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	
SHIVANGI NEGI	1NT16CS104	
SUMEDHA RAGHU	1NT16CS117	

Date:

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	
SHIVANGI NEGI	1NT16CS104	
SUMEDHA RAGHU	1NT16CS117	

Date:

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 rock formations. A chunk of rock or an unconsolidated deposit that can yield a usable quantity of water is called an aquifer. 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 kilometers 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]. Groundwater reservoir is a complicated system that is faced with either natural or unnatural factors caused by human activity. For the prediction of groundwater level fluctuation for different natural conditions and usage rates are of great importance for the use and management 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.

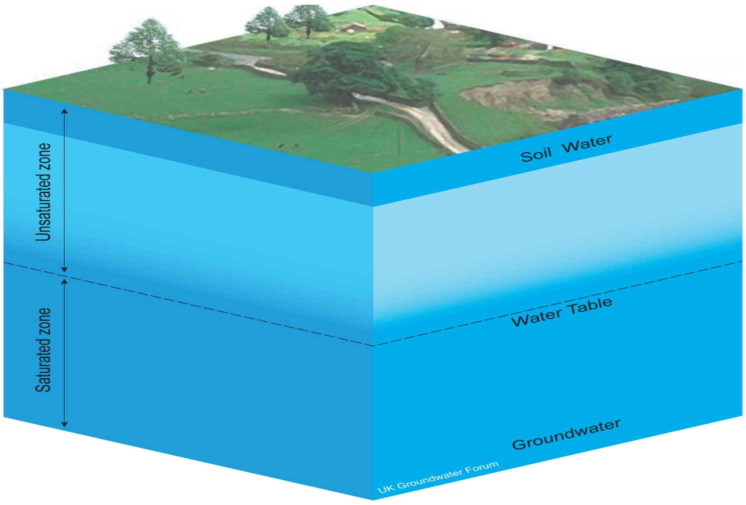


Figure 1.1: Groundwater

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

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 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 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.

Brief History of Technology/Concept

Machine learning (ML) is a class of algorithms that permits software applications to become more accurate in predicting outcomes while not being explicitly programmed[1]. Machine learning focuses on the development of computer programs that will access data and use it to learn for themselves. The main objective is to allow the computers to learn automatically without human intervention and adjust actions accordingly. Machine learning algorithms are usually classified as supervised or unsupervised.

- **Supervised machine learning algorithms**are applied to what has been learned in the past to new data using labelled examples to predict future events. The system can provide target values for any new input after enough training. The learning algorithm can also compare its output with the intended output and find errors in order to modify the model accordingly.

- **Unsupervised machine learning algorithms** are used when the knowledge accustomed to train is neither classified nor labelled. Unsupervised learning analyses how systems can infer a function to describe a hidden structure from unlabelled data. The system explores the data and can draw inferences from datasets to describe hidden structures from unlabelled data.

- **Semi-supervised machine learning algorithms** fall in between supervised and unsupervised learning, since they use both labelled and unlabelled data for training – typically a small amount of labelled data and a large amount of unlabelled data.

The systems that use this method can improve learning accuracy. Usually, semi-supervised learning is chosen when the acquired labelled data requires skilled and relevant resources in order to train it to learn from it.

- **Reinforcement machine learning algorithms** are a learning methodology that interacts with its surroundings by producing actions and discovers errors or rewards. Trial and error search and delayed reward are the foremost relevant features of reinforcement learning. This methodology permits machines to automatically determine the ideal behaviour within a particular context so as to maximize its performance.

Most commonly used machine learning models are:

- **Regression** : In machine learning, this model is used to predict the outcome of an event based on the relationship between variables obtained from the dataset.
- **Classification** : It is a supervised learning approach in which the computer program learns from the data input given to it and then uses this learning to classify the output into categorical form.
- **Decision trees** : These models use observations about certain actions and identify an optimal path for arriving at a desired outcome.
- **K-means clustering** : This model groups a specified number of data points into a specific number of groupings based on similar characteristics.
- **Neural networks** : These deep learning models utilize large amounts of training data to identify correlations between many variables to learn to process incoming data in the future.For example:Artificial Neural Networks(ANN).

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 an essential component of suitable water resources management. Water table forecasting plays an important role in the management of groundwater resources in agricultural regions where there are drainage systems in river valleys.

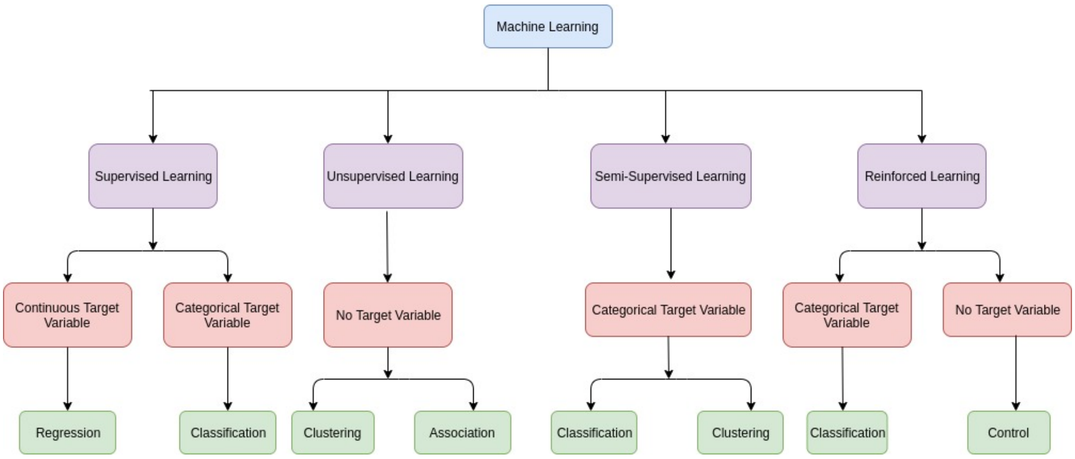


Figure 1.2: Groupings of Machine Learning
Figure 1.2 depicts the various modules that fall under the broad classiffication of Machine Learning

Such forecasts may be useful in coastal regions in planning the use of groundwater and surface water that help maintain the natural water table gradient to protect seawater intrusion or water logging condition.

Research Motivation and Problem Statement

Research Motivation

Groundwater levels are an indicator for groundwater present in aquifers and shallower water tables. Groundwater decline is a real-world serious problem in many parts of the Nation and the world. The water levels in aquifers is often not constant. Groundwater levels rely on recharge from infiltration of precipitation so when a drought hits the land surface or if exploitation takes place it can impact the water levels below ground.

Contributions of groundwater to streams, rivers, lakes and wetlands play a crucial role in maintaining surface water quantity, quality, temperature and all vitals that maintain the health of aquatic ecosystems. 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

The primary object of our project is to model ground water table using regression trees, random forest, SVR based on historical data and also to predict the water level fluctuations in the nearby areas using ANN techniques.

Research Objectives

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. Using previous years rainfall data along with groundwater level data, to make better predictions.
3. Given the previous years' groundwater level data to predict future groundwater level data.
4. Using different algorithms along with selected groundwater dataset, to achieve more accuracy of prediction than the existing system.

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 it 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 valuable nature of the resources, 29% of groundwater blocks are semi-critical, critical, or overexploited, and the situation is deteriorating rapidly (2004 nationwide assessment.) Moreover, aquifers are depleting in the most populated and economically productive areas. Climate change will further strain groundwater resources[2]. Monitoring water table fluctuations is essential and it is even more important to predict the groundwater level to plan for the future 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], Artificial Neural Networks (ANN) were used to predict the groundwater level in the Dawu Aquifer of Zibo in Eastern China. 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 49 X 104 m3, 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 analysis 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 takes into account 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, with the earlier data for training and the later data for testing. 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 took into account 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. Both the training results and the testing results agreed well with the observed data, which indicates that the model can effectively describe the relationship between the groundwater level fluctuations and the main influencing factors.

In [4], This study forms a part of the river Godavari delta system in East Godavari district of Andhra Pradesh in India. Geographically, Central Godavari Delta, is located between 16.25 N to 16.55 N latitude and 81.44E to 82.15 E longitude with its hydrological boundaries as the river Gowthami. The ANN models purpose here will be to forecast the water levels up to 4 months in advance. These forecasts might be useful in combined use planning of groundwater and surface water in the coastal areas that help maintain the natural water table gradient to protect intrusion of seawater or water logging condition. It is extremely important to understand the spatial and temporal variations of the water level for the management of same in coastal areas. ANNs have been proven to be effective in modeling virtually any nonlinear function to an arbitrary degree of accuracy.

The project employs a standard back propagation algorithm for training, and the number of hidden neurons is optimized by trial and error procedure. It started with two hidden neurons initially, and the number of hidden neurons increased up to 10 with a step size of 1 in each trial. For each set of hidden neurons, the network has to be trained in batch mode inorder to minimise the mean square error. In order to check any over-fitting during training, a cross validation was performed by keeping track of the efficiency of the fitted model. The training is stopped when there was no significant improvement in efficiency, and the model was then tested for its generalization properties. 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 network after standardization (subtracting monthly mean and dividing it by the standard deviation of the corresponding month) to remove the cyclicity or periodicity in the data. The scalability is limited between 0 and 1as the activation function warrants. The total available data has been divided into two sets, calibration and validation set: the model is trained using data for 6 years (1981–1986) and validated on the rest of the data (1987–1989). The final structure of the ANN model for Munganda observation well is: 8 input neurons, 3 hidden neurons and 1 output neuron and for Cheyyeru observation wellis: 10 input neurons, 2 hidden neurons and 1 output neuron. The resulting waterlevel plots obtained from the model are analyzed statistically using different indices employed for performance analysis of models. The goodness of fit statistics considered are the root mean square error (RMSE) between the computed and observed runoff, coefficient of correlation (CORR), average absolute relative error (AARE) and percentage error in deepest level estimation (%EDLF). The analysis shows that the water level at Munganda well at any time period has a significant correlation with the water level at Kattunga well at a lag of 1 time step (month). The difference between the observed and predicted water levels, is used for assessment of the model developed and is presented for Munganda and Cheyyeru. This error plot helps understanding whether the model is predicting the increase or decrease in levels. Although good results are obtained or Munganda observation well up to 4 months ahead forecasts, the model performance is found 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 such as RMSE(Root mean squared errors), MAE(Mean arithmetic error) and Coefficient of efficiency R^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 BP algorithm and is propagated back to the input layer through the hidden layer. The lower the values of RMSE, MAE and R^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 commonly used forecasting method. An attempt was made in this study to investigate the use of an improved BPANN model for prediction of groundwater levels. The results of the study suggest that the BPANN model is reliable for modeling of groundwater levels for forecasting purpose in this area.

The levels estimated, provided the corresponding salinity data, was accurate enough for most applications. According to the authors Yue Chen and Yuhong Li, in the paper [6], the lower reaches of Tarim River is taken as the study area. They have proposed a grey correlation analysis and cloud generator (GCA-CG) based groundwater level prediction model. The most important characteristic feature of the model used here is the observation data that contains uncertainty is taken into account. Grey relational analysis (GRA), which is also called as Deng's Grey Incidence Analysis model, was developed by a Chinese Professor Julong Deng of Huazhong University of Science and Technology. A grey system means that a system in which part of information is known and part of the information is unknown. Since uncertainty always exists, one is always somewhere in the middle, somewhere between the extremes, somewhere in the grey area.

Grey analysis then provides a 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 fluctuations with large rises and falls without many typical distribution patterns. Applying statistical methods was shown to not achieve many useful conclusions. The grey correlation analysis remedied this defect found in existing statistics when it was applied in the content of systems analysis. It was applied to cases of various sample sizes and distributions with a relatively small amount of computation. The characteristics used for this model were very useful to generate uncertainty estimates of groundwater levels, even when limited information about the soil is available. Although prediction errors and were often large, the groundwater levels predicted were highly accurate.

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 sample, 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 analysed 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.

Considering groundwater level dynamic series length and peak mutation characters, the least squares support vector machine arithmetic based on peak value identification was proposed. Aiming at parameter optimization, training and speed test of support vector machine arithmetic, a least square support vector machine groundwater level dynamic forecasting model based on chaos optimization peak value identification was established. [7]

In [8], Radial basis function (RBF) neural network is used to predict groundwater table levels, which often shows complex nonlinear characteristic. The traditional RBF training algorithm based on gradient descent optimization method can only obtain the partial/local optimums solution. Furthermore, humans selecting the structure of RBF neural network leads to blindness and expends much time. Therefore differential evolution (DE) algorithm was adopted to automatically search the weight of output layer, the center of RBF and the network width. In order to improve the population's diversity and the ability of escaping from the local optimum, a self-adapting crossover probability factor was presented. Furthermore, a chaotic sequence based on logistic map was employed to self-adaptively adjust mutation factor, which can improve the convergence of DE algorithm. A radial basis function (RBF) is a real-valued function whose value depends only on the distance between the input and some fixed point known as the centre 'c'. The sums of radial basis functions are used to approximate given functions. This process can also be interpreted as a simple kind of neural network. Differential Evolution (DE) is a population-based optimizer that generates perturbations given the current generation. Instead of generating vectors using samples from a predefined probability functions, DE perturbs vectors using the scaled difference of two randomly population vectors. Differential Evolution produces a trial vector which competes against population vector of the same index. Once all the trial vector have been tested, the survivors of the pairwise competitions become the parents for the next generation in the evolutionary cycle.

A chaotic sequence based on logistic map has been employed to self-adaptively adjust mutation factor, which can improve the convergence of DE algorithm. Compared with the traditional RBF neural network, DE trained RBF neural network is shown to be more robust and can greatly improve the convergence speed and precision of groundwater table prediction.

According to the paper by authors Jihong Qu, Yuepeng Li and Juan Zhou, Groundwater table often shows complex nonlinear characteristic. Back Propagation (BP) neural network is increasingly used to predict groundwater table. Manual selection of the structure of BP neural network has blindness and expends much time, so differential evolution (DE) algorithm was adopted to automatically search BP neural network weight matrix and threshold matrix[9]. This paper combines the methods of Differential Evolution and Backpropagation methods to predict the ground water table levels.

In the neural network that makes use of the Backpropagation algorithm, each neuron of the network operates by taking the sum of its weighted inputs and passing the result through a nonlinear activation function.

This is done to capture the errors, which are fed back as the input every iteration. This is done to minimize the errors and improve the accuracy of the model. The BP algorithm is successful but it is shown to have some disadvantages. The selection of the learning factor and inertial factor affects the convergence of the BP neural network which is determined by experience. Man-made selection of the structure is said to have blindness and expends much time. DE has been successfully applied to the optimization including non-linear models. In a population of potential solutions to an optimization problem within an n-dimensional search space, a fixed number of vectors are randomly initialized, then evolved overtime to explore the search space and to locate the minima of the objective function[9]. A self-adapting crossover probability factor is shown to improve the population's diversity and the ability of escaping from the local optimum.

Compared to the traditional BPANN, DE trained BP neural network is said to be more robust, and improves the convergence speed and precision of groundwater table prediction.

In [10], an attempt is made to predict monthly groundwater level fluctuations using integrated wavelet and support vector machine modeling. The discrete wavelet transform with two coefficients (db2 wavelet) is adopted for decomposing the input data into wavelet series. These series are further used as input variables in different combinations for Support Vector Regression(SVR) model to forecast groundwater level fluctuations. Wavelets are a mathematical expression which decomposes the original time series into various components. The wavelet components thus obtained are very helpful for improving the forecasting capability of a model by capturing useful information at various levels. Wavelet Transforms Proved To Perform Better Compared to the traditional Fourier transforms. In this study wavelet analysis is used to decompose the time series of groundwater depths into various components. The decomposed components are thus used as inputs for the SVR model. WA-SVR model has been implemented which is a combination of wavelet and support vector regression to predict the groundwater level variations for three observation wells in the city of Visakhapatnam,India. According to the paper the prediction of groundwater is very complex and highly nonlinear in nature as it depends upon many complex factors such as precipitation, evapotranspiration, soil characteristics and topography of the watershed. It is observed that the WA-SVR model is able to capture the underlying dynamics of the groundwater level variations and forecast even when there is a sudden change in groundwater levels in the consecutive months. Further the WA-SVR model is able to forecast accurately the higher and lower peaks groundwater levels in the testing period when compared to the other models. The forecast performance is evaluated using the Normalized Mean Square Error (NMSE), Root Mean Square Error(RMSE), Mean Absolute Percentage Error (MAPE), Nash-Sutcliffe Efficiency Coefficient (Ec) and Correlation Coefficient (R2).

The performance of the WA-SVR model is compared with SVR, ANN and also with the traditional Auto Regressive Integrated Moving Average(ARIMA) models. LIBSVM toolbox is used to develop the SVR models for predicting the groundwater level. ANN's used the Levenberg–Marquardt(LM) algorithm for training as it is the fastest algorithm. SPSS 13 is used in developing the ARIMA model for predicting the groundwater levels.

For the Sivajipalem, Madhurawada and Gullalapalem observation wells, the maximum error in the best WA-SVR model are 17%, 50% and 7% and for the best SVR model are 27%, 133% and 10% and for the best ANN model are 33%, 118% and 12% and for the best ARIMA model are 71%, 167% and 26% respectively. Thus, based on the performance criteria and maximum error in predictions obtained for the study area it can be concluded that the WA-SVR model is a superior alternative to SVR, ANN and ARIMA models to forecast the Groundwater levels. It also mentions on how a multivariate time series analysis can be done considering either by adding effective D series and A series or separately giving D's and A's as the input to the SVR model.

In [11], The focus on quantitative estimates of groundwater temporally and spatially. Data set is taken from Groundwater Survey and Development Agency (GSDA), an agency of Government of Maharashtra established in 1972. They have performed analysis of groundwater level data in three districts of Maharashtra - Thane, Latur and Sangli. Analysis for data of more than 100 observation wells in each of these districts and developed seasonal models to represent the groundwater behavior. Three different type of models were developed-periodic, polynomial and rainfall models. While periodic and polynomial models capture trends on water levels in observation wells, the rainfall model explores the correlation between the rainfall levels and water levels. The periodic and polynomial models are developed only using the groundwater level data of observation wells while the rainfall model also uses the rainfall data. It also explores how over extraction can also cause problems such as sinking of land and water quality issues such as fluoride and arsenic.

To make a yearly model the data of over 30 years is fold into a single year. The groundwater data set is fit into periodic, linear interpolation, spline interpolation and polynomial models. For developing initial polynomial models he has only used observation well data. These were the time stationary model. In the next model i.e. the rainfall model, along with observation well data, the rainfall data for developing the models has also been used. These models are developed using only those year data in which both rainfall data and observation well data is available. Periodic models developed with original point shows unpredictable behaviour at times when no data is present. Modeling from January to December is a bad choice of model, as modelling the sudden rise in water level after monsoon is difficult. Using synthetic data generated to overcome the problem of sparsity does not help much and moreover is not a true reflection of the water levels. After the heavy rains in June-September it is expected the water levels in observation wells to drop continuously till the next monsoon. In polynomial the monotonically decreasing behaviour of water level in an observation well is captured using the observation well data. The comparison is done with periodic model for Thane with it being the only periodic model and is developed using only the original points (i.e. no interpolated points are used), as the other periodic models are developed using synthetic data. The comparison of R2 values should be made in cases when degree of freedom is the same for both the models. On comparing R2 values of periodic model with function F1(x) to R2 values of polynomial model with degree 2 (both have 3 degrees of freedom) the following points are observed :

- The R2 value increases for all bore wells in case of polynomial model Out of 92 dug wells the R2 value increases in 70 dug wells in case of polynomial model. From these observations we conclude that polynomial model are better than the periodic model. For developing the rainfall model, observation well data is taken as degree 2 polynomial and rainfall as linear. The observation well data is folded into single year and mapped to June-May period as done in the case of polynomial model. It is concluded that using the rainfall for the year again improves the quality of fit.

- On an average, the increase in R2 value is 0.12-0.16 in case of Latur and Sangli, but only 0.02-0.05 in case of Thane. This led to conclude that rain has a more prominent role in Latur and Sangli and not so much in Thane. Currently, models are developed using rainfall from current year, but previous year rain may also have influence on groundwater levels. In future this aspect can be explored to see if including previous years rain have an impact or not. Instead of using rainfall, use infiltration to model groundwater, but this would require accurate estimate of rainfall, evapotranspiration, runoff which is not easily available.

In [12], The data-driven models attempt to identify a direct mapping between the inputs and outputs of the system. The focus of this study is the application and comparison of three data driven models (i.e., ANN, SVM and M5 model tree) for forecasting short-term groundwater levels in the Shule river basin situated in Gansu province, China. The Shule River basin is one of the extremely dry regions in China with the characteristics of very arid continental climate, low precipitation, low runoff coefficient, sandstorm, and high evaporation capacity. The ANN architecture that is used for regression or prediction is the Multi-Layer Perceptron (MLP) network which has a layered architecture. A layer typically consists of a number of neurons. Directed synapses connect each neuron in one layer to every neuron in the next layer. Each synapse is assigned with a weight. The information regarding the behavior of training set data is stored in terms of synapses weights. The sigmoid activation function was used for both the hidden and output neurons. Support Vector Regression (SVR) is used to describe regression with SVM. Given the training set, SVR maps each input into a high dimensional feature space via a nonlinear function and then performs a linear regression in this feature space to find a function that can best approximate the actual output value with an error tolerance. The M5 model tree algorithm technique combines a conventional decision tree with the possibility of generating linear regression functions at the 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.

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 thorough analysis is conducted concerning the prediction of groundwater levels of Ljubljana polje aquifer. Machine learning methodologies are implemented using strongly correlated physical parameters as input variables. The goal in such a scenario would be to predict groundwater levels based on temporal data inputs (historic groundwater and surface water level data, weather data and forecasts,land-use, groundwater withdrawal and other anthropogenic data) and outputs (groundwater level). Three different data sets from Ljubljana (Slovenia) and Skiathos (Greece): Groundwater information, pump sensor data from Skiathos and weather data. Groundwater dataset contains data from 518 stations comprising 28 regions which measure groundwater levels. Data are collected since 1960 (with median frequency of one day); however, there are some stations that started operating later, or operated intermittently, which means some data are not available. Weather data are refreshed once per day and contain temperatures (daily average, minima and maxima), location data, precipitation, snow blanket, new snow blanket, cloud cover and sun duration.

Work involves training a model that will be able to predict continuous groundwater level values based on a set of related attributes (weather data, available historic values of groundwater levels, etc.). Such a problem belongs to the field of supervised learning, more specifically—it is a regression problem.

The initial experiments aimed at predicting absolute value of groundwater levels. This proved to be an inefficiently defined problem, since absolute water level depends strongly on long-term historic processes, which we cannot easily grasp with limited attribute vectors. Therefore, target value is not absolute groundwater level, but rather the changes in groundwater levels. Groundwater level prediction is a regression problem. Based on available data (i.e., weather data, weather predictions, people-behavior model prediction, etc.) the algorithm is trying to generate the best possible continuous predictions for groundwater level change on a particular day. Regression trees are an algorithm based on decision trees. When learning,the attribute space is segmented into many different subspaces, where each particular subspace represented by a tree leaf has a value, which might be obtained simply by averaging all the samples from training set that belong to that leaf or by introducing another model (often linear regression) at each final node.

Regression trees work well in ensembles. Each regression tree is trained with a particular sub-sample of a data set (different attributes and different samples from original dataset are used). Final value is given by an average over the whole ensemble. Gradient boosting also utilizes an ensemble of weak learners (usually trees) to provide final prediction, but it stacks them additively. In the first stage the algorithm approximates target values. Random forests algorithm is also implemented. All models tested are multivariate and the predictors inserted are: cloud cover, snow blanket, new snow blanket, precipitation, sunlight duration, and average, maximum and minimum temperatures. A variable delay between groundwater change and the predictors was introduced to simulate the actual dynamics of aquifer recharge. Cloud cover and precipitation were chosen to better predict the target attribute. Gradient boosting resulted as the best fitting method with R2 ¯ 0.644 and RMSE ¯ 2.11 times (10 minus 4).

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 of the Python and R programming languages for scientific computing (data science, machinelearning applications, large-scale data processing ,predictive analytics, etc.), that aims to simplify package management and deployment. Package versions are managed by the package management systemconda. Navigator can search for packages on Anaconda Cloud or in a local Anaconda Repository, install them in an environment, run the packages and update them. It is available for Windows, mac OS and Linux.

The following applications are available by default in Navigator :

- JupyterLab
- Jupyter Notebook
- QtConsole
- Spyder
- Glueviz
- Orange
- Rstudio
- Visual Studio Code

TensorFlow Library

TensorFlow is a free and open-source software library for dataflow and differentiable programming across a range of tasks. It is a symbolic math library, and is also used for machine learning applications such as neural networks. It is used for both research and production at Google. TensorFlow was developed by the Google Brain team for internal Google use. It was released 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 can be used to achieve all of these applications.The reason for its popularity is the ease with which developers can build and deploy applications. Moreover, Tensorflow was created with processing power limitations in mind. The library can be run on computers of all kinds, even on smartphones.

Python in machine learning

Python is widely considered as the preferred language for teaching and learning ML (Machine Learning). Few simple reasons are :

- It's simple to learn. As compared to c, c++ and Java the syntax is simpler and Python also consists of a lot of code libraries for ease of use.
- Though it is slower than some of the other languages, the data handling capacity is great.
- Open Source – Python along with R is gaining momentum and popularity in the Analytics domain since both of these languages are open source.

Hardware Requirements

The groundwater level prediction system is implemented in Anaconda, and to run Anaconda on the system, the following are the recommended 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, sub-modules 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.

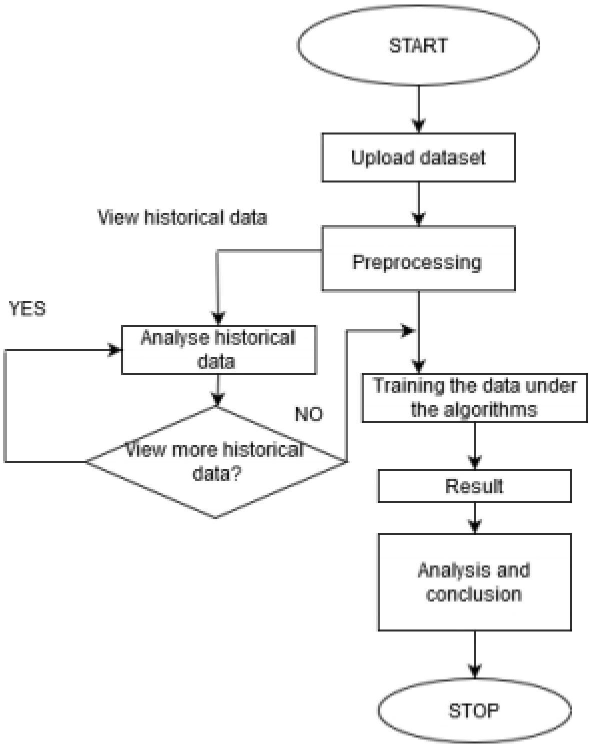


Figure 4.1: Architectural design of the system

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.

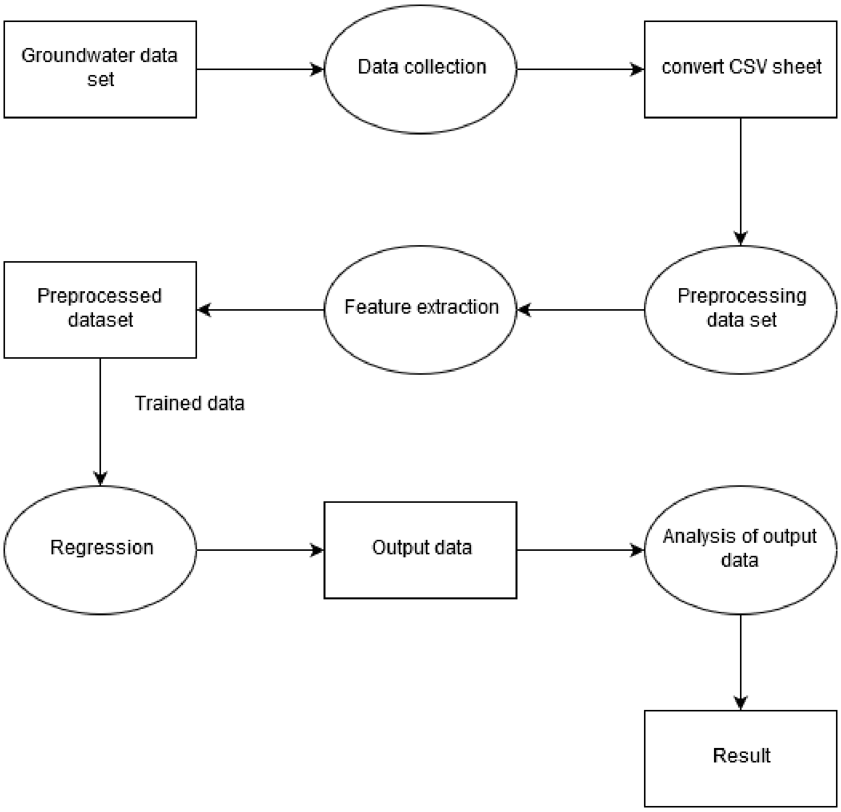


Figure 4.2: Dataflow diagram

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.

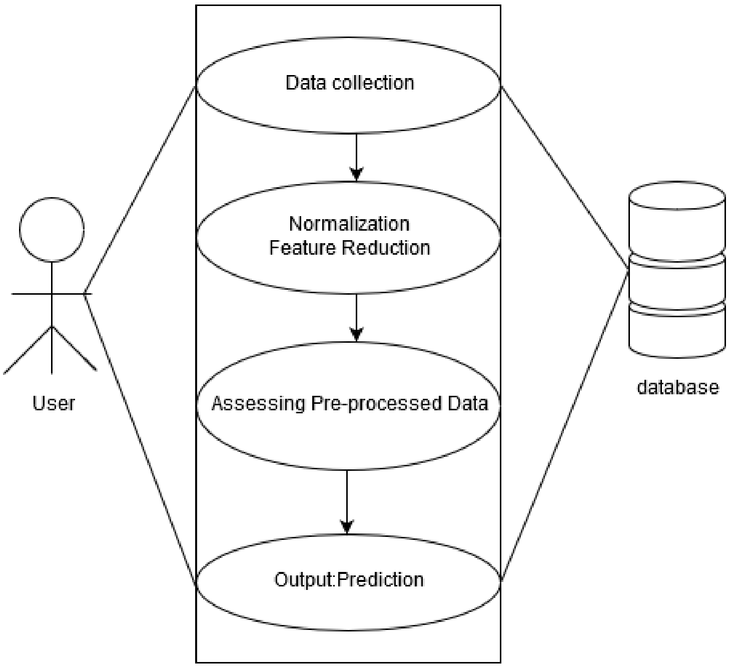


Figure 4.3: Use-case diagram

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** : Spiral model is one of the most important Software Development Life Cycle models, which provides support for Risk Handling. The spiral model combines the idea of iterative development with the systematic, controlled aspects of the waterfall model. Each phase of Spiral Model is divided into four quadrants :

  1. **Objectives determination and identify 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. **Identify and resolve Risks** : During the second quadrant all the possible solutions are evaluated and the risks associated with it are identified to select the best possible solution.

  3. **Develop next version of the Product** : During the third quadrant, the identified features are developed and tested. The next version of the software is available at the end of third quadrant.

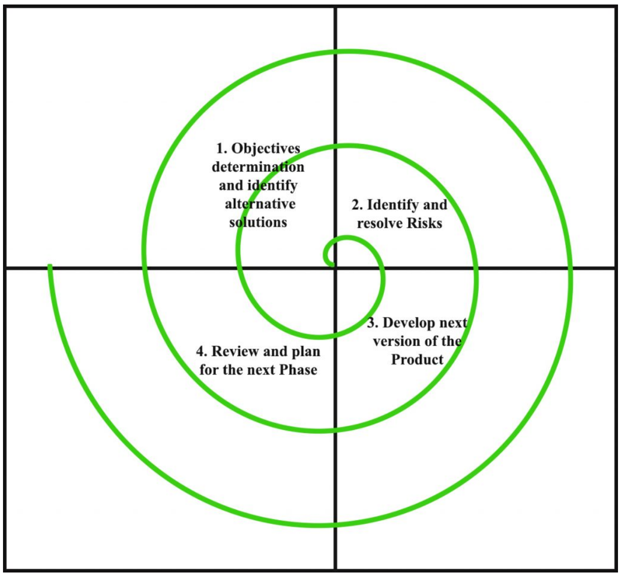

1. Objectives determination and identify alternative solutions
2. Identify and resolve Risks
3. Develop next version of the Product
4. Review and plan for the next Phase

Figure 5.1: Spiral model

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

4. **Review and plan for the next Phase** : In the fourth quadrant, the so far developed version of the software is evaluated. In the end, planning for the next phase is started.

- **Random forest** : Random forest belongs to the class of Supervised Learning algorithm which uses ensemble learning method for classification and regression. Random forest is a type of bagging technique that involves training each decision tree on a different data sample where sampling is done with replacement. Random forest regressor is one such bagging algorithm. A random forest regressor is a meta estimator that uses some kind of splitting criterion to measure the quality of a split and fits a number of classifying decision trees on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting.

```
From sklearn.ensemble import RandomForestRegressor
# Create the model with 100 trees
model = RandomForestClassifier(n_estimators=100
bootstrap = True
max_features = ’sqrt’)
```

Fit on training data

model.fit(train, train_labels)

- **Step 1** : First, start with the selection of random samples from a given dataset.
- **Step 2** : Next, this algorithm will construct a decision tree for every sample. Then it will get the prediction result from every decision tree.
- **Step 3** : In this step, voting will be performed for every predicted result.
- **Step 4** : At last, select the most voted prediction result as the final prediction result.

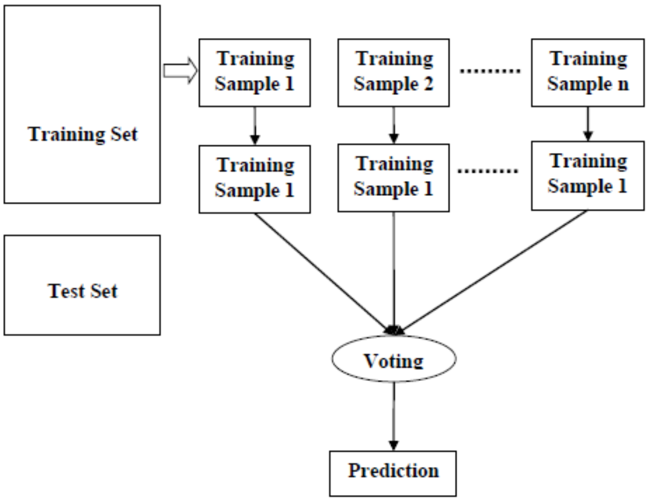


Figure 5.2: Random Forest
The Figure 5.2 depicts the working of Random Forest model in Machine Learning

- **Gradient Boosting** : In gradient boosting machines (GBMs), the learning method consecutively fits new models to provide an accurate estimate of the response variable. The principle idea behind this algorithm is to construct the new base-learners to be maximally correlated with the negative gradient of the loss function, associated with the whole ensemble. Gradient boosting involves three elements :

  1. A loss function to be optimized.
  2. A weak learner to make predictions.

3. An additive model to add weak learners to minimize the loss function.

A benefit of the gradient boosting framework is that the same boosting algorithm can be used for each loss function that might be used. Commonly used weak learners in gradient boosting are decision trees. The trees are added one at a time, and existing trees in the model are not changed. A gradient descent procedure is used to minimize the loss when adding 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)
```

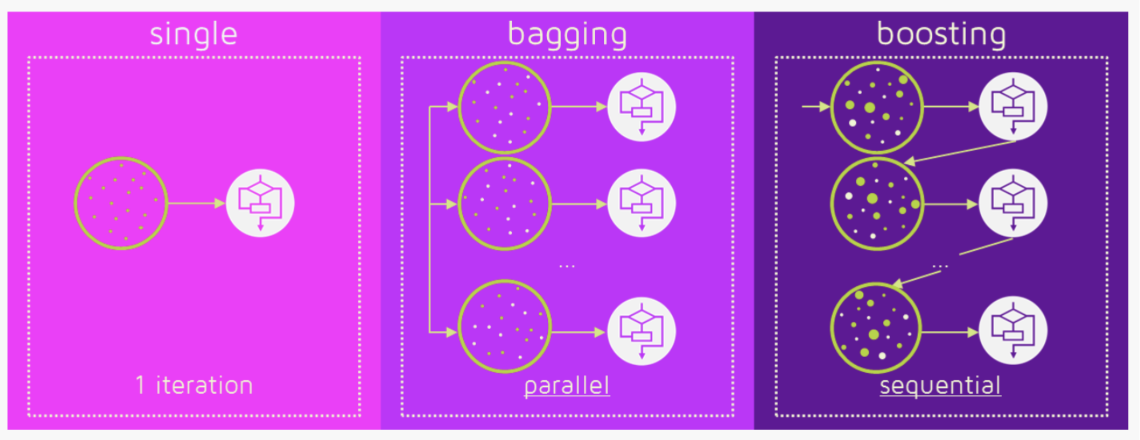


Figure 5.3: Gradient Boosting model
The Figure 5.3 depicts the working of Gradient Boosting model in Machine Learning

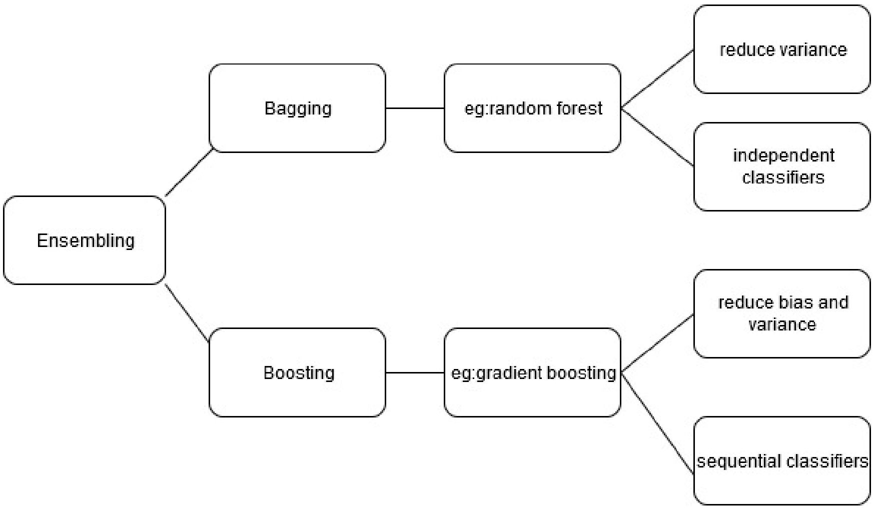


Figure 5.4: Random Forest vs Gradient Boosting
The Figure 5.4 is a comparison of Random Forest model and Gradient boosting model in Machine Learning

- **LSTM** : Long short-term memory (LSTM) units are units of a recurrent neural network (RNN)[15]. 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 :

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

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

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

Here f_t, i_t and o_t can be described as forget gate, input gate and output gate respectively. The matrices W_q contains the weight of the input and U_q contains recurrent connections 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.

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.

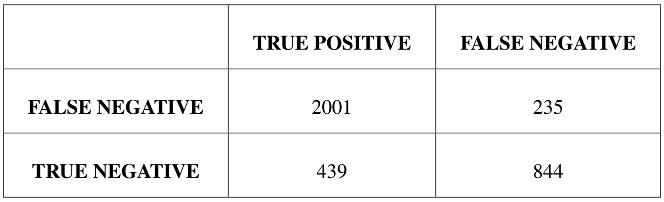
Pre-processing

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

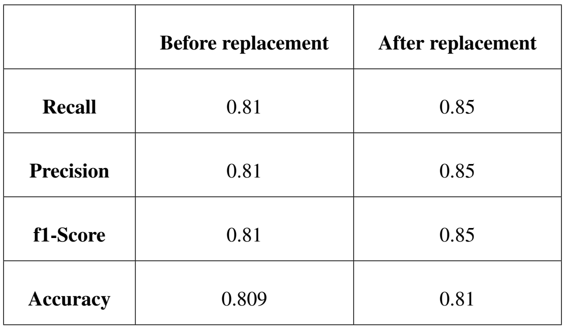
Table 5.3: Confusion matrix from Random Forest Regression
Table 5.3 is a tabular representation of the Confusion matrix from Random Forest Regression.

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.

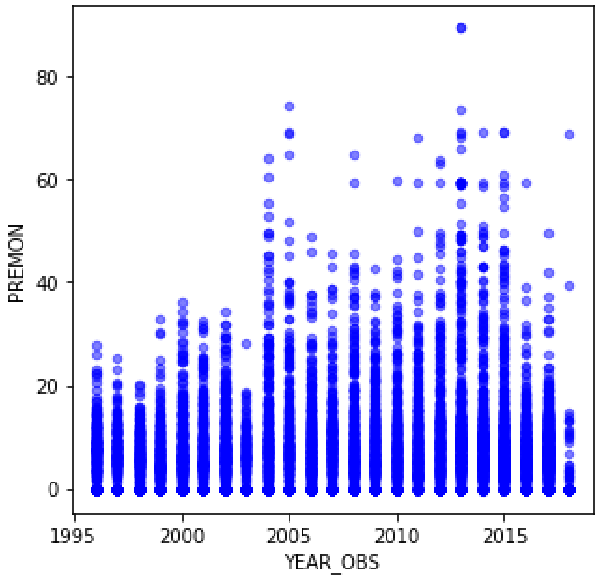


Figure 6.1: Pre-Monsoon Ground water levels year wise
The Figure 6.1 is a year wise scatter plot depicting depletion in groundwater levels during the pre-monsoon season.

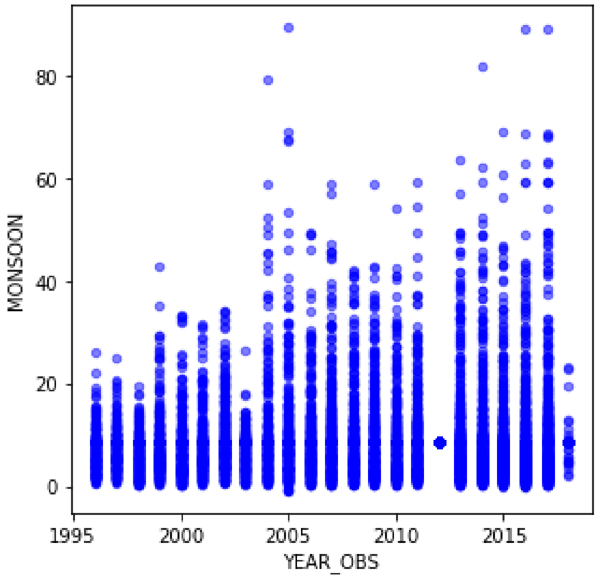


Figure 6.2: Monsoon Ground water levels year wise
The Figure 6.2 is a year wise scatter plot depicting depletion in groundwater levels during the Monsoon season.

A benefit of using gradient boosting is that after the boosted trees are constructed, we can determine importance scores for each attribute. These importance scores are available in the feature_importances_ member variable of the trained model. Importance is calculated for a single decision tree by the amount that each attribute split point improves the performance measure. The feature importances are then averaged across all of the the decision trees within the model. Feature importance gives you a score for each feature of your data, the higher the score more important or relevant is the feature towards your output variable.

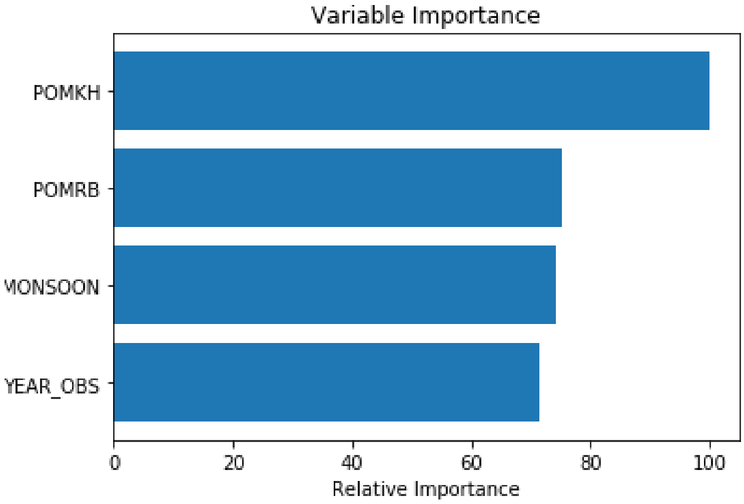


Figure 6.3: Variable Importance
The Figure 6.3 is a plot of variable feature importance.

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.

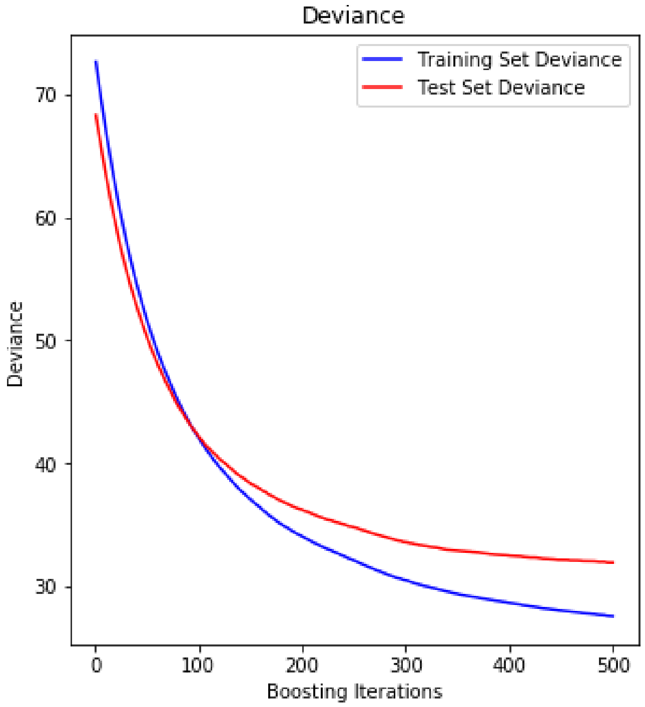
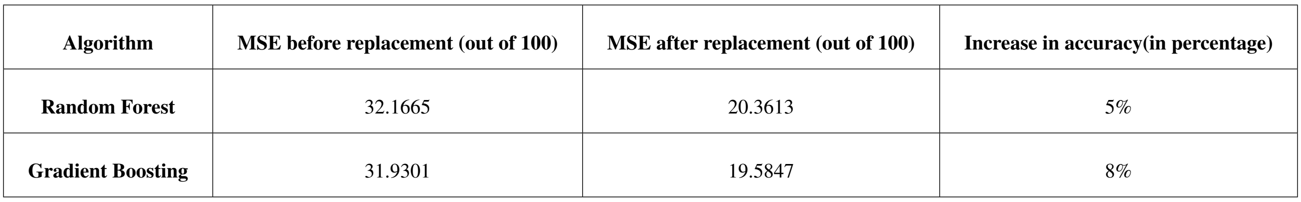


Figure 6.4: Test vs Prediction deviation
The Figure 6.4 is a plot of Test vs Prediction deviation for Gradient Boosting.

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.

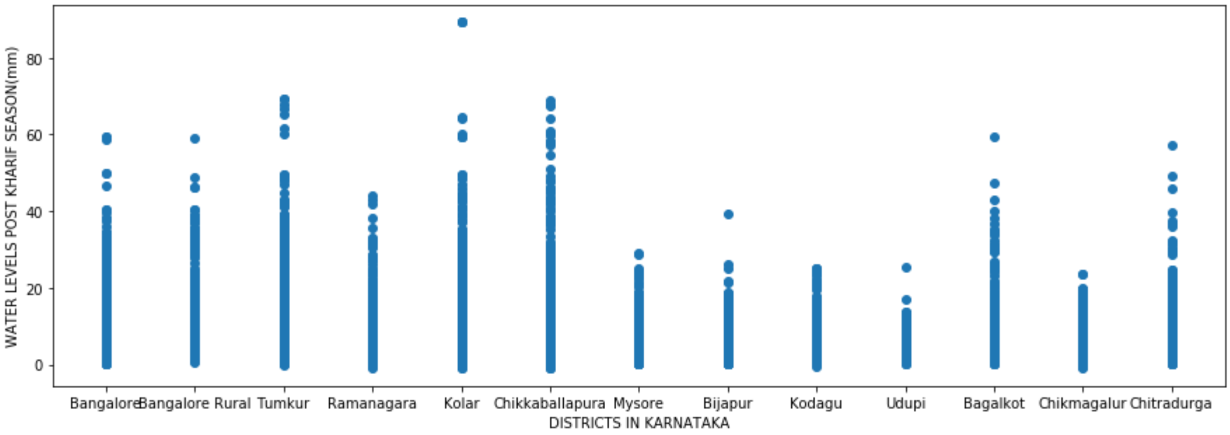


Figure 6.5: Groundwater level distribution in various districts
Figure 6.5 plots the Groundwater level distribution in various districts

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.

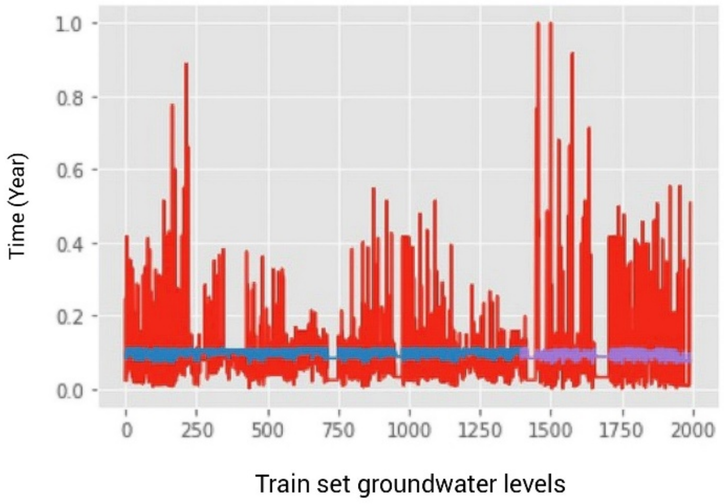


Figure 6.6: LSTM prediction using train set
Figure 6.6 is a plot that describes LSTM prediction using train set values

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.

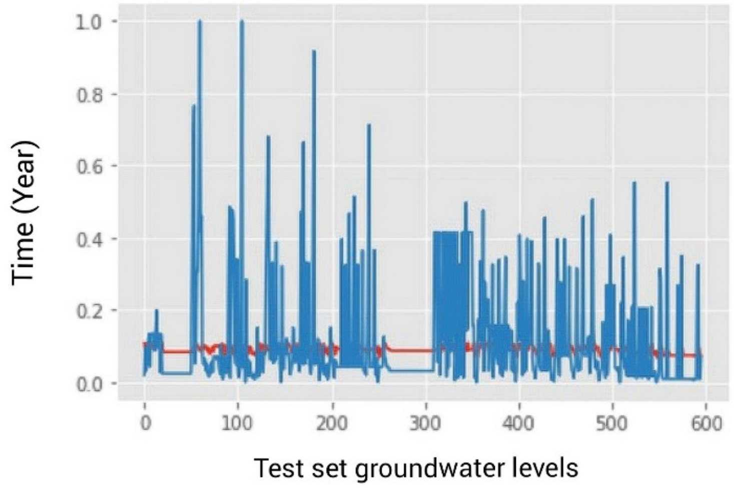


Figure 6.7: Groundwater level distribution in various districts
Figure 6.7 is a plot that describes LSTM prediction using test set values

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

CONCLUSION

In this paper, we propose a groundwater level forecasting system. The study areas were various aquifers and observation wells in several districts across Karnataka, where time-series data were collected. 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 ground-water levels, with considerably good accuracy and lower value MSE. It is practically possible to develop groundwater forecasting models using these data-driven approaches. The methodology presented in this study can be easily applied to other parts of the world as well, irrespective of the hydrogeological settings. Thus, the findings demonstrated in this paper are useful to the currently existing public and private water body organizations as well as research communities of our nation involved in groundwater management and protection.

Bibliography

[1] What is Machine Learning? A definition (As of 05/11/2019) https://expertsystem.com/machine-learning-definition/

[2] India Groundwater: a Valuable but Diminishing Resource (As of 05/11/2019) https://www.worldbank.org/en/news/feature/2012/03/06/india-groundwater-critical-diminishing

[3] Groundwater Level Predictions Using Artificial Neural Networks (Dec 2002) MAO Xiaomin , SHANG Songhao, LIU Xiang

[4] Groundwater Level Forecasting in a Shallow Aquifer Using Artificial Neural Network Approach Purna C. Nayak1, Y. R. Satyaji Rao1 And K. P. Sudheer2(2006)

[5] Application of Back-Propagation Artificial Neural Network Models for Prediction of Groundwater Levels: Case study in Western Jilin Province, China Zhongping (2008)

[6] GCA-CG Based Groundwater Level Prediction With Uncertainty in Lower Reaches of Tarim River (2009) Yue Chen and Yuhong Li

[7] Groundwater level Dynamic prediction based on Chaos Optimization and Support Vector Machine (2009) Jin Liu, Jian-xia Chang Wen-ge Zhang

[8] RBF Neural Network Using Improved Differential Evolution for Groundwater Table Prediction (2010) ZHOU Juan, WEN Zhonghua, QU Jihong

[9] Improved Differential Evolution Based BP Neural Network for Prediction of Groundwater Table* (2010) Jihong Qu, Yuepeng Li and Juan Zhou

[10] An integrated wavelet-support vector machine for groundwater level prediction in Visakhapatnam,India(2014)

[11] Temporal Models for Groundwater Level Prediction in Regions of Maharashtra Dissertation Report -Lalit Kumar(2014) Ch. Suryanarayana a,n, Ch.Sudheer b, VazeerMahammood c, B.K.Panigrahi d

[12] Prediction of Groundwater Level for Sustainable Water Management in an Arid Basin Using Data driven Models (2015) Mutao Huang1, and Yong Tian

[13] Prediction of groundwater level dynamics using Artificial Neural Network (May 2015) J.Colins Johnny, M.C.ShashikKumar, K.Sivadevi, M.Kirubakaran

[14] Groundwater Modeling with Machine Learning Techniques: Ljubljana polje Aquifer Klemen Kenda, Matej Čerin, Mark Bogataj, Matej Senožetnik, Kristina Klemen, Petra Pergar, Chrysi Laspidou and Dunja Mladenić (2018)

[15] Long short-term memory (1997) Hochreiter, S.; Schmidhuber, U.