RESEARCH ARTICLE OPEN ACCESS

Rainfall Prediction Using Classification and Clustering Complex Data Science Models with Geological Significance

Deepak Sharma [1], Dr. Priti Sharma [2]

[1] Research Scholar, [2] Assistant Professor
Department of Computer Science & Applications, M. D. University, Rohtak-India

ABSTRACT

Prediction of rainfall is indeed an important task in today's world. Agricultural sector largely depends upon the water concluded from the rainfall for filling up the need of water for crop cultivation. Abnormal rainfall always leads to the crisis for the agricultural sector. The effect of unfavourable rainfall can be minimized by accurately predicting it and so on time. In this way on can figure out a plan by knowing the somehow accurate value of the rainfall and can dodge the bullet. In this paper some of the best and widely accepted techniques are discussed in depth. After critical analysis of the techniques some significant improvement are proposed with can be useful for enhancing the accuracy.

Keywords: Data mining, Bayesian Classifier, Clustering, Rain fall prediction, Linear Regression Technique, K-fold, Weather predictions, Multiple Regression Technique.

I. INTRODUCTION

Indian economy depends largely on agriculture which is roughly 20.5% of the total GDP. Due to poor irrigation facility, most of the agriculture tasks depends on rain [5]. Rainfall affects the crop yield which ultimately effects the economy of the country. It is very essential to accurately forecast rainfall because if the forecast is not matching the demands some other arrangements can be done for harvesting of the crops so that the overall affect can be neutralized. The techniques available for forecasting of rainfall can be divided in to two categories. One is Dynamic approach and the other named as empirical approach [9].

In dynamical approach, Physical models are generated on the basis of combination of equations. These physical models after analysing the initial atmospheric conditions forecast the progression of global climate [12]. In empirical approach deep investigation of historical climate data and its dependency on a collection of many different attributes of atmosphere.

Meteorological/climatic data mining is a type of data mining that deals with large meteorological data with the intent of finding hidden patterns, so that the retrieved information can be used for learning. Weather is one of the climatic data that contains knowledge .Rainfall is the most important climatic element which impacts both agricultural & non-agricultural sectors. Therefore prediction of rainfall becomes an important issue for country's economy. Precipitation expectation is a vital piece of climate forecast.

All manuscripts must be in English. These guidelines include complete descriptions of the fonts, spacing, and related information for producing your proceedings manuscripts.

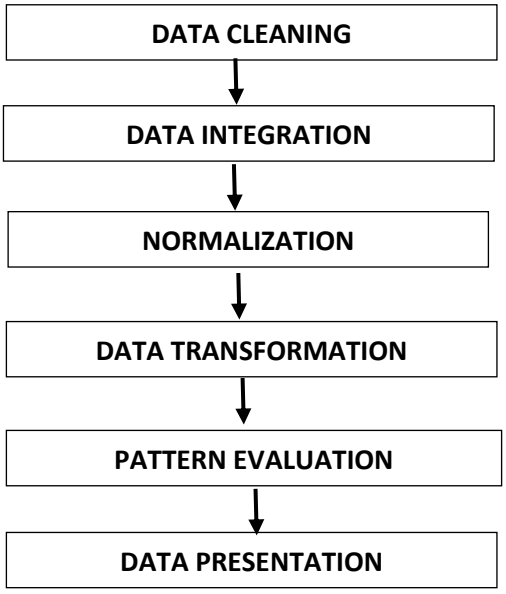

Fig. 1 Process of data mining

Data mining refers to the task of analysing great deal of information with the intent of finding hidden patterns and trends that don't seem to be directly apparent from summarized data. Data mining and information extraction is becoming more and more necessary and helpful because the quantity and complexity of information is quickly increasing.

Data mining normally involves four categories of tasks: Classification arranges the information into predefined groups, clustering-is comparable to classification, however the groups don't seem to be predefined, that the algorithmic rule can try and cluster similar things together, Regression-tries to seek out a function that

models the information with the smallest amount of error and Association rule learning searches for relationships between variables. Data mining has been outlined as-the nontrivial extraction of implicit, previously unknown, and potentially useful data. It employs numerous techniques like supervised or unsupervised learning techniques, so as automate the retrieval of information and derive patterns which may be used for prediction.

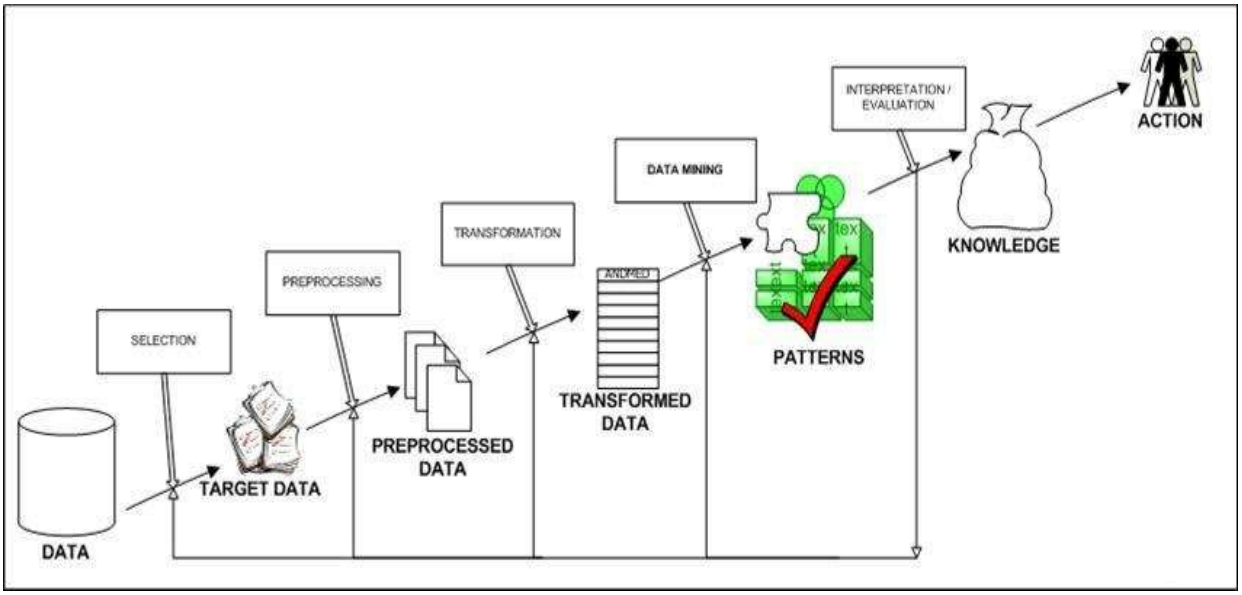

Fig. 2 Outline of data mining

Data mining is performed on information represented in quantitative, textual, or transmission forms. Data mining applications will use a spread of parameters to look at the information. They will include numerous association patterns wherever one event is connected to a different event, like buying a tooth paste and buying tooth brush, sequence or path analysis(patterns wherever one event ends up in another event, like returning of joyous sessions and buying of cloths), classification (identification of recent patterns), prediction (discovering patterns from that one will build affordable predictions relating to future activities), and cluster(finding and visually documenting groups of previously unknown facts).

In this paper, some of the best known work in the prediction of rainfall is discussed in depth. Some of the best known techniques are compared and models are critically analysed in a tabled format to help a better understanding of the field.

II. EXISTING WORK

Critical analysis of the exiting work is the most important part of any research work because it leads us to the anomalies and challenges of the research work which was already faced by the researchers. By knowing these challenges one can simply know where to hit the target and it will be easy to formulate a new model based on the limitations of the exiting one.

Chandrasegar Thirumalai et al. [16] proposed a model for the prediction of rainfall using linear regression. There are mainly three crop seasons RABI, KHARIF AND ZAYAD. In this work, Rainfall for one crop season was predicted by the use of actual rainfall of another season.

In Linear regression, a linear model of dependent and independent variables was made and after that value of dependent variable was predicted based on the value of independent variable. In this work, using the actual values of the RABI season, values for other two season which are ZAYAD and KHRIF were predicted.

TABLE 1.
CROPS OF DIFFERENT SEASONS [16]

ZAYAD (MARCH TO JULY)	KHRIF (JULY TO OCTOBER)	RABI (OCTOBER TO MARCH)
Sugar cane	Rice	Wheat
Cucumber	Sorghum	Oats
Rapeseed	Groundnut	Onion
Sunflower	Jowar	Tomato
Rice	Soya bean	Potato
Cotton	Bajra	Peas
Oilseeds	Jute	Barley
Watermelon	Maize	Linseed
muskmelon	Cotton	Mustard oil seeds
	Hemp	Masoor
	Tobacco ragi	
	Millet	
	Arhar	

Wassamon Phusakulkajorn et al. [19] proposed a rainfall prediction model using Artificial Neural Network and wavelet decomposition for the prediction of daily rainfall with the help of previous data. This work shows the evidence that Artificial Neural Network (ANN) has the fine ability for the representation of complex nonlinear relations using input and output variables.
The performance of the model which is based on the artificial neural network are expressed in terms of R^2 and Root Mean Square Error (RMSE). Rainfall prediction is indeed a complex process because it depends on large categories of complex non-linear data. Artificial neural network (ANN) has the ability to deal with complex non-linear data.
The model proposed in this work has the capacity of prediction of rainfall for 4 consecutive days with the prediction accuracy of R2 =0.8819 and RMSE=4.6912mm.

Niketa Gandhi et al. [17] proposed a model for the prediction of yield of cultivated rice for the state of Maharashtra using Bayesian Network (Classifier). A total of 27 Districts of the state of Maharashtra were taken for consideration. Also collection of data is done by deeply analysing publicly available government records about the yield of rice crop.

TABLE 2.
PARAMETER UNDER CONSIDERATION [17]

PARAMETER UNDER CONSIDERATION	UNIT FOR MEASUREMENT
Precipitation	Mm
Minimum Temperature	Celsius
Average Temperature	Celsius
Maximum Temperature	Celsius
Production of rice	Tonnes
Yield	Tonnes/Hectare
Area	Hectares

Nikhil Sethi et al. [15] proposed a rainfall prediction model using empirical statistical technique (Multiple linear regression). The data set used in this research work contains 30 years of data ranging from year 1973 to year 2002. Some of the focused attribute of the data set are rainfall, precipitation, vapour pressure, average temperature and cloud cover. This model mainly focused for UDAIPUR CITY, Rajasthan, India.

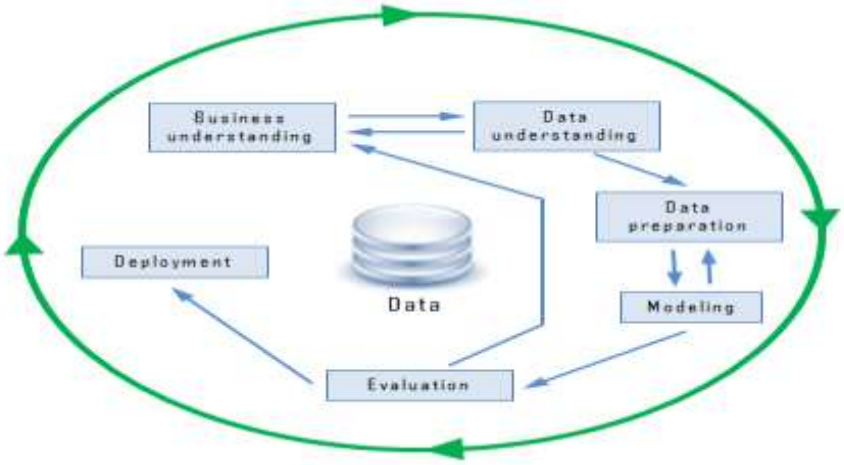

Figure. 3 Outline of the model used [15]

This model forecasts monthly rainfall for the month of July in mm. In this work Empirical statically technique is used. After pre-processing of collected data the next step is to reduce the predicators which have high level of inter correlation otherwise it will adversely affect the ability of the predictive model. After that Next step was to train the model using training data. Technique used to training the model is Linear Regression techniques. In this work Rainfall prediction is done only for UDAIPUR city (A small area). Model can be applied on other geographical location and its robustness should be tested.

Suvidha Jambekar et al. [18] proposed a model for predicting the future production of different crops after analyzing some very crucial parameter of existing data using multiple linear regression, Random forest regression and Multivariate Adaptive Regression Splines. Data set used in this work was a secondary data set which is collected from officially available government record of the authorities.

The data set collected contains the data for almost 64 years ranging from 1950 to 2013[20]. It is shown in the work that the performance of multivariate adaptive regression splines (Earth) was better compared to multiple linear regression. Parameter which are taken under consideration for the research work are-

TABLE 3.
PARAMETER UNDER CONSIDERATION [18]

PARAMETER UNDER CONSIDERATION	UNIT FOR MEASUREMENT
Mean temperature	Degree Celsius
Area	Million Hectares
Area under irrigation	Percentage
Production	Million Tonnes
Yield	Tonnes/ Hectares

III. CRITICAL ANALYSIS OF EXISTING WORK

In this section of this paper, a more crisp cum highlighted tabular analysis is given. Few research gaps are outlined which might increase the efficiency and improve the accuracy of already existing models.

TABLE 4.
ANALYSIS OF EXISTING WORK

Pub. Year	Title of the Paper	Techniques Used	Research Gaps
2014	Exploiting Data Mining Technique for Rainfall Prediction [15]	Empirical statistical technique (Multiple Linear Regression)	Rainfall prediction is done for UDAIPUR city only (A small area).
2017	Heuristic Prediction of Rainfall Using Machine Learning Techniques [16]	Linear Regression	Only last one year of data is used to train the model which is very small for identifying relationships accurately between the variables.
2016	Predicting Rice Crop Yield Using Bayesian Networks [17]	Bayesian Network (Classifier)	This model focuses on only Rice crop, a more generic and robust model can be made.
2018	Prediction of Crop Production in India Using Data Mining Techniques [18]	Multiple linear regression, Random forest regression and Multivariate Adaptive Regression Splines	Less number of independent variables are used for the prediction of the dependent variable.
2009	Wavelet-Transform Based Artificial Neural Network For Daily Rainfall Prediction in Southern Thailand [19]	Wavelet decomposition and Artificial Neural network (ANN)	This model fails to predict rainfall in some selected areas.

IV. PROPOSED APPROACH

In this section, after observing the results and accuracy of the models some approaches have been proposed to make better model in respect of accuracy and computations.

There is always a chance of improvement in any techniques or model. In complex data analysis both efficiency and accuracy are important factor. One should not over ignore efficiency for better accuracy and vice versa.

TABLE 5.
PROPOSED APPROACHES

Pub. Year	Title of the Paper	Authors	Future scope
2014	Exploiting Data Mining Technique for Rainfall Prediction [15]	Nikhil Sethi, Dr.Kanwal Garg	Model can be applied on other geographical location and its robustness should be tested.
2017	Heuristic Prediction of Rainfall Using Machine Learning Techniques [16]	Chandrasegar Thirumalai, M Lakshmi Deepak, K Sri Harsha, K Chaitanya Krishna	Accuracy of the model can be improved by removing outliers efficiently during the pre-processing phase.
2016	Predicting Rice Crop Yield Using Bayesian Networks [17]	Niketa Gandhi, Owaiz Petkar, Leisa J. Armstrong	Accuracy of model can be improved by selecting more data attributes.

2018	Prediction of Crop Production in India Using Data Mining Techniques [18]	Suvidha Jambekar, Shikha Nema, Zia Saquib	Accuracy of the prediction can be improved by selecting more appropriate number of predictors for predicting the predictant.
2009	Wavelet-Transform Based Artificial Neural Network For Daily Rainfall Prediction in Southern Thailand [19]	Wassamon Phusakulkajorn, Chidchanok Lursinsap, Jack Asavanant	Accuracy of the model can be improved by improving the data pre-processing phase.

V. CONCLUSION

"Data is the new fuel" have you heard this phrase? Probably yes. Data mining is one of the most trending topics of today's world. In this paper models for the prediction of rainfall has been discussed and critically analyzed. Furthermore futuristic ideas have been suggested to improve the limitations of the model and increase the accuracy. In this way the possibilities for creation of new and better model are induced which ultimately results in better perdition.

This paper not only contains some of the best and trusted approaches of modern day data mining for rainfall prediction but also contain some of the finest work from recent times and some modern ideas to improve the accuracy and performance of these modern approaches.

REFERENCES

[1] Aswini.R, Kamali.D, Jayalakshmi.S, R.Rajesh, "Predicating Rainfall and forecast weather sensitivity using data mining techniques", International Journal of Pure and Applied Mathematics, Vol. 119 No. 14 2018.

[2] Chowdari K.K, Dr. Girisha R, Dr. K C Gouda, "A study of rainfall over India using data mining", International Conference on Emerging Research in Electronics, Computer Science and Technology – 2015.

[3] Sandeep Kumar Mohapatra, Anamika Upadhyay, Channabasava Gola, "Rainfall prediction based on 100 years of Meteorological Data", IEEE-2017.

[4] Tharun V.P, Ramya Prakash, S. Renuga Devi, "Prediction of Rainfall Using Data Mining Techniques", 2nd International Conference on Inventive Communication and Computational Technologies, IEEE-2018.

[5] Abishek.B, R.Priyatharshini, Akash Eswar M, P.Deepika, "Prediction of Effective Rainfall and Crop Water Needs using Data Mining Techniques", International Conference on Technological Innovations in ICT For Agriculture and Rural Development, IEEE-2017.

[6] Chandrasegar Thirumalai, M Lakshmi Deepak, K Sri Harsha, K Chaitanya Krishna, "Heuristic Prediction of Rainfall Using Machine Learning Techniques", International Conference on Trends in Electronics and Informatics ICEI 2017.

[7] inghui Qiu, Peilin Zhao, Ke Zhang, Jun Huang, Xing Shi, Xiaoguang Wang, Wei Chu, "A Short-Term Rainfall Prediction Model using Multi-Task Convolutional Neural Networks", International Conference on Data Mining, IEEE-2017.

[8] Valmik B Nikam, B.B.Meshram, "Modeling Rainfall Prediction Using Data Mining Method: A Bayesian Approach", Fifth International Conference on Computational Intelligence, Modelling and Simulation, IEEE-2013.

[9] A.Geetha, Dr. G.M.Nasira, "Data Mining for Meteorological Applications: Decision Trees for Modeling Rainfall Prediction", International Conference on Computational Intelligence and Computing Research, IEEE-2014.

[10] Soo-Yeon Ji, Sharad Sharma, Byunggu Yu, Dong Hyun Jeong, "Designing a Rule-Based Hourly Rainfall Prediction Model, IEEE IRI August 8-10, 2012.

[11] R. Sukanya, K. Prabha," Comparative Analysis for Prediction of Rainfall using Data Mining Techniques with Artificial Neural Network", International Journal of Computer Sciences and Engineering, ISSN: 2347-2693, IJCSE-2017.

[12] Fahad Sheikh, S. Karthick, D. Malathi, J. S. Sudarsan, C. Arun, "Analysis of Data Mining Techniques for Weather Prediction", Indian Journal of Science and Technology, Vol 9(38), ISSN (Print): 0974-6846, IJST-2016.

[13] Ramsundram N, Sathya S, Karthikeyan S, "Comparison of Decision Tree Based Rainfall Prediction Model with Data Driven Model Considering Climatic Variables", Irrigation Drainage Sys Eng, an open access journal ISSN: 2168-9768, 2016.

[14] Bhaskar Pratap Singh, Pravendra Kumar, Tripti Srivastava, Vijay Kumar Singh, "Estimation of Monsoon Season Rainfall and Sensitivity Analysis Using Artificial Neural Networks", Indian Journal of Ecology (2017) 44 (Special Issue-5): 317-322.

[15] Nikhil Sethi, Dr.Kanwal Garg, "Exploiting Data Mining Technique for Rainfall Prediction", (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (3), 2014, 3982-3984.

[16] Chandrasegar Thirumalai, M Lakshmi Deepak, K Sri Harsha, K Chaitanya Krishna, "Heuristic Prediction of

Rainfall Using Machine Learning Techniques", International Conference on Trends in Electronics and Informatics - ICEI 2017.

[17] Niketa Gandhi, Owaiz Petkar, Leisa J. Armstrong, "Predicting Rice Crop Yield Using Bayesian Networks", Intl. Conference on Advances in Computing, Communications and Informatics (ICACCI), Sept. 21-24, 2016, Jaipur, India.

[18] Suvidha Jambekar, Shikha Nema, Zia Saquib, "Prediction of Crop Production in India Using Data Mining Techniques", 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA).

[19] Wassamon Phusakulkajorn, Chidchanok Lursinsap, Jack Asavanant, "Wavelet-Transform Based Artificial Neural Network for Daily Rainfall Prediction in Southern Thailand", ISCIT 978-1-4244-4522-6/09 2009 IEEE.

[20] Deepak Sharma, Dr. Priti Sharma "Rain Fall Prediction using Data Mining Techniques with Modernistic Schemes and Well-Formed Ideas", International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-9 Issue-1, November 2019.

AUTHORS PROFILE

Deepak Sharma has completed his M.tech from C-DAC: Centre for Development of Advanced Computing, Ministry of Communications and Information Technology, Government of India affiliated from Guru Gobind Singh Indraprastha University, Delhi. He is pursuing Ph.D. in Computer Science at M. D. University, Rohtak since 2018. He has published more than 25 publications in various journals/ magazines of national and international repute. His main research areas include Data mining, Mobile Adhoc Network (MANET), wireless sensor network (WSN) and Internet of things (IoT).

Dr. Priti Sharma MCA, Ph.D. (Computer Science) is working as an Assistant Professor in the Department of Computer Science & Applications, M.D. University, Rohtak. She has published more than 50 publications in various journals/ magazines of national and international repute. She is engaged in teaching and research from the last 12 years. Her area of research includes Data mining, Big data, Software Engineering, Machine Learning.