iJRASET
International Journal For Research in
Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 8 **Issue:** IX **Month of publication:** September 2020

DOI: https://doi.org/10.22214/ijraset.2020.31515

Advanced Ground Water Level Prediction using KNN and Random Forest Algorithm

D. Suganthi[1], A.V. Vishnupriya[2], M. Kaviyarasan[3], S. Vigneshwaran[4]
[1]*Assistant Professor,* [2, 3, 4]*UG Student, Department of Computer Science and Engineering, RVS College of Engineering and Technology, Coimbatore*

Abstract: The main objective of this project is to predict groundwater levels in various areas under circumstances. In order to predict and forecast the ground water levels various machine learning techniques has been used in this project. India has enrolled a basic fall in groundwater levels running somewhere in the range of 75 and 85 percent, as indicated by a checking report by the Central Ground Water Board (CGWB). Groundwater levels in different pieces of the nation are declining a result of ceaseless withdrawal because of reasons, for example, expanded interest for crisp water for different utilizations, impulses of precipitation, expanded populace, industrialization, and urbanization. The southern conditions of Kerala, Telangana, and Pondicherry recorded a decay of 40 to 46 percent. Andhra Pradesh and Tamil Nadu, which is confronting an enormous water emergency, saw a 60 percent tumble. In order to implement this application KNN and Random Forest algorithm are used for prediction and forecasting. In this project we procured a dataset from the trusted resource for analysis the ground water levels in various areas. Using ANN the data training can be implemented. Using clustering method unwanted and irrelevant data will be removed from the dataset. The analysis will be done from the pre processed dataset while implementing random forest algorithm. In this project random forest algorithm play a major role for predicting the ground water level. Here random forest algorithm method analysis various factors and whole attributes from the dataset. It analysis major fields like annual rainfall, sol type, temperature, humidity, industrial areas, number of bore wells, lakes, ponds and etc. All the consideration fields analyzed together to plot and predict the present and future ground water level. All the information can be shown in chart and graph for visual representation.
Keywords: Prediction, Forecasting, KNN algorithm, Random Forest, Data set, Visual representation,

I. INTRODUCTION

The problem definition is to analyse and to conducted concerning the prediction of groundwater levels using machine learning techniques. Machine learning methodologies are implemented using strongly correlated physical parameters as input variables. The results show that data-driven modelling approaches can perform sufficiently well in predicting groundwater level changes. Different evaluation metrics confirm and highlight the capability of these models to catch the trend of groundwater level fluctuations in various areas. Despite the overall adequate performance, further investigation is needed towards improving their accuracy in order to be comprised in decision making processes to attain best performance.

II. METHODOLOGY

Soil salinity spreads in more than 100 countries, and no continent is completely free from salinity. The level of salinity problem varies trans-country and even within the country at different locations, landforms, and irrigated agriculture regions to farmers' fields. Local climatic, environmental, and management conditions determine the salinity problem. Current global estimates reveal over one billion ha area affected to various degrees of soil salinization [1]. The main objective of this study was to assess the impacts of watershed structures and irrigation water use on streamflow and groundwater levels, which in turn affect availability of water for the Cheyenne Bottoms Wildlife Refuge Management area [2]. Considerable uncertainty occurs in the parameter estimates of traditional rainfall–water level transfer function noise (TFN) models, especially with the models built using monthly time step datasets. This is due to the equal weights assigned for rainfall occurring during both water level rise and water level drop events while estimating the TFN model parameters using the least square technique [3]. Prediction of groundwater levels with reasonable accuracy is essential for sustainable groundwater resource management. This is especially critical in arid and semi-arid regions where the groundwater resource is highly utilized for various needs, such as in agriculture, industry and municipal sectors. Time series models are often used for prediction of groundwater levels based on historical data [4,5]. The goal of a formal mathematical optimisation-based groundwater management model is to achieve a specified objective in the best possible manner within the various limiting restrictions.

The limiting restrictions are derived from managerial considerations and physical behaviour of the system [6]. The combined use of simulation and optimisation techniques have been demonstrated to be powerful and useful methods in determining planning and management strategies for optimal development and operation of groundwater systems [7,8].

III. ABOUT ALGORITHM WORKING

A. KNN Algorithm

The k-nearest neighbours (KNN) algorithm is a simple, easy-to-implement supervised machine learning algorithm that can be used to solve both classification and regression problems. The KNN algorithm assumes that similar things exist in close proximity. In other words, similar things are near to each other. It is standard practice to represent the output (label) of a classification algorithm as an integer number such as 1, -1, or 0. In this instance, these numbers are purely representational. Mathematical operations should not be performed on them because doing so would be meaningless.

B. KNN Algorithm Working Model

1) Load the data
2) Initialize K to your chosen number of neighbours
3) For each example in the data
   a) Calculate the distance between the query example and the current example from the data.
   b) Add the distance and the index of the example to an ordered collection
4) Sort the ordered collection of distances and indices from smallest to largest (in ascending order) by the distances
5) Pick the first K entries from the sorted collection
6) Get the labels of the selected K entries
7) If regression, return the mean of the K labels
8) If classification, return the mode of the K labels

C. Random Forest Algorithm

Random forests is an ensemble learning algorithm. The basic premise of the algorithm is that building a small decision-tree with few features is a computationally cheap process. If we can build many small, weak decision trees in parallel, we can then combine the trees to form a single, strong learner by averaging or taking the majority vote. In practice, random forests are often found to be the most accurate learning algorithms to date. The pseudocode is illustrated below. The algorithm works as follows: for each tree in the forest, we select a bootstrap sample from S where S (i) denotes the ith bootstrap. Here the decision-tree using a modified decision-tree learning algorithm taken as sample. The algorithm is modified as follows: at each node of the tree, instead of examining all possible feature-splits, we randomly select some subset of the features f ⊆ F. where F is the set of features. The node then splits on the best feature in f rather than F. In practice f is much, much smaller than F.

D. Random Forest Working Model

A training set S := (x1, y1), . . . ,(xn, yn), features F, and number of trees in forest B.

1) Function RandomForest (S , F)
2) H ← ∅ Sub set creation for tree creation
3) For i ∈ 1, . . . , B do For each tree repeat same process
4) S (i) ← A bootstrap sample from S
5) hi ← RandomizedTreeLearn(S (i) , F)
6) H ← H ∪ {hi}
7) End for
8) Return H
9) end function
10) Function RandomizedTreeLearn(S , F)
11) At each node:
12) f ← very small subset of F from the available trees
13) Split on best feature in f – Tree consolidation for result generation
14) Return the learned tree for various values
15) End function

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 8 Issue IX Sep 2020- Available at www.ijraset.com

E. Analysis and Prediction

Data training process are one of the important roles in the machine learning technique. In this module the dataset will be analyzed and a internal memory will be created for all the fields. One factor will affect another factor in data training process. All the affecting factors will be considered as the major factor. Only trained dataset will be sent to the prediction process. Here KNN and random forest algorithm has been used for data prediction. KNN and random forest algorithm is a statistical measurement used in this project for prediction and forecasting. And other disciplines that attempts to determine the strength of the relationship between one dependent variable (usually denoted by Y) and a series of other changing variables (known as independent variables). random forest algorithm takes a group of random variables, thought to be predicting Y, and tries to find a mathematical relationship between them. This relationship is typically in the form of a straight line (linear regression) that best approximates all the individual data points. In random forest algorithm the separate variables are differentiated by using numbers with subscripts. In the end of the process the prediction result will be generated. All the generated results will be shown in graph and charts. Using MS Chart tools the result will be displayed. X and Y values should the affected factors with the generated result. Changing values in the dataset may affect the predicted result.

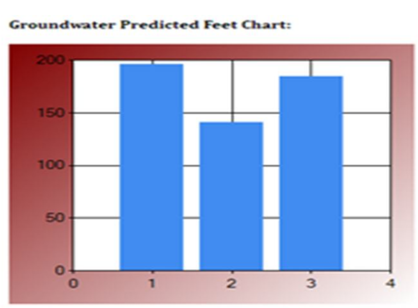
district	area	soiltype	agri	urban	industry	resi	avgrain_cm	humi_per	temp_c	windflow_kmph	bores	availfeet	avgtree	river	pond	lake	Approve	Month	Survey_id
1	4718	Clay Mixture	yes	no	no	no	174	64	35	9	77	140	6816	0	6	1	2019	10	6787
1	4928	Loamy	yes	no	no	no	188	51	30	8	74	179	6515	1	1	1	2018	1	9446
1	7278	Loamy	yes	no	no	no	175	39	30	18	92	194	6914	0	4	0	2018	1	4867
1	3179	Silty	yes	no	no	no	141	45	32	11	95	209	10117	0	6	0	2017	5	6955
1	6034	Silty	yes	no	no	no	157	53	28	16	84	148	11439	0	6	1	2018	7	9453
1	2308	Silty	yes	no	no	no	182	59	31	16	90	180	11589	1	0	1	2018	12	3297
1	6648	Clay Mixture	yes	no	no	no	196	63	29	7	72	148	12522	1	2	0	2018	9	3910
1	184	Clay Mixture	yes	no	no	no	140	38	30	12	91	197	9230	1	6	1	2017	7	3563
1	4102	Loamy	yes	no	no	no	164	46	33	6	74	195	12682	0	5	1	2019	6	8839
1	3407	Clay Mixture	yes	no	no	no	173	68	34	8	70	199	12682	0	1	1	2017	4	4676
1	7013	Silty	yes	no	no	no	159	45	35	17	82	152	11330	0	4	0	2019	2	8791
1	791	Loamy	yes	no	no	no	178	38	32	7	86	170	12575	0	0	0	2019	10	8383
1	1665	Clay Mixture	yes	no	no	no	161	56	29	17	82	147	11333	1	1	0	2019	1	7797
1	1325	Loamy	yes	no	no	no	192	48	33	17	70	159	9589	1	3	1	2019	8	5965
1	6675	Clay Mixture	no	yes	no	no	141	36	33	15	296	257	6961	1	2	2	2017	3	6516
1	394	Loamy	no	yes	no	no	130	55	33	15	220	345	6543	1	2	2	2017	9	5038
1	4876	Loamy	no	yes	no	no	134	71	28	9	179	246	5067	0	0	1	2019	5	9140
1	4688	Silty	yes	no	no	no	180	49	32	10	85	178	10278	0	0	0	2017	8	5340
1	2416	Loamy	yes	no	no	no	150	54	29	16	90	209	12876	0	1	2	2018	4	5813
1	4747	Silty	yes	no	no	no	184	63	30	8	75	149	12752	0	2	0	2018	2	4085
1	3382	Silty	no	no	yes	no	110	35	28	5	629	442	1618	0	0	0	2019	3	2839
1	468	Silty	no	no	yes	no	101	36	33	7	371	449	3329	1	2	0	2017	11	9490
1	2729	Clay Mixture	no	no	yes	no	117	75	28	17	649	513	3387	0	2	1	2019	4	5044
1	1954	Clay Mixture	no	no	yes	no	111	39	28	15	447	411	4118	1	2	1	2019	11	7302
1	2934	Clay Mixture	no	no	yes	no	113	75	31	8	681	597	1367	1	1	0	2019	10	5290

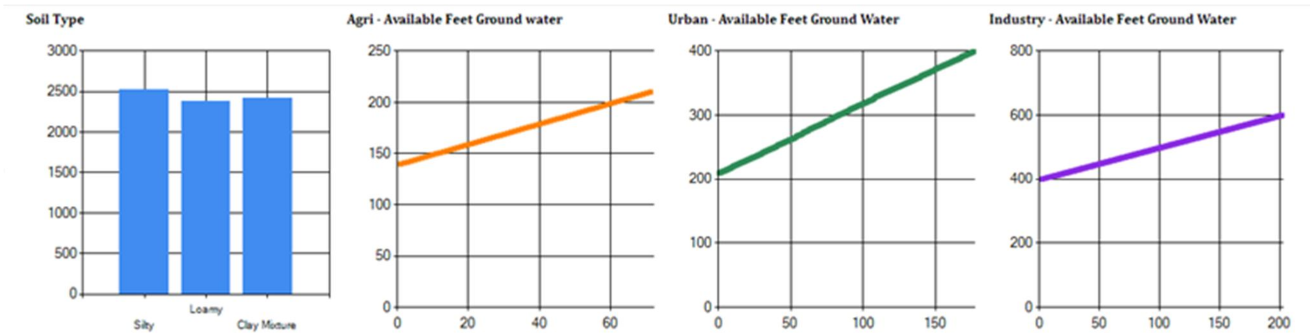
Fig :View Data Set

F. Result

The auto-correlation dataset based analysis results were then used to design an ANN model for the groundwater level prediction. The training results agreed well with the observed data. However, the testing results did not agree as well, which indicates that although the groundwater level is strongly time dependent, the well extraction, the precipitation and the river infiltration also strongly influence the groundwater level fluctuation and should be considered in the prediction model.

Random forest with ANN model was also used to analyze the data by taking into account not only the time dependence of the groundwater level, but also the main groundwater recharge and discharge factors. Both the training results and the testing results agree well with the observed data, which indicates that the model can describe the relationship between the groundwater level fluctuations and their main influence factors. Therefore, the model can be used to predict groundwater level in accurate.

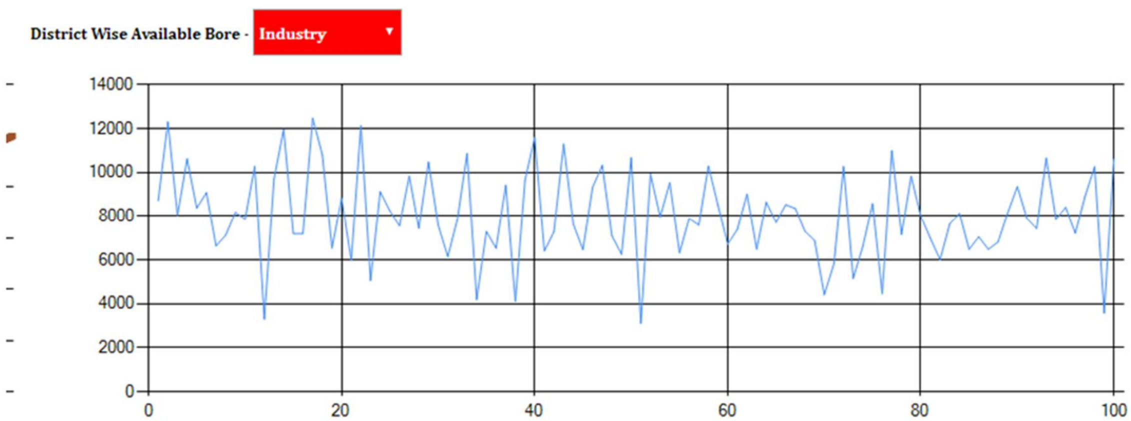




IV. CONCLUSION

Thus this project has been executed successfully and the output has been verified. All obtained outputs are according to committed in abstract. Initially more problems occurred during the algorithm implementation. As mentioned above here KNN and Random Forest has been implemented successfully. These two algorithms are working perfectly in the Application. And these algorithms will works on independent process too. These features will make this project more successful and efficient.

The drinking water crisis in is reaching alarming proportions. It might very soon attain the nature of global crisis. Hence, it is of utmost importance to preserve water for human beings. In many places there is unnecessary wastage of water due to exceed usage.



REFERENCE

[1] Cai Shuying, Yang Jinzhong. “Recent developments of the study on prediction methods of regional groundwater level and soil salinity”. China Rural Water and Hydro power, 1997, (sl) : 106-110 . (in Chinese)
[2] Ramireddygari S R, Sophocleous M A, Koelliker J K, et al. “Development and application of a comprehensive simulation model to evaluate impacts of watershed structures and irrigation water use on streamflow and groundwater” : The case of Wet Walnut Creek Watershed, Kansas, USA. J. Hydro. , 2000, 236(3): 223-246 .
[3] Tankersley C D, Graham W D, Hatfield K. “Comparison of univariate and transfer function models of groundwater fluctuations”. Water Resour. Res. , 1993, 29(10) : 3517-3533.
[4] Bierkens M F P. “Modeling water table fluctuations by means of a stochastic differential equation.” Water Resour. Res. , 1998, 34(10) : 2485-2499.
[5] Bierkens M F P , Knotters M , Van Geer F C. “Calibration of transfer function-noise models to sparsely or irregular observed time series”. Water Resour. Res. , 1999, 35(6): 1741-1750.
[6] Knotters M , De Gooijer J G. “TARSO modeling of water table depths”. Water Resour. Res. , 1999, 35(3): 695-705 .
[7] Philip D W. Neural Network Models : “Theory and Projects (2nd Edition). London: Springer-Verlag London Limited”, 1997.
[8] Yuan Zengren. “The Artificial Neural Networks and Its Application”. Beijing; Tsinghua University Press, 1999. (in Chinese)
[9] Morshed J, Kaluarachchi J J. “Application of artificial neural network and genetic algorithm in flow and transport simulations”. Adv. Water Resour. , 1998.
[10] Mukhopadhyay A. “Spatial estimation of transmissivity using artificial neural network”. Ground Water, 1999.