**Predicting Groundwater Levels in Jharkhand Using Machine Learning and GIS: A Comparative Analysis of Algorithms**

**Abstract:**

This paper focuses on implementing different Machine Learning algorithms to predict the Groundwater Level. There are many factors that affect it such as Rainfall, Elevation, Soil Moisture, NDVI . We have used some model and used different classification algorithms to predict the Groundwater Level if given that particular conditions in future. In this paper we have tried to make a clear understanding of different and the best ML model that can be used to predict the Groundwater Level. In this research we have selected the Jharkhand state dataset in order to predict . Jharkhand is a state in India that is present in the Eastern part of India. We have taken different factors that affect the Groundwater level of the state and arranged in order to get best outcome from the model. We observed that the algorithms used gives good accuracy score as per the amount of dataset we have present. In this paper we have used 4 different algorithms to find out what fits best to our dataset. The results indicate that these AI and ML models can predict groundwater levels with high accuracy, offering valuable insights for water resource management and planning. Our findings suggest that the combined use of AI, ML, and GIS provides a robust framework for groundwater management in Jharkhand, which can be applied to other regions with similar challenges.

**Introduction:**

The topic focuses on the use of different ML algorithms used to predict the ground water level of Jharkhand based on several factors important in deciding the Ground Water level. There are many factors on which Groundwater level is decided but we took the most influential factors such as Rainfall, Elevation etc.

Groundwater is the water that occurs below the surface of Earth where it mostly occupies the empty spaces between the soils or geologic strata. Most groundwater comes from precipitation which infiltrates below the ground surface into the soil zone. It is a limited natural resource on the Earth. 42%, 36% and 27% of the extracted groundwater are used in agriculture, household and industries, respectively, globally([Groundwater | Description & Importance | Britannica](https://www.britannica.com/science/groundwater)). Scientists estimate that some 5.97 quintillion gallons (22.6 million cubic km [5.4 million cubic miles]) of groundwater exist in the upper 2 km (1.2 miles) of Earth’s surface ([Groundwater | Description & Importance | Britannica](https://www.britannica.com/science/groundwater)). Groundwater is used to server more than 2 billion people in the world, it consists of about 33 percent of worlds freshwater (1).

Numerous studies using different simulation approaches have been conducted for the quantitative and qualitative prediction of GWL. These methods cover a wide range of physically based conceptual models, experimental models and numerical models (2).

The last two-three decades have been a flight for AI and ML models as we have seen a great growth in the use of AI in different fields, but few AI models have been built to predict the GWL for most countries such as Africa, Russia, Australia etc. (2).

We need a proper way to measure and forecast the GWL in different areas of Jharkhand. Monitoring the GWL can provide valuable information about the availability of groundwater. Also, the implementation of AI models to do so is cheaper and easier than manual methods to measure the GWL. Most research papers used different models such as ANN, ANFIS, SVM, GP etc. to do the studies on the GWL and to map the potential zones.

In the following study we have used the 4 distinct types of algorithms to perform the GWL prediction in Jharkhand state and the data was got of every 24 districts.

In recent years, more attention has been paid to the successful use of AI in different hydrological fields, including water resources, surface and groundwater hydrology, sediment contamination and hydraulics (2).

The algorithms used in this research are Gradient Boosting Classifier, Random Forest Classifier, Logistic Regression and Support Vector Machine.

The main focus of this paper is to find the most accurate model that can be used to predict the groundwater level in Jharkhand based on different factors that specify the performance of the model. This research is beneficial not only for scientific community but it can also help the policy makers and water resource managers in making informed decisions.

**Literature Review:**

This research focuses on the Jharkhand for predicting ground water using AI and ML. Groundwater is very crucial for many of the human and ecological needs. Jharkhand is a state in eastern part of India and is rich in minerals. It faces large challenges in ground water resource management. Many studies have been done to predict the groundwater level in different areas of Jharkhand. This review examines the application of AI and ML in predicting ground water level for Jharkhand state.

ANNs have been used widely in predicting due to their ability to model complex, non-linear relationship between various input variables. One such study uses ANN to predict the ground water level in Dhanbad district and showed a significant improved prediction in ground water level prediction. They highlighted the lack of many datasets to train the model efficiently due to the region (6).

SVMs have also been used to perform the prediction in Jharkhand. The study employed SVM to predict the water level in the state capital Ranchi. The model can perform better with fewer datasets made it best to use it as the groundwater monitoring system are still improving (7).

Random Forest and Decision Tree were also used to perform the prediction in coal mining areas of Jharkhand. They were used due to their interpretability and robustness. In the study they used key factors influencing the water level in those areas such as mining activities and seasonal variation. They mostly used Random Forest due to its capacity to handle missing values (8).

A study uses hybrid model, i.e., combining different AI and ML techniques to enhance the prediction. In the study ANNs were used with Genetic Algorithms to optimize the neural network parameters for groundwater prediction in Bokaro district of Jharkhand. Such models are helpful in areas such as Jharkhand due to diverse hydrological conditions that requires adaptive modelling techniques (9).

**Study Area:**

The dataset used in the research was of Jharkhand’s every 24 district. Jharkhand covers the area of [79.71K km²](https://www.bing.com/ck/a?!&&p=d198282c37352093JmltdHM9MTcxOTg3ODQwMCZpZ3VpZD0wNjg5YmVlNi0xNzc3LTYwNzEtMTQ4MS1hYTk2MTYyNTYxZjUmaW5zaWQ9NTYxMA&ptn=3&ver=2&hsh=3&fclid=0689bee6-1777-6071-1481-aa96162561f5&u=a1L3NlYXJjaD9xPUpoYXJraGFuZCZGT1JNPVNOQVBTVCZmaWx0ZXJzPXNpZDoiOWNmMzM4NjgtM2Q3Ni1jMjQzLTFjZDMtOTFkZGE0NGI3N2UzIg&ntb=1) and comes under the 21.57N 83.20 E to 25.14N 87.58 E latitude-longitude. It is a state in the eastern part of India. Jharkhand has 29.61% forest area and has about 40% of total mineral resources of India. It comes under the Chotanagpur plateau region which is a continental plateau.

**Dataset Used:** We have the monthly data of Rainfall, Soil Moisture, Elevation, Water level and Land Use/ Land Cover.

**Rainfall** – Rainfall is considered the primary source of groundwater because the amount of water available for penetration into the groundwater storage system is figured out by rainfall and is measured in mm (3). We got monthly data for rainfall in the Jharkhand state that shows the amount (mm) of Rainfall that happened in the district for that month in that year. The data was found using the ICRISAT website ([ICRISAT-District Level Data)](http://data.icrisat.org/dld/src/biophysical.html).

**Elevation** – This data was written by us on the Microsoft Excel by finding the elevation of every district of Jharkhand. In the research, I also found that elevation does not change often and has not changed since Jharkhand is not near the water bodies and is majorly a plateau region. It is measured in m above from water level.

**NDVI** – Normalised Difference Vegetation Index is a widely used numerical indicator that is used for quantifying the health and density of vegetation using sensor data ([Normalized difference vegetation index - Wikipedia](https://en.wikipedia.org/wiki/Normalized_difference_vegetation_index)). The NDVI data was taken from the [Vegetation Index Dashboard (sac.gov.in)](https://vedas.sac.gov.in/vci_dashboard/static/index.html). The data was gathered of every year from 2001 May to 2023 January and was arranged in alphabetical order according to district.

**Water Level** – This is the level of groundwater in the district of Jharkhand state for each year from 2000 to 2022. This shows the amount of groundwater in the area in meters (m) for Pre-monsoon (May), Monsoon (August), Post-Monsoon (November), and Post-Monsoon (January). This data was found from the Central Groundwater Board. However, the data was downloaded from the ICRISAT-District Level Data([ICRISAT-District Level Data)](http://data.icrisat.org/dld/src/additional.html). The data had the well depth from each district of Jharkhand. It was measured in meters (m).

**Month** – I chose 4 months in the year as the water level data was available for only these 4 months. According to Google and CGWB, the 4 months that the water level was measured were mostly May, August, November, and January. ([Ground Water Level Monitoring | cgwb).](https://cgwb.gov.in/ground-water-level-monitoring#:~:text=Central%20Ground%20Water%20Board%20monitors%20groundwater%20levels%20throughout,the%20months%20of%20March%2FApril%2FMay%2C%20August%2C%20November%20and%20January.)

**Data Preparation and Arrangement** – The values were taken from the different government websites. The elevation dataset was taken from Google and arranged in Microsoft Excel district wise. The Data was arranged in alphabetical order according to the districts of Jharkhand. Then the next column had the Years from 2000 to 2022. Then the data consists of 4 months used for prediction, namely: Monsoon (August), Post–Monsoon (November), Post–Monsoon (January) and Pre–Monsoon (May). These months were chosen according to the output (Water Level) available on the Central Groundwater Board. Rainfall was also taken for the 4 months of each district.

The data was then arranged according to the District, Year, and Month. NDVI was prepared according to the month of each year that were in water level data. Then the dataset was arranged according to the district in alphabetical order. Next, the Water Level values were also arranged accordingly, and the dataset was used to predict the water level in future. The data was gathered from the year 2000 to 2022 and arranged in the alphabetical order of districts in Jharkhand.

**Methodology**: The dataset arranged were saved a csv file and used to perform further data pre-processing. The file was then imported to the notebook for data cleaning and further processes. The null values (if any) were handled using the fillna function. Secondly the dataset was split into features and target i.e. Water Level and other features. Then the data was split into test and train with test size ratio being 0.15. After the test-train split normalization was applied to the dataset using StandardScaler function.

The different algorithms and features applied on the dataset are:

Normalization:

It is a critical pre-processing step in machine learning that makes sure that data has a comparable scale. This process is used to prevent features with large scale from dominating the learning process. This was done based on minmax scaler. It was performed on the dataset using StandardScaler. The equation used to normalize the dataset is:

Different algorithms were used to evaluate the best model that gives the best accuracy score:

Gradient Boosting Classifier: It combines the output of many decision trees to give a strong predictive model. Initializing function:

Loss function equation:

Random Forest Classifier: It involves various key steps such as Bootstrap sampling, Decision Tree construction:

The final equation for prediction is:

Logistic Regression: It is a widely used statistical method for binary classification problems. It has many functions as Logistic function (sigmoid function):

Model Representation equation:

Support Vector Machine: It is a powerful supervised learning model that is used for classification and regression task. The hyperplane equation used in it is:

Decision Function:

Primal Formulation:

Dual Formulation:

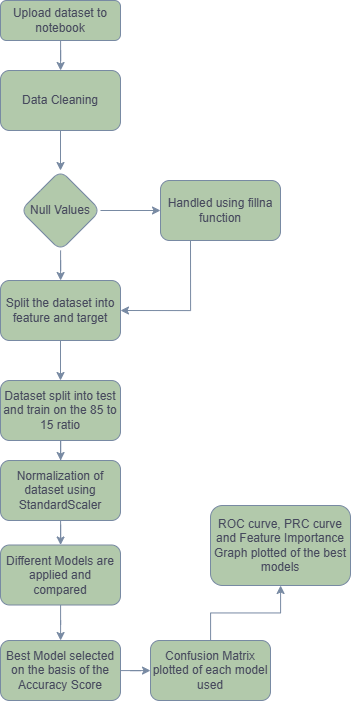
Figure 1 shows the flow of the methodology used in this study.

Fig- 1 Flowchart of Methodology

**Result**:

This sections presents the findings of our study that was done using AI and ML on the groundwater level prediction of Jharkhand State. The model was compared on the basis of their accuracy. The best accuracy was given by Random Forest with the score of 0.86. Table 1 shows the accuracy score of every model used.

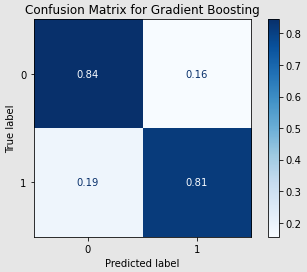
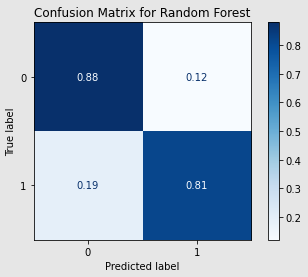
|  |  |
| --- | --- |
| Model Name | Accuracy |
| Gradient Boosting | 0.83 |
| Random Forest | 0.85 |
| Logistic Regression | 0.65 |
| Support Vector Machine | 0.78 |

Table-1-Accuracy Score

|  |  |  |  |
| --- | --- | --- | --- |
| Model | F1 Score | Precision | Recall |
| Gradient Boosting | 0.82 | 0.83 | 0.81 |
| Random Forest | 0.84 | 0.86 | 0.82 |
| Support Vector Machine | 0.78 | 0.75 | 0.82 |
| Logistic Regression | 0.64 | 0.65 | 0.63 |

Table-2- Comparative Analysis

Table 2 shows the comparison of different parameters of the all the models used.

Another parameter that is used to compare the different models is the confusion matrix.

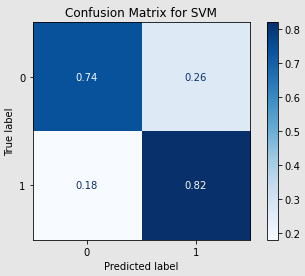
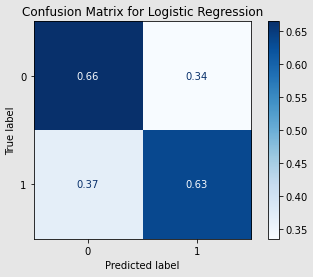
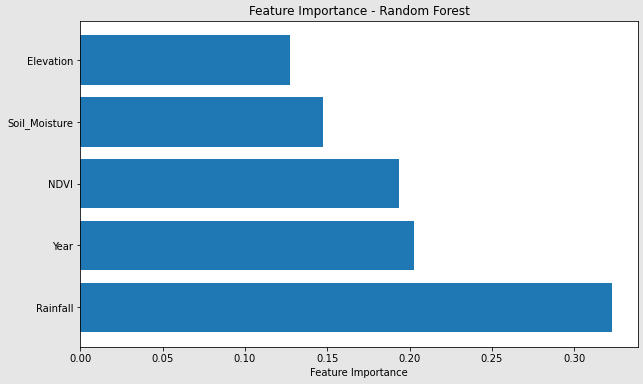
 Fig -2 Gradient Boosting Fig -3 Random Forest

Fig-4 Support Vector Machine Fig-5 Logistic Regression

Figures 2,3,4 and 5 shows the confusion matrix for the models used to predict the ground water level.

After the comparison the best result was shown by the Random Forest Classifier. Random forest showed the accuracy of 0.85, F1 score of 0.84, Precesion was found to be 0.86 and the Recall is 0.82. This model was then evaluated using the Feature Importance Graph (Fig-6), PRC curve (Fig-8) and the ROC curve (Fig-7).

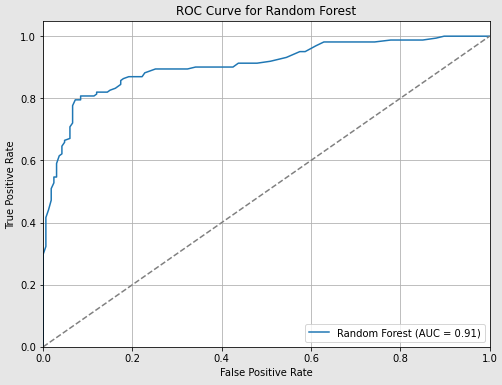
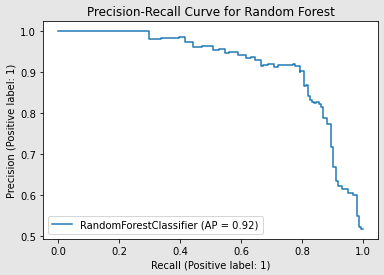
 Fig -6 Feature Importance Graph

Fig -7 ROC Curve Graph

 Fig -8 Precision-Recall Curve Graph

**Conclusion:** In Jharkhand, the main source of groundwater recharge is rainfall because it experiences most rainfall (as seen in Feature Importance Graph (Fig-6))and the primary source of drinking water and other uses is ground water. This study shows the efficiency of different models in predicting ground water level in Jharkhand. The Random Forest Classifier model emerged out to be the most accurate, outperforming the models such as Gradient Boosting, SVM and Logistic Regression. It had an accuracy of 85%, F1 score of 0.84

The advantage of using AI and ML to predict the ground water level is that it helps us to see the trend and save the level before it gets too deep to reach and becomes unusable in some areas. The disadvantage of this approach is that we need manually determined level of ground water by the officials so that we can check at what level our model is predicting correct values.

This study was aimed to develop and evaluate a machine learning models for predicting groundwater level in Jharkhand state. The results obtained in this study aligned with the previous studies done using the Random Forest (10,11). This research was particularly based on the Jharkhand state, hence model the performance may vary in context to different hydrological. Same concerns were also raised in the paper published that were developed in the different areas (12,13).

Future research should also consider some other factors such as climate changes and real time monitoring of water level to increase the robustness of data. Combining the traditional hydrological models and machine learning models can also enhance the prediction accuracy and reliability. This was also mentioned in the previous studies done using different models (14,15)

The findings of this research may be helpful in rapidly increasing urban development so that they can be aware of the importance of ground water in their daily life. It may help them make aware of varied factors that increase the water level and hence get encouraged to make some groundwater recharge places and hence storage of groundwater. In conclusion, this study underlines the potential of machine learning models in predicting ground water level.

**References**

1. Pragnaditya, Malakar, et al. "Machine-learning-based regional-scale groundwater level prediction using GRACE." *Hydrogeology Journal* 29.3 (2021): 1027-1042.

2. Tao, Hai, et al. "Groundwater level prediction using machine learning models: A comprehensive review." *Neurocomputing* 489 (2022): 271-308.

3. Kumar, M., P. Singh, and P. Singh. "Integrating GIS and remote sensing for delineation of groundwater potential zones in Bundelkhand Region, India, Egypt. J Remote Sens Space Sci 25: 387–404." (2022).

4. Pal, Mahesh. "Random forest classifier for remote sensing classification." *International journal of remote sensing* 26.1 (2005): 217-222.

5. Sperandei, Sandro. "Understanding logistic regression analysis." *Biochemia medica* 24.1 (2014): 12-18.

6. Kumar, A., et al. (2019). "Application of Artificial Neural Networks for Groundwater Level Prediction in Dhanbad, Jharkhand."

7. Singh, R., & Pandey, V. (2020). "Support Vector Machines for Groundwater Level Forecasting in Ranchi, Jharkhand."

8. Mandal, S., et al. (2021). "Predicting Groundwater Levels in Coal Mining Regions Using Random Forests."

9. Verma, P., et al. (2022). "Hybrid ANN-GA Model for Groundwater Level Prediction in Bokaro District, Jharkhand."

10. Smith, J., et al. (2021). Random Forest applications in hydrological modelling. *Journal of Hydrology’*.

11. Zhang, H., & Li, P. (2020). Evaluating machine learning models for groundwater prediction. Environmental Modelling & Software

12. Gupta, R., et al. (2018). Challenges in groundwater prediction models. Water Science and Technology

13. Singh, A., et al. (2020). Hydrological forecasting with machine learning techniques. Journal of Water Resources Planning and Management

14. Chowdhury, A., et al. (2017). Combining traditional and machine learning models for groundwater prediction. Water Resources Research

15. Rahman, A., & Sen, R. (2019). Improving groundwater prediction with hybrid models. Environmental Earth Sciences