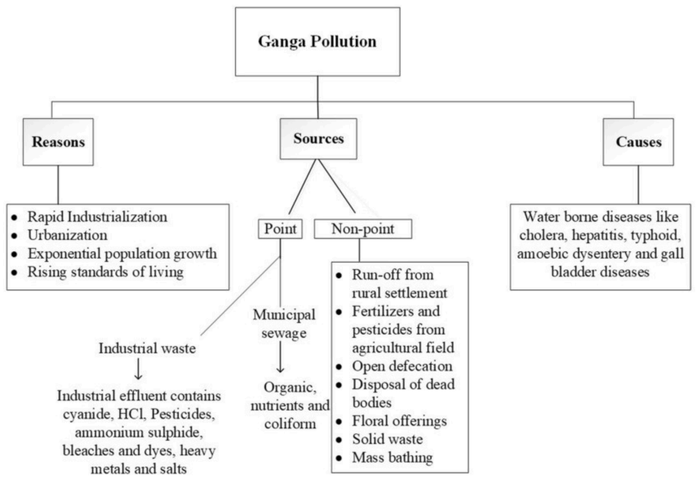
P.V.Hemachandra (RA2211003011040)
D.Sree Charanya (RA2211003011046)
N.Rohan Teja (RA2211003011052)

DATA SCIENCE

BATCH 12
MENTOR : DR.S.PADMINI
ASSOCIATE PROFESSOR
CTECH

WATER PREDICTION AND CONTAMINATION DETECTION

Clean Water And Sanitation (SDG 6)



Station Name	pH	BOD (mg/L)	DO (mg/L)	TC (MPN/100mL)	Date	Location
Anoopshahar	7.1	3.0	9.78	540	2020-03-25	Uttar Pradesh
Farrukabad	7.1	3.0	8.70	2200	2020-03-25	Uttar Pradesh
Rajghat, Kannauj	8.37	3.0	9.35	4700	2020-03-25	Uttar Pradesh
Bithoor, Kanpur	7.8	1.17	7.66	4100	2020-03-25	Uttar Pradesh
Jajmau, Kanpur	7.64	1.79	8.18	14000	2020-03-25	Uttar Pradesh
Assi Ghat, Varanasi	6.58	2.2	5.0	17000	2020-03-25	Uttar Pradesh
Malviya Bridge, Varanasi	8.05	1.4	7.62	17000	2020-03-25	Uttar Pradesh
Patna	7.63	30.13	0.25	1700	2020-03-25	Bihar

PROBLEM STATEMENT

The **Ganga River** faces severe pollution from industrial waste, sewage, and human activities, while traditional monitoring methods are slow and reactive, delaying contamination detection. **The National Mission for Clean Ganga (NMCG)**, under the **Namami Gange Program**, along with committees like the **Empowered Task Force (ETF)** and **State Ganga Committees**, oversee cleanup efforts but rely on outdated data collection methods. This project supports NMCG's mission by using data science and machine learning for real-time water quality monitoring, contamination prediction, and automated alerts, enabling faster decision-making and a sustainable Ganga River.

CHALLENGES

1. **Slow and Manual Monitoring** – Water testing takes too long because it's done manually.
2. **Delayed Contamination Detection –** Pollution is found too late, after damage has occurred.
3. **Lack of Real-time Analysis –** Traditional methods analyze past data but don't show current water quality.
4. **No Predictive Capability –** Authorities can't predict pollution in advance to take early action.
5. **Absence of Automation –** Without automated systems, responses to contamination are delayed.

STEPS TO SOLVE

1. **Problem Definition**
- Predict and detect water contamination using machine learning.
- Focus on key parameters: pH, BOD, DO, TC.
2. **Data Collection**
- Gather real-time data from 14 CPCB stations.
- Extract historical water quality data (pre-, during, and post-lockdown).
3. **Data Cleaning & Preprocessing**
- Handle missing values with imputation.
- Normalize data for consistency.
- Use SMOTE to balance datasets.
4. **Exploratory Data Analysis (EDA)**
- Visualize trends and correlations between parameters.
- Detect anomalies and seasonal variations.
5. **Model Building & Evaluation**
- Use SVR-GA, ANN, and Random Forest for prediction.
- Evaluate models with R^2 Score, RMSE, and MAE.
- Optimize with hyperparameter tuning.
6. **Deployment & Real-time Monitoring**
- Build a dashboard/app for real-time alerts.
- Automate predictions for environmental agencies.
- Integrate IoT sensors for continuous monitoring.

SUGGESTIONS

1. Use **IoT Sensors** for continuous, real-time data collection.
2. Implement **deep learning models** for better prediction accuracy.
3. Develop a **mobile app** for real-time data and contamination alerts.
4. Expand the dataset with more **monitoring stations and parameters.**
5. Collaborate with **government bodies** to influence policies for pollution control.

CONCLUSION

In conclusion, using data science and machine learning to monitor the Ganga River is a big improvement over old methods. By collecting real-time data, using prediction models, and adding automation, the project helps take quick action to protect the river. Adding more data, using **IoT sensors**, and working with **government bodies** makes the system even better, supporting sustainable water management.