NITTE MEENAKSHI INSTITUTE OF TECHNOLOGY

AN AUTONOMOUS INSTITUTION, AFFILIATED TO VISVESVARAYA TECHNOLOGICAL UNIVERSITY, BELGAUM,
APPROVED BY AICTE AND GOVT.OF KARNATAKA)

Department of Computer Science and Engineering

PROJECT REPORT

on

"Water Table Analysis using Machine Learning"

Submitted in partial fulfillment of the requirement for the award of Degree
of
Bachelor of Engineering in Computer Science and Engineering
Submitted By:

AISHWARYA S. KULKARNI	1NT16CS007
SHIVANGI NEGI	1NT16CS104
SUMEDHA RAGHU	1NT16CS117

Under the Guidance of
Dr. Vijaya Shetty S.
Associate Professor, Dept. of CSE,NMIT

Department of Computer Science and Engineering
2019-20

NITTE MEENAKSHI INSTITUTE OF TECHNOLOGY

AN AUTONOMOUS INSTITUTION, AFFILIATED TO VISVESVARAYA TECHNOLOGICAL UNIVERSITY, BELGAUM, APPROVED BY AICTE AND GOVT.OF KARNATAKA)

Department of Computer Science and Engineering

CERTIFICATE

This is to certify that the Project titled **"Water Table Analysis using Machine Learning"** is an authentic work carried out by **Aishwarya S. Kulkarni (1NT16CS007)**, **Shivangi Negi (1NT16CS104)** and **Sumedha Raghu (1NT16CS117)** bonafide students of Nitte Meenakshi Institute of Technology, Bangalore in partial fulfilment for the award of Degree of ***Bachelor of Engineering*** in COMPUTER SCIENCE AND ENGINEERING of the Visvesvaraya Technological University, Belgaum during the academic year **2019-2020**. It is certified that all corrections and suggestions indicated during the internal assessment has been incorporated in the report. This project has been approved as it satisfies the academic requirement in respect of project work presented for the said degree.

Internal Guide	**Signature of HOD**	**Signature of Principal**
Dr. Vijaya Shetty S. Associate Professor, Dept. of CSE, NMIT, Bangalore	Dr. Thippeswamy M. N. Professor, HoD, Dept. of CSE, NMIT, Bangalore	Dr. H. C. Nagaraj Principal, NMIT, Bangalore

Name of Examiner	Signature of Examiner
1.	1.
2.	2.

DECLARATION

We hereby declare that:

- The project work is our original work.
- This project work has not been submitted for the award of any degree or examination at any other University/College/Institute.
- This project work does not contain other persons' data, pictures, graphs or other information, unless specifically acknowledged as being sourced from other persons.
- This project work does not contain other persons' writing, unless specifically acknowledged as being sourced from other researchers. Where other written sources have been quoted, then: a) their words have been re-written but the general information attributed to them has been referenced; b) where their exact words have been used, their writing has been placed inside quotation marks, and referenced.
- This project work does not contain text, graphics or tables copied and pasted from the Internet, unless specifically acknowledged and the source being detailed in the References section.

NAME	USN	Signature
AISHWARYA S. KULKARNI	1NT16CS007	Aishwarya
SHIVANGI NEGI	1NT16CS104	
SUMEDHA RAGHU	1NT16CS117	

Date: 16 July 2020

ACKNOWLEDGEMENT

The satisfaction and euphoria that accompany the successful completion of any task would be incomplete without the mention of the people who made it possible, whose constant guidance and encouragement crowned our effort with success. We express our sincere gratitude to **Dr. H. C. Nagaraj**,Principal, Nitte Meenakshi Institute of Technology for providing facilities.

We wish to thank our HOD, **Dr. Thippeswamy M.N** for the excellent environment created to further educational growth in our college. We also thank him for the invaluable guidance provided which has helped in the creation of a better project.

We hereby like to thank our guide, **Dr. Vijaya Shetty S.**, Associate Professor, Department of Computer Science & Engineering on her periodic inspection, time to time evaluation of the project and help to bring the project to the present form.

Thanks to our Departmental Project coordinators. We also thank all our friends, teaching and non-teaching staff at NMIT, Bangalore, for all the direct and indirect help provided for the completion of the project.

NAME	USN	Signature
AISHWARYA S. KULKARNI	1NT16CS007	Aishwarya
SHIVANGI NEGI	1NT16CS104	
SUMEDHA RAGHU	1NT16CS117	

Date: 16 July 2020

ABSTRACT

Machine learning is an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed. Machine learning focuses on the development of computer programs that can access data and use it learn for themselves. Groundwater level is the depth below the earth's surface that is saturated with water, or the level to which groundwater would rise in a well that is drilled in a confined (pressurized) aquifer. Ground water levels, as of 2019, in India, are said to be depleting at alarming rates in most of the states. 21 major cities of India are expected to run out of groundwater as soon as 2020, affecting around 100 million people. It is important to keep the ground water levels in check for sustainable usage of water resources. Quantification of the groundwater recharge is a basic prerequisite for efficient groundwater resource development and this is particularly vital for India due to prevalent semi-arid and arid climate.

Contents

List of Figures

List of Tables

Chapter 1

INTRODUCTION

Background

Groundwater is characterized as the water present in the sections between the soil pore spaces and within the cleavages of earth formations. A chunk of rock or a non-consolidated deposit that can produce a viable amount of water is described as an aquifer. The depth at which soil pore ruptures and gaps in rock turn inundated with water is determined as the water table [1]. In several arid and semi-arid areas, groundwater has emerged as a vital source of water required for domestic, irrigation, urban, and industrial activities. India is a substantial consumer of groundwater in the world using 230 cubic kilometers of groundwater per year, which is equivalent to the fourth of the world total [2]. Therefore, sustainable development of groundwater resources is essential for precise quantitative analysis, which is necessary for India due to its prevalent semi-arid and arid climate. Continuous monitoring of groundwater levels is essential to prevent the misuse of groundwater resources that can usher to local water rationing, limiting in agricultural yields, wells going dried- up or generating unpredictable groundwater quality variations, differences in flow behaviours of groundwater emerging in the inflow of meagre quality water and seawater intervention in coastal areas [3]. Groundwater reservoir is a complex structure which is challenged with either natural or artificial factors that derives from human actions. For the prediction of groundwater level fluctuation for different natural conditions, usage rates are of great importance for the operation and administration of groundwater resources. The water levels, if forecasted prior, might facilitate the executives to plan better, the groundwater usage. In this research, we focus on observation wells from various districts of Karnataka. Tradi-

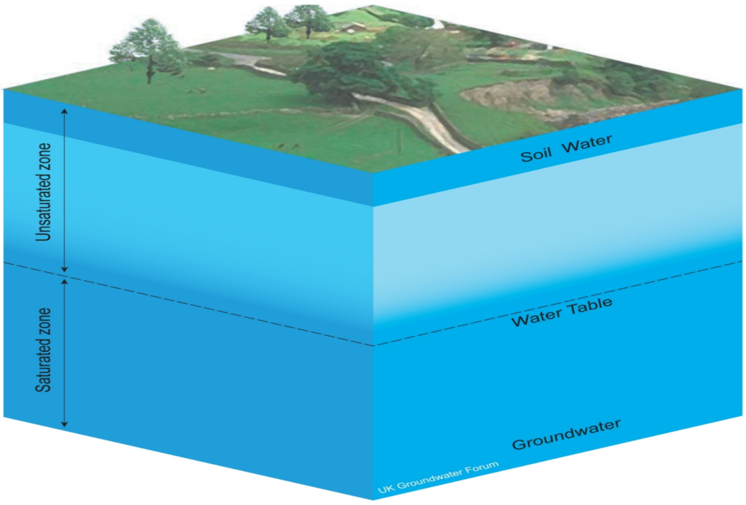


Figure 1.1: Groundwater

In the Figure 1.1, Ground water process and Ground water storage is depicted

tionally, process-based models are frequently used to perform groundwater simulation and predications, which rely on spatial data of the observed system dynamics. Although, they are not suitable in several arid and semi-arid areas as a consequence of insufficient data. Meanwhile, in data-driven modeling with machine learning methodology, our model attempts to establish a direct relationship among the inputs and outputs of the system without having any knowledge about the interior structure of the physical process [4]. The focus here is to use temporal data as inputs (historic groundwater level, weather, and rainfall data) to learn the best approximation of the groundwater level values [4].

Recurrent neural networks (RNNs), a technique of deep learning, are a prominent choice for designing groundwater time series data due to their ability to retain a memory of previous network conditions, but they face challenges in acquiring long term dependencies within variables as weights associated with the network reaches to zero or turn exceedingly large during training the model [5]. LSTM, a class of RNN can prevent these training problems by eradicating unessential information being routed to future model states while keeping a memory of relevant past events [5]. LSTM networks have lately incorporated to model the groundwater table in an inland agricultural area of China on a monthly time step basis.

In our study, we are training models based on a set of related attributes to generate optimal predictions for the missing groundwater level values in our dataset using binary classification and later applying data-driven techniques to evaluate the performance of our models. Although the growing applications of data-driven methods in surface water problems, there are hardly any studies associated to groundwater in arid and semi-arid areas [4]. Therefore, the emphasis of this study is on the implementation of data-driven models with machine learning (i.e., RF AND GB) and deep learning (i.e., LSTM) and comparison of two ensemble methods (i.e., RF and GB) for forecasting groundwater levels in Karnataka, India.

Brief History of Technology/Concept

Machine learning (ML) is a class of algorithms that permits software system packages to become more precise in predicting results whereas not being explicitly programmed [6]. Machine learning specializes on the event of computer programs which will be able to obtain information and use it to learn on their own. The fundamental objective is to permit the computer systems to be informed automatically without human intervention and alter actions consequently [6]. Machine learning algorithms are usually classified as these two types: supervised or unsupervised.

- **Supervised machine learning algorithms** requires the data scientist to use new data based on what has been previously learnt in the past by using labelled instances to predict future events or values. The system can deliver target values for any new input after enough training. The learning rule also compares its output with the supposed output and identify errors with the purpose to adjust the model appropriately [6].

- **Unsupervised machine learning algorithms** are used once the knowledge accustomed to train is neither classified nor labelled. Unsupervised learning analyses a way for the systems to deduce an operation such that it explains a hidden patterns from unlabelled data. The system examines the data and can derive a conclusion from datasets to explain hidden patterns from unlabelled data to group them into subsets [6].

- **Semi-supervised machine learning algorithms** lies in the middle ground between the performance of supervised and efficiency of unsupervised learning as they both use labelled and unlabelled data for training, generally a small number of labelled data and a great number of unlabelled data. This algorithm helps in identifying the size of the dataset and then apply it to new unlabelled data [6].

- **Reinforcement machine learning algorithms** are a learning methodology that interacts with its surroundings by generating actions and figuring out errors or positive rewards for the ultimate goals. Trial and error search and overdue rewards are the primary significant features of reinforcement learning. This methodology permits machines to automatically specify the standard conduct within a particular situation so as to optimize its performance [6].

Most commonly used machine learning models are:

- **Regression** : In machine learning, this model is based on the relationship between variables obtained from the dataset which determines the outcome of an event.
- **Classification** : It is a supervised learning technique which classifies the output into categorical form after the computer program learns from the data input given to it.
- **Decision trees** : This is a type of supervised learning algorithm that uses views about specific actions and determines an optimal way for arriving at an intended result for classification problems [7].
- **K-means clustering** : This model takes into consideration a specified number of data points such that it can group them into a specific number of groupings based on similar characteristics [7].
- **Neural networks** : These are deep learning models that employ large amounts of training data in order to identify correlations between different variables to learn how to process incoming data in the future. For example: Artificial Neural Networks (ANN) [7].

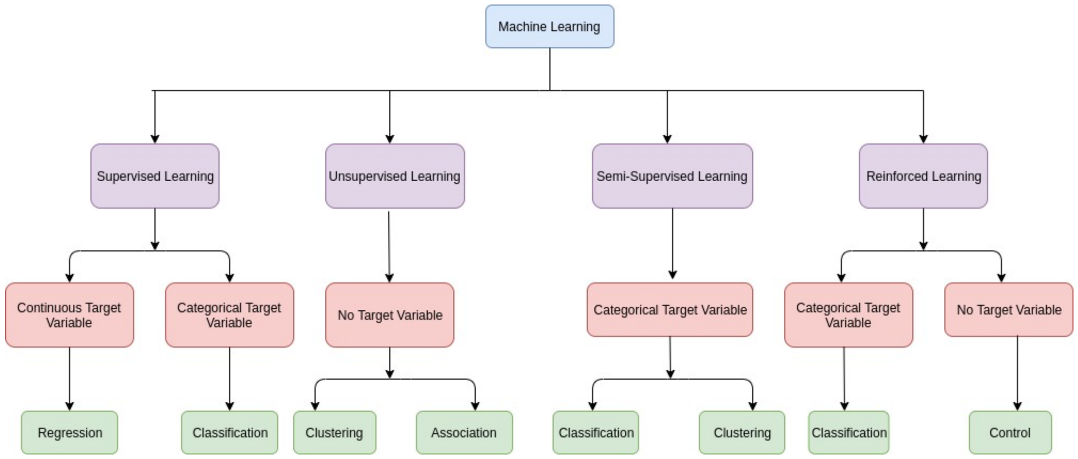


Figure 1.2: Groupings of Machine Learning
Figure 1.2 depicts the various modules that fall under the broad classiffication of Machine Learning

Applications

In the field of hydrology, groundwater would be useful for identifying locations that are vulnerable to groundwater depletion under the influences of a changing climate and increased groundwater demand. It is a vital element of suitable water resources management. Forecasting of water table plays a crucial role in managing groundwater resources in agricultural regions where there are sewage networks in river valleys [8]. Such predictions may be effective in coastal regions in organizing the use of groundwater and surface water to protect seawater intrusion or water logging condition by maintaining the natural water table grade.

Research Motivation and Problem Statement

Research Motivation

Groundwater levels are an indicator for groundwater present in aquifers and shallower water tables. Groundwater decrease is a real-world serious issue in several parts of the Nation and the world. The levels of water in aquifers is often not constant. Groundwater levels rely on recharge from seepage of precipitation even when drought hits the land surface or if exploitation takes place it can affect the water levels below ground [9].

Contributions of groundwater to streams, rivers, lakes and wetlands play a major part in holding surface water quality, quantity, temperature and all vitals that maintain the health of marine plants and animals [10]. The consumption of water increases every day with the growth in population, and urbanization. The groundwater level is going down day by day. Measurement and analysis of groundwater level is needed for maintaining groundwater availability.

Hence, for the management of the groundwater level, our project aims to build a model that is required to predict the groundwater levels in the future, with the currently available information.

Statement of the Problem

For the management of the groundwater level, this project aims to build a model that is required to predict the groundwater levels in the future, with the currently available data.

The primary objective of the project is to model Groundwater Table using Random forest, Gradient boosting and LSTM (long short term memory) based o historical data.

Research Objectives

Objectives

The main objectives of the project are :

1. Given the groundwater level data for observation wells, to predict the missing grounwater level data.

2. Given the previous years groundwater level data to predict future groundwater level data.

3. Using different algorithms along with selected groundwater dataset, to achieve more accuracy of prediction than the existing system.

4. Using previous years' rainfall data along with groundwater level data, to make better predictions.

Summary

This chapter elucidates initially on the very definition of groundwater and its existence. Next it moves on to the applications of groundwater in daily life and its purposes. Further, the need for monitoring of groundwater and the consequences of its excessive consumption is mentioned. Reading about this helps one understand the solemnity of the situation and how necessary actions are needed to be taken as quickly as possible. After that it throws light on the technology used for the project that is Machine Learning. The history of machine learning and development since its inception is highlighted. Then it has explained about the types of machine learning models and the commonly existing models. This summarizes the brief understanding of the technology. Further, it mentions the applications of machine learning in the field of hydrology which is accompanied by the research objectives of our project and finally the problem statement. The following chapters include detailed explanation on the work carried out in the project.

Chapter 2

LITERATURE SURVEY

Introduction

Despite the resources being valuable, 29% of groundwater blocks are semi- critical, critical, or overexploited, and the situation is worsening rapidly .Majority of the exhaustion of aquifers is in the densely populated and economically resourceful areas. Change in climate further strains the resources. Monitoring water table fluctuations is essential, and it is even more important to predict the groundwater level to plan needs.

A complex web of factors that determines groundwater levels, which are: Rainfall, aquifers, precipitation levels, seasonal changes, patterns of groundwater storage, water extraction, quality of water, Atmospheric pressure, evapo-transipration levels, type of area and water yield levels.

The following literature survey on groundwater levels consists of papers that talk about previously implemented projects on analysing and preserving groundwater levels, around different regions in the world.

Related Work

A survey of various papers were conducted, some which had similar approaches while some had a rather indirect and a unique approach.

In [3], In the Dawu Aquifer of Zibo in Eastern China, Artificial Neural Networks (ANN) were used to forecast the groundwater level. To predict groundwater flow and table fluctuations two types of approaches are used, which can be divided into two classes i.e. deterministic approaches and stochastic approaches. Deterministic approaches include the analysis of the groundwater level using water balance methods, analytical methods, and numerical simulation based on the theory of groundwater dynamics. Stochastic approaches include regression analysis, time series analysis, stochastic differential equations, etc.

There are about 158 wells pumping water for industrial use with daily use of $49X104m^3$, which make up the main groundwater consumption and lower the groundwater level. The increase in groundwater level depends on the precipitation and river infiltration and groundwater use. These factors were considered in the groundwater level prediction model. Monthly observation data from June 1988 to May 1998, including the groundwater level, precipitation, and flow from the Taihe Reservoir, and groundwater use were used for the groundwater level prediction. The Auto-Correlation Analysis was used to understand the relationship between historical groundwater level fluctuations and the present fluctuations. The auto-correlation coefficient and the autocorrelation chart were used to understand the time-dependent character of the groundwater level. The results of the auto-correlation anal- ysis indicate that the monthly groundwater level is related to the level of the previous month, and is also related to the level two months before and to that of the same month one year before. The results of autocorrelation analysis were used to design the ARANN model to predict the monthly groundwater level. It considers the input vector consisting of, the groundwater level one month ago, two months ago, and the same month of the previous year. The output vector consisting only of the predicted groundwater level of the current month. The observed data is divided into two parts for before and after the May 1996,for both training and testing accordingly. This model considers only the impact of the historical groundwater level on the present level, which may result in over-estimation of the groundwater level for certain years. Therefore, a RARANN model was developed which considered not only the time dependence of the groundwater level, but also the main groundwater recharge and discharge factors like the usage of water, precipitation, and the river infiltration.

In this model, the input vector consisting of 4 components, i. e. , the monthly precipitation, monthly groundwater use, monthly water recharge from the Zi River and the groundwater level one month ago. The output vector contained only the groundwater level of the current month. The training and testing results fit well with the observed data, which is proof that the model effectively describe the connection between the changes in groundwater level and the pressing factors.

In [4], This study is of the Godavari river delta system in East Godavari district of Andhra Pradesh situated in South India. Topographically, the Delta, lies between 16.25 N to 16.55 N latitude and 81.44E to 82.15 E longitude with its water borders for the river Gowthami. The ANN models purpose here will be to predict the hydro levels for four months in advance. These might be required in combined use planning of groundwater and surface water in the coastal areas that helps keep up the natural water table gradient to protect from incoming of seawater or water logging condition. It is extremely important to understand the different variations of the water level for the management of same near the oceans and seas.

For the project, Back propagation algorithm is used for training, where through trial and error hidden neurons is optimized. Beginning with 2 hidden neurons initially,it increased up to 10 having step size of 1 in each trial. For each set of hidden neurons, the network must be trained in batch mode for each of the hidden neuron sets so that the mean square error is minimised. For checking over-fitting during training, a cross validation was carried while tracking efficiency of the fitted model. When there was no significant improvement in efficiency,training is closed. A sigmoid function is used as the activation function in both hidden and output layers. As the sigmoid transfer function has been used in the model, the input-output data have been scaled appropriately to fall within the function limits.

The data is used for training the algorithm after standardization to get rid of the cyclicity or invariability in the data. The scalability is limited to 0-1as the activation function warrants. The entire dataset has been split into two , calibration, and validation set: the model is trained under data of six years (1981–1986) and validated on the rest (1987–1989). The final representation of the ANN model for Munganda observation well is: 8 input , 3 hidden and 1 output neuron respectively and for Cheyyeru observation well is: 10 input, 2 hidden and 1 output neuron, respectively. The ensuing water levels obtained from the model are studied by various indices employed for performance caliberation of models. The wellbeing of fit statistics included are coefficient of correlation (CORR), average absolute relative error (AARE) ,root mean square error (RMSE) between the computed and observed runoff, and percentage error in deepest level estimation (%EDLF). The analysis shows that the water levels at Munganda well whenever seen provides a significant correlation with the water level at Kattunga well with a lag of 1-time step (month). The difference in the observed and predicted water levels, is useful when assessing the model developed . This kind of error plotting helps understand if the model is predicting the increase/decrease in levels. Although required results are obtained on Munganda observation for 4 months ahead forecasts, the model performance is seen to crash in 2 months for lead forecast of Cheyyeru observation well.

In [5], Backpropagation algorithm has been used in a Feed-forward artificial neural network to forecast hydrological variations. Error estimation methods like RMSE(Root mean squared errors), MAE(Mean arithmetic error) and Coefficient of efficiency $R\hat{}2$ to improve the model's accuracy. These methods are used to minimize the errors by iteratively analysing the errors and improve the model's accuracy.

The data used for training and testing purposes of the model are the factors in measuring ground water level quality which are complex and nonlinear. The models used in this paper are all stochastic models and have gradient descent due to time series analysis involved in its implementation. In the ANN, it contains hidden layers in the network.
These hidden layers are used to capture the non-linearity of data. This is done using the Backpropagation algorithm used in Artificial neural networks. Backpropagation computes the way the errors are compute on the output side of the network. These errors are captured by the Backpropagation algorithm and is propagated back to the input layer via the hidden layer. The lower the values of RMSE, MAE and $R\hat{}2$ values, the model is more accurate.

The data used in the analysis is collected from irrigation wells in agricultural fields of a village in China. Time-series analysis is used for forecasting the ground water levels and it uses the monthly average of the groundwater table of these wells. The BPANN is a famously implemented forecasting method. Improvisation of BPANN is attempted in this study. Results suggested that the BPANN model is implied for modeling of groundwater levels for forecasting purpose in this area.

The levels estimated, provided the respective salinity information, which was needed for most applications. In [6], the lower end of Tarim River is taken as the study area. They have proposed a gray correlation analysis and cloud generator (GCA-CG) based groundwater level prediction model. The observation data that contains uncertainty is most crucial characteristic feature. A grey system is a system in which part of information is known and part unknown.

Grey analysis then provides clear set of statements about system solutions. This GRA is used with ground water table prediction since the data is nonlinear and random. This means that the model that is used, predicts the fuzzy and random values that are required for ground water analysis. Because of human influences, most records showed variations of large ups and downs i.e. not having many typical distribution patterns. Implementing statistical ways were shown to not give much useful conclusions. The grey correlation analysis remedied this flaw that existed in present statistics when it was applied in the context of systems analysis. It was applied to cases of different sizes and distributions with a relatively lesser computation. The characteristics used for this model were especially useful to estimate uncertainty of groundwater levels, even during limited information about the soil composition is at disposable.

In [7], Least Squares Support Vector Machine(LSSVM) for dynamic Ground Water level forecasting in the Hetao district used for irrigation in Inner Mongolia. The factors included in this Groundwater level forecasting are random, fuzzy, and nonlinear. Support Vector Machine is used on a small set of samples and is known as small-sample machine learning. In the LSSVM method, it is checked for errors, whether it exceeds the threshold or not for a non-linear regression function. Chaos Optimization Arithmetic is used in Simulated

Annealing method. The Hetao region in Mongolia has a lot of illegally constructed wells that are used for irrigation. These irrigation wells are analyzed based on various factors such as rainfall data, aquifer data, precipitation, seasonal changes, quality of water, water extraction patterns, area type and other such factors affecting the groundwater level. The model used to predict the ground water levels in this paper has high precision due to the usage of the LSSVM. The models used in this paper are all stochastic models due to time series analysis involved in its implementation. In this paper, the theory of support vector machine in small-sample machine learning theory was introduced into dynamic prediction of groundwater level.

Looking into water level dynamics in terms of length as well as peak mutation characters, the least squares support vector machine arithmetic which is relying on peak value identification was pro- posed [7].

In [8], Radial basis function (RBF) neural network which displays complex nonlinear characteristic is used to predict groundwater table. The rudimentary RBF training Algorithm could only obtain the partial/local optimums solution and was based on gradient descent optimization method . Interestingly, selecting the structure of RBF neural network causes confusion and wastes much time. Therefore, differential evolution (DE) algorithm came to existence to automatically search the weight of output layer, the center of RBF and the network width. A self- adapting crossover probability factor was presented to improve the quantitative diversity and the chance of escaping from the local optimum, Furthermore, to improve the convergence of DE algorithm, a chaotic sequence based on logistic map was employed to self-adaptively adjust mutation factor.

Based on logistic map a chaotic sequence has been deployed to adjust mutation factor, which can optimize the convergence of DE algorithm. When compared to the traditional RBF neural network, DE trained RBF neural network is more robust and improves the convergence speed and precision of groundwater table prediction.

According to the paper, Water tables usually displays complex nonlinear characteristic. Therefore Back Propagation (BP) neural network is used here. Manual checking of the

structure of this kind of neural network time consuming, hence to automatically search BP neural network weight matrix and threshold matrix differential evolution (DE) algorithm was adopted .This paper combines the methods of Differential Evolution and Backpropagation methods to predict the ground water table levels.
The BP algorithm is successful, but it is shown to have some disadvantages. The selection of the inertial factor and learning factor affects the convergence of the BP neural network which is otherwise found by experience. Man-made selection of the structure is said to have blindness and expends much time. DE applied to the non-linear models for optimization is proved to be successful. In a crowd of possible solutions for a problem inside an n-dimensional search space, a stagnant count of vectors are initialized in random, then brought u overtime to understand the search space and locate the minima of the objective function [9].

Distinguishing with the traditional BPANN, DE trained BP neural network is said to be more robust, and improves the accuracy and precision of groundwater table prediction.

In [10], It is a trial to forecast groundwater level fluctuations monthly using integrated wavelet and modeling using support vector machine. The discrete wavelet transform with two coefficients (db2 wavelet) is utilized here for deconstructing the input data into respective wavelet series. These series are further utilized as input variables for different combinations for Support Vector Regression(SVR) model to forecasting. Wavelets are decomposing the original time series into various parts. The wavelet components are immensely helpful for increasing the capability of a model by capturing useful information at various levels. WA-SVR model introduced has a combination of wavelet and support vector regression(SVR) to predict the groundwater level variations for 3 water wells.

In [11], The light is on quantitative measurement of groundwater temporally and spatially. Data is taken from Groundwater Survey and Development Agency (GSDA). Analysis of ground- water level data is done in more than 100 observation wells in three districts of Maharashtra - Thane, Latur and Sangli. Three different type of models -periodic, polynomial and rainfall models. While periodic and polynomial models are trends on water levels in wells, the rainfall model is for the correlation between the former and latter . It also tells how excessive extraction causes sinking of land and water quality issues in Visakhapatnam, South India. The paper states that prediction of groundwater is extraordinarily complex and highly nonlinear in nature as it relies on different factors like evapotranspiration, soil characteristics ,precipitation, and topography of the water- shed. Further the WA-SVR model is able to predicts crests and troughs in the testing period when compared to the other models. The forecast performance is evaluated using the Absolute Percentage Error (MAPE), Root Mean Square Error(RMSE), Normalized Mean Square Error (NMSE), x,Correlation Coefficient (R2) and Root Mean Square Error(RMSE).

The results of this model are compared with SVR, ANN and Auto Regressive Integrated Moving Average(ARIMA) models.

Based on the maximum error in predictions and performance criteria it can be concluded that the WA-SVR model is a superior to other models to forecast the Groundwater levels. It also says about a multivariate time series analysis(D and A) to be considered for analysis by adding the two series or concurrently.

In [12], A direct mapping between the inputs and outputs of the system is identified by the data-driven techniques. The emphasis of this paper is the application and comparison of three data-driven techniques (i.e., SVM, ANN, and M5 model tree) for predicting short-term groundwater levels in the Shule river basin located in Gansu province of China. The Shule River basin has a very arid continental climate with low rainfall, low runoff coefficient, sandstorm, and high evaporation capacity. It is one of the driest areas of China. The ANN architecture that is used for prediction is the Multi-Layer Perceptron (MLP) network which is made up of a layered framework. A layer usually comprises of a number of neurons to which synapses are connected. Each neuron in one particular layer is connected to every other neuron in the next layer. Each synapse is associated with a weight. The information regarding the action of training set data is retained in the form of synapses weights. The sigmoid activation function was applied for both the hidden neurons as well as the output neurons. Support Vector Regression (SVR) is a type of regression that employs SVM. Given the training set, SVR maps each input to a high dimensional feature space by means of a nonlinear function and thereafter carries out a linear regression in this feature space to determine a function that can best estimate the actual output value with a fault tolerance. The M5 model tree model integrates a conventional decision tree so as to the prospect of creating a linear regression function at the end of a tree (i.e; leaves).

The predictive accuracy of the various models was evaluated using four numerical indicators i.e. the correlation coefficient (R), the root mean squared error (RMSE), and the Nash–Sutcliffe efficiency coefficient (NSE). The groundwater level and runoff have different units and their values do not represent the same quantities, so the normalization of data within a uniform range is essential. The overall performance of the models are found acceptable based on the high correlation efficiency. The M5 model holds the best performance in the testing period. The forecasted groundwater levels produced by the three models versus the measured values at the two stations in the testing period shows that there are a relatively good agreement between the simulated and observed groundwater level for all three models. The fact that the three data-driven models ran independently and generated good forecasting results was made to prove that the system can be applied to the real world scenarios.

The predictive accuracy of the various models was evaluated using four numerical indicators i.e. the root mean squared error (RMSE), the correlation coefficient (R), and the Nash–Sutcliffe efficiency coefficient (NSE). The values of water table and runoff have different units and don't depict the exact quantities, therefore, the standardization of data under a mean range is crucial. The overall outcomes of the models are reliable as it is based upon the high correlation efficiency. During the testing performance, the M5 model maintains the optimum performance. The forecasted formations produced by the three techniques in comparison with the measured values at the two stations within the testing period shows that there is a comparatively good understanding between the simulated and observed groundwater level for all the mentioned models. The actual fact that these data-driven methodologies worked irrespective of each other and produced great forecasting results that made us prove that the system can be employed to real-world problems.

In [13], a Feed Forward Artificial Neural Network (ANN) architecture has been designed and trained to learn the past water table fluctuations, to predict the future groundwater level in the wells of Upper Kodaganar basin. Monthly water level data of all these wells for the period of January 2004 to December 2014 has been collected from the Public Works Department (PWD) and weather data is collected from Indian Meteorological Department(monthly rainfall and monthly temperature). There might be missing observations in the data set, to find a missing value in a graph, the polynomial function or the combination of polynomial function that best represents the graph is used. Spline polynomial method also called spline is used to interpolate the non linear discontinuities, it is a common curve fitting strategy used to determine the missing values. In the next step the normalization was carried out to evenly distribute all the input data so that it falls in the range 0 to 1. A three-layer feed-forward neural network having an input layer (24 input neurons), one hidden layer and one output layer, is constructed for this study in MATLAB.

The normalized data is fed into the constructed ANN architecture, and trained by employing Levenberg- Marquardt training algorithm due to its ability to achieve convergence quickly and is way faster than the usual gradient-descent back-propagation algorithm. Then the model is calibrated and the trained ANN is now capable of water level prediction. The constructed ANN model is validated by comparing the predicted values and the Field Observed values based on the statistical indicators such as MAE (Mean Absolute Error), MSE (Mean Square Error), RMSE (Root Mean Squared Error) and correlation value.

In [14], a detailed analysis is carried out regarding the prediction of groundwater levels of Ljubljana polje aquifer. Machine learning techniques are applied using strongly correlated physical parameters as input data. The aim here, in such a situation, would be to predict groundwater levels based on temporal data inputs (historic groundwater and surface water level data, land-use, weather data, groundwater level reductions, and other anthropogenic data) and outputs (groundwater levels). Three different data sets from Ljubljana (Slovenia) and Skiathos (Greece) were used here. Those are groundwater information, pump sensor data from Skiathos, and weather data. Groundwater dataset contains data from 518 stations composing of 28 areas that evaluate the groundwater levels. Data collected are from 1960 onwards though, there are some stations that started operating later, or operated eventually, indicating some missing data. Weather data covers temperatures (daily average, minima, and maxima), location data, new snow blanket, cloud cover, precipitation, snow blanket, and sun duration. Work involves training a model based on a set of related attributes (weather data, available historic values of groundwater levels, etc.) that will be able to predict continuous groundwater level values. This type of problem belongs to the area of supervised learning, more precisely, it is a regression problem.

The early experiments focus on predicting the absolute value of groundwater levels. This shows to be a poorly explained problem since absolute water level depends clearly on long-term historic procedures. Therefore, the target value is not the absolute groundwater level, but it is the value of change in groundwater levels. As said before, it is a regression problem and based on available data (i.e., weather data, weather predictions, people-behavior model prediction, etc.) the algorithm is trying to produce the best feasible continuous predictions for groundwater level change on a certain day. Regression trees are based on decision trees. When learning, the attribute group is broken down into several different subgroups, in which each particular subgroup is indicated by a tree leaf that holds a value, which might be generated simply by averaging all the sample values from the training set that belongs to that leaf.

Regression trees are a good ensemble technique. Every regression tree is trained with a certain sub-sample of a data set. Gradient boosting also employs an ensemble of weak learners to generate the final output, but it piles them additively. In the beginning, the algorithm estimates target values. A random forests algorithm is also implemented. A variable stall in between the groundwater change and the predictors was established to incite the actual dynamics of aquifer recharge. Precipitation and cloud cover parameters were chosen as the attributes to predict the target value. Gradient boosting resulted as the best ensemble method with R2 - 0.644 and RMSE - 2.11 times.

Summary

This chapter highlights the series of algorithms used to form the model. The algorithms have been transformed over the years. The popular algorithms used these days are Random forrest, regression trees, SVR etc. It also gives insights to the shortcomings of different papers.

Chapter 3

SYSTEM REQUIREMENTS SPECIFICATION

General Description

System requirement specification is obtained by providing the appropriate platform to implement the system. It is the elaborative conditions which the system need to attain. Moreover, it provides a thorough understanding of the system on what to do, without having any conditions for the system on how to do. The specification gives out the implementation base or the plan and restricts for the outside visible characters.

Product Perspective

Anaconda

Anaconda is a free and open-source distribution containing R programming and Python for data science, machine learning applications, large-scale data processing, predictive analytics, and more, that aims to simplify package management and deployment. Conda Navigator searches the packages on the Anaconda Cloud or in a local repository, installs it in the user's environment, runs these packages, and updates them. It is available for Windows, Mac OS, and Linux.

Some applications that are available by default in Navigator include Visual Studio Code, Rstudio, JupyterLab, Spyder, Jupyter Notebook, QtConsole, Orange, and Glueviz.

TensorFlow Library

TensorFlow is an open-source software library for dataflow and differentiable programming across a range of tasks. It is a symbolic math library and is used in Machine Learning for applications such as neural networks. TensorFlow was developed by the Google Brain team for internal Google use. It was released for public use under the Apache License 2.0 on November 9, 2015.

FEATURES :

1. Responsive construct
2. Flexible
3. Easily trainable
4. Open source
5. Feature columns
6. Availability of statistical distributions
7. Layered components
8. Parallel neural network training
9. Visualizer(with tensor board)
10. Event logger(with tensor board)

Tensorflow is used to achieve all of the applications. The reason for its popularity is the ease with which developers can build and deploy applications. Moreover, Tensorflow has processing power limitations taken into consideration. The library is run on computers of all kinds, even on smartphones.

Hardware Requirements

The groundwater level prediction system is implemented in Anaconda, and to run Anaconda on the system, the following are the system requirements :

- Processors: Intel Core i5 processor which runs at 2.60 GHz, and has 8 GB of RAM
- Hard-Disk space: 1 GB to 1.5 GB

- Operating Systems: Windows 10, Windows 8, Ubuntu

Software Requirements

- Programming language: Python
- Included development tools: anaconda, anaconda-env, Jupyter Notebook
- Libraries: PIP , NumPy, scikit, tensorflow, keras

Functional Requirements

- The required parameters of the function is gathered from dataset and fit into the model.
- Regression techniques using Random forest, SVM, Regression trees and Gradient boosting.
- Finding the water levels in nearby areas implementing the ANN-Levenberg- Marquardt training algorithm.
- Accuracy is calculated based on performance between the algorithms.

Non-Functional Requirements

- Assessing the correct working of the system for different algorithms, with data during training and prediction stages making it reliable and capable.
- Comparing algorithms to determine which provides the best performance and also which is easy to maintain.
- Getting performance evaluated on the basis of its usage.
- Maintaining consistency of data during retrieval of data set for prediction.

Summary

This chapter includes the resources that will be required by our system to implement our Machine Learning model. For the success of any software project proper intelligent use of available hardware and software is important.

Chapter 4

DESIGN

High level design is the one used to design the software requirements. In this chapter the complete system design is generated that shows integrated system of the modules, submodules and the flow of the data between them. The errors done here will be modified in the coming processes.

Architectural Design

In Figure 4.1, the proposed system is implemented in python. The groundwater dataset is gathered from various aquifers and observational wells of regions across India. The data set is split into training and testing data and further feature extraction is carried. Further the processed data is fit into the various algorithms like Random forest, Regression trees, SVR in order to predict the ground water levels and their fluctuations. The output obtained can used for further analysis and necessary actions can be taken by the concerned authorities.

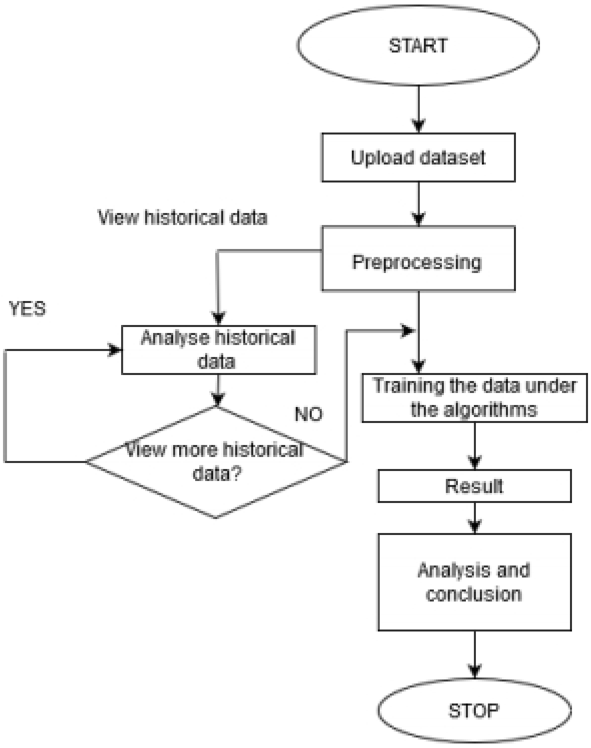


Figure 4.1: Architectural design of the system
The Figure 4.1 is the flowchart for the architectural design of the proposed system

Dataflow Diagram

It is the process which is explained in detail like how data flows between the different processes. Comprises of the input, process and ouput. After each process data flown between system is specified, hence called dataflow model. Figure 4.2 indicates 5 different modules namely data collection, feature extraction, regression, analysis and result display. The data set is gathered made to undergo all the above processes before finally displaying the conclusion.

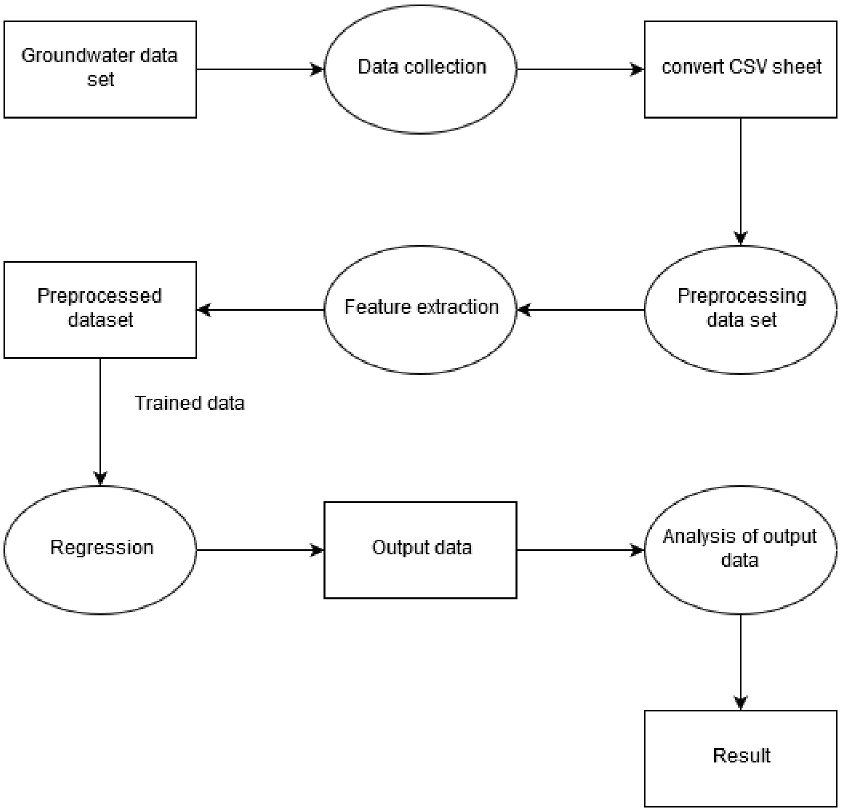


Figure 4.2: Dataflow diagram

The Figure 4.2 is the flowchart for the flow of data in the proposed system

Use case Diagram

The relationship among the user and the system is shown in Figure 4.3. A name within the ellipse indicates the use cases. Each use case represents the functionality of the system. A stickman notation is used for actor, with the name being placed below and solid line connects the actor and use cases. The actor actively takes part by collecting data, reprocessing the data using various techniques, assessing it and finally reading the predicted outcomes.

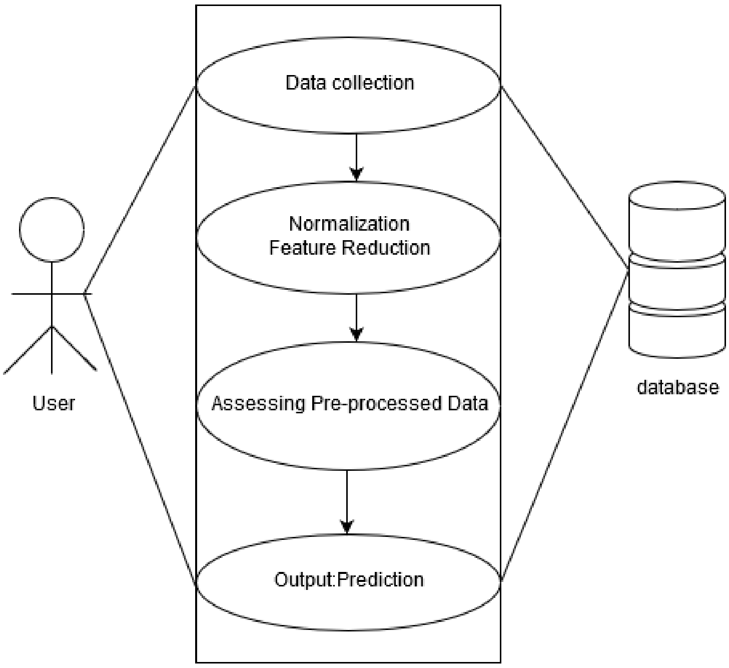


Figure 4.3: Use-case diagram

The Figure 4.3 is the flowchart for the use-case design of the proposed system

Chapter 5

IMPLEMENTATION

The working of the system is defined in this chapter. All the modules are interconnected and the system is made to run as a whole. It is the stage in which the model comes in real working.

Methodology

- **Spiral Model** : The Spiral model is an important Software Development Life Cycle model. It is very helpful in terms of support for Risk Handling and uncertainty. The spiral model combines the iterative development approach with the systematic and controlled aspects of the waterfall model. Each phase of the spiral model is divided into four phases, like a quadrant :

  1. **Determine objectives and recognize alternative solutions** : The phase starts with identifying the objectives and gathering the requirements like system requirements, subsystem requirements, and unit requirements are all done in this phase.
  2. **Find and resolve Risks** : In the second phase, all possible solutions are assessed and the risks affiliated with it are identified to finalize the best possible solution.
  3. **Development of the next version** : During the third phase, the distinctive features are developed and tested. The new, later version of the software is then available at the end of the third phase.
  4. **Review and planning for the next phase** : In the fourth and final phase of a

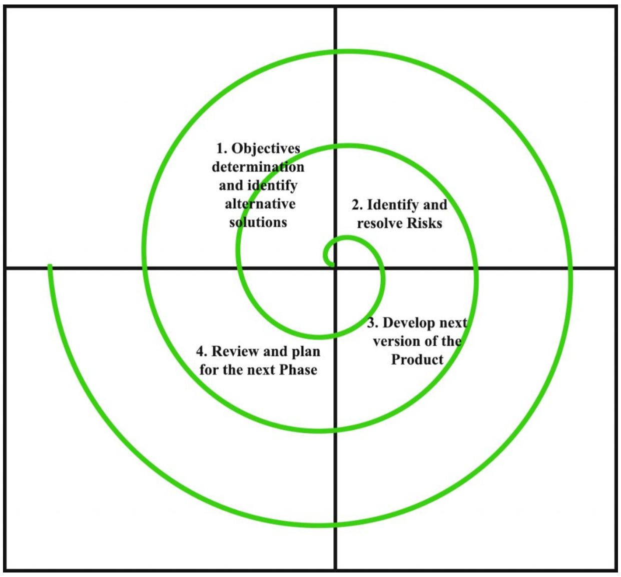

1. Objectives determination and identify alternative solutions
2. Identify and resolve Risks
3. Develop next version of the Product
4. Review and plan for the next Phase

Figure 5.1: Spiral model
The Figure 5.1 depicts the working of Spiral model for the proposed system

cycle, the later version of the software is evaluated. Towards the end, planning for the next phase is begun.

- **Random forest** : Random forest algorithm falls under Supervised Learning techniques, which uses the ensemble learning method for classification and regression problems. Random forest is a type of bagging technique that makes use of decision trees to train data, on a different set of data where With replacement sampling is done. A random forest regressor, one such bagging algorithm, is a meta estimator that makes use of a certain splitting criterion to estimate the characteristic of the split and fits classifying decision trees on multiple sub-samples of the dataset and uses an averaging method to enhance the predictive accuracy and limit over-fitting.

```
From sklearn.ensemble import RandomForestRegressor as RFC
# Create 100 trees for the model
mdl = RFC(n_estimators=100
bootstrap = True
max_features = 'sqrt')
# Fit on training data
```

mdl.fit(train, train_labels)

- **Step 1** : Start with the collection of random samples from a given dataset.
- **Step 2** : This algorithm is then used to create a decision tree for every sample. It will then get the prediction outcome from every decision tree.
- **Step 3** : Voting is then performed for every predicted outcome.
- **Step 4** : Select the most voted prediction result as the conclusive prediction result.

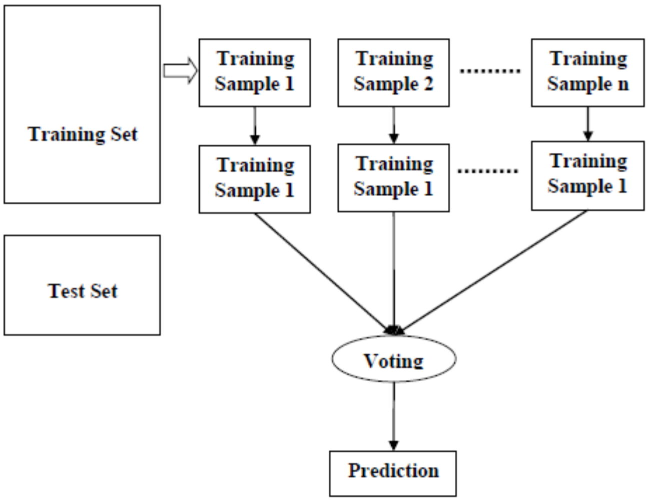


Figure 5.2: Random Forest
The Figure 5.2 depicts the working of Random Forest model in Machine Learning

- **Gradient Boosting** : In the Gradient Boosting algorithm (GBMs), the training method continuously fits new models to provide a precise estimation of the response variable. The system behind this algorithm is to create the new base-learners to be maximally correlated with the negative gradient of the loss function, correlated to the whole ensemble, to minimize the loss or errors. Gradient boosting consists of three elements:

  1. A loss function which is to be optimized.
  2. A weak learner to obtain predictions.

3. An additive model to add the weak learners, in order to reduce the loss function.

The advantage of the Gradient Boosting structure is that the very boosting algorithm can be used for every loss function that is to be used. Regularly used weak learners in gradient boosting are decision trees. The trees are appended one at a time, and existing trees in the model are not replaced. A gradient descent procedure is used to decrease the loss when appending trees.

```
From sklearn.ensemble import GradientBoostingRegressor
model = GradientBoostingRegressor(verbose=True)
print("Gradient boosting model before replacing with predicted values : ")
gbr1=model.fit(X_train, y_train)
print(gbr1)
```

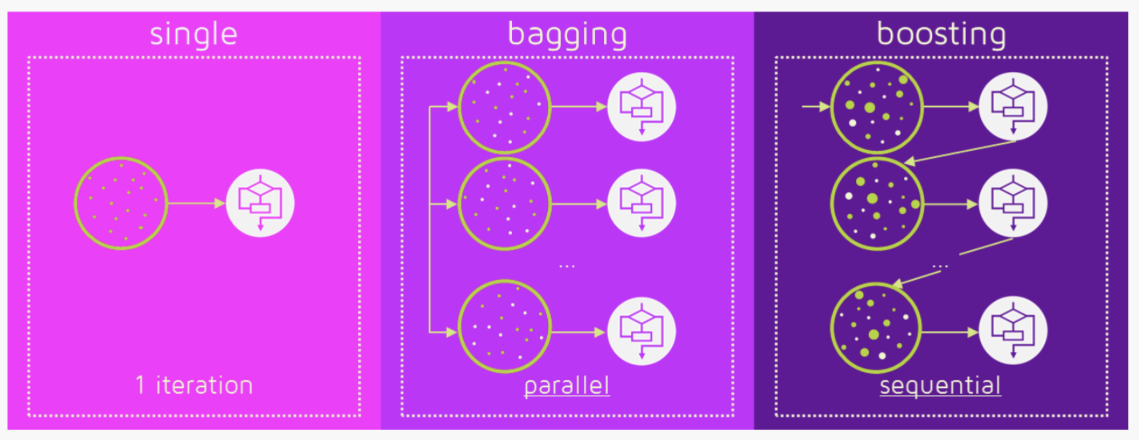


Figure 5.3: Gradient Boosting model
The Figure 5.3 depicts the working of Gradient Boosting model in Machine Learning

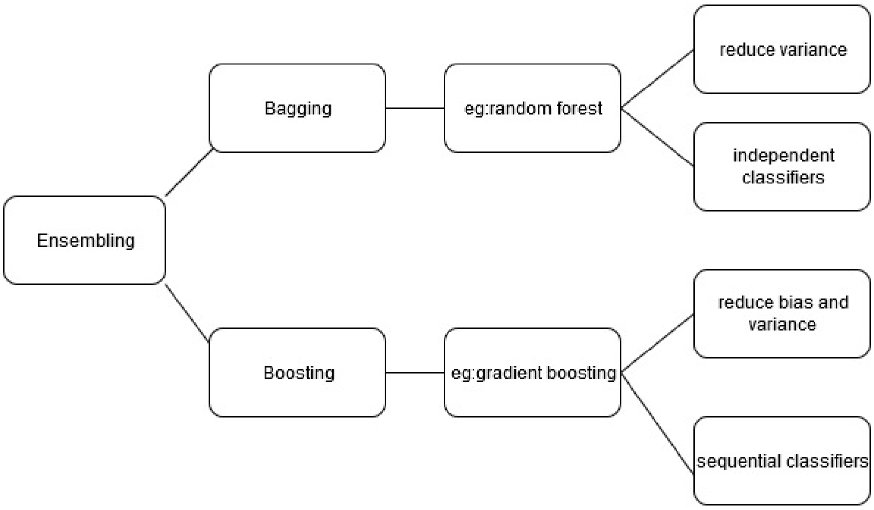


Figure 5.4: Random Forest vs Gradient Boosting
The Figure 5.4 is a comparison of Random Forest model and Gradient boosting model in Machine Learning

- **LSTM** : Long short-term memory (LSTM) blocks are units of a recurrent neural network (RNN)[15]. An RNN which is constituted of LSTM units forms an LSTM network. A standard LSTM unit consists of a cell, an input gate, a forget gate, and an output gate. The cell retains values over an arbitrary time period and the three gates direct the flow of information in and out of the cell. LSTM networks are well-suited for predictions, to classify and to process data based on time series data as there can be lags of unknown duration between crucial events in a time series algorithm. The forward pass of LSTM network is in the below equations :

$$f_t = \sigma_g(W_f x_t + U_f h_{t-1} + b_f) \tag{5.1}$$

$$i_t = \sigma_g(W_i x_t + U_i h_{t-1} + b_i) \tag{5.2}$$

$$o_t = \sigma_g(W_o x_t + U_o h_{t-1} + b_o) \tag{5.3}$$

Here f_t, o_t and i_t can be defined as forget gate, output gate, and input gate respectively. The matrix W_q contains the weight of the input and the matrix U_q contains recurrent connections and σ_g is the sigmoid activation function used in the LSTM network. The network output is calculated by stacking a fully pertinent layer on top of the LSTM cell. The product of the output layer is the prediction of the groundwater level for the coming season.

Dataset

Table 5.1 represents the dataset that has been used. The pre-processing is done on this dataset. The dataset consists of groundwater levels in the pre-monsoon season as well as the Post Rabi and Post Kharif crop seasons. It also consists of location data including the well-code, its district, state, site name, and site type.

SITE_TYPE	WLCODE	YEAR_OBS	MONSOON	POMRB	POMKH	PREMON
BORE WELL	W05243	2018	19.6 3	18.11	NA	NA
DUG WELL	W24336	2018	NA	NA	4.13	3.72
BORE WELL	W05497	2018	23	24.57	NA	NA
BORE WELL	W06424	2018	NA	59.32	NA	NA
BORE WELL	W21201	2018	NA	59.3	NA	NA
BORE WELL	W05727	2018	NA	21.06	37.5	NA
BORE WELL	W05731	2018	NA	7.64	9.1	NA

Table 5.1: Dataset
Table 5.1 is a tabular representation of a snippet of the Dataset being used.

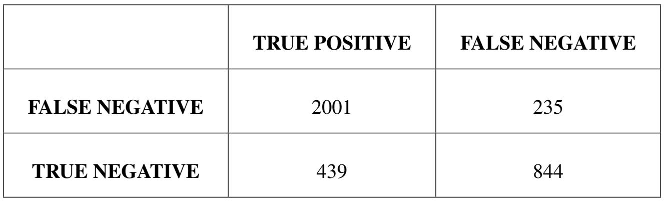
Pre-processing

The data gathered is compiled from various government and water board sources. The dataset initially required a significant amount of pre-processing. The techniques used on the groundwater dataset are data cleaning, data reduction ,and checking feature importance. The dataset compiled contained a large amount of NaN(not a number) values, due to which data was not usable. We first replaced the NaN values with 0s and then used an imputer function to alternate the 0s with the mean value of each column. Certain parameters like well code and site name were then dropped as it was not required for any of the methods employed in this project. The attributes used in the dataset are described in the Table 5.2.

ATTRIBUTE	ATTRIBUTE DESCRIPTION
STATE	Name of the state where the observational well is located
DISTRICT	Name of the district where the observational well is located
TEH_NAME	Name of the tehsil(administrative area) where the observational well is located
BLOCK_NAME	Name of the sub-regions in the district where the observational well is located
LAT	Latitude of the observational well location
LON	Longitude of the observational well location
SITE_NAME	Name of the site of the observational well
SITE_TYPE	Type of the site i.e. bore-well or dug-well
WLCODE	A specific unique number assigned to each observational well
YEAR_OBS	Year during which the observation has been recorded
MONSOON	Groundwater levels during the monsoon
POMRB	Groundwater levels post monsoon during the rabi crop season(October to November)
POMKH	Groundwater levels post monsoon during the kharif crop season(June to October)
PREMON	Groundwater levels before the monsoon season

Table 5.2: Dataset Description
Table 5.2 is a tabular representation of the attributes of the Dataset.

The groundwater dataset initially contained 14 parameters [Table 1], and these parameters had a lot of NaN values. During pre-processing, we replaced the NULL values 0. Using an imputer function, we replaced the 0 valued data with the mean value calculated for each column. The feature importance of the parameters containing numerical values was checked. We applied the random forest and gradient boosting algorithms on the dataset both before and after replacing the 0 valued data with the mean imputation. The accuracies of the two algorithms before and after the replacement of the null values were also calculated. We divided the dataset as a 70 : 30 train-test split and applied the two techniques to determine accuracies, Mean squared errors, Confusion matrices, and classification reports. For the random forest algorithm, we used the threshold value of 8.19, deliberated using the collective mean of all the groundwater values, to divide the data into 0 and 1 classes where any value below the threshold is classified as 0 and anything above the threshold is classified as 1. This division was created for the test and predicted classes to identify the data relationship between the true-positive, true-negative, false positive 4 and false negative, as shown in Table 5.3.



	TRUE POSITIVE	FALSE NEGATIVE
FALSE NEGATIVE	2001	235
TRUE NEGATIVE	439	844

Table 5.3: Confusion matrix from Random Forest Regression
Table 5.3 is a tabular representation of the Confusion matrix from Random Forest Regression.

In the test split that contained 3519 values, the following was observed. We then calculated the accuracy and MSE using the random forest regression algorithm. For gradient boosting algorithm, we made use of gradient boosting regressor function from the python libraries to calculate the accuracy and MSE of both before and after replacing the null values. These algorithms were also used to predict the missing data in the dataset. We then implemented an RNN network, LSTM i.e. Long Short Term Memory, to predict future groundwater levels, and to check the MSE values for training and test values, where we also check the accuracy of the predicted values compared to its test and train set values. By implementing the algorithms, we yield useful and necessary results.

Chapter 6

RESULTS

On extracting data from various sources, the data is cleaned and pre-processed, and is then fed to the random forest regressor model. To measure the effectiveness of the model the Confusion Matrix [Table 5.3] is provided. Let Truly Positive be TF, Truly Negative - TN, Falsely Positive - FP, Falsely Negative - FN, Actual results - $(TP + FP)$ and Predicted results - $(TP + FN)$; Based on the same it is found that the $Recall : (TP/Predictedresults)$, $Precision : (TP/Actualresults)$,$f1 - score : (2 * (Precision * Recall)/Precision + Recall)$ and $Accuracy : (TP + TN/TP + TN + FN + FP)$ to have increased from 0.81, 0.81, 0.81, 0.809 to 0.85,0.85, 0.85 and 0.81 respectively after replacing the null values with mean values. Thus, we understand that all the measures of the confusion matrix have been in increased in their values which asserts the replacement. The change in values due to replacement is shown in Table 6.1

	Before replacement	**After replacement**
Recall	0.81	0.85
Precision	0.81	0.85
f1-Score	0.81	0.85
Accuracy	0.809	0.81

Table 6.1: Change in parameters on replacement of Null values with Mean values
Table 6.1 represents the change in model parameters after replacing the null values with mean values.

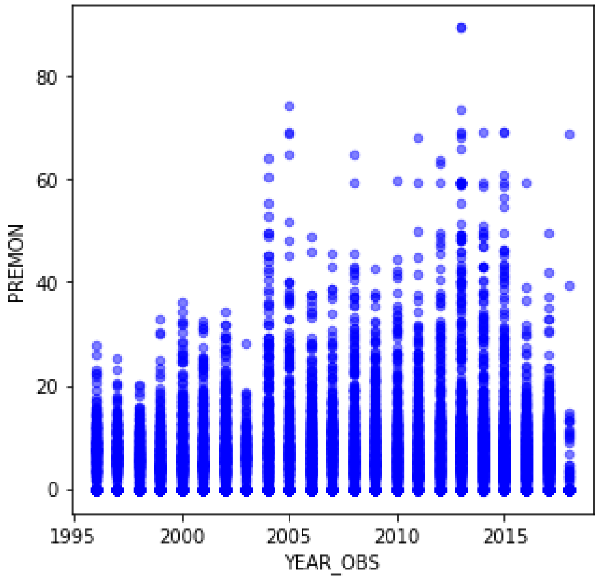


Figure 6.1: Pre-Monsoon Ground water levels year wise
The Figure 6.1 is a year wise scatter plot depicting depletion in groundwater levels during the pre-monsoon season.

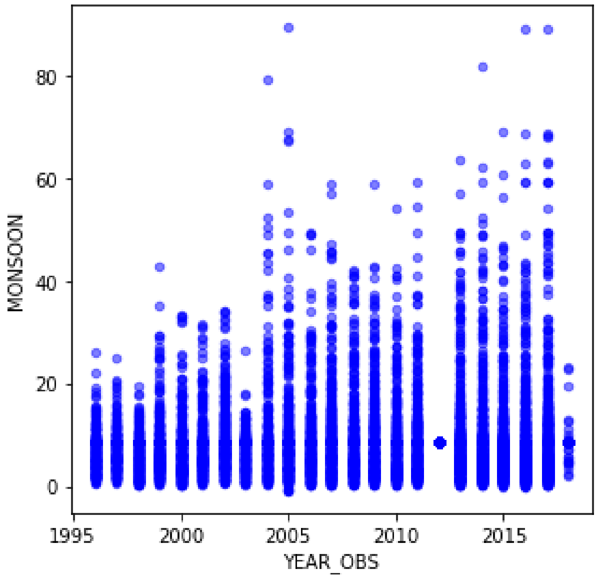


Figure 6.2: Monsoon Ground water levels year wise
The Figure 6.2 is a year wise scatter plot depicting depletion in groundwater levels during the Monsoon season.

An advantage of using gradient boosting is that once the boosted trees are built, importance scores can be determined for each attribute. The feature_importances_ member variable of the trained model stores the importance scores. Importance is determined for every decision tree by the degree that each attribute split point enhances the performance measure. The feature importances are then averaged across all of the decision trees within the model. Feature importance gives a score for each feature of the data where the higher score implies that the feature is more important or relevant towards the output variable.

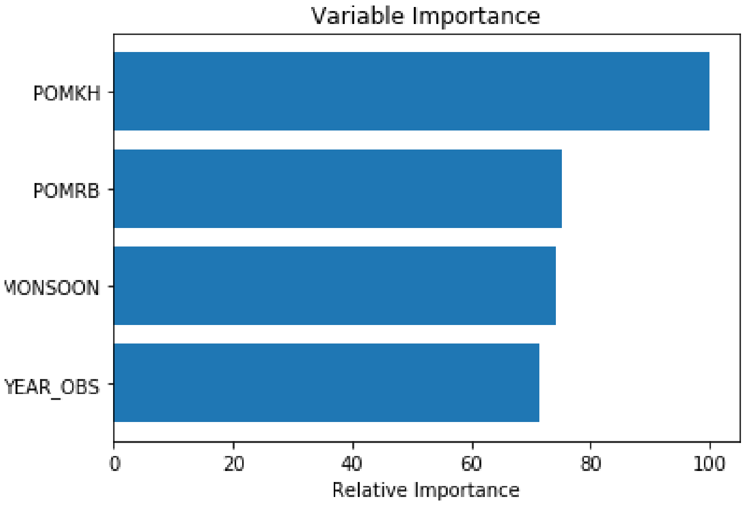


Figure 6.3: Variable Importance
The Figure 6.3 is a plot of variable feature importance.

Figure 6.3 The above plot shows POMKH (post monsoon kharif) has the highest importance. It is the top feature contributing to the predictions of the model and YEAR_OBS(years observed) has the lowest importance. When training a tree we can compute how much each feature contributes to decreasing the weighted impurity, which in case of regression trees is variance.

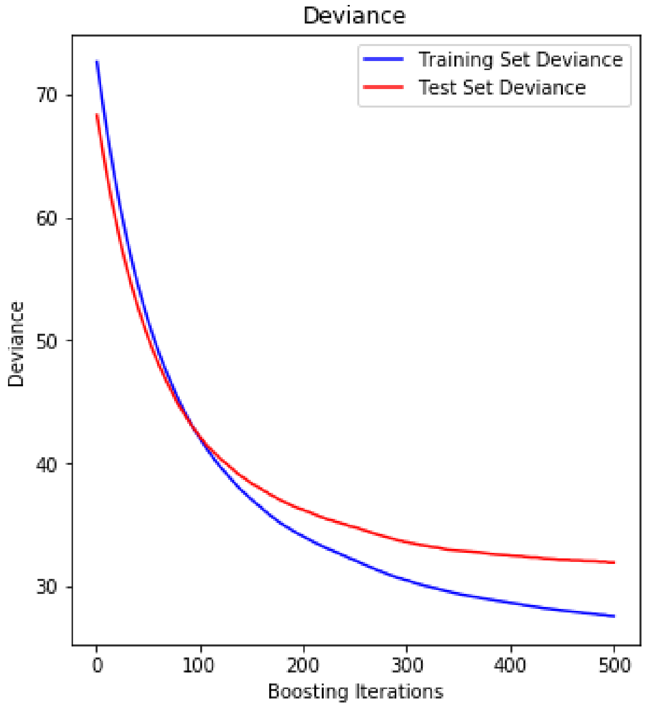


Figure 6.4: Test vs Prediction deviation
The Figure 6.4 is a plot of Test vs Prediction deviation for Gradient Boosting.

On implementing gradient boosting we again have two conditions i.e. with the replacement of null values with mean and without replacement and we find that the MSE from 31.93 to 19.58 when replaced. In gradient boosting, it is found that in each iteration, we fit a base learner to the negative gradient of the loss function and later multiply the prediction with a constant and add it to the value from the previous iteration. From implementing the two algorithms, we observe that the MSE and Accuracy using the gradient boosting yields the leading results for the data set. Table 6.2. shows the Comparison between Random Forest Regression and Gradient Boosting. In the Gradient Boosting algorithm, we also checked for the training and test set deviance, where deviance is the goodness-of-fit statistic for a statistical model and we found that the deviance from the actual value is very minimal as seen in Figure 6.4.

An RNN has also been implemented for prediction of future groundwater levels using the LSTM algorithm, which shows the improvement in the Root Mean Squared Error (RMSE). It is used to compare the predicted values with the existing one is in the data set which is learnt by the Machine Learning model that has been used.

Algorithm	MSE before replacement (out of 100)	MSE after replacement (out of 100)	Increase in accuracy(in percentage)
Random Forest	32.1665	20.3613	5%
Gradient Boosting	31.9301	19.5847	8%

Table 6.2: Comparison between Random Forest and Gradient Boosting
Table 6.2 represents the Comparison between Random Forest and Gradient Boosting

Out of every 100 data values, we find that the RMSE score of the LSTM model for trained data compared to predicted values scored a 7.87, whereas for the tested data compared 5 to predicted values has scored a 13.36, which indicates how close the observed data points are to the model's predicted values. The overall performance of the models is found to be acceptable based on the high correlation efficiency.

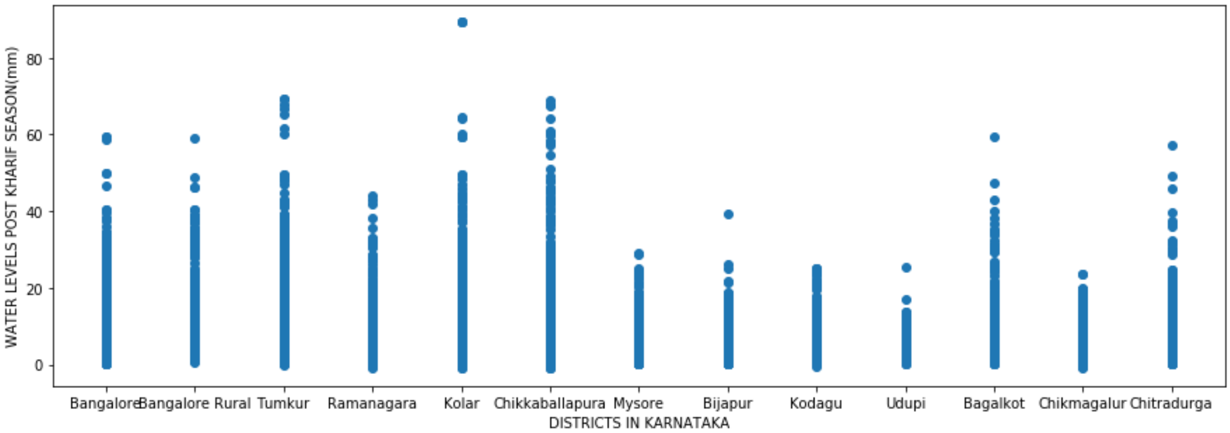


Figure 6.5: Groundwater level distribution in various districts
Figure 6.5 plots the Groundwater level distribution in various districts

The Figure 6.5 is a scatterplot that depicts the groundwater level distribution in the 13 districts : Bangalore, Bangalore rural, Tumkur, Ramanagara, Kolar, Chikkaballapura, Mysore, Bijapur, Kodagu, Udupi, Bagalkot, Chikmagalur and Chitradurga. It can be observed that regions such as Bijapur and Mysore receive lesser rainfall and therefore have lower groundwater levels than when compared to districts such as Kolar and Chikkaballapura. We can therefore conclude that these trends are able to predict when a district is in critical condition with respect to groundwater levels.

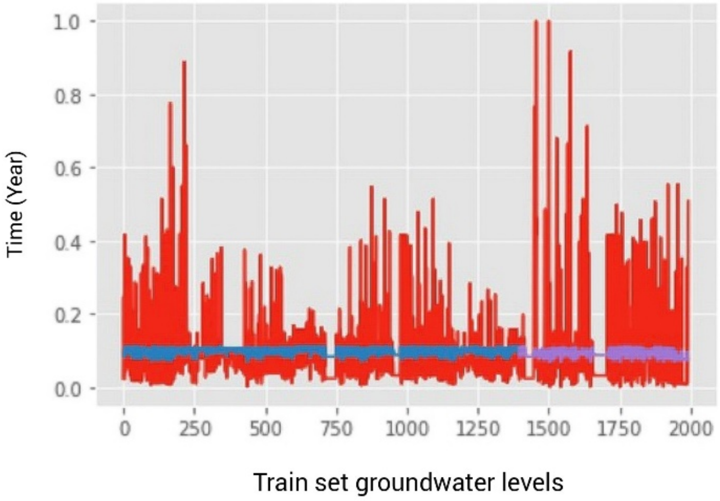


Figure 6.6: LSTM prediction using train set
Figure 6.6 is a plot that describes LSTM prediction using train set values

The Figure 6.6 is a graph comparing the train set values of the dataset (depicted in red) to the future values predicted by the LSTM network (depicted in blue) over the course of a year.

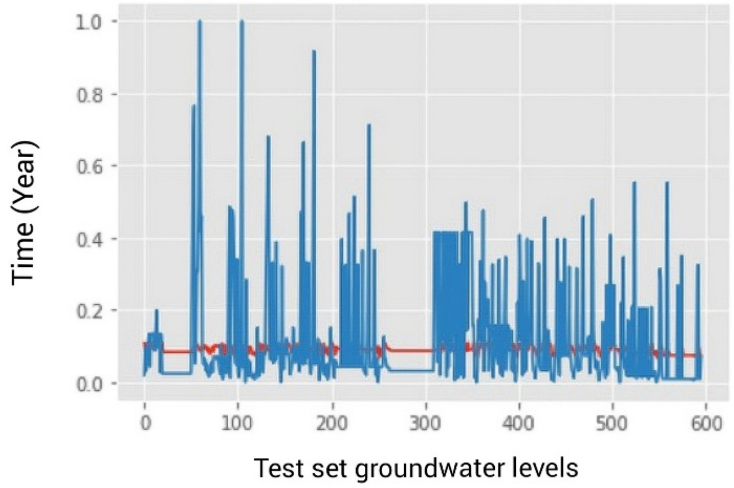


Figure 6.7: Groundwater level distribution in various districts
Figure 6.7 is a plot that describes LSTM prediction using test set values

Similarly, The Figure 6.7 is a graph comparing the test set values of the dataset (depicted in red) to the future values predicted by the LSTM network (depicted in blue) over the course of a year. It can be observed that the performance of the LSTM network is better with respect to test set values than with respect to the train set values. The overall performance of the models is found acceptable built on high correlation efficiency.

Chapter 7

IMPACT OF THE PROJECT TOWARDS SOCIETY AND ENVIRONMENT

The significance of groundwater for the survival of humans cannot be emphasized enough. Groundwater is a major source of drinking water for a majority of the people in India. As the population increases, so does the requirement for water. This has led to water scarcity and is only beginning to get more critical. The inadvisability of the situation has also been aggravated by the problem of water contamination. India is inclining towards a freshwater emergency due to poor administration of water resources and environmental degradation. This crisis is evident in many parts of India, in terms of the scale of the crisis as well as its intensity.

Many approaches and researches have been carried out to improve the conditions of this dreadful situation. A solution to try and predict the water levels is provided in the project. From our developed model water levels of a certain region can be predicted based on the rainfall and also based on the previous year's water levels. An example of this is that it will help the municipal corporation of a city to make decisions on which buildings to be given permission to dig bore wells as it is known that all bores are drying very easily in more recent times. On a more broader aspect, knowing this will help the concerned authorities like the water board and other hydrology experts understand how much longer the water sources can exist in an area before it runs empty. Based on that they can take much needed action to ensure sustainable development and preservation of water.

Chapter 8

CONCLUSION

In this project, we designed a groundwater level forecasting system. The study areas were various aquifers and observational wells in several districts across Karnataka. Three data-driven methodologies are tested based on various Machine Learning algorithms; namely, random forests, gradient boosting, and LSTM. The system is regarding using past groundwater levels data in the pre-monsoon and monsoon seasons as well as the Post Rabi and Post Kharif crop seasons to predict the groundwater levels for the missing values in the dataset, as well as for future usage.

Analysis of the results indicated that the designed gradient boosting model provided a good prediction of groundwater levels, with considerably good accuracy and lower value MSE. In practice, we were able to develop a groundwater level forecasting system using the above mentioned data-driven approaches. The results show that gradient boosting model performed better than random forest model in predicting the groundwater levels. The methodology and findings demonstrated in this project are useful to the currently existing public and private water body organizations as well as the research communities of our nation involved in groundwater management and protection.

Bibliography

[1] Expert System, "What is Machine Learning? A definition" , machine-learning-definition, 6 May 2020

[2] The World Bank, "India Groundwater: a Valuable but Diminishing Resource", india-groundwater-critical-diminishing, 6 March 2012

[3] X. Mao, S. Shang and X. Liu, "Groundwater level predictions using artificial neural networks," in Tsinghua Science and Technology, vol. 7, no. 6, pp. 574-579, Dec. 2002.

[4] Nayak, P.C., Rao, Y.R.S. & Sudheer, K.P. Groundwater Level Forecasting in a Shallow Aquifer Using Artificial Neural Network Approach. Water Resour Manage 20, 77–90 (2006)

[5] Z. Yang, W. Lu, Y. Long and P. Li, "Application of Back-Propagation Artificial Neural Network Models for Prediction of Groundwater Levels: Case study in Western Jilin Province, China," 2008 2nd International Conference on Bioinformatics and Biomedical Engineering, Shanghai, 2008, pp. 3203-3206, doi: 10.1109/ICBBE.2008.1130.

[6] Y. Chen and Y. Li, "GCA-CG Based Groundwater Level Prediction with Uncertainty in Lower Reaches of Tarim River," 2009 International Conference on Energy and Environment Technology, Guilin, Guangxi, 2009, pp. 589-592, doi: 10.1109/ICEET.2009.380.

[7] J. Liu, J. Chang and W. Zhang, "Groundwater Level Dynamic Prediction Based on Chaos Optimization and Support Vector Machine," 2009 Third International Conference on Genetic and Evolutionary Computing, Guilin, 2009, pp. 39-43, doi: 10.1109/WGEC.2009.25.

[8] J. Zhou, Z. Wen and J. Qu, "RBF Neural Network Using Improved Differential Evolution for Groundwater Table Prediction," 2010 International Conference on

Computational Intelligence and Software Engineering, Wuhan, 2010, pp. 1-4, doi: 10.1109/CISE.2010.5676973.

[9] Jihong Qu, Yuepeng Li and Juan Zhou, "Improved differential evolution based BP neural network for prediction of groundwater table," 2010 Third International Symposium on Knowledge Acquisition and Modeling, Wuhan, 2010, pp. 36-39, doi: 10.1109/KAM.2010.5646232. Jihong Qu, Yuepeng Li and Juan Zhou

[10] Ch. Suryanarayana, Ch. Sudheer, Vazeer Mahammood, B.K. Panigrahi, An integrated wavelet-support vector machine for groundwater level prediction in Visakhapatnam, India, Neurocomputing, Volume 145, 2014, Pages 324-335, ISSN 0925-2312

[11] "Temporal Models for Groundwater Level Prediction in Regions of Maharashtra" Dissertation Report - Lalit Kumar(June 2012) Ch. Suryanarayana a,n, Ch.Sudheer b, Vazeer-Mahammood c, B.K.Panigrahi d

[12] Mutao Huang and Yong Tian, "Prediction of Groundwater Level for Sustainable Water Management in an Arid Basin Using Data-driven Models", Proceedings of the 2015 International Conference on Sustainable Energy and Environmental Engineering, 2015, pp. 134-137, 10.2991/seee-15.2015.33

[13] Johnny, Colins & Sashikkumar, M. & Sivadevi, K & Muniraj, Kirubakaran. (2015). Prediction of groundwater level dynamics using Artificial Neural Network.

[14] Kenda, Klemen & Senozetnik, Matej & Klemen, Kristina & Mladenić, Dunja. (2018). Groundwater Modeling with Machine Learning Techniques: Ljubljana polje Aquifer. Proceedings. 2. 10.3390/proceedings2110697.

[15] Hochreiter, Sepp & Schmidhuber, Jürgen. (1997). Long Short-term Memory. Neural computation. 9. 1735-80. 10.1162/neco.1997.9.8.1735.

Appendices

Project Abstracts 2019-20	
1.Department:CSE	
2.Year:2020	**3.Semester : 8**
4. Student Name **5. USN**	1. Aishwarya Kulkarni 1NT16CS007 2. Shivangi Negi 1NT16CS104 3. Sumedha Raghu 1NT16CS117
6. Name of the Project Guide	Dr. Vijaya Shetty
7. Domain	Environment Conservation
8. Title/Topic of the Project	Water Table Analysis using Machine Learning (ML)
9. Abstract(Not more than 100-150words)	
Groundwater has always been the primary water resource in arid and semi-arid areas. Monitoring water table fluctuations is essential for predicting the groundwater levels to outline the future needs. In this study, a detailed analysis is carried out on the prediction of groundwater levels in Karnataka. Nonlinear data-driven models (i.e.; random forest (RF) and gradient boosting (GB)) along with a variant of standard RNNs, Long Short-Term Memory (LSTM) was proposed to predict groundwater level variations. The prediction ability of these models was probed and evaluated using yearly groundwater level data scraped together from observation wells located in various districts across Karnataka in India. The statistical parameters: correlation coefficient (R), Mean Square Error (MSE), accuracy, precision, recall, f1 score, and support were used to assess the performance of these models. These evaluation measures emphasize the capability of these models to keep up with the shift in groundwater levels. The results reveal that our proposed model worked better using the GB algorithm, resulting into good accuracy and less MSE.	

10. Objectives	The main objectives of our project are: 1. Given the groundwater level data for observation wells, to predict the missing groundwater level data. 2. Given the previous years' groundwater level data to predict future groundwater level data. 3. Using different algorithms along with selected groundwater dataset, to achieve more accuracy of prediction than the existing system. 4. Using previous years’ rainfall data along with groundwater level data, to make better predictions.
11. Deliverables	1. Determine groundwater levels in various districts across Karnataka. 2. Analysis of groundwater level data to provide accurate information for groundwater level management. 3. Measurement and analysis of groundwater level helps in maintaining groundwater availability.
12. Status of the Project (Kindly provide correct status of the project)	

Nitte Meenakshi Institute of Technology

(AN AUTONOMOUS INSTITUTION AFFILIATED TO VISVESVARAYA TECHNOLOGICAL UNIVERSITY, BELGAUM)
PB No. 6429, Yelahanka, Bangalore 560-064, Karnataka
Telephone: 080- 22167800, 22167860, Fax: 080 - 22167805

Department of Computer Science and Engineering

a. Publication Status Ongoing **b. Productized** No **c. Patent Filed** No

d. Plagiarism Check Yes **e. Best project status Describe** Done

Signature of the Students

1. Aishwarya Kulkarni : Aishwarya
2. Shivangi Negi :
3. Sumedha Raghu :

Signature of the Guide

Signature of the Head

Water Table Analysis using Machine Learning

Aishwarya Kulkarni[1] ,Shivangi Negi[1],Sumedha Raghu[1] and Dr. Vijaya Shetty[2]

[1] Department of Computer Science and Engineering ,Nitte Meenakshi Institute of Technology

Bangalore 50064,India

[2]Department of Computer Science and Engineering ,Associate Professor,

Nitte Meenakshi Institute of Technology Bangalore 50064,India

email:vijayashetty.s@nmit.ac.in; ⌄ skay4kulkarni@gmail.com;

negishivangi3@gmail.com;sumedharaghu@gmail.com

Abstract. Groundwater is the primary water resource in arid and semi-arid areas. Monitoring water table fluctuations is essential for predicting the groundwater levels to outline the future needs. In this study, a thorough analysis is carried out on the prediction of groundwater levels in Karnataka. Nonlinear data-driven models (i.e.; random forest (RF) and gradient boosting (GB)) along with a variant of standard RNNs (Long Short-Term Memory LSTM) were proposed to predict groundwater level variations. The prediction ability of these models was probed and evaluated using yearly groundwater level data scraped together from observation wells located in various districts of Karnataka in India. The statistical parameters Correlation coefficient (R), Mean Square Error (MSE), precision, recall, f1 score, support, and accuracy were used to evaluate the performance of these models. These evaluation measures highlight the capability of models to keep up with the shift in groundwater levels.

Keywords: groundwater levels; machine learning; data-driven; random forest; gradient boosting; long short-term memory

1 Introduction

Groundwater is characterized as the water present in the sections between the soil pore spaces and within the cleavages of rock formations. The depth at which soil pore fractures and gaps in rock become saturated with water is determined as the water table [1]. In several arid and semi-arid areas, groundwater has emerged as an important source of water required for domestic, irrigation, urban, and industrial activities. India is a substantial consumer of groundwater in the world using 230 cubic kilometres of groundwater per year, which is equivalent to the quarter of the global total [2]. Therefore, sustainable development of groundwater resources is essential for precise quantitative analysis, which is necessary for India due to its prevalent semi-arid and arid climate. Constant monitoring of groundwater levels is important to prevent the misuse of groundwater resources that can usher to local water rationing, limiting in agricultural yields, wells going dried-up or generating unpredictable groundwater quality changes, variations in flow patterns of groundwater emerging in the inflow of meagre quality water and seawater intervention in coastal areas [3]. The water levels, if forecasted prior, might facilitate the executives to plan better, the groundwater usage. In this research, we focus on observation wells from various districts of Karnataka.

Traditionally, process-based models are often employed to perform groundwater simulation and predications, which rely on spatial data of the observed system dynamics. However, they are not suitable in several arid and semi-arid areas as a consequence of insufficient data. Meanwhile, in data-driven modelling with machine learning methodology, our model attempts to establish a direct relationship among the inputs and outputs of the system without having any knowledge about the interior structure of the physical process [4]. The focus here is to use temporal data inputs (historic groundwater level, weather, and rainfall data) to learn the best approximation of the groundwater level values.

Recurrent neural networks (RNNs), a technique of deep learning, are a prominent choice for designing groundwater time series data due to their ability to retain a memory of previous network conditions, but they face challenges in acquiring long term dependencies within variables as weights associated with the network reaches to zero or become exceedingly large during model training [5]. LSTM, a class of RNN can avoid these training problems by eliminating unnecessary information being passed to future model states while retaining a memory of relevant past events [5]. LSTM networks have recently incorporated to model the groundwater table in an inland agricultural area of China on a monthly time step basis [6].

In our study, we are training models based on a set of related attributes to generate optimal predictions for the missing groundwater level values in our dataset using binary classification and later applying data-driven techniques to evaluate the performance of our models. Despite the growing applications of data-driven approaches in surface water problems, there are hardly any studies related to groundwater in arid and semi-arid areas [4]. Therefore, the emphasis of this study is on the implementation of data-driven models with machine learning (i.e., RF AND GB) and deep learning (i.e., LSTM) and comparison of two ensemble methods (i.e., RF and GB) for forecasting groundwater levels in Karnataka, India.

2 Literature Survey

Various works are being implemented in the field of hydrology and groundwater study which also includes prediction analysis. The right form of data extraction is highly necessary for an accurate model [13], A complex web of factors that determines groundwater levels, which are: Rainfall, aquifers, precipitation levels, seasonal changes, groundwater storage patterns, water extraction. As highlighted in [7], [8], [9], Artificial Neural Networks (ANN) is ideal for forecasting based on the implementation of data from wells and shallow aquifers, respectively. Further, different algorithms are analysed [10]. Least Squares Support Vector Machine (LSSVM) was used for dynamic forecasting of wells in Mongolia. The paper [11] helped understand the wavelet integration with the support vector machine (SVM),which is a regression model that is being implemented to check the fluctuations in the water level. In [12], The focus is on quantitative estimates of groundwater temporally and spatially. They have performed a study of groundwater level data in three districts of Maharashtra - Thane, Latur, and Sangli. Analysis of data of above 100 observation wells in each of these districts and developed seasonal models to represent the groundwater behaviour. Three different types of models were developed-periodic, polynomial, and rainfall models. In [13], a detailed survey was conducted concerning the prediction of groundwater levels of Ljubljana polje aquifer. Three different datasets from Ljubljana (Slovenia) and Skiathos (Greece): weather pump sensor data from Skiathos and, information. This paper processed the problem as a regression problem and hence implemented regression trees. They have highlighted the optimization when using ensembles by using random forests. Gradient boosting was also implemented. From the literature survey, neural networks were found to be widely used for predictive analysis. Hence for this Gradient Boosting and Random Forest are the suitable algorithms to be used for predicting the missing values and also compare the two algorithms to see which has better accuracy. LSTM (Long short-term memory) which is RNN (Recurrent neural network) model is considered for predicting future values.

3 System Architecture

The presented system as shown in Fig 1, is developed using machine learning and deep learning application using which the dataset has been generated later analysed. The groundwater dataset is gathered from various aquifers and observational wells of several districts across Karnataka. The dataset which comprises of the multiple parameters of groundwater conditions is split into testing and training data. Firstly, cleaning of data is carried out by alternating the empty cells with the mean set of values. To compare the difference between a dataset with NULL values and that without one, the data is fit into the two algorithms that are Random forest and Gradient boosting. The loss functions MSE (Mean Square Error) is presented on the models. Also, the output obtained is plotted which highlights the most important parameter among all the existing parameters. Further, the LSTM (Long short-term memory network) is applied which is an RNN (Recurrent neural network) model to predict future water level values These results are intended to be given to the respective authority for necessary actions.

Data pre-processing is the approach wherein data is gathered from the specific account using the XLS sheet. Since the gathered data is unstructured and not well defined, the pre-processing techniques are used which include normalization, data cleaning, dimension reduction. This pre-processed data is reviewed and analysed several times by considering all of its parameters before fitting it into the model. Next, the data is subjected to the prediction analyser using the algorithms that include gradient boosting, random forest, and LSTM namely. Finally, the outcome is compared to know how accurate the prediction is, also the result of the data cleaning and the replaced values is represented as graphs.

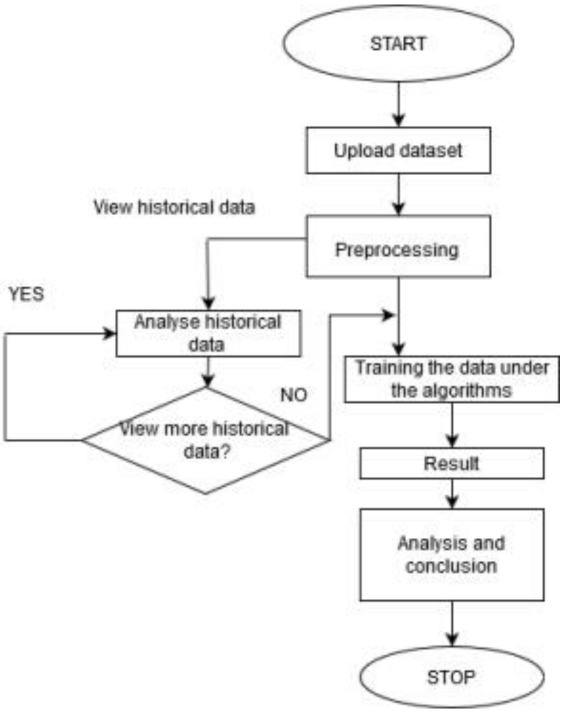


Fig. 1:System Architecture of Groundwater level prediction

4 Dataset

Table 1 represents the attributes in a given dataset. The pre-processing is done on this dataset. The dataset consists of groundwater levels in the pre-monsoon season as well as the Post Rabi and Post Kharif crop seasons. It also consists of location data including the well-code, its district, state, site name, and site type. Fig 2 shows the dataset that has been used.

SITE_TYPE	**WLCODE**	**YEAR_OBS**	**MONSOON**	**POMRB**	**POMKH**	**PREMON**
BORE WELL	W05243	2018	19.63	18.11	NA	NA
DUG WELL	W24336	2018	NA	NA	4.13	3.72
BORE WELL	W05497	2018	23	24.57	NA	NA
BORE WELL	W06424	2018	NA	59.32	NA	NA
BORE WELL	W21201	2018	NA	59.3	NA	NA
BORE WELL	W05727	2018	NA	21.06	37.5	NA
BORE WELL	W05731	2018	NA	7.64	9.1	NA

Fig. 2:Dataset

4.1 Pre-Processing

The data gathered is compiled from various government and water board sources. The dataset initially required a significant amount of pre-processing. The techniques used on the groundwater dataset are data cleaning, data reduction ,and checking feature importance. The dataset compiled contained a large amount of NaN(not a number) values, due to which data was not usable. We first replaced the NaN values with 0s and then used an imputer function to alternate the 0s with the mean value of each column. Certain parameters like well code and site name were then dropped as it was not required for any of the methods employed in this project.

ATTRIBUTE	ATTRIBUTE DESCRIPTION
STATE	Name of the state where the observational well is located
DISTRICT	Name of the district where the observational well is located
TEH_NAME	Name of the tehsil(administrative area) where the observational well is located
BLOCK_NAME	Name of the sub-regions in the district where the observational well is located
LAT	Latitude of the observational well location
LON	Longitude of the observational well location
SITE_NAME	Name of the site of the observational well
SITE_TYPE	Type of the site i.e. bore-well or dug-well
WLCODE	A specific unique number assigned to each observational well
YEAR_OBS	Year during which the observation has been recorded
MONSOON	Groundwater levels during the monsoon
POMRB	Groundwater levels post monsoon during the rabi crop season(October to November)
POMKH	Groundwater levels post monsoon during the kharif crop season(June to October)
PREMON	Groundwater levels before the monsoon season

Table 1: Dataset Description

Then a feature importance was calculated for the four features YEAR_OBS, MONSOON, POMKH AND POMRB that contain groundwater level values. The plot shows POMKH(post monsoon kharif) (Fig 3) has the highest importance. It is the top feature contributing to the model, while (year of observation) has the lowest importance. When training trees one can compute the quantity, the feature contributes to decreasing the weighted impurity, which in case of regression trees is variance.

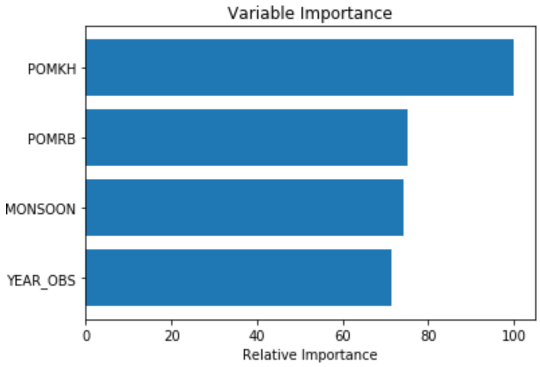


Fig. 3: Feature importance after pre-processing

5 Algorithms

Random Forest :

Random forest is one of the ensemble learning methodology that combines the concepts of classification and regression tasks with the help of multiple decision trees and a technique named bagging (Bootstrap Aggregation) with some additional degree of randomization.

Given a training set Z = z1, ..., zn with responses Y = y1, ..., yn, bagging repeatedly (M times) chooses a random sample with replacement and fits trees to generate new training sets:

For m = 1, ..., M :

- Sample, with replacement, n training examples from Z, Y called Z_m , Y_m
- Train a classification or regression tree f_m on Z_m , Y_m

Bootstrap aggregation uses the following equation to predict unseen samples by averaging prediction from individual regression trees :

$$\hat{f} = \frac{1}{M}\sum_{m=1}^{M} f_{\text{m}}(z') \qquad (1)$$

where $\hat{f}$ represents the prediction for unseen samples.

An individual decision tree is considered as a weak predictor (low bias, high variance) but is fast enough to build. More trees give you a more robust model and prevent overfitting. When modelling, the data is resampled with replacement and for each sampling, a new classifier is trained. A new object type is classified based on new attributes and each tree votes a classification. The forest selects the classifications having the most votes of all the other trees in the forest.

In general, different classifiers overfit the data differently, and via voting, those disparities are averaged out.

Gradient Boosting :

Gradient boosting is a power technique for building predictive models. Gradient Boosting is about taking a model that by itself is a weak predictive model and combining that model with other models of the same type to produce a more accurate model. The idea is to compute a sequence of simple decisions trees, where each successive tree is built for the prediction residuals of the preceding tree.

In gradient boosting the weak model is called a weak learner. The term used when combining various machine learning models is an ensemble. The weak learner in XGBoost is a decision tree. The ensemble works in a forward stage-wise manner by adding classifiers one at a time such that the upcoming classifier is trained to improvise the previously trained ensemble, instituting a weak learner to improve the faults of the already present weak learners.

LSTM :
Long short-term memory (LSTM) units are units of a recurrent neural network (RNN)[5]. An RNN composed of LSTM units is often called an LSTM network. A common LSTM unit is composed of a cell, an input gate, an output gate and a forget gate. The cell remembers values over arbitrary time intervals and the three gates regulate the flow of information into and out of the cell. LSTM networks are well-suited to classifying, processing and making predictions based on time series data, since there can be lags of unknown duration between important events in a time series.

The forward pass of LSTM network is in the below equations:

$$fo_{\mathrm{t}} = \sigma_{\mathrm{g}}(\mathrm{W}_{fo}\, x_t + \mathrm{C}_{\mathrm{fo}}\, \mathrm{h}_{\mathrm{t-1}} + \mathrm{b}_{\mathrm{fo}}) \qquad (2)$$
$$ip_{\mathrm{t}} = \sigma_{\mathrm{g}}(\mathrm{W}_{ip}\, x_t + \mathrm{C}_{\mathrm{ip}}\, \mathrm{h}_{\mathrm{t-1}} + \mathrm{b}_{\mathrm{ip}}) \qquad (3)$$
$$op_{\mathrm{t}} = \sigma_{\mathrm{g}}(\mathrm{W}_{op}\, x_t + \mathrm{C}_{\mathrm{op}}\, \mathrm{h}_{\mathrm{t-1}} + \mathrm{b}_{\mathrm{op}}) \qquad (4)$$

Here fo_{t} , ip_{t} and op_{t} can be described as forget gate, input gate and output gate, respectively. The matrices Wq contains the weight of the input and Cq contains recurrent connection and σ_g is the sigmoid activation function. used in the LSTM network. The network output was calculated by stacking a fully connected layer on top of the LSTM cell. The product of the output layer is the forecast of the groundwater level for the coming season.

6 Methodology

The groundwater dataset initially contained 14 parameters [Table 1], and these parameters had a lot of NaN values. During pre-processing, we replaced the NULL values 0. Using an imputer function, we replaced the 0 valued data with the mean value calculated for each column. The feature importance of the parameters containing numerical values was checked. We applied the random forest and gradient boosting algorithms on the dataset both before and after replacing the 0 valued data with the mean imputation. The accuracies of the two algorithms before and after the replacement of the null values were also calculated. We divided the dataset as a 70 : 30 train-test split and applied the two techniques to determine accuracies, Mean squared errors, Confusion matrices, and classification reports.

For the random forest algorithm, we used the threshold value of 8.19, deliberated using the collective mean of all the groundwater values, to divide the data into 0 and 1 classes where any value below the threshold is classified as 0 and anything above the threshold is classified as 1. This division was created for the test and predicted classes to identify the data relationship between the true-positive, true-negative, false positive 4 and false negative, as shown in Table 2. In the test split that contained 3519 values, the following was observed. We then calculated the accuracy and MSE using the random forest regression algorithm. For gradient boosting algorithm, we made use of gradient boosting regressor function from the python libraries to calculate the accuracy

	TRUE POSITIVE	FALSE NEGATIVE
FALSE NEGATIVE	2001	235
TRUE NEGATIVE	439	844

Table 2 : Confusion matrix from Random Forest Regression

and MSE of both before and after replacing the null values. These algorithms were also used to predict the missing data in the dataset. We then implemented an RNN network, LSTM i.e. Long Short Term Memory, to predict future groundwater levels, and to check the MSE values for training and test values, where we also check the accuracy of the predicted values compared to its test and train set values. By implementing the algorithms, we yield useful and necessary results.

7 Results

On extracting data from various sources, the data is cleaned and pre-processed, and is then fed to the random forest regressor model. To measure the effectiveness of the model the Confusion Matrix [Table 2.] is provided. Let Truly Positive be TF, Truly Negative-TN ,Falsely Positive-FP ,Falsely Negative -FN, Actual results-(TP+FP) and Predicted results- (TP+FN); Based on the same it is found that the Recall(TP / Predicted results), Precision (TP/Actual results) ,f1-score (2*(Precision*Recall)/Precision+Recall) and Accuracy(TP+TN/TP+TN+FN+FP) to have increased from 0.81, 0.81, 0.81, 0.809 to 0.85,0.85, 0.85 and 0.51 respectively after replacing the null values with mean values. Thus, we understand that all the measures of the confusion matrix have been in increased in their values which asserts the replacement.

On implementing gradient boosting we again have two conditions i.e. with the replacement of null values with mean and without replacement and we find that the MSE from 31.93 to 19.58 when replaced. In gradient boosting, it is found that in each iteration, we fit a base learner to the negative gradient of the loss function and later multiply the prediction with a constant and add it to the value from the previous iteration. From implementing the two algorithms, we observe that the MSE and Accuracy using the gradient boosting yields the leading results for the data set. Table 3. shows the Comparison between Random Forest Regression and Gradient Boosting. In the Gradient Boosting algorithm, we also checked for the training and test set deviance, where deviance is the goodness-of-fit statistic for a statistical model and we found that the deviance from the actual value is very minimal as seen in Fig 4.

Algorithm	**MSE before replacement (out of 100)**	**MSE after replacement (out of 100)**	**Increase in accuracy (in percentage)**
Random Forest	32.1665	20.3613	5%
Gradient Boosting	31.9301	19.5847	8%

Table 3: Comparison between Random Forest and Gradient Boosting

We have also implemented an RNN for prediction of future groundwater levels using the LSTM algorithm, which shows the improvement in the Root Mean Squared Error (RMSE). We can compare the predicted values with the existing one is in the data set which is learnt by the Machine Learning model that we have used.

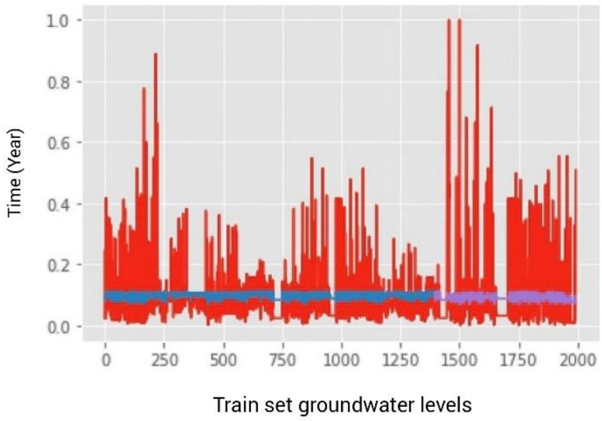


Fig. 6: LSTM prediction using train set

The Fig 6 is a graph comparing the train set values of the dataset (depicted in red) to the future values predicted by the LSTM network (depicted in blue) over the course of a year.

Similarly, The Fig 7 is a graph comparing the test set values of the dataset (depicted in red) to the future values predicted by the LSTM network (depicted in blue) over the course of a year. We can observe that the performance of the LSTM network is better with respect to test set values than with respect to the train set values. The overall performance of the models is found acceptable built on high correlation efficiency.

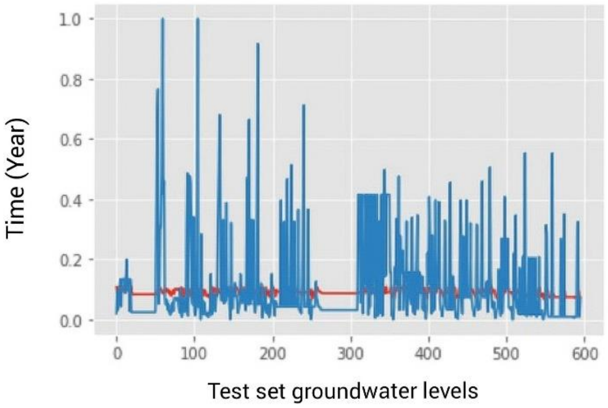


Fig. 7: LSTM prediction using test set

8 Conclusion

In this paper, we propose a groundwater level forecasting system. Three data-driven methodologies are tested based on various Machine Learning algorithms; namely, random forests, gradient boosting, and LSTM. The system is regarding using past groundwater levels data in the pre-monsoon and monsoon seasons as well as the Post Rabi and Post Kharif crop seasons to predict the groundwater levels for the missing values in the dataset, as well as for future usage.

Analysis of the results indicated that the designed gradient boosting model provided a good prediction of groundwater levels, with considerably good accuracy and lower value MSE. The methodology and findings demonstrated in this paper are useful to the research community of our nation involved in groundwater management and protection.

9 References

[1] Richard Greenburg (2005). *The Ocean Moon: Search for an Alien Biosphere*. Springer Praxis Books.

[2] *The World Bank,* India Groundwater: a Valuable but Diminishing Resource, *india-groundwater-critical-diminishing,* 6 March 2012

[3] " *Groundwater Governance.* "Thematic Papers on Groundwater". April 2016. PDF File.

[4] Mutao Huang and Yong Tian, "Prediction of Groundwater Level for Sustainable Water Management in an Arid Basin Using Data-driven Models", Proceedings of the 2015 International Conference on Sustainable Energy and Environmental Engineering, 2015, pp. 134-137, 10.2991/seee-15.2015.33

[5] Hochreiter, Sepp & Schmidhuber, Jürgen. (1997). Long Short-term Memory. Neural computation. 9. 1735-80. 10.1162/neco.1997.9.8.1735

[6] Zhang, Jianfeng & Zhu, Yan & Zhang, Xiaoping & Ye, Ming & Yang, Jinzhong. (2018). Developing a Long Short-Term Memory (LSTM) based Model for Predicting Water Table Depth in Agricultural Areas. Journal of Hydrology. 561. 10.1016/j.jhydrol.2018.04.065.

[7] X. Mao, S. Shang and X. Liu, "Groundwater level predictions using artificial neural networks," in Tsinghua Science and Technology, vol. 7, no. 6, pp. 574-579, Dec. 2002.

[8] Nayak, P.C., Rao, Y.R.S. & Sudheer, K.P. Groundwater Level Forecasting in a Shallow Aquifer Using Artificial Neural Network Approach. *Water Resour Manage* 20, 77–90 (2006)

[9] Z. Yang, W. Lu, Y. Long and P. Li, "Application of Back-Propagation Artificial Neural Network Models for Prediction of Groundwater Levels: Case study in Western Jilin Province, China," 2008 2nd International Conference on Bioinformatics and Biomedical Engineering, Shanghai, 2008, pp. 3203-3206, doi: 10.1109/ICBBE.2008.1130.

[10] J. Liu, J. Chang and W. Zhang, "Groundwater Level Dynamic Prediction Based on Chaos Optimization and Support Vector Machine," 2009 Third International Conference on Genetic and Evolutionary Computing, Guilin, 2009, pp. 39-43, doi: 10.1109/WGEC.2009.25.

[11] Ch. Suryanarayana, Ch. Sudheer, Vazeer Mahammood, B.K. Panigrahi, An integrated wavelet-support vector machine for groundwater level prediction in Visakhapatnam, India, Neurocomputing, Volume 145, 2014, Pages 324-335, ISSN 0925-2312

[12] "Temporal Models for Groundwater Level Prediction in Regions of Maharashtra"

Dissertation Report - Lalit Kumar(June 2012)

Ch. Suryanarayana a,n, Ch.Sudheer b, VazeerMahammood c, B.K.Panigrahi d

[13] Kenda, Klemen & Senozetnik, Matej & Klemen, Kristina & Mladenić, Dunja. (2018). Groundwater Modeling with Machine Learning Techniques: Ljubljana polje Aquifer. Proceedings. 2. 10.3390/proceedings2110697.

Vijaya Shetty S

ORIGINALITY REPORT

16%	10%	8%	10%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	www.mdpi.com Internet Source	2%
2	Submitted to Universiti Malaysia Pahang Student Paper	2%
3	Submitted to University of Florida Student Paper	2%
4	Klemen Kenda, Matej Čerin, Mark Bogataj, Matej Senožetnik, Kristina Klemen, Petra Pergar, Chrysi Laspidou, Dunja Mladenić. "Groundwater Modeling with Machine Learning Techniques: Ljubljana polje Aquifer", Proceedings, 2018 Publication	2%
5	www.datasciencecentral.com Internet Source	1%
6	www.newgenapps.com Internet Source	1%
7	www.cse.iitb.ac.in Internet Source	1%

8	Submitted to Imperial College of Science, Technology and Medicine Student Paper	1%
9	www.csauthors.net Internet Source	1%
10	Submitted to Macquarie University Student Paper	<1%
11	Stefan Ungureanu, Vasile Topa, Andrei Cziker. "Industrial load forecasting using machine learning in the context of smart grid", 2019 54th International Universities Power Engineering Conference (UPEC), 2019 Publication	<1%
12	S Vijava Shetty, G A Karthik, M Ashwin. "Symptom Based Health Prediction using Data Mining", 2019 International Conference on Communication and Electronics Systems (ICCES), 2019 Publication	<1%
13	Fan Zhang. "A hybrid structured deep neural network with Word2Vec for construction accident causes classification", International Journal of Construction Management, 2019 Publication	<1%
14	Submitted to National College of Ireland Student Paper	<1%

15	Submitted to University of Dallas Student Paper	<1%
16	Submitted to University of Sheffield Student Paper	<1%
17	Submitted to Aliah University Student Paper	<1%
18	Submitted to Stefan cel Mare University of Suceava Student Paper	<1%
19	Mohammad Taghi Sattari, Rasoul Mirabbasi, Reza Shamsi Sushab, John Abraham. "Prediction of Groundwater Level in Ardebil Plain Using Support Vector Regression and M5 Tree Model", Groundwater, 2018 Publication	<1%

Exclude quotes	On	Exclude matches	< 10 words
Exclude bibliography	On		

Nitte Meenakshi Institute of Technology

(AN AUTONOMOUS INSTITUTION AFFILIATED TO VISVESVARAYA TECHNOLOGICAL UNIVERSITY, BELGAUM)
PB No. 6429, Yelahanka, Bangalore 560-064, Karnataka
Telephone: 080- 22167800, 22167860, Fax: 080 - 22167805

Department of Computer Science and Engineering

Couse name: Major Project

Course Code: 14CSP84

Project Title: Water table analysis using machine learning

Self-Assessment of PO-PSO Attainment. – Project Work.

Programme Outcomes (PO)	**Justification**
PO1. Engineering knowledge:	Machine learning technology
PO2. Problem analysis:	Determine the ground water levels in various districts of Karnataka
PO3. Design/development of solutions:	Build a machine learning model using data driven techniques
PO4. Conduct investigations of complex problems:	1. Designing the dataset from the data collected via various sources. 2. Removing null values from the generated Datasets. 3. Applying different methodologies to predict Groundwater levels.
PO5. Modern tool usage:	Anaconda, Spyder, various python libraries
PO6. The engineer and society:	Helped contribute what we have learnt in our engineering to the society
PO7. Environment and sustainability:	Will contribute to sustainable development of water levels and monitoring digging of bore wells
PO8. Ethics:	Responsibility, respecting each other, honesty
PO9. Individual and team work:	Individually all 3 of us divided the literature survey and the understanding of three algorithms equally (later explained to each other). The implementation, presentation and word report as a team effort
PO10. Communication:	Updated each other through face to face discussion, mails, WhatsApp and hangouts.

PO11. Project management and finance:	Project was completed with help and guidance of our guide
PO12. Life-long learning:	Importance of Team work, patience and to be hard and smart working

Programme Specific Outcomes (PSO)	**Justification**
PSO 1: Professional Skills	Teamwork, time management, project management
PSO 2: Problem Solving Skills	researching, analytical, decision making
PSO 3: Ethics and Career Development	Honesty, patience, confidence, improved oral skills

Signature of the Guide