Assignment-1

AI/ML & Applications (MDS-302)

M.Tech (Data Sciences)

SUBMITTED TO:
Dr. Saif Ali

SUBMITTED BY:
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

1. Introduction

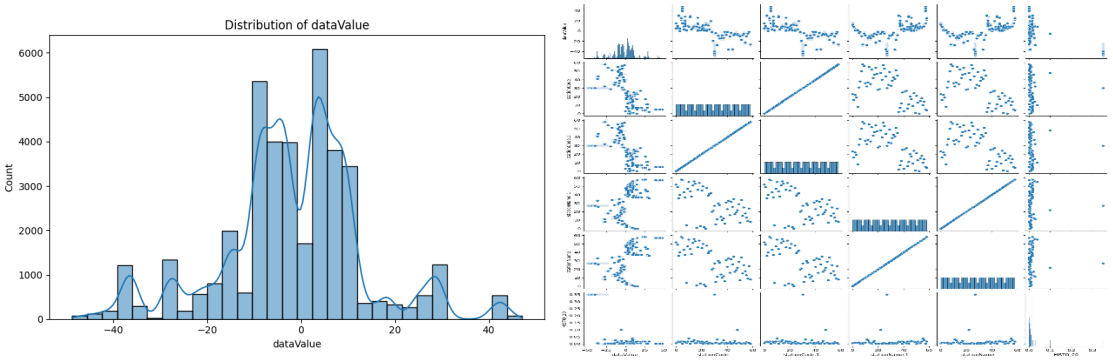
- Objective: Identify key drivers of groundwater depletion and predict groundwater levels at unsampled locations.
- Groundwater is a critical resource for agricultural, industrial, and domestic purposes, especially in rapidly urbanizing regions. Understanding the factors that influence groundwater levels 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.

2. Methodology

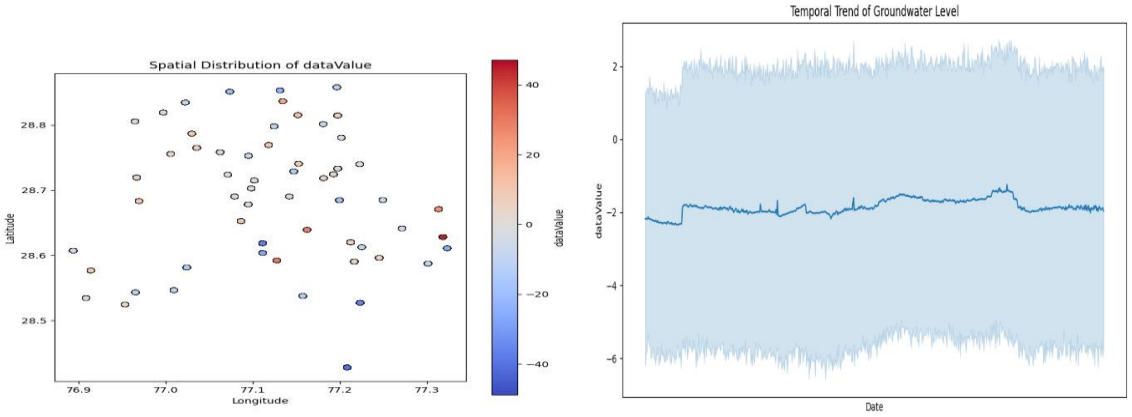
- **Data:** Pre-processed groundwater dataset with 175 columns including the target 'data Value' and 39000+rows.
- **Dependent variable**: data Value (Ground Water Level)
- **Independent variables**: 58 selected features (BIC criterion)
- **Spatial unit**: District, latitude and longitude
- **Temporal 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 R^2, RMSE.
- **Model Diagnostics**: Residuals analysed for patterns.
- **EDA**: Scatter plots, pair plots, and spatial-temporal plots analysed to observe trends.

3. Exploratory Data Analysis (EDA)

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



- Spatial and temporal trends were observed.

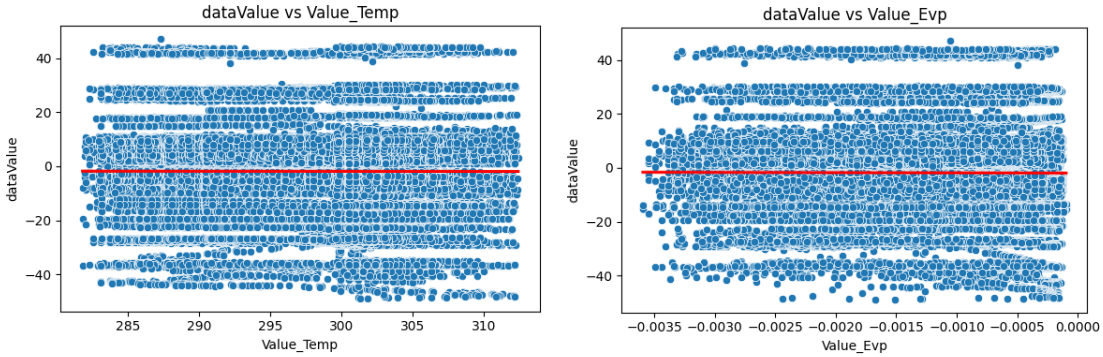


- **Summary Statics**: Computed for all features to identify ranges, distribution and central tendencies.

	A	B	C	D	E	F	G	H	I
1		**count**	**mean**	**std**	**min**	**25%**	**50%**	**75%**	**max**
2	date	39530.0	45534.5	193.414571192336...	45200.0	45367.0	45534.5	45702.0	45869.0
3	village	39530.0	29.0	17.0296017682723...	0.0	14.0	29.0	44.0	58.0
4	stationNa...	39530.0	29.0	17.0296017682723...	0.0	14.0	29.0	44.0	58.0
5	dataValue	39530.0	-1.8577254360362765	15.1358441037300...	-48.80...	-8.828...	-2.2985	6.46736363...	47.0895
6	id	39530.0	225413.89830508476	197.798981356421...	22493...	22521...	22549...	225555.0	22564...
7	subdistric	39530.0	8.35593220338983	6.73416926898451...	0.0	1.0	8.0	13.0	22.0
8	district	39530.0	4.423728813559322	3.00964635295361...	0.0	3.0	3.0	7.0	10.0
9	stationCo...	39530.0	29.0	17.0296017682723...	0.0	14.0	29.0	44.0	58.0
10	latitude	39530.0	28.689813559322037	0.101141116142699	28.4278	28.6072	28.6903	28.7694	28.8581
11	longitude	39530.0	77.12314576271186	0.10628292321268...	76.8939	77.0356	77.1311	77.1994	77.323
12	wellDepth	39530.0	51.958474576271186	19.42087977543538	15.55	49.0	49.0	49.0	178.0
13	Value_at...	39530.0	98388.30460429579	707.3485402643782	96508....	97752....	98483...	99004.2128...	99665...
14	Value_Evp	39530.0	-0.00112396734888...	0.00081506547901...	-0.003...	-0.001...	-0.000...	-0.00046352...	-9.621...
15	Value_Rain	39530.0	0.001032396639641...	0.00361192396815...	0.0	3.5514...	2.024...	0.00049594...	0.097...
16	Value_Te...	39530.0	297.4062515911138	7.12826893728139...	281.84...	291.07...	299.2...	302.676452...	312.4...

4. Model Assumptions

Linearity: Relationships between predictors and target are linear.



No Perfect Multicollinearity: Checked correlations among predictors.

VIF ≈ 1 → No multicollinearity

VIF < 5 → Acceptable, usually fine

VIF ≥ 5 → Potential multicollinearity issue

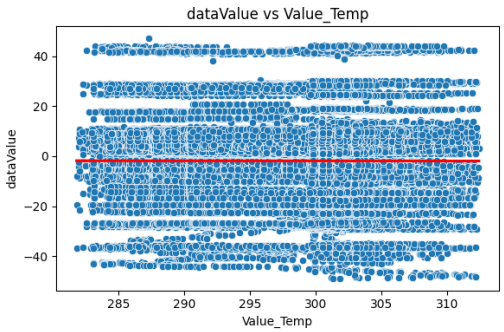
VIF ≥ 10 → Serious multicollinearity problem, feature needs attention

There is some features VIF values in below pic:

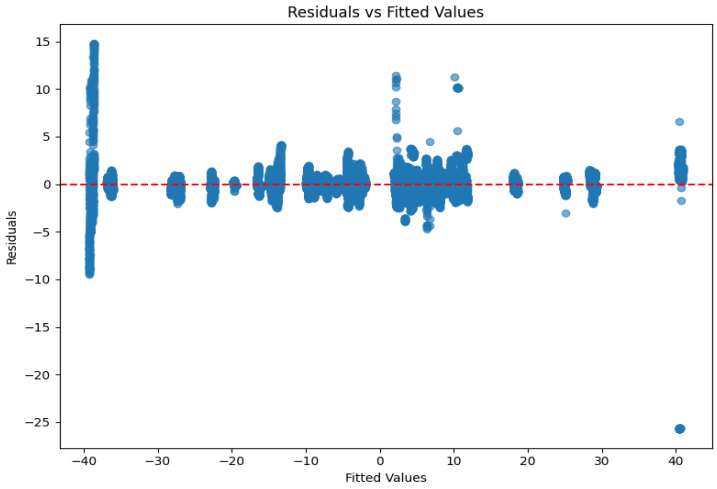
167	Value_Sm1	11.808873896663833
168	Value_atm_pre	9.359134608584789
169	Value_Temp	6.215853181317188
170	date	5.873867536679541
171	Value_Sm3	5.323647044025617
172	Value_Evp	3.613515636701771
173	Value_Rain	1.53144020452512
174	Rainfall (mm)_2025	0.0

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.



5. Model Selection

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

6. 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:**
    - 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:

- $R^2 = 1$: Perfect prediction

- $R^2 = 0$: Model predicts no better than the mean
- $R^2 < 0$: 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

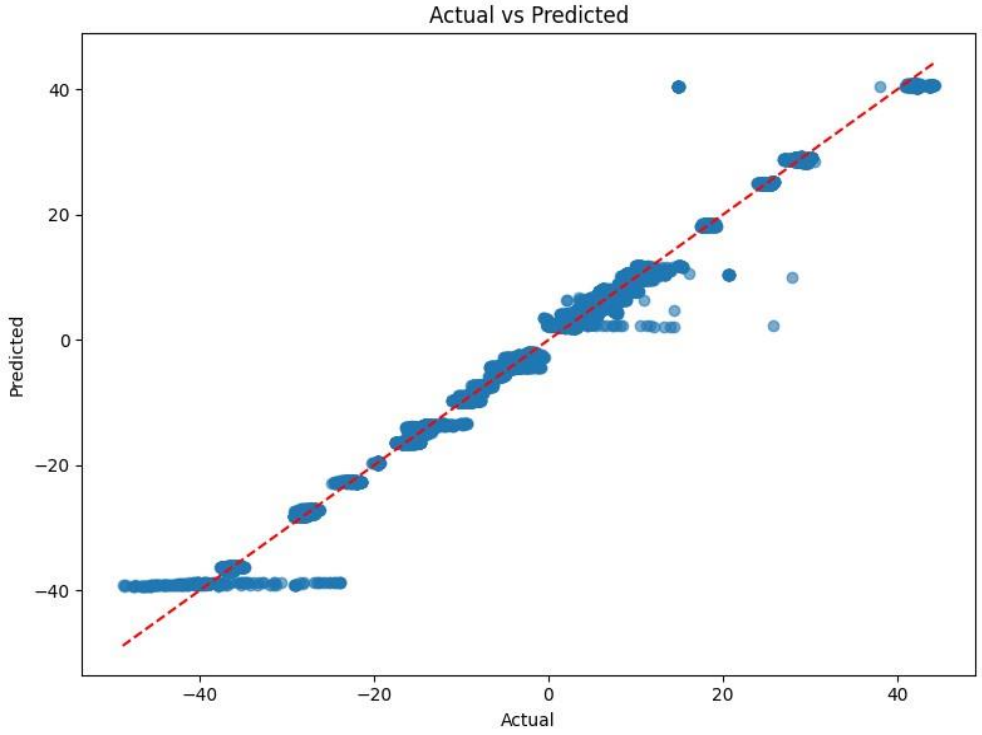
Model prediction Metrics on Basis of BIC model Selection:

Statics	**values**
R_squared(testing data)	0.9868
RMSE	1.7404
MSE	3.0291

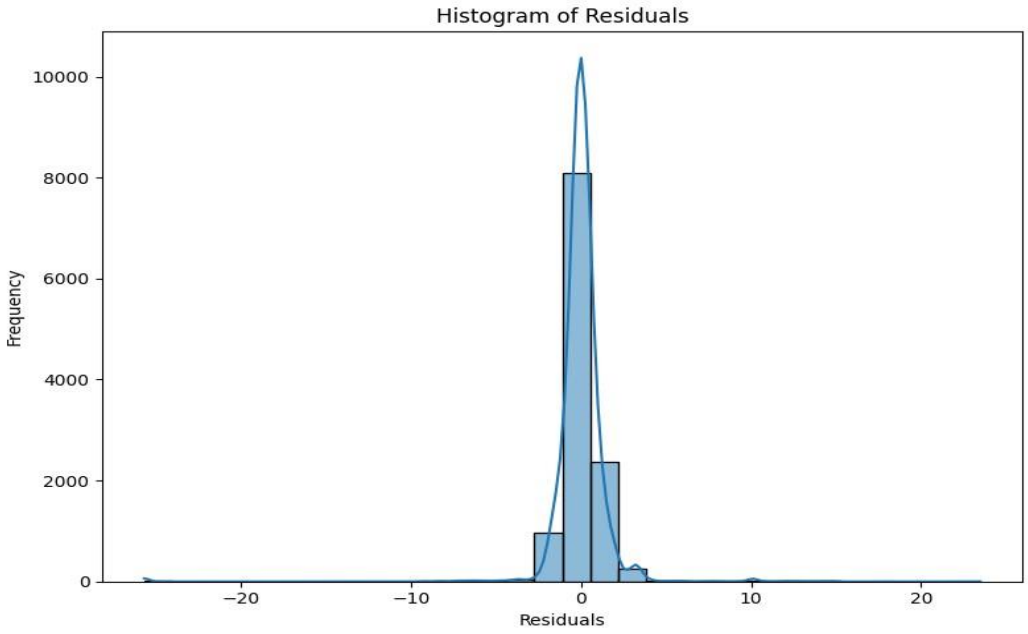
Model performance was evaluated using regression metrics:

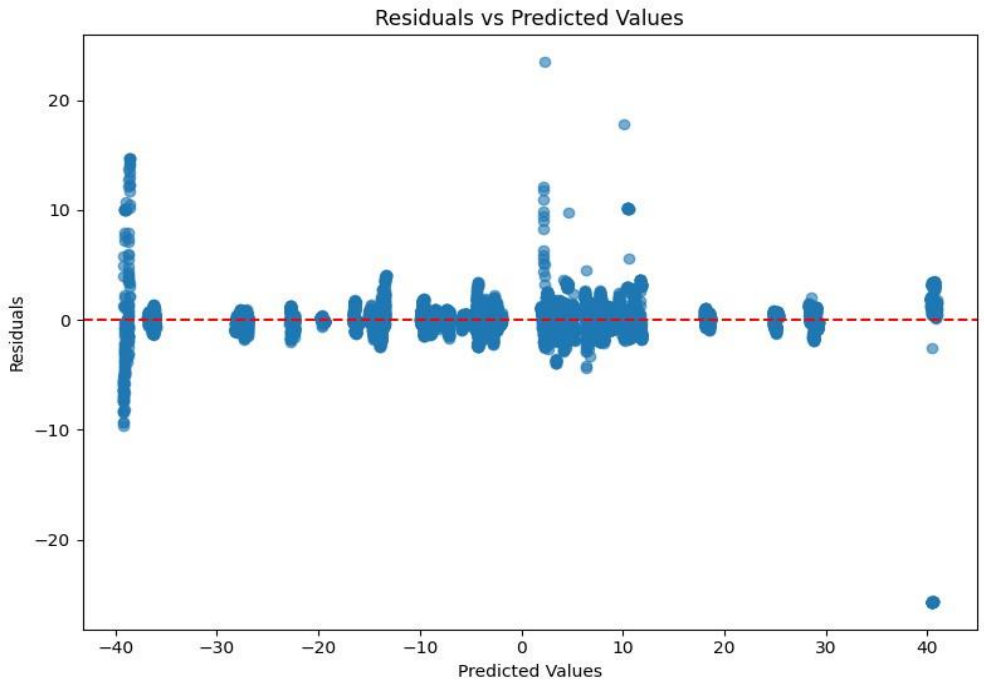
- **Mean Squared Error (MSE):** 3.0291, indicating the average squared deviation between actual and predicted groundwater levels.
- **R^2 Score:** 0.9868, showing that approximately 98.68% of the variance in groundwater levels was explained by the model.

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.

7. Significant Features & Interpretation

Significant factors and impact outcomes:

	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:

	A	B
1	Metric	Value
2	R_squared(training)	0.99039004655034...
3	Adj_R_squared	0.99037020925100...
4	F_stat	49925.64913486291
5	F_pvalue	0.0
6	Num_Predictors	58.0

Model Prediction Metrics:

	A	B
1	Metric	Value
2	R-squared(testing data)	0.98682877119432...
3	RMSE	1.74044178360589...
4	MSE	3.02913760212128...
5		

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.

8. 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.

9. 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