JOURNAL OF CRITICAL REVIEWS

ISSN- 2394-5125 VOL 7, ISSUE 04, 2020

Improving the Analysis rate of Ground Level Water Prediction using different Machine learning models

M.Saravanan[1] , Nagoor Meeran.A.R[2]

[1]SRM Institute of Science and Techology,Kattankulathur
[2]Shinas College of Technology ,Oman
e-mail – saran84gct@gmail.com, nagoor.meeran@shct.edu.om

Received: 14 Feb 2020 Revised and Accepted: 25 March 2020

ABSTRACT: Water scarcity becoming major problem in many countries like india, Recent problems of critical shortage of water in so many parts of our country due to population, lack of water harvesting plans, excess wastage of water and extreme pollution of water bodies, bringing inaccurate estimation of ground water available in these parts. Even though many water resoureces available majorly dependent on ground water,So estimation of groud level water becomig difficulty in some occations for the goverment, Even many authors proposed water resource managment ,Still we facing challenge in our country for prediction of ground level water based on the various factors like rain fall,inorder to address this issue in this work developed analysis of various machine learning models to predict the ground level water for the data set available from the goverment side.This work deals with dataset considered from india official site and hydrological parameters consideration for different states of India. We performed a data cleaning step to make sure that the dataset used had no null values, so that it could provide the best optimal result in the final stage. After preprocessing the dataset , we performed training on different machine learning models such as SVM (Support Vector Machine), KNN (K-nearest neighbors), Gausian Naive Bayes(GNB), Logistic Regression, Decision-Tree Classifier Algorithm.All the models trained individualy and tested with the different dataset and finaly accuracy of individual model compared with each other. From the comparision of all models Decision Tree Classifier model provides the best improved prediction results. Hence, it can be used to analyze groundwater level properties, so that the government can take preventive measures beforehand.

KEYWORDS:

I. INTRODUCTION

1.1 GENERAL OVERVIEW

Water present below the ground surface level mainly comprises two types of zones - saturated and the unsaturated zone. In the monsoon season, when the rainfall occurs, some part of that rainfall water gets infiltrated into the ground [1]. Out of this water getting infiltrated into the ground, some of its part gets stored in the topmost layer's pore spaces of soil. This soil layer is right beneath the land surface, & contains both water and air, and is called an unsaturated zone. Once the soil pores are entirely filled with water, then it percolates further down via the fracture and gap in the rocks. All soil pores get completely filled with water at a certain depth, this layer is known as Saturated Zone. The saturated zone is the water table, & the water in this layer is known as groundwater.

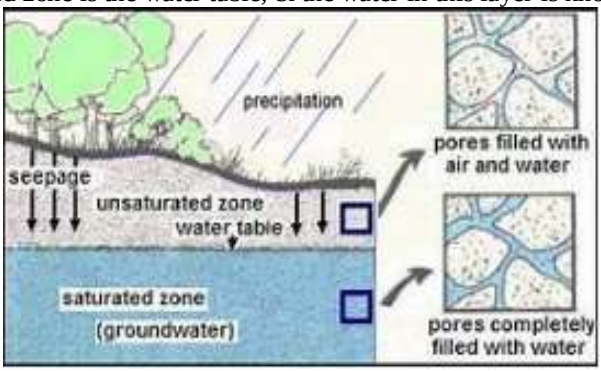

Figure-1: Groundwater Aquifers

ISSN- 2394-5125 VOL 7, ISSUE 04, 2020

1.2 EVOLUTION

In recent years, there has been regressive research and attempts to study the properties of ground water content and groundwater level analysis [2]. We will be describing various approaches that have been implemented before and evolution in this field with the help of a study which reviews significant projects in the field beginning with "Groundwater table depth analysis for Gaza city", study of Groundwater Management Index to gather information about groundwater level and content properties[3],etc.In this work try to implement idea to develop a system that can perform analysis of groundwater level properties for Indian states in general.
Our aim is to analyse ground water level through estimation of the magnitude of hydrological parameters, and then implement the concept of supervised learning to train our machine to check the water level study, so that we can suggest the best suitable machine learning algorithm that is used to predict the final groundwater level situation beforehand. So that the government can take preventive measures for the situation beforehand. We used parameters to design the forecasting model, & evaluate its scope to predict groundwater level. The input parameters of groundwater level's forecasting will be derived using Time Series Analysis (TSA). We further using Supervised Machine Learning to train our model for analysis of water table depth with different combinations of hydrological parameters.

II. RELATED WORK

In this paper [3], the authors have used forecasting as a data mining method to study rainfall scale andwater production in Deir EL-Balah, Gaza. It suggested that lack and fluctuation in the rainfall level and increased population in the city may cause a major crisis in Gaza, Palestine, since water resources are dwindling coupled with an increase in population. It stated that the major source of water in Gaza is Groundwater. Hence, with an increase in groundwater demand, and decreasing rainfall (i.e the prominent source of groundwater), will lead to sudden depletion in groundwater wells, and cause a jump in the salinity rate. In this paper [4], the author performed various experiments to establish Groundwater Management Index (GMI) for "Szuchun Creek" Hot Springs observation site in Taiwan city and any many other such sites. They studied the fluctuations in the groundwater level for all these above mentioned places. They further combined the results gathered from this stage with the information of the different features of hot spring creation, & management to improve hot spring's GMI. In this paper [5], the author elaborated that the groundwater table (WTD, Water Table Depths) is crucial to determine turf vegetation, & GHG emissions from peatlands, which is why they explained that the study of water table depth is very important for agricultural purposes and climatic changes perspective. They explained in a turf area that is partially decayed plants, moisture content in the topmost layer of soil is majorly affected by hydrostatic equilibrium with water table depth. They collected a dataset projected by the satellite "Sentinel-1", that depicted C-band radar information with increased spatial and temporal resolution. In the results, they predicted the final hydrostatic equilibrium range for the corresponding sites and concluded radar satellite study technique is effective with the accuracy level but suggested that more research work needs to be done in this area for better advancements.This paper [6], suggests that groundwater is a crucial source of drinking water across the world and therefore the awareness of the presence of content and dynamics has turned into very important control measures for water pollution and sustainable development. In addition, quantitative information about the level location and amount of surface water is a crucial variable for accurate hydrological modelling of groundwater table depth systems. They suggested the concept of Surface nuclear magnetic resolution (SNMR), that can help in providing solutions to such problem statements by estimating quantitative water content and pore parameters of sub water surfaces. This paper [7] focuses on various problems the groundwater sector has been experiencing in recent years. The groundwater flow model provides information about water balance and helps in over consumption of water resources. The authors of this paper developed a prototype model expert system for the completion of groundwater modelling technique, also known as ALAES. In this paper[8] authors developed a system predict monthly groundwater levels and their results based on the wavelet (WA) integrated WA-SVM model to performs better than the models auto regressive integrated moving average (ARIMA), ANN and SVM. The authors [9] in this paper showed the prototype ability of data based generalized non-linear relationship hydro(geo)logical, meteorological and climatic input variables and groundwater.

III. TRAINING ON PREDICTION MODELS

In this study different Machine learning predition stechniques,these models works based on the data set from the previous experiences are colleted and feeded in the machine corresponding to some tasks to produce output for other sets of tasks with improved performance. In other words, machine learning is a method to analyze the data in order to build an automated analytical model. It is based on the concept of learning a system from particular data, to identify relevant patterns out of it, and then to predict output in order to minimize human interaction. Supervised learning is one of the types in which the process of machine learning can be performed. It

ISSN- 2394-5125 VOL 7, ISSUE 04, 2020

involves building up a function that produces output for a particular input based on it's learning experience of input-output pairs. It produces a function from training data to produce output for larger master dataset. In other words, supervised learning can be explained as a technique in which a new model gets training on a labelled dataset shown in fig-1. In this work dataset mainly comprises various hydrological parameters with their water level situation and different prediction models evaluated based on the dataset finaly selected the most appropriate model for predicting the ground level water.

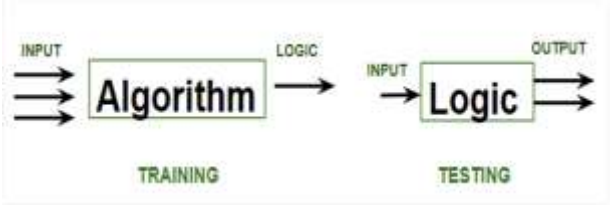

Figure 2 - Machine Learning Process

3.1 Data Cleaning and Pre-Processing

We have collected dataset that has proper details of various hydrological parameters with the corresponding water level situation[10].After the dataset collection process, the next step is to perform data pre processing steps. The term data pre processing is an essential pre step in any model evaluation,it helps to handles data efficiency, remove error, remove blank spaces and replace null values in the dataset. In other words, data preprocessing is the step that transforms raw data into clean data. Because, in the primary stage, when data is collected through different sources, it exists in raw form, which is not suitable for accurate results.

3.2 Feature Engineering

Feature engineering is the technique of extracting domain properties of the data to produce the feature properties that enables machine learning algorithms to work accurately and predict results with better accuracy show in fig-3. Hence, if the completion process of feature engineering is done properly, it tends to improve the efficiency rate of machine learning algorithms by creating the important features from the given dataset. In this study, different feature properties examined and then split the data set into test data set and training data set, implemented the concept of decision tree classifier to extract the most important feature that is most likely to affect the final outcome of machine learning algorithms with the changes in the initial parameters.

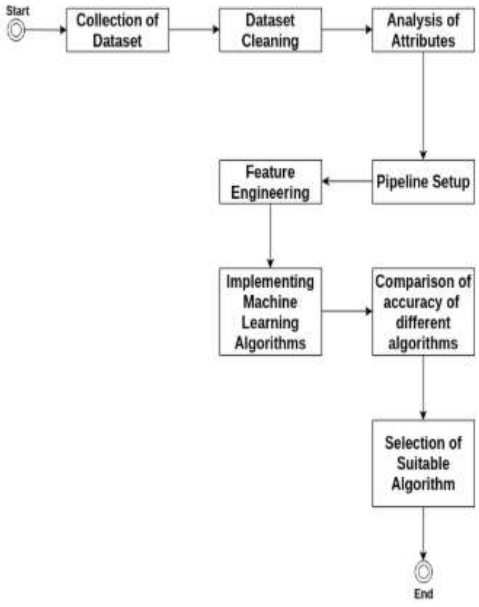

Figure 3 - Block Diagram

3.3 Support Vector Machine(SVM)

SVMs come under supervised learning models, utilized in order to analyze data that can be further used for classification and regression analysis. SVM model [10] is a representation of the points in space and is mapped in such a way that the points are put into separate categories by a plane that is as wide as possible. Any addition of new points are mapped are predicted and added to the category they rightfully belong to based on the side of the plane they would fall. SVM can work with any size of datasets to provide efficient results as desired by the users. Support Vector Machine (SVM) is commonly used for classification problem statements show in fig-4.

ISSN- 2394-5125 VOL 7, ISSUE 04, 2020

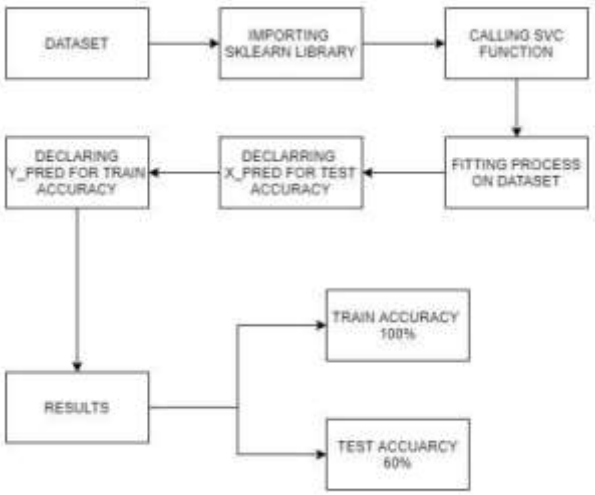

fig 4- Blockdiagram SVM

In this work, we have used the sklearn library to import the libraries termed as SVC and accuracy scores. A variable termed as clf is declared, in which SVC function is called with kernel as a parameter, we have used kernel type as linear. Then fitting is performed on our dataset using a fitting method. Two variables named as x_pred and y_pred were declared to predict the train and test accuracy shown in fig-4.

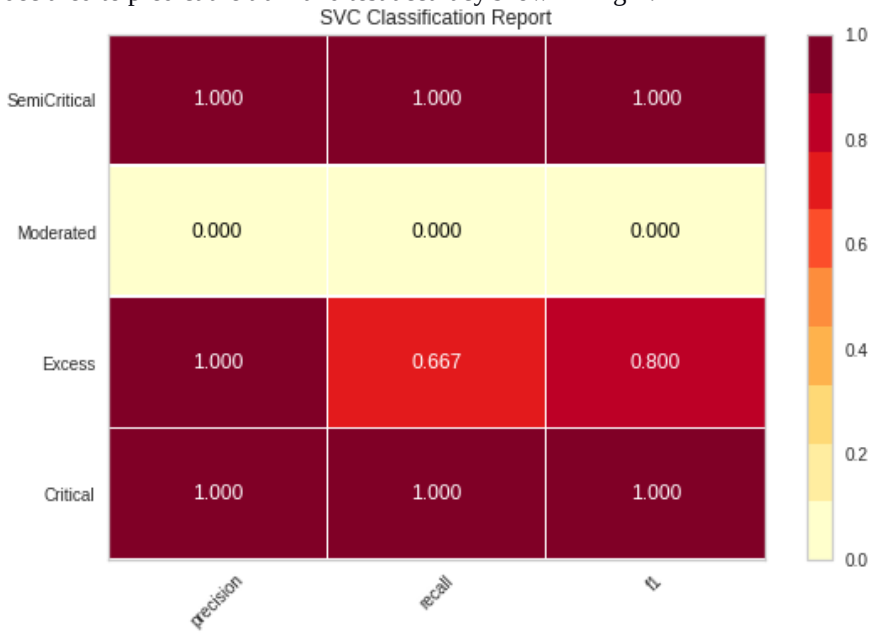

Fig 5- Classification Report for Support Vector Machine

Fig-5 shown Classification report helps in calculating the quality of the results predicted by the corresponding model ,from the model out come results obtained turned out correct and incorrect values based on the dataset. This values can be identifed from involves Precision value, recall value, f1- score values shown in fig- 5. The term precision equation-1 can be understood as the skill of the Classifier used to not to mention positive attributes that are actually negative. The term recall value is the skill of the classifier implemented to calculate all the positive instances correctly it shown in equation-2.The term F1- score represents a weighted harmonic mean of precision value and recall values show in equation-3.

$$Precision = TP/(TP + FP$$

Equation-1

$$Recall = TP/(TP+F$$

Equation-2

TP --> True Positives values FP--> False Positives values.

TP--> True Positive values FN--> False Negative values.

ISSN- 2394-5125 VOL 7, ISSUE 04, 2020

F1 Score = 2*(Recall * Precision) / (Recall + Precisic

Equation-3

3.4 KNN Algorithm

KNN is the easiest prediction model, categorized as Supervised Learning algorithms. K-Nearest Neighbour algorithm presumes the similarity between the test data and training dataset and then looks for similar properties and features for incoming new data that is to be processed. This algorithm stores all the information regarding available dataset and then distinguishes the new dataset on the basis of available information similar features[12]. Thus, this algorithm helps in classifying new dataset into appropriate categories. The KNN algorithm in the training phase only saves the dataset and when it gets a new data request, then only it categorizes that data into a category that is much similar to the previous data properties.

In this work sklearn.neighbors library to import K NeighborsClassifier. A variable termed as knn is declared, in which the K Nearest neighbor function is called with no of neighbors as its parameter, in our project it is 3. Then we perform fitting on our dataset by using the parameters as X_train, y_train. Variables termed as accuracy_train and accuracy_test have been declared to calculate the train and test accuracy. Our dataset showed an accuracy of 87.5 for the train dataset and 40% for the test data.

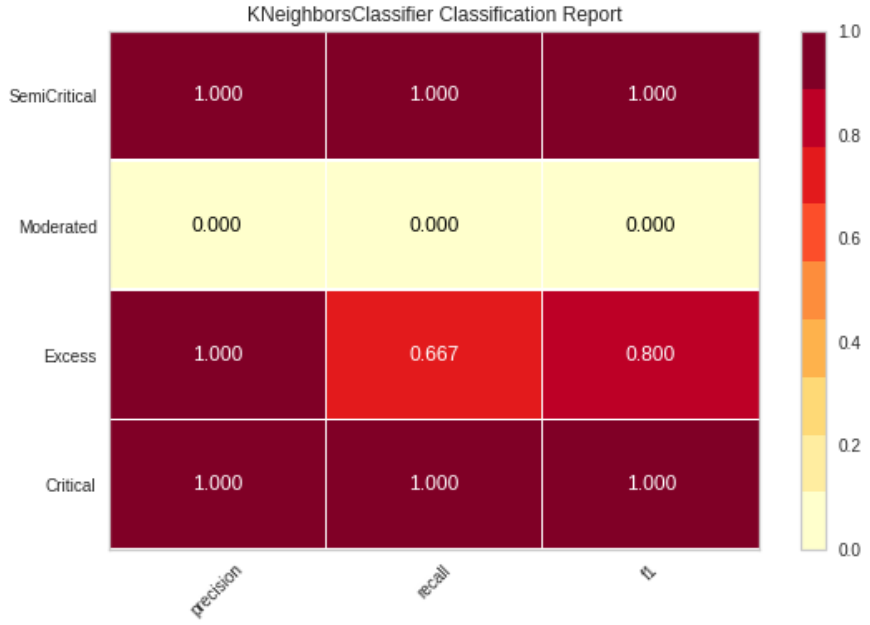

Fig 6- Classification Report for KNN algorithm

Precision value, recall value, f1- score values shown in fig- 6

3.5 Logistic Regression

Logistic regression (LR) is one of the most common algorithms used in the Machine Learning field [13], and can be categorized into the Supervised Learning technique category. It helps in the prediction of the dependent variable by taking reference of certain independent variables. It helps in predicting the result for categorical dependent variables. Therefore, the result obtained must be either categorical or discrete value, which means either it is Yes or No, 0 or 1, True or False, etc. This algorithm is mainly used for solving classification problems and is a very significant prediction algorithm because it can create probability factors to classify new datasets.

ISSN- 2394-5125 VOL 7, ISSUE 04, 2020

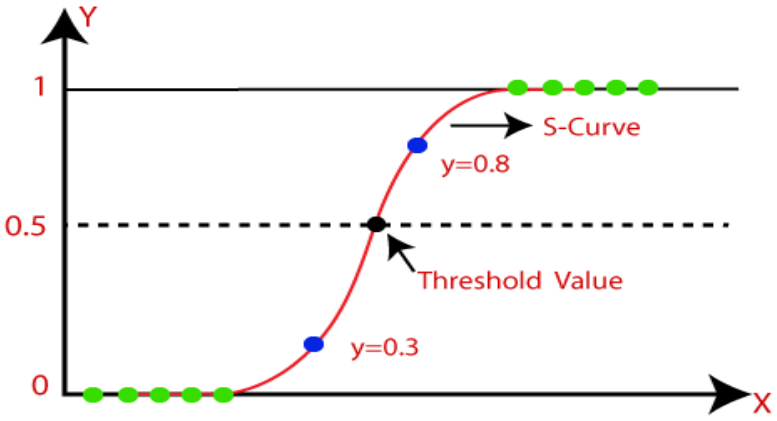

Figure 7 - Sigmoid Function

Logistic Function also known as Sigmoid function shown in fig-7 helps in mapping predicted values with corresponding probability factor. The range of the result of this function is from 0 to 1, which suggests that it cannot exceed this limit, hence it forms the curve in "S" shape. This S-type curve is termed as Sigmoid Function or Logistic Function.

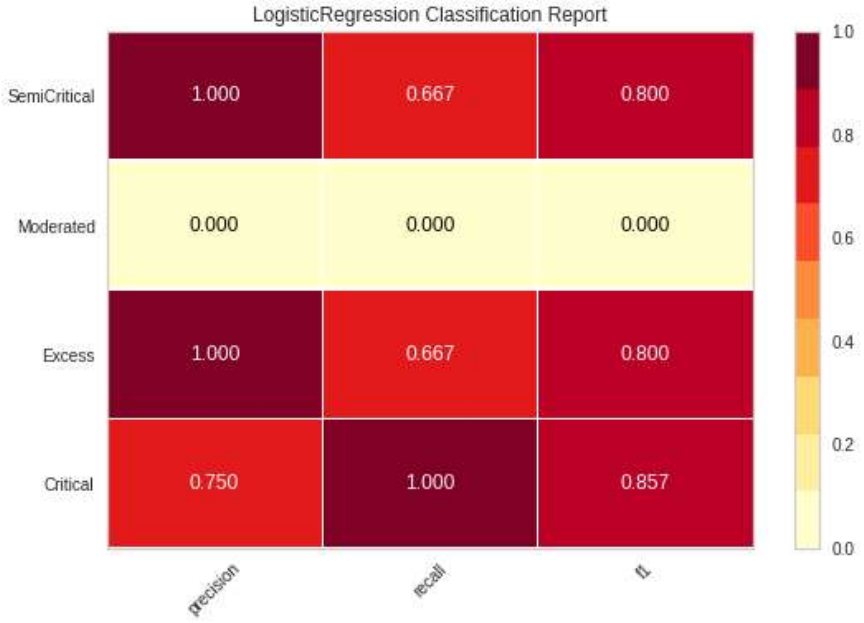

Fig 8- Classification Report for Logistic Regression

Precision value, recall value, f1- score value for Semicritical, Moderated, Excess, Critical situation shown in fig-8.

3.6 Gaussian Naive Bayes

Gaussian Naive Bayes algorithm can be classified as supervised learning algorithm[14], that is mainly based on Bayes theorem and used to focus on classification problem statements. It is majorly used for a type of text classification problem that comprises high-dimensional training data.It is also known as a probabilistic classifier algorithm, which basically means it gives results taking the probability factor of an object into account. The common examples of Gaussian Naive Bayes algorithms are Sentimental analysis, classification articles[14], etc. Bayes' theorem is commonly known as Bayes' Rule or Bayes' law, that helps in finding the probability factor of an event with the previously available knowledge. The formula for theorem is shown in fig-9.

$$P(A|B)=\frac{P(B|A)P(A}{P(B)}$$

Fig-9: Bayes Equation

where,

ISSN- 2394-5125 VOL 7, ISSUE 04, 2020

1. P(A|B) - Posterior probability: The probability of hypothesis A on the observed event B.
2. P(B|A) - Likelihood probability: The probability of the evidence given that the probability of a hypothesis is true.
3. P(A) - Prior Probability: The probability of hypothesis before observing the evidence.
4. P(B) is known as Marginal Probability: Probability of Evidence.

In this work , we used the sklearn library to import the Gaussian and Multinomial Naive Bayes. A variable termed as gnb is created in order to call the GaussianNB()function. A variable termed as y_pred is where the model fitting and prediction was done and finally the prediction was performed on the test data. Classification report for Gaussian Naive Bayes shown in fig-10

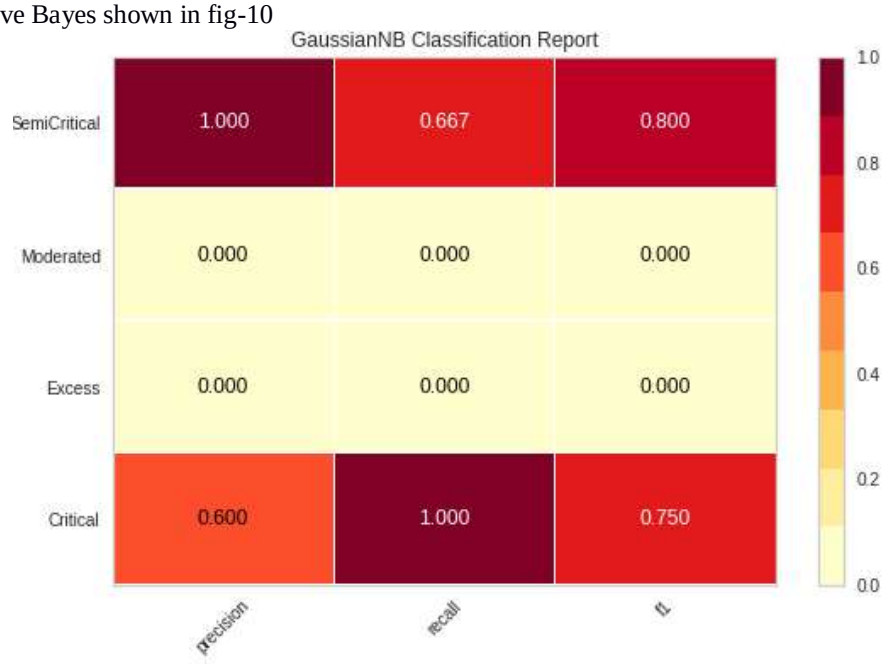

Fig 10- Classification Report for Gaussian Naive Bayes

Precision value, recall value, f1- score value for Semicritical, Moderated, Excess, Critical situation shown in fig-10

3.7 Decision Tree Classification

Decision Tree can be categorized under Supervised learning technique,which mainly focuses on solving classification and regression problem statements[15]. In this work used the sklearn library[15] to import tree. A variable termed as clf is declared to call the Decision tree classifier function. Then the prediction is performed on our dataset and it is stored in a variable called y_pred. From the dataset classification report for Decision Tree shown in fig-11.

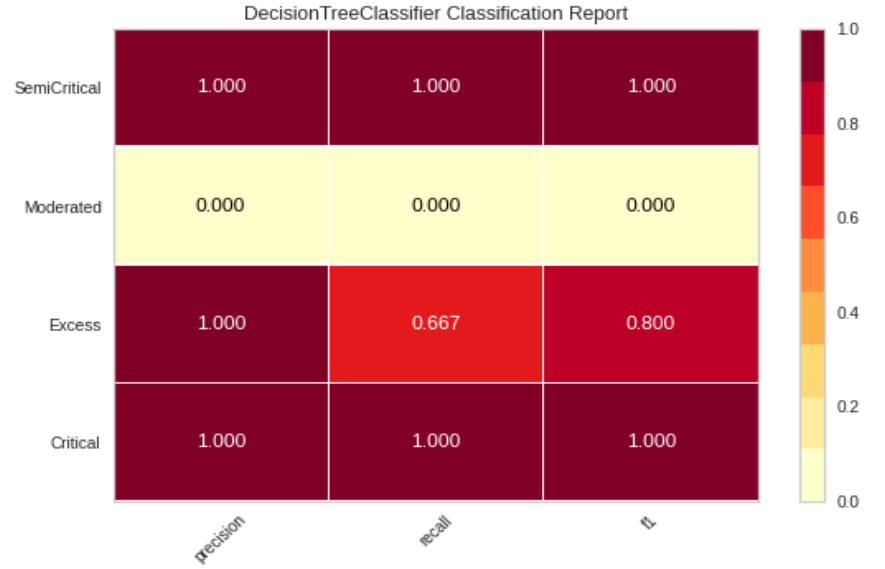

Fig 11- Classification Report for Decision Tree Classifier

ISSN- 2394-5125 VOL 7, ISSUE 04, 2020

Precision value, recall value, f1- score value for Semicritical, Moderated, Excess, Critical situation shown in fig-11

IV. RESULTS AND DISCUSSION

In this review we implemented five different prediction algorithms to study their respective accuracy and performance for the given dataset. These models evluated with training and testing accuracy. All the models trainig accuracy provided more than 100% and testing accuracy shown in table-1 and fig-12.

ML ALGORITHM	ACCURACY
K-NN	40%
SUPPORT VECTOR MACHINE	60%
LOGISTIC REGRESSION	78%
DECISION TREE CLASSIFIER	88.88%
GAUSSIAN NAIVE BAYES	55.55%

Table 1- Machine Learning Algorithms Accuracy comparison

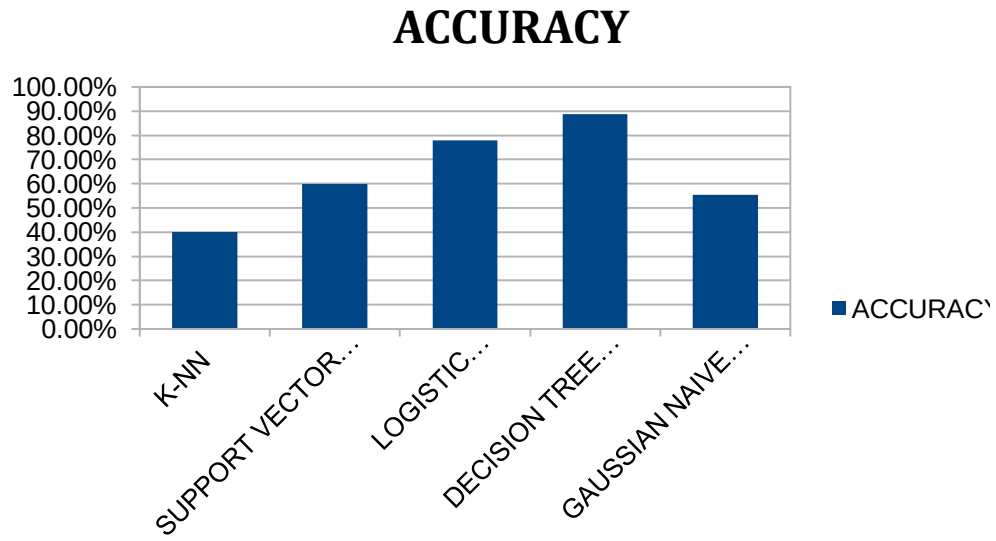

Figure 12: Accuracy Comparision

Based on the accuracy for training and testing models Decision-Tree Classifier Algorithm perform better thanSVM (Support Vector Machine), KNN (K-nearest neighbors), Gausian Naive Bayes(GNB) and Logistic Regression models.

V. CONCLUSION

In this paper we reviewed the different prediction algorithms and provided the one best model to apply in future for the dataset. Ground water level analysis system mostly depends on the rain fall, small data set Decision tree perform good results but size of larger master dataset need to develop to find the best machine learning

ISSN- 2394-5125 VOL 7, ISSUE 04, 2020

algorithm that will produce the optimal result for hybrid parameters.We still working on ensemple techniques for large dataset.The Recent problems of critical shortage of water in so many parts of Tamil Nadu, Andhra Pradesh and Karnataka they can predict early from this model from various hydrological parametrs and government can take preventive measures beforehand to avoid water scarcity problems in future.

VI. REFERENCES

[1] https://www.usgs.gov/special-topic/water-science-school/science/aquifers-and-groundwater?qt-science_center_objects=0#qt-science_center_objects

[2] https://www.lenntech.com/groundwater/properties.htm

[3] Amra, I.A.A. and Maghari, A.Y., 2018, October. Forecasting groundwater production and rain amounts using ARIMA-Hybrid ARIMA: Case study of Deir El-Balah City in GAZA. In *2018 International Conference on Promising Electronic Technologies (ICPET)* (pp. 135-140). IEEE.

[4] Chen, W.P., Wang, J.T., Kan, C.C. and Lin, H.I., 2018, November. Establishment of a Groundwater Management Index for the Szuchung Creek Hot-Spring. In *2018 IEEE International Conference on Advanced Manufacturing (ICAM)* (pp. 155-158). IEEE.

[5] Asmuß, T., Bechtold, M. and Tiemeyer, B., 2018, July. Towards monitoring groundwater table depth in peatlands from Sentinel-1 radar data. In *IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium* (pp. 7793-7796). IEEE.

[6] Chen, B., Li, J., Hu, X. and Liu, Y., 2018. Surface NMR responses of typical 3-D water-bearing structures evaluated by a vector finite-element method. *IEEE Transactions on Geoscience and Remote Sensing*, *56*(10), pp.5626-5635.

[7] Alaoui, M.M., Kacimi, I. and Ouazar, D., 2019, April. A knowledge based framework for groundwater modeling. In *2019 5th International Conference on Optimization and Applications (ICOA)* (pp. 1-7). IEEE.2

[8] Suryanarayana, C., Sudheer, C., Mahammood, V., and Panigrahi, B. K. (2014). An integrated wavelet-support vector machine for groundwater level prediction in Visakhapatnam, India. Neurocomputing, 145, 324-335.

[9] Sahoo, S., Russo, T.A., Elliott, J. and Foster, I., 2017. Machine learning algorithms for modeling groundwater level changes in agricultural regions of the US. *Water Resources Research*, *53*(5), pp.3878-3895.

[10] https://data.gov.in/catalog/rainfall-india?filters%5Bfield_catalog_reference%5D=1090541&format=json&offset=0&limit=6&sort%5Bcreated%5D=desc

[11] Ben-Hur, A. and Weston, J., 2010. A user's guide to support vector machines. In *Data mining techniques for the life sciences* (pp. 223-239). Humana Press.

[12] Srivastava, T., 2018. Introduction to k-Nearest Neighbors: A powerful Machine Learning Algorithm (with implementation in Python & R). *Analyticsvidhya. Com*.

[13] Peng, C.Y.J., Lee, K.L. and Ingersoll, G.M., 2002. An introduction to logistic regression analysis and reporting. *The journal of educational research*, *96*(1), pp.3-14.

[14] Huang, Y. and Li, L., 2011, September. Naive Bayes classification algorithm based on small sample set. In *2011 IEEE International conference on cloud computing and intelligence systems* (pp. 34-39). IEEE.

[15] Garreta, R. and Moncecchi, G., 2013. *Learning scikit-learn: machine learning in python*. Packt Publishing Ltd.