Indian Journal of Data Mining (IJDM)
ISSN: 2582-9246 (Online), Volume-3 Issue-2, November 2023

Collection of Weather Data from Authentic Websites and Secondary Data Sources for Rainfall Prediction

Deepak Sharma, Priti Sharma

Abstract: ***The field of data mining and machine learning has been grown many folds from the last two decades. Almost every other problem can be solved using data mining and this becomes the most tempting part of it for the scientist and researchers all over the world. Data mining can be viewed as a process of discovering knowledge. This discovery of knowledge starts with the collection of data and ends with the acquired knowledge in the form of patterns. Data collection lays the foundation for the process of knowledge discovery. In this paper, various secondary data sources from where data can be collected for rainfall prediction are deeply studied and analyzed. Some of these authentic websites and secondary data sources are NCDC (National climate data center), Kaggle, Datahub.io, UCI machine learning repository, Earth Data etc. The data collected from these secondary data sources for rainfall prediction have been critically analyzed and compared on the parameters of Accuracy, Completeness, reliability, relevance, and timeliness.***

Keywords: ***Data Mining, Data Collection, Secondary Data Sources, Weather Data, Rainfall Prediction, Machine Learning.***

I. INTRODUCTION

Clive Humby, a notable mathematician, made the remark "DATA IS THE NEW OIL" in 2006. This notion was not widely recognized or relatable at the time, but now every multinational corporation craves data, and data is selling like crude oil. [1] The term "data mining" refers to the process of extracting relevant information from vast amounts of data. Because knowledge and meaningful patterns must be extracted from large data sets. Knowledge mining is a more appropriate name for this process. [2] Many researchers referred to data mining as knowledge discovery from data, or KDD. The most important step in the knowledge discovery process is data mining. This process is represented in Figure 1. [3]. Conventional categorizations of data mining methods include supervised and unsupervised approaches. Unsupervised techniques lack training data, but supervised techniques divide the data set into two parts: a training data set and a testing data set. [4]

The training data set is used to train the model, and the testing data set is used to validate the model. Regression and classification issues are resolved using supervised data mining techniques. The final prediction result for a classification problem is a class label, whereas the final prediction result for a regression problem is a numerical number. [5]

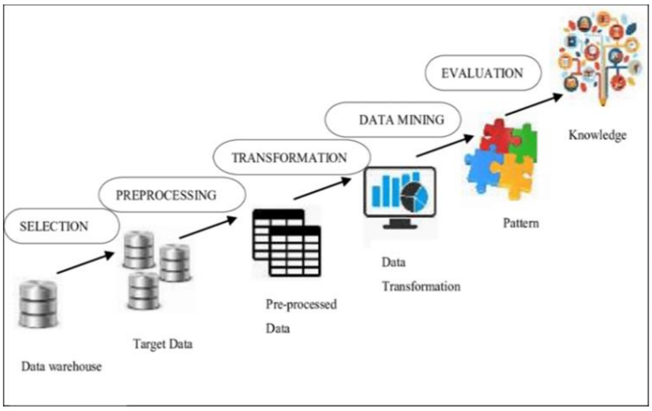

Figure 1: Process of Knowledge discovery (KDD)

The process of knowledge discovery starts with the process of acquiring related data. [6] Data which id to be used in the knowledge discovery process, is of two types i.e., Primary and Secondary data. Primary data, is the data which is particularly collected for a research work whereas secondary data, is already collected data which is available and can be used by other persons as well. [7] In this paper, various secondary data sources from where weather data set can be collected are discussed in depth and compared on the basis of accuracy, completeness, reliability, relevance and timeliness.

II. SECONDARY DATA SOURCES FOR WEATHER DATA

Data is the most powerful weapon of modern world. In this era of high accessibility one can find many secondary data sources available on the internet. [8] Some of most reliable and famous sources are National climate data center (NCDC), Kaggle, UCI Machine, Learning Repository, Earth Data, Google Dataset Search and Datahub.io. These secondary data sources are discussed one by one in this section. [9]

1. **National Climate Data Center (NCDC):** This weather database is managed by National centers for environmental information (NCEI).

Manuscript received on 09 May 2023 | Revised Manuscript received on 19 May 2023 | Manuscript Accepted on 15 November 2023 | Manuscript published on 30 November 2023.
*Correspondence Author (s)
Deepak Sharma*, Department of Computer Science & Applications, Maharshi Dayanand University, Rohtak (Haryana), India. E-mail: erdeepaksharmabwn@gmail.com, ORCID ID: https://orcid.org/0000-0002-7490-557X
Dr. Priti Sharma, Assistant Professor, Department of Computer Science & Applications, Maharshi Dayanand University, Rohtak (Haryana), India. E-mail: pritish80@yahoo.co.in

The world meteorological organization (WMO) has structured a world weather watch program in which weather data for over 9000 stations has been shared between various countries. [10] This database contains daily weather data for approximately 18 meteorological elements. [11] This database is free available on the website of National climate data center and data can be used freely for noncommercial, research and educational purpose without any registration.

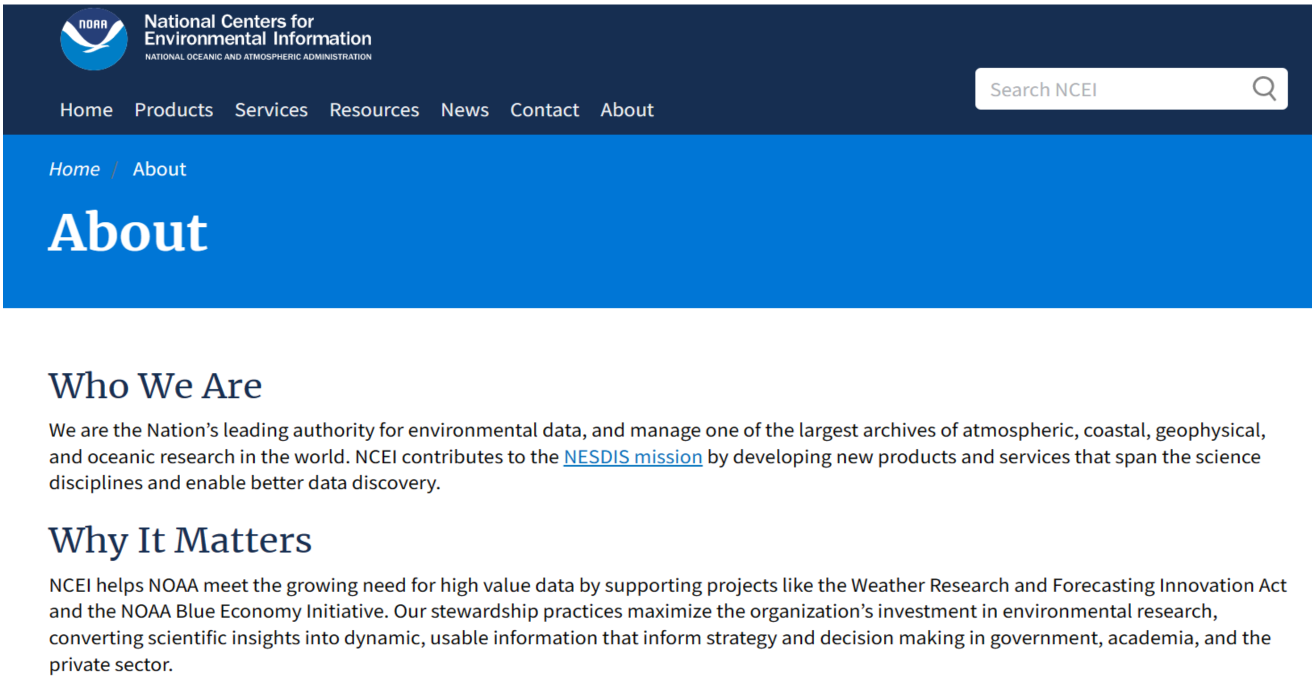

Figure 2: National Climate Data Center (NCDC)

2. **Kaggle:** Kaggle provides free data sets for use in machine learning. It works as a community hub. It was initially launched in 2010. [12] It does require a registration before downloading the dataset in order to gain some insight about the users. It contains many datasets and cover many of the data science topics.

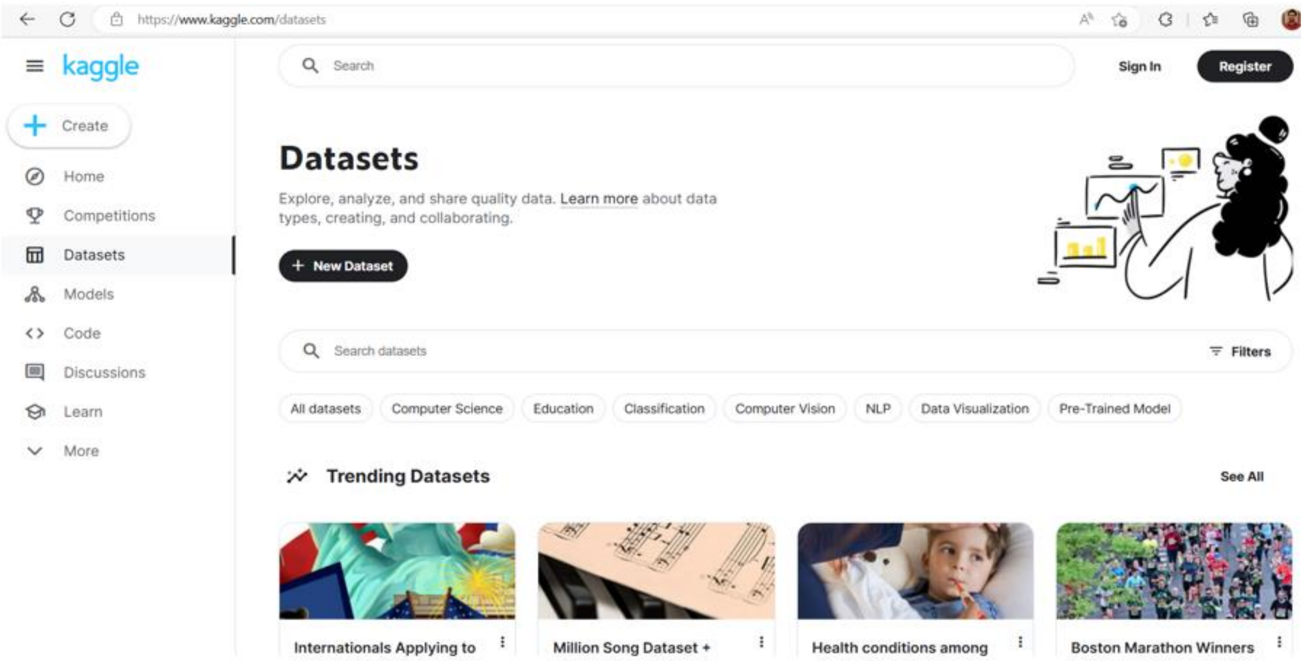

Figure 3: Kaggle

3. **UCI Machine learning repository:** UCI machine learning repository is managed by university of California Irvine. [13] It contains categorized data set of various machine learning problems such as classification, regression, and clustering etc. It provides data for free without any registration for students, teachers, and researchers. [14]

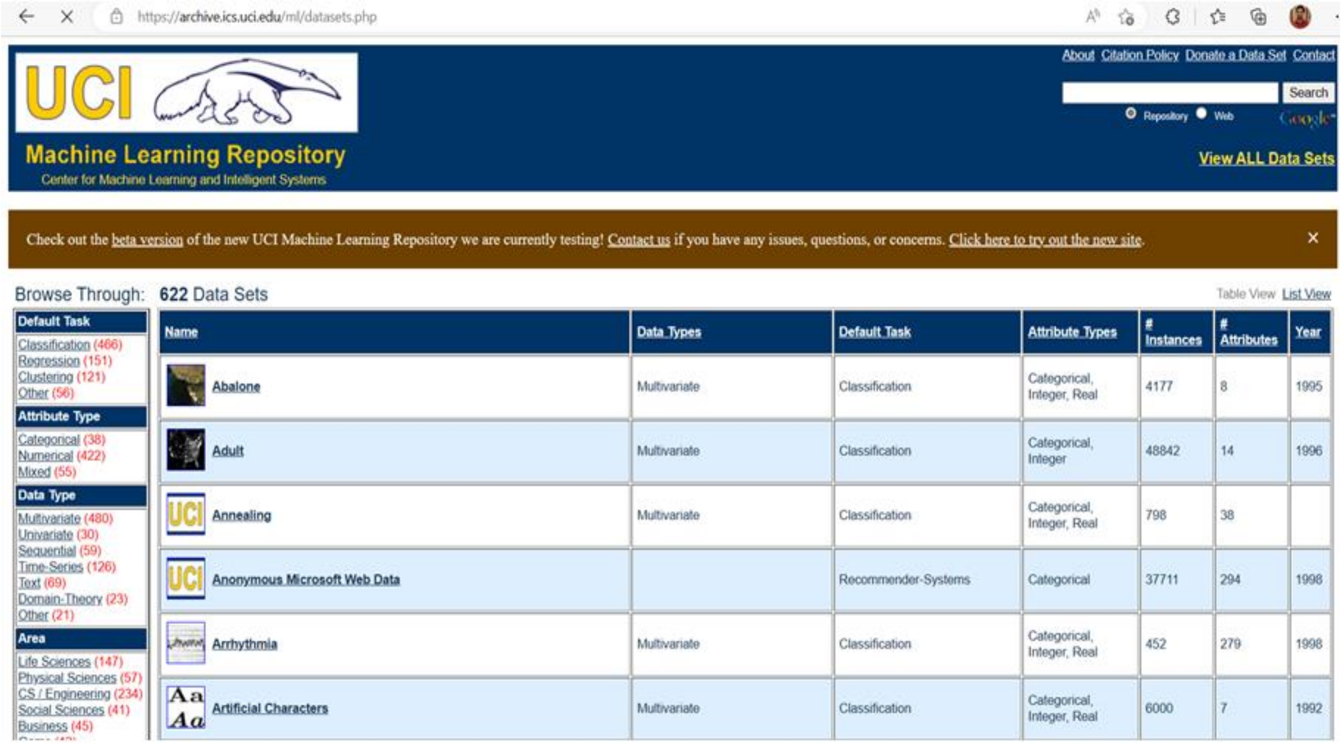

Figure 4: UCI Machine learning repository

4. **Earth Data:** Earth data contains data set related to various weather and climate measurements. It is managed by NASA. [15] Data is available freely without ay registration and anyone can download it with a basic internet connection. This secondary data source was first available in 1994 and becomes a popular choice in a short duration of time. It provides data regarding many atmospheric observations. [16]

Figure 5: Earth Data

5. **Google Data search:** Google data search has similarities with the google search engine. It comes into picture in 2018 and instantly become a hit like any other google product. [17] It provides cumulative data from many sources.

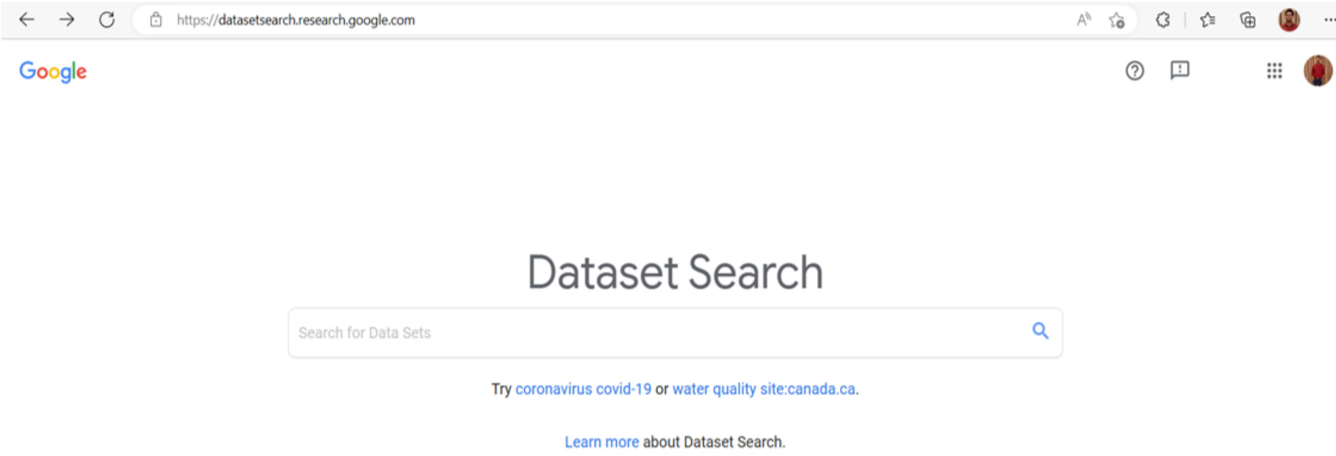

Figure 6: Google Data search

6. **Datahub.io:** It contains data mainly related to business and finance in order to help in decision making. It also contains data related to climate change and weather forecast. [18]

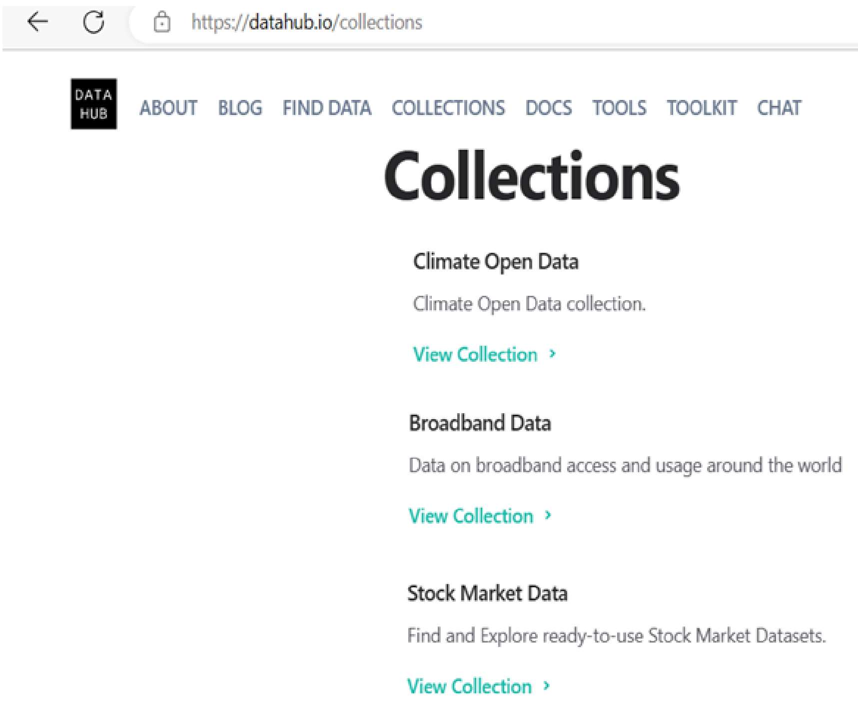

Figure 7: Datahub.io

III. COMPARATIVE ANALYSIS

Secondary data sources	Type of Data	Accessibility	Accuracy	Completeness	Reliability	Relevance
National climate data center (NCDC)	Weather related	Free, no registration required	Excellent	Above Average	Very Good	Highly relevant
Kaggle	Miscellaneous	Free, Registration required	Very Good	Average	Good	Relevant
UCI Machine Learning Repository	Machine learning, Others	Free, no registration required	Very Good	Average	Average	Relevant
Earth Data	Earth Science	Free, no registration required	Good	Average	Good	Relevant
Google Dataset Search	Miscellaneous	Free, Registration required	Very Good	Average	Good	Relevant
Datahub.io	Business and Finance	Free, no registration required	Good	Average	Average	Less Relevant

Retrieval Number:100.1/ijdm.B1626113223
DOI:10.54105/ijdm.B1626.113223
Journal Website: www.ijdm.latticescipub.com

Published By:
Lattice Science Publication (LSP)

IV. CONCLUSION AND FUTURE SCOPE

Collecting and storing of data is a very tedious and cumbersome process. It requires many resources and plaining. [19] Also, providing data free for research and non-commercial purpose help the research community. Weather data is difficult to capture with less errors and storing it is also a challenging task. [20] In this paper, many different secondary data sources have been discussed in great details and compared for weather data. The secondary data sources are compared on the basis of accessibility, accuracy, completeness, Reliability and relevance. These data sources are considered as top data sources among researchers and academicians and are very famous in public domain. [21] The National climate data center (NCDC) is considered as best secondary data source for collecting weather related data for rainfall prediction among the other secondary data sources discussed in this paper.

DECLARATION

Funding/ Grants/ Financial Support	No, I did not receive.
Conflicts of Interest/ Competing Interests	No conflicts of interest to the best of our knowledge.
Ethical Approval and Consent to Participate	No, the article does not require ethical approval and consent to participate with evidence.
Availability of Data and Material/ Data Access Statement	Not relevant.
Authors Contributions	All authors have equal participation in this article.

REFERENCES

1. 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. [CrossRef]
2. 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. [CrossRef]
3. 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. [CrossRef]
4. 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.
5. 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.
6. 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.
7. 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. [CrossRef]
8. 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. [CrossRef]
9. K C Gouda, Libujashree R,Priyanka Kumari,Manisha Sharma, Ambili D Nair, "An Approach for Rainfall Prediction using Soft Computing" International Journal of Engineering Trends and Technology (IJETT) – Volume 67 Issue 3 - March 2019 ISSN: 2231-5381. [CrossRef]
10. Moulana Mohammed, Roshitha Kolapalli, Niharika Golla, Siva Sai Maturi, "Prediction of Rainfall Using Machine Learning Techniques" International Journal of Scientific & Technology Research Volume 9, Issue 01, January 2020 ISSN: 2277-8616.
11. Deepali Patil, Shree L.R. Tiwari, Abhishek Jain, Shree L.R. Tiwari, Aniket Gupta, "Rainfall Prediction using Linear approach & Neural Networks and Crop Recommendation based on Decision Tree" International Journal of Engineering Research & Technology (IJERT) ISSN: 2278-0181 Vol. 9 Issue 04, April-2020. [CrossRef]
12. Sudha Mohankumar and Valarmathi Balasubramanian, "Identifying Effective Features and Classifiers for Short Term Rainfall Forecast Using Rough Sets Maximum Frequency Weighted Feature Reduction Technique", Journal of Computing and Information Technology, Vol. 24, No. 2, June 2016, 181–194 DOI: 10.20532/cit.2016.1002715 [CrossRef]
13. Balamurali Ananthanarayanan, Siva Balan, Anu Meera Balamurali and Karthika Balamurali, "Efficient Dissemination of Rainfall Forecasting to Safeguard Farmers from Crop Failure Using Optimized Neural Network Model" International Journal of Intelligent Engineering and Systems, Vol.10, No.1, 2017 DOI: 10.22266/ijies2017.0228.05. [CrossRef]
14. 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). [CrossRef]
15. 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.
16. N. Tyagi and A. Kumar, "Comparative analysis of backpropagation and RBF neural network on monthly rainfall prediction," Proc. Int. Conf. Inven. Comput. Technol. ICICT 2016, vol. 1, 2017.
17. N. Mishra, H. K. Soni, S. Sharma, and A. K. Upadhyay, "A Comprehensive Survey of Data Mining Techniques on Time Series Data for Rainfall Prediction," J. ICT Res. Appl., vol. 11, no. 2, p. 168, 2017. [CrossRef]
18. 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 (Online): 2278-3075, Volume-9 Issue-1, 2019, Page no. 258-263. [CrossRef]
19. Deepak Sharma, Dr. Priti Sharma. "Rainfall Prediction Using Classification and Clustering Complex Data Science Models with Geological Significance". International Journal of Computer Science Trends and Technology (IJCST) V8 (5): Page (39-44) Sep - Oct 2020. ISSN: 2347-8578. www.ijcstjournal.org.Published by Eighth Sense Research Group
20. Kalyankar MA, Alaspurkar SJ, "Data Mining Technique to Analyze the Meteorological Data", IJARCSSE, vol. 3(2), pp. 114–118, 2013.
21. Browning. K. A., "The mesoscale data base and its use in mesoscale forecasting", Quarterly Journal of the Royal Meteorological Society, vol. 115:488, pp. 717-762, 1989. [CrossRef]

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 currently pursuing a Ph.D. in Computer Science at M. D. University, Rohtak. 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 15 years. Her area of research includes Data Mining, Big Data, Software Engineering, Machine Learning.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of the Lattice Science Publication (LSP)/ journal and/ or the editor(s). The Lattice Science Publication (LSP)/ journal and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.