JCRES
Journal of Current Research in Engineering and Science
Bi-Annual Online Journal (ISSN : 2581 - 611X)

Volume7-Issue1, January 2024
Paper: 33

Comprehensive Agricultural Drought Assessment

Gayatri Ajith[1]
Department of Data Science and Business Systems
SRM Institute of Science and Technology
Kattankulathur, Tamilnadu, India
ga5975@srmist.edu.in

Snehha S[1]
Department of Data Science and Business Systems
SRM Institute of Science and Technology
Kattankulathur, Tamilnadu, India
ss7389@srmist.edu.in

Priyadarsini K[2*]
Department of Data Science and Business Systems
SRM Institute of Science and Technology
Kattankulathur, Tamilnadu, India
priyadak@srmist.edu.in

***Abstract*— A drought in agriculture refers to a shortage of moisture in relation to atmospheric droughts and meteorological conditions, as well as the effects these have on agricultural output and economic viability. Although the linkages are complex, regional research including a range of aspects relating to the moisture content of the soil is necessary to completely comprehend drought and it is extremely complicated. A sincere diagnosis of land use changes and the measures that correlate to it, must thus be developed. The analysis will be based on the average soil moisture of the United States by identifying the latitude, longitude and slope of each city recognized using FIPS. Machine learning algorithms are one of several techniques for evaluating meteorological drought using drought indices, and they are useful in assessing systems. The different methods which we are going to compare and scrutinize are DT, KNN, Naïve Bayes, Random forest and SVM**

Keywords- DT, KNN, Naïve Bayes, Random forest, SVM,Drought

I. Introduction

Droughts are long periods of low precipitation that are a component of the natural environmental cycle and can occur everywhere on Earth. This disaster is characterized by a gradual deterioration in water supply due to a lack of precipitation. Drought may have catastrophic effects on human health, agricultural production, the economy, energysources, and the natural world. Droughts are the primary factor of livestock and agricultural losses worldwide, impacting an estimated 55 third of the population every year. People's ability to provide for themselves is threatened, the likelihood of sickness and mortality is increased, and the refugee crisis is spurred by drought.
About 40% of the global population faces water constraintsand by 2030, drought might force 700 million people to relocate.

Because of greenhouse gasses, previously dry places are getting dryer, while previously wet ones are becoming wetter. This implies that in arid regions, where temperaturesare rising, water evaporates more rapidly, increasing the chance of drought or extending its duration. Eighty percent to ninety percent of all natural catastrophes in the previous 10 years have been caused by disasters, plagues, tropical cyclones, extreme weather events, and severe storms. It is sometimes difficult to pinpoint the exact beginning and conclusion of a drought, in contrast to more spectacular weather events. The initial signs of a famine may be hard toknow, and it may be several weeks or months already when people recognize that one has commenced. Similarly, it maybe difficult to determine whether a scarcity has ended.
Droughts can endure for moments, seasons, or even years. Droughts can endure for two decades or more in some areas. The longer a famine lasts, the greater the impact on people'slife

Some theories of meteorological drought classify dry intervals as such depending on the percentage of days with wetness below a set threshold, for instance. Tropical rainforests, humid subtropical temperatures, and humid mid-latitude climates are examples of environments where precipitation occurs year-round, making this statistic applicable exclusively there. A determination that relied on the number of occasions below a given threshold for precipitation would be absurd in these regions. Precipitation variations can also be defined in terms of their relationship tomonthly, seasonal, or yearly norms

II. Literature Survey

Pavan Mohan Neelamraju used a 117-year dataset from the Indian meteorological department to estimate the meteorological drought in India[6] They used a variety of analysis techniques, such as support vector machines (SVM) and decision trees (DT). When they examined both accuracylevels, they discovered that DT was more accurate than SVM.

Potential of Deep Learning in drought assessment by extracting information from hydrometeorological precursors was carried out by Rajib Maity;Mohd Imran Khan; Subharthi Sarkar; Riya Dutta; Subhra Sekhar Maity; ManaliPal;[10] by taking

Journal of Current Research in Engineering and Science
JCRES
Bi-Annual Online Journal (ISSN : 2581 - 611X)

meteorological dataset obtained from ERA5 and analyzed using CNN, SVR and DL based models. The DL based model outperformed among the othermodels

Drought prediction using machine learning algorithm was carried out by Aishwarya M Iyengar,Deepika, Kanthi, Utkarsha Bharat, Mitaigar in 2019.[5] They self acquired the dataset from the field and used Support Vector Machine

to propose a model, they were successful in developing amodel for prediction of when drought will take place.

Mustafa A. Alawsi, Salah L. Zubaidi, and Nabeel Saleem, Khalid Hashim, Saad Al-Bdairi, and Nadhir Al-Ansari conducted a review and assessment of the hybrid techniquesand data pre-processing for drought forecasting.[8] They employed a variety of methods, including decision trees, regression, random forests, and artificial neural networks (ANNs). They also integrated other models to produce hybrid models, and they calculated the MAE and RMSE values for each. MAE, RMSE, and R2 values were compared, and it was determined that hybrid models are superior.
Puyu Fenga, Bin Wangb, De Li Liub, and Qiang Yu conducted machine learning-based integration of remotely sensed drought indicators in South-Eastern Australia in 2019 using MODIS satellite data.[16] It was discovered that the combination of RF and bias-correction

III. METHODOLOGY

A. Data Insights

For the inspection to be consistent and reliable, a large amount of data is usually required. A big dataset ensures that your investigation can filter out irregular data and therefore that your size of the sample is realistic. It also ensures that observed patterns are not flukes and can account for change efforts. Forecasting, the practice of generating predictions about the future given the past, may also benefit from data. Here we have collected the dataset having weather and soil data. By using the US country FIPS code, you will draw the latitude, longitude, elevation and slopes.

We will be spotting 6 slopes based on the ranges.

Slope1= 0 % ≤ slope ≤ 0.5 %
Slope2= 0.5 % ≤ slope ≤ 2 %
Slope3= 2 % ≤ slope ≤ 5 %
Slope4= 5 % ≤ slope ≤ 10 %
Slope5= 10 % ≤ slope ≤ 15 %
Slope6= 15 % ≤ slope ≤ 30 %

LONGITUDE

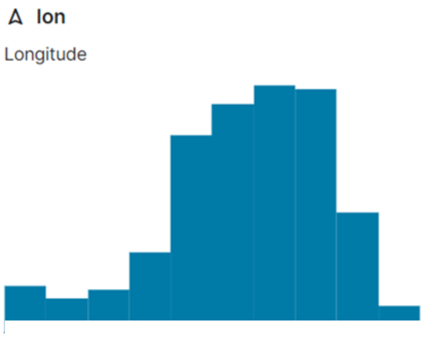


Fig.2 Longtitude of the land

GRAPHS LATITUDE

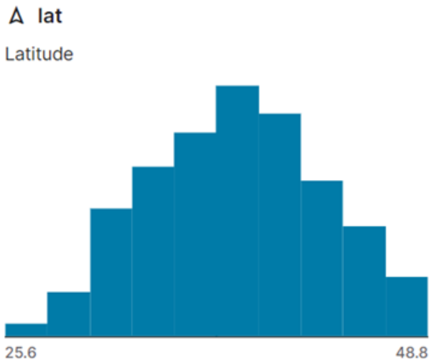


Fig.1 Latitude of the land

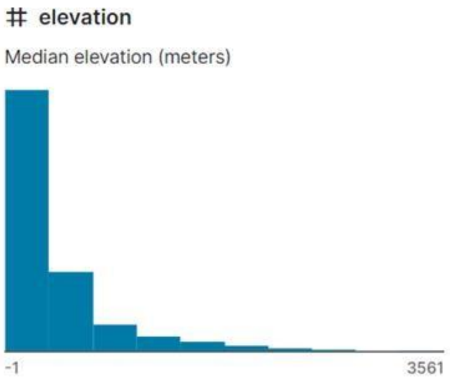


Fig.3 Elevation of the land

Volume7-Issue1, January 2024
Paper: 33

SLOPE1

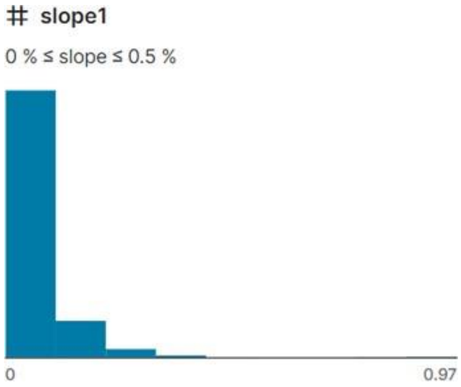


Fig 4. Slope of 0-0.5

SLOPE2

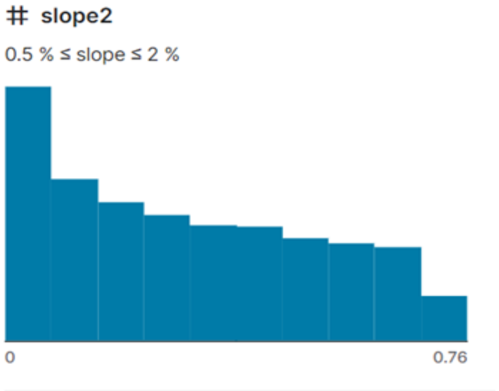


Fig 5 Slope of 0.5-2

SLOPE3

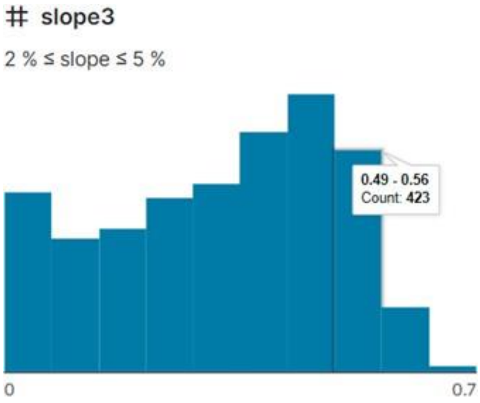


Fig 6 Slope of 2-5

SLOPE4

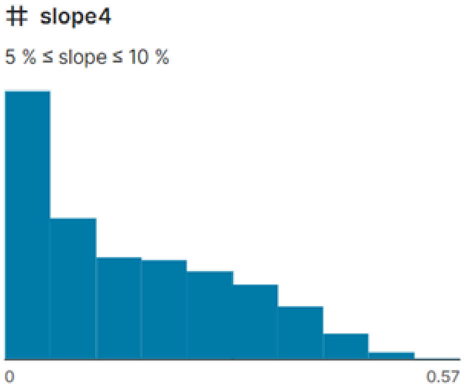


Fig 7 Slope of 5-10

SLOPE5

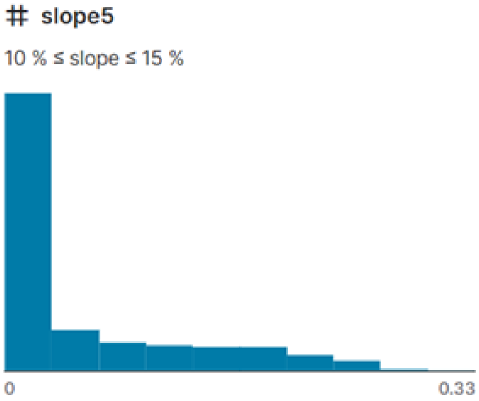


Fig 8 Slope of 5-10

SLOPE6

F

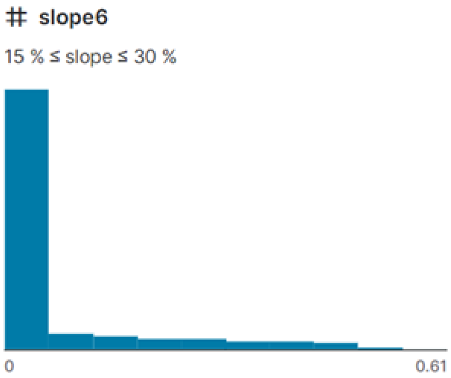


Fig 9. Slope of 15-30

B. Model Development

Step 1: First will be importing the required libraries likesklearn, pandas, numphy, imblearn, matplotlib etc

Step 2: Followed by that we read the input data and initialexploation and data wrangling(cleaning) is done.

Step 3: Exploratory data analysis (Univariate and Bivariateanalysis is done)

Step 4: The next is we have identify the outliers and removevalues beyond the standard outlier limit.

Step 5: Extracting of independent and dependent variables isdone

Step 6: Correlation between independent variables is foundfor feature selection.

Step 7: Then comes the splitting and training the data.Step 8: Standardizing the data.

Step 9: Fixing class imbalance occurs by Upsampling using SMOTE, Downsampling using neighborhood cleaning rule and Downsampling using near miss.

Step 10: PCA and LA Dimensionality reduction occurs.

Step 11: Finally the model is developed using KNN, DT, Naïve Bayes, Random forest and SVM

KNN Model

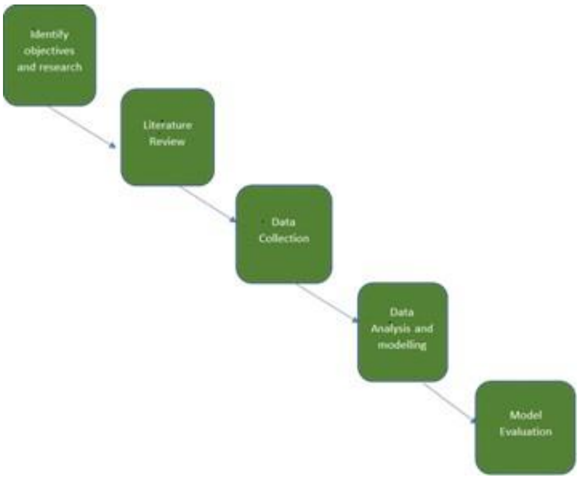


Fig 10 Flowchart

A non-parametric, supervised learning classifier, the k- nearest neighbours algorithm, provides predictions or classifications about how a single data point will be categorized. In order to categorize a query point, the k- nearest neighbour approach seeks out the point's closest neighbours. Due to the KNN algorithm's remarkably accurate predictions, it can compete with the most precise models.. The accuracy of the forecasts is influenced by themeasurement of distance.

The distance between (x1, y1) and (x2, y2) in Euclideanspace is,

$$d(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

Decision Tree

A decision tree is a supervised learning technique that is used in both classification and regression. Its tree-like structure consists of a root , branches, internal, and leaf nodes. The leaf nodes or terminal nodes, which are represented by both types of nodes, perform assessments based on the attributes that are available to produce homogeneous groupings. The dataset's leaf nodes act as representations for all results. Since they divide complex data into more digestible bits, decision trees are very usefulfor machine learning and data analytics.

Information gain estimates the degree of uncertainty that acertain feature reduces and decides which characteristic should be used as a decision node or root node.

$$Information\ Gain\ =\ E(Y)\ -\ E(Y|X)$$

Entropy, represented by the letter E, is effectively a count of the contaminants contained in a node. The amount of contamination reveals how random our data are. A "yes" or a "no" response should be given when there is a pure sub- split.

$$E(S)\ =\ -p_{(+)}\log p_{(+)}\ -\ p_{(-)}\log p_{(-)}$$

S is a subset of the training example, and p+ and p- are probabilities of positive and negative classes, respectively.

Naïve Bayes

One of the fundamental machine learning methods that aids in categorizing issues is the Naive Bayes approach. The simplicity and efficiency of the Naive Bayes algorithm are well-known. Model creation and prediction can be done more quickly with this method. Any ML model should be created using the Bayes theorem. It is used to determine thelikelihood of a hypothesis with prior knowledge. Yes or No,depending on the conditional likelihood

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

in which P(A|B) is the Probability of hypothesis A based on the observed event B, and P(B|A) is the posterior probability likelihood of the information provided that a hypothesis is likely to be correct.

Random Forest

One popular strategy for classifying and predicting data is the supervised machine learning method known as "random forest".The tasks that include regression and classification benefit from its improved performance. While using the Random Forest Algorithm to tackle regression issues, you must be aware of the mean squared error in order to identifyhow your data branches(MSE).

$$MSE = \frac{1}{N}\ \sum_{i=1}^{N}(fi - yi)^2$$

here N is total number of data points with fi is the model'soutput, and yi is the data point's actual value.
SVM
Regression, classification, and outlier identification all involve support vector machines (SVMs), a class of supervised learning techniques. SVM selects the points andvectors that are at extreme to construct the hyperplane. Thesupport vectors used to represent these severe events give the Support Vector Machine technique its name.

$$h(\boldsymbol{x}_i) = sign(\ \sum_{j=1}^{s} \alpha_j\ y_j\ K(\boldsymbol{x}_j, \boldsymbol{x}_i) + b\ \ )$$

$$K(\boldsymbol{v}, \boldsymbol{v}') = \exp(\frac{||\boldsymbol{v} - \boldsymbol{v}'||^2}{2\gamma^2})$$

Xi is to be predicted. Each data of xj's (-1 or +1) is given bythe yj. Aj is the constant for every xj.b is a single numerical constant . S stands for support vectors. The K is a kernel function.

C. Performance evaluation

The performance of a machine/deep learning model can beevaluated using the following parameters:

Accuracy: It is the total of all actual results as compared to expectations. It suggests that all model predictions are takeninto account when dividing the total forecasts and are taken to be accurate. The formula to calculate it is

$$\text{Accuracy} = \frac{(\text{TP} + \text{TN})}{(\text{TP} + \text{FP} + \text{TN} + \text{FN})}$$

Precision: It is the ratio of the number that was truly predicted or that was genuinely forecasted to all of the actual predictions. the model's ability to forecast each trueclass properly. The formula to calculate it is:

$$Precision = \frac{TP}{TP + FP}$$

The percentage of all data samples for a class—the positiveclass that a machine learning model correctly identifies as being a member of that class is known as recall. The calculation method is as follows:

$$\textbf{Recall} = \frac{\text{True Positive}}{\text{Predicted Results}}$$

F-score: The F-measure or F1-score is a measure of precision and recall that takes into account harmonic average. While FP and FN are both significant, accuracyand recall are taken into account using a single score.

$$F_1\text{–score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}}$$

IV. Results

The aim of the project was to compare different models for drought analysis and propose the best method for it. We used various methods such as Decision tree, KNN and Random Forest. Physically measurable indices such as precipitation, streamflow and groundwater. These indices can be taken with real time value data. The data we carried our project included information about the soil in different area and it had information such as latitude longitude median elevation slope etc and depending upon all these categories the best model was found out.

A. Decision Tree

Various algorithms were carried out in Decision Tree suchas DT without resampling, DT with SMOTE Upsampling and LDA, DT with Near miss downsampling and LDA etc. The best accuracy in DT was with DT without resampling with an accuracy of 0.7627

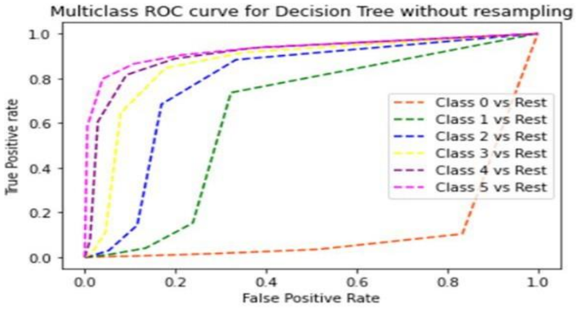


Fig 11 Decision Tree

Hyperparameter Tuning was also carried on

B. KNN algorithm

In case of KNN, the highest accuracy was in KNN withoutresampling with the highest accuracy of 0.798

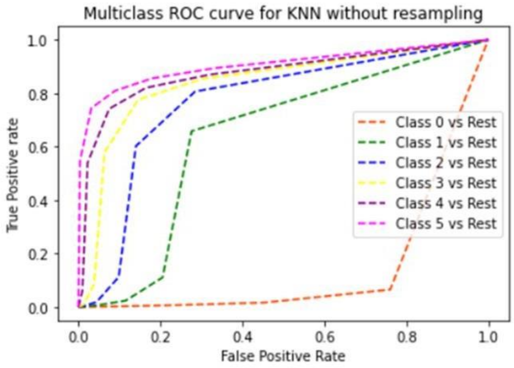
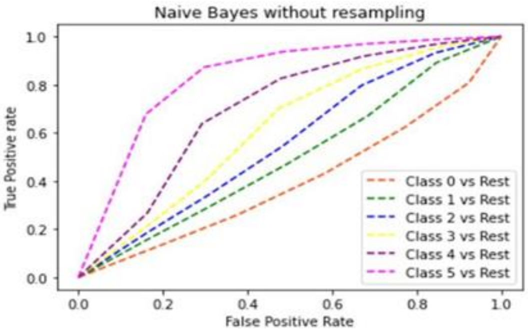


Fig 12 KNN

Hyperparameter tuning was carried out for the same

C. Naves Bayes

Naves Bayes had an accuracy of only 0.585



D. Random forest

Random forest without resampling had the highest accuracyof 0.808 among all the models

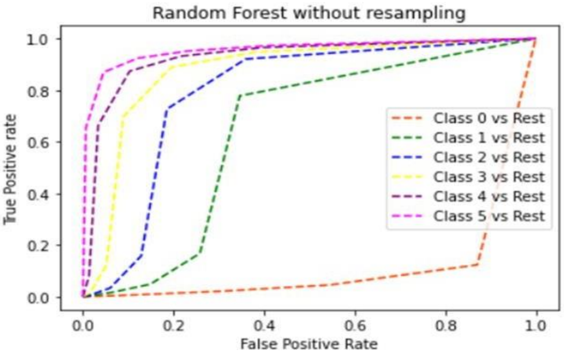


Fig 14 Random forest

E. SVM

SVM with near miss dawnsampling had the least accuracyof 0.29

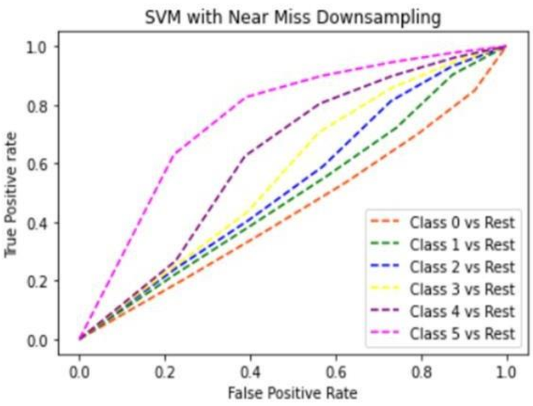


Fig.15 SVM

V. CONCLUSION

Agriculture production is a key component of a country's economy. Therefore, it is crucial to assess and forecast draught. By enabling them to take the necessary actions at theright time, this evaluation will be able to save a significant number of crops. A precise model that aids in draught assessment can be created using techniques from decision tree methods. The best model out of KNN, DT, SVM and Naïve Bayes is Random Forest without resampling with an accuracy of 0.80 and precision of 0.79 recall of 0.80 and FI Score of 0.798. So in further analysis it can be done with random forest only.

VI. FUTURE SCOPE

Since now the data covers only US land, the drought can bepredicted only in that country, but we can do the same for soil from all over the world. For this we need to get more data from around the world. By increasing the dataset, the model will also have higher accuracy and reliability. Also this research is shortlisting the best model making it easy to analyse. In this way soil from all around the world can be analysed just using Random Forest and drought can be predicted easily. This will help in improving farmers' lives.It can be also carried on by partnering with NGOs that helpfarmers.

REFERENCES

[1] Basack S, Bhattacharya AK, Maity P (2013) A coastal groundwater management model with Indian case study. Proc Inst Civ Eng Water Manage 167(WM3):126–140

[2] Bachmair S, Stahl K, Collins K, Hannaford J, Acreman M, Svoboda M, Knutson C, Smith KH, Wall N, Fuchs B, Crossman ND, Ian Overton C (2016) Drought indicators revisited: the need for a wider consideration of environment and society. Wires Water 3(4):516–536

[3] Aksoy H, Cetin M, Eris E, Burgan HI, Cavus Y, Yildirim I, Sivapalan M (2021) Critical drought intensity-duration-frequency curves based on total probability theorem-coupled frequency analysis. Hydrol Sci J 66(8):1337–1358

[4] Altin TB, Saris F, Altin BN (2020) Determination of drought intensity in Seyhan and Ceyhan River Basins, Turkey, by hydrological drought analysis. Theor Appl Climatol 139:95–107

[5] Aishwarya M Iyengar ,Deepika K ,Kanthi Utkarsha Bharat ,Mitaigar Divya ,Vaidehi M , "Drought Prediction using Machine Learning Algorithm" , International Journal of Advances in Computer Science and Cloud Computing (IJACSCC) , pp. 1-6, Volume-7,Issue-1(2019)

[6] Meteorological Drought Assessment in India using Machine Learning Technique, International Journal of Science and Research (IJSR) 10(7):592 - 598(July 2021)

[7]. The Use of Time Series Analysis Techniques in Forecasting Meteorological Drought February 1974

[8]. Drought Forecasting: A Review and Assessment of the Hybrid Techniques and Data Pre-Processing by Mustafa A. Alawsi ,Salah L. Zubaidi, Nabeel Saleem Saad Al-Bdairi,,Nadhir Al-Ansari and Khalid HashimHydrology 2022, 9(7), 115; .

[9] Nhu, V.H., Zandi, D., Shahabi, H., Chapi, K., Shirzadi, A., Al- Ansari, N., Singh, S.K., Dou, J., Nguyen, H. Comparison of Support Vector Machine, Bayesian Logistic Regression, and Alternating Decision Tree Algorithms for Shallow

Landslide Susceptibility Mapping along a Mountainous Road in the West of Iran. Appl. Sci. 2020,

[10] Rajib Maity, Mohd Imran Khan ,Subharthi Sarkar ,Riya Dutta,Potential of Deep Learning in drought assessment by extracting information from hydrometeorological precursors Journal of Water and Climate Change 12(14–15) OI:10.2166/wcc.2021.062., 2021

[11] Vapnik, V. N and Cortes, C., 1995, Support Vector Networks. Machine Learning,

[12] Taghi Sattari, M., Anli, A.S., Apaydin. H. and Kodal, S., Decision trees to determine the possible drought periods in Ankara.Atmosfera [online]. 2012

[13] Pereira, L.S., Cordery, I. and Iacovides, I., 2002, Coping with water scarcity, UNESCO IHP VI, Technical Documents in Hydrology

[14] Nadarajah, S., 2009, A bivariate Pareto model for drought.Stoch. Environ. Res. Risk Assess

[15] Shewale, M.P. and Shravan Kumar, 2005, Climatological Features of Drought Incidences in India, Climatology No. 21, Indian Meteorological Department, Govt. of India

[16] Comparison of two model approaches in the Zambezi river basin with regard to model reliability and identifiability, Hydrol. Earth Syst. Sci., 10(3), 339-352, doi:10.5194/hess10-339-2006

[17] Adams, D. K., and A. C. Comrie (1997), The north American monsoon, Bull. Am. Meteorol. Soc., 78(10), 2197-2213

[18] Bosch, D. D., V. Lakshmi, T. J. Jackson, M. Choi, and J. M. Jacobs (2006), Large scale measurements of soil moisture for validation of remotely sensed data: Georgia soil moisture experiment of 2003, Journal of Hydrology, 323(1-4), 120-137

[19] Cordery, I., and M. McCall (2000), A model for forecasting drought from teleconnections, Water Resour. Res., 36(3), 763-768, doi:10.1029/1999wr900318

[20] Machine learning-based integration of remotely-sensed drought factors can improve the estimation of agricultural drought in South-Eastern Australia,Puyu Feng, Bin Wang, De Li Liu , Qiang Yu.2019

[21] Heim , R., Jr. (2002), A review of twentieth-century drought indices used in the United States, Bull. Am. Meteorol. Soc., 83, 1149-1165.

[22] Ledieu, J., P. De Ridder, P. De Clerck, and S. Dautrebande (1986), A method of measuring soil moisture by time-domain reflectometry, J.Hydrol., 88(3), 319-328

[23] Luo, L., and E. F. Wood (2007), Monitoring and predicting the 2007 U.S. drought, Geophys. Res. Lett., 34(22), L22702, doi:10.1029/2007gl031673.

[24] Mishra, A. K., and V. P. Singh (2010), A review of drought concepts, J. Hydrol., 391(1–2), 202- 216,

[25] Kumar, V., and U. Panu (1997), Predictive assessment of severity of agricultural droughts based on agro-climatic factors JAWRA Journal of the American Water Resources Association, 33(6), 1255-1264.