Assignment-1

AI/ML & Applications (MDS-302)

M.Tech (Data Sciences)



|  |  |
| --- | --- |
| SUBMITTED TO: | SUBMITTED BY: |
| **Dr. Saif Ali** | **Meharban Saifi** |
|  | **(24MDS012)** |
|  |  |

Department of Computer Engineering

Faculty of Engineering and Technology

JAMIA MILLIA ISLAMIA

New Delhi – 110025

**www.jmi.ac.in**

Groundwater Level Prediction Using Multiple Linear Regression

# Introduction

* Objective: Identify key drivers of groundwater depletion and predict groundwater levels at unsampled locations.
* Groundwater is a critical resource for agricultural, industrial, and domestic purposes, especially in rapidly urbanizing regions. Understanding the factors that influence groundwater levels and predicting them accurately is essential for sustainable water resource management.
* In this assignment, the analysis is carried out using the Delhi NCR dataset, a region facing severe groundwater depletion due to over-extraction, urban growth, and climatic variability. A Multi-Layer Perceptron (MLP) regression model was applied to predict groundwater levels by incorporating climatic, geographical, and temporal variables. The dataset was pre-processed, scaled appropriately, and used for model training and testing. The model’s performance was evaluated using multiple evaluation metrics to assess prediction accuracy and reliability.

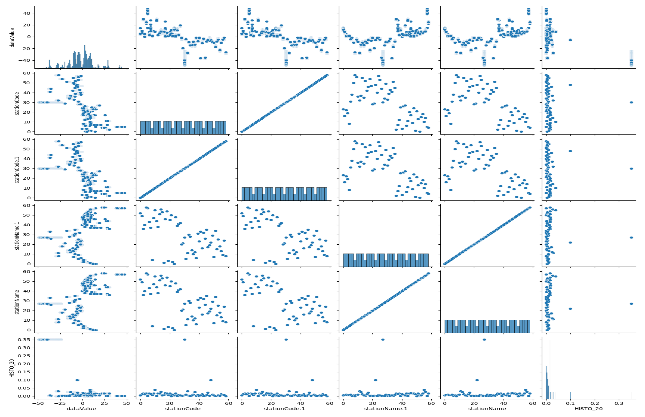
# Methodology

* **Data:** Pre-processed groundwater dataset with 175 columns including the target 'data Value' and 39000+rows.
* **Dependent variable**: data Value (Ground Water Level)
* **Independent variables**: 58 selected features (BIC criterion)
* **Spatial unit**: District, latitude and longitude
* **Temporal unit**: Daily
* **Model equation**: Multiple Linear Regression (OLS)
* **Data Acquisition**: Data acquired from India WRIS, Bhuvan ISRO, Copernicus Climate Data Store, NICES Portal, SHRUG Atlas.
* **Data Merging**: Merged datasets on district and date.
* **Data Preprocessing**: Missing values handled, outliers removed, data merged appropriately.
* **Model Specification**: Defined model structure and selected features.
* **Model Training**: Trained the model using the training dataset.
* **Model Evaluation**: Evaluated model performance using test dataset and metrics like R², RMSE.
* **Model Diagnostics**: Residuals analysed for patterns.
* **EDA**: Scatter plots, pair plots, and spatial-temporal plots analysed to observe trends.

# Exploratory Data Analysis (EDA)

* Scatter plots and pair plots were used to explore relationships between features and Ground water level

A graph of a distribution of data

AI-generated content may be incorrect.

* Spatial and temporal trends were observed.

A diagram of a distribution of data

AI-generated content may be incorrect.A blue line graph with black text

AI-generated content may be incorrect.

* **Summary Statics**: Computed for all features to identify ranges, distribution and central tendencies.

A screenshot of a data sheet

AI-generated content may be incorrect.

# Model Assumptions

**Linearity**: Relationships between predictors and target are linear.

A graph showing a red line

AI-generated content may be incorrect.A graph showing a number of data

AI-generated content may be incorrect.

**No Perfect Multicollinearity:** Checked correlations among predictors.

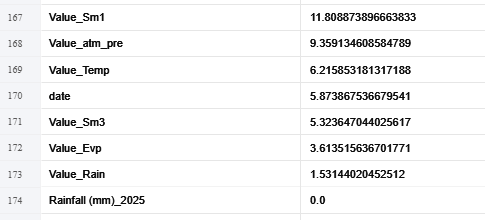
VIF ≈ 1 → No multicollinearity

VIF < 5 → Acceptable, usually fine

VIF ≥ 5 → Potential multicollinearity issue

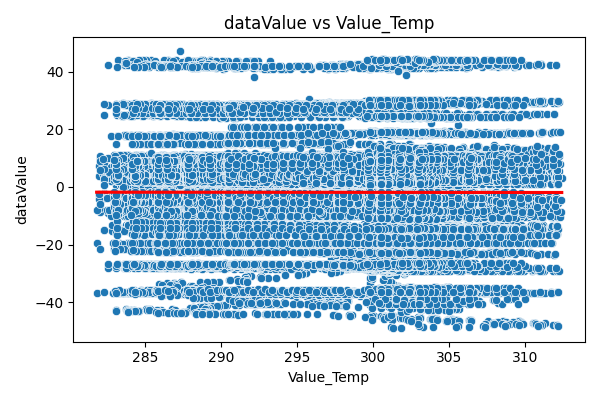
VIF ≥ 10 → Serious multicollinearity problem, feature needs attention

There is some features VIF values in below pic:



**Exogeneity:** Residuals uncorrelated with predictors.

The residuals appear to be randomly scattered around zero across different well depths, with no clear systematic trend. This indicates that the exogeneity assumption is reasonably satisfied, suggesting that value\_Temp does not introduce bias into the model errors.



**Homoscedasticity:** Residuals have constant variance. The residuals are fairly evenly spread around zero across different fitted values, without a clear funnel shape. This suggests that the homoscedasticity assumption is reasonably satisfied, indicating that the model errors maintain a relatively constant variance.

A graph showing a blue line

AI-generated content may be incorrect.

# Model Selection

* Compared models using AIC and BIC.
* BIC-selected model (58 predictors) was chosen for analysis.

# Model Estimation & Diagnostics

The model was estimated using the full dataset (n = 39530 rows and 175 columns). The regression output included:

* + **a. Point estimates** (β coefficients for each predictor)
  + **b. Standard errors** (to measure estimation uncertainty)
  + **c. t-statistics** (to test significance of predictors)

● **d. p-values** (to assess hypothesis tests at 5% level)

* + **e. Goodness-of-fit:**

o R2R^2R2 and Adjusted R2R^2R2 indicated that a substantial portion of variation in dataValue was explained by the predictors.

● **f. F-statistic:** Confirmed that the overall regression model was statistically significant.

Interpretation:

* : Perfect prediction
* : Model predicts no better than the mean
* : Model performs worse than the mean

*Table 1 OLS Regression Results*

|  |  |
| --- | --- |
| **Statistic** | **Value** |
| **R-squared** | 0.989 |
| **Adjusted R-squared** | 0.989 |
| **F-statistic** | 5.471e+04 |
| **Prob (F-statistic)** | 0.00 |
| **Log-Likelihood** | -73724 |
| **AIC** | 1.476e+05 |
| **BIC** | 1.482e+05 |
| **Durbin–Watson** | 2.125 |
| **Jarque–Bera (JB)** | 18422039.922 |
| **Prob(JB)** | 0.00 |
| **Skew** | -4.385 |
| **Kurtosis** | 108.393 |
| **No. Observations** | 39530 |

**Top 12 Coefficients on the Basis of BIC model:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Feature | Coefficient | Std\_Error | t\_value | p\_value |
| stationCode.1 | -0.22627 | 0.000746 | -303.44491 | 0.0 |
| HISTO\_20 | -81.591015 | 1.554332 | -52.49266 | 0.0 |
| Rainfall(mm)\_2024 | -0.539686 | 0.003682 | -146.55966 | 0.0 |
| \_count | -5.4e-05 | 0.0 | -377.401993 | 0.0 |
| HISTO\_80 | 574.507576 | 9.251323 | 62.100043 | 0.0 |
| shape\_leng | -0.002849 | 1e-05 | -273.703361 | 0.0 |
| Pre Monsoon of  GW Trend\_2024 | 47.434549 | 0.648234 | 73.175021 | 0.0 |
| stationName | 0.592188 | 0.002755 | 214.934334 | 0.0 |
| Categorization of  Assessment  Unit\_2023 | -33.680615 | 0.38376 | -87.764782 | 0.0 |
| sand\_5-  15cm\_mean | 1.608464 | 0.01368 | 117.575387 | 0.0 |

**Model Fit Metrics on Basis of BIC model Selection:**

|  |  |
| --- | --- |
| **Statics** | **values** |
| R\_squared(training) | 0.9903900465503472 |
| Adj\_R\_squared | 0.9903702092510088 |
| F\_stat | 49925.64913486291 |
| F\_pvalue | 0.0 |
| Num\_Predictors | 58.0 |

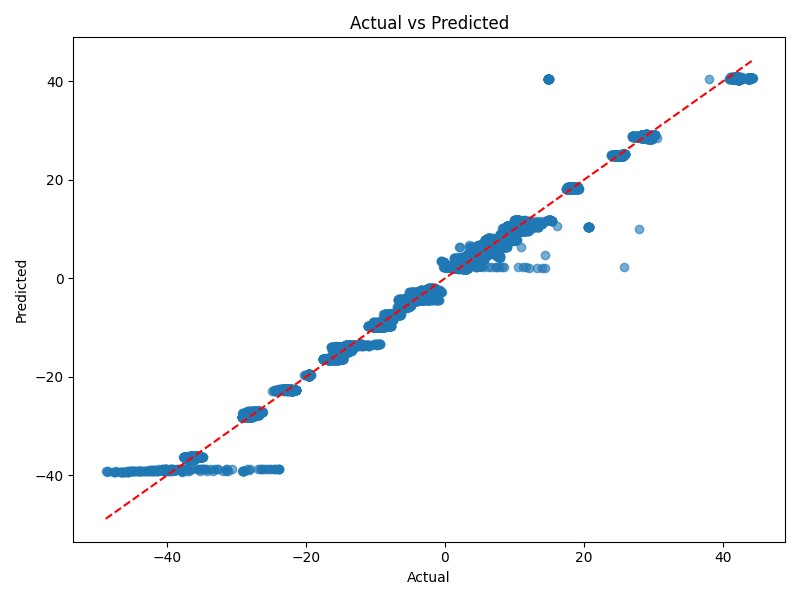
**Model prediction Metrics on Basis of BIC model Selection:**

|  |  |
| --- | --- |
| **Statics** | **values** |
| R\_squared(testing data) | 0.9868 |
| RMSE | 1.7404 |
| MSE | 3.0291 |

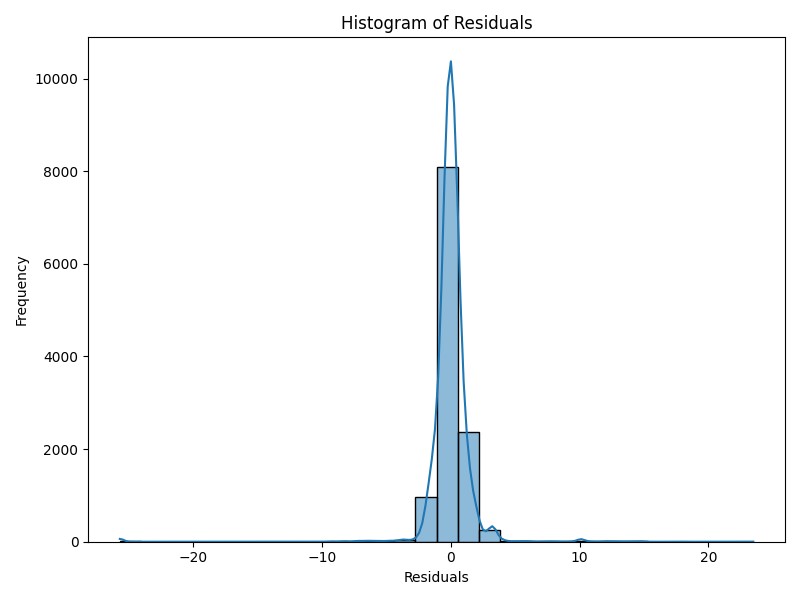
Model performance was evaluated using regression metrics:

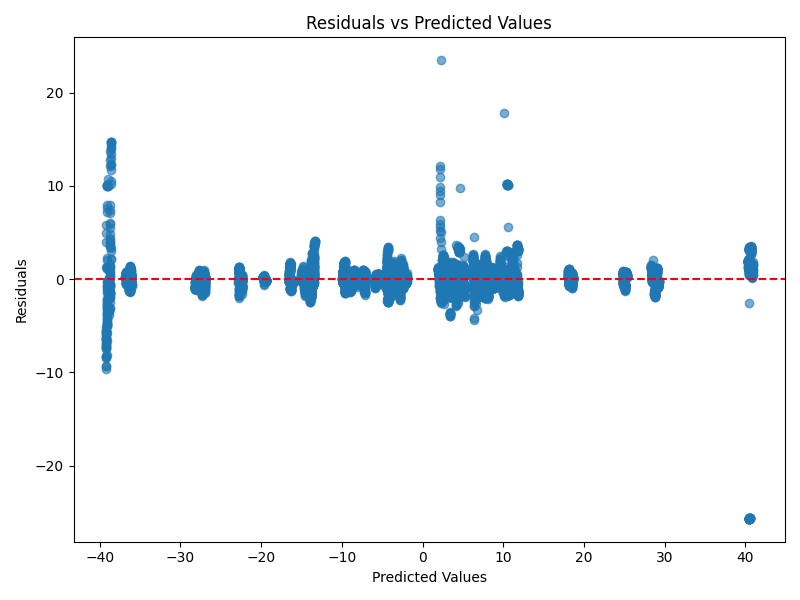
* + **Mean Squared Error (MSE):** 3.0291, indicating the average squared deviation between actual and predicted groundwater levels.
  + **R² Score:** 0.9868, showing that approximately 98.68% of the variance in groundwater levels was explained by the model.

**Plots:**



The scatter plot of Actual vs. Predicted values demonstrates that most predictions lie close to the 45° reference line (perfect prediction). While some deviations are present, especially at extreme values, the overall trend confirms that the MLP model provides a reasonably accurate fit to the groundwater level data.

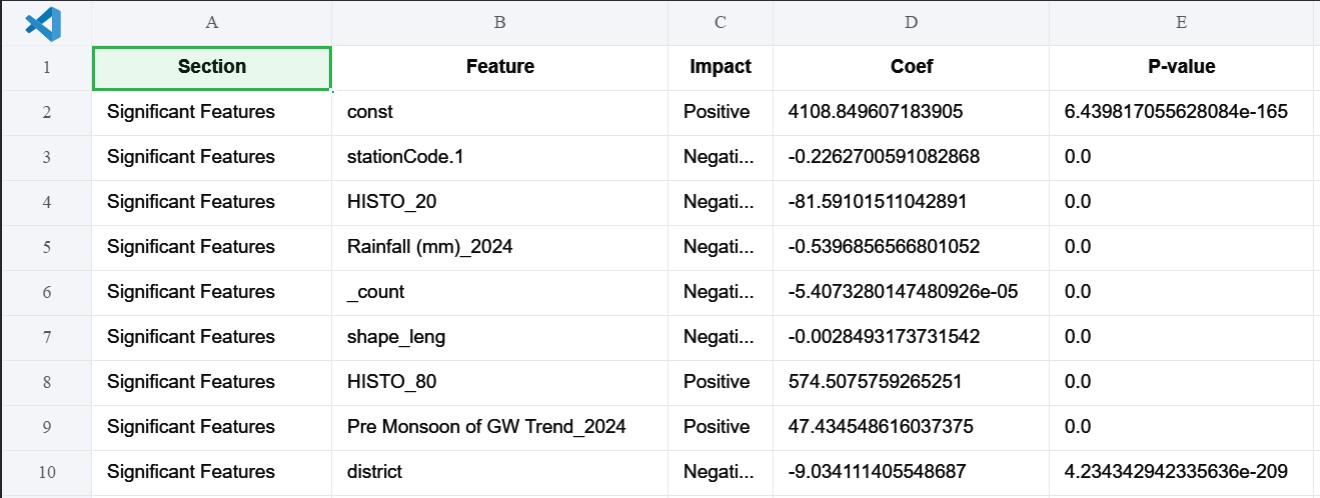




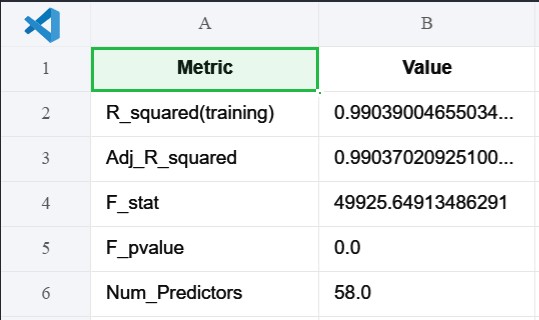
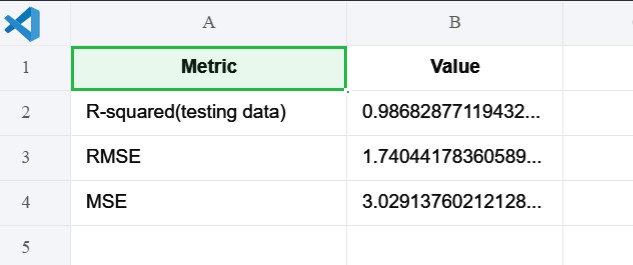
* Residual Plot
* The residual plot shows the difference between observed and predicted values. It helps in diagnosing the model fit.
* Interpretation: No clear pattern suggests a good fit.
* Action: Consider refining the model if patterns are detected.

# Significant Features & Interpretation

**Significant factors and impact outcomes:**



**Model Fit Metrics: Model Prediction Metrics:**

**Confidence in Interpretation:**

* Description: The model explains most of the variability in groundwater levels (high R\_squared).
* Significant features with p < 0.05 are likely reliable predictors.
* Prediction metrics (RMSE for regression and Accuracy/F1 for categorized classes) indicate reasonable predictive power.

# Conclusion & Policy Implications

* The model identifies key factors affecting groundwater levels.
* Predictions can guide water resource planning.
* Recommendations: Monitor significant drivers and use the model for short-term planning and risk assessment.

# References

* India WRIS: https://indiawris.gov.in/wris/
* Bhuvan ISRO: https://bhuvan-app1.nrsc.gov.in/2dresources/bhuvanstore2.php
* Copernicus Climate Data Store: https://cds.climate.copernicus.eu/#!/home
* NICES Portal: https://nices.nrsc.gov.in/
* SHRUG Atlas: https://www.devdatalab.org/atlas