga River and its tributaries, emphasizing the need for strategic conservation and management efforts to preserve this vital water resource. (Kumar et al., 2021).

Another study conducted an extensive assessment of water quality in the Ganga River, focusing on nine sampling locations along the river's stretch from Haridwar to Garhmukteshwar during both post-monsoon and pre-monsoon periods in 2014-15. The findings indicate that the river's water quality is not suitable for direct drinking due to significant pollution. Water quality indices, including NSFWQI and CPI, revealed severe pollution levels in both seasons, with a more pronounced deterioration observed in the pre-monsoon months, possibly due to reduced dilution and increased untreated wastewater discharge. The study also highlighted the potential health risks associated with heavy metal contamination in the river. Recommendations were made for proper water treatment before consumption. Moreover, the study emphasized the importance of precise water quality assessment using comprehensive indices like NSFWQI and CPI, which incorporate multiple physicochemical parameters. These results provide valuable insights for researchers, water quality monitoring authorities, and policymakers involved in Ganga River conservation efforts. (Chaudhary et al., 2017, pp 53-60).

In the study of Water Quality Monitoring in the River Ganga, several key studies in the field of Non-Linear Data Analytics and water quality assessment can be noted. A seminal work by Maier and Dandy (2000) highlighted the early potential of AI models for water quality prediction, setting a foundation for subsequent research. Chau (2006) emphasized the versatility of AI models, albeit primarily focusing on coastal water quality, while Solomatine and Ostfeld (2008) addressed data-driven models in river basin management, albeit with a limited scope predating 2008. Nicklow et al. (2010) explored evolutionary algorithms for water supply and quality systems, offering insights into planning and management. Raghavendra and Deka (2014) delved into the application of support vector machines (SVM) for groundwater quality modelling, showcasing SVM's potential in hydrology. These studies collectively provide valuable insights into the evolving landscape of AI-driven water quality research, serving as a foundation for the current research on the River Ganga's water quality using Non-Linear Data Analytics (Tiyasha et al, 2020).

In the context of Water Quality Monitoring for the River Ganga and the application of Non-Linear Data Analytics, prior research reveals a dual approach. Some methodologies offer the potential to enhance monitoring and decision-making. In the context of the River Ganga, leveraging Non-Linear Data Analytics can further advance our understanding and management of water quality issues (Faculty of Computer and Mathematical Sciences, UniversitiTeknologi MARA Cawangan Kelantan, Malaysia).

In another study, researchers have applied Support Vector Machines (SVM) to predict Dissolved Oxygen (DO) levels in lakes, emphasizing the importance of factors like pH, temperature, and conductivity in achieving accurate predictions. SVM demonstrated robustness and precision in water quality predictions. Additionally, a constructed wetland study in Malaysia showcased SVM's performance in predicting Water Quality Index (WQI), outperforming neural networks and simplifying calculations. Furthermore, in the Kinta River of Malaysia, an Artificial Neural Network (ANN) approach using multi-layer perceptron (MLP) was found to be highly effective for predicting WQI. This comprehensive and reliable technique encourages the use of ANN-based approaches in the field. A comparative study of ANN algorithms for water quality prediction in the River Ganga highlighted the effectiveness of supervised learning techniques, emphasizing the use of MLP backpropagation and gradient descent adaptive algorithms. These studies collectively underscore the potential of Non-Linear Data Analytics for water quality monitoring in the context of the River Ganga (Giri, A. et al, 2014).

Another study having a model, QUAL2E, has been widely used for evaluating stream water quality parameters. It has demonstrated its effectiveness in different settings and is considered a valuable tool for regulatory and policy decision-making. The study focuses on applying QUAL2E to assess water quality in the Yamuna River, particularly for dissolved oxygen (DO) and biological oxygen demand (BOD) during low-flow conditions. It also incorporates uncertainty analysis and explores the impact of remedial measures on water quality. This comprehensive analysis utilizes QUAL2E to predict water quality and visualize the effects of different measures to improve the river's water quality in an easily understandable format. The study area encompasses a critical 25 km stretch from Wazirabad to the Okhla barrage, heavily impacted by industrial and sewage discharges. The river receives substantial wastewater loads, leading to water quality deterioration. The model calibration and input variables, including hydraulic constants, BOD decay constants, and settling rates, are considered in detail. The study emphasizes the importance of understanding and managing water quality in the River Yamuna, a crucial resource for the region (Paliwal et al, 2007).

The study was conducted in Rishikesh, Uttaranchal, along the banks of the River Ganga, assessing water quality from Mini Goa Beach (upstream) to Bhardwaj Sthal (downstream) in December 2008. Twenty samples were collected based on prominent

activities like bathing, washing, and sewage/wastewater discharge. On-site analysis included parameters like pH, electrical conductivity (EC). The study also evaluated parameters relevant to irrigation suitability, including residual sodium carbonate (RSC), soluble sodium percentage (SSP), sodium adsorption ratio (SAR), permeability index (PI), Kelley's ratio (KR), and magnesium hazard (Mg Haz.). Results showed varying values for these parameters, indicating the suitability of the water for different uses. The study provides comprehensive data on the physical, chemical, and biological properties of the river water in Rishikesh, which can be essential for assessing its usability for various purposes (Haritash et al, 2016).

The Analysis System and Its Setup

Proposed System

Material Used

As discussed above, the system consists of many components namely-

1. Arduino Uno: Arduino Uno is used to collect and convert all the Analog signals coming from all the 4 sensors to Digital signals for our interpretation.

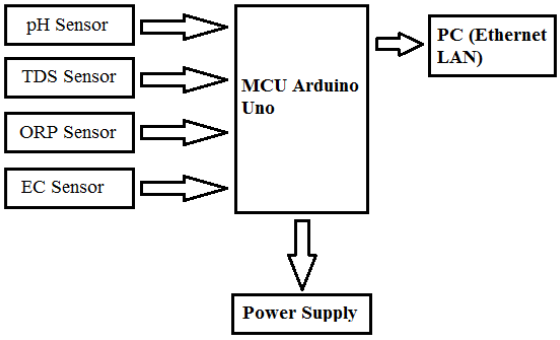
To expand the Arduino Uno to have an Ethernet port for transferring data over to cloud an Ethernet W5100 Shield is used.

2. Raspberry pi: Raspberry pi is being used as the main computer with all the required software (like Arduino IDE) to execute all the code. Raspberry Pi is also being used to power the Arduino Uno and all the other sensors in the setup.
3. Sensors:
   a) pH sensor: pH sensor helps to know how much acidic/ alkaline water is.
   b) TDS sensor: TDS sensor helps to measuring the Hardness of water.
   c) ORP sensor: ORP sensor helps to measuring the potential of water to oxidise or reduce any substance.
   d) Electrical Conductivity sensor: Electrical Conductivity sensor helps us in measuring the amount of salt concentration in the given water sample.
4. Other Components used:
   a) Bus bar / Mini Bread Board
   b) Raspberry Pi power adapter
   c) Jumper Wires
   d) Micro HDMI to HDMI Cable USB Type B to USB Type A Cable
   e) San Disk 16 GB Micro SD Card (Installed in Raspberry Pi)
   f) Ethernet Cable
   g) Enclosure Box

Set up: For measuring water turbulence, pH, and temperature appropriately, turbidity sensors, pH sensors, and temperature sensors that are directly connected to the

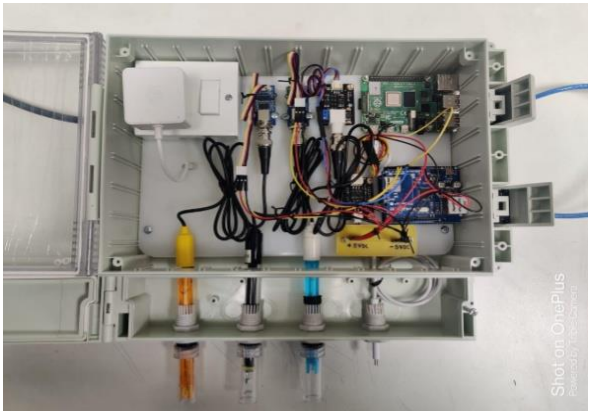
microcontroller are employed. Data is gathered by the microcontroller and processed with the system through LAN.Architecture of the system can be illustrated better in figure 1.

Figure 1: *Architecture of the system*



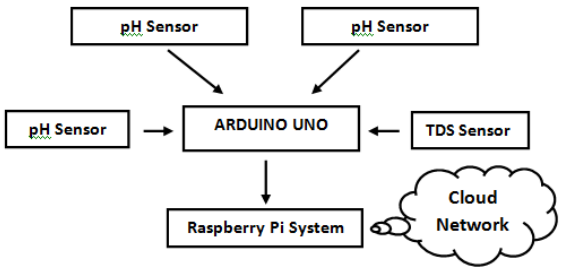
A picture of the working model with all the sensors and other components is shown in the figure 2.

Figure 2: *Set up of the monitoring systemWorking*



Sensors send analog signals to the Arduino Uno microcontroller. The Arduino microcontroller converts this analog signal to digital data. Raspberry Pi is used as the main computer which receives all data from the Arduino microcontroller which sends data to theCloud database. Figure 3 shows the flow of data in the system and to the cloud network.

Figure 3: *Flow of data in the system network*



Water Sampling: Water samples were collected over a period of months to determine the water's quality on Mehendi Ghat. Water samples are taken in accordance with ISO 5667-6.

The samples were then examined using the pH, TDS, ORP, and EC sensors. The table 2 below displays the information from water samples.

Table 2

Data collected by monitoring system

Sample Date	TDS Sensor	pH Sensor	EC Sensor	ORP Sensor
04-05-20	206.69	7.51	0.74	77.6
07-08-20	186.09	7.32	0.71	68.22
14-08-20	194.33	7.1	0.74	85.44
21-08-20	161.7	6.68	0.71	106.9
28-08-20	187.84	7.94	0.71	78.25
11-09-20	191.58	6.82	0.71	97.52

The Prediction System

The approach involved a combination of data collection, pre-processing, machine learning, and web development techniques. This section provides an overview of the approach undertaken to achieve the project objectives.

Dataset

Data collection

The first step in the approach involved acquiring a dataset of water quality parameters from a reliable and publicly available source. The dataset consisted of measurements obtained from various locations, which encompassed parameters such as ph, Hardness, Solids, Chloramines, Sulphate, Conductivity, Organic carbon, Trihalomethanes and Turbidity. There were 3000+ records for all the parameters present in the dataset. While the specific data was not collected directly from the Mehandi Ghat on the River Ganga, it was sourced from an authoritative open-source database that provides comprehensive information on water quality. This approach ensured a diverse and representative dataset, enabling a thorough analysis of water quality variations. By utilizing an established dataset, the project leveraged existing knowledge and contributed to the broader understanding of water quality analysis.

Data Pre processing

Once the data was collected, it underwent pre-processing to prepare it for further analysis. This involved handling missing values, outliers, and noise, as well as performing data normalization or scaling to ensure all parameters were on a similar scale. Data pre-processing was crucial to enhance the quality and reliability of the data used for model training.

Machine Learning Model Development

To develop the classification model for determining water portability, a machine learning approach was employed. The pre-processed data was divided into training and testing sets. Various machine learning algorithms, including logistic regression, decision trees, random forests, and support vector machines (SVM), were evaluated and compared for their effectiveness in classifying water samples as potable or non-potable. The model was trained on the training set and evaluated using appropriate performance metrics.

Cleaning of Dataset

To ensure the accuracy and reliability of the dataset, a cleaning process was conducted to address abnormalities such as null values and outlier values. This section focuses on the cleaning of the dataset for null values and outlier values. Cleaning the Dataset for Null Values Null values, or missing values, can introduce challenges during the analysis process. To handle null values, the following approach was employed:

1. The dataset was examined to identify the presence of null values in each parameter.
2. For each parameter with null values, a replacement strategy was implemented. In this case, the mean value for the respective parameter across the entire dataset was used as a replacement for the null values.
3. By replacing the null values with the mean value, the overall statistical characteristics of the dataset were preserved, and the impact of missing values on the analysis was mitigated.

The cleaning process for null values ensured that the dataset remained complete and suitable for further analysis and modelling.

Handling Outliers

Outliers refer to data points that deviate significantly from the rest of the dataset. In our analysis, we observed the presence of outliers and recognized their importance in accurate parameter analysis. Therefore, the decision was made to retain outliers in the dataset. Retaining outliers allowed us to consider and analyse extreme values that might provide valuable insights into the water quality parameters. Additionally, removing outliers could potentially lead to biased results and inaccurate conclusions. By acknowledging the presence of outliers and retaining them in the dataset, we aimed to conduct a comprehensive and robust analysis of the water quality parameters. The cleaning process for null values and the retention of outliers ensured that the dataset was prepared appropriately, enabling accurate analysis and interpretation of the parameters.

Statistical Analysis of Dataset

The dataset underwent statistical analysis to derive insights into its characteristics. Two key statistical measures, the mean and standard deviation, were calculated for each parameter. Based on these statistical analyses, the following observations are made:

1. **Hardness:** The average hardness of the water samples is approximately 7.08, with a minimum value of 0 and a maximum value of 14. Hardness refers to the concentration of calcium and magnesium ions in the water.
2. **Solids:** The average concentration of solids in the water is about 196.37 mg/L. The minimum value is 47.43 mg/L, and the maximum value is 323.12 mg/L. Solids in water can include various minerals, salts, and organic matter.
3. **Chloramines:** The average chloramines concentration is around 7.12 mg/L. The minimum and maximum values are 0.352 mg/L and 13.127 mg/L, respectively. Chloramines are disinfectants commonly used in water treatment.
4. **Sulfate:** The average sulfate concentration is approximately 333.78 mg/L. The sulfate levels range from a minimum of 129 mg/L to a maximum of 481.03 mg/L.
5. **Conductivity:** The average conductivity of water is about 426.21 μS/cm. Conductivity is a measure of the water's ability to conduct electricity and is related to the concentration of dissolved salts and minerals.
6. **Organic Carbon:** The average organic carbon content is 14.28 mg/L. Organic carbon represents the presence of natural organic matter in the water.
7. **Trihalomethanes:** The average concentration of trihalomethanes is 66.40 μg/L. Trihalomethanes are byproducts of water disinfection and can potentially be harmful in high concentrations.
8. **Turbidity:** The average turbidity value is 3.97 NTU (Nephelometric Turbidity Units). Turbidity measures the cloudiness or haziness of water due to suspended particles.
9. **Potability:** The data indicates a binary classification of water samples into "potable" and "non-potable" (1 for potable and 0 for non-potable). Around 39.01% of the samples are labeled as potable (1), and 60.99% are labeled as non-potable (0).

A summary of statistical analysis is provided in the following table 3.

Table 3

A summary of statistical analysis

SN	Parameter	Mean	Minimum	Maximum	Standard error
1	ph	7.08	0.0	14	1.59
2	Hardness	196.37	47.44	323.12	32.88
3	Solids	22014.09	321.95	61227.19	8769.58
4	Chloramines	7.13	0.35	13.12	1.59
5	Sulfate	334.78	129	481.4	41.42
6	Conductivity	426.2	181.4	753.34	81.83

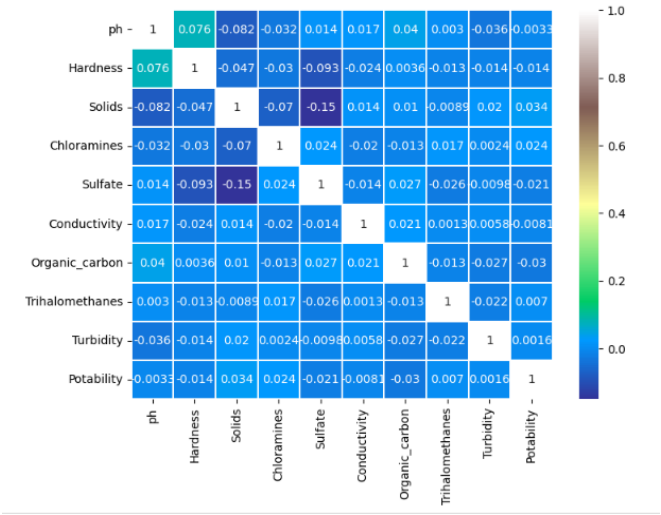
7	Organic Carbon	14.29	2.2	28.30	3.3
8	Trihalomethanes	66.39	0.74	124	16.17
9	Turbidity	3.97	1.45	7.73	1

By conducting statistical analysis, we gained insights into the central tendencies and variability of the dataset. These findings contribute to our understanding of the water quality parameters and provide a foundation for further analysis and interpretation.

Correlation between Parameters

To understand the relationships and dependencies between the various parameters in the dataset, a correlation analysis was performed. Correlation analysis helps us determine how strongly and in what direction variables are related. We prepared the heat map of the above parameters as seen in the figure 4.

Figure 4: *Heat map showing correlation in different parameters of the collected dataset*



Based on the correlation heat map, it can be concluded that there is no significant direct correlation (0.7 or 70 percent) between any two parameters. All the parameters are mainly independent of each other, highlighting their individual importance in the analysis. Feel free to customize the text and the inclusion of the correlation heat map analysis according to your specific dataset and project requirements.

Data Requirements for Training the Model

For training the model, we divide the randomized dataset into two parts: the training set and the testing set.

1) ***Training Set:*** The training set comprises 70 percent of the dataset, which corresponds to approximately 19,000 rows (out of 33,000). This subset of the data is used to train the machine learning models. During training, the models learn patterns and relationships between the input features (water quality parameters) and the target variable (portable or not portable).

2) ***Testing Set:***The testing set consists of the remaining 30 percent of the dataset, which amounts to approximately 8,000 rows. This subset is kept separate and not used during the training phase. Instead, it is used to evaluate the performance of the trained models on unseen data. By assessing the models on the testing set, we can estimate their ability to generalize and make accurate predictions on new, unseen instances.

By partitioning the dataset into training and testing sets, we ensure that our machine learning models are trained on a diverse range of data and can be evaluated on independent samples. This helps in assessing the performance and reliability of the models in classifying water quality as portable or not portable.

Decision Tree Classifier

The Decision Tree Classifier is a powerful and interpretable machine learning algorithm. It creates a tree-like model of decisions and their possible consequences. The classifier recursively splits the data based on the values of different features to create the most effective classification. We trained our model using decision tree classifier. In our case, we utilize the Decision Tree Classifier as the machine learning models for training. These models are suitable for classification tasks and have shown promising results in similar domains. During the training phase, the models analyze the relationships between the input features (pH, temperature, turbidity, dissolved oxygen, conductivity, total organic carbon, color) and the target variable (portable or not portable). By identifying patterns and correlations in the training data, the models adjust their internal parameters to make accurate predictions. The trained models learn to classify unseen instances based on the learned patterns. They become capable of predicting the portability of water samples based on their respective parameter values.

Testing the Model

After training the machine learning models using the training dataset, we evaluate their performance on two types of data: the data used for training and the remaining data reserved for testing.

1) Testing on Training Data: In order to assess how well the models generalize, we test them on the same dataset that was used for training. This step helps us understand the model's ability to learn and predict the ground truth labels accurately on familiar data. By evaluating the models on the training data, we can gain insights into their training performance and check for any signs of overfitting.
2) Testing on Unseen Data: To evaluate the real-world performance of the trained models, we use the remaining data that was not included in the training process. This data, which was set aside during the partitioning step, serves as a proxy for unseen instances. By testing the models on this independent dataset, we can assess their ability to generalize and make accurate predictions on new, unseen data points. This step helps us measure the

models' performance in real-world scenarios and provides insights into their effectiveness in classifying water quality as portable or not portable.

Accuracy

1) The data upon which the model was trained Upon testing the model on the data again we found that the model gave nearly 92.2 percent accuracy.
2) The rest of the data we left for testing Upon testing the model on the testing dataset we found that the model gave nearly 64 percent accuracy.

Table 4

Performance Metrics

Metric	Value
Train Accuracy	92.2 %
Test Accuracy	64%

Results

The study conducted a comprehensive analysis of water quality parameters, including TDS, pH, and ORP, with a particular focus on their trends during the early winter season (October to November) in 2021. As illustrated in Figure 5, the results indicated a notable surge in the values of these parameters during this specific period.

The primary factor contributing to this observed increase is likely the elevated concentration of microbial organisms in the Ganga River during the post-monsoon season (October to November). These microorganisms can significantly influence the chemical composition of the water and are a key determinant of water quality. The monitoring system employed in this study saves the collected data in the cloud, providing a convenient means of accessing and displaying real-time information through an interactive interface. This interface, accessible remotely via the internet, offers users the ability to visualize historical data as well as real-time measurements. Figure 6 provides an example of the user-friendly interface designed for this purpose.

Figure 5: *Trends in values of pH over months of year 2021*

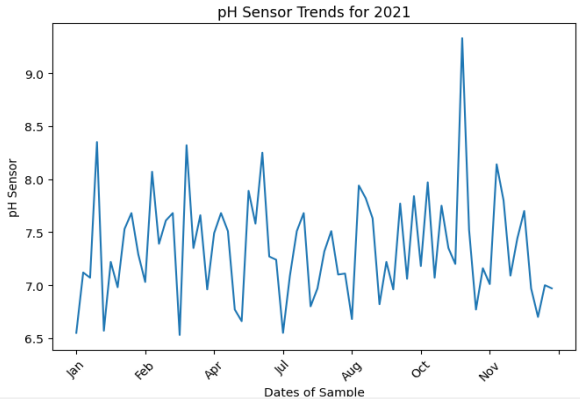


Figure 6: *Interface for monitoring of historical data or in real time*

Building upon these findings, the study proposes the integration of an alert system based on machine learning models. This system can generate timely alerts when parameter values exceed recommended norms. The research also demonstrated the effectiveness of a Decision Tree Classifier in predicting water portability. The classifier exhibited promising results, showcasing its potential utility in water quality assessment and prediction. It's important to note that researchers and practitioners can select the most suitable classifier for their specific needs, taking into account factors like interpretability, accuracy, and the characteristics of their dataset. Future research endeavors may explore alternative classification algorithms and feature selection techniques to further enhance the prediction of water portability and ensure the safety of water resources.

Conclusions

The proposed system not only outperforms existing solutions but also presents results in a more user-friendly and comprehensive manner, making it a significant advancement in water quality monitoring technology. Its potential impact extends beyond individual use; concerned authorities, environmental agencies, and researchers can leverage this system to actively monitor parameter variations and water quality at Mehandi Ghat, Kannauj, in real-time. This proactive approach allows for early intervention and mitigation measures, preventing situations from deteriorating and ensuring the protection of vital water resources.

The data collected through this advanced system serves as a valuable resource for further research, analysis, and advancements in the field of water quality monitoring and environmental conservation. Researchers can use this extensive dataset to gain insights into long-term trends and patterns, facilitating a deeper understanding of water quality dynamics. Furthermore, the adaptability of this system makes it suitable for deployment in various locations worldwide, thereby contributing to a global effort to study and manage the water quality of rivers and water bodies. Customizing the system's standards according to the specific geographical regions of interest ensures its effectiveness and relevance in diverse ecosystems. In conclusion, the proposed system not only enhances our ability to monitor water quality effectively but also paves the way for more informed decision-making and sustainable water resource management on both local and global scales.

References

Bureau of Indian Standards. (2015). *Drinking Water specification* (2ndrevision) Amendment no.1, June 2015to IS 10500:2012 Drinking water specification. http://cgwb.gov.in/Documents/WQ-standards.pdf.

Rana, Garima, Ahmad, F., &Vidyarthi, A.(2022). Study of Water Quality of Ganga River and Its Suitability for Mass Ritualistic Bathing before Kumbh Mela-2021 at Haridwar in Uttarakhand, India. VivekandDublish. *Journal of Indian Water Works Association*, 23–29.

Chauhan, A., &Singh, S. (2011). *Evaluation of Ganga Water for drinking purpose by water quality index at Rishikesh, Uttarakhand, India*.

Bhutiani, R., Khanna, D. R., Kulkarni, D. B., &Ruhela, M.(2016). Assessment of Ganga river ecosystem at Haridwar, Uttarakhand, India with reference to water quality indices. *Applied Water Science*, *6*(2), 107–113.https://doi.org/10.1007/s13201-014-0206-6

Veesommai, C., &Kiyoki, Y.(2015). River water-quality analysis: "critical contaminate detection", "classification of multiple-water-quality-parameters values" and "real-time notification" by rspa processesInternational Electronics Symposium (IES), *2015*. https://doi.org/10.1109/ELECSYM.2015.7380842

Salah, M., Chowdurya, U., Bin Emranb, T., Ghosha, S., Pathaka, A., ManjurAlama, M., Hossaind, S., &Anderssonc, K. (2019). IoTbasedreal-time riverwaterqualitymonitoringsystem. The 16th International Conference on Mobile Systems and Pervasive Computing (MobiSPC), August19-21, 2019. Halifax. https://doi.org/10.1016/j.procs.2019.08.025

Wang, Y., Zhang, W., Engel, B. A., Peng, H., Theller, L., Shi, Y., &Hu, S.(2015)**.08.003**. A fast mobile warning system for water quality emergency risk in ungauged river basins. *Environmental Modelling and Software*, *73*, 76–89.https://doi.org/10.1016/j.envsoft

Kothari, V., Vij, S., Sharma, S. K., &Gupta, N.(2021). Correlation of various water quality parameters and water quality index of districts of Uttarakhand. *Environmental and Sustainability Indicators*, *9*, 100093. https://doi.org/10.1016/j.indic.2020.100093

Jalal, D., &Ezzedine, T.(2019). Toward a Smart Real Time Monitoring System for Drinking Water Based on Machine Learning. In International Conference *on* Software, *Telecommunications and* ComputerNetworks *(SoftCOM)*, IEEE,*2019*(pp. 1–5). https://doi.org/10.23919/SOFTCOM.2019.8903866

Kumar Jha, M., Kumari Sah, R., Rashmitha, M. S., Sinha, R., Sujatha, B., &Suma, K. V.(2018). Smart water Monitoring System for Real-time water quality and usage monitoring. *Proceedings of the International ConferenceonInventiveResearchinComputingApplications (ICIRCA2018), IEEE Xplore Compliant Part Number: CFP18N67-ART*; ISBN: 978-1-5386-2456-2. https://doi.org/10.1109/ICIRCA.2018.8597179

Saravanan, K., Anusuya, E., Kumar, R., &Son, L. H.(2018). Real-time water quality monitoring using Internet of Things in SCADA. *Environmental Monitoring and Assessment*, *190*(9), 556. https://doi.org/10.1007/s10661-018-6914-x

Mishra, A.(2010). Assessment of water quality using principal component analysis: A case study of the river Ganges. *Journal of Water Chemistry and Technology*, *32*(4), 227–234. https://doi.org/10.3103/S1063455X10040077

Kumar, A., Taxak, A. K., Mishra, S., &Pandey, R.(2021). Long-term trend analysis and suitability of water quality of River Ganga at Himalayan hills of Uttarakhand, India. *Environmental Technology and Innovation*, 22.https://doi.org/10.1016/j.eti.2021.101405

Chaudhary, M., Mishra, S., &Kumar, A.(2017). Estimation of water pollution and probability of health risk due to imbalanced nutrients in River Ganga, India. *International Journal of River Basin Management*, *15*(1), 53–60.https://doi.org/10.1080/15715124.2016.1205078

Tiyasha, T., Tung, T. M., &Yaseen, Z. M.(2020). A survey on river water quality modeling using artificial intelligence models: 2000–2020. *Journal of Hydrology*, *585*, 124670. https://doi.org/10.1016/j.jhydrol.2020.124670

Faculty of Computer and Mathematical Sciences, UniversitiTeknologi MARA, Kelantan, C., Bharu, K. K., Sireh, L., &Bharu, K.p. 15050.Kelantan, Malaysia.

Giri, A.et al.(2014). Comparison of Artificial Neural network algorithm for water quality prediction of River Ganga. *Environmental Research Journal*, *8*(2), 55–63.

Paliwal, R., Sharma, P., & Kansal, A. (2007). Water quality modeling of the river Yamuna (India) using QUAL2E-UNCAS. *Journal of Environmental Management*, *83*(2), 131–144. https://doi.org/10.1016/j.jenvman.2006.02.003

Haritash, A. K., Gaur, S., &Garg, S.(2016). Assessment of water quality and suitability analysis of River Ganga in Rishikesh, India.*Applied Water Science*, *6*(4), 383–392. https://doi.org/10.1007/s13201-014-0235-1

To cite this article:

Patro, B. D. K., Sharma, S. Bajpai, A. (2023**).** Water quality monitoring of river ganga using non-linear data analytics. *Mathematics Education Forum Chitwan, 8* (1), 76-92. https://doi.org/10.3126/mefc.v8i1.60477