A Minor Project Report

on

Water Quality Prediction and Monitoring System

by

Shreya Nikam

22BIT047

Keya Padia

22BIT048

Under the Guidance of

Dr. Parth Thakar

HOD, ECE

Submitted to

Department of Information and Communication Technology,

School of Technology,

Pandit Deendayal Energy University, Gandhinagar

MAY 2024-2025

CERTIFICATE

This is to certify that the seminar report entitled **Water Quality Prediction and Monitoring System** submitted by **Nikam Shreya-22BIT047** and **Keya Padia-22BIT048** and has been conducted under the supervision of **Dr. Parth Thakar, HOD, Department of EC**, and is hereby approved for the partial fulfilment of the requirements for the award of the degree of Bachelor of Engineering in the Department of **Information and Communication Technology** at Pandit Deendayal Energy University, Gandhinagar. This work is original and has not been submitted to any other institution for the award of any degree.

Sign:

Name of Guide:

Designation:

Department:

School of Technology,

Pandit Deendayal Energy University, Gandhinagar

Sign:

Name of Examiner:

Designation:

Department:

School of Technology,

Pandit Deendayal Energy University, Gandhinagar

DECLARATION

We hereby declare that the minor project report entitled Water Quality Prediction and Monitoring System is the result of our own work and has been written by us. This report has not utilized any language model or natural language processing artificial intelligence tools for the creation or generation of content, including the literature survey.

The use of any such artificial intelligence-based tools was strictly confined to the polishing of content, spell checking, and grammar correction after the initial draft of the report was completed. No part of this report has been directly sourced from the output of such tools for the final submission.

This declaration is to affirm that the work presented in this report is genuinely conducted by us and to the best of our knowledge, it is original.

Nikam Shreya-22BIT047

Keya Padia-22BIT048

Information and Communication Technology Department,

School of Technology,

Pandit Deendayal Energy University,

Gandhinagar

Date:

Place:

ACKNOWLEDGEMENT

We want to sincerely thank everyone who helped us with our project, Water Quality Prediction and Monitoring System. Without their constant support, direction, and inspiration, this work would not have been feasible.

We want to express our sincere gratitude to our mentor for his guidance, insightful criticism, and technical knowledge that shaped this project. His recommendations were instrumental in improving the water quality prediction models' precision and scientific dependability.

The opportunity to work with real-world water datasets - performing data preprocessing, feature engineering, exploratory analysis, and machine learning model development - helped us gain deeper insights into environmental data science.

We want to thank the use of modern tools and technologies that made our work much easier. Power BI dashboards helped us make interactive and useful visualizations to keep an eye on trends and understand water quality across locations. Machine learning algorithms like Random Forest and XGBoost were very helpful in predicting important water quality indicators.

LIST OF FIGURES

TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION AND OBJECTIVE

1.1 INTRODUCTION

The most important resource for the sustainability of life is water, but it is becoming more and more threatened. For the vast majority of people, particularly in countries like India, groundwater is their main source of drinking water, but is currently experiencing both contamination and depletion. Groundwater quality has declined as a result of rapid industrialization, agricultural runoff (fertilizers and pesticides), and inappropriate waster disposal, making it a cause of health risk.

In the early years, water quality was checked by collecting samples from different wells and sending them to chemical labs for testing. This process is expensive and time consuming, which often leads to delay in analysis. Measuring parameters like Flourine, Uranium and Total Hardness needs special equipment and trained workers. Because of this, many rural areas often go unmonitored for long periods, leading to health issues among people such as Fluorosis, kidney damage from Uranium, or digestive problems from too much hardness.

In the modern world, Artificial Intelligence (AI) and Machine Learning (ML) are the solutions that are transforming the era of Industry 4.0. ML models can act as Virtual Sensors by using historical data and determines non-linear relationships between different chemical parameters. Such models have the potential to calculate parameters that are hard to measure (such as Hardness or WQI) based on easily available and inexpensive inputs (such as pH and Electrical Conductivity).

This project aims to develop a data-driven system that monitors and predicts groundwater quality. It uses a large dataset from the Central Ground Water Board (CGWB), Government of India, covering the years 2020 to 2023. The study contains a vast geographical area, utilizing data from major Indian states including Andhra Pradesh, Telangana, Karnataka, Tamil Nadu, Punjab, and Uttar Pradesh, ensuring the model is tested across diverse hydro-geological terrains**.** The target is to create a low-cost and high-accuracy prediction model. The system uses data from 2020 to 2022 to train strong algorithms and tests accuracy on 2023 data. This work connects raw data collection to useful health insights.

1.2 OBJECTIVES

The primary point of the study is to research the historical data on the groundwater in 2020- 2022 to learn how the water quality has changed throughout the years in different districts. We are especially seeking concealed trends in the critical chemical parameters, such as pH, Chloride, and Nitrate, to construct a useful Low-Cost Estimator. The main concept is to develop a machine learning model that is capable of making predictions of expensive, or hard-to-measure, values (e.g., Total Hardness and the Water Quality Index (WQI)) with just simple, easy-to-obtain inputs. In doing so, we will be in a position to reduce the use of expensive, time-consuming lab tests, making it significantly cheaper and quicker to maintain a constant water safety check.

In this project, in addition to prediction, high emphasis is laid on the aspect of the health of people that is at stake due to the contamination of groundwater. We specifically focus on the high-risk areas or a hotspot of dangerous substances such as Uranium and Fluoride, which are usually not noticed until it is too late. In order to make sure that our system actually works in the real world, and not that the given system is merely a theoretical experiment, we put our models to the test intensively on a totally new, unknown dataset of the year 2023. Such validation is an essential step, as it will show that the model is able to process new data and that it is sufficiently dependable to become a tool of actual on-ground monitoring.

Lastly, we felt like making these technical discoveries helpful to the real-life decision makers. In order to have the evolution between sophisticated data analysis and action in real life, we came up with an interactive visualization dashboard. It is a tool that simplifies the difficult forecasts into an easy, pictorial format that classifies the areas as safe, poor, or unsafe. It can be viewed as a useful decision-support tool, enabling the government officials and policymakers to immediately identify the problem areas and make a decision on where the resources and interventions are most necessary to ensure that the health of the population is safeguarded.

CHAPTER 2: LITERATURE REVIEW

2.1 OVERVIEW OF GROUNDWATER QUALITY ASSESSMENT

The concept of groundwater quality measurement has been greatly enhanced during the past decades. Traditionally, the water quality monitoring used to be a reactive process- it was not tested until a health outbreak or some other noticeable change in the water taste and color. Nevertheless, the existing literature highlights the proactive strategy due to the fact that, in many cases, the degradation of groundwater is considered to be irreversible when the aquifers are contaminated. This has been supported by several years of study of agencies such as the Central Ground Water Board (CGWB) and the World Health Organization (WHO), which have long since concluded that the chemistry of ground waters is not fixed; both geogenic (rock-water-interaction) and anthropogenic (human activity such as agriculture and industry) processes play a role.

In the recent circumstances in the Indian setting, some worrying findings have been reported: the management of traditional biological contamination (bacteria) is being established, but chemical contamination is growing. Scientists have reported that the breakneck urbanization and excessive exploitation of groundwater sources have changed the natural chemical equilibrium of aquifers and increased the dissolved salts and heavy metal content. Such a change requires the transfer of the analysis of certain hydro-chemical parameters and not only an overview purity test.

2.2 HYDRO GEO-CHEMICAL PARAMETERS: CATIONS AND ANIONS

Much of the available literature is devoted to the basic ionic equilibrium of water. The level of concentration of major cations (positive ions) and anions (negative ions) determines, to a great extent, the potability of the groundwater.

2.2.1 Major Cations (Calcium, Magnesium, Sodium, Potassium)

Calcium (Ca2+) and Magnesium (Mg2+) are the major determinants of the hardness of water, which is a fundamental theme of this study.

Calcium (Ca2+) & Magnesium (Mg2+): These are the most common ions within natural water, which is usually obtained through the leaching of limestone, dolomite and gypsum. They are vital to human well-being in low amounts, but in high amounts, they cause hard water. It is proven in literature that hardness above 300 mg/L is associated with more scaling on the pipes, inefficiency of the soaps, and may be associated with urolithiasis (kidney stones). In contrast, in some epidemiological studies, a very low concentration has been associated with an increased risk of cardiovascular disease, indicating a Goldilocks zone of these cations.

Sodium (Na^+) & Potassium (K^+): Sodium level is often employed in literature as a measure of salinity. High sodium levels are considered a critical risk factor among hypertensive patients and may make water inappropriate in terms of water irrigation application because it influences soil permeability (Sodium Adsorption Ratio). Although typically in smaller amounts, potassium is a highly important biological indicator; its abrupt increase is typically an indicator of agricultural fertilizer pollution (potash) or sewage overflow, and is a useful tracer in environmental investigations.

2.2.2 Major anions (chloride, sulphate, nitrate, Bicarbonate)

Though cations are the ones that determine hardness, the anions are usually the easiest indicators of sources of pollution.

Chloride (Cl^-) & Sulphate (SO_4^{2-}): Chloride commonly appears in hydro-chemical literature as a conservative tracer, i.e., a substance that does not readily react chemically. Sewage pollution or sea intrusion in coastal aquifers is specifically identified by high chloride levels. In the same way, Sulphates can be either identified as a result of industrial discharge or as a result of the decomposition of minerals that contain a lot of Sulphur. In excess, they both give a bitter taste and laxative effects to the consumers.

Nitrate (NO_3^-): One of the most widely documented anthropogenic pollutants, Nitrate is nearly solely associated with agricultural effluents (urea fertilizers) and leaching of septic tanks. Literature is in agreement about the health hazards of nitrates, especially the so-called Blue Baby Syndrome (Methemoglobinemia) in infants. Nitrate levels greater than the acceptable levels of 45 mg/L are becoming a cause of alarm in most of the Indian districts, and this is an indication of contamination that is caused by human activities.

Bicarbonate (HCO_3^-): This is the anion that symbolizes the Alkalinity of the water. It is commonly geogenic and is formed as a result of silicate rock weakening. It is a buffer that ensures the pH of the water. It has been shown that though Bicarbonate by itself is not lethal, it is important in regulating the solubility of other toxic metals such as Uranium.

2.3 THE CONTAMINANTS: FLUORIDE AND URANIUM

The modern literature has come to include aspects beyond the normal ions and given particular emphasis on the trace contaminants that are extremely harmful to health even in microscopic concentrations (parts per billion).

Fluoride (F 1-): India lies in a region of high fluoride. Widespread geochemical surveys have recorded that the loss of fluoride through granitic rocks results in fluorosis, which is a debilitating disorder of the teeth and bone. It is an archetypal geogenic contaminant, i.e., it is a natural contaminant, but its concentration is inflated by the lowering of water tables.

Uranium (U): This is a comparatively new area of Indian groundwater research. Recent reports have detected that the granitic aquifers of Telangana, Andhra Pradesh, and Punjab have considerable uranium contamination. Uranium is also a dual threat, unlike the other parameters, as it is both chemically toxic (to the kidneys) and radiologically dangerous. According to literature, a high concentration of Bicarbonate will promote the leaching of Uranium in rocks to the groundwater. Although it is a hazardous substance, Uranium is not a regular component of water tests and poses a great deficiency in the general health control.

2.4 WATER QUALITY INDEX (WQI) FRAMEWORK

It is hard to communicate the complicated information about chemistry to people. In order to address this, researchers came up with the Water Quality Index (WQI). The WQI is an accumulator of various parameters (pH, TDS, Hardness, etc.) into one unitless value, typically a value between 0 and 100.

The most widely cited method in the literature is the so-called Weighted Arithmetic Method, which is used by standards such as the Bureau of Indian Standards (BIS). It weighs the parameter using their perceived health risk, i.e., Fluoride, Nitrate, weight is higher than Chloride since the former is more dangerous. Such a single-score system enables policymakers to see contamination hotspots without having to know the complex chemistry of such instances.

CHAPTER 3: RESEARCH GAPS AND PROBLEM STATEMENT

3.1 RESEARCH GAPS

Although the literature on the subject of groundwater quality is not wanting, there is a gaping criticality on the feasibility of the monitoring approaches. Most of the available water quality prediction models relying on Machine Learning still demand so-called expensive inputs, such as the Calcium (Ca^{2+}) and the Magnesium (Mg^{2+}), to predict the Total Hardness or the Water Quality Index (WQI). This forms a paradox: since a field worker already needs to send a sample to a laboratory and determine the concentration of Calcium and Magnesium precisely, he or she could as well determine Hardness directly. No reliable models of Virtual Sensors have been developed that can predict these complicated parameters based only on low-cost and quickly measurable field data such as pH, Electrical Conductivity (EC) and Total Dissolved Solids. Through the maintained reliance on wholesome lab reports as outputs, present studies tend to provide a real-time monitoring alternative that is actually cost-effective in rural or resource-challenged regions.

Secondly, the standard water quality measurements have an inherent blind spot regarding toxicity. The classic Water Quality Indices (WQI) is highly biased towards general physico-chemical indices such as hardness, salinity, and pH, which often entirely omits trace heavy metals. Uranium and Fluoride contamination is a colossal silent health epidemic in the particular case of Indian groundwater, especially in granitic geology, such as Telangana and Karnataka. This notwithstanding, most predictive models and standard monitoring protocols incorporate these elements as outliers or simply neglect them as they cannot be quantified. As a result, a source of water may be considered as a Good one by the standard WQI paradigm but as a rationalistically hazardous source because of high contents of Uranium. Models are urgently needed that can incorporate these concealed contaminants into the safety narrative and no longer view them as niche issues.

Lastly, the gap between scholarly data science and environmental policy is significant and needs to be bridged. Most of the studies end with statistical measures, giving an R-squared or an accuracy score, and do not convert these difficult numbers into easily applicable decisions. The data is usually stuck in inflexible tables or difficult-to-understand Python notebooks that cannot be comprehended by the local government officials. The use of pipeline beyond prediction to interactive and geospatial visualization is under-researched. It is necessary to bridge this gap, and the key to it is the lack of an intuitive dashboard that will differentiate regions with the help of categories: safe or action required. Only in this case, even the best machine learning model will be a mere academic activity with minimal effect on the work of the real management of the population.

3.2 PROBLEM STATEMENT

The main problem of the groundwater surveillance lies in the logistical and economic challenges of the measurement of the complicated chemical parameters. Although basic water quality parameters can be immediately obtained using cheap handheld measurement devices, such as pH and Electrical Conductivity (EC), such critical parameters as Total Hardness and the general Water Quality Index (WQI) are at the moment titrated in the laboratory to determine the concentration of Calcium and Magnesium salts. Such dependency on laboratory analysis poses a big setback in terms of regular testing, particularly in remote or rural areas. Field workers have to decide between costly and lab tests that are rarely performed, and more frequent but partial on-site tests. This is why large masses of people keep drinking water that has not been properly tested in months and just because it is too expensive and time-consuming to test the complete picture of water quality, in order to do it regularly.

To make this problem even larger is the lack of correlation between standard safety ratings and real toxicity. The standard water quality reports do not usually look into trace contaminants such as Uranium and Fluoride, using reports aimed at a general assessment of potability. This causes a false sense of security in which water can be considered to be drinkable because it has low salinity or NP levels, whereas in reality, it may have levels of radioactive or geogenic poisons that are dangerous. Moreover, the sheer volumes of past-based hydro-chemical data that are gathered by organizations such as the CGWB usually lie in stagnant reports. The existing systems are severely short of systems that take an active approach by using this historical data to forecast the present risk. The issue, then, does not simply lie in gathering information, but in turning cheap and easily obtainable field data into correct, comprehensive safety warnings without having to put in place a full-scale laboratory facility on an individual basis.

CHAPTER 4: METHODOLOGY ADOPTED

4.1 OVERVIEW OF THE PROPOSED FRAMEWORK

The paper is systematic and data-driven in order to evaluate and forecast the quality of groundwater. The methodology has five different phases, which are: (1) Data Collection and Harmonization, (2) Preprocessing and Imputation, (3) Feature Engineering and Ionic Balancing, (4) Machine Learning Model Development (The "Low-Cost Estimator") and (5) Validation on Unseen Data. The general process flow will be aimed at modeling a real-life situation with historical data used to train the system and predict the next water quality with minimal inputs.

4.2 DATA COLLECTION AND ACQUISITION

The raw hydro-geochemical data were obtained from the Central Ground Water Board (CGWB), Ministry of Jal Shakti, Government of India. The data covers four years in a row:

- Training Set (2020-2022): This is a longitudinal dataset that has high physicochemical records utilized to train machine learning algorithms.
- Testing Set (2023): A rigidly held-out dataset that is solely used in the ultimate validation of the model's predictive power in a real-world environment.

The raw files were received in unequal Excel (.xlsx) files with unequal column names of years (e.g., Total Hardness in 2020 with the abbreviation TH and Total Hardness in 2021). To make all these files into Comma Separated Values (.csv) and to standardize the schema to a common format, a custom Python script was written to convert them.

4.3 PREPROCESSING AND CLEANING OF DATA

The real-life environmental data is also inherently noisy. To be model robust, the following preprocessing mechanism was strictly followed:

1. Normalization of Schemas: All column names were changed to standard scientific notation, including (e.g., changing gemsid w to S. No. and calc_ca to Ca (mg/L)).

2. Missing Values (Imputation):
   Geospatial Data: rows in which the Latitude or Longitude were missing were eliminated because

such geospatial visualization is one of the fundamental goals.

Chemical Parameters: Linear Interpolation was used to fill in missing values in critical columns (pH, EC, Cl, etc.), and Linear Interpolation was used to fill in certain areas of the table where data was missing. The reason why this technique was selected instead of mean imputation is that the characteristics of groundwater tend to have a spatial or temporal gradient.

Trace Elements: Trace elements such as Uranium (U ppb) that could not be interpolated because of the sparsity of the data were imputed with the global mean of the dataset to ensure that no data was lost.

3. Conversion of Data Types: Coercion of all chemical parameters to numeric types was done to remove non-numeric artifacts (e.g., text entries that had "BDL" - Below Detectable Limit written in them).

4.4 FEATURE ENGINEERING

To enhance the model's ability to predict complex parameters without direct measurements, domain-specific features were engineered based on hydro-chemical principles:

1. Total Anions (Proxy Variable):

$$Total\_Anions = Cl^- + SO_4^{2-} + NO_3^- + HCO_3^- + CO_3^{2-}$$

   Justification: Based on the principle of electrical neutrality, the sum of anions acts as a strong predictor for the sum of cations (Calcium + Magnesium), allowing the model to estimate hardness indirectly.

2. Alkalinity:

$$Alkalinity = HCO_3^- + CO_3^{2-}$$

   Justification: This feature helps the model distinguish between Carbonate (Temporary) Hardness and Non-Carbonate (Permanent) Hardness.

3. EC-to-Chloride Ratio:

$$Ratio = \frac{EC}{Cl^- + 1}$$

Justification: This ratio helps the model differentiate whether high conductivity is caused by salinity (NaCl) or by hardness-causing minerals, refining the prediction accuracy.

4.5 FORMULATION OF WATER QUALITY INDEX (WQI)

The weighted Arithmetic Method of calculating a "Ground Truth" WQI followed the Bureau of Indian Standards (BIS IS 10500:2012).

- Unit Weights (Wi): The allocations are inverse to the standard permissible limits (Si Soutions): Parameters that were less tolerant (such as Fluoride and Nitrate) were given greater weights to show their health outcome.
- Classification: The obtained scores of WQI were grouped into five levels of safety, which are Excellent (0-25), Good (26-50), Poor (51-75), Very Poor (76-100), and Unsafe (>100).

4.6 MACHINE LEARNING MODEL DEVELOPMENT

The key innovation of this approach is the Low-Cost Estimator.

1. Algorithm Selection:
    - Random Forest Regressor (RF): It was chosen due to the possibility of having non-linear relationships and the avoidance of overfitting. Environmental data of high variance is especially effective in the ensemble nature of RF.
    - XGBoost Regressor: This model is introduced as a competent model to compare the performance with other models that use gradient boosting as a basis for potentially better performance.
2. Target Variables:
    - Model A: Predicting the Total Hardness.
    - Model B: Model to predict Water Quality Index (WQI).
3. Introduction Feature Selection (The "Blind" Approach):
    - Importantly, the models have been trained without including Calcium (Ca 2+) and Magnesium (Mg 2+) as input features.
    - It was necessary to model the Total Hardness with proxies that are easy to measure (pH, EC, TDS, Anions and Trace Metals). This is modelling a field environment, a low-cost case where Ca/Mg titration is not available.
4. Hyperparameter Tuning:

- Random Forest model was optimized to the best of its performance, where n_estimators=400 and maxdepth=14 were selected as a balance between bias and variance.

4.7 VALIDATION STRATERGY:

The method involves a two-step validation, which is rigorous:

- Train-Test Split(80/20): The historical data (2020-2022) was subjected to this split to test the internal consistency using such measures as R-Squared (R 2), Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE).
- Unseen Data Testing: The trained model was fitted on the 2023 data. This step confirms that the model can predict the future water quality on the basis of the past trends, hence the generalization of its time.

4.8 VISUALIZATION AND DASHBOARD DEVELOPMENT

The last step in the methodology was to create an interactive Business Intelligence (BI) dashboard in Microsoft Power BI to make the difference between complex statistical predictions and policy decisions.

Data Modeling: Power BI was used to import the clean dataset and the results that were predicted into the application. DAX (Data Analysis Expressions) measures were designed to reflect dynamic measurements like the "Count of Unsafe Locations" and the ratio of certain chemicals (e.g., the ratio of EC/CL).

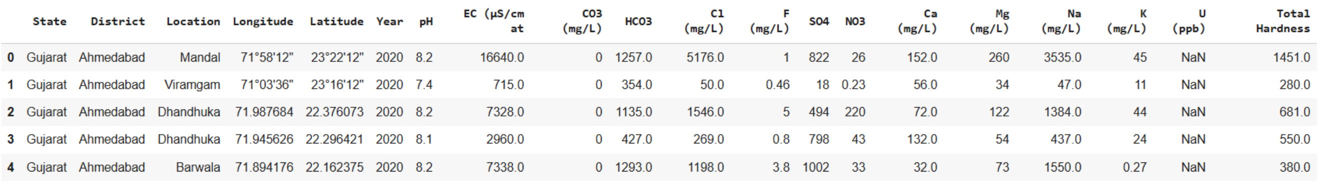
Report Architecture: The visualization was defined into three separate layers of analysis:

1. Regional Overview: A geospatial distribution study that determined contamination hotspots state-wise using filled maps.
2. Hydro-Chemical Analysis: A closer look at the ionic equilibrium, through the constant of Anions and Alkalinity to express the approaches to hardness origin.
3. Safety & Toxicity Profile: A health-risk report of dangerous outliers such as Uranium and Fluoride in relation to the WHO/BIS safety limits.

CHAPTER 5: DETAILS OF WORK EXECUTION

5.1 DATASET COLLECTION AND DESCRIPTION

Groundwater dataset containing pH, EC, ions (HCO_3, Cl, SO_4, NO_3), cations (Ca, Mg, Na, K), Total Hardness, etc. Latitude, Longitude, District, Year captured from various water sampling locations.



	State	District	Location	Longitude	Latitude	Year	pH	EC (µS/cm at	CO3 (mg/L)	HCO3	Cl (mg/L)	F (mg/L)	SO4	NO3	Ca (mg/L)	Mg (mg/L)	Na (mg/L)	K (mg/L)	U (ppb)	Total Hardness
0	Gujarat	Ahmedabad	Mandal	71°58'12"	23°22'12"	2020	8.2	16640.0	0	1257.0	5176.0	1	822	26	152.0	260	3535.0	45	NaN	1451.0
1	Gujarat	Ahmedabad	Viramgam	71°03'36"	23°16'12"	2020	7.4	715.0	0	354.0	50.0	0.46	18	0.23	56.0	34	47.0	11	NaN	280.0
2	Gujarat	Ahmedabad	Dhandhuka	71.987684	22.376073	2020	8.2	7328.0	0	1135.0	1546.0	5	494	220	72.0	122	1384.0	44	NaN	681.0
3	Gujarat	Ahmedabad	Dhandhuka	71.945626	22.296421	2020	8.1	2960.0	0	427.0	269.0	0.8	798	43	132.0	54	437.0	24	NaN	550.0
4	Gujarat	Ahmedabad	Barwala	71.894176	22.162375	2020	8.2	7338.0	0	1293.0	1198.0	3.8	1002	33	32.0	73	1550.0	0.27	NaN	380.0

Fig 5.1.1: Ground Water Quality Data (Top 5 columns)

5.2 DATA PREPROCESSING

Actions performed: Data Merging, Converting to Numeric Values, Handling missing values, Renaming column names, Creating engineered fields:

- Total_Anions= CL+SO4+NO3+HCO3+CO3
- Alkalinity= HCO3+ CO3
- EC_Cl_Ratio=EC/CL

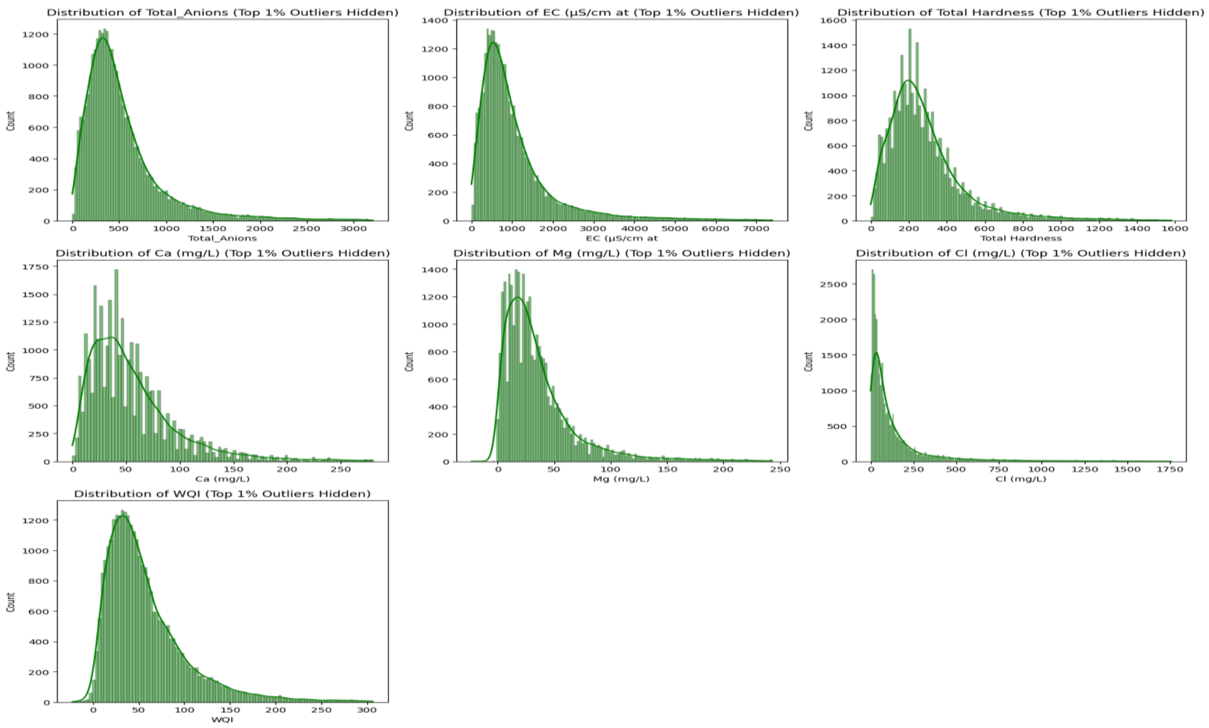


Fig 5.2.1: Distribution Plots of Key Water Quality Parameters (Top 1% Outliers Removed)

5.3 WATER QUALITY INDEX (WQI) CALCULATION

Let Vi = measured value of parameter i

Let Si = standard permissible value

Let Videal,i = ideal value

Let n = number of parameters

1. Quality Rating (Q_i)

$$Q_i = \frac{(V_i - V_{ideal,i})}{(S_i - V_{ideal,i})} \times 100$$

2. Unit Weight (W_i)

$$W_i = \frac{1}{S_i}$$

3. Weighted Value

$$Q_i W_i$$

4. Final WQI

$$\mathrm{WQI} = \frac{\sum_{i=1}^{n} Q_i W_i}{\sum_{i=1}^{n} W_i}$$

5. WQI Classification

$$\mathrm{WQI\_Level} = \begin{cases} \text{Excellent}, & \text{if WQI} \leq 25 \\ \text{Good}, & 25 < \text{WQI} \leq 50 \\ \text{Poor}, & 50 < \text{WQI} \leq 75 \\ \text{Very Poor}, & 75 < \text{WQI} \leq 100 \\ \text{Unsafe}, & \text{WQI} > 100 \end{cases}$$

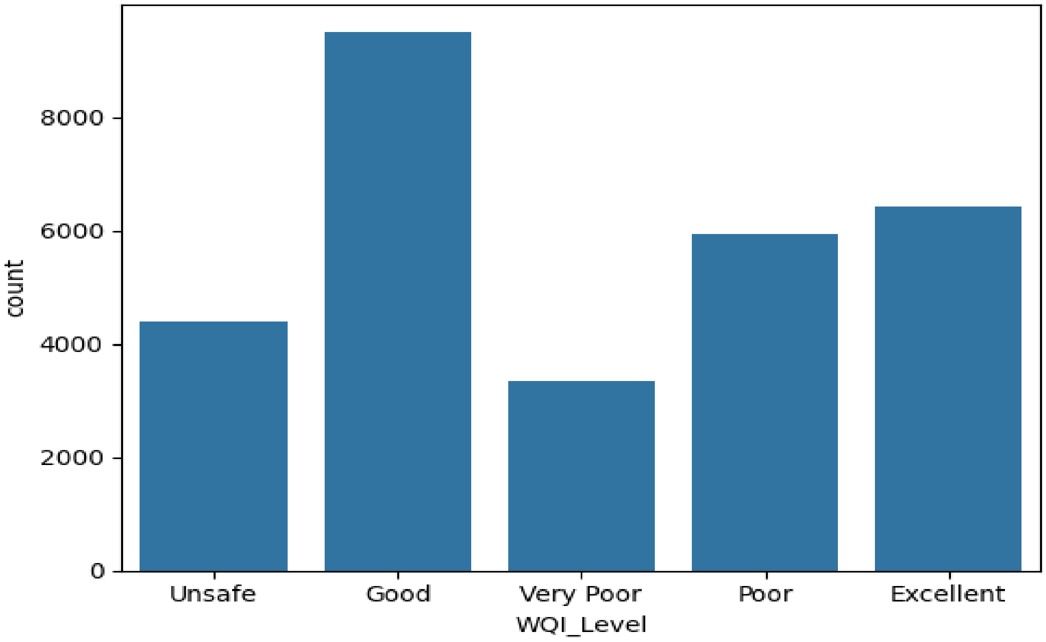


Fig 5.3.1: Histogram of WQI_Level

5.4 FEATURE ENGINEERING

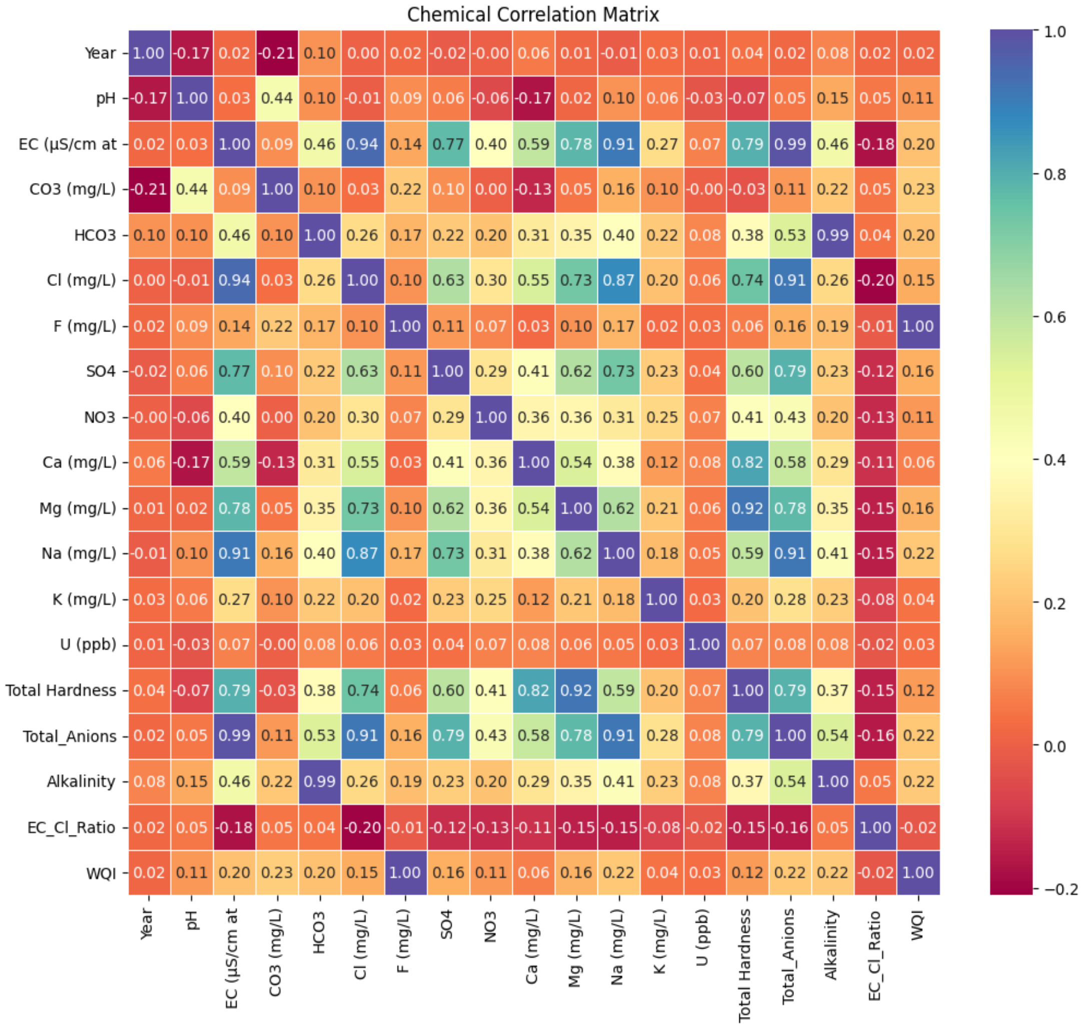


Fig 5.4.1: Chemical Correlation Matrix

Key Observations:

EC (Electrical Conductivity)

Shows **very strong correlation** with:

- **Cl (0.94)**
- **SO_4 (0.77)**
- **Na (0.91)**
- **Total Anions (0.99)**
- **WQI (0.20)** (weak–moderate)

This confirms that EC is primarily driven by dissolved ions, especially chloride, sulphate, and sodium.

Total Hardness

Strong correlations with:

- **Ca (0.82)**
- **Mg (0.79)**
- **Total Anions (0.79)**

Moderate correlation with:

- **SO_4 (0.64)**
- **EC (0.64)**

Hardness mainly depends on Ca and Mg, and also relates to total ionic concentration.

WQI (Water Quality Index)

Moderate correlations with:

- **EC (0.20)**
- **Cl (0.23)**
- **SO_4 (0.16)**
- **Na (0.22)**
- **Total Hardness (0.14)**

WQI increases when ionic content increases, but the relationship is not extremely strong (as expected due to weighting).

Total Anions

Extremely strong correlation with:

- **EC (0.99)**
- **SO_4 (0.70)**
- **Cl (0.70)**
- **Na (0.73)**
- **Total Hardness (0.79)**

Total anions represent overall mineralization and strongly align with conductivity and hardness.

pH

- Very weak correlations with all parameters (<0.20).
- pH remains stable and does not strongly influence mineral content.

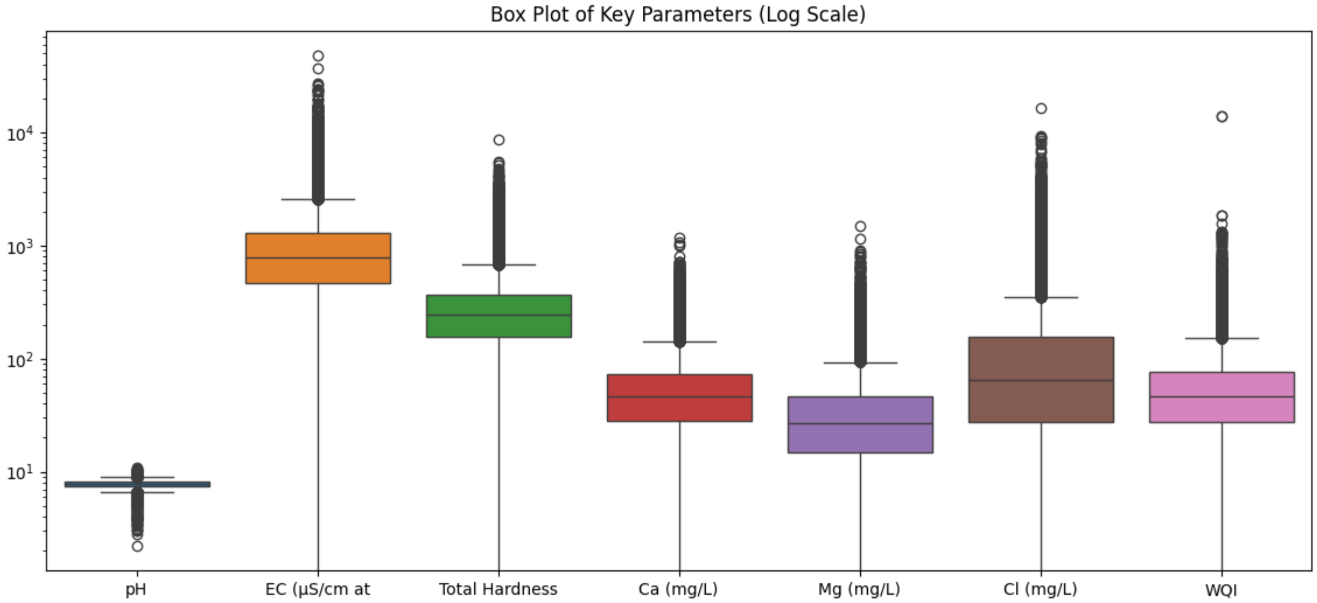


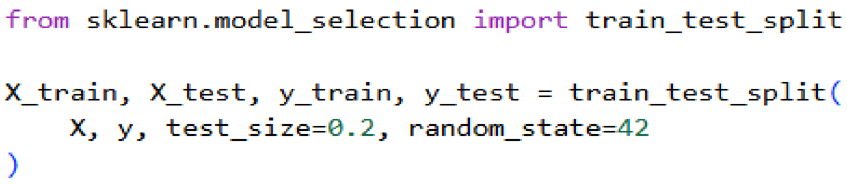
Fig 5.4.2: Box Plot of Key Parameters on Log scale

Key Observations:

- Water is essentially neutral, and pH is closely clustered with very little extreme variation.
- High variability in mineral content across locations is indicated by the wide ranges and numerous outliers displayed by EC, Total Hardness, Cl, Ca, Mg, and WQI.
- In certain samples, the most prevalent ionic contamination is caused by EC and chloride.
- A wide range of soft, hard, and extremely hard water sources can be found in total hardness.
- WQI varies greatly, indicating regional variations in overall water quality.

5.5 MACHINE LEARNING MODEL DEVELOPMENT

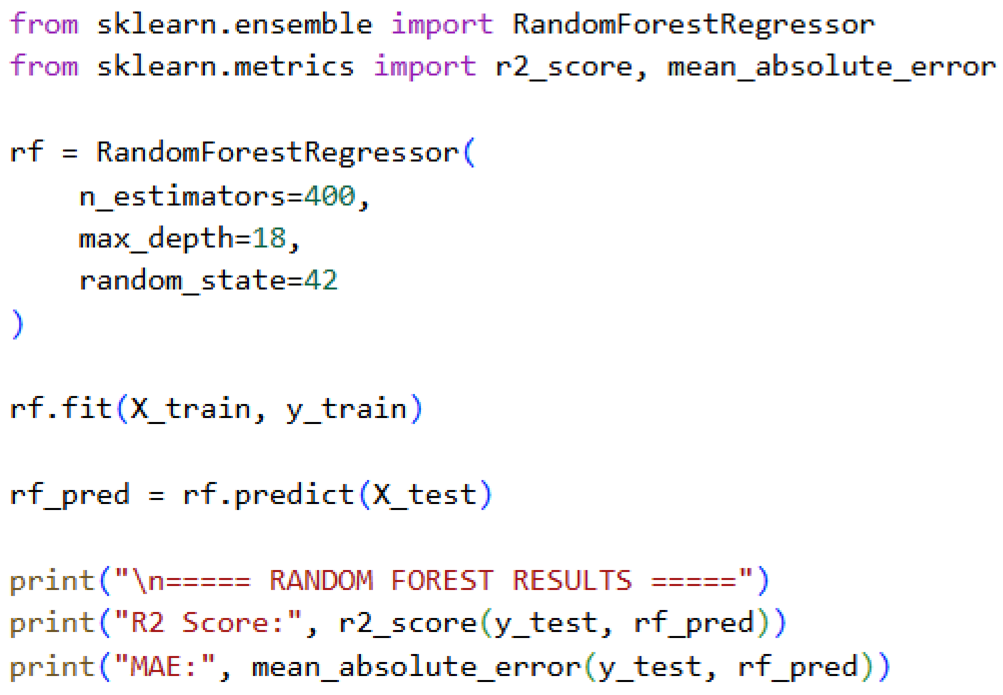
5.5.1 Train-Test Split



```
from sklearn.model_selection import train_test_split

X_train, X_test, y_train, y_test = train_test_split(
    X, y, test_size=0.2, random_state=42
)
```

Fig 5.5.1: Train -Test Split

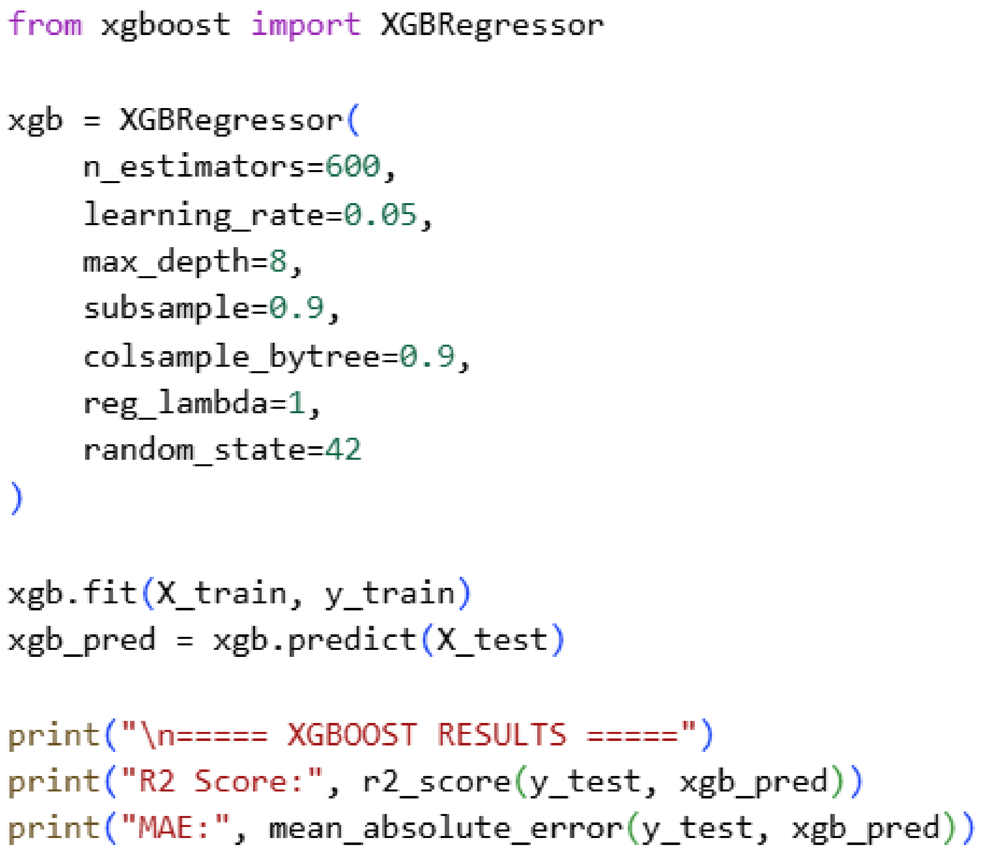
5.5.2 Random Forest Regressor



```
from sklearn.ensemble import RandomForestRegressor
from sklearn.metrics import r2_score, mean_absolute_error

rf = RandomForestRegressor(
    n_estimators=400,
    max_depth=18,
    random_state=42
)

rf.fit(X_train, y_train)

rf_pred = rf.predict(X_test)

print("\n===== RANDOM FOREST RESULTS =====")
print("R2 Score:", r2_score(y_test, rf_pred))
print("MAE:", mean_absolute_error(y_test, rf_pred))
```

Fig 5.5.2: Random Forest Regressor

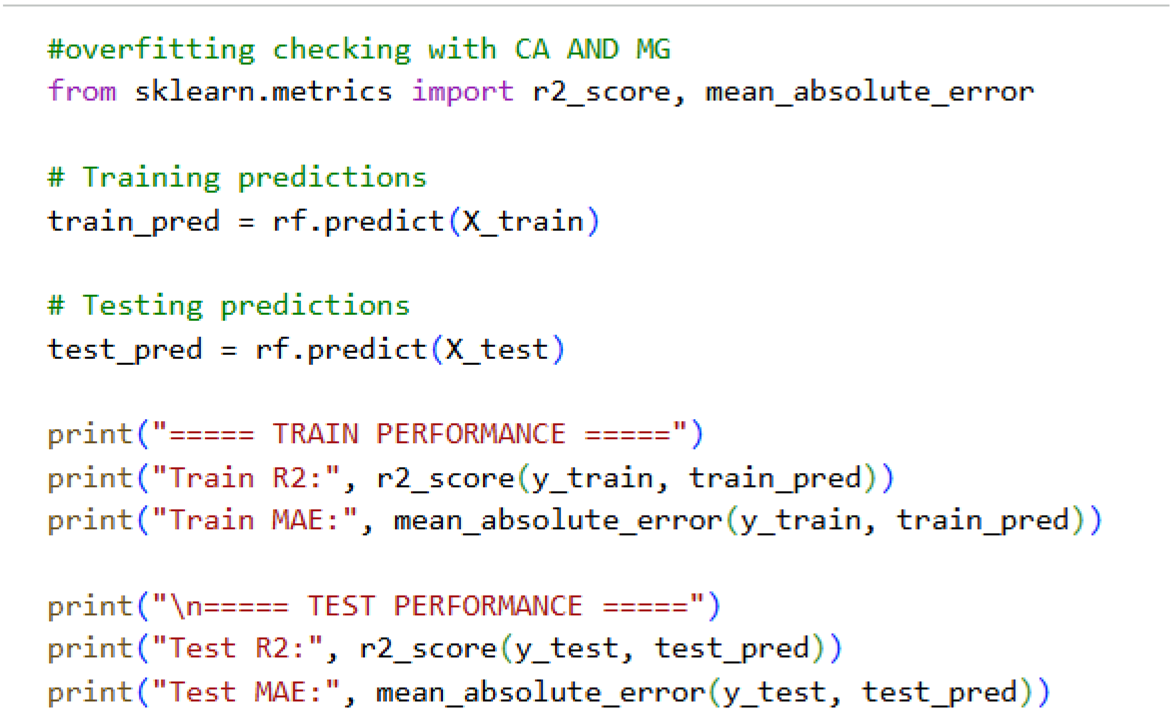
5.5.3 XGBoost Regressor



```
from xgboost import XGBRegressor

xgb = XGBRegressor(
    n_estimators=600,
    learning_rate=0.05,
    max_depth=8,
    subsample=0.9,
    colsample_bytree=0.9,
    reg_lambda=1,
    random_state=42
)

xgb.fit(X_train, y_train)
xgb_pred = xgb.predict(X_test)

print("\n===== XGBOOST RESULTS =====")
print("R2 Score:", r2_score(y_test, xgb_pred))
print("MAE:", mean_absolute_error(y_test, xgb_pred))
```

Fig 5.5.3: XGBoost Regressor

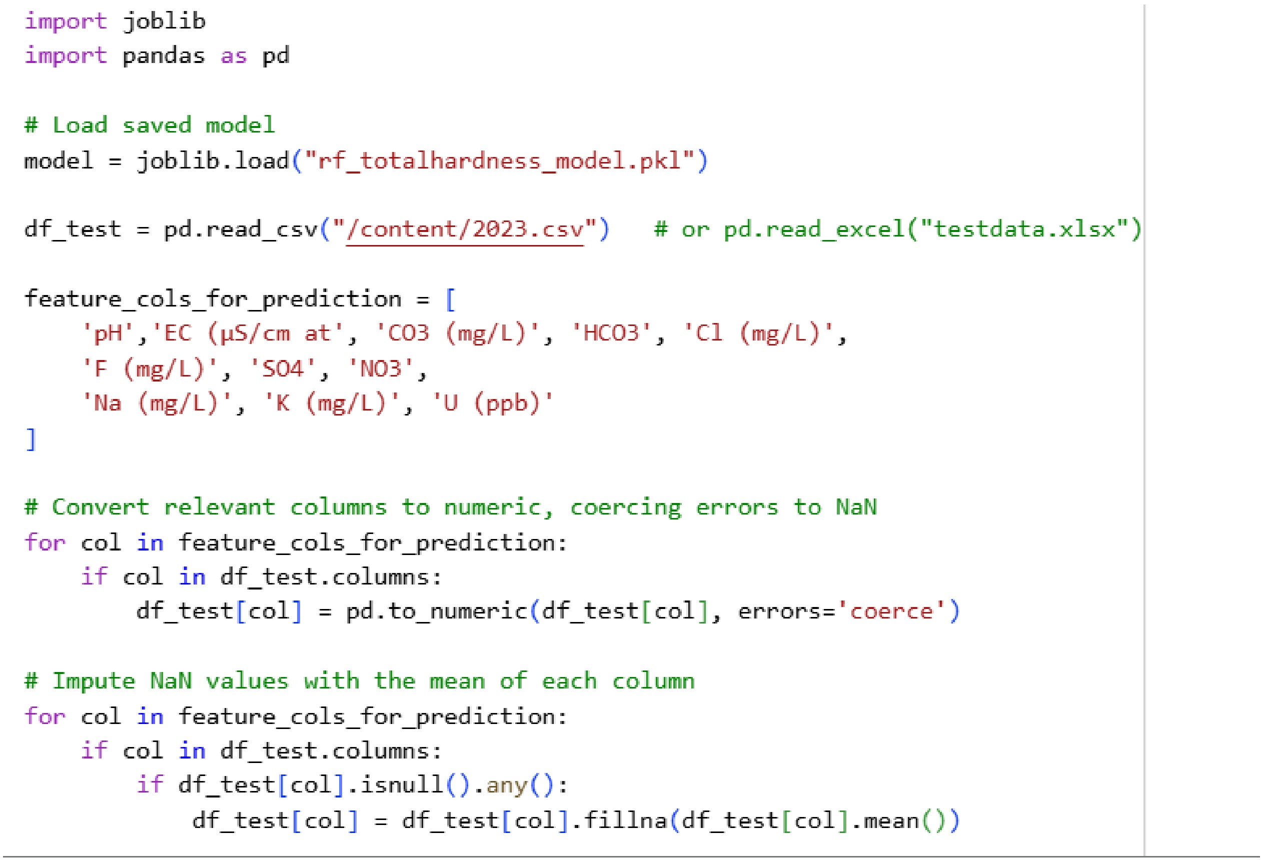
5.6 OVERFITTING CHECK



```
#overfitting checking with CA AND MG
from sklearn.metrics import r2_score, mean_absolute_error

# Training predictions
train_pred = rf.predict(X_train)

# Testing predictions
test_pred = rf.predict(X_test)

print("===== TRAIN PERFORMANCE =====")
print("Train R2:", r2_score(y_train, train_pred))
print("Train MAE:", mean_absolute_error(y_train, train_pred))

print("\n===== TEST PERFORMANCE =====")
print("Test R2:", r2_score(y_test, test_pred))
print("Test MAE:", mean_absolute_error(y_test, test_pred))
```

Fig 5.6.1: Overfitting check by comparing train and test predictions

5.7 PREDICTION ON NEW UNSEEN TEST DATASET



```
import joblib
import pandas as pd

# Load saved model
model = joblib.load("rf_totalhardness_model.pkl")

df_test = pd.read_csv("/content/2023.csv")   # or pd.read_excel("testdata.xlsx")

feature_cols_for_prediction = [
    'pH','EC (µS/cm at', 'CO3 (mg/L)', 'HCO3', 'Cl (mg/L)',
    'F (mg/L)', 'SO4', 'NO3',
    'Na (mg/L)', 'K (mg/L)', 'U (ppb)'
]

# Convert relevant columns to numeric, coercing errors to NaN
for col in feature_cols_for_prediction:
    if col in df_test.columns:
        df_test[col] = pd.to_numeric(df_test[col], errors='coerce')

# Impute NaN values with the mean of each column
for col in feature_cols_for_prediction:
    if col in df_test.columns:
        if df_test[col].isnull().any():
            df_test[col] = df_test[col].fillna(df_test[col].mean())
```

Fig 5.7.1: Prediction on 2023 Dataset

Explanation:

The code first loads the trained Random Forest model and then imports the new test dataset. It cleans the data by converting all necessary columns to numeric values and replacing any missing entries with the average of that column. After preparing the data, the model is used to generate Total Hardness predictions for the new samples.

5.8 POWER BI DASHBOARD DEVELOPMENT

The last stage of the implementation was to convert the machine learning predictions and the hydro-chemical data into an interactive decision-support system with the help of Microsoft Power BI. The process was carried out in three consecutive steps, which were: Data Integration, Modeling (DAX) and Report Architecture.

5.8.1 Data Integration and Transformation.

- The cleaned data (2020-2022clean.csv) and the file with the results of the prediction (2023_Predicted.csv) were included in the Power BI workspace.
- Query Editor: This was used to check the types of data (e.g., confirm that Latitude and Longitude were treated as Geospatial Data to be rendered on a map).
- Column Formatting: The chemical parameters were given appropriate scientific units (e.g., mg/L, µS/cm) so that the axes of visualizations were labeled correctly.

5.8.2 DAX Implementation and Data Modeling.

To facilitate dynamic analysis, Data Analysis Expressions (DAX) were authored to generate calculated measures and logic flags that change according to user slicers.

- Unsafe Location Counter: This is a measure that was invented to count dynamically wells that are not safe.

  Count_Unsafe = CALCULATE(COUNT(Rows), WQI_Level IN {"Poor", "Unsafe"})

- Dynamic Safety Classification: One conditional column was coded to classify the districts according to the level of Uranium (Safe < 30 ppb vs. Unsafe > 30 ppb).
- Traffic Light Logic: WQI levels (Green: Safe, Red: Unsafe) were mapped to hex color codes to provide standardized conditional formatting of all the charts.

5.8.3 Dashboard Architecture/Navigation.

The system was designed with three different analytical modules (pages), which were linked together through page navigator buttons, to create the appearance of an app-like interface:

- Page 1: Overview of the Regional Groundwater:
    - Written for executive summaries.
    - Implementation: The Geospatial Severity Heatmap was implemented based on latitude/longitude data to visualize the exact well locations. State and Year had slicers to enable temporal and regional filtering.
- Page 2: Hydro-Chemical Analysis:
    - Intended to be used scientifically in deep dives.
    - Implementation: Making Cation-Anion Balance charts (Stacked Bar Visuals) to visualize the correlation between Hardness (Ca/Mg) and Pollutants (Cl/SO4). The EC-Hardness correlation was confirmed by the addition of linear regression scatter plots.
- Pages 3-4: Toxic Contaminants and Safety:
    - Health risk assessment design.
    - Execution: Constructed Gauge Charts with set highests (e.g., Fluoride Max = 2.5 mg/L) to track a certain level of toxicity. A Horizontal Bar Chart was set with the settings of "Top 10 Worst Districts" ranked in order of Uranium concentration.

CHAPTER 6: RESULTS AND DISCUSSION

A number of significant analytical and predictive findings were generated by the developed Water Quality Prediction and Monitoring System. The Water Quality Index (WQI), feature engineering, machine learning modeling, and statistical analysis were used to arrive at these conclusions.

6.1 WATER QUALITY INDEX (WQI) COMPUTATION

- WQI values were calculated for all samples using the weighted arithmetic index method.
- The dataset showed a wide spread of WQI values, ranging from Excellent to Unsafe, indicating significant spatial variation in water quality.
- High WQI values were primarily associated with elevated levels of EC, Cl^-, SO_4^{2-}, Na^+, and Total Hardness.

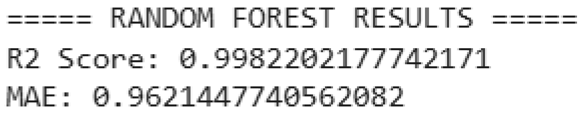
6.2 EXPLORATORY AND STATISTICAL DATA ANALYSIS

- The majority of chemical parameters, such as EC, Cl, Total Hardness, Ca, and Mg, were found to be right-skewed in distribution plots, suggesting the existence of high-value outliers.
- A large variation in ionic concentrations was confirmed by boxplots (log scale), indicating mixed hydro chemical characteristics across sampling locations.

The correlation matrix displayed:

- EC has a strong correlation with Cl (0.94), Na (0.91), SO_4 (0.77), and Total Anions (0.99).
- Ca (0.82) and Mg (0.79) have a strong correlation with total hardness.
- WQI has a moderate correlation with EC, Cl, SO_4, and total hardness, indicating that mineral makeup is the primary cause of declining water quality.
- The majority of variables showed little correlation with pH, suggesting that acidity and alkalinity stayed mostly constant.

6.3 MACHINE LEARNING BASED PREDICTION OF WATER QUALITY INDEX



```
===== RANDOM FOREST RESULTS =====
R2 Score: 0.9982202177742171
MAE: 0.9621447740562082
```

Description:

Using important parameters like pH, EC, HCO_3^-, Cl^-, F^-, SO_4^{2-}, NO_3^-, Ca^{2+}, Mg^{2+}, Na^+, and Total Hardness, a Random Forest regression model was trained to predict the Water Quality Index (WQI).
With an R2 score of **0.9982,** which indicates an almost perfect fit to the data, and a very low MAE of **0.96**, which indicates minimal prediction error, the model performed exceptionally well. These findings demonstrate that the chosen water quality parameters can be used to estimate WQI with high accuracy.

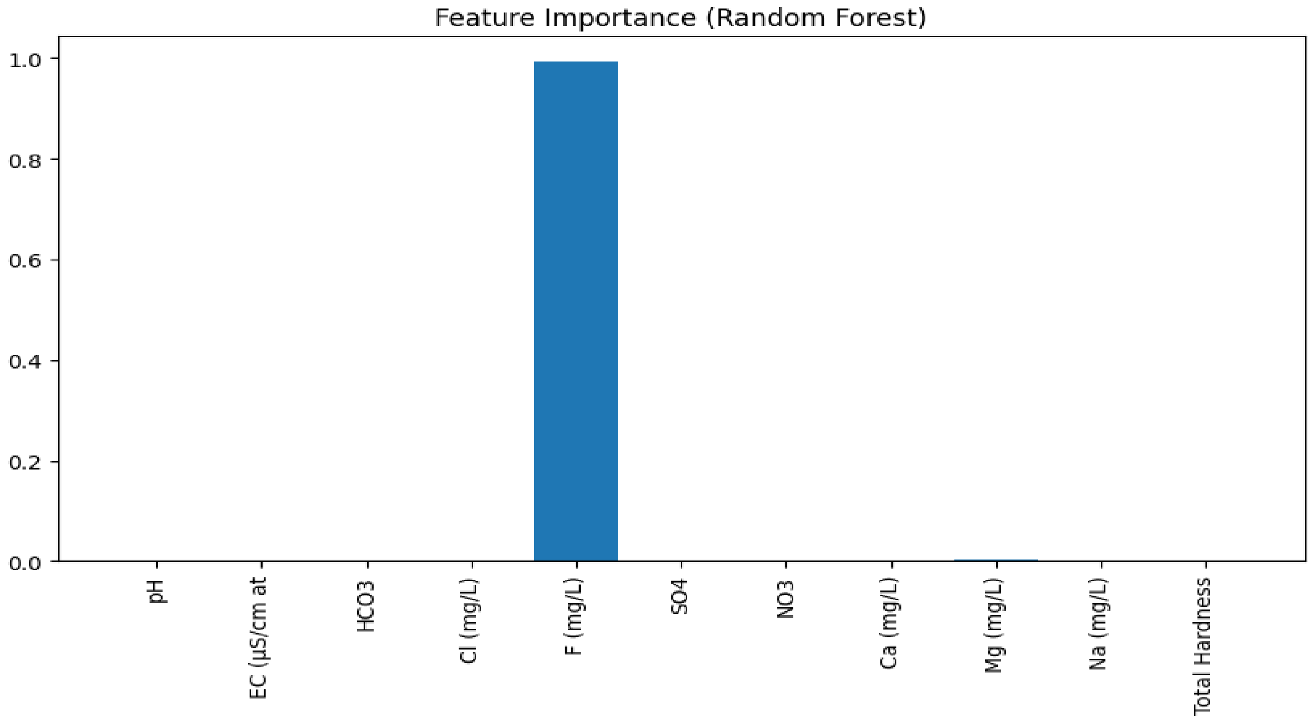
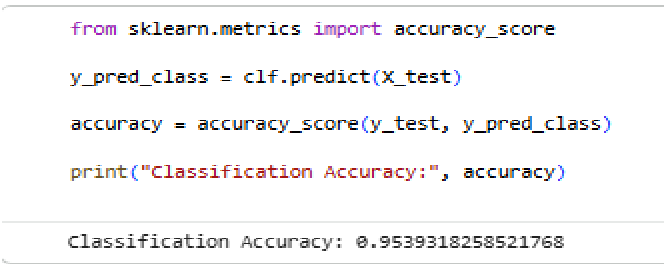


Fig 6.3.1: Feature Importance Graph for WQI

Description:

The feature importance plot for the Random Forest WQI prediction model shows that **Fluoride (F^-)** is the dominant predictor, contributing almost entirely to the model's decision-making. Other parameters such as pH, EC, Total Hardness, Ca, Mg, and major ions have **negligible influence** in comparison.
This indicates that, within the dataset used, variations in WQI are primarily driven by fluoride levels, while other chemical parameters contribute very minimally.



```
from sklearn.metrics import accuracy_score

y_pred_class = clf.predict(X_test)

accuracy = accuracy_score(y_test, y_pred_class)

print("Classification Accuracy:", accuracy)
```

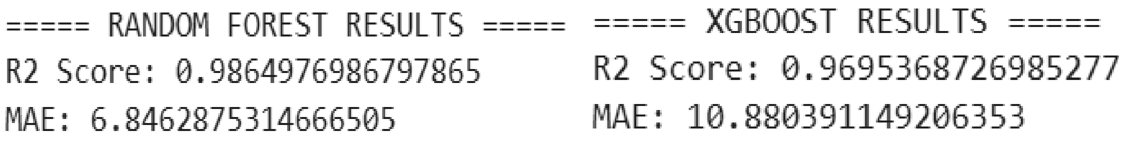
```
Classification Accuracy: 0.9539318258521768
```

Fig 6.3.2: Random Forest Classification Accuracy

Description:

A Random Forest classifier was trained to predict the WQI category (Excellent, Good, Poor, Very Poor, Unsafe) using the selected water quality parameters. The model achieved a **classification accuracy of 95.93%**, indicating that it can reliably classify water quality levels based on the chemical features provided.

6.4 MACHINE LEARNING BASED PREDICTION OF TOTAL HARDNESS



```
===== RANDOM FOREST RESULTS =====    ===== XGBOOST RESULTS =====
R2 Score: 0.9864976986797865         R2 Score: 0.9695368726985277
MAE: 6.8462875314666505              MAE: 10.880391149206353
```

Description:

A complete set of physicochemical parameters, such as pH, EC, carbonate/bicarbonate, major ions (Cl^-, SO_4^{2-}, NO_3^-), cations (Ca^{2+}, Mg^{2+}, Na^+, K^+), fluoride, and trace U levels, were used to train both XGBoost and Random Forest regression models to predict Total Hardness.

XGBoost demonstrated excellent predictive ability with an R2 of **0.9695** and an MAE of **10.88.**

Random Forest was the more accurate model for estimating Total Hardness, with an R2 of **0.986**5 and a lower MAE of **6.84**.

Overall, both models can accurately estimate Total Hardness from the chosen water quality parameters, with Random Forest performing better, as shown by the high R2 values and low prediction errors.

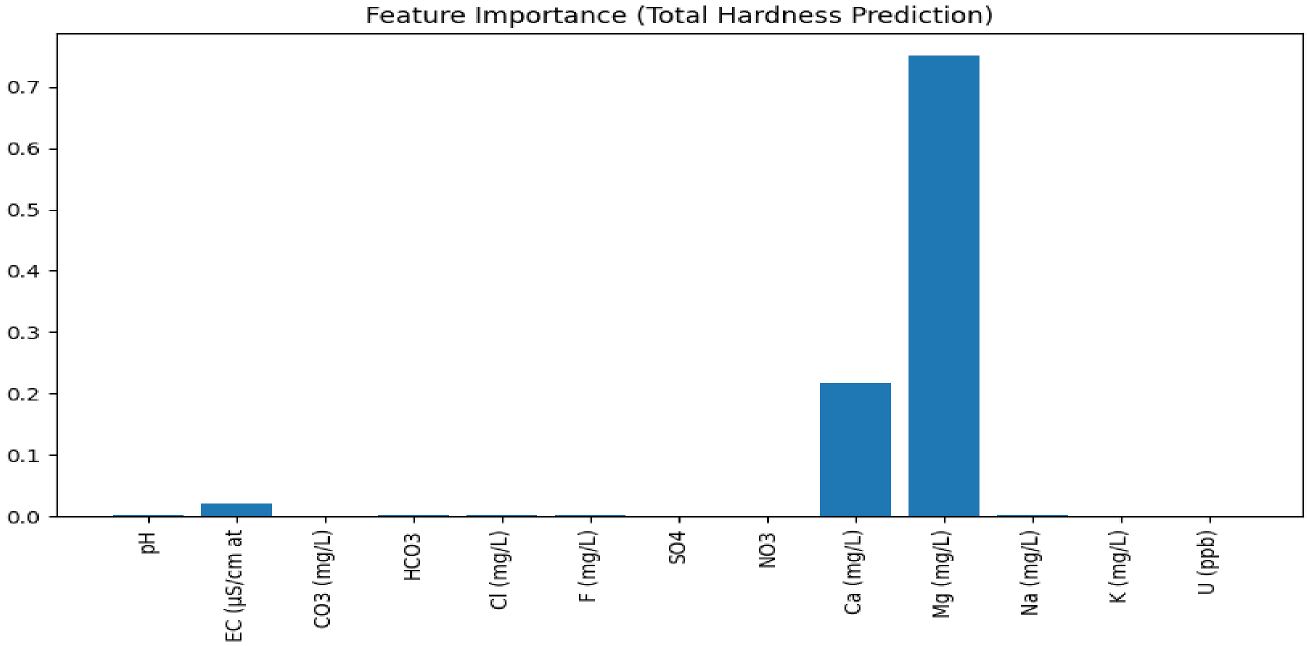
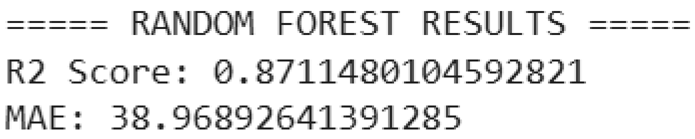


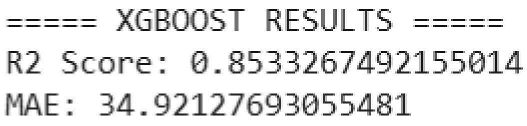
Fig 6.4.1: Feature Importance Graph for Total Hardness

Description:

The feature importance plot for Total Hardness prediction shows that Calcium (Ca^{2+}) and Magnesium (Mg^{2+}) are the dominant contributors, with magnesium having the highest importance. EC has a very small influence, while all other parameters including pH, chloride, sulphate, sodium, and nitrate contribute minimally. This confirms that Total Hardness is primarily controlled by Ca and Mg concentrations, consistent with standard hydrochemical principles.



```
===== RANDOM FOREST RESULTS =====
R2 Score: 0.8711480104592821
MAE: 38.96892641391285
```



```
===== XGBOOST RESULTS =====
R2 Score: 0.8533267492155014
MAE: 34.92127693055481
```

Description:

The Random Forest model trained without calcium and magnesium achieved an R^2 score of 0.87 and an MAE of 38.97 mg/L, XGBoost achieved R2 score of 0.85 and an MAE of 34.91 mg/L. This indicates that the model still predicts Total Hardness reasonably well using only indirect ionic indicators such as EC, HCO_3^-, Cl^-, SO_4^{2-}, NO_3^-, Na^+, K^+, and U levels. Although accuracy decreases compared to using Ca and Mg, the performance remains strong, showing that hardness can be estimated using low-cost parameters.

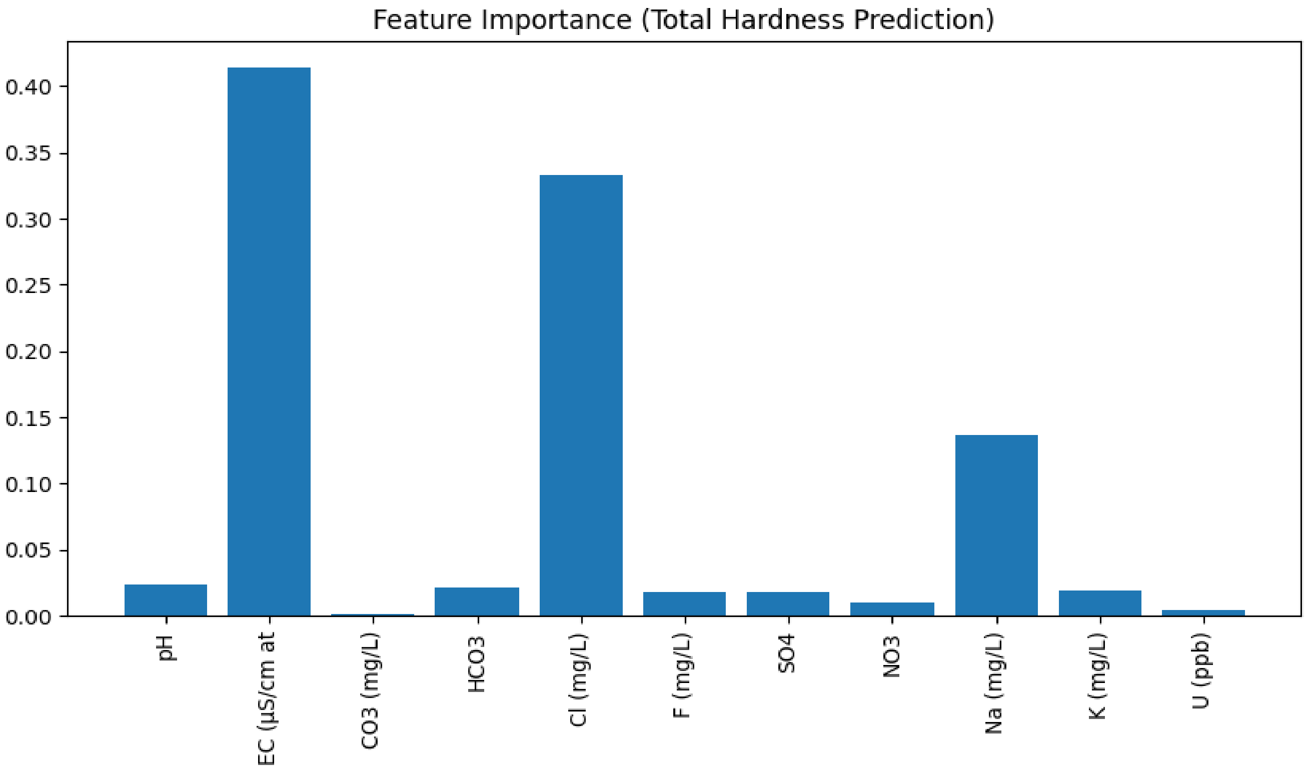


Fig 6.4.2 Feature Importance Graph for Total Hardness without CA and MG

Description:

The feature importance plot shows that Electrical Conductivity (EC) and Chloride (Cl^-) are the strongest predictors of Total Hardness when calcium and magnesium are removed. Sodium (Na^+) also contributes moderately. Other parameters such as pH, HCO_3^-, SO_4^{2-}, NO_3^-, fluoride, potassium, and uranium have minimal influence.

6.5 MODEL PERFORMANCE VISUALIZATION

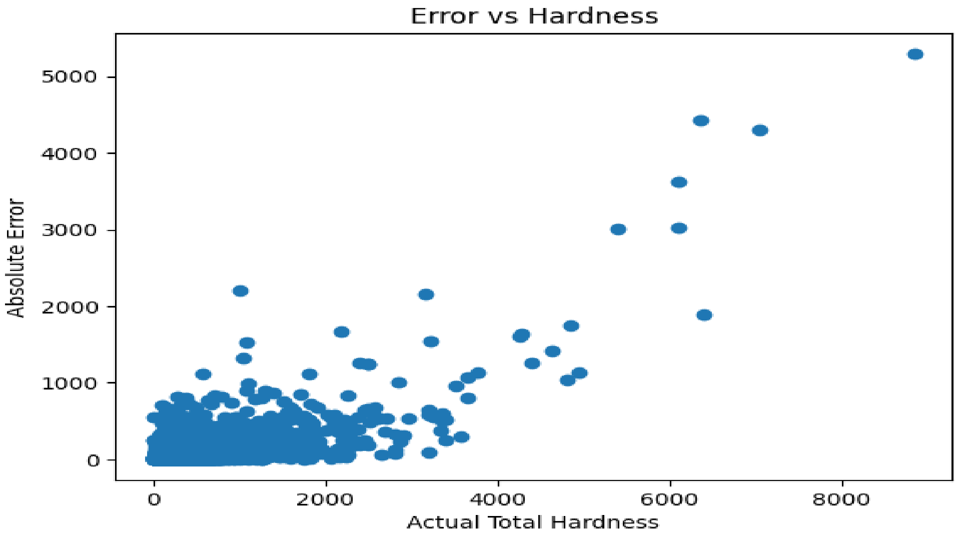


Fig 6.5.1: Error vs Hardness

Description:

The scatter plot shows how well the model's predictions match the actual Total Hardness values. Most of the data points fall close to the diagonal line, meaning the model is usually very accurate. While a few samples with very high hardness show larger differences, the overall pattern makes it clear that the model is doing a good job of estimating hardness across the dataset.

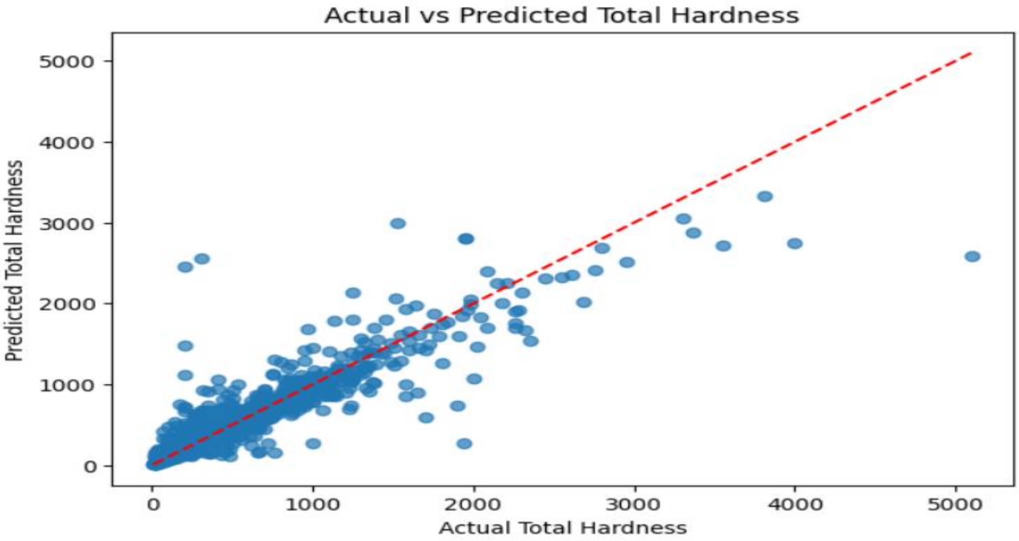


Fig 6.5.2: Actual vs Predicted Total Hardness

Description:

This plot compares the actual Total Hardness values with the values predicted by the model. The majority of the points lie close to the red diagonal line, showing that the model's predictions closely match the real measurements. A few points deviate at higher hardness levels, but overall the model demonstrates strong predictive accuracy.

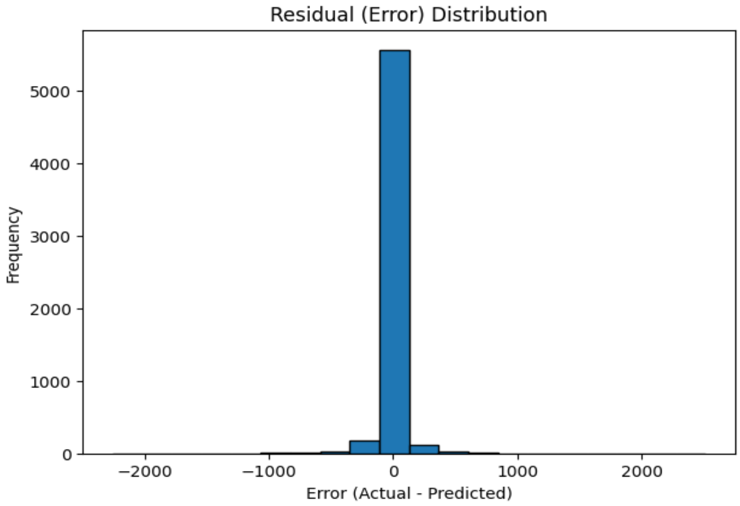


Fig 6.5.3: Error Distribution Plot

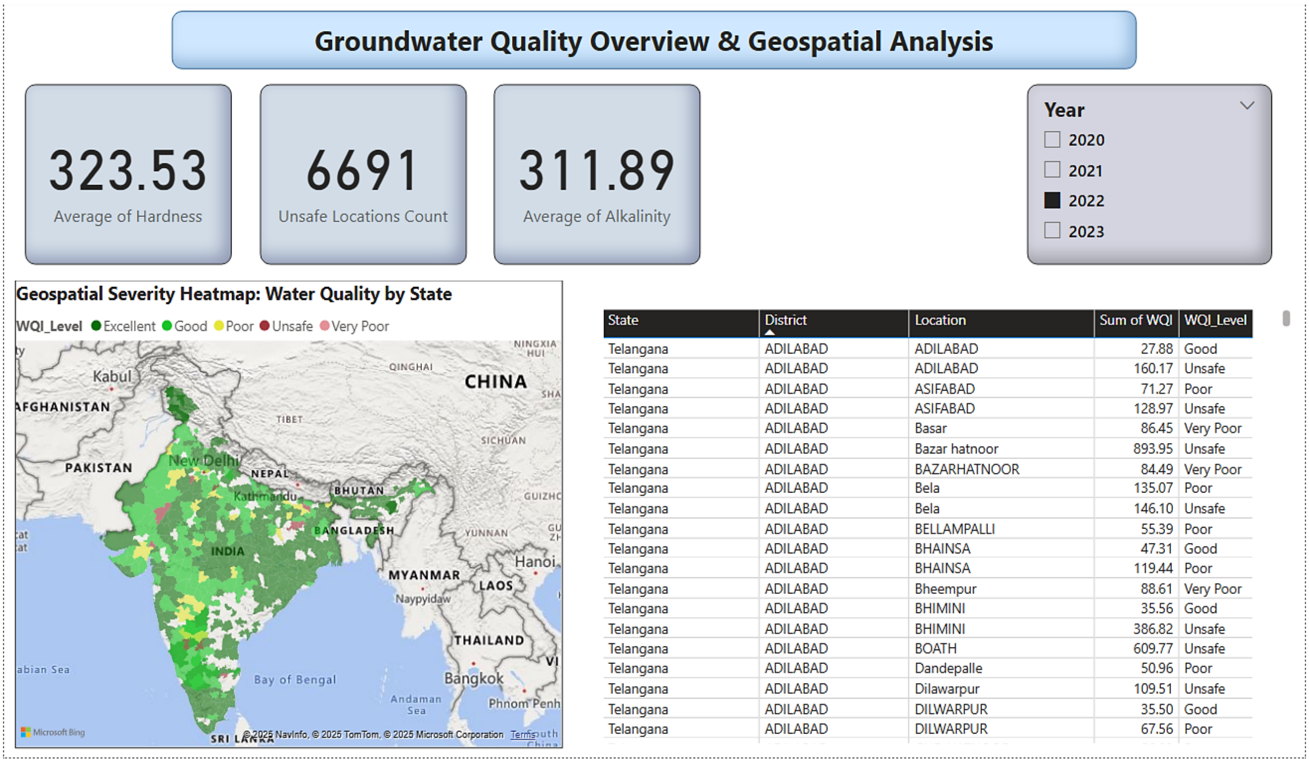
Description:

The residual distribution shows that most prediction errors are tightly clustered around zero, indicating that the model generally predicts Total Hardness accurately. Only a small number of samples show larger positive or negative errors, suggesting a few outliers. Overall, the narrow, centered distribution reflects good model performance and low systematic bias.

6.6 INTERACTIVE DASHBOARD ANALYSIS (POWER BI)

A key outcome of this project is the deployment of the "Groundwater Quality Monitoring System," an interactive dashboard that translates the ML predictions into visual insights. The results from the three reporting modules are discussed below.

6.6.1 Groundwater Quality Overview and Geospatial Analysis



Groundwater Quality Overview & Geospatial Analysis

323.53
Average of Hardness

6691
Unsafe Locations Count

311.89
Average of Alkalinity

Year
□ 2020
□ 2021
■ 2022
□ 2023

Geospatial Severity Heatmap: Water Quality by State

WQI_Level ● Excellent ● Good ● Poor ● Unsafe ● Very Poor

State	District	Location	Sum of WQI	WQI_Level
Telangana	ADILABAD	ADILABAD	27.88	Good
Telangana	ADILABAD	ADILABAD	160.17	Unsafe
Telangana	ADILABAD	ASIFABAD	71.27	Poor
Telangana	ADILABAD	ASIFABAD	128.97	Unsafe
Telangana	ADILABAD	Basar	86.45	Very Poor
Telangana	ADILABAD	Bazar hatnoor	893.95	Unsafe
Telangana	ADILABAD	BAZARHATNOOR	84.49	Very Poor
Telangana	ADILABAD	Bela	135.07	Poor
Telangana	ADILABAD	Bela	146.10	Unsafe
Telangana	ADILABAD	BELLAMPALLI	55.39	Poor
Telangana	ADILABAD	BHAINSA	47.31	Good
Telangana	ADILABAD	BHAINSA	119.44	Poor
Telangana	ADILABAD	Bheempur	88.61	Very Poor
Telangana	ADILABAD	BHIMINI	35.56	Good
Telangana	ADILABAD	BHIMINI	386.82	Unsafe
Telangana	ADILABAD	BOATH	609.77	Unsafe
Telangana	ADILABAD	Dandepalle	50.96	Poor
Telangana	ADILABAD	Dilawarpur	109.51	Unsafe
Telangana	ADILABAD	DILWARPUR	35.50	Good
Telangana	ADILABAD	DILWARPUR	67.56	Poor

Fig 6.6.1: Geospatial Overview Dashboard displaying National Groundwater Quality and Severity Heatmap

The Groundwater Quality Overview dashboard is a high-level monitoring tool that aims at policymakers who need to evaluate compliance with water safety in the region. Three high-impact Key Performance Indicators (KPIs) anchor the interface, giving an immediate view of the health of the ecosystem as the national Average Total Hardness, the cumulative Unsafe Locations Count, and the Average Alkalinity.

These indicators can be used to quickly determine the overall water quality change toward the BIS levels by either improving or worsening.

The main ingredient is the Geospatial Severity Heatmap that assigns states and districts a color depending on their level of Water Quality Index (WQI), with the highest level of Water Quality being the green color and the worst being the red color. This visual allows viewers to recognize geographical clusters of contamination instantly with no need to filter complicated data volumes. A Detailed Compliance Table, offering a highly detailed list of particular locations, arranged by the degree of contamination, supports this point. The site has interactive capabilities, including the Year Slicer that lets the user switch between datasets (2020-2023) to see historic trends, and the table can be drilled down to see which specific villages make up the "Unsafe" statistics.

The above figure is an example of dataset 2022, in which the dashboard of the snapshot of 2022 shows a critical overage in the mineral composition, where the Average Total Hardness (323.53 mg/L) exceeds desirable values. The granular data reveals that there are harsh foci of hotspots in the Adilabad district (Telangana) where the values of the WQI were extremely high (e.g., over 800), meaning that on the one hand, regional averages could seem constant, whereas on the other hand, it is clear that such a risk is acute by a few of the population since the foci are local.

6.6.2 Contamination And Health Safety

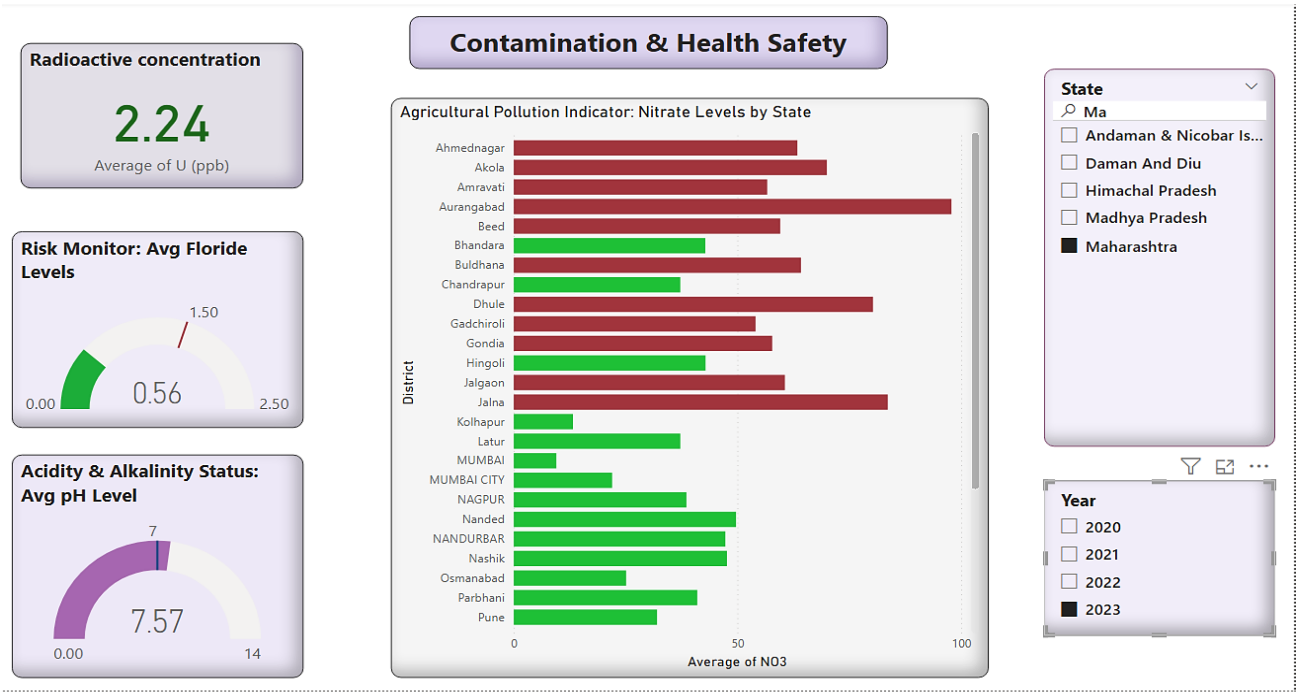
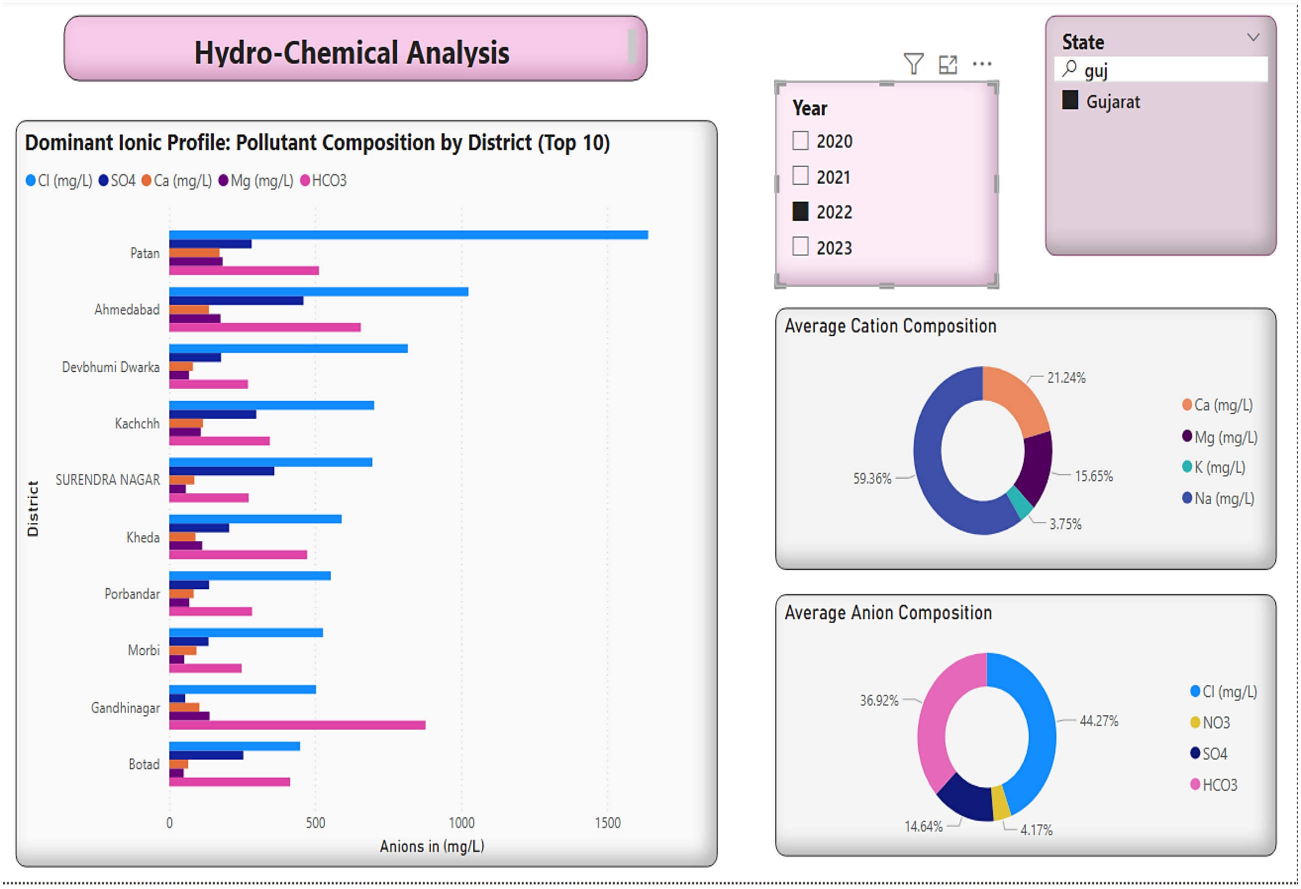


Fig 6.6.2: Toxic Contaminants & Health Safety Dashboard focusing on Chemical Toxicity and Agricultural Pollution.

The "Contamination & Health Safety" dashboard shifts the focus from general mineral content to specific toxicity factors that pose immediate health risks. The interface has two main Safety Gauges for Fluoride and pH. These gauges provide an instant "Health Check" against WHO/BIS standards. Health officials can quickly check if water is causing fluorosis (Skeletal Risk) or is corrosive (pH instability). There is also a dedicated Radioactive Concentration KPI that monitors Uranium levels, addressing the growing threat of geogenic contamination in granitic areas. The central visualization is the Agricultural Pollution Indicator, a bar chart that tracks Nitrate (NO3) levels by district. Since Nitrate is well-known as a "fertilizer footprint," this chart effectively identifies regions affected by excessive agricultural runoff or sewage contamination.

The above figure shows a clear divide between geogenic and human-made pollution in Maharashtra. While natural contaminants like Uranium (2.24 ppb) and Fluoride (0.56 mg/L) are well within safe limits, Nitrate pollution is severe. The districts of Aurangabad and Jalna show Nitrate levels above 80 mg/L, which is nearly double the safe limit of 45 mg/L (marked in Red). In contrast, urban coastal districts like Mumbai remain safe (Green), clearly linking intensive inland agriculture to groundwater issues.

6.6.3 Hydro Chemical Analysis



Year
□ 2020
□ 2021
■ 2022
□ 2023

State
guj
■ Gujarat

Fig 6.6.3: Hydro-Chemical Analysis Dashboard illustrating Ionic Balance and Pollutant Composition

The dashboard of the analysis of hydro-chemical gives an in-depth scientific approach to the exact ionic composition of the groundwater to further the planning and disposal of safety ratings, as it presents the factors that drive the contamination as the reasons. The report contains two major pie charts that provide an overview of the Average Cation and Anion Composition, which confirms the concept of electrical neutrality that is vital in the hydro-geological studies. The Cation chart indicates the domination of positive ions, including Sodium (Na) and Calcium (Ca), which proves that it is the salinity and hardness that will be the main geological forces affecting water quality in the regions under study. At the same time, the Anion chart indicates that the predominant negative ions are Chloride (Cl) and Bicarbonate (HCO3), which indicate a combination of both the saline effect and the natural alkalinity.

The key graphic, the Dominant Ionic Profile: Pollutant Composition by District, is the critical disaggregation of the sources of contamination of the top 10 most impacted districts. The chart separates anthropogenic and geogenic pollution by isolating certain ions present. For example, districts Patan and Ahmedabad have disproportionately large levels of Chloride (Blue bars), which indicates a strong salinity problem or more serious problems with industrial discharge. Other districts, on the contrary, exhibit a high Bicarbonate (Pink bars), which is a symptom of temporary hardness produced by the weathering of natural rocks. This type of granular chemical fingerprinting enables the authorities to select the appropriate remediation process, like Reverse Osmosis in the case of salinity and simple softening in the case of carbonate hardness, instead of using a one-size-fits-all approach.

CHAPTER 7: CONCLUSION AND FUTURE SCOPES

7.1 CONCLUSION

This project's Water Quality Prediction and Monitoring System effectively combines machine learning, scientific analysis, and interactive visualization to evaluate groundwater quality in an effective and data-driven way. The system successfully categorized water samples into groups ranging from Excellent to Unsafe by calculating the Water Quality Index (WQI), exposing notable regional differences in water quality.

Extensive exploratory data analysis through distribution plots, boxplots, and correlation matrices showed the impact of dissolved ions such as chloride, sulphate, sodium, calcium, and magnesium on both WQI and Total Hardness. Outlier analysis helped improve the reliability of model training and aided the identification of highly mineralized or contaminated zones.

Predictive analysis relied heavily on machine learning models. In the case of prediction of Total Hardness, Random Forest and XGBoost regressors were among the best performers, whereas Random Forest had the highest accuracy. Based on ionic indicators such as EC, chloride, sodium, and alkalinity, the model delivered a good predictive ability even when the calcium and magnesium parameters were excluded. High accuracy of WQI level classification by means of Random Forest showed the potential for automated water quality categorization. These results show that the system is suitable for low-cost field deployment because both hardness and WQI can be reliably estimated using properly measurable parameters.

Power BI dashboards greatly improved the presentation of the results through an interactive, real-time visual display. The resulting dashboards allowed exploration of trends in districts, comparison of chemical parameters, monitoring of WQI categories, and identification of hotspots of poor water quality. The integration of spatial mapping, charts, filters, and time-based analytics made it highly intuitive and useful to decision-makers, researchers, and water management authorities.

Overall, the project demonstrates that combining environmental data science with machine learning and dynamic visualization tools can create a comprehensive and scalable water monitoring solution. The system not only evaluates present water quality but also provides predictive capabilities that can support early warning systems, resource planning, and sustainable water management. With further enhancements such as IoT sensor integration or real-time data streaming, this solution can evolve into a powerful platform for continuous groundwater surveillance and public health protection.

7.2 FUTURE SCOPES

1. Integration of IoT Sensors for Real-Time Monitoring

- Low-cost IoT devices can be deployed in the field for continuous measurement of key parameters like pH, EC, temperature, turbidity, and dissolved ions.
- These live readings can be streamed to the system to generate instant WQI scores and alerts.

2. Development of Real-Time Alert and Notification System

- Threshold-based alarms can be built to notify users whenever water quality becomes bad or unsafe.
- Such notifications can be delivered through SMS, mobile apps, and email in order to increase community awareness and enable prompt action.

3. Predictive Forecasting of Water Quality Trends

Time-series modeling can be implemented to predict:

- Seasonal variation in hardness
- Long-term groundwater quality trends
- Probability of contamination events

4. Extension to Surface Water Bodies

The methodology can be extended to rivers, lakes, ponds, and reservoirs for broader environmental applications.

5. Mobile and Web-Based Application Development

- An easy-to-use mobile or web-based platform will be developed, through which citizens, researchers, and authorities are able to:
- Upload data View dashboards Check WQI instantly Access the water quality reports region-wise.

REFERENCES:

1. Brown, R. M., McClelland, N. I., Deininger, R. A., & Tozer, R. G. (1970). *A Water Quality Index—Do We Dare?* Water & Sewage Works, 117(10), 339–343.

2. Horton, R. K. (1965). *An Index Number System for Rating Water Quality.* Journal of Water Pollution Control Federation, 37(3), 300–306.

3. Bureau of Indian Standards (BIS). (2012). *Indian Standard IS 10500:2012 – Drinking Water Specification.* New Delhi, India.

4. World Health Organization (WHO). (2017). *Guidelines for Drinking-Water Quality* (4th ed.). Geneva: WHO Press.

5. Tyagi, S., Sharma, B., Singh, P., & Dobhal, R. (2013). Water Quality Assessment in Terms of Water Quality Index. *American Journal of Water Resources*, 1(3), 34–38.

6. Singh, S., & Kamal, V. (2017). Assessment of Groundwater Quality Using WQI Method. *International Journal of Environmental Sciences*, 7(2), 45–52.

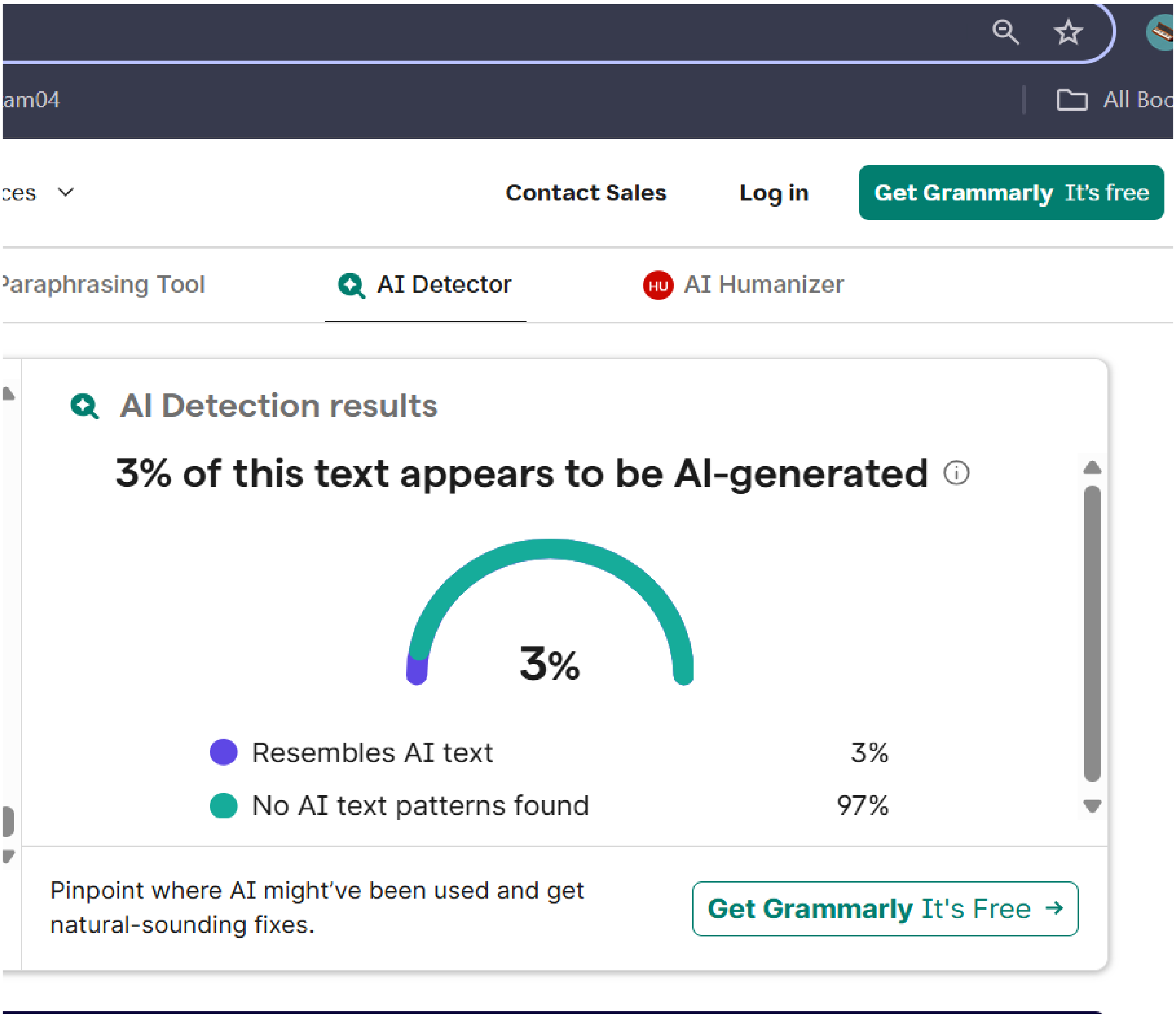
7. Kouadri, F., & Chabaca, M. (2020). Predicting Water Quality Using Machine Learning Techniques. *Journal of Hydrology*, 586, 124881.

8. Li, L., et al. (2021). Machine Learning Based Prediction of Groundwater Quality Parameters. *Environmental Monitoring and Assessment*, 193(6), 1–16.

9. Jha, M. K. (2019). Experimental and Machine Learning-Based Modeling of Water Quality. *Journal of Water and Health*, 17(5), 754–768.

10. Central Ground Water Board (CGWB). “Ground Water Quality.” Accessed via CGWB official portal. Available at: https://cgwb.gov.in/en/ground-water-quality

PLAGIARISM CHECK:



AI Detection results

3% of this text appears to be AI-generated

3%

● Resembles AI text	3%
● No AI text patterns found	97%

Pinpoint where AI might've been used and get natural-sounding fixes.