Groundwater Level Prediction Using Multiple Machine Learning Techniques

Kotari Jayanth
Department of Computer Science Engineering
Koneru Lakshmaiah Education
Foundation
Vaddeswaram, Andhra Pradesh,
India
jayanthkotari2001@gmail.com

Mahammad Abdul Rawoof
Department of Computer Science Engineering
Koneru Lakshmaiah Education
Foundation
Vaddeswaram,Andhra Pradesh,
India
mohammadrawoof8@gmail.com

Sai Prasanna Vamsi Perni
Department of Computer Science Engineering
Koneru Lakshmaiah Education
Foundation
Vaddeswaram,Andhra Pradesh,
India
pernivamsi123456@gmail.com

Valasapalli Mounika
Department of Computer Science Engineering
Koneru Lakshmaiah Education
Foundation
Vaddeswaram,Andhra Pradesh,
India
vmounika@kluniversity.in

Abstract

Predicting groundwater levels is important for managing water resources sustainably. Accurate predictions can aid in the efficient allocation of water resources, preventing over-extraction, and minimizing environmental impacts. In the present study, we employ ML (Machine Learning) algorithms to predict groundwater levels based on historical data. We utilize a comprehensive dataset comprising groundwater level measurements, meteorological data, and other relevant parameters. The dataset is preprocessed to handle missing values and ensure its suitability for ML. Four ML algorithms, including SVM (Support Vector Machine), Logistic Regression, KNN (K-Nearest Neighbors), and Gradient Boosting, are employed for groundwater level prediction. These algorithms are trained on historical data and evaluated using appropriate metrics. Our experiments reveal the effectiveness of ML in predicting groundwater levels. We present comparative results for the four algorithms, including accuracy scores and predictive capabilities. Additionally, we visualize the model performance through confusion matrices.The comparative analysis highlights the strengths and weaknesses of each algorithm in groundwater level prediction. We discuss how these findings impact water resource management and the viability of embedding predictive models into decision support systems.

Keywords— *Groundwater level prediction,ML, SVM, Gradient Boosting, KNN, water resource management.*

I. INTRODUCTION

In India, groundwater is a priceless natural resource that is essential for meeting the country's enormous population's water demands as well as for delivering water to numerous industries. Similar to many other countries, India relies heavily on groundwater for a range of purposes, making it a critical component of the nation's water resources infrastructure.[1]

Groundwater Usage in India:Groundwater is a significant source of water for household use in millions of houses around the country. It is a lifeline for rural areas where people depend on wells and groundwater for drinking and household chores. In urban areas, public water supply systems often extract groundwater to meet the demand of residents.Indian agriculture heavily relies on groundwater for irrigation. Farmers use wells and tube wells to access groundwater to irrigate crops, especially during dry seasons.

Groundwater is essential to the procedures and operations of many commercial and industrial companies.[2] Industries such as manufacturing, textiles, and food processing rely on groundwater as a crucial resource. Groundwater also contributes to sustaining natural ecosystems. It helps maintain springs, sustains water in marshlands, ponds, and swamps, and feeds streams, bays, and rivers, which are essential for the environment.

Water Scarcity and Drought Mitigation: India faces challenges related to water scarcity and recurring droughts in various regions. Planning and managing water supplies during these emergencies depends on being able to predict when groundwater will be available.Groundwater is integral to the Green Revolution in India, which transformed the agricultural landscape. Effective management and prediction of groundwater availability are critical to maintaining agricultural productivity and ensuring food security.

Much research has been carried out utilizing different simulation techniques to forecast GWL. These techniques include conceptual models with a physical foundation, experimental models [3, 4, 5,] and numerical models.AI (Artificial intelligence) models have been prominent during the last two decades as possible substitutes for traditional numerical models in GWL simulation.

Fig. 1 illustrates a roadmap, highlighting geographical regions with extensive GWL modeling studies and areas that remain relatively unexplored [2]. It also spotlights four major countries with robust GWL modeling research, while the black zone represents regions where AI applications have yet to gain prominence. Nearly 70% of areas have not tapped into GWL modeling, either due to an abundance of surface water resources or lower population densities, such as polar regions or parts of Russia.

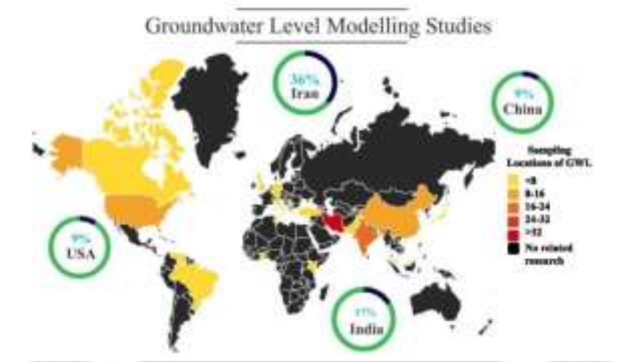


Fig.1. A map showing the locations of GWL data sampling around the world in a certain region without any associated study on GWL modeling using AI models.

II. LITERATURE REVIEW

The study titled *"A ML Approach to Predict Groundwater Levels in California Reveals Ecosystems at Risk,"* [6] conducted by Tanushree Biswas, Melissa M. Rohde, Leah S. Campbell, Ian W. Housman, Jeanette K. Howard, and Kirk R. Klausmeyer, provides valuable insights into the relationship between groundwater levels and GDEs (Groundwater-Dependent Ecosystems) in California, USA. The study's findings reveal significant challenges facing GDEs in California. Over a 35-year period, groundwater levels decreased significantly in 44 percent of GDEs while rising in just 28 percent.Notably, losses in groundwater levels have accelerated recently, especially from 2003 to 2019 [Rohde et al.].Addressing Data Gaps: The absence of comprehensive groundwater monitoring well data is a challenge. However, the authors' ML model, which utilizes satellite-based remote sensing, climate data, Landsat imagery, and field-based groundwater data, offers a promising solution. This model provides valuable insights into groundwater conditions across GDEs, enabling state and local water agencies to fill critical data gaps, assess potential impacts, and enhance sustainable groundwater management policies in California [Rohde et al.].

In another study titled *"Groundwater level prediction using ML algorithms in a drought-prone area,"* [7] conducted by Quoc Bao Pham and his team, the authors address the critical issue of groundwater level (GWL) prediction in drought-prone regions. Recognizing the importance of groundwater for various sectors, they assess7ML models' performance using historical GWL data and climate variables. Their findings reveal that Bagging-RF and Bagging-RT models outperform others, providing accurate GWL predictions during both training and testing stages. This research offers valuable insights for policymakers, enabling informed decisions for sustainable groundwater resource management in areas vulnerable to droughts.

The study titled *"Groundwater level prediction based on a combined intelligence method for the Sifangbei landslide in the Three Gorges Reservoir Area"*[8] by Taorui Zeng and colleagues introduces an innovative predictive model. This model combines the MIC ("Maximum Information Coefficient") algorithm and the LSTM ("Long Short-Term Memory") model to dynamically forecast GWL, considering the time lag associated with triggering factors. Conducted in the 3 Gorges Reservoir area of China, the research emphasizes the temporal and spatial variations in GWL influenced by factors such as accumulated reservoir and rainfall water levels. By integrating the strengths of MIC, GWO optimization, and LSTM, the proposed MIC-GWO-LSTM model demonstrates superior accuracy in GWL prediction, offering valuable insights for monitoring and early warning systems in landslide-prone regions like the Three Gorges Reservoir Area.

In the study titled "*Deep Learning-Based Forecasting of Groundwater Level Trends in India: Implications for Crop Production and Drinking Water Supply"*[9] by Pragnaditya Malakar and colleagues, the authors tackle the challenge of predicting groundwater level trends in India, a crucial factor for drinking water supply andcrop production. Utilizing data from GRACE ("Gravity Recovery and Climate Experiment"), the WaterGap model, and in situ observations from a vast network of monitoring wells, the study emphasizes the dominance of groundwater withdrawal (GWW) over groundwater recharge (GWR) in influencing groundwater storage changes. To address this, the authors employ deep learning techniques, including RNN ("recurrent neural networks"), FNN ("Feed-Forward Neural Networks"), and LSTM, to simulate as well as forecast groundwater levels. The results highlight LSTM's superior performance, projecting declining groundwater levels in various regions of India and emphasizing the potential for improved water supply and sustainable agriculture for India's vast population.

In the study titled *"Deep learning shows declining groundwater levels in Germany until 2100 due to climate change"* [8] by Andreas Wunsch and colleagues, the authors investigate climate change's direct effect on groundwater resources in Germany throughout the 21stcentury. Employing an MLtechnique on the basis ofCNNs (convolutional neural networks), they analyze data from 118 sites across Germany, considering various RCP ("Representative Concentration Pathway") scenarios (2.6, 4.5, 8.5). The study focuses solely on meteorological inputs, excluding uncertain anthropogenic factors like groundwater extractions. According to the results, groundwater levels have significantly decreased, especially under the RCP8.5 scenario.

These drops are more evident in northern and eastern Germany due to a geographic pattern. The study also shows that the yearly cycle exhibits greater unpredictability and protracted low groundwater levels near the end of the century, underlining the possible difficulties caused by climate-induced groundwater changes in the area.

In the study titled *"Groundwater level forecasting with artificial neural networks: a comparison of long short-term memory (LSTM), convolutional neural networks (CNNs), and non-linear autoregressive networks with exogenous input (NARX)"*[9] by Andreas Wunsch, Stefan Broda, and Tanja Liesch, the authors address the challenge of forecasting groundwater levels in India. They explore various factors affecting groundwater storage (GWS) changes in the region, including GRACE-derived GWS, GWR modeled using the WaterGap model, and groundwater withdrawal (GWW). The study employs a range of ANN (Artificial Neural Network)models, including FNN, RNN, and LSTM, utilizing data from a dense network of monitoring wells over a significant period (1996-2018). The results highlight the superior performance of the LSTM model, demonstrating its effectiveness in forecasting groundwater levels. The study also identifies statistically significant declining trends in groundwater levels in certain regions, emphasizing the potential implications for crop production and drinking water supply for India's large population.

III. METHODOLOGY

A. Data Collection:

Gather a comprehensive dataset that includes groundwater level measurements, meteorological data, and other relevant parameters. The data consists of attributes like Total Rainfall: The entire quantity of rainfall is shown in this column. It's a continuous variable.Natural discharge during non-monsoon season: The natural groundwater outflow during the non-monsoon season is shown in this column. Net annual groundwater availability: This column represents the net annual availability of groundwater. It's a continuous variable and is typically calculated based on factors like rainfall, discharge, and other sources.Irrigation: This column represents the amount of groundwater used for irrigation purposes.

Industrial and Domestic uses: This column represents the amount of groundwater used for domestic and industrial purposes. TotalUsage: This column represents the total usage of groundwater, including both irrigation and domestic/industrial uses. Projected demand for industrial and domestic uses up to 2025: This column represents the "projected demand for groundwater for domestic and industrial purposes up to the year 2025.

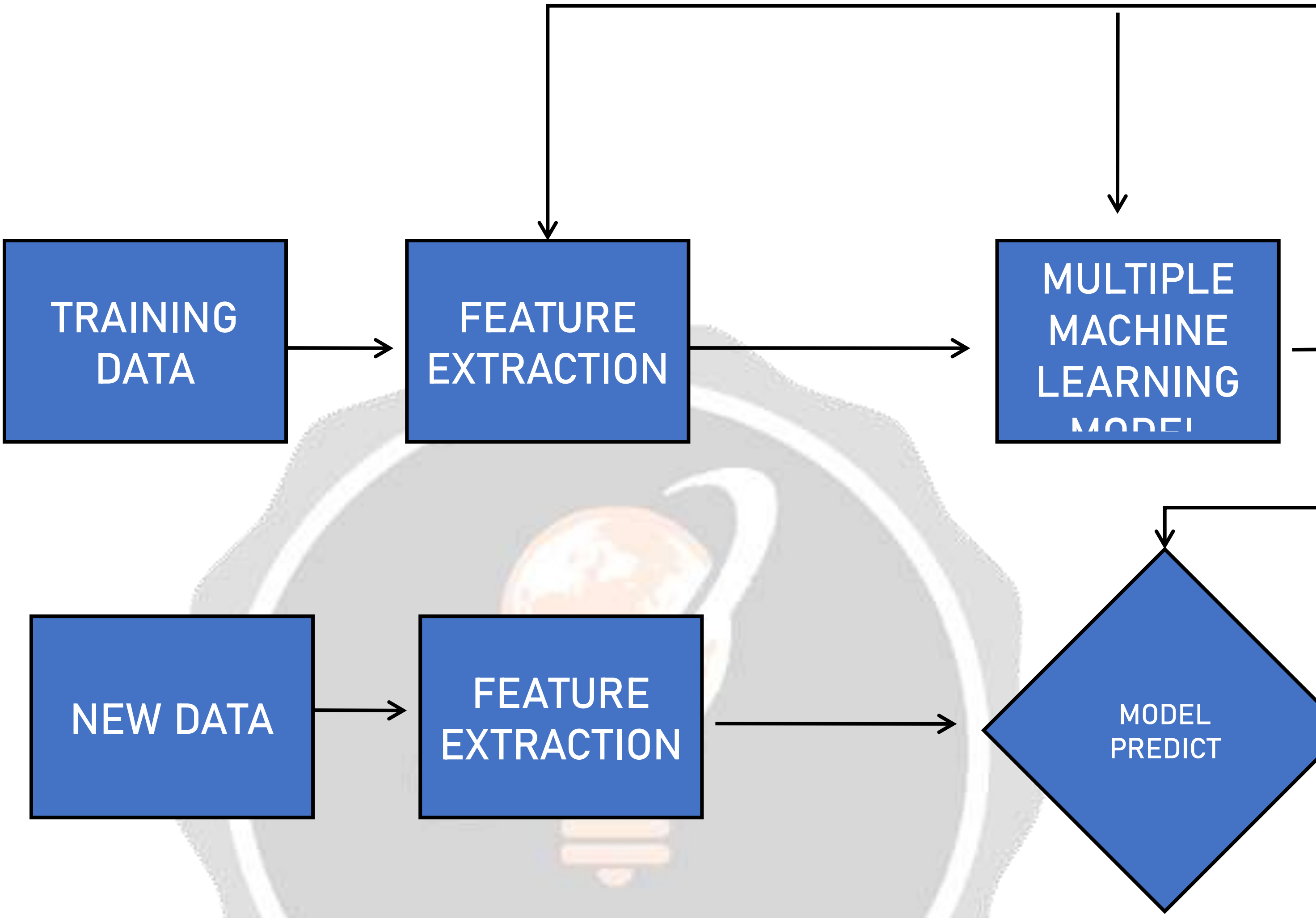


Fig:2:Block Diagram

Groundwater availability for future irrigation use: The availability" of groundwater for potential irrigation usage is shown in this column.Situation: This column represents the situation or status of groundwater availability in different regions. It's a categorical variable with values such as 'EXCESS,''SEMICRITICAL,''MODERATED,' and 'CRITICAL.'Each row in the dataset represents a different region or location, and the values in each column provide information about groundwater availability, usage, and related factors for that region.The data appears to be related to the assessment of groundwater resources in different regions and the potential issues related to groundwater availability and usage, as indicated by the 'Situation' column.

B.Data Preprocessing:

Handle missing values, remove outliers, normalize or scale features, and encode categorical variables to prepare the data for ML.LabelScikit-encoder learns class is used to translate category labels into numerical values. Each distinct category or class in the categorical column is given a special number (label).Once the categorical data is encoded into numerical form, it can be used as input features for ML models.There are training and testing sets created from the dataset.Feature scaling is applied to standardize the features.

Various data visualizations are performed, including pie charts, pair plots, and distribution plots.These visualizations help in understanding the relationships and distributions within the dataset.A count plot is created using Seaborn to visualize the distribution of values in the 'Situation' column.This plot shows the number of occurrences of each unique category in the 'Situation' column.It helps you understand the balance or imbalance of data across different situations.

Pie charts Fig: 3 are created using Matplotlib to visualize the distribution of categories in different columns. The pie chart shows the distribution of categories for another set of labels ('excess,' 'moderated,' 'semi-critical,' 'critical').Pie charts are useful for displaying the relative proportions of categories within a dataset.

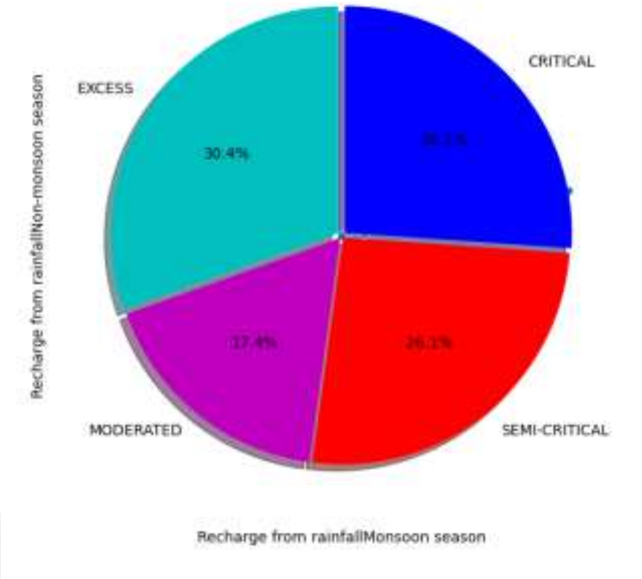


Fig: 3:Pie chart shows the distribution of categories

Line plots Fig: 4 is created using Matplotlib to visualize trends over a continuous range. In this case, line plots are used to display the trends of 'Total_Rainfall,' 'Net annual groundwater availability ‘and’ Total_Usage' over some unspecified range.Line plots are suitable for showing how values change over a continuous scale, such as time.

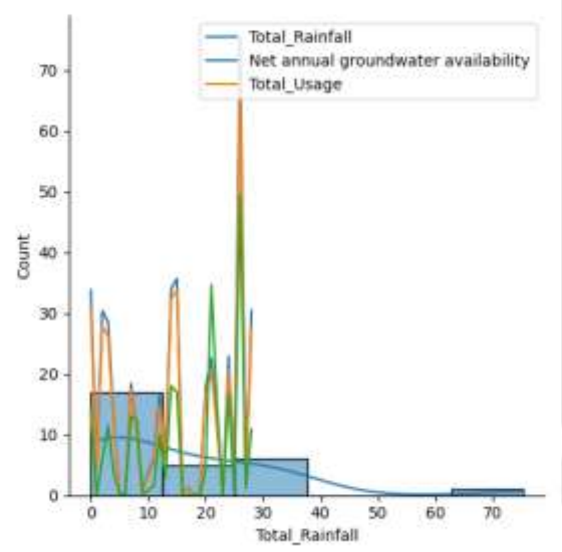


Fig: 4 Line plots for trends over a continuous range

Distribution plots (histograms) are created using Seaborn's displot function to visualize the distribution of specific numerical features, such as 'Recharge from the rainfallMonsoon season,' 'Total_Rainfall,' 'Net annual groundwater availability,' etc.These plots show the frequency distribution of values within each feature, helping you understand their distributions and potential outliers.

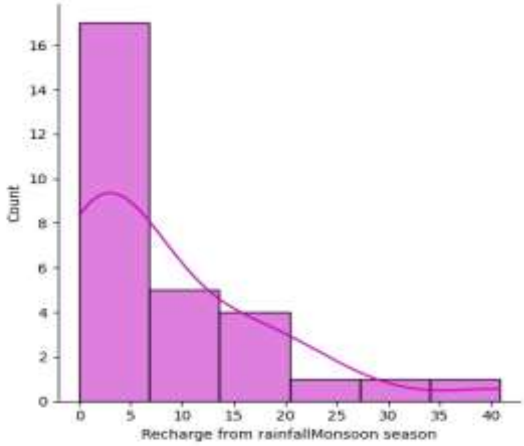


Fig: 5:Distribution plotsto visualize the distributionRecharge from rainfall Monsoon season

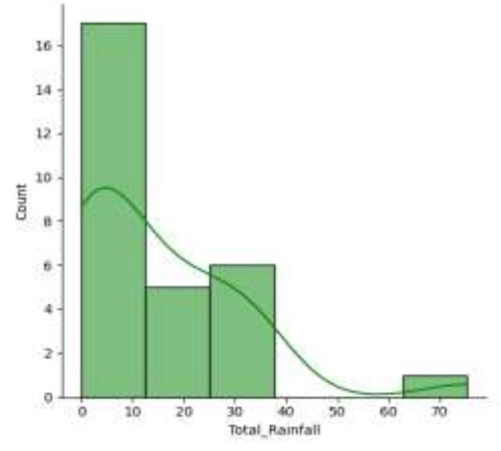


Fig: 6:Distribution plotsto visualize the distributionTotal_Rainfall

C. Feature Selection/Engineering:

Identify the most relevant variables that influence groundwater levels to improve model efficiency and interpretability.Feature selection is a crucial step in building predictive models. It involves selecting the most relevant and informative features while discarding those that do not contribute much to the model's accuracy or interpretability. The decision to remove specific columns should be guided by domain knowledge and a thorough understanding of the dataset and the modeling task at hand.

D. Model Training:

For each selected algorithm, the training process involves using the training dataset's features (independent variables) and the corresponding groundwater level measurements (dependent variable). The algorithm learns the underlying patterns and relationships in the data through an optimization process.

This process varies depending on the algorithm: Logistic regression: It models the relationship between independent variables and the probability of belonging to a specific class.SVM: Finds the hyperplane that best separates data into classes.Gradient Boosting: Ensemble method that builds trees sequentially to correct errors made by previous trees.K-Nearest Neighbors (KNN): Data points are categorized using a non-parametric approach based on the majority class of their k-nearest neighbors.

E.Model Testing:

Model testing refers to the phase where you evaluate the trained ML model's performance on a separate testing dataset that it hasn't seen during training.To determine how effectively the model generalizes to new, untested data, you utilize measures like precision, recall, F1-score, and accuracy during testing. The model's performance on several classes in the testing dataset is broken down in depth using the confusion matrix and classification report.

F.Comparative Analysis:

Comparative analysis involves comparing the performance of different ML algorithms to determine which one is the most suitable for your groundwater level prediction task.

IV. RESULT AND ANALYSIS

A crucial tool in assessing classification models is a confusion matrix. By summing the counts of the model's true positive, true negative, false positive, and false negative predictions, this table aids in the visualization of an ML model's performance.

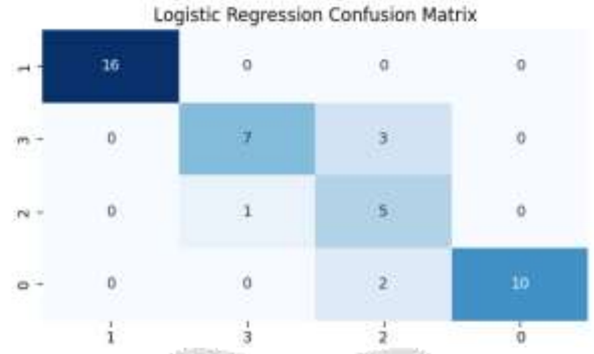


Fig: 7:confusion matrix of logistic regression

The above graph shows the confusion matrix of logistic regression and this is main metrics for performance evaluation.

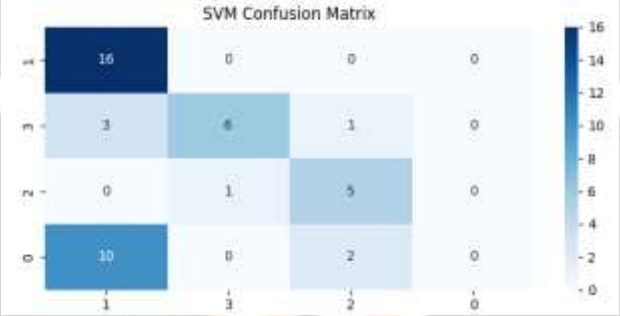


Fig: 7:confusion matrix of SVM

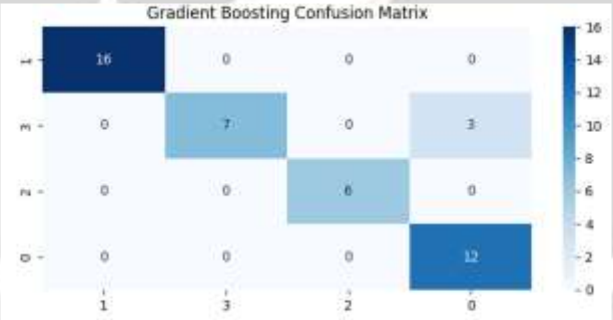


Fig: 7:confusion matrix of Gradient Boosting

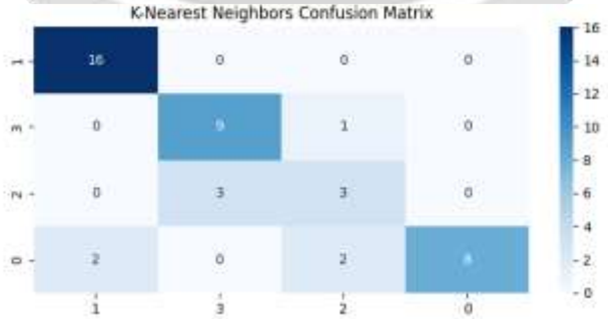


Fig: 7:confusion matrix of KNN

By the above confusion matrix comparative bar graph is made with the accuracy value using a formula.

$$Accuracy = \frac{Number\ of\ Correctly\ Predicted\ Instances}{Total\ Number\ of\ Instances} \times 100\%$$

‘Number of Correctly Predicted Instances’ is the count of instances for which the model's predictions match the actual true values. ‘Total Number of Instances’ represents the entire dataset's size, including both the correctly and incorrectly predicted instances.

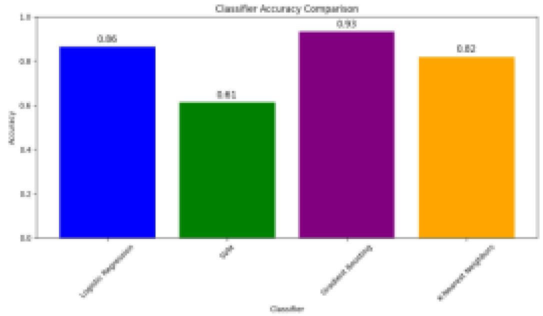
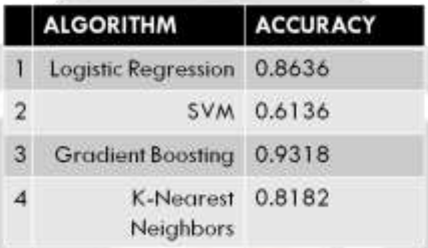


Fig: 7:Comparative analysis



	ALGORITHM	ACCURACY
1	Logistic Regression	0.8636
2	SVM	0.6136
3	Gradient Boosting	0.9318
4	K-Nearest Neighbors	0.8182

Table: 1 Accuracy comparison

A classification model's accuracy is a measure of how effectively it forecasts the appropriate class labels. Gradient Boosting outperformed the other algorithms evaluated in terms of accuracy, accurately categorizing more instances in your dataset than any other method.

V. Conclusion

In conclusion, this groundwater resource assessment and prediction project has employed rigorous and systematic methodologies to collect, preprocess, and analyze relevant data for the purpose of forecasting groundwater levels. By carefully selecting and training ML models and conducting comprehensive testing, we have identified the Gradient Boosting algorithm as the most suitable for accurate and reliable predictions. This project holds significant value for regions reliant on groundwater resources, as it provides a data-driven approach to inform water resource management decisions and make sure the sustainable usage of this vital natural asset.

References

[1] Hussein, Eslam A., Christopher Thron, Mehrdad Ghaziasgar, Antoine Bagula, and Mattia Vaccari. 2020. "Groundwater Prediction Using Machine-Learning Tools" Algorithms 13, no. 11: 300. https://doi.org/10.3390/a13110300

[2] Hai Tao; Jasni Mohamad Zain; Mohammed Majeed Hameed “Groundwater level prediction using ML models: A comprehensive review” 2022. 10.1016/j.neucom.2022.03.014

[3] P.K. Gupta, B. Yadav, B.K. Yadav Assessment of lnapl in subsurface under fluctuating groundwater table using 2d sand tank experiments J. Environ. Eng., 145 (9) (2019), p. 04019048

[4] A. Izady, K. Davary, A. Alizadeh, A.N. Ziaei, A. Alipoor, A. Joodavi, M.L. Brusseau A framework toward developing a groundwater conceptual model Arab. J. Geosci., 7 (9) (2014), pp. 3611- 3631

[5] J. Xue, Z. Huo, F. Wang, S. Kang, G. Huang Untangling the effects of shallow groundwater and deficit irrigation on irrigation water productivity in arid region: New conceptual model Sci. Total Environ., 619 (2018), pp. 1170-1182

[6] Rohde, M. M., Biswas, T., Housman, I. W., Campbell, L. S., Klausmeyer, K. R., & Howard, J. K. (Year). A ML Approach to Predict Groundwater Levels in California Reveals Ecosystems at Risk.

[7] Pham, Q.B., Kumar, M., Di Nunno, F. et al. Groundwater level prediction using ML algorithms in a drought-prone area. Neural Comput & Applic 34, 10751–10773 (2022). https://doi.org/10.1007/s00521-022-07009-7

[8] Zeng, T., Yin, K., Jiang, H. et al. Groundwater level prediction based on a combined intelligence method for the Sifangbei landslide in the Three Gorges Reservoir Area. Sci Rep 12, 11108 (2022). https://doi.org/10.1038/s41598-022-14037-9

[9] Malakar, P., Mukherjee, A., Bhanja, S. N., Sarkar, S., Saha, D., & Ray, R. K. (2021). Deep learning-based forecasting of groundwater level trends in india: implications for crop production and drinking water supply. ACS ES&amp;T Engineering, 1(6), 965-977. https://doi.org/10.1021/acsestengg.0c00238

[10] Beg AH, Islam MZ (2016) Advantages and limitations of genetic algorithms for clustering records. In: 2016 IEEE 11th Conference on Industrial Electronics and Applications (ICIEA). IEEE, pp. 2478–2483

[11] Di Nunno F, Granata F (2020) Groundwater level prediction in Apulia region (Southern Italy) using NARX neural network. Environ Res 190:110062.

[12] Fallah-Mehdipour E, Haddad OB, Mariño MA (2013) Prediction and simulation of monthly groundwater levels by genetic programming. J Hydro-Environ Res 7(4):253–260

[13] Gong Y, Zhang Y, Lan S, Wang H (2016) A comparative study of artificial neural networks, support vector machines and adaptive neuro fuzzy inference system for forecasting groundwater levels near Lake Okeechobee, Florida. Water Resour Manag 30(1):375–391

[14] Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH (2009) The WEKA data mining software: an update. ACM SIGKDD Explor Newslett 11(1):10

[15] Kasiviswanathan KS, Saravanan S, Balamurugan M, Saravanan K (2016) Genetic programming based monthly groundwater level forecast models with uncertainty quantification. Model Earth Syst Environ 2:27