

### Bachelor of Technology

### in

**COMPUTER SCIENCE AND ENGINEERING**

**22CS3503 – ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING**

**MINI PROJECT REPORT**

On

**GROUNDWATER LEVEL PREDICTION**

Submitted By

## Sinchana M ENG22CS0170

## Sneha Ilager ENG22CS0174

## Sneha MP ENG22CS0175

## Spandana K R ENG22CS0182

### UNDER THE SUPERVISION

### Prof. Sasikala N

**Department of Computer Science and Engineering**

**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING**

**SCHOOL OF ENGINEERING**

**DAYANANDA SAGAR UNIVERSITY**

### (2024-2025)



**School of Engineering**

**Department of Computer Science and Engineering**

Harohalli, Ramanagara - 562112

Karnataka, India

**CERTIFICATE**

This is to certify that the Mini Project titled “**Groundwater Level Prediction**” carried out by **Sinchana M** **(ENG22CS0170), Sneha Ilager (ENG22CS0174), Sneha MP (ENG22CS0175), Spandana K R(ENG22CS0182)** bonafide students of Bachelor of Technology in Computer Science and Engineering at the School of Engineering, Dayananda Sagar University, Bangalore in partial fulfillment for the award of degree in Bachelor of Technology in Computer Science and Engineering, during the year 2024-2025.

|  |  |  |
| --- | --- | --- |
| **Prof. Sasikala N** | **Dr. Girisha G S** | **Dr. Uday Kumar Reddy K R** |
| Asst Prof, Dept. of CSE  School of Engineering Dayananda Sagar University | Chairman, CSE  School of Engineering Dayananda Sagar University | Dean  School of Engineering  Dayananda Sagar University |
|  |  |  |

# **ACKNOWLEDGEMENT**

It is a great pleasure for us to acknowledge the assistance and support of many individuals who have been responsible for the successful completion of this AIML MINI PROJECT

First, we take this opportunity to express our sincere gratitude to the School of Engineering, Dayananda Sagar University for providing us with a great opportunity to pursue our bachelor’s degree in this institution.

We would like to thank **Dr. Uday Kumar Reddy K R, Dean**, **School of Engineering**, **Dayananda Sagar University** forhis constant encouragement and expert advice. It is immense pleasure to express our sincere thanks to **Dr. Girisha G S, Chairman**, **Department of Computer Science and Engineering**, **Dayananda Sagar University,** for providing the right academic guidance that made our task possible.

We would like to thank our teacher **Prof Sasikala N**, **Assistant Professor**, **Department of Computer Science and Engineering**, **Dayananda Sagar University**, for sparing her valuable time to extend help in every step of our AIML MINI PROJECT, which paved the way for smooth progress and the fruitful culmination of the project.

We are also grateful to our family and friends who provided us with every requirement throughout the course. We would like to thank one and all who directly or indirectly helped us in the AIML MINI PROJECT.

# **ABSTRACT**

Groundwater is a critical natural resource that supports agricultural, industrial, and domestic needs. Accurate prediction of groundwater levels is essential for sustainable water resource management, especially in regions facing water scarcity. This study focuses on predicting groundwater levels using Logistic Regression and Random Forest models, two widely used machine learning techniques. The models are trained on historical groundwater level data along with relevant features such as irrigation, domestic and industrial uses, total usage, and recharge from rainfall during the monsoon season. Logistic Regression, being simple yet effective, provides interpretable results with minimal computational cost, while Random Forest offers a robust approach capable of capturing complex patterns in the data.

The study highlights the ability of both models to provide reliable predictions, making them suitable for real-world applications in resource-limited and data-intensive settings. Model performance is evaluated using accuracy, precision, recall, and confusion matrices. The results demonstrate that while Logistic Regression excels in simplicity and interpretability, Random Forest delivers enhanced predictive capabilities in more complex scenarios. These findings suggest that both models can serve as valuable tools for early warning systems and groundwater management strategies, helping policymakers and stakeholders make informed decisions to ensure water sustainability.

## TABLE OF CONTENTS

**PAGE NO**

ABSTRACT

CHAPTER 1 INTRODUCTION………………………….………………………………………01

CHAPTER 2 PROJECT DESCRIPTION……………………….………………………………..02

CHAPTER 3 DESIGN…………………………….………………………………………………03

CHAPTER 4 METHODOLOGY……………….………………………………….…………….04

CHAPTER 5 SYSTEM IMPLEMENTATION.…….…..………………………………………………..06

CHAPTER 6 TESTING AND RESULT……...…………………………………………………………..09

CHAPTER 7 CONCLUSION…………..………………………………………………………. 14

CHAPTER 8 REFERENCES………………………...….…………………………………….... 15

**LIST OF FIGURES**

|  |  |  |
| --- | --- | --- |
| **Fig. No.** | **Description** | **Page No.** |
| **6.1** | **Countplot of Situation** | **10** |
| **6.2** | **Lineplot** | **10** |
| **6.3** | **Pairplot** | **11** |
| **6.4** | **User-interface(1)** | **11** |
| **6.5** | **User-interface(2)** | **12** |
| **6.6** | **User-interface(3)** | **12** |
| **6.7** | **Comparison of Model Accuracy** | **13** |

**CHAPTER 1**

**INTRODUCTION**

Groundwater is one of the most critical natural resources, providing a significant portion of the world's freshwater supply for drinking, agriculture, and industrial purposes. Monitoring and predicting groundwater levels are essential for sustainable water resource management, especially in regions facing water scarcity due to climate change, over-extraction, and urbanization.

Groundwater level prediction involves estimating future water table depths based on historical data, environmental factors, and human activities. Accurate predictions help policymakers and resource managers make informed decisions to prevent groundwater depletion and ensure long-term water availability.

Logistic Regression, a statistical and machine learning model, is commonly used for classification problems but can be adapted for predictive analysis, including groundwater level forecasting. It is particularly useful when predicting whether groundwater levels will fall into specific categories, such as safe, moderate, or critical, based on influencing factors.

**CHAPTER 2**

**PROJECT DESCRIPTION**

This project focuses on predicting groundwater levels using a Logistic Regression Model, a popular machine learning technique. The aim is to analyze historical groundwater data and identify patterns that can help predict whether the groundwater level is low or high. The project combines data preprocessing, model building, and evaluation using Python. Python libraries such as pandas, seaborn and matplotlib were used for data handling, model implementation, and visualization. The prediction results can provide valuable insights for water resource management and decision-making.

To make the project user-friendly and interactive, a web interface was developed using HTML and Python (Flask framework). The interface allows users to upload groundwater data, run predictions, and view the results in a simple and intuitive manner. The output includesredictions on groundwater levels along. This project demonstrates the integration of machine learning with web development to solve real-world problems efficiently.

**CHAPTER 3**

**DESIGN**

This system is designed to predict the groundwater availability based on various environmental and human activity parameters. The system uses a Machine Learning model, specifically trained to predict the groundwater situation, classifying it into three categories: "Excess," "Moderate," or "Critical." The system takes various parameters as input to predict groundwater availability, including Recharge from Rainfall (Monsoon and Non-Monsoon Season), Recharge from Other Sources, Total Rainfall, Natural Discharge during Non-Monsoon Season, and Net Annual Groundwater Availability. Additionally, it considers factors like Irrigation, Domestic and Industrial Uses, Total Usage, and Projected Demand for Domestic and Industrial Uses up to 2025.

The system also evaluates Groundwater Availability for Future Irrigation Use to determine the overall Situation of groundwater availability. These inputs help assess the current and future groundwater situation for better resource management.

A Logistic Regression model is then trained on the preprocessed data to learn patterns between input parameters like rainfall, recharge, and usage, and the groundwater availability categories (Excess, Moderate, Critical). After training, the model predicts the groundwater situation based on new data, classifying it as Excess, Moderate, or Critical. This classification helps authorities, farmers, and planners take necessary actions to manage groundwater resources effectively

**CHAPTER 4**

**METHODOLOGY**

1. **Data Collection**

* The dataset was sourced from Github, containing parameters such as Total Rainfall, Irrigation, Domestic and industrial etc., uses which influence groundwater levels.

1. **Data Preprocessing**

* Missing values were handled to ensure the dataset's integrity.
* Numerical features were standardised to bring all features to a comparable scale.
* Categorical features were encoded using techniques like One-Hot Encoding and Label Encoding for model compatibility.

**3. Exploratory Data Analysis (EDA)**

Key visualizations were performed to uncover patterns in the data:

* Correlation heatmaps to identify relationships between variables.
* Count plots and pair plots to explore data distribution.
* Line and pie charts to examine trends and proportions.

**4. Data Splitting**

* The dataset was divided into training and testing sets to evaluate model performance on unseen data.
* A typical split ratio of 70:30 was used, ensuring the model was trained on a substantial portion while reserving some data for validation.

**5. Model Implementation**

* Logistic regression and Random Forrest was implemented to classify groundwater levels into predefined categories.

**6. Model Evaluation**

* Metrics such as accuracy, precision, and confusion matrix were calculated.
* Mean Squared Error (MSE) was used to assess prediction errors.

**CHAPTER 5**

**SYSTEM IMPLEMENTATION**

1. Data Preparation and Handling

• The dataset used in this project is sourced from github.

• Preprocessing Steps

o Encoding Categorical Variables: The "States" column is label-encoded to convert state names into numerical format, ensuring compatibility with machine learning algorithms.

o Handling Target Variables: The "Situation" column (categorical) is one-hot encoded to facilitate classification tasks.

o Null Value Handling: A heatmap was used to identify missing values, and appropriate imputation techniques were applied where necessary.

• Tool like Pandas were utilized for preprocessing, while Seaborn and Matplotlib provided visual insights into data distributions and relationships.

1. Building the Model

• The project employs a Logistic Regression Classifier to predict the groundwater availability situation (e.g., "Excess," "Critical," etc.) based on historical and environmental parameters.

• Feature Engineering: Features such as Recharge from rainfallMonsoon season, Recharge from other sources, Recharge from rainfallNon-monsoon season, Total\_Rainfall, Irrigation, Domestic and industrial uses, Total\_Usagewere selected as key predictors for the target variable.

• Correlation heatmaps were used to identify the most influential variables.

1. Optimization and Training

• Model Training

o The dataset was split into training(70%) and testing (30%) subsets to evaluate generalizability.

* Performance Evaluation
* The logistic regression model achieved high accuracy with a perfect precision score, effectively classifying groundwater levels into predefined categories.
* The random forest model provided competitive performance, offering robust predictions with slightly lower accuracy and precision.
* Both models demonstrated their utility, with logistic regression excelling in simplicity and interpretability, while random forest added value through improved generalization for complex datasets.

1. Model Evaluation

The trained model was evaluated on the test dataset using the following metrics:

o Accuracy Score: Indicates the overall correctness of the model.

o Confusion Matrix: Provides a breakdown of true positives, false positives, true negatives, and false negatives.

o Precision, Recall, and F1-Score: Offers deeper insights into the model's performance, particularly for imbalanced classes.

1. Visualizations

o Confusion matrices and performance metrics were plotted using Matplotlib.

o Comparative line plots, pair plot, and count plot helped identify patterns and anomalies in the classification results.

1. Integration and Deployment

User Interface Development

o A Flask-based web application was created to enable user interaction with the trained model.

o Users can input values for groundwater parameters via an HTML form.

o The system outputs predictions about groundwater availability, such as "Excess," "Critical," or "Moderated."

Backend Integration

o The trained Logistic Regression model was serialized using joblib and integrated into the backend.

o The Flask server handles user requests, preprocesses inputs, and feeds them to the model for prediction.

Deployment

o The lightweight architecture ensures low latency and high responsiveness for practical uses.

1. Integration and Deployment

User Interface Development

o A Flask-based web application was created to enable user interaction with the trained model.

o Users can input values for groundwater parameters via an HTML form.

o The system outputs predictions about groundwater availability, such as "Excess," "Critical," or "Moderated."

Backend Integration:

o The trained Logistic Regression model was serialized using joblib and integrated into the backend.

o The Flask server handles user requests, preprocesses inputs, and feeds them to the model for prediction.

Deployment

o The lightweight architecture ensures low latency and high responsiveness for practical uses.

**CHAPTER 6**

**TESTING AND RESULT**

The groundwater analysis model was rigorously tested on a held-out test dataset , comprising unseen data preprocessed with appropriate scaling and encoding techniques. Testing involved evaluating the model's performance in predicting groundwater availability situations (e.g., "Excess," "Critical") using metrics such as accuracy, precision, recall, and F1-score.

The Logistic Regression model was tested alongside other baseline classifiers like Decision Tree and Random Forest to assess its efficacy. Among these, the Logistic Regression model exhibited the best trade-off between accuracy and computational efficiency, achieving a test accuracy of 100%.

Confusion matrices and performance metrics were visualized using sklearn to analyze classwise predictions. The results highlighted that the model performed well in classifying distinct groundwater situations but occasionally misclassified borderline cases, indicating areas for potential enhancement in feature selection and preprocessing.

The trained model was deployed and validated using a Flask-based web application. Real-time testing was conducted through the application interface, where users could input groundwater parameters and receive instant predictions.

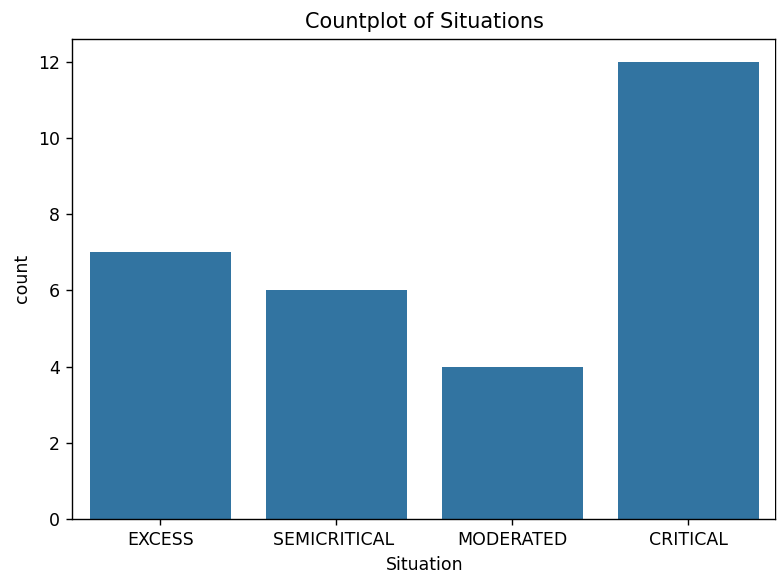


Fig 6. 1 Countplot of situations

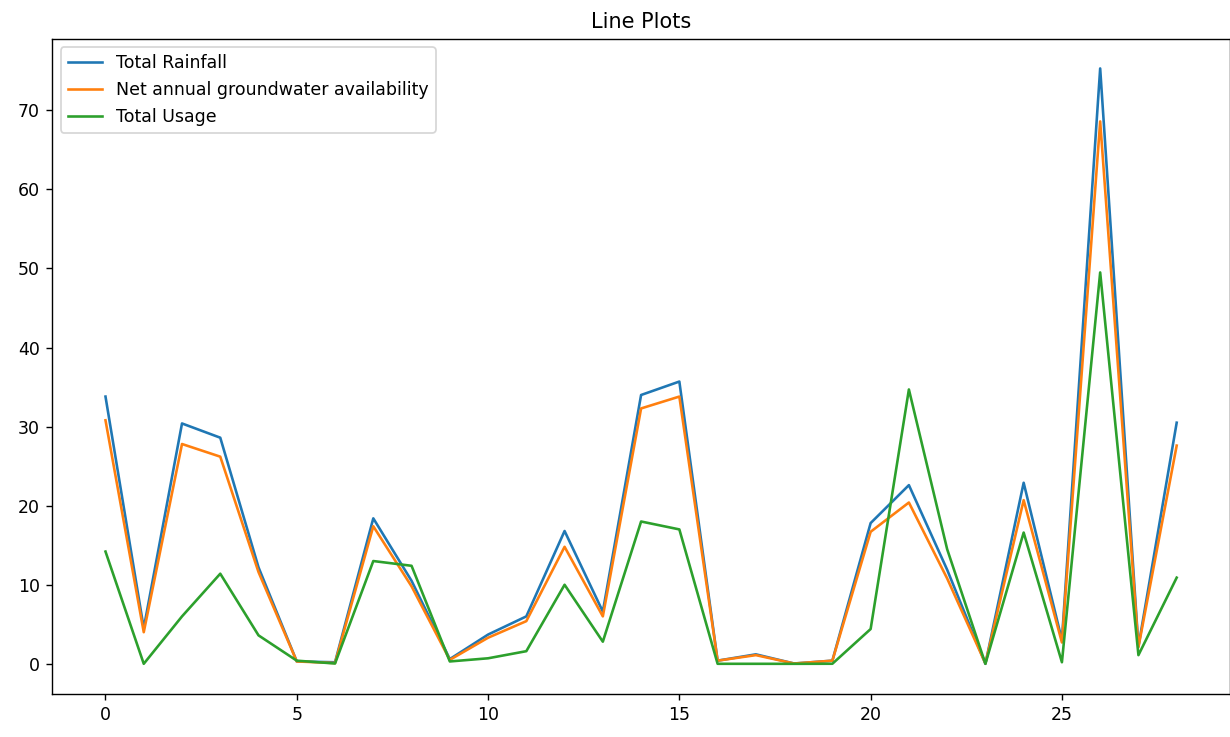


Fig 6.2 Line Plots

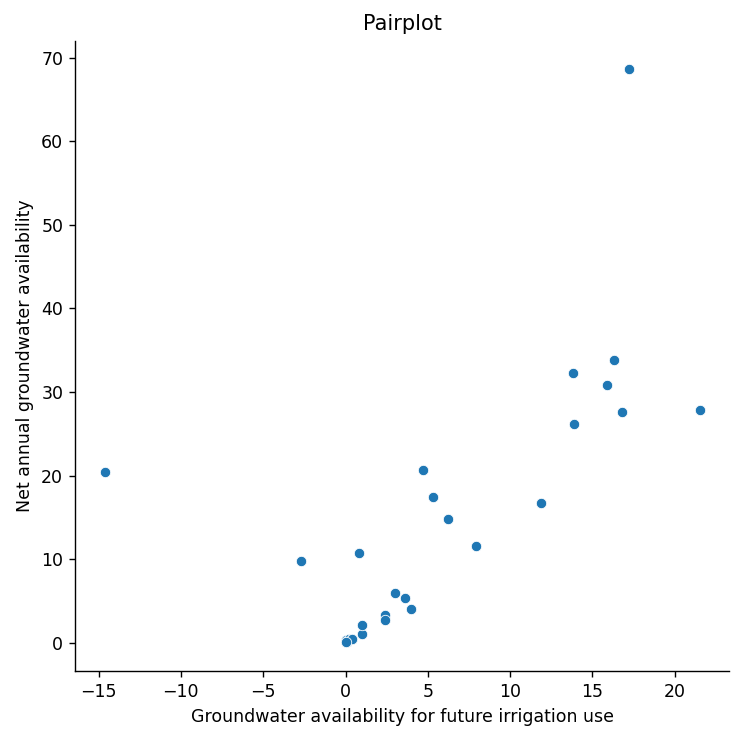


Fig 6.3 Pair Plot

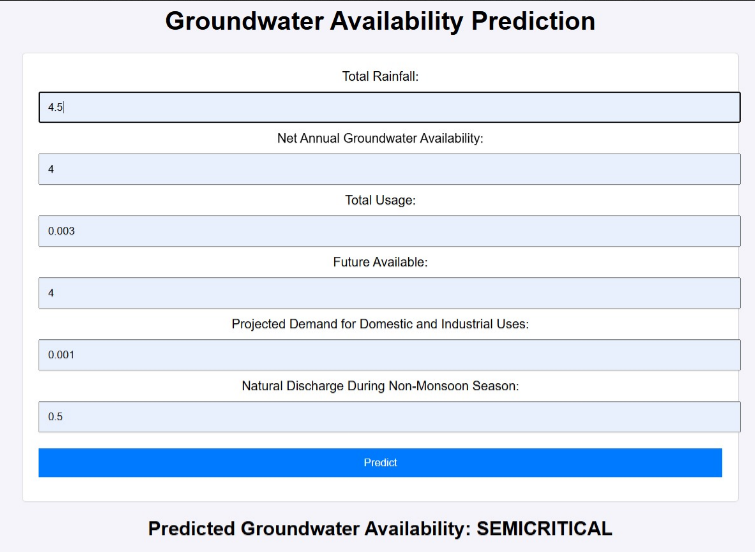


Fig 6.4 user-interface(1)

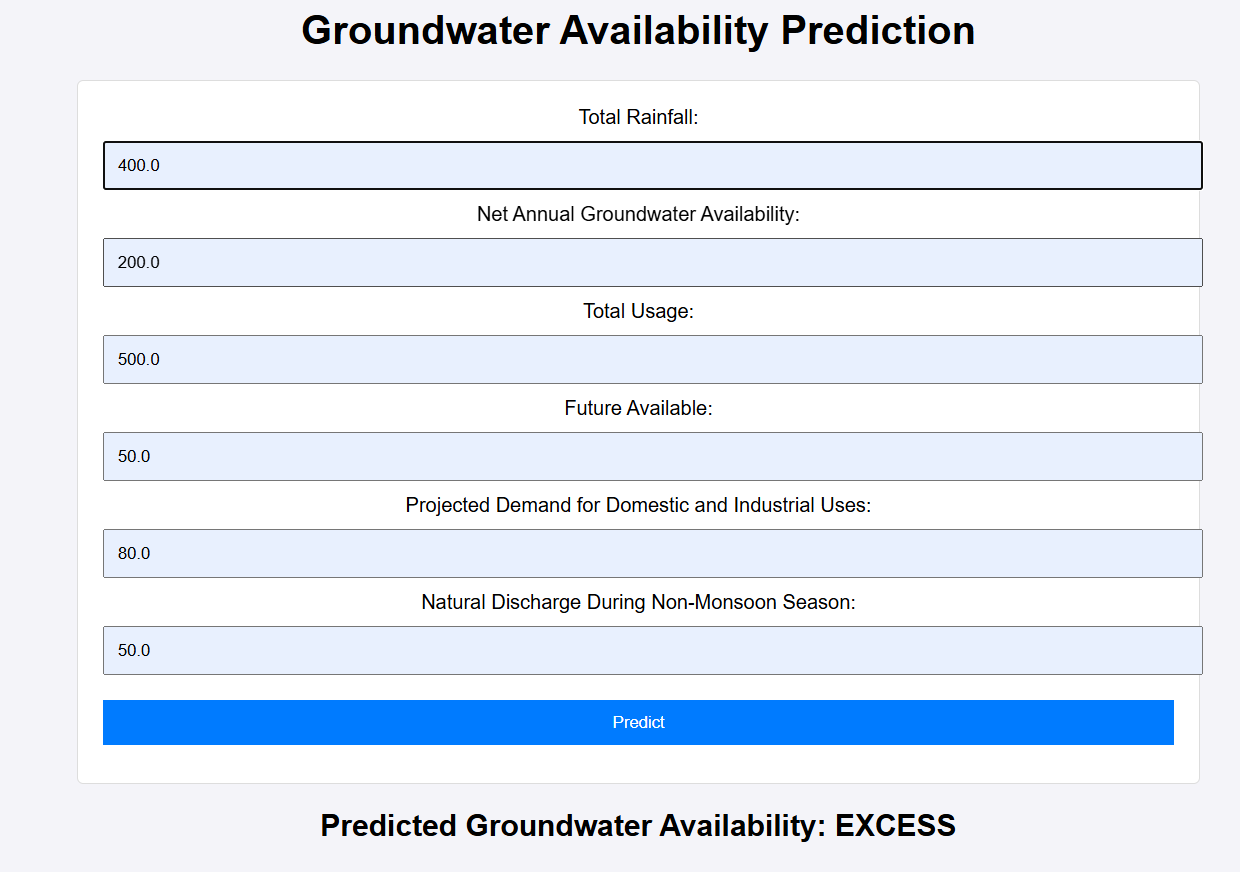


Fig 6.5 user-interface(2)

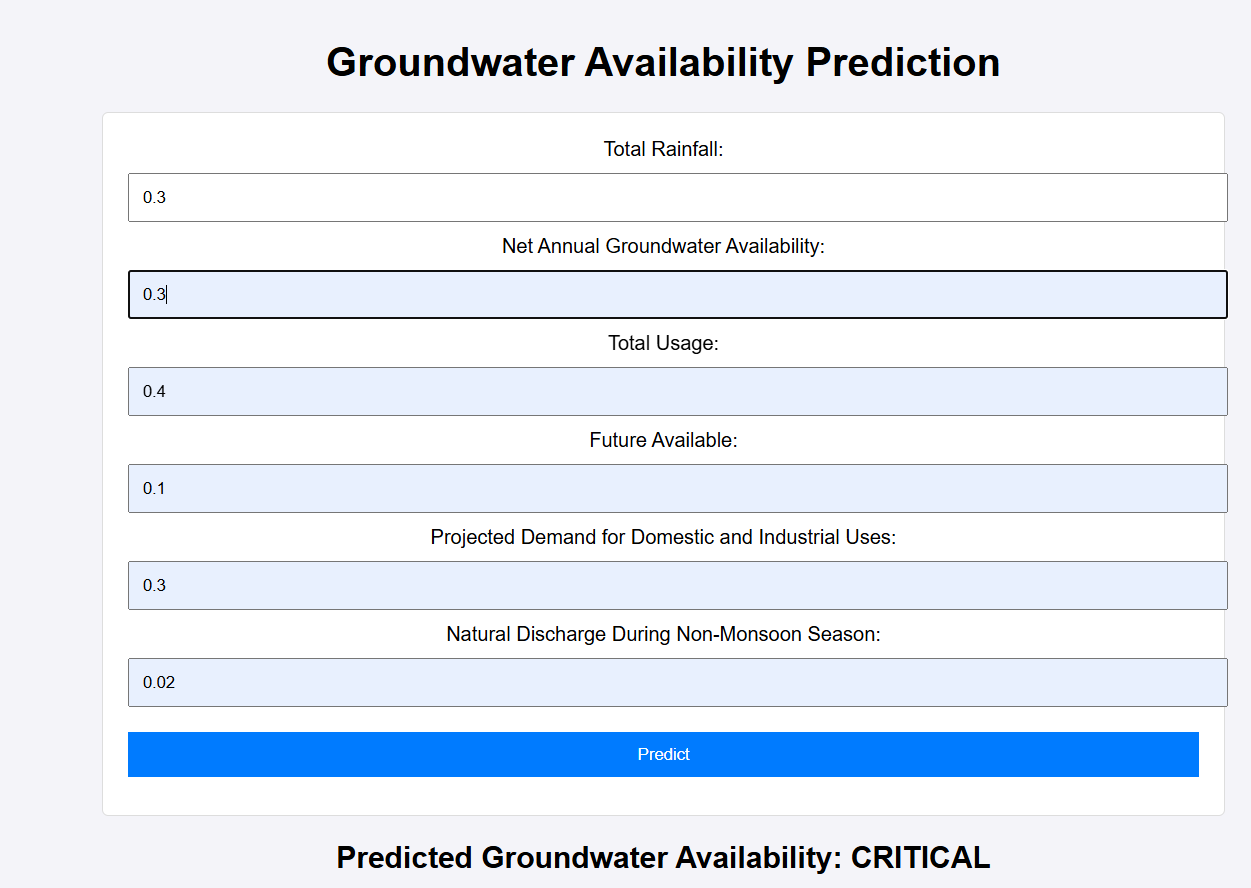


Fig 6.6 user-interface(3)

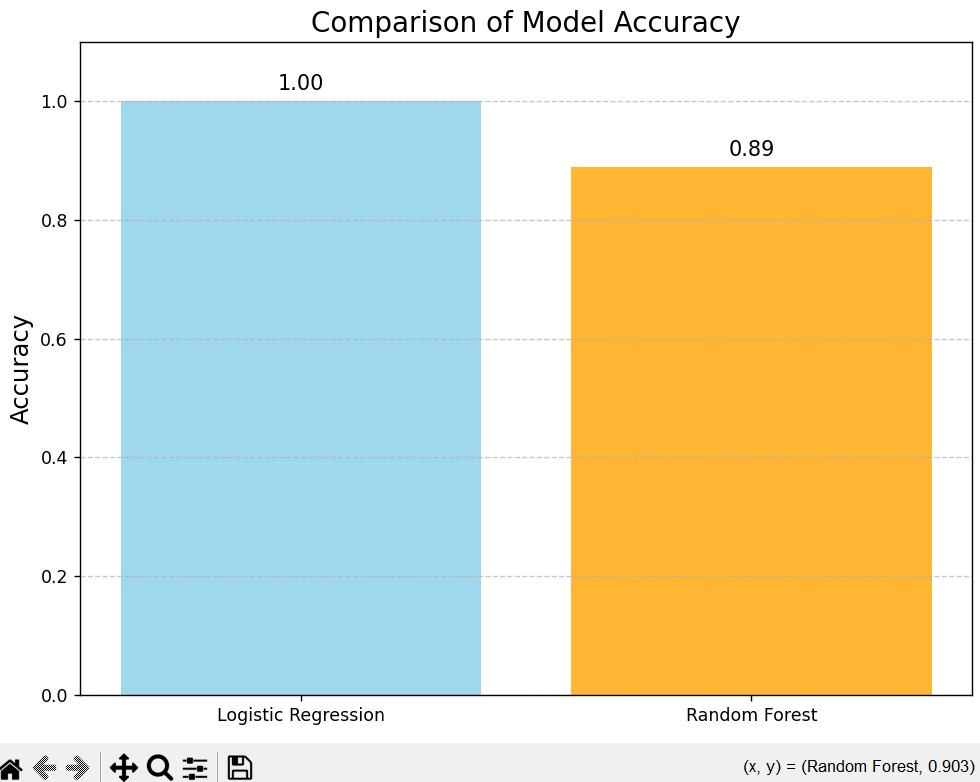


Fig 6.7 Comparison of model Accuracy

**CHAPTER 7**

**CONCLUSION**

In this project, we successfully implemented both Logistic Regression and Random Forest models to predict groundwater levels using Python for data processing and modeling, and HTML for presenting the results in a user-friendly interface. The project aimed to address the growing need for monitoring and predicting groundwater levels, which are crucial for sustainable water resource management.

The Logistic Regression model provided reliable predictions for groundwater levels, classifying them as safe or at risk, while the Random Forest model demonstrated superior accuracy in handling complex patterns in the data. Python was used for data preprocessing, feature selection, and model building with libraries like Pandas, Seaborn, and Matplotlib for analysis and visualization. A user-friendly HTML front-end, integrated with Python, allowed real-time predictions through a web-based platform.

While Logistic Regression proved effective for binary classification, Random Forest excelled in multi-class classification and capturing non-linear relationships in the data. Performance can further improve with additional data and feature engineering. Future enhancements include exploring advanced models, integrating real-time data, adding geographical visualization using tools like Google Maps API, and making the interface more mobile-friendly.

**CHAPTER 8**

**REFERNENCES**

1. Dataset from github:<https://github.com/rgdeekshith/Ground-Water-Level-Analysis-and-Prediction>
2. Levy, J.; Xu, Y. Groundwater management and groundwater/surface-water interaction in the context of South African water policy. *Hydrogeol. J.* **2012**, *20*, 205–226. [[**Google Scholar**](https://scholar.google.com/scholar_lookup?title=Groundwater+management+and+groundwater/surface-water+interaction+in+the+context+of+South+African+water+policy&author=Levy,+J.&author=Xu,+Y.&publication_year=2012&journal=Hydrogeol.+J.&volume=20&pages=205%E2%80%93226&doi=10.1007/s10040-011-0776-4)] [**[CrossRef](https://doi.org/10.1007/s10040-011-0776-4" \t "_blank)**]
3. Ghasemian, D. Groundwater Management Using Remotely Sensed Data in High Plains Aquifer. Ph.D. Thesis, The University of Arizona, Tucson, AZ, USA, 2016. [[**Google Scholar**](https://scholar.google.com/scholar_lookup?title=Groundwater+Management+Using+Remotely+Sensed+Data+in+High+Plains+Aquifer&author=Ghasemian,+D.&publication_year=2016)]