Groundwater Level Prediction Using Multiple Linear Regression

1. Scenario

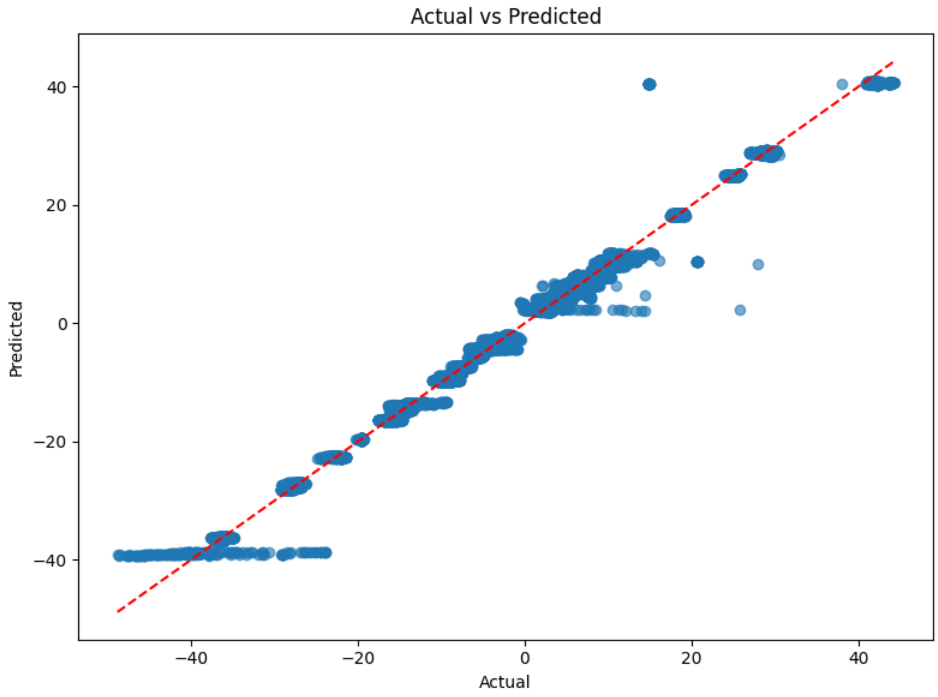
- Delhi NCR
- Objective: Identify key drivers of groundwater depletion and predict groundwater levels at unsampled locations.

2. Methodology

- **Data:** Preprocessed groundwater dataset with 175 columns including the target 'data Value'.
- **Dependent variable**: dataValue
- **Independent variables**: 58 selected features (BIC criterion)
- **Spatial unit**: District
- **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 R^2, RMSE.
- **Model Diagnostics**: Residuals analyzed for patterns.
- **EDA**: Scatter plots, pair plots, and spatial-temporal plots analyzed to observe trends.

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

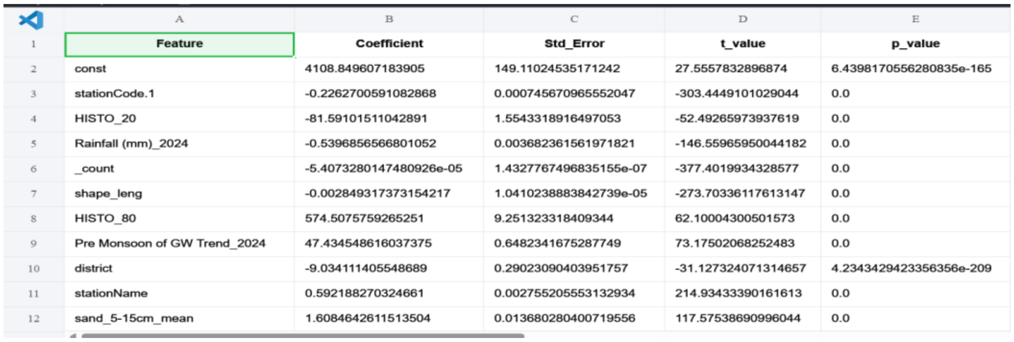
- Linearity: Relationships between predictors and target are linear.
- No Perfect Multicollinearity: Checked correlations among predictors.
- Exogeneity: Residuals uncorrelated with predictors.
- Homoscedasticity: Residuals have constant variance.

5. Model Selection

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

6. Model Estimation & Diagnostics

- Top 12 Coefficients:



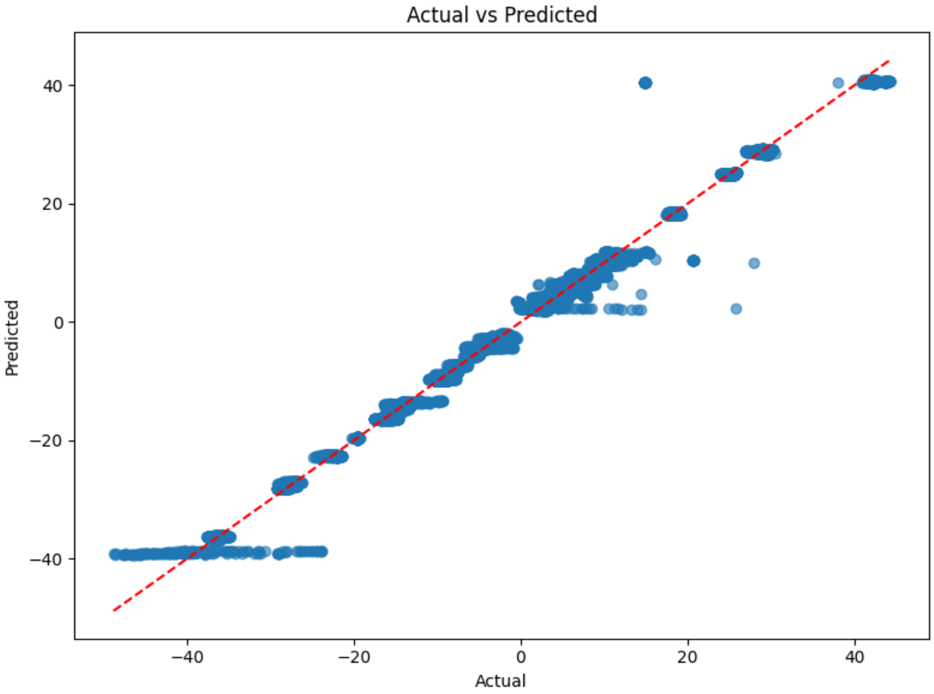
	A	B	C	D	E
1	**Feature**	**Coefficient**	**Std_Error**	**t_value**	**p_value**
2	const	4108.849607183905	149.11024535171242	27.5557832896874	6.4398170556280835e-165
3	stationCode.1	-0.2262700591082868	0.000745670965552047	-303.4449101029044	0.0
4	HISTO_20	-81.59101511042891	1.5543318916497053	-52.49265973937619	0.0
5	Rainfall (mm)_2024	-0.5396856566801052	0.003682361561971821	-146.55965950044182	0.0
6	_count	-5.4073280147480926e-05	1.4327767496835155e-07	-377.4019934328577	0.0
7	shape_leng	-0.002849317373154217	1.0410238883842739e-05	-273.70336117613147	0.0
8	HISTO_80	574.5075759265251	9.251323318409344	62.10004300501573	0.0
9	Pre Monsoon of GW Trend_2024	47.434548616037375	0.6482341675287749	73.17502068252483	0.0
10	district	-9.034111405548689	0.29023090403951757	-31.127324071314657	4.2343429423356356e-209
11	stationName	0.592188270324661	0.002755205553132934	214.93433390161613	0.0
12	sand_5-15cm_mean	1.6084642611513504	0.013680280400719556	117.57538690996044	0.0

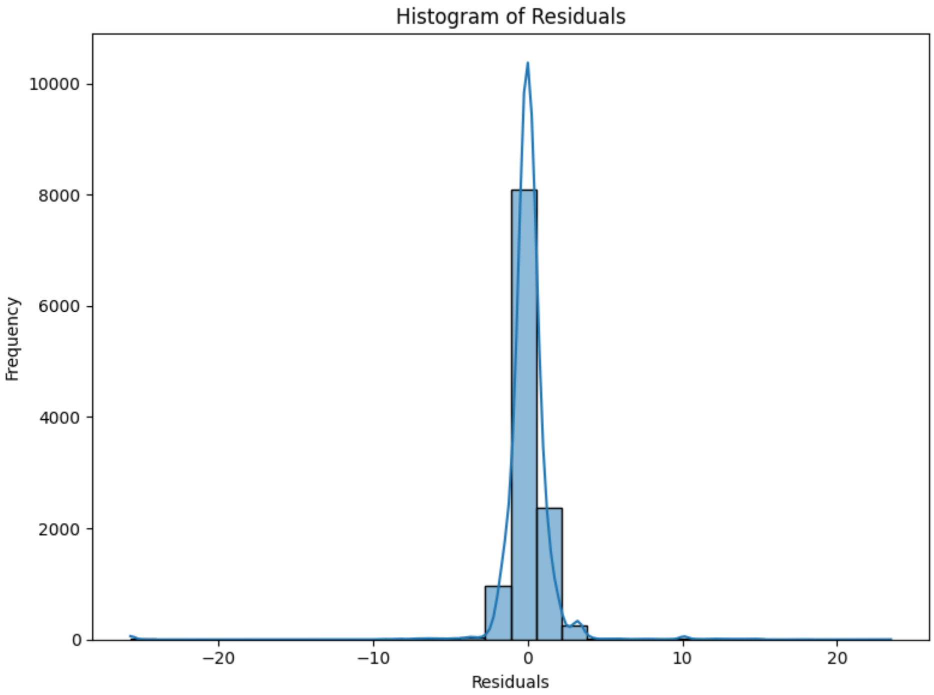
Model Fit Metrics:

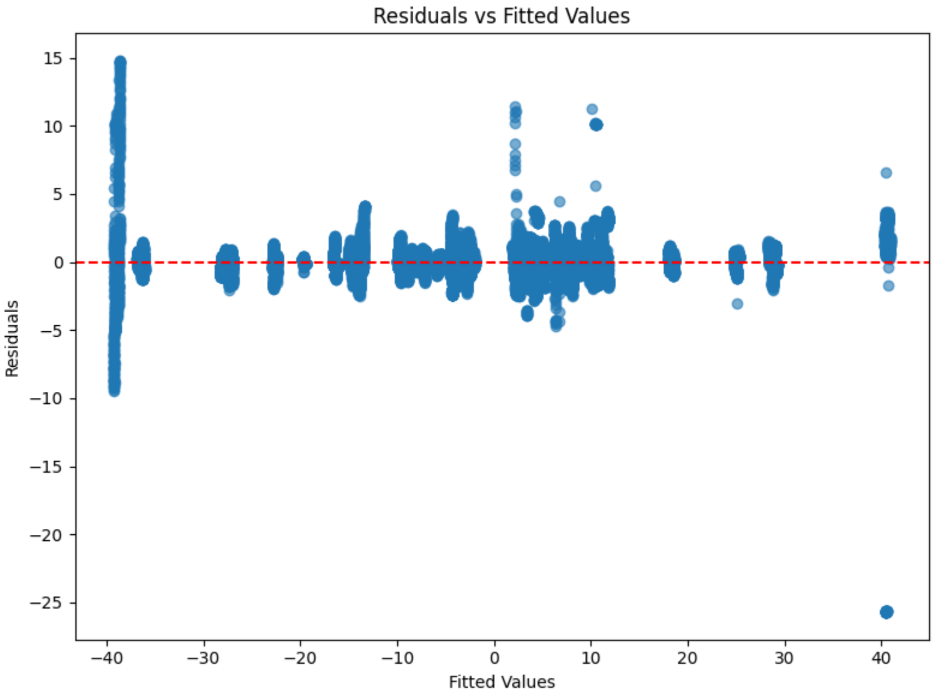
- R_squared(training): 0.9904
- Adj_R_squared: 0.9904
- F_stat: 49925.6491
- F_pvalue: 0.0000
- Num_Predictors: 58.0000

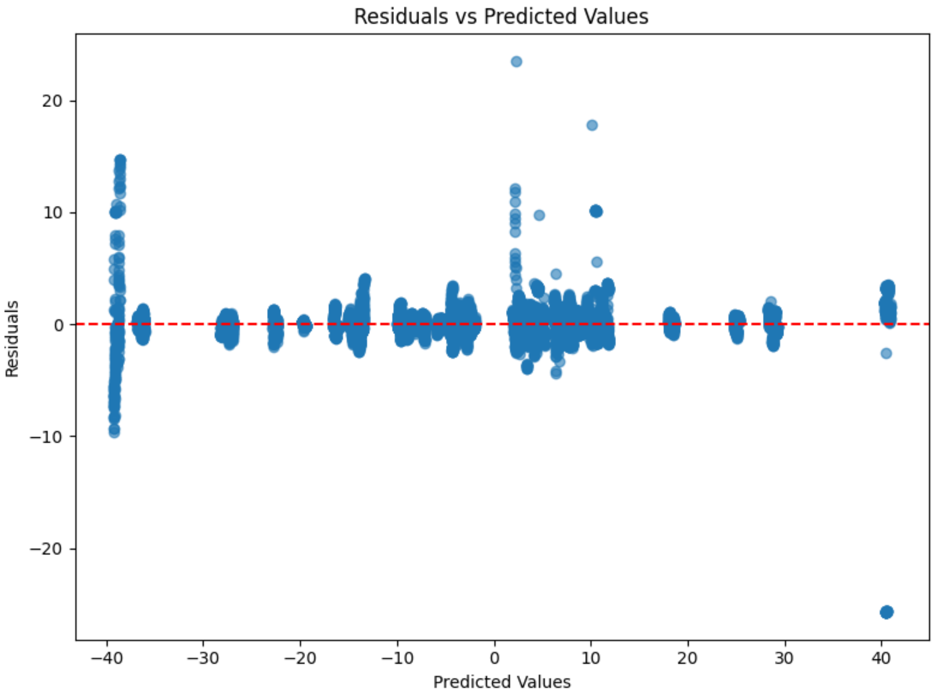
7. Predictions & Evaluation

- R-squared(testing data): 0.9868
- RMSE: 1.7404
- MSE: 3.0291
- Plots:



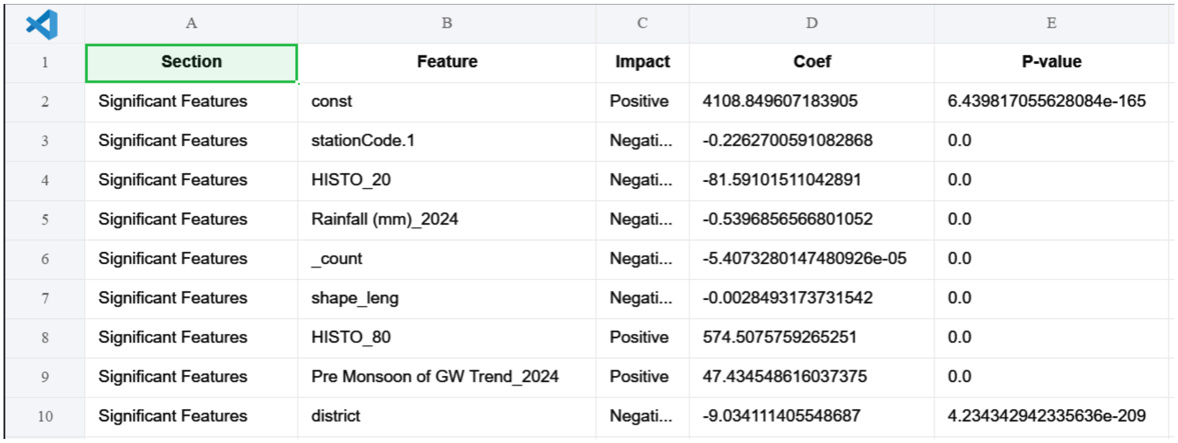

Histogram of Residuals
Frequency
10000
8000
6000
4000
2000
0
−20
−10
0
10
20
Residuals


Residuals vs Fitted Values
Residuals
15
10
5
0
−5
−10
−15
−20
−25
−40
−30
−20
−10
0
10
20
30
40
Fitted Values



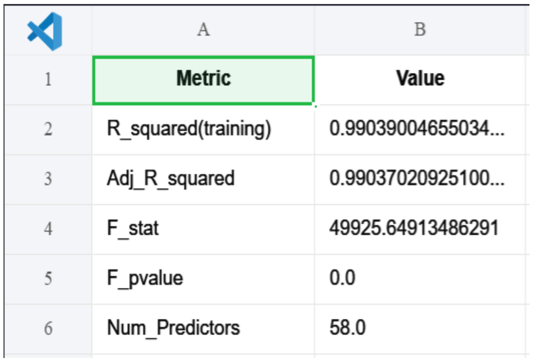
- Residual Plot
- The residual plot shows the difference between observed and predicted values. It helps in diagnosing the model fit.
- Interpretation: No clear pattern suggests a good fit.
- Action: Consider refining the model if patterns are detected.

8. Significant Features & Interpretation



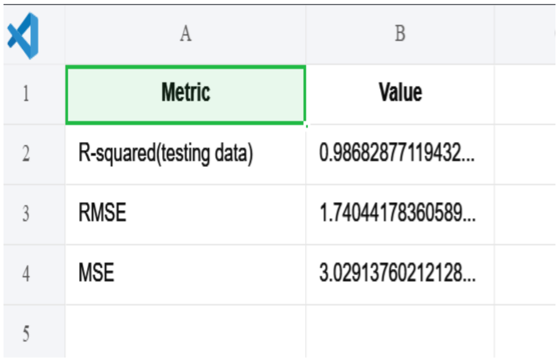
	A	B	C	D	E
1	**Section**	**Feature**	**Impact**	**Coef**	**P-value**
2	Significant Features	const	Positive	4108.849607183905	6.439817055628084e-165
3	Significant Features	stationCode.1	Negati...	-0.2262700591082868	0.0
4	Significant Features	HISTO_20	Negati...	-81.59101511042891	0.0
5	Significant Features	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

Model Fit Metrics:



Metric	Value
R_squared(training)	0.99039004655034...
Adj_R_squared	0.99037020925100...
F_stat	49925.64913486291
F_pvalue	0.0
Num_Predictors	58.0

Model Prediction Metrics:



Metric	Value
R-squared(testing data)	0.98682877119432...
RMSE	1.74044178360589...
MSE	3.02913760212128...

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.

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