Analysis and Prediction of Water Quality and Contaminant levels in River Ganga and their Effect on Surface Water.

Kkrishna Saxena & Laasya Reddy
Registration No - 210953172 & 210953170

Abstract

The Ganga River spans across nearly a quarter of India's landmass (Koshy), nurturing a rich tapestry of ecosystems within this densely populated territory. This research assesses Ganga River water quality and its impact on surface water by employing various machine learning models to predict contamination levels. Utilizing data spanning 2017-2022 containing crucial parameters like Dissolved Oxygen (DO), Biological Oxygen Demand (BOD), Fecal Coliform (FC), Faecal Streptococci (FS), pH, station details, and years, models such as Linear Regression, Random Forest, Deep Learning, and Polynomial Regression are evaluated. Evaluation metrics such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), R-squared, and accuracy percentages provide insights into model performance. The study's findings highlight the models' effectiveness in comprehending water quality dynamics and their significance in informing environmental management strategies for preserving the Ganga River.

Keywords : Ganga River, Water quality Index, Contamination prediction, Surface Water

Chapter 1. Introduction

The Ganga River, often revered as the lifeline of India, holds profound ecological, cultural, and socioeconomic significance. Spanning nearly a quarter of India's landmass, this iconic river traverses a diverse landscape, supporting various ecosystems within an intricately interconnected environment. However, the sustenance of these ecosystems faces a formidable challenge due to escalating pollution levels stemming from anthropogenic activities.

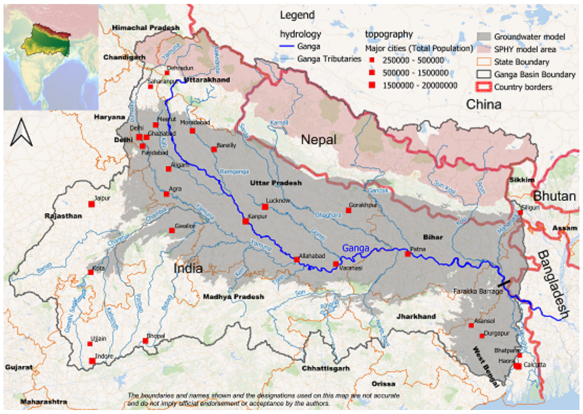


Figure 1. Ganga river basin map. Background based on Wikimedia unlabeled layer. (Vat)

This research endeavors to comprehensively assess the water quality of the Ganga River and its consequential impact on surface water within a densely populated region as shown in Figure 1. The study utilizes a robust dataset sourced from data.gov.in, spanning the years 2017 to 2022 to employ Random Forest, Linear and Polynomial Regression models. This dataset encompasses an extensive array of vital water quality parameters, including Dissolved Oxygen (DO), Biological Oxygen Demand (BOD), Fecal Coliform (FC), Faecal Streptococci (FS), pH levels, station details, and other pertinent attributes.

Due to the lack of recent and comprehensive data specifically designed for deep learning techniques, we utilized an alternative dataset spanning 1996 to 2006. Although this dataset differs in time from the present, it serves as a foundational resource for both training and evaluating our deep learning model. The dataset includes information on various surface contaminant quantities such as station codes, types, states, districts, and extensive number of contaminants and their quantities.

This choice was made because there is a scarcity of recent extensive data specifically suited for deep learning applications. The adopted dataset, while temporally different, allows us to develop and assess our deep learning model for predicting future surface contaminant levels based on historical data. This adaptive strategy allows us to make meaningful predictions despite the temporal gap in the data, offering a pragmatic solution for addressing the challenges posed by the availability of relevant information.

The compilation of the primary dataset from data.gov.in ensures a credible and standardized source, directly obtained from the Government of India, laying the groundwork for a rigorous and meticulous analysis. Citations from established sources and governmental repositories substantiate the authenticity and reliability of the data employed in this study. Furthermore, scholarly works, scientific publications, and environmental reports augment our understanding of the Ganga River's historical context, environmental challenges, and the necessity for comprehensive water quality assessments.

By using these datasets, this research aims to employ various machine learning models, including Linear Regression, Random Forest, Deep Learning, Polynomial Regression, and Support Vector Regression, to predict contamination levels. These models play a pivotal role in extrapolating insights into water quality dynamics, facilitating informed decision-making for environmental management strategies aimed at preserving and rejuvenating the Ganga River.

In summary, this paper seeks to delve into the intricacies of Ganga River water quality, leveraging datasets sourced from data.gov.in. Through the amalgamation of empirical data, credible citations, and advanced analytical methodologies, this research aims to offer a comprehensive understanding of water quality parameters and their implications for the Ganga River's ecosystem, paving the way for informed and sustainable conservation efforts

Chapter 2. Methodology

2.1. Dataset Description:

Our first dataset, which was used to run Linear and polynomial regression, support vector machine and random forest algorithms encompasses various attributes, each contributing essential information for the evaluation of environmental conditions.
The "Year" attribute signifies the temporal dimension spanning 2017 to 2022, indicating the specific year of observation. The "Station Code" provides a unique identifier for the monitoring station, aiding in precise referencing. The "State" attribute delineates the geographical location of the station. The "Station Name" offers a contextual identifier, specifying the monitored location.
Crucial water quality parameters are detailed in attributes such as Dissolved Oxygen levels, Biological Oxygen Demand, Fecal Coliform and Faecal Streptococci counts per 100 milliliters and pH level. This dataset, sourced from data.gov.in, underpins our exploration into machine learning algorithms, serving as a foundation for understanding the intricate relationships among these parameters and facilitating predictive analyses for water quality assessments.

The dataset under consideration for employing the Deep Learning Model spans surface water quality measurements from various monitoring stations, covering the years 1996 to 2006 and was also sourced from data.gov.in.

The dataset encompasses diverse attributes, including station-specific details such as codes, names, and types, along with geographical information such as state, district, block, basin, and sub-basin names. Additionally, temporal aspects are captured through date collection records.

The dataset provides a comprehensive array of contaminant quantities, featuring measurements of various elements and compounds like silver, aluminum, arsenic, boron, barium, biochemical oxygen demand, calcium, cadmium, chlorides, cyanide, and numerous others. Specific parameters related to water quality, such as pH, dissolved oxygen levels, electrical conductivity, turbidity, fecal streptococci and coliforms are also included.

The dataset comprises a rich set of information, ranging from chemical compositions to biological indicators. This detailed dataset serves as a valuable resource for understanding and predicting surface contaminant levels in surface water bodies, facilitating the development and evaluation of deep learning models for water quality assessment and prediction of River Ganga.

2.2. Data Preprocessing:

For the datasets employed in our analysis, an essential preprocessing stage was undertaken to ensure the integrity and usability of the data for subsequent modeling on both datasets.

Firstly, we examined the dataset spanning the years 2017 to 2021, focusing on river water quality parameters. Following the initial loading of the dataset using Pandas (“pandas documentation — pandas 2.1.3 documentation”), we executed a crucial step to handle any potential non-numeric values within the dataset. Employing the Pandas apply function, we coerced any non-numeric entries to NaN, effectively homogenizing the data type. Subsequently, any missing values were imputed by replacing them with the mean of their respective columns. This process ensures a consistent and complete dataset, laying the groundwork for the application of various machine learning models such as Linear Regression, Polynomial Regression, Random Forest, and SVM.

Similarly, for the second dataset spanning the years 1996 to 2006, which encompasses over 8000 entries, a more comprehensive preprocessing pipeline was implemented. Initially, the dataset was loaded using Pandas with specific encoding to accommodate its characteristics. To enhance data quality, columns with more than 40% null values were systematically removed, preserving the structural integrity of the dataset. Non-numeric values were systematically converted to NaN, and any missing values were imputed by replacing them with the mean, maintaining data consistency. Subsequently, we curated a list of contaminants from the columns of interest, focusing on the 10th column onward, since the first 10 columns are station, location and temporal related attributes. This preprocessing sequence, ensures the dataset's readiness for the subsequent application of deep learning models. These steps are a necessary approach to prepare diverse datasets for meaningful analysis and modeling.

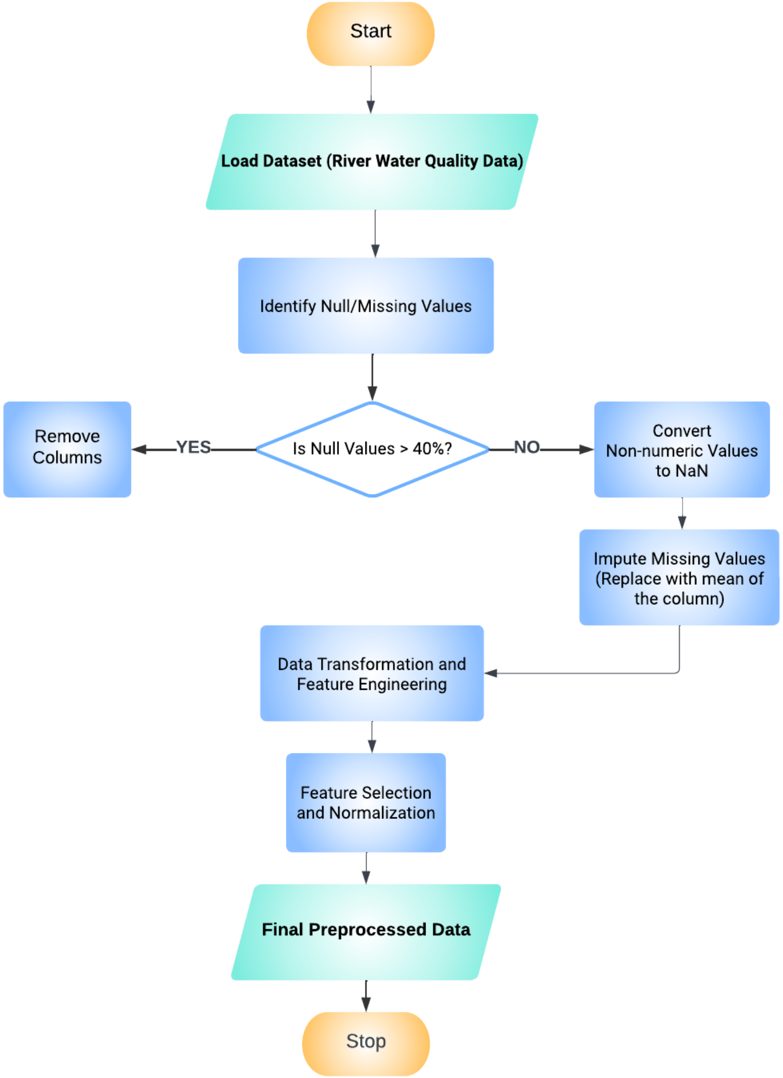

Start
Load Dataset (River Water Quality Data)
Identify Null/Missing Values
Remove Columns
YES
Is Null Values > 40%?
NO
Convert Non-numeric Values to NaN
Impute Missing Values (Replace with mean of the column)
Data Transformation and Feature Engineering
Feature Selection and Normalization
Final Preprocessed Data
Stop

Figure 2.2.1 Flow diagram

2.3. Machine Learning Models:

The *Random Forest Regressor*, a powerful machine learning model employed in our project, operates by constructing many decision trees during the training phase. Each tree is built using a subset of the dataset's features and samples. During prediction, the model aggregates the outputs of these individual trees, yielding a robust and accurate estimation. In our specific implementation, we utilized key water quality parameters—Dissolved Oxygen (DO), Biological Oxygen Demand (BOD), Fecal Coliform (FC), Faecal Streptococci (FS), and pH—as features for predicting contaminant levels. The model's hyperparameters, such as the number of trees (n_estimators), maximum depth of the trees (max_depth), and minimum samples required to split an internal node (min_samples_split), were fine-tuned to optimize performance. The accuracy of the model was assessed using the Mean Absolute Error (MAE), providing a clear indication of prediction quality. Visualizing the results through a scatter plot further elucidates the model's predictive capabilities, showcasing the alignment between actual and predicted water contamination levels.

Linear Regression serves as a foundational model, representing the relationship between features and contamination levels through a straight line. It simplifies the analysis but may overlook intricate patterns. *Polynomial Regression* introduces flexibility by incorporating curves into the line, the degree of which is determined by the polynomial's degree. In this section, both models were applied to the same dataset, considering parameters like Dissolved Oxygen, Biological Oxygen Demand, Fecal Coliform, Faecal Streptococci, and pH. The objective was to evaluate and compare their predictive capabilities for the year 2022. We constructed scatter plots to obtain a visual comparison between predicted and actual values, aiding in the assessment of each model's efficacy.

Deep Learning operates as a potent tool in our water contamination prediction project, demonstrating its proficiency in discerning intricate patterns within the dataset. This method leverages a neural network architecture, specifically a Sequential model, designed for regression tasks. The neural network is constructed with an input layer and a hidden layer containing 120 neurons utilizing the Rectified Linear Unit (ReLU) activation function. The model's output layer, tailored for regression, consists of a single neuron. Over the course of 100 epochs, the model learns from the training data, refining its internal parameters to capture the underlying relationships between the selected features and the water contamination levels. Deep Learning approach harnesses the capacity of neural networks to automatically learn complex relationships within the water quality dataset.

2.4. Model Training, Evaluation and Validation:

Our Random Forest model used a state-wise river water quality dataset from 2017 to 2022 divided into training and testing sets using a 90:10 ratio. The training involved creating an ensemble of decision trees, and the model's effectiveness was evaluated by predicting contamination levels for the year 2022. We assessed its performance using Mean Absolute Error (MAE),

$$MAE = \frac{\sum_{i=1}^{n}\left|y_i - x_i\right|}{n} = \frac{\sum_{i=1}^{n}\left|e_i\right|}{n} \text{ (“Mean absolute error”)}$$

MAE = mean absolute error
y_i = prediction
x_i= true value
n= total number of data points

a measure of prediction accuracy. Visualizing the predicted versus actual contamination levels provided the way to validate the model's reliability. A scatter plot with actual and predicted values was constructed for validating the results produced by our model.

Both Linear and Polynomial Regression models were implemented on the 2017-2021 dataset divided into training and testing sets using an 80:20 ratio, focusing on capturing linear and non-linear relationships between water quality parameters. Evaluation centered around predicting contamination levels for the year 2022 and comparing them with actual values. For validation we use Mean Absolute Error (MAE) served as a key metric for assessing the accuracy of predictions, with visual representations aiding in understanding the models' effectiveness.

The Deep Learning model, implemented through a neural network architecture, showcased its ability to discern intricate patterns in the extensive 1996-2006 dataset. Trained over 100 epochs, the model learned the complex relationships between selected features and contamination levels. Evaluation metrics included Root Mean Squared Error (RMSE) and R-squared, providing insights into prediction accuracy.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}\left(x_i - \hat{x}_i\right)^2}{N}}$$ (“Root-mean-square deviation”)

$RMSE$ = root-mean-square deviation
i = variable i
N = number of non-missing data points
x_i = actual observations time series
$\widehat{x}_i$ = estimated time series
The model's practical applicability was further assessed by calculating the accuracy percentage within a specified tolerance level for validation.

2.5. Software and Tools:

For this data-driven investigation into water contamination, we relied on various essential tools and libraries. *Pandas*, used for data manipulation in Python, was essential for seamless data handling and preprocessing. We opted for Python due to its versatility and extensive support in data science.

In the data prep phase, tools like *NumPy* ("NumPy user guide") and *Scikit-learn* ("scikit-learn 1.3.2 documentation") played key roles. NumPy, a core package for scientific computing in Python, helped with numerical operations, while Scikit-learn provided standardized preprocessing methods and feature scaling. *Matplotlib* came in handy for visualizing data distribution and trends.

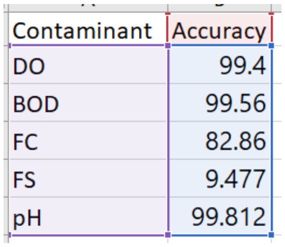
For model development, our choice of libraries aligned with the algorithms we chose to implement. *Scikit-learn* facilitated the implementation of traditional machine learning models (Random Forest, Linear Regression, Polynomial Regression). For the more complex Deep Learning model, *Keras* (built on *TensorFlow*) offered a high-level neural network API. This comprehensive toolkit enabled us to explore datasets thoroughly and deploy diverse models for robust analysis.

Chapter 3. Results and Discussions

3.1. Linear Regression:

In the evaluation of Ganga River water contaminants using linear regression models, several parameters exhibited noteworthy predictive accuracies. Dissolved Oxygen (DO), Biological Oxygen Demand (BOD), and pH levels showcased highly accurate predictions, with the models achieving accuracies of approximately 99.4%, 99.56%, and 99.81%, respectively. These results suggest a robust alignment between the model's estimations and the actual concentrations of these contaminants, highlighting the efficacy of linear regression in forecasting these crucial water quality indicators.

Conversely, the models for Fecal Coliform (FC) and Faecal Streptococci (FS) displayed significantly lower predictive accuracies. The FC model exhibited an accuracy of around 82.86%, whereas the FS model's accuracy drastically declined to a mere 9.48%. It's probable that the linear regression model might not capture the intricate relationships or hidden patterns inherent in the FS data due to its complex nature. Faecal Streptococci, being a bacteriological parameter, might involve multifaceted interactions affected by numerous unaccounted environmental and ecological factors. These outcomes indicate notable discrepancies between predicted and actual contaminant levels, emphasizing the limitations of linear regression in effectively estimating FC and FS concentrations based on the provided parameters. These results underscore the complexity of predicting these particular contaminants solely through the chosen attributes and suggest the need for alternative modeling techniques or inclusion of additional influential variables to enhance predictive accuracy.



Contaminant	Accuracy
DO	99.4
BOD	99.56
FC	82.86
FS	9.477
pH	99.812

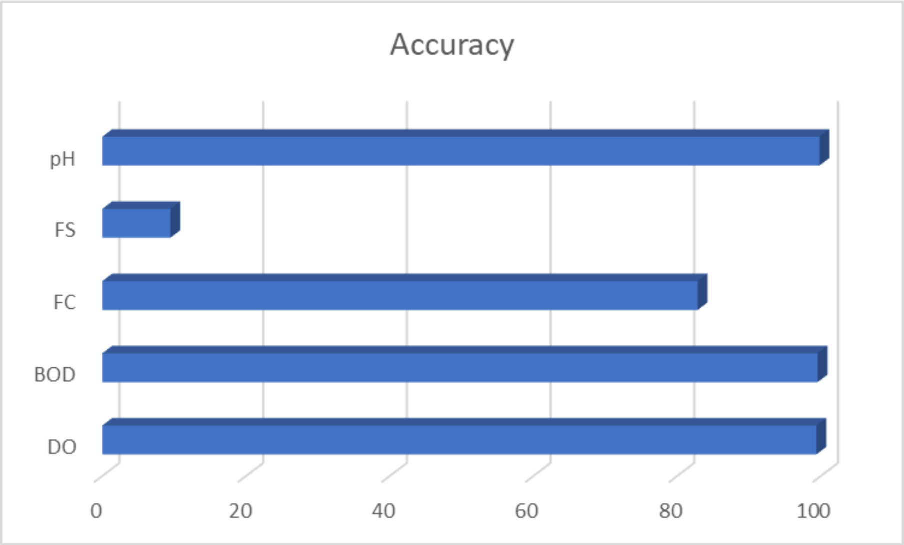


Figure 3.1.1. Performance of Linear Regression on each contaminant prediction

3.2. Random Forest:

The Random Forest Regressor exhibited robust predictive capabilities across various contaminants within the Ganga River water quality dataset. Dissolved Oxygen (DO) and pH demonstrated exceptional accuracy percentages of approximately 99.13% and 99.57%, respectively, showcasing the model's precision in estimating these crucial parameters. These high accuracies underscore the model's effectiveness in understanding their pivotal roles in assessing water quality and environmental health. Additionally, Biochemical Oxygen Demand (BOD) achieved a remarkable accuracy of 97.46%, emphasizing the model's capability in evaluating organic pollution levels accurately.

Furthermore, the model performed admirably for Faecal Streptococci (FS) with an accuracy of approximately 94.06%, indicating its adeptness in predicting this contaminant, which is significant for understanding microbial pollution. While Fecal Coliform (FC) showed a slightly lower accuracy of around 80.87%, this moderate result suggests the complexity in accurately estimating FC levels in the river's ecosystem. Nonetheless, overall, the Random Forest Regressor demonstrated strong predictive potential for most contaminants, providing valuable insights into the Ganga River's water quality profile. These results highlight the model's ability to discern and predict the impact of various contaminants, contributing significantly to the understanding of the river's environmental health.

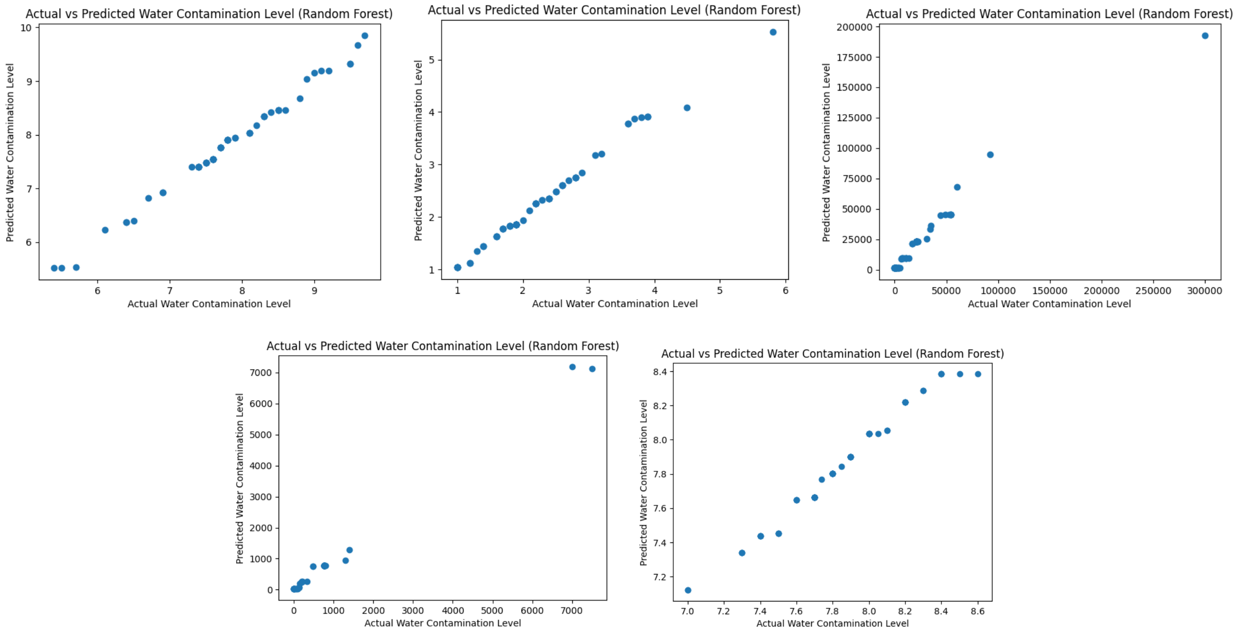
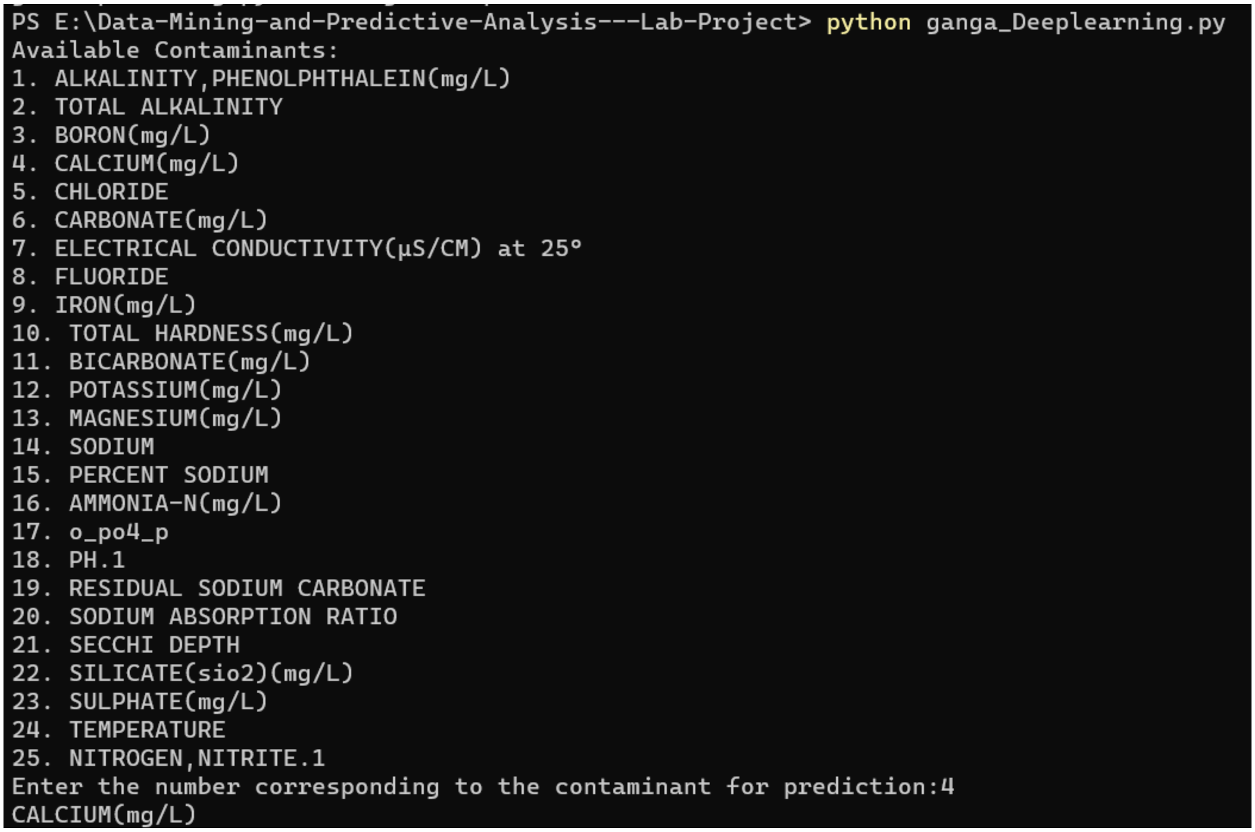


Figure 3.2.1 to 3.2.5 Performance of Random Forest Regressor on each contaminant

3.3. Deep Learning:

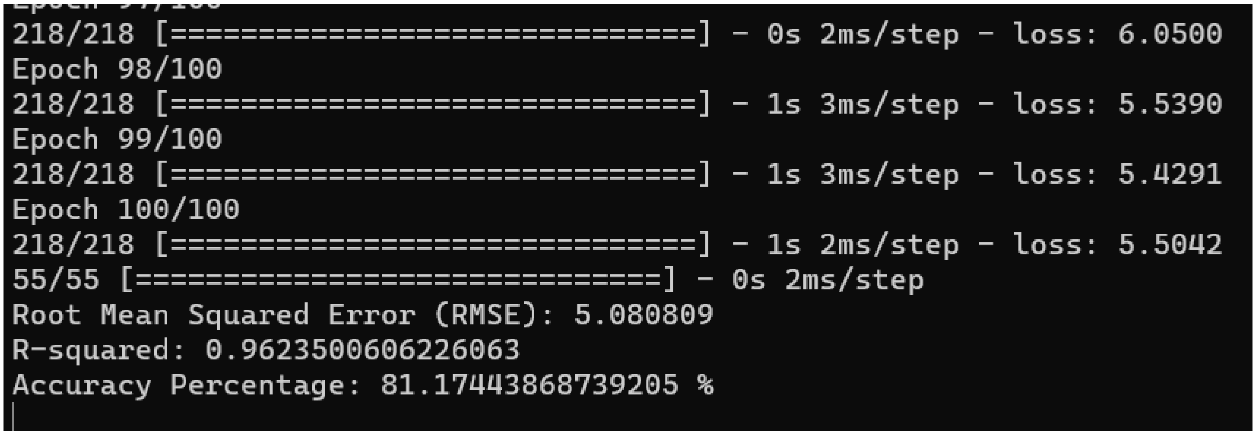
The predictive performance of our Deep Learning model exhibits variability across different contaminants. Figure 3.1 illustrates the diverse set of contaminants under consideration. As part of our analysis, users have the flexibility to select any contaminant from this comprehensive list for model evaluation. The accuracy of predictions is contingent upon the distinctive characteristics of each contaminant and the inherent intricacies within the dataset.

It's crucial to recognize that the accuracy percentages vary due to the distinct nature of the samples. For instance, contaminants like Fluoride and Calcium demonstrate considerable variation over short distances, presenting a challenging scenario for accurate predictions. This leads to a lower accuracy percentage for these specific contaminants. This observation underscores the importance of understanding the unique challenges posed by each contaminant when interpreting the model's predictive capabilities.



```
PS E:\Data-Mining-and-Predictive-Analysis---Lab-Project> python ganga_Deeplearning.py
Available Contaminants:
1. ALKALINITY,PHENOLPHTHALEIN(mg/L)
2. TOTAL ALKALINITY
3. BORON(mg/L)
4. CALCIUM(mg/L)
5. CHLORIDE
6. CARBONATE(mg/L)
7. ELECTRICAL CONDUCTIVITY(µS/CM) at 25°
8. FLUORIDE
9. IRON(mg/L)
10. TOTAL HARDNESS(mg/L)
11. BICARBONATE(mg/L)
12. POTASSIUM(mg/L)
13. MAGNESIUM(mg/L)
14. SODIUM
15. PERCENT SODIUM
16. AMMONIA-N(mg/L)
17. o_po4_p
18. PH.1
19. RESIDUAL SODIUM CARBONATE
20. SODIUM ABSORPTION RATIO
21. SECCHI DEPTH
22. SILICATE(sio2)(mg/L)
23. SULPHATE(mg/L)
24. TEMPERATURE
25. NITROGEN,NITRITE.1
Enter the number corresponding to the contaminant for prediction:4
CALCIUM(mg/L)
```

Figure 3.3.1. Providing input to our model, selecting which contaminant to predict.



```
Epoch 97/100
218/218 [==============================] - 0s 2ms/step - loss: 6.0500
Epoch 98/100
218/218 [==============================] - 1s 3ms/step - loss: 5.5390
Epoch 99/100
218/218 [==============================] - 1s 3ms/step - loss: 5.4291
Epoch 100/100
218/218 [==============================] - 1s 2ms/step - loss: 5.5042
55/55 [==============================] - 0s 2ms/step
Root Mean Squared Error (RMSE): 5.080809
R-squared: 0.9623500606226063
Accuracy Percentage: 81.17443868739205 %
```

Figure 3.3.2 Accuracy and Evaluation metrics for our deep learning model for the contaminant Calcium

The Root Mean Squared Error (RMSE) of 5.080809 implies that, on average, the model's predictions for Calcium contamination deviate by approximately 5.08 units from the actual values. While this signifies a degree of prediction error, the high R-squared value of 0.9623500606226063 highlights the model's robust ability to capture and explain 96.24% of the variability in Calcium levels. This indicates a strong correlation between the predicted values generated by the model and the actual observed values. The combination of a relatively low RMSE and a high R-squared value underscores the effectiveness and accuracy of the model in predicting Calcium contamination levels

The Accuracy Percentage of 81.17% provides valuable insights into the model's effectiveness in predicting Calcium contamination levels within a defined tolerance level. This metric indicates that, on average, the model's predictions align with the actual values in approximately 81.17% of instances, demonstrating a commendable level of accuracy. However, it's crucial to acknowledge the specific challenges associated with Calcium, particularly its known variations over short distances.

The observed accuracy percentage reflects the model's adeptness in navigating these challenges, showcasing a high degree of precision in predicting Calcium contamination. This performance metric, coupled with an understanding of Calcium's inherent complexities, underscores the model's overall proficiency and reliability in providing accurate predictions for this specific contaminant

Despite the overall strong, positive, linear relationship observed in the scatter plot (Figure 3.3.3), where points cluster tightly around the trend line, indicating a robust correlation, there is a noteworthy finding. The model tends to slightly overestimate water pollution levels, as indicated by the consistently higher predicted values compared to the actual levels. This discrepancy suggests a systematic bias in the model, potentially influenced by unaccounted variables or model complexity.

The slight spread in the data, where not all points align perfectly on the trend line, underscores the inherent variability in water pollution levels. This variability could arise from factors not considered in the model, emphasizing the complexity of real-world environmental dynamics. Despite these nuances, the clear trend and strong correlation affirm the model's utility in providing accurate predictions.

However, the overall trend is clear: the predicted water pollution level is a good predictor of the actual water pollution level.

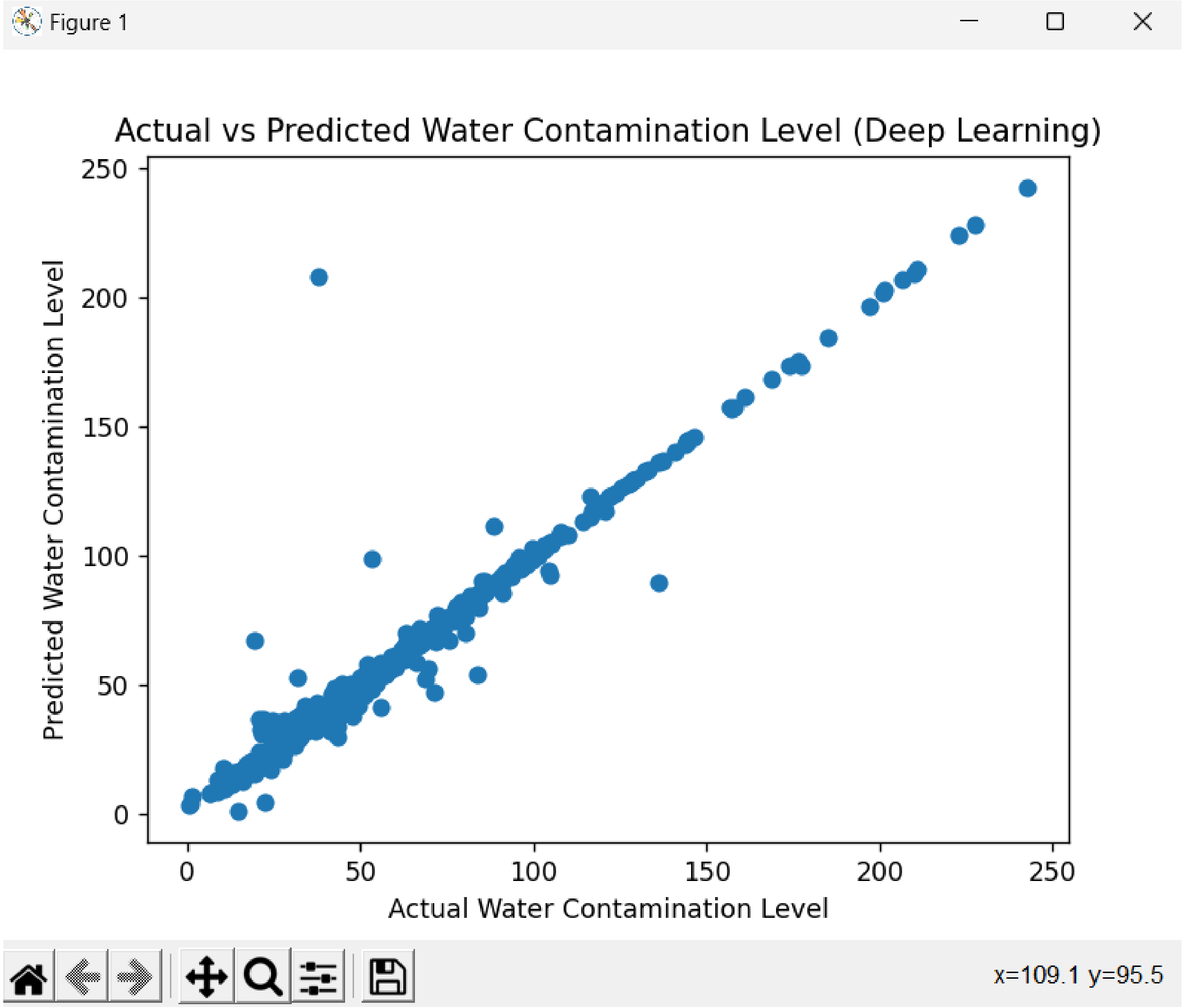
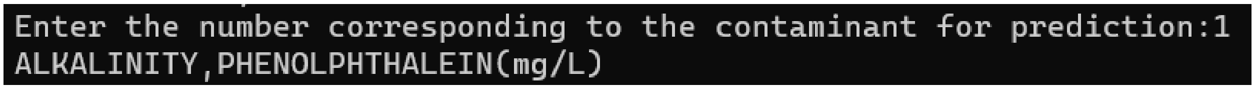


Figure 3.3.3. Scatter plot to validate and visualize the predicted and actual values for the same.

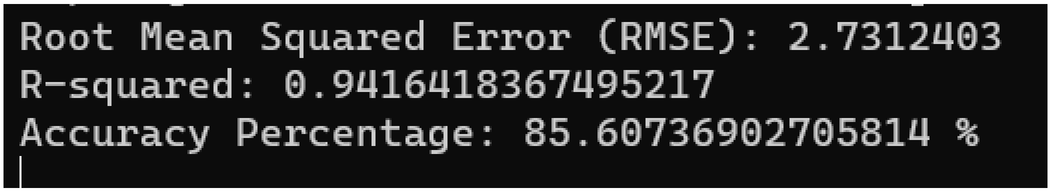
Applying the model individually to each contaminant provides valuable insights, akin to our previous analysis. The Root Mean Squared Error (RMSE), R-squared, and accuracy percentage for predicting each contaminant are depicted below:

ALKALINITY, PHENOLPHTHALEIN is a water quality parameter measured in milligrams per liter (mg/L). It indicates the alkalinity of water using phenolphthalein, providing insights into the water's buffering capacity against acidic substances.



```
Enter the number corresponding to the contaminant for prediction:1
ALKALINITY,PHENOLPHTHALEIN(mg/L)
```

Figure 3.3.4. Selection of contaminant 1



```
Root Mean Squared Error (RMSE): 2.7312403
R-squared: 0.9416418367495217
Accuracy Percentage: 85.60736902705814 %
```

Figure 3.3.5. Evaluation of model for alkalinity, phenolphthalein

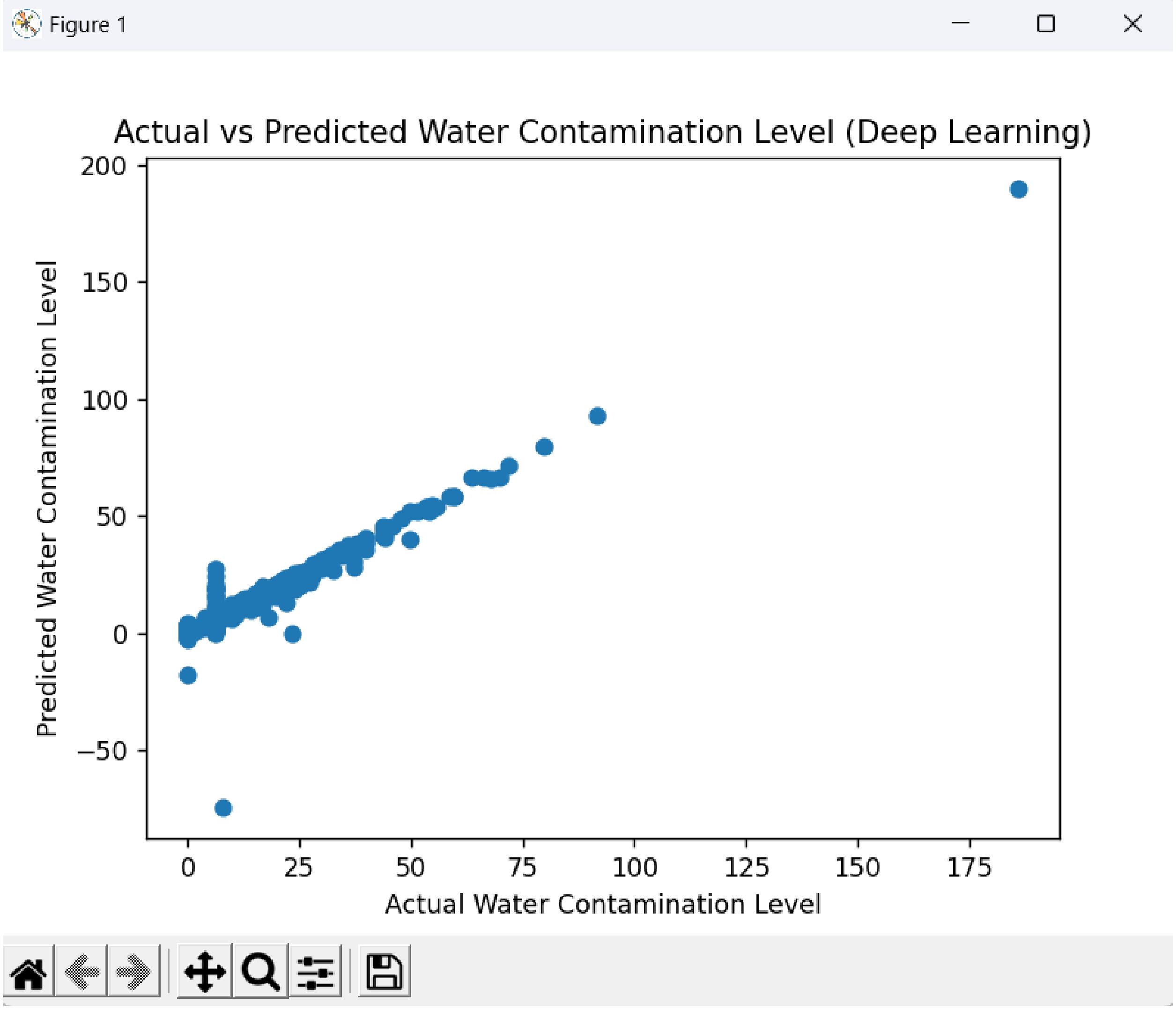
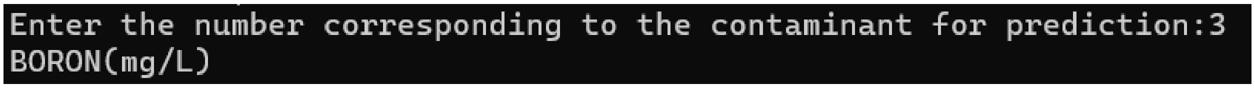


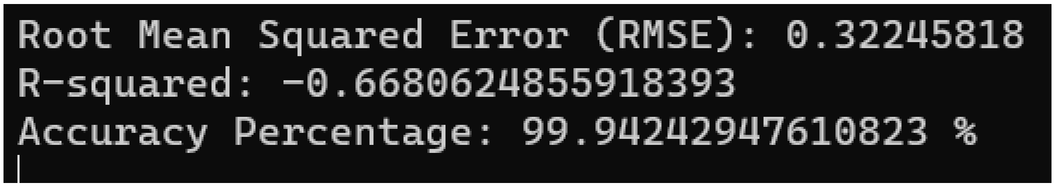
Figure 3.3.6. Scatter plot for alkalinity, phenolphthalein

BORON(mg/L)-measured in milligrams per liter (mg/L), is a water quality parameter that assesses the concentration of boron in water. Boron levels in water are significant as elevated concentrations may have ecological and health implications, making it crucial for monitoring and regulatory purposes.



```
Enter the number corresponding to the contaminant for prediction:3
BORON(mg/L)
```

Figure 3.3.7. Selection of contaminant



```
Root Mean Squared Error (RMSE): 0.32245818
R-squared: -0.6680624855918393
Accuracy Percentage: 99.94242947610823 %
```

Figure 3.3.8. Evaluation for Boron

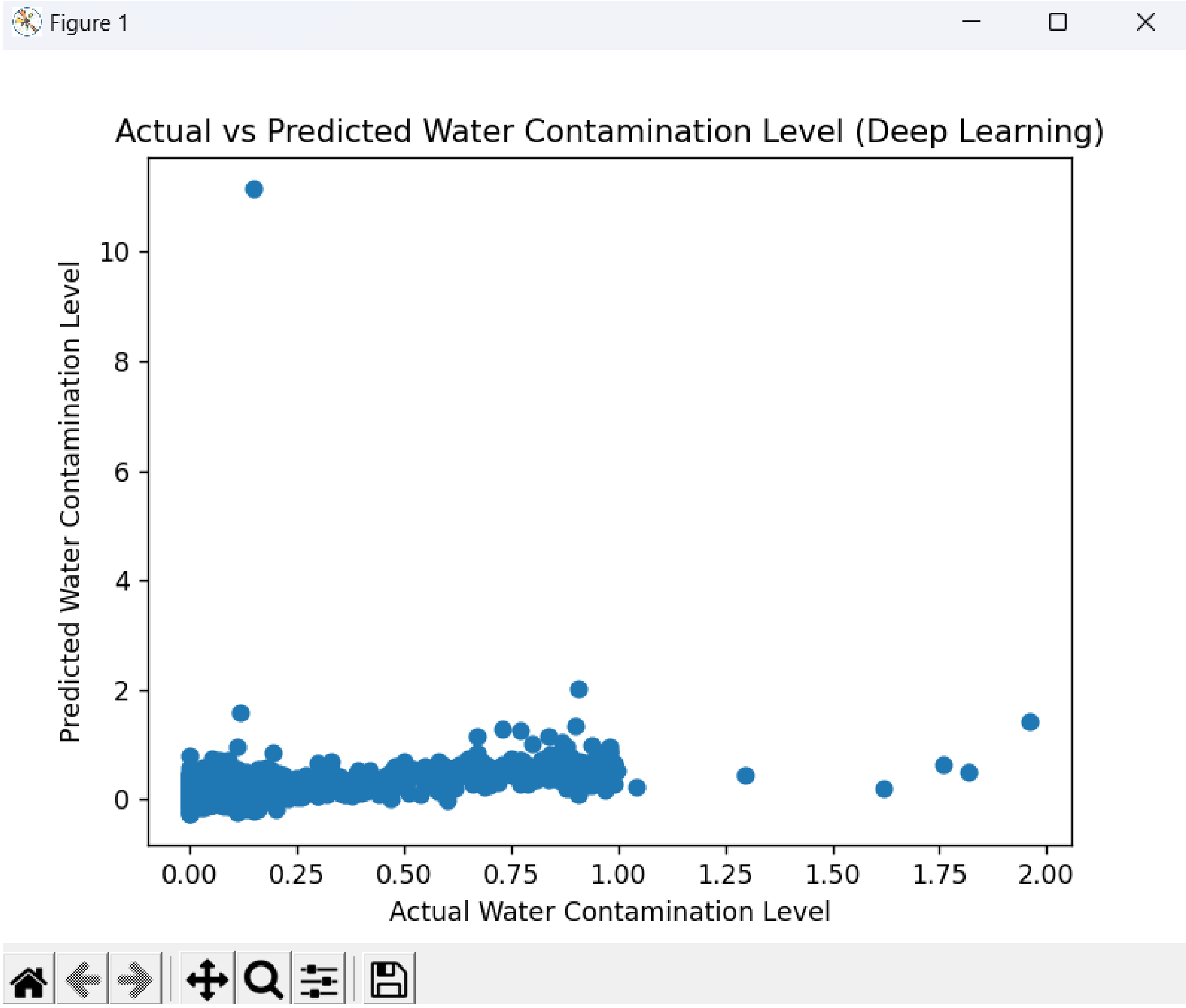
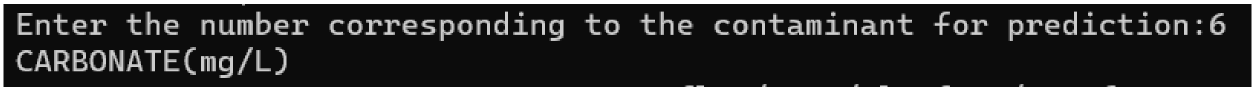


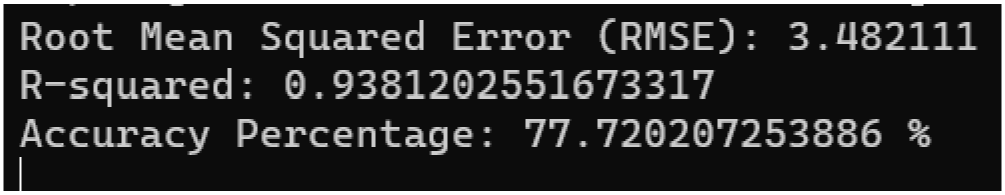
Figure 3.3.9. Scatter Plot for Boron

CARBONATE(mg/L) measured in milligrams per liter (mg/L), is a water quality parameter indicating the concentration of carbonate ions in water. Carbonate levels are crucial in assessing water alkalinity and can influence the overall chemistry and buffering capacity of water. Monitoring it is essential for understanding water quality, particularly in terms of its impact on pH and potential reactions with other constituents.



```
Enter the number corresponding to the contaminant for prediction:6
CARBONATE(mg/L)
```

Figure 3.3.10. Selection of contaminant



```
Root Mean Squared Error (RMSE): 3.482111
R-squared: 0.9381202551673317
Accuracy Percentage: 77.720207253886 %
```

Figure 3.3.11. Evaluation metrics for Carbonate

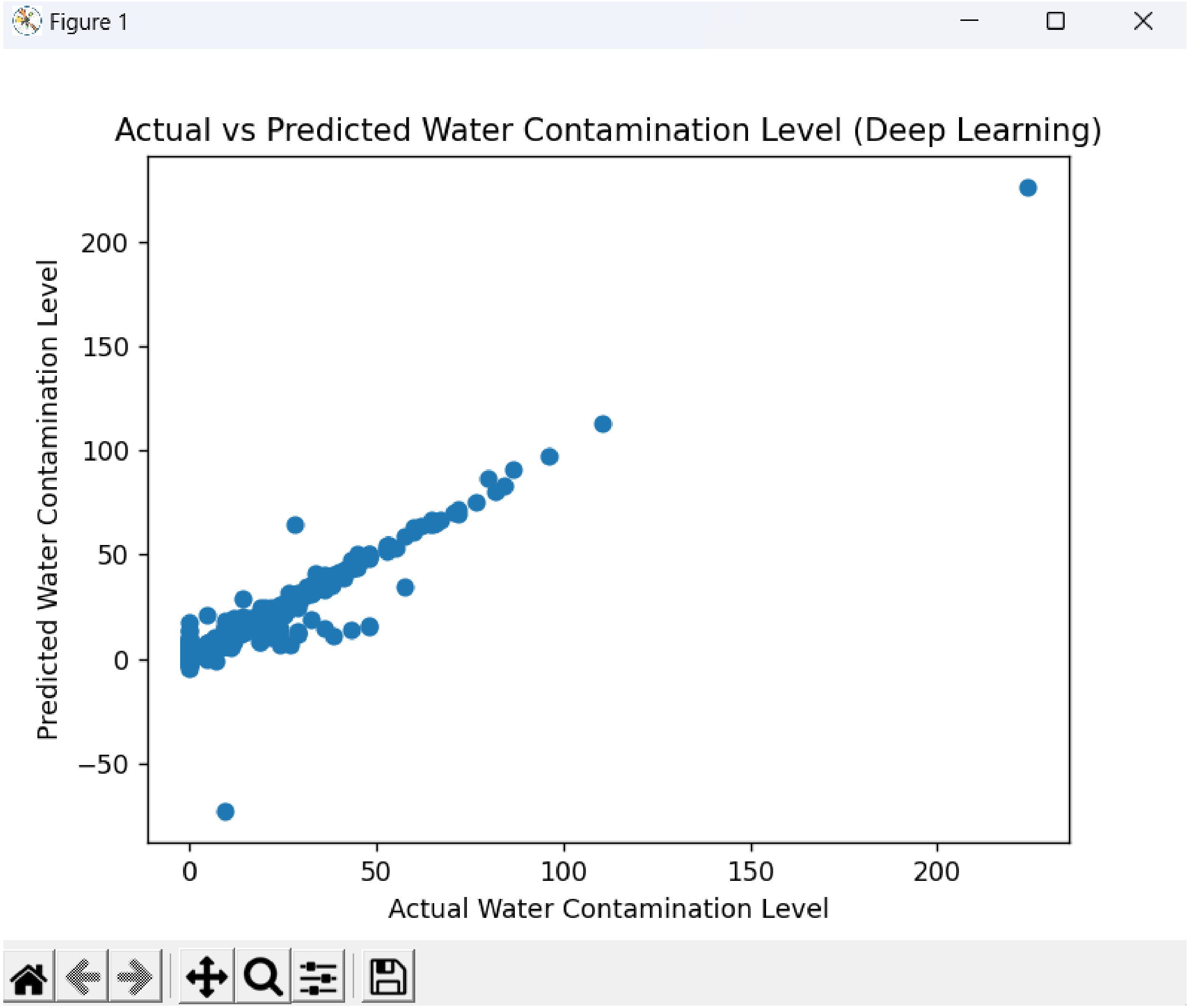
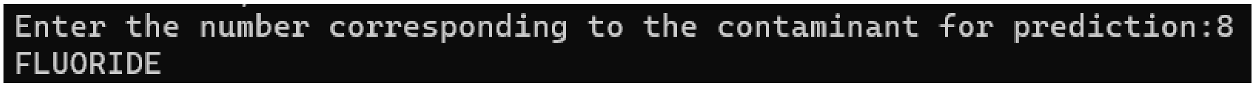


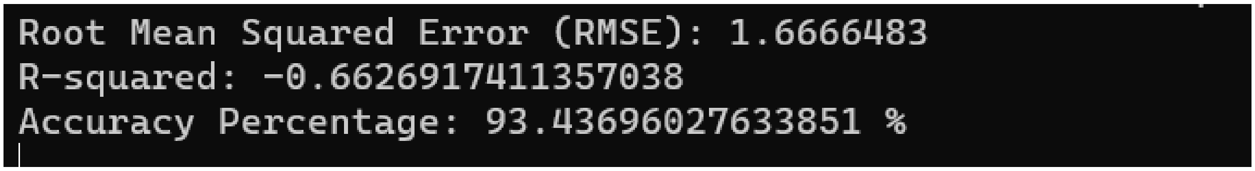
Figure 3.3.12. Scatter plot for Carbonate

FLUORIDE measured in milligrams per liter (mg/L) in water quality, is a crucial parameter for dental health and overall water safety. Excessive levels of Flouride can lead to dental and skeletal fluorosis. Monitoring its levels in water is essential to ensure it meets regulatory standards for both dental health protection and preventing adverse health effects associated with excessive fluoride intake.



```
Enter the number corresponding to the contaminant for prediction:8
FLUORIDE
```

Figure 3.3.13. Selection of contaminant



```
Root Mean Squared Error (RMSE): 1.6666483
R-squared: -0.6626917411357038
Accuracy Percentage: 93.43696027633851 %
```

Figure 3.3.14. Evaluation metrics for Flouride

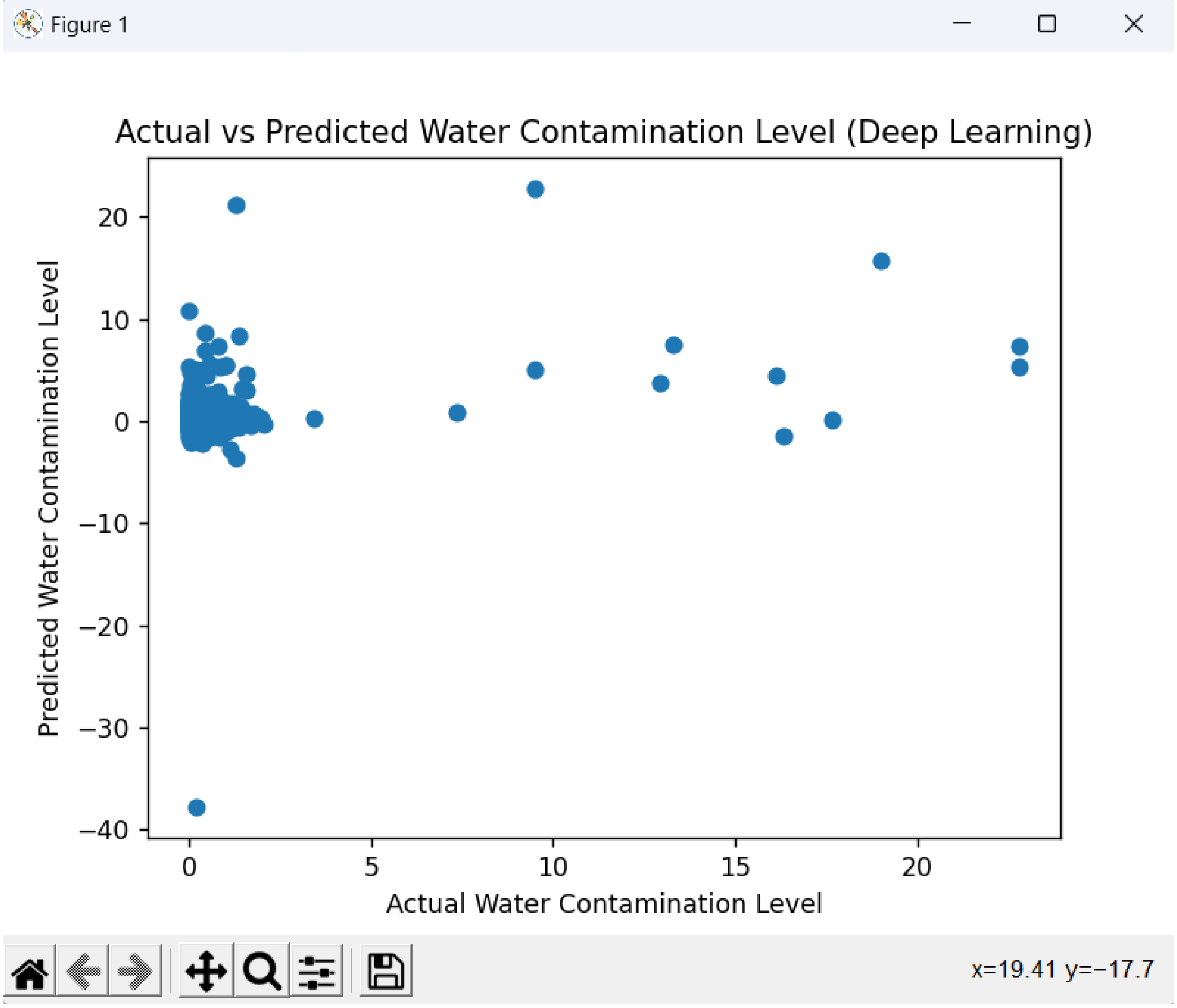
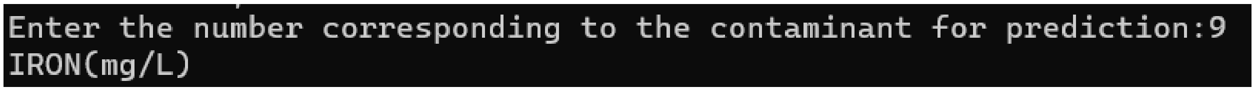


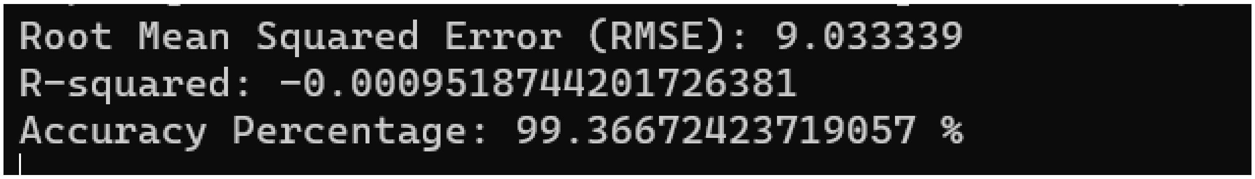
Figure 3.3.15. Scatter plot for Flouride

IRON(mg/L), measured in milligrams per liter (mg/L) in water quality assessments, indicates the presence of iron ions in water. Elevated levels of iron may lead to aesthetic issues like discoloration, taste, and odor concerns. Additionally, high iron concentrations can contribute to corrosion in plumbing systems.



```
Enter the number corresponding to the contaminant for prediction:9
IRON(mg/L)
```

Figure 3.3.16. Selection of contaminant



```
Root Mean Squared Error (RMSE): 9.033339
R-squared: -0.0009518744201726381
Accuracy Percentage: 99.36672423719057 %
```

Figure 3.3.17. Evaluation metrics for Iron

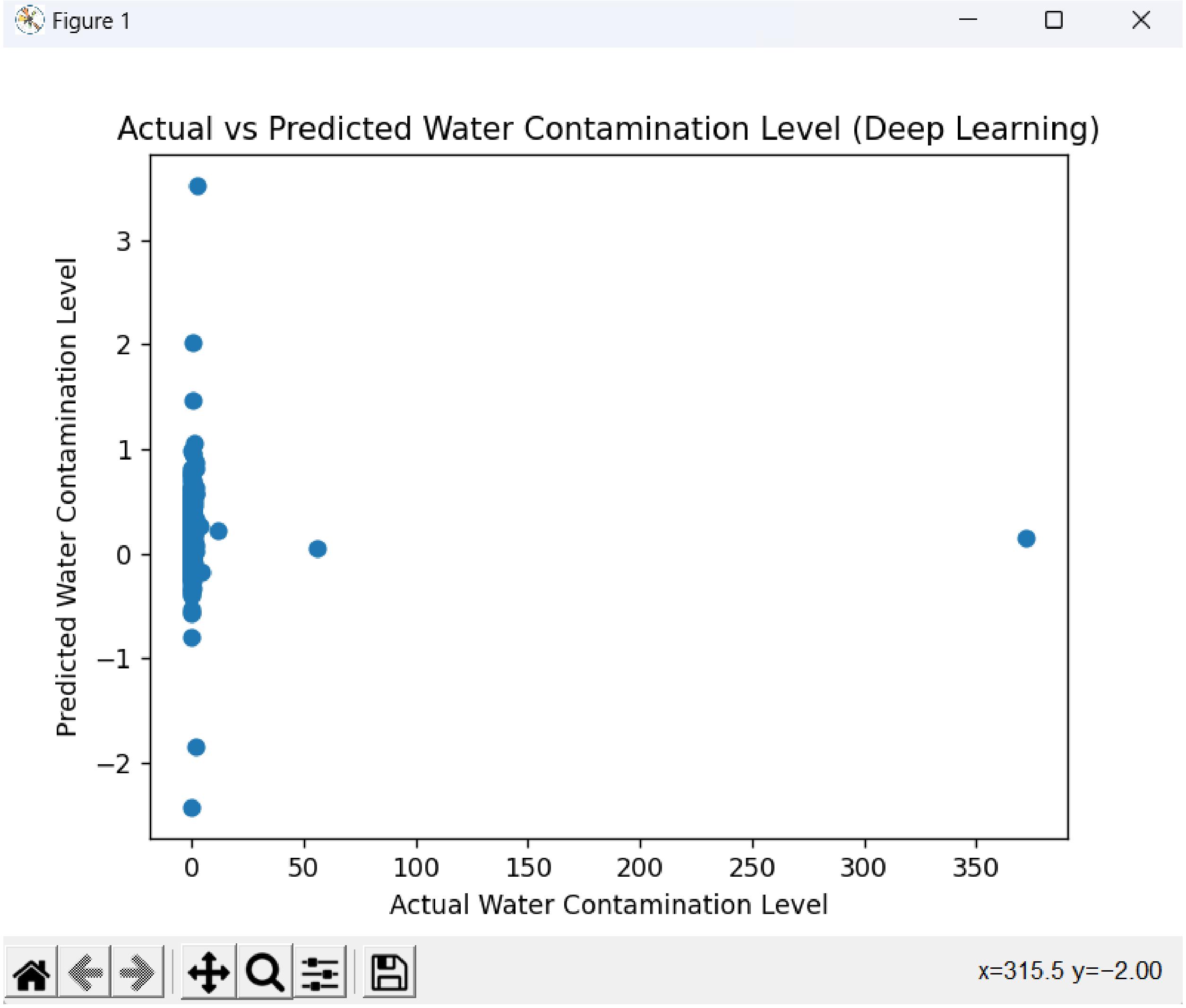
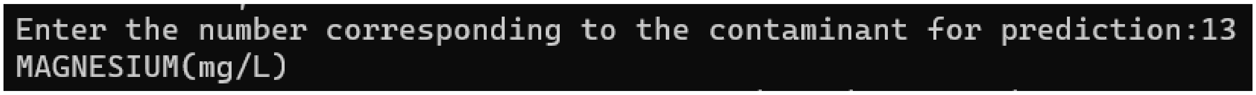


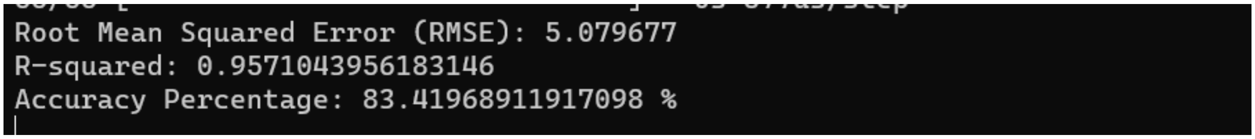
Figure 3.3.18. Scatter plot for Iron

MAGNESIUM(mg/L) concentration in water, measured in milligrams per liter (mg/L), is a key indicator of water hardness. It contributes to the overall mineral content of water and influences its taste and texture. Mg levels are crucial in evaluating the suitability of water for various purposes, such as drinking, agriculture, and industrial applications.



```
Enter the number corresponding to the contaminant for prediction:13
MAGNESIUM(mg/L)
```

Figure 3.3.3.19. Selection of contaminant



```
Root Mean Squared Error (RMSE): 5.079677
R-squared: 0.9571043956183146
Accuracy Percentage: 83.41968911917098 %
```

Figure 3.3.20. Evaluation metrics for Mg

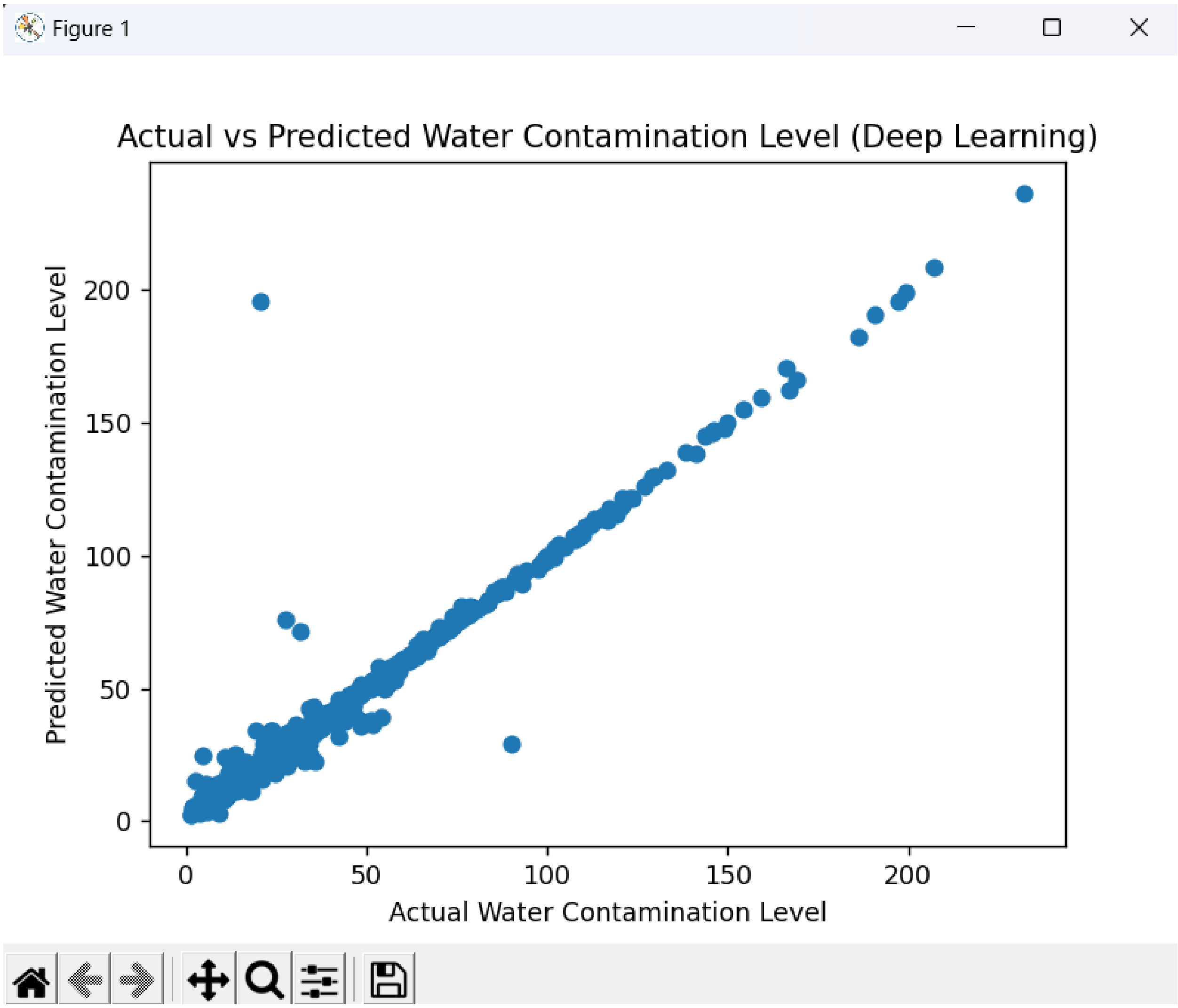
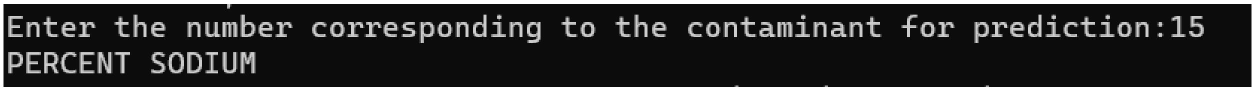


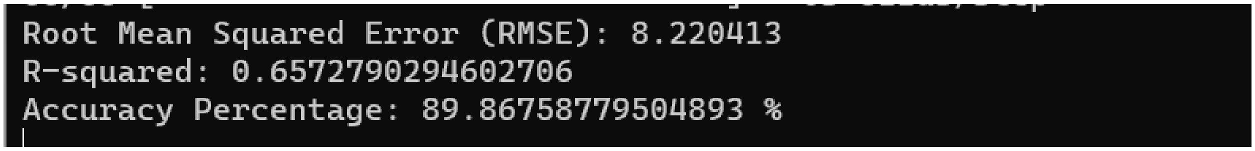
Figure 3.3.21. Scatter Plot for Mg

PERCENT SODIUM represents the proportion of sodium relative to the total cation content in water. It is a crucial measure in assessing the sodium hazard associated with irrigation water. Elevated levels of percent sodium can affect soil structure and plant growth, making it an essential consideration in agricultural practices.



```
Enter the number corresponding to the contaminant for prediction:15
PERCENT SODIUM
```

Figure 3.3.22. Selection of contaminant



```
Root Mean Squared Error (RMSE): 8.220413
R-squared: 0.6572790294602706
Accuracy Percentage: 89.86758779504893 %
```

Figure 3.3.23 Evaluation Metrics for Percent Sodium

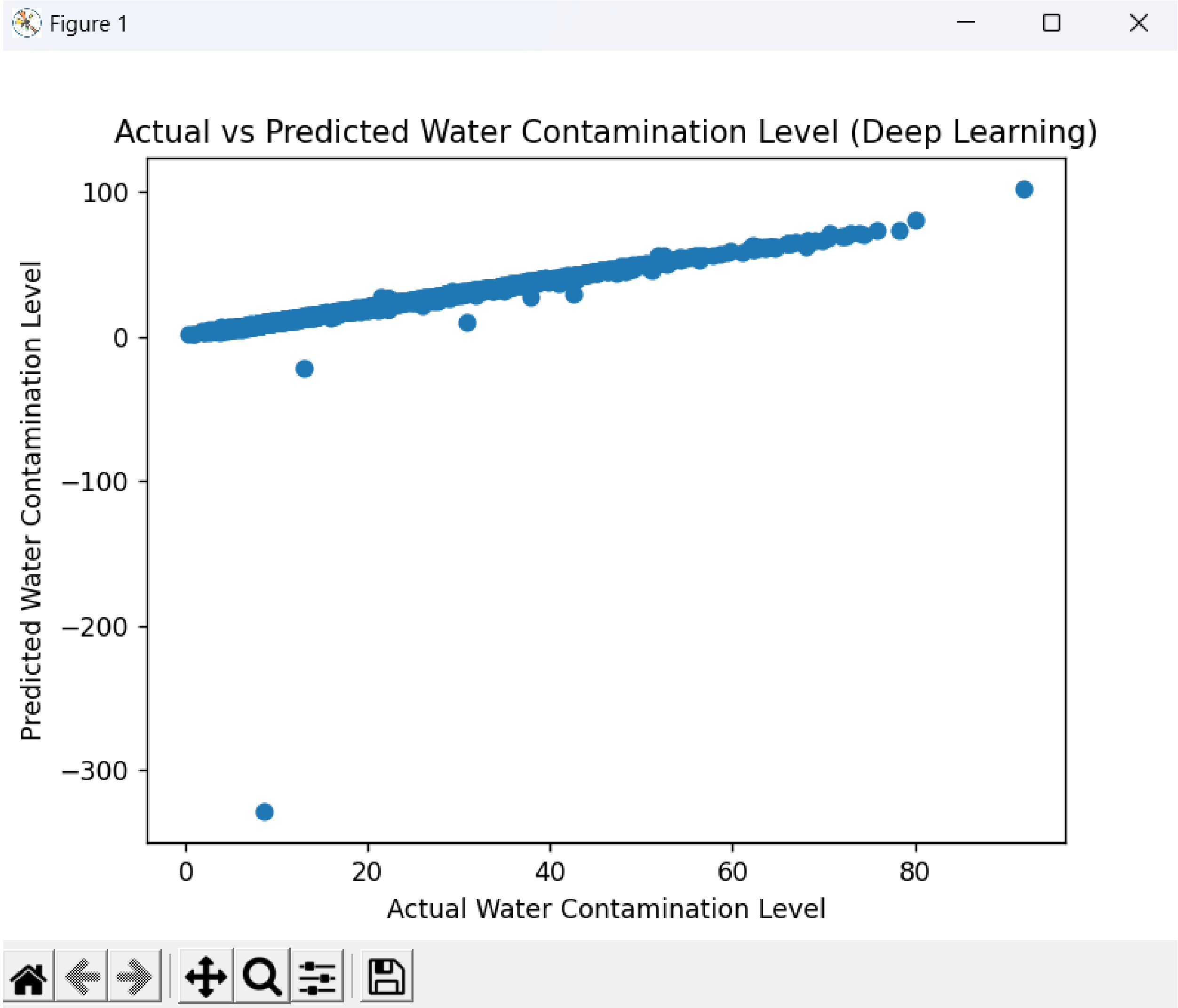
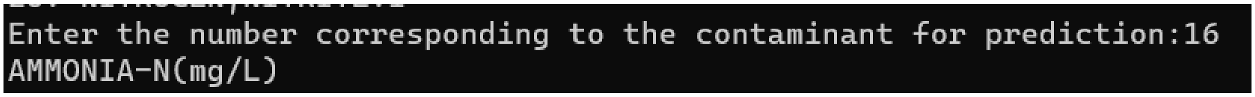


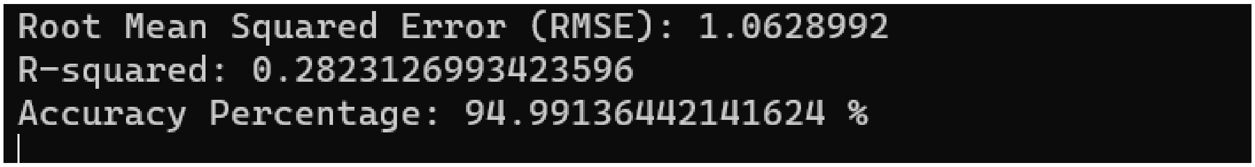
Figure 3.3.24 Scatter Plot for Percent Sodium

AMMONIA-N(mg/L) Monitoring this parameter is crucial for assessing water quality, as excessive ammonia levels can negatively impact aquatic ecosystems, leading to issues such as nutrient imbalances and harmful algal blooms. Controlling ammonia-N levels is essential for maintaining a healthy and balanced aquatic environment.



```
Enter the number corresponding to the contaminant for prediction:16
AMMONIA-N(mg/L)
```

Figure 3.3.25. Selection of contaminant



```
Root Mean Squared Error (RMSE): 1.0628992
R-squared: 0.2823126993423596
Accuracy Percentage: 94.99136442141624 %
```

Figure 3.3.26 Evaluation Metrics for NH4

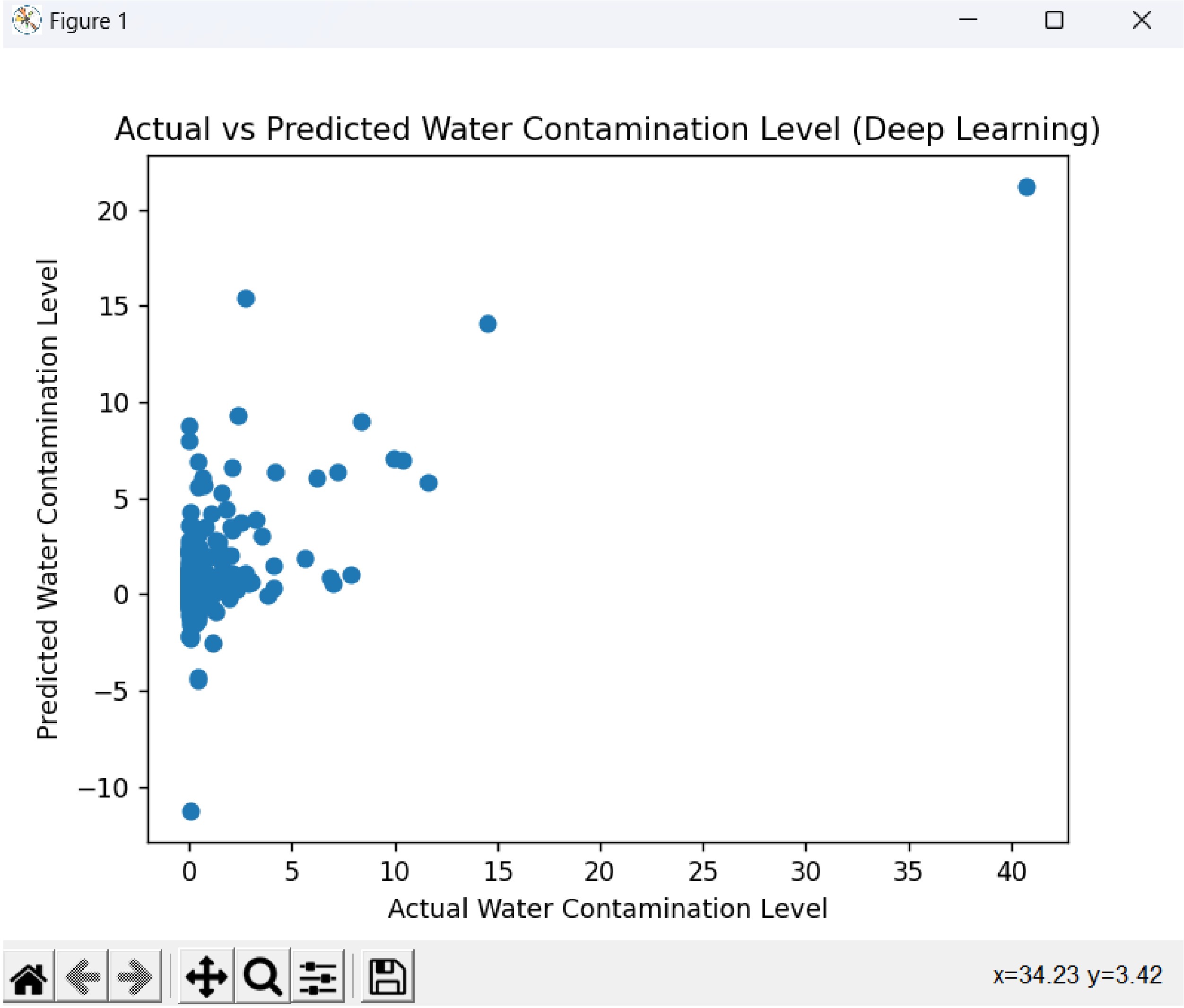
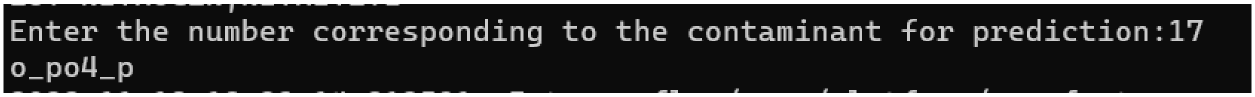


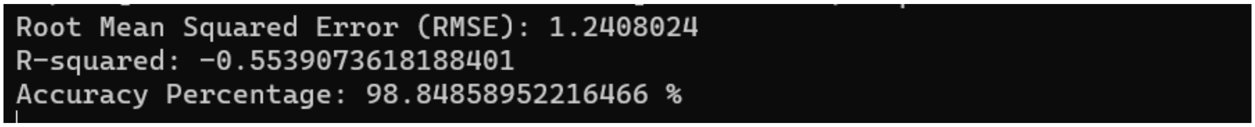
Figure 3.3.27 Scatter plot for NH4

O_PO4_P: This parameter refers to orthophosphate phosphorus, a form of phosphorus present in water. Orthophosphate is a key component in assessing nutrient levels and potential eutrophication in aquatic systems. Eutrophication occurs when excess nutrients, particularly phosphorus and nitrogen, lead to the overgrowth of algae and other aquatic plants, negatively impacting water quality.



```
Enter the number corresponding to the contaminant for prediction:17
o_po4_p
```

Figure 3.3.28. Selection of the Contaminant



```
Root Mean Squared Error (RMSE): 1.2408024
R-squared: -0.5539073618188401
Accuracy Percentage: 98.84858952216466 %
```

Figure 3.3.29. Evaluation metrics of O_po4_p

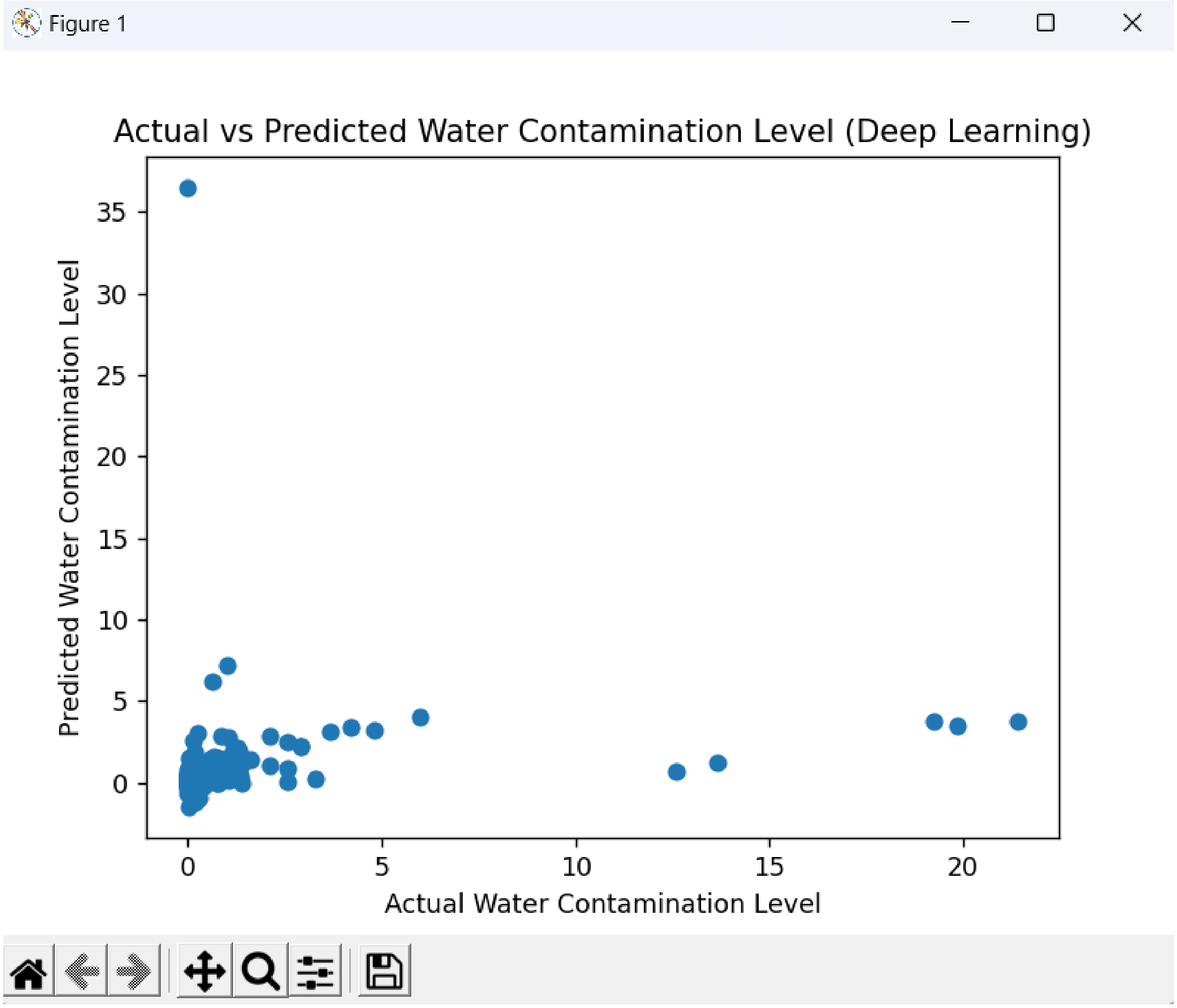
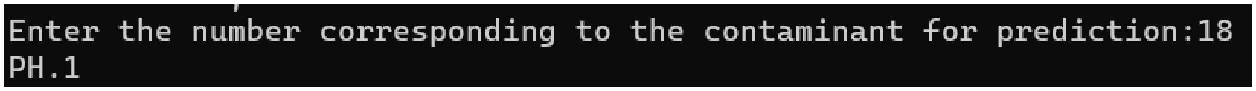


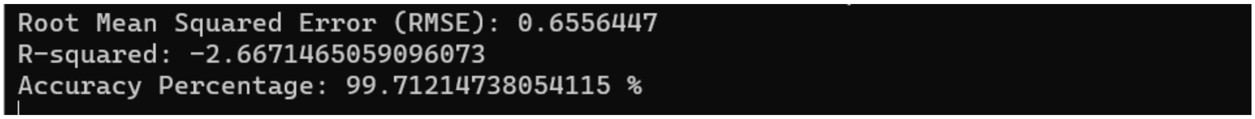
Figure 3.3.30. Scatter Plot of O_po4_p

PH.1 influences nutrient availability, chemical reactions, and the overall health of aquatic organisms. Maintaining appropriate pH levels is essential for supporting diverse aquatic ecosystems and ensuring water quality for various uses, including drinking water, industrial processes, and ecological preservation. Regular monitoring of pH provides insights into the stability and balance of aquatic environments.



```
Enter the number corresponding to the contaminant for prediction:18
PH.1
```

Figure 3.3.31. Selection of contaminant



```
Root Mean Squared Error (RMSE): 0.6556447
R-squared: -2.6671465059096073
Accuracy Percentage: 99.71214738054115 %
```

Figure 3.3.32 Evaluation Metrics for PH

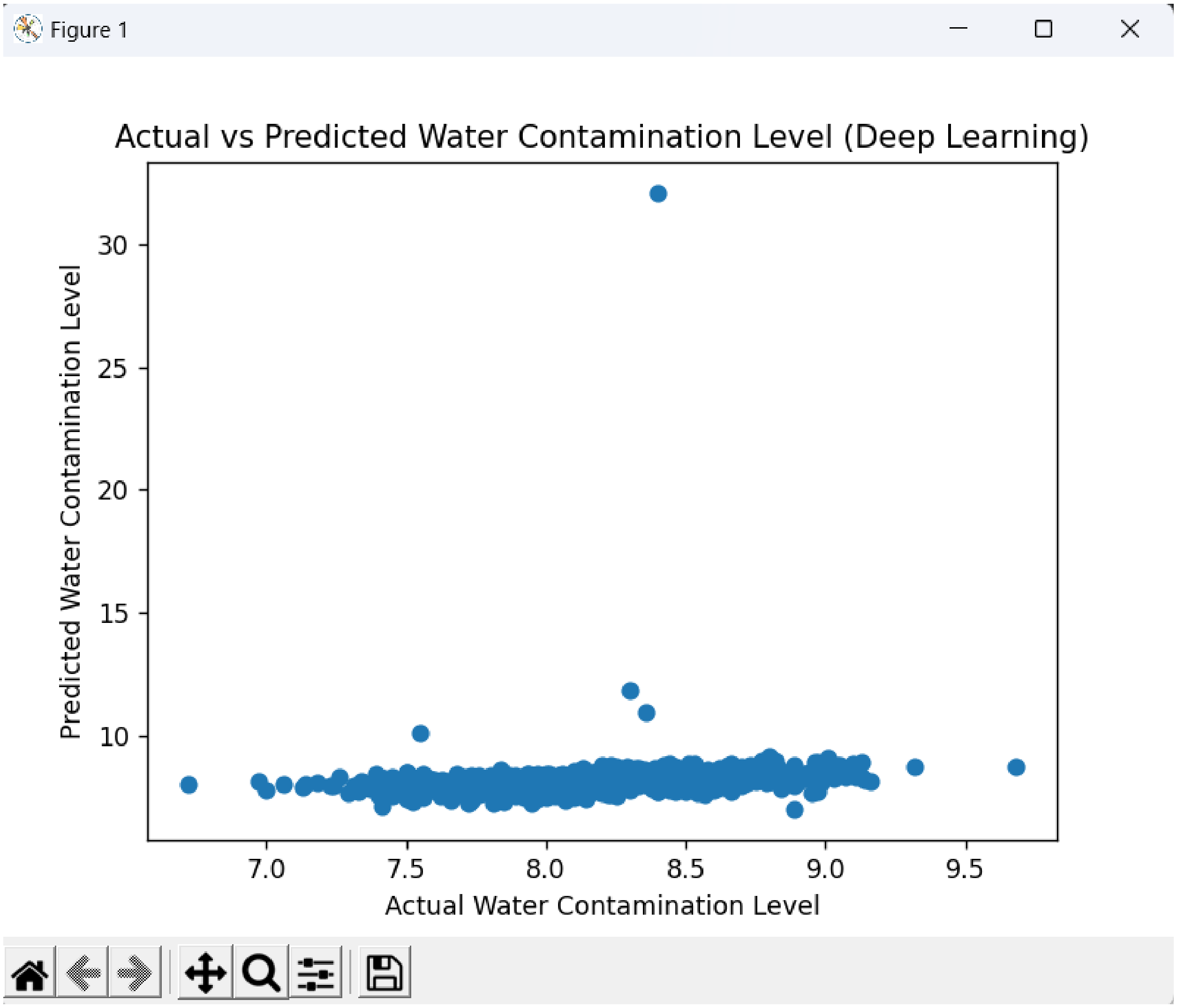
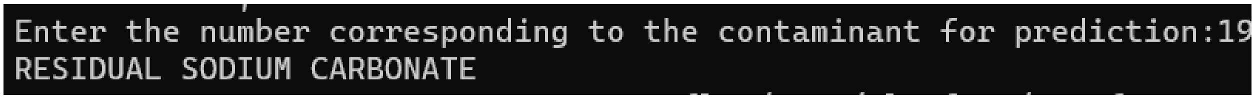


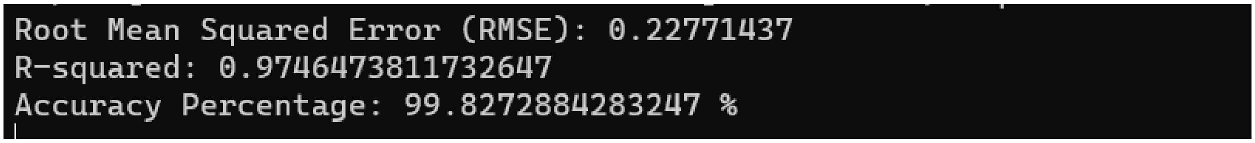
Figure 3.3.33. Scatter plot for PH

RESIDUAL SODIUM CARBONATE is a water quality parameter that measures the potential for water to cause soil sodicity. It indicates the concentration of sodium carbonate remaining after neutralizing the water with calcium and magnesium. Elevated RSC levels can lead to soil degradation, affecting agricultural productivity and necessitating proper water management practices.



```
Enter the number corresponding to the contaminant for prediction:19
RESIDUAL SODIUM CARBONATE
```

Figure 3.3.34. Select contaminant



```
Root Mean Squared Error (RMSE): 0.22771437
R-squared: 0.9746473811732647
Accuracy Percentage: 99.8272884283247 %
```

Figure 3.3.35 Evaluation Metrics for Residual Sodium Carbonate

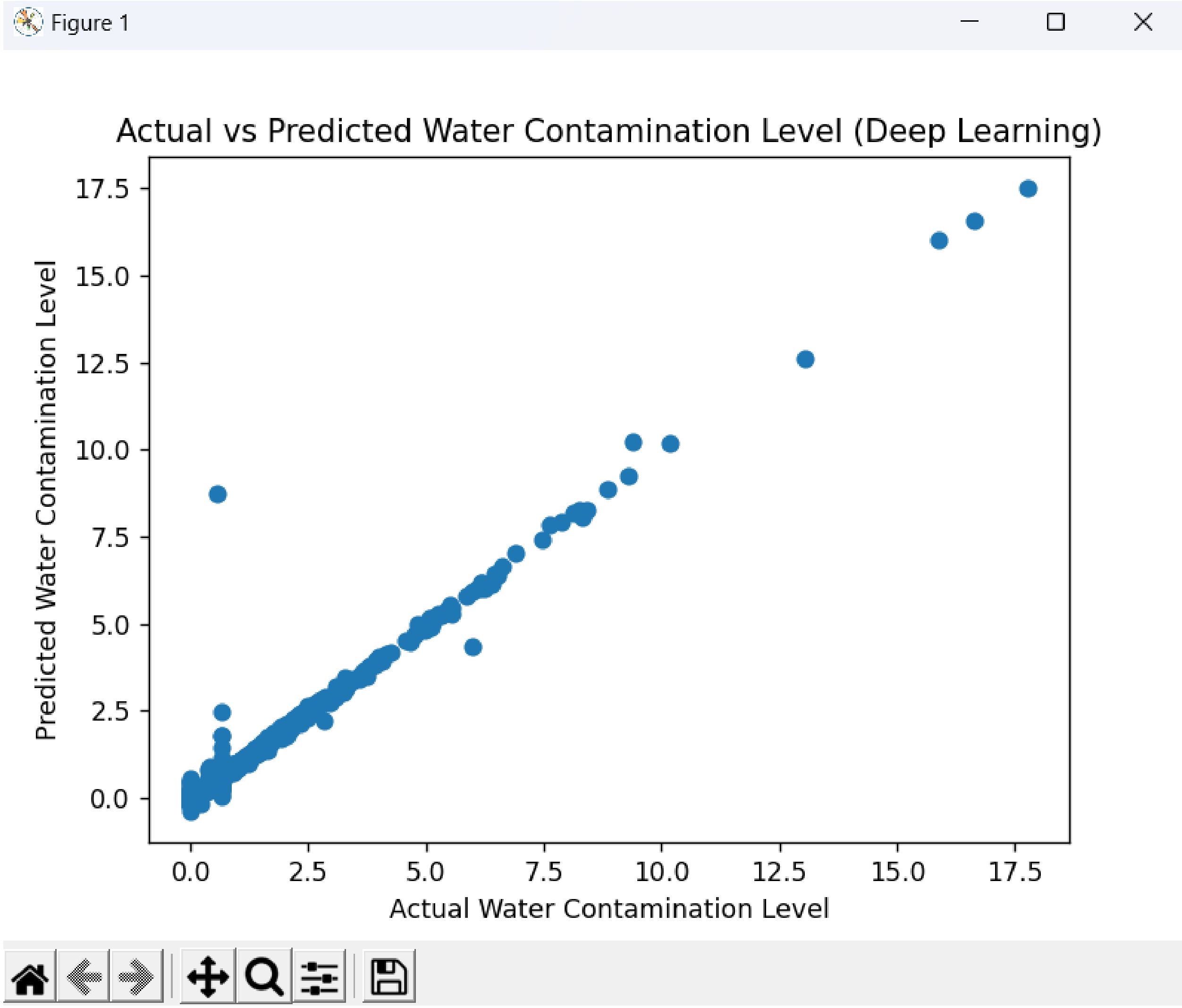
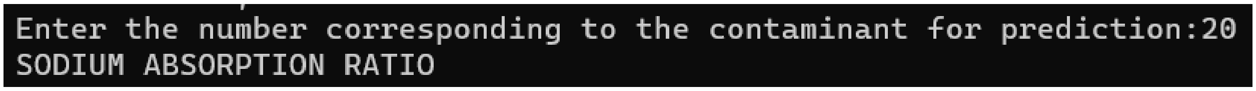


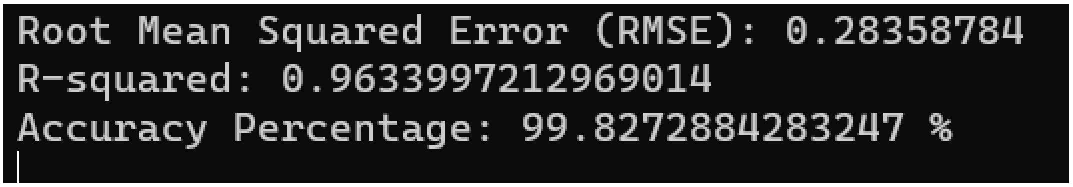
Figure 3.3.36. Scatter Plot for Residual Sodium Carbonate

SODIUM ABSORPTION RATIO (SAR) is a measure of the suitability of water for irrigation. It assesses the sodium hazard in water by considering the ratio of sodium to calcium and magnesium. High SAR values indicate an increased risk of soil dispersion and reduced water permeability, potentially leading to soil degradation.



```
Enter the number corresponding to the contaminant for prediction:20
SODIUM ABSORPTION RATIO
```

Figure 3.3.37 select contaminant



```
Root Mean Squared Error (RMSE): 0.28358784
R-squared: 0.9633997212969014
Accuracy Percentage: 99.8272884283247 %
```

Figure 3.3.38 evaluation metrics for SAR

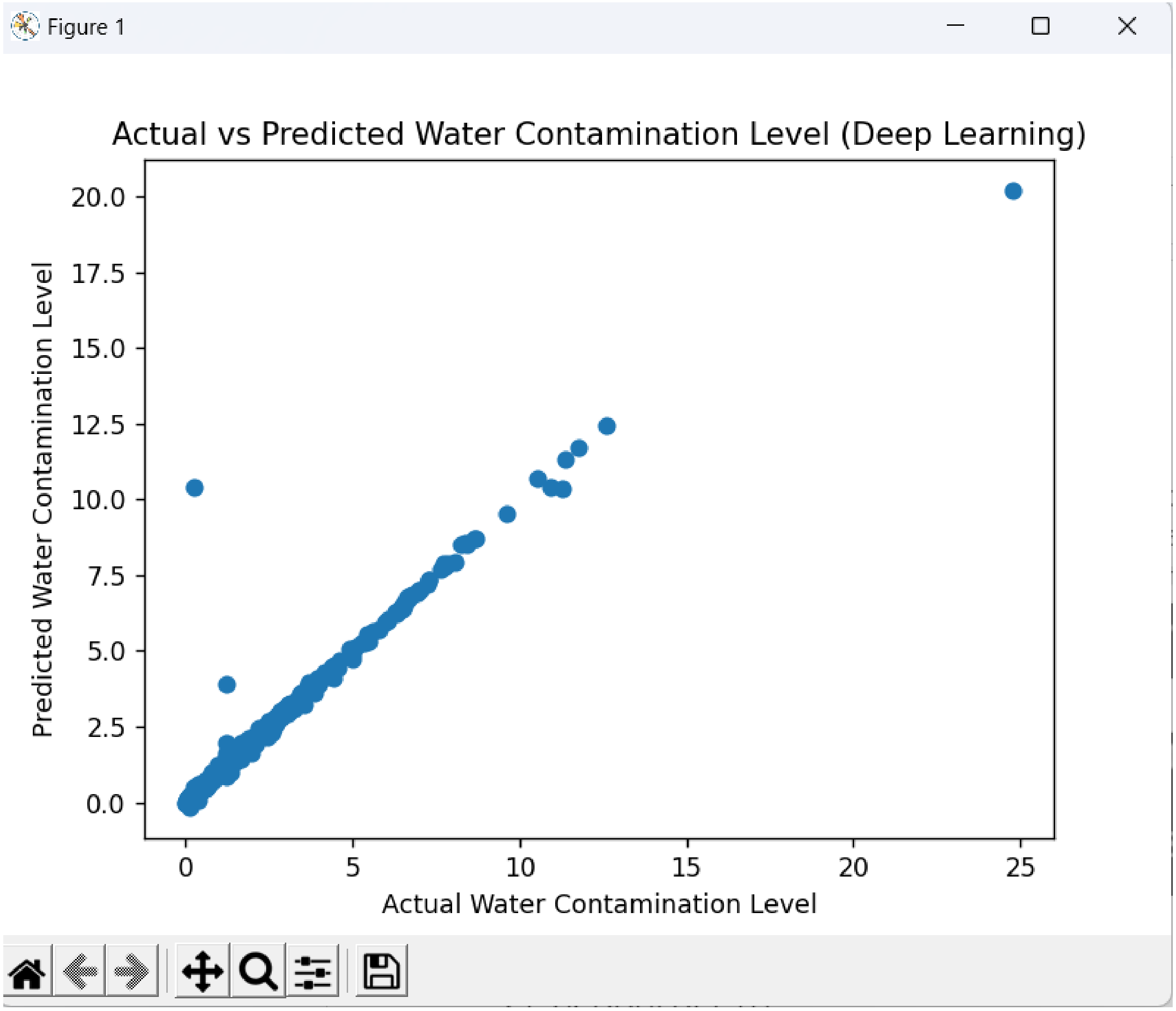
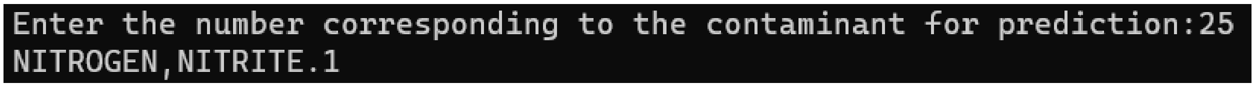


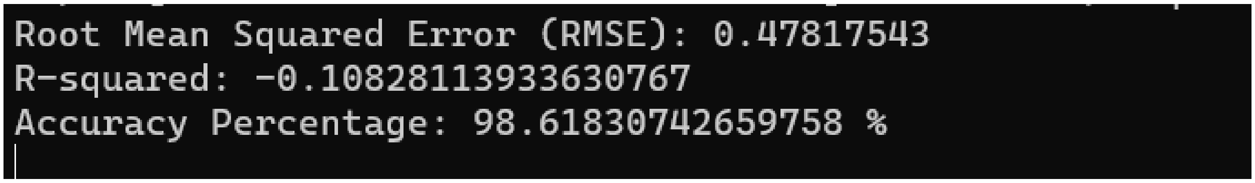
Figure 3.3.39. Scatter plot for SAR

NITROGEN,NITRITE.1 (NO2-) is a form of nitrogen in water and is a key parameter in assessing water quality. Elevated levels of nitrite can indicate contamination and pose a threat to aquatic ecosystems. Nitrite is a byproduct of the nitrogen cycle and excessive concentrations may result from agricultural runoff or wastewater discharge.



```
Enter the number corresponding to the contaminant for prediction:25
NITROGEN,NITRITE.1
```

Figure 3.3.40. Selection of contaminant



```
Root Mean Squared Error (RMSE): 0.47817543
R-squared: -0.10828113933630767
Accuracy Percentage: 98.61830742659758 %
```

Figure 3.3.41. Evaluation metrics for Nitrite

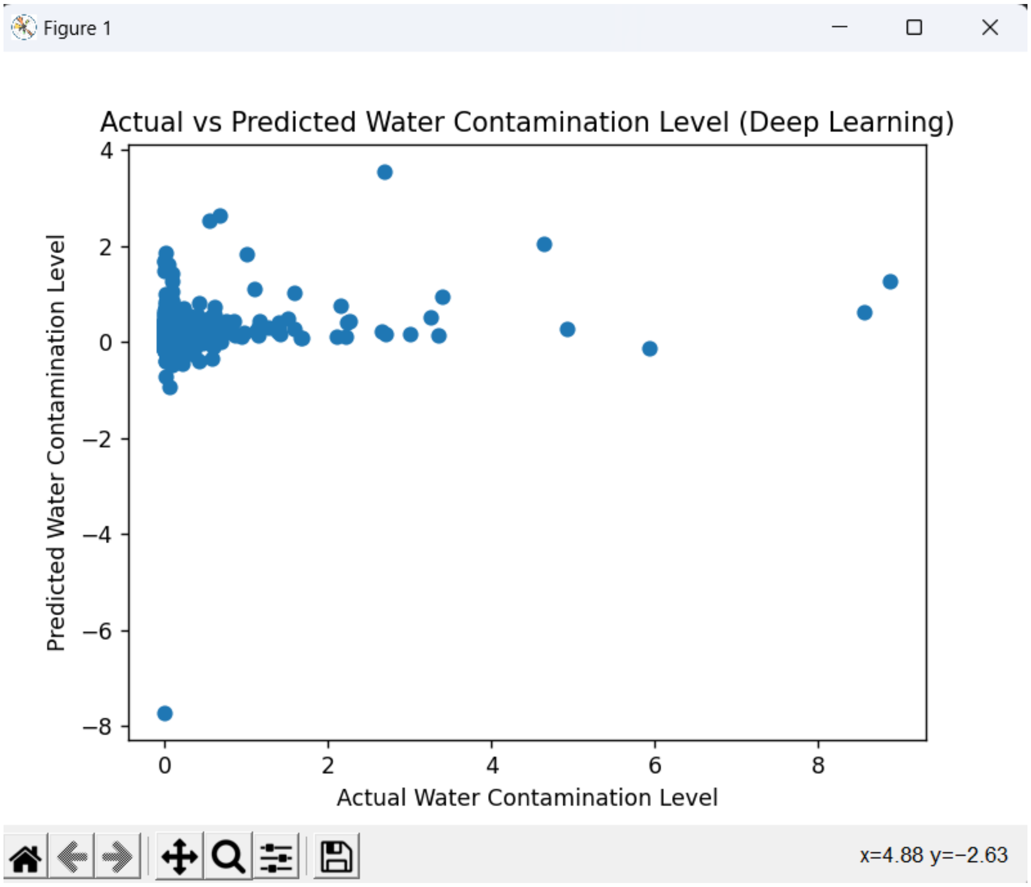


Figure 3.3.42. Scatter plot for Nitrite

Certain contaminants are excluded from the analysis due to their lower prediction accuracy, which may result from factors such as a high volume of missing values, skewed data distribution, or other contributing reasons. The focus is placed on contaminants with a prediction accuracy exceeding 75 percent, ensuring a more reliable and meaningful assessment of water quality in the presented analysis.

3.4. Polynomial Regression

The implementation of polynomial regression with a consistent degree of 3 across various contaminants resulted in impressive predictive accuracies and relatively low mean absolute errors (MAE). The model demonstrated robust performance across the parameters evaluated: Dissolved Oxygen (DO) with an MAE of 0.04 and accuracy of 99.47%, Biological Oxygen Demand (BOD) with an MAE of 0.03 and accuracy of 98.42%, Fecal Coliform (FC) with an MAE of 0.33 and perfect accuracy at 100.00%, Faecal Streptococci (FS) with an MAE of 0.49 and accuracy of 99.93%, and pH with an MAE of 0.01 and accuracy of 99.86%. The consistent use of a polynomial degree of 3 facilitated fair and uniform assessments across the contaminants, allowing for a standardized comparison of their predictive capabilities.

Employing the same polynomial degree for all contaminants not only ensured a level playing field for evaluation but also facilitated a standardized assessment of their predictive performance. The close approximation to the actual contaminant values, as evidenced by the low MAEs for DO, BOD, FC, FS, and pH, showcased the reliability of the polynomial regression model with degree 3. This uniform approach provides valuable insights into their predictive accuracies while maintaining a balanced evaluation framework for these parameters.

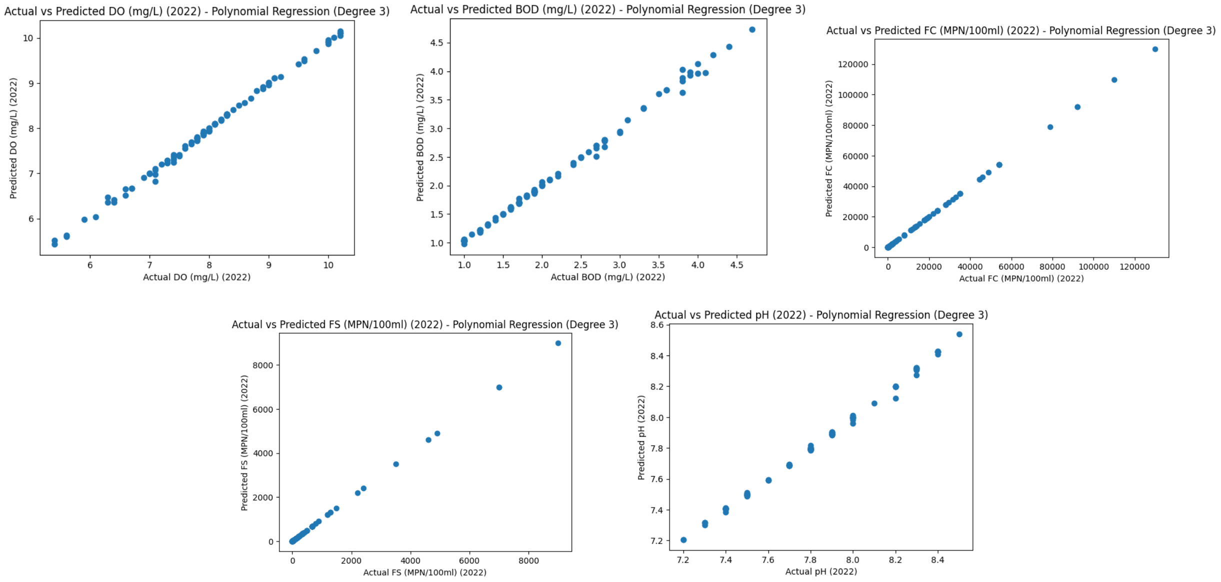


Figure 3.4.1 to 3.4.5 Performance of Polynomial Regression on each contaminant

Chapter 4. Conclusion

The evaluation of various predictive models, including Linear Regression, Random Forest, Deep Learning, and Polynomial Regression, provides a comprehensive understanding of Ganga River water contaminants' predictive accuracy. Linear regression models achieved exceptional accuracies for Dissolved Oxygen (DO), Biological Oxygen Demand (BOD), and pH, showcasing strong alignment between predicted and actual values, indicating their effectiveness in forecasting these parameters. However, the limitations surfaced for Fecal Coliform (FC) and Faecal Streptococci (FS), highlighting the complexities in predicting these contaminants solely based on selected attributes.

Conversely, the Random Forest Regressor demonstrated robust predictive capabilities across most contaminants, revealing precision in estimating DO, BOD, pH, and even for Faecal Streptococci (FS), while facing challenges in accurately estimating Fecal Coliform (FC). Deep Learning, while displaying variable accuracy across contaminants, showcased strong prediction for some parameters, such as PH and Nitrite, while struggling with Fluoride, Carbonate and Calcium due to their varying nature with short distances, impacting accuracy.

Polynomial Regression with a consistent degree of 3 delivered impressive predictive accuracies for DO, BOD, FC, FS, and pH. The uniform approach allowed for standardized comparison, showcasing close approximations to actual values, signifying the model's reliability.

Challenges in our study included data gaps, especially in recent years, and the intricate nature of certain contaminants affecting prediction accuracy. To improve, integrating more recent and comprehensive data and exploring advanced techniques for complex contaminants is crucial.

In conclusion, while each model had its strengths and limitations in predicting Ganga River contaminants, a combined approach leveraging the strengths of these models could yield a comprehensive understanding. Integrating these models' insights into real-time monitoring systems can facilitate timely interventions and policy decisions to ensure Ganga's environmental health. By harnessing the models' predictive abilities, policymakers and environmentalists can implement targeted strategies, including pollution mitigation measures and adaptive management practices, leading to the effective preservation and rejuvenation of the Ganga River ecosystem. This research serves as a foundation for future interdisciplinary studies, promoting a holistic approach towards sustainable water resource management and preserving the cultural and ecological significance of the Ganga River.

References

1. Koshy, Jacob. “Why is it difficult to clean up the Ganga?” *The Hindu*, 21 October 2017, https://www.thehindu.com/sci-tech/energy-and-environment/why-is-it-difficult-to-clean-up-the-ganga/article19896952.ece. Accessed 18 November 2023.
2. Vat, Marnix & Boderie, Pascal & Bons, Kees & Hegnauer, Mark & Hendriksen, Gerrit & van Oorschot, Mijke & Ottow, Bouke & Roelofsen, Frans & Sankhua, R & Sinha, S.K. & Warren, Andrew & Young, William. (2019). Participatory Modelling of Surface and Groundwater to Support Strategic Planning in the Ganga Basin in India. Water. 11. 2443. 10.3390/w11122443.
3. https://data.gov.in/resource/surface-water-quality-uttar-pradesh-1996-2006
4. Kadam, A.K., Wagh, V.M., Muley, A.A. *et al.* Prediction of water quality index using artificial neural network and multiple linear regression modelling approach in Shivganga River basin, India. *Model. Earth Syst. Environ.* 5, 951–962 (2019). https://doi.org/10.1007/s40808-019-00581-3
5. Kumar, A., Matta, G. & Bhatnagar, S. A coherent approach of Water Quality Indices and Multivariate Statistical Models to estimate the water quality and pollution source apportionment of River Ganga System in Himalayan region, Uttarakhand, India. *Environ Sci Pollut Res* 28, 42837–42852 (2021). https://doi.org/10.1007/s11356-021-13711-1
6. “Root-mean-square deviation.” Wikipedia, https://en.wikipedia.org/wiki/Root-mean-square_deviation. Accessed 18 November 2023.
7. “Mean absolute error.” *Wikipedia*, https://en.wikipedia.org/wiki/Mean_absolute_error. Accessed 18 November 2023.
8. “pandas documentation — pandas 2.1.3 documentation.” *Pandas*, https://pandas.pydata.org/docs/. Accessed 18 November 2023.

9. “User guide: contents — scikit-learn 1.3.2 documentation.” *Scikit-learn*, https://scikit-learn.org/stable/user_guide.html. Accessed 18 November 2023.
10. “NumPy user guide — NumPy v2.0.dev0 Manual.” *NumPy*, https://numpy.org/devdocs/user/index.html#user. Accessed 18 November 2023.
11. Butu, A. W., et al. "The impacts of poor solid waste management practices on Ala river water quality in Akure, Nigeria." Global Journal of Earth and Environmental Science 5.2 (2020): 37-50.
12. Vellingiri, J., et al. "Strategies for classifying water quality in the Cauvery River using a federated learning technique." International Journal of Cognitive Computing in Engineering 4 (2023): 187-193.
13. N. Radhakrishnan and A. S. Pillai, "Comparison of Water Quality Classification Models using Machine Learning," 2020 5th International Conference on Communication and Electronics Systems (ICCES), Coimbatore, India, 2020, pp. 1183-1188, doi: 10.1109/ICCES48766.2020.9137903
14. Ghadai, Madhusmita, et al. "Artificial neural network and weighted arithmetic indexing approach for surface water quality assessment of the Brahmani river." Glob. Nest J 24.4 (2022): 562-568.
15. Kumar, Siddharth, and Jayadeep Pati. "Assessment of groundwater arsenic contamination level in Jharkhand, India using machine learning." Journal of Computational Science 63 (2022): 101779.
16. Shukla, A. K., et al. "Water Quality Modelling and Parameter Assessment Using Machine Learning Algorithms: A Case Study of Ganga and Yamuna Rivers in Prayagraj, Uttar Pradesh, India." Environmental Processes and Management: Tools and Practices for Groundwater. Cham: Springer International Publishing, 2023.
17. Bhardwaj, A., Dagar, V., Khan, M.O. et al. Smart IoT and Machine Learning-based Framework for Water Quality Assessment and Device Component Monitoring. Environ Sci Pollut Res 29, 46018–46036 (2022). https://doi.org/10.1007/s11356-022-19014-3