Gound Water Level Prediction Using Machine Learning Concepts

Valluri Keerthi Ram[1], Kamma Ajay Nageswara Rao[2], Abhay Gokavarapu[3]
Department of Computer Science and Engineering
Amrita School of Computing, Bengaluru
Amrita Vishwa Vidyapeetham, India
keerthiramvalluri@gmail.com, kammaajay11@gmail.com, abhaygokavarapu@gmail.com

***Abstract*—Groundwater depletion was always a problem to sustainable development and water security. This study uses historical hydrological and meteorological data to forecast groundwater levels based on a number of traditional ML models, including Random Forest, Gradient Boosting and Support Vector Regression. Once again, Random Forest delivered the best forecasting accuracy of all the models tested. Overall, Random Forest provides a viable and affordable solution for forecasting groundwater levels. Rainfall prediction will be used in future studies to potentially improve model performance further.**

***Index Terms*—Groundwater prediction, Machine learning, Random Forest, Regression models, Water management**

I. Introduction

Groundwater is one of the most critical sources of freshwater for businesses, agriculture, and human beings. It is the main or unique water supply for many locations world wide (i.e., mostly and particularly in drought area zones). However, groundwater levels have increasingly dropped due to unsustainable extraction, uncontrolled use, and climate variability. This problem has been exacerbated by ignorance, bad planning, and lack of forecasting / prediction tools. Therefore, anticipating groundwater availability has become fundamental for water resource managers and policymakers.

Depletion of groundwater has far-reaching effects. Crop failure, decreased access to drinking water, higher pumping costs, land subsidence, and even groundwater contamination can result from falling water tables. Small-scale farmers and rural communities are especially at risk because they frequently rely solely on wells for domestic and irrigation needs. From a societal standpoint, water insecurity can exacerbate health problems, migration, and economic instability. In order to guarantee food security, water equity, and long-term environmental sustainability, this issue must be resolved.

Piezometers and observation wells are used for physical measurements in traditional groundwater monitoring. Despite their accuracy, these methods are labor-intensive, expensive to maintain, and have a limited spatial coverage. As a result, numerous statistical and empirical models have been created to examine trends in groundwater. These methods, however, frequently fall short in capturing intricate, nonlinear relationships between influencing factors like land use, temperature, humidity, and historical water table data.

This study suggests using machine learning algorithms to model and forecast groundwater levels using historical data in order to get around the drawbacks of traditional approaches. The emphasis is on supervised learning methods that can model nonlinear relationships between several variables, such as RFR(Random Forest Regression), GB(Gradient Boosting) and SVR(Support Vector Regression). Historical datasets containing groundwater levels as well as pertinent hydrological and meteorological parameters are used to train these models. The ultimate objective is to create a reliable, affordable, and deployable prediction system that can assist in groundwater management decision-making.

This research focuses on machine learning methods that are better suited for my work environment, as they need less data, are easier to implement, and more interpretable, rather than relying on deep learning and other more complex black-box methods like many other recent studies do. With this approach, accuracy and transparency are both achieved, therefore, it can be more readily applied by water boards, NGOs, and local governments.

The best way to increase the ground water level prediction is to incorporate the rainfall prediction as a future extension of this work. Rainfall prediction can significantly improve the accuracy of groundwater forecasting as it is a major source for groundwater recharge, therefore, it is vital for the model to account surface and subsurface water interactions. This will enable the system to improve in real-time and help develop a sophisticated water prediction system that integrates environmental dynamics.

II. LITERATURE SURVEY

M.B. Raj et al. (2024) conducted a single comprehensive evaluation of machine learning methods for predicting groundwater levels. The authors focused strictly on the assessment of the supervised learning models' performance to accurately predict groundwater levels. They showcased the two-pronged approach on the importance of algorithm selection and apply-pre-processing of data as a means of improving performance. In the authors' view, machine learning approaches, if implemented appropriately, can provide effective and adaptable solutions for managing water resources, especially in areas

where conventional hydrological models are ineffective as a result of limited data [1].

T. Zeng et al. (2022) applied a hybrid model of Artificial Intelligence (AI) for groundwater level forecasting. The authors used AI techniques to capture the nonstationary, nonlinear variations of groundwater levels using SVR(Support Vector Regression modelling) and EMR(Empirical Mode Decomposition modelling) methods. The authors showed that the application of physically based and data-driven approaches could enhance forecasting accuracy in areas with complex geological conditions, supporting risk management and landslide early warning systems using forecasting data to predict risk [2].

F. Feng et al. (2024) conducted a study regarding the effectiveness of traditional ML models and DL architectures in predicting groundwater levels. It featured models including LSTMs and CNNs, in addition to traditional ML heuristics such as Decision Trees, Random Forests, and Support Vector Machines. It was noted by the authors that although traditional models were cheaper to compute, deep learning approaches, through the use of LSTMs, performed significantly better in time-series data with long-range dependencies in comparison to the simpler algorithms. Such research exemplifies the increasing influence of deep learning algorithms in hydrology [3].

V. Kumar et al. (2023) investigated how various machine learning models can be utilized to forecast rainfall in urban municipalities. While the primary purpose was not rainfall forecasting but groundwater modeling, the research results can still be utilized because of the highly correlated nature of rainfall trends and groundwater recharge. They evaluated model performance for models including Gradient Boost, Random Forest, and Deep Neural Networks and provided a performance baseline across the urban settings. One caveat is that the respective spatial and temporal characterizations/histories of the sites greatly affect model performance accuracy in these environmental forecasting applications [4].

N. Mungale et al. (2024) did an extensive comparison study between conventional machine learning models and deep learning models for dynamic rainfall forecasting. They used real-world environmental data and utilized models such as XGBoost, ANN, and LSTM networks. They found that the deep learning models, specifically LSTM, produced better accuracy in forecasting over the long term, while conventional machine learning models were better equipped to provide short-term forecasts to their computational efficiency and other overheads. Thus, the findings of this research will be useful for forecasting groundwater level, particularly in instances that rainfall is the dominant contributor for recharge [5].

T. V. Tran et al. (2025) exemplified a new way to reconstruct and predict groundwater time series with machine learning in the Allertal area of Germany. They took on the issue of missing data, which is an ongoing issue for groundwater datasets, using imputation procedures and regression-based machine learning models, including Random Forest and Gradient Boosting. The authors demonstrated to some extent the validity of these models to identify seasonal and interannual variation in groundwater levels, and provided useful information about how machine learning can enhance approaches to water resource planning and resilience in these temperate climate zones [6].

A. Jari et al., (2023) described the application of individual and ensemble machine learning models to map groundwater potential zones in the arid area of Tan-Tan in Morocco. The authors utilized input parameters such as geology, slope, rainfall, soil type, and land use, and applied models such as Random Forest, Logistic Regression, and ensemble methods to produce groundwater potential maps. The research showed the utility of ensemble models by higher accuracy and generalization. The implications of the findings are critical for groundwater exploration and sustainable utilization in arid and water-scarce regions [7].

A. Ali et al. (2024) developed an advanced machine learning model that proposed using Transformer models for forecasting groundwater levels in the Thames Basin, London. The authors investigated more standard models (i.e., GRU, LSTM) as well as more recent deep architectures and illustrated that Transformers improved long-horizon prediction ability much better than other models. The authors emphasized the importance of high reliability in the temporal observations and outside features such as weather and land use in ultimately improving model performance. This work represents one of the first example of demonstrating Transformers can be used in hydrogeology for time-series forecasting, and thus opens up additional avenues for research [8].

G. Tuysuzoglu et al. (2023) have recommended the use of an ensemble machine learning model to improve rainfall prediction, using the K-Star algorithm. The study aims to improve prediction of rainfall patterns by introducing K-Star, a lazy learning algorithm that integrates instance-based learning with distance functions based on entropy in an ensemble configuration. The models were evaluated against a range of individual learners and superior models were detected, especially where the meteorological data was noisy and irregular. The authors commented on the strength of ensemble methods to promote generalization and robustness, particularly with chaotic weather systems. Overall, this research makes a rich contribution to model selection for hydrological variables that may indirectly impact elevated levels of groundwater since rainfall is often a critical driver of recharge [9].

P. N. Triveni et al. (2023) offered a detailed review of rainfall forecasting methods, highlighting the application of Multiple Linear Regression (MLR) and Artificial Neural Networks (ANN). They integrated the findings from several studies and weighed the advantages and shortcomings of MLR (a linear statistical technique) and ANN (a non-linear adaptive model) in describing spatiotemporal patterns of rainfall. The review highlighted that as MLR is simple to interpret and beneficial for linear relationships, ANN is better at discovering sophisticated, non-linear relationships inherent in meteorological data. The authors concluded hybrid models fusing statistical and machine learning techniques hold great promise. Although the paper is a review, it presents the basics important for researchers of groundwater level forecasting, considering

rainfall plays a core role in hydrological cycles [10].

III. METHODOLOGY

Q1.) We trained a classification model using selected features from the dataset. The dataset was split into a training dataset and a test dataset. After fitting the model the confusion matrices were created for the training and test datasets in order to calculate precision, recall, F1-score, and accuracy. The models learning behaviour was deemed underfit, regularfit and overfit based on the training and test score.

Q2.) We executed a price prediction task using a regression model trained using historical data. After we made our price predictions on the test set, we calculated RMSE(Root Mean Squared Error) , regression metrics, MAPE(Mean Absolute Percentage Error), Mean Squared Error (MSE) and R^2 score. The metrics were analyzed in order to examine the prediction accuracy and model performance.

Q3.) We generated 20 random points of data (with 2 features (X and Y) between 1 and 10), we used prescribed rules to assign Class 0 (Blue) or Class 1 (Red) to each point. Then we created a scatter plot to see how the points were laid out, and if the points were separated between the classes in the 2D feature space.

Q4.) Using a k-Nearest Neighbors (k=3) algorithm trained with the 20 point dataset, a test set of 10,000 grid points (where X and Y are from 0 to 10 at 0.1 intervals) was classified, and then we plotted our output classified test points to look at the class boundaries and spatial process decision regions.

Q5.) We repeated the Q4 classification task using multiple values of k=1,3,5,10. For each k, we generated a new scatter plot to examine how the decision boundaries shifted. This illustrated how model complexity and smoothness of classification vary as a function of k.

Q6.) We created two numerical features from the project dataset in a similar way that was done in Q3–Q5, we used a small subset of our dataset for training and then a dense grid for testing. From there we once again used a kNN classifier, and visualised the impact of various k values on the class boundary and model behaviour using scatter plots.

Q7.)To determine the optimal value of k for the kNN classifier, we utilized GridSearchCV to cross-validate k values across a range of k values. We selected the best performing k based on cross-validation accuracy. Final model performance was evaluated with a confusion matrix and classification report to verify performance.

IV. RESULTS

From observing Table 1 and Table 2 we can say that it is overfit because the accuracy of Table -1 is higher than the accuracy of Table-2.

From observing Table 3 we can say that relationship between the selected features (Latitude, Longitude, Well-Depth) and Pre-monsoon-2022 is not strong or not linear.

Q9.)Best k value found: 1

Best cross-validation accuracy: 99.56

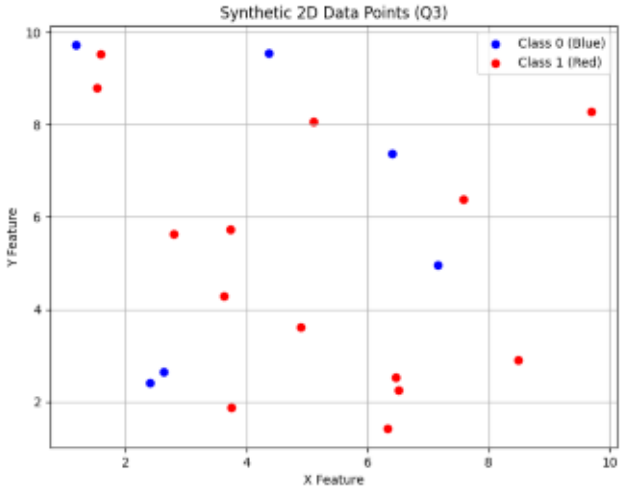

Fig. 1. Scatter plot of synthetic 2D data points used for classification in Q3.

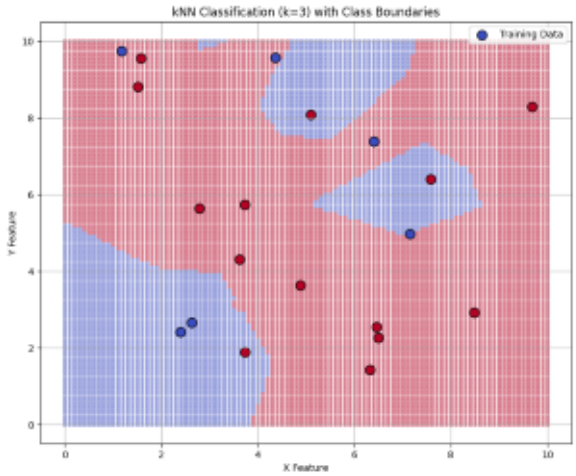

Fig. 2. Decision boundaries generated by k-NN classifier (k = 3) on the synthetic dataset.

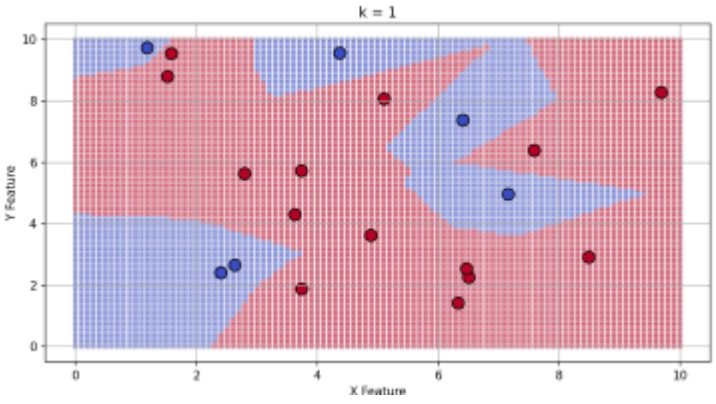

Fig. 3. Comparison of k-NN decision boundaries for different values of k=1

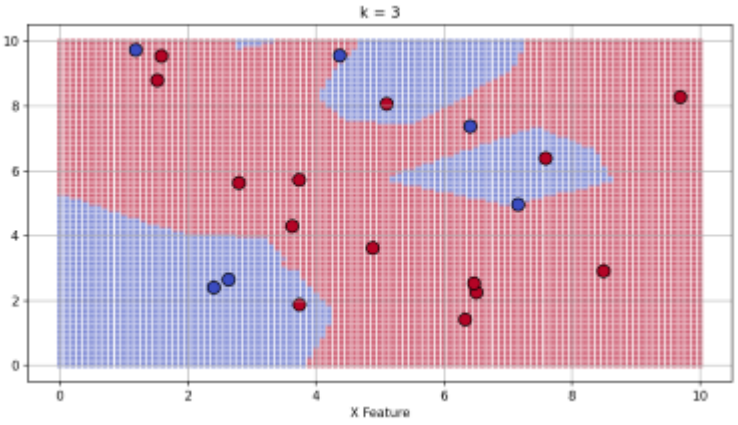

Fig. 4. Comparison of k-NN decision boundaries for different values of k=3

TABLE I
TRAIN CLASSIFICATION REPORT

Class	Precision	Recall	F1-Score	Support
0	0.92	0.92	0.92	2237
1	0.87	0.87	0.87	1390
Accuracy			0.90	3627
Macro Avg	0.90	0.90	0.90	3627
Weighted Avg	0.90	0.90	0.90	3627

TABLE II
TEST CLASSIFICATION REPORT

Class	Precision	Recall	F1-Score	Support
0	0.86	0.85	0.85	958
1	0.76	0.78	0.77	597
Accuracy			0.82	1555
Macro Avg	0.81	0.81	0.81	1555
Weighted Avg	0.82	0.82	0.82	1555

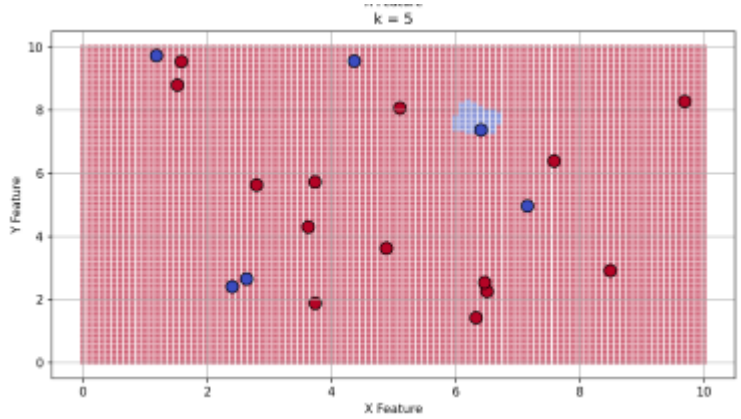

Fig. 5. Comparison of k-NN decision boundaries for different values of k=5

TABLE III
REGRESSION EVALUATION METRICS

Metric	Value
MSE	438.99
RMSE	20.95
MAPE	120.36%
R^2	0.37

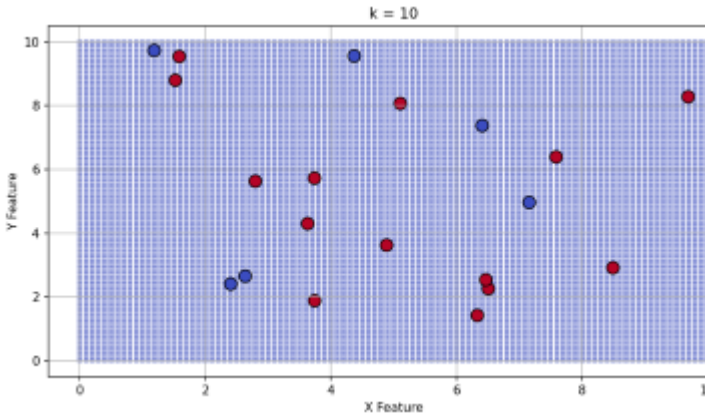

Fig. 6. Comparison of k-NN decision boundaries for different values of k=10

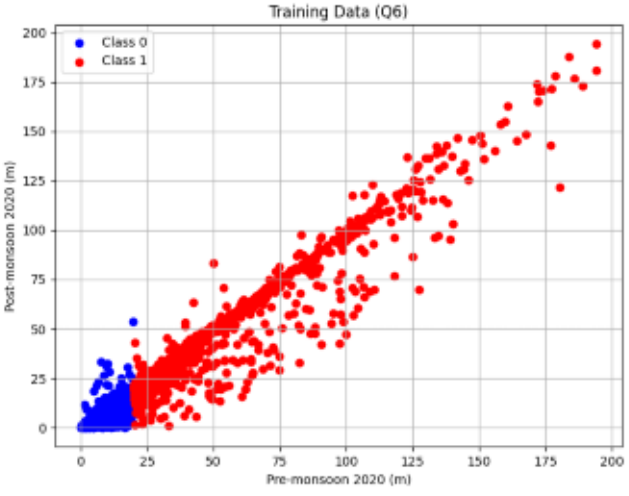

Fig. 7. Scatter plot showing real-world data classification for Q6 based on Pre-monsoon and Post-monsoon groundwater levels.

Test set accuracy: 100

The synthetic 2D data shows two separate classes (Class 0 is blue and Class 1 is red) spread out in the feature space. Between the two classes, there is moderate overlap, but most of the points in the feature space are spatially grouped. This spatial grouping of points suggests that k-NN, a distance-based classifier, is able to effectively discriminate two classes as long as they are spatially separated. The notation of the distributions in this 2D data is perfect for visually observing how class boundaries establish given a number of different k-values.(For Fig 1).

The k-NN classification result for k = 3 shows clear, non-linear decision boundaries between the two classes. The blue shaded regions are the predicted Class 0 and the red shaded regions are the predicted Class 1. The classifier has adapted well to the local structure of the data, sufficient for the nature of the distribution. For k = 3 at this point the model strikes a reasonable balance between flexibility and generalization.(for Fig 2)

The decision boundaries for k = 1, 3, 5 and 10 had many similarities and many differences. Notice that the smaller k-values such as k = 1 had more complexity and therefore more flexibility in their boundaries. So while the model with k = 1 looked great on the training data, it is not desirable because of how it is formally overfitting the training data. In general, as k increased, the boundaries became smoother and therefore increasingly generalized, especially at k = 10, where the regions fell into very large parts of the planes and less influenced by noise in the training data. The visual interpretation of these boundaries as it pertains to model complexity and generalization provides a poignant example of this trade-off in k-NN.(For Fig 3,4,5,6)

The plot in Q6 shows a set of real-world groundwater level data with Pre-monsoon and Post-monsoon 2020 levels as features. The classes are well separated along a diagonal trend meaning there could be a strong linear relationship between the two features. Class 0 (blue) points are concentrated at the lower values while Class 1 (red) points are diffused across higher ranges. This suggests the two features chosen are good at differentiating the two groundwater levels.(For Fig 7)

REFERENCES

[1] M. B. Raj, C. Priya (2024). GROUND WATER LEVEL PREDICTION USING MACHINE LEARNING. International Research Journal of Modernization in Engineering Technology and Science.

[2] T. Zeng, K. Yin, H. Jiang, X. Liu, Z. Guo and D. Peduto (2022). Groundwater level prediction based on a combined intelligence method for the Sifangbei landslide in the Three Gorges Reservoir Area. Scientific Reports, 12(1). .

[3] F. Feng, H. Ghorbani and A. E. Radwan (2024). Predicting groundwater level using traditional and deep machine learning algorithms. Frontiers in Environmental Science, 12. h

[4] V. Kumar, N. Kedam, K. V. Sharma,K. M. Khedher and A. E. Alluqmani (2023). A comparison of machine learning models for predicting rainfall in urban metropolitan cities. Sustainability, 15(18), 13724.

[5] N. Mungale and J. Shinde (2024, January 12). Rainfall Forecasting: A Comparative Analysis of Deep Learning and Machine Learning Models with Application to Environmental Data.

[6] T. V. Tran, A. Peche, R. Kringel, k. Brömme and S. Altfelder (2025). Machine Learning-Based Reconstruction and Prediction of groundwater time Series in the Allertal, Germany. Water, 17(3), 433.

[7] A. Jari, E. M. Bachaoui, S. Hajaj, A. Khaddari, Y. Khandouch,, A. E. Harti, A. Jellouli and M. Namous (2023). Investigating machine learning and ensemble learning models in groundwater potential mapping in arid region: case study from Tan-Tan water-scarce region, Morocco. Frontiers in Water, 5.

[8] A. Ali, A. Ahmed and M. Abbod (2024, November 28). Groundwater level predictions in the Thames Basin, London over extended horizons using Transformers and advanced machine learning models.

[9] G. Tuysuzoglu, K. U. Birant and D. Birant (2023). Rainfall prediction using an ensemble machine learning model based on K-Stars. Sustainability, 15(7), 5889.

[10] P. N. Triveni, Dr. G. JawaherlalNehru, Dr. R. Santhoshkumar and S. BavanKumar (2023). A REVIEW ON RAINFALL PREDICTION USING MACHINE LEARNING ALGORITHMS: MLR AND ARTIFICIAL NEURAL NETWORK. In ResMilitaris: Vol. vol.13 (Issue n°4) [Journal-article].