Groundwater Level Prediction Using Multiple Linear Regression

1. Scenario

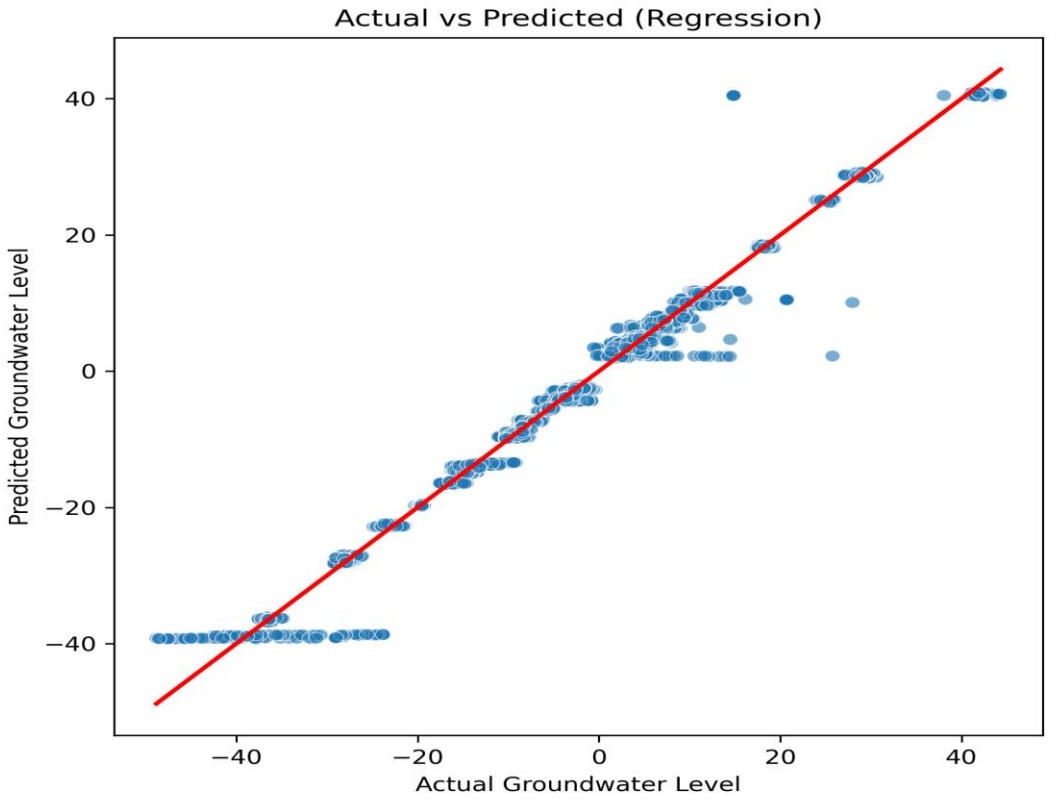
- Delhi NCR
- Objective: Identify key drivers of groundwater depletion and predict groundwater levels at unsampled locations.

2. Methodology

- **Data:** Preprocessed groundwater dataset with 175 columns including the target 'data Value'.
- **Dependent Variable**: data value
- **Independent Variables**: BIC criterion used (58 features).
- **Model**: Multiple Linear Regression (OLS).
- **Temporal unit**: Daily/Monthly/Yearly
- **Spatial unit**: District, Latitude, Longitude etc.
- **Data Acquisition**: Data acquired from India WRIS, Bhuvan ISRO, Copernicus Climate Data Store, NICES Portal, SHRUG Atlas.
- **Data Preprocessing**: Missing values handled, outliers removed, data merged appropriately.

3. Exploratory Data Analysis (EDA)

- Scatter plots and pair plots were used to explore relationships between features and groundwater levels.
- Spatial and temporal trends were observed.
- Data was cleaned and outliers removed.



4. Model Assumptions

- Linearity: Relationships between predictors and target are linear.
- No Perfect Multicollinearity: Checked correlations among predictors.
- Exogeneity: Residuals uncorrelated with predictors.
- Homoscedasticity: Residuals have constant variance.

5. Model Selection

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

6. Model Estimation & Diagnostics

- Top 10 Coefficients:

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

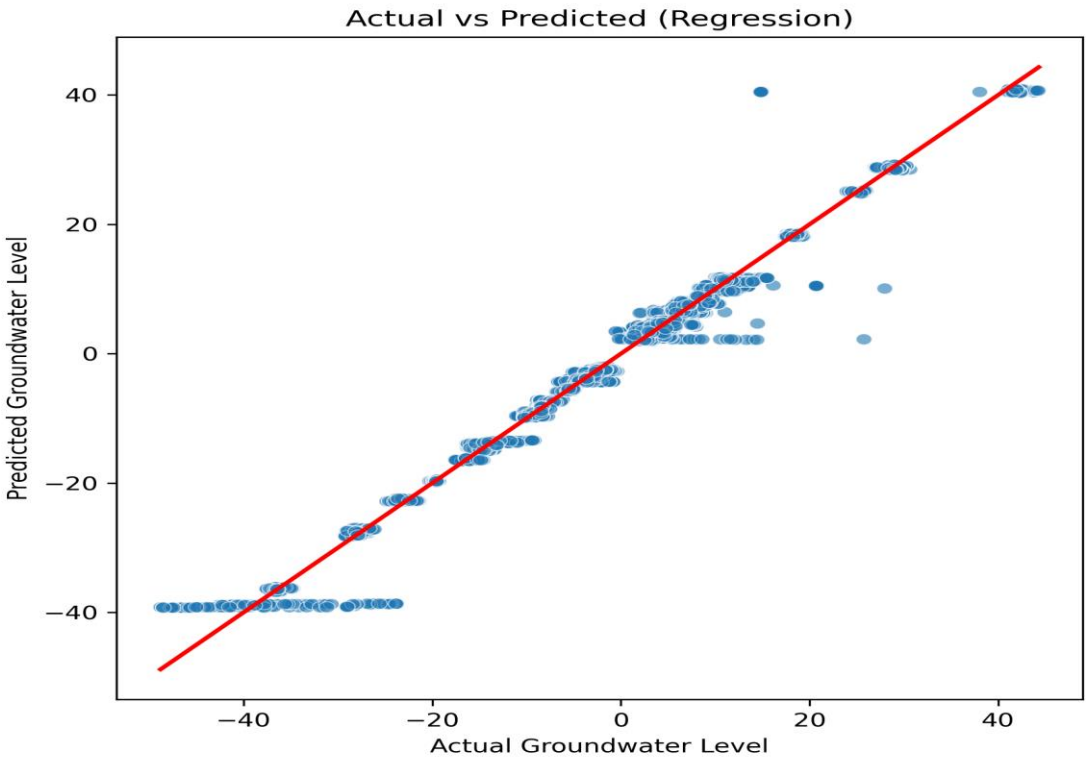
Model Fit Metrics:

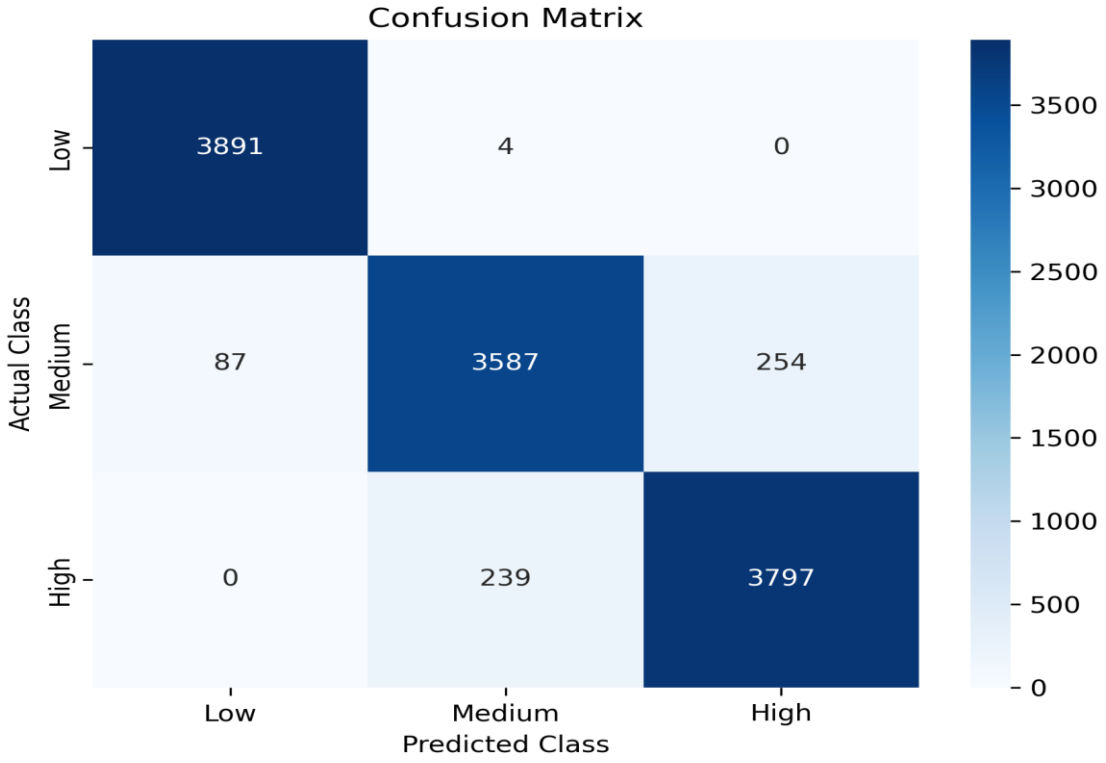
- R_squared(training): 0.9904
- Adj_R_squared: 0.9904
- F_stat: 49925.6491
- F_pvalue: 0.0
- Num_predictor: 58.0

7. Predictions & Evaluation

- Accuracy: 0.9508
- Precision: 0.9505
- Recall: 0.9508
- F1-score: 0.9505
- R-squared (test): 0.9868
- RMSE (regression): 1.7404
- MSE (regression): 3.0291

- Plots:


Actual vs Predicted (Regression)
Predicted Groundwater Level
Actual Groundwater Level
40
20
0
−20
−40
−40
−20
0
20
40


Confusion Matrix
Actual Class
Predicted Class
Low
Medium
High
3891
4
0
87
3587
254
0
239
3797
3500
3000
2500
2000
1500
1000
500
0

8. Significant Features & Interpretation

Top 12 Rows of Significant Features:

	A	B	C	D	E
1	**Section**	**Feature**	**Impact**	**Coef**	**P-value**
2	Significant Features	const	Positive	4108.849607183905	6.439817055628084e-165
3	Significant Features	stationCode.1	Negati...	-0.2262700591082868	0.0
4	Significant Features	HISTO_20	Negati...	-81.59101511042891	0.0
5	Significant Features	Rainfall (mm)_2024	Negati...	-0.5396856566801052	0.0
6	Significant Features	_count	Negati...	-5.4073280147480926e-05	0.0
7	Significant Features	shape_leng	Negati...	-0.0028493173731542	0.0
8	Significant Features	HISTO_80	Positive	574.5075759265251	0.0
9	Significant Features	Pre Monsoon of GW Trend_2024	Positive	47.434548616037375	0.0
10	Significant Features	district	Negati...	-9.034111405548687	4.234342942335636e-209
11	Significant Features	stationName	Positive	0.592188270324661	0.0
12	Significant Features	sand_5-15cm_mean	Positive	1.6084642611513504	0.0

64 2. Model Fit Metrics:

65 - R_squared(training data fit): 0.9904

66 - Adj_R_squared: 0.9904

67 - F_stat: 49925.6491

68 - F_pvalue: 0.0000

69 - Num_Predictors: 58.0000

71 3. Prediction Metrics:

72 - Accuracy: 0.9508

73 - Precision: 0.9505

74 - Recall: 0.9508

75 - F1-score: 0.9505

76 - R-squared (predict data): 0.9868

77 - RMSE (regression): 1.7404

78 - MSE (regression): 3.0291

Confidence in Interpretation:

- 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.

9. 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.

10. 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