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

Pre-Processing and Normalization of the Historical Weather Data Collected from Secondary Data Source for Rainfall Prediction

Deepak Sharma, Priti Sharma

Abstract: ***In the twenty first century, data analysis has become the talk of the town. Almost every company or organization depends on data analysis for taking future decision. The most important step in data analysis after data collection is the preprocessing of the collected data. The main aim of data analysis is to find meaningful pattern by processing large amount of data. In data preprocessing, the inconsistency of collected data has been removed. After storing data for a relatively longer period, it becomes noisy and inconsistent. While measuring various parameter due to error in the instrument or human error, the value become incorrect or invalid. It is necessary to remove the invalid data otherwise it will deflect the results and produce error in the prediction. In this work preprocessing of the weather data has been analyzed for rainfall prediction using data mining.***

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

I. INTRODUCTION

The center piece of the art of data mining is data itself and is largely responsible for the fortune of the process of discovering new knowledge. [1] The initial stages of data mining also called as knowledge discovery process includes preprocessing of data. [2],[3] This is considered as the second most important step after data collection. It includes cleaning the raw data either collected from a primary data source or secondary data source. Need of preprocessing arises because of the presence of noises, outliers, and missing values in the raw data. [4] Data contains ambiguity and inconsistency is not recommend to use directly for the process of data mining. Noisy and inconsistent data leads to less accurate predictions with relatively less precision. Therefore, preprocessing of data is an indispensable part of the whole processes. [5] In this paper, preprocessing of weather data of having 9500 instances approximately collected from secondary data source i.e., National climate data center has been showcased. [6],[7] The weather data contains daily values of 11 atmospheric weather attributes for the district Hissar of Haryana. The various steps taken for the analysis and preprocessing of the collected data for rainfall prediction has been discussed in the further sections of this paper. [8]

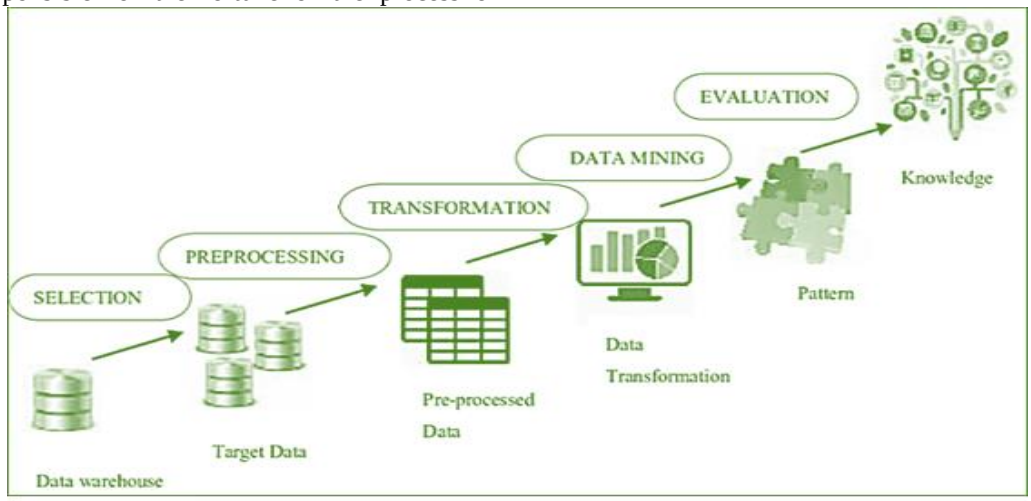

Figure 1: Process of Knowledge discovery (KDD)

Manuscript received on 26 May 2023 | Revised Manuscript received on 13 June 2023 | Manuscript Accepted on 15 November 2023 | Manuscript published on 30 November 2023.
*Correspondence Author(s)
Deepak Sharma*, Research Scholar, Department of Computer Science and Applications, Maharshi Dayanand University, Rohtak (Haryana), India. E-mail: erdeepaksharmabwn@gmail.com, ORCID ID: 0000-0002-7490-557X
Dr. Priti Sharma, Assistant professor, Department of Computer Science and Applications, Maharshi Dayanand University, Rohtak (Haryana), India. E-mail: pritish80@yahoo.co.in

© The Authors. Published by Lattice Science Publication (LSP). This is an open access article under the CC-BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/)

II. PREPROCESSING AND NORMALIZATION OF DATA

Before preprocessing, the detailed description of data set collected from the secondary data source has been given in the table 1 and table 2. [9] The original data set contains some noisy data with irrelevant symbols and missing values. Data set of January 1988 of Hisar before pre-processing is shown in the figure 2 and 3. Rows containing missing values are removed from the data set. [10], [11]

© Copyright: All rights reserved.

Table 1: Description of Data set before preprocessing

Attributes	Description	Instances
9	(1988-2022) 35 Years	9500

Table 2: Description of attributes present in the Data set before preprocessing

S. No.	Attribute Name	Data Type
1	Station Code	Numeric
2	Date	Date
3	Temperature (TEMP)	Numeric
4	Dew point (DEWP)	Numeric
5	Sea level pressure (SLP)	Numeric
6	Visibility (VISIB)	Real
7	Wind speed (WDSP)	Numeric
8	Maximum sustained wind speed (MXSPD)	Numeric
9	Maximum Temperature (MAXT)	Numeric
10	Minimum Temperature (MINT)	Numeric
11	Precipitation Amount (PRCP)	Real

Table 3: Description of missing values present in the Data set for each attribute

S. No.	Attribute name	Missing values
1	TEMP (Temperature)	9999.9
2	DEWP (Dew point)	9999.9
3	SLP (Sea Level pressure)	9999.9
4	VISIB (Visibility)	999.9
5	WDSP (Wind Speed)	999.9
6	MXSPD (Maximum Sustained wind speed)	999
7	PRCP (Precipitation)	99.99

STATION	DATE	TEMP	DEWP	SLP	VISIB	WDSP	MXSPD	MAX	MIN	PRCP
42131099999	01-01-1988	60.7	46.4	1018.8	1.4	0.5	1.9	76.6	46.4	0
42131099999	02-01-1988	61.9	49.8	1019.8	1.7	0.7	1.9	75.6	47.7	0
42131099999	03-01-1988	60.6	49.2	1018.8	1.4	0.6	1.9	74.8	48.2	0
42131099999	04-01-1988	61.7	49.6	1016	1.6	2.8	4.1	75.6	50.4	0
42131099999	05-01-1988	60.6	54.1	1016.5	2	2.1	5.1	70.9	50.9	0
42131099999	06-01-1988	62.2	56.4	1017.3	0.6	1.2	1.9	74.3	52.2	0
42131099999	07-01-1988	59.6	50.1	1017.4	1.2	1.4	4.1	9999.9	45.9	0
42131099999	08-01-1988	59.5	48.1	1018.5	1.9	1.2	5.1	73.9	45.3	0
42131099999	09-01-1988	60.7	48.9	1018.1	1.9	0.6	1.9	76.3	46.4	0
42131099999	10-01-1988	58.3	49.6	1018.6	1.3	0.2	1	72.7	47.1	0
42131099999	11-01-1988	60.1	53.5	1018.8	1.6	1.5	4.1	73.4	48.2	0
42131099999	14-01-1988	58	52.7	1016	1.5	0.5	1.9	69.1	46.8	0.02
42131099999	15-01-1988	60	50.9	1017.6	1.3	1.5	4.1	70.9	48.2	0
42131099999	16-01-1988	9999.9	47.3	1019.8	1.7	2.2	5.1	71.6	45.5	0

Figure 2: Data set of January 1988 of Hisar before pre-processing

STATION	DATE	TEMP	DEWP	SLP	VISIB	WDSP	MXSPD	MAX	MIN	PRCP
42131099999	17-01-1988	59.2	45.8	1018.6	999.9	1.4	2.9	74.5	44.4	0
42131099999	18-01-1988	56.5	45	1017.7	1.9	0.7	4.1	76.1	44.1	0
42131099999	19-01-1988	62.1	47.1	1015	2	2.4	5.1	78.3	43.3	0
42131099999	20-01-1988	59.9	48.4	1014.5	2.2	1.4	1.9	79.3	42.6	0
42131099999	21-01-1988	65.9	52.6	1010.3	2.1	3.2	6	78.3	49.1	0.12
42131099999	22-01-1988	60	52.7	1015	1.9	2.6	6	72.3	49.5	0
42131099999	23-01-1988	56.3	47.4	1015.2	2	1.9	999	69.3	44.2	0
42131099999	24-01-1988	54.6	44.3	1014.4	1.8	1.3	4.1	68.9	40.6	G
42131099999	25-01-1988	54.1	43.3	1016	2.1	1.8	2.9	69.8	40.8	0
42131099999	26-01-1988	57	41.4	1017.6	2	1.8	5.1	72	40.8	0
42131099999	27-01-1988	55.7	41.7	1017.2	2.1	1.7	4.1	74.8	39.6	0
42131099999	28-01-1988	60.4	9999.9	1017.3	2.2	2.2	6	73	50	0.02
42131099999	29-01-1988	60.8	45.1	1016.3	1.9	2.8	6	73.2	46.2	0.03
42131099999	30-01-1988	64.6	45.9	1016.2	2.2	1.5	1.9	77.2	46.8	0

Figure 3: Data set of 1988 of Hisar before pre-processing

© Copyright: All rights reserved.

III. DATA AFTER PREPROCESSING

Before preprocessing the total number of instances in the data set are 9500 and after removal of rows having missing values and outliers the number of instances remains 8695. [12] Also, there are total 11 attributes in the original data set but two attribute station code and date are found as redundant attributes and hance removed from the data set which makes the total number of attributes as 9. [13] This remaining data set is error free and consistent. Using a cleaned and error free data set is supremely necessary. Some rows of cleaned data set after preprocessing have been shown in the figure 4. [14]

Table 4: Description of Data set after preprocessing

Attributes	Description	Instances
9	(1988-2022) 35 Years	**8695**

Table 5: Description of attributes present in the Data set after preprocessing

S. No.	Attribute Name	Data Type
1	Temperature (TEMP)	Numeric
2	Dew point (DEWP)	Numeric
3	Sea level pressure (SLP)	Numeric
4	Visibility (VISIB)	Real
5	Wind speed (WDSP)	Numeric
6	Maximum sustained wind speed (MXSPD)	Numeric
7	Maximum Temperature (MAXT)	Numeric
8	Minimum Temperature (MINT)	Numeric
9	Precipitation Amount (PRCP)	Real

STATION	DATE	TEMP	DEWP	SLP	VISIB	WDSP	MXSPD	MAX	MIN	PRCP
42131099999	01-01-1988	60.7	46.4	1018.8	1.4	0.5	1.9	76.6	46.4	0
42131099999	02-01-1988	61.9	49.8	1019.8	1.7	0.7	1.9	75.6	47.7	0
42131099999	03-01-1988	60.6	49.2	1018.8	1.4	0.6	1.9	74.8	48.2	0
42131099999	04-01-1988	61.7	49.6	1016	1.6	2.8	4.1	75.6	50.4	0
42131099999	05-01-1988	60.6	54.1	1016.5	2	2.1	5.1	70.9	50.9	0
42131099999	06-01-1988	62.2	56.4	1017.3	0.6	1.2	1.9	74.3	52.2	0
42131099999	08-01-1988	59.5	48.1	1018.5	1.9	1.2	5.1	73.9	45.3	0
42131099999	09-01-1988	60.7	48.9	1018.1	1.9	0.6	1.9	76.3	46.4	0
42131099999	10-01-1988	58.3	49.6	1018.6	1.3	0.2	1	72.7	47.1	0
42131099999	11-01-1988	60.1	53.5	1018.8	1.6	1.5	4.1	73.4	48.2	0
42131099999	14-01-1988	58	52.7	1016	1.5	0.5	1.9	69.1	46.8	0.02
42131099999	15-01-1988	60	50.9	1017.6	1.3	1.5	4.1	70.9	48.2	0
42131099999	18-01-1988	56.5	45	1017.7	1.9	0.7	4.1	76.1	44.1	0
42131099999	19-01-1988	62.1	47.1	1015	2	2.4	5.1	78.3	43.3	0
42131099999	20-01-1988	59.9	48.4	1014.5	2.2	1.4	1.9	79.3	42.6	0
42131099999	21-01-1988	65.9	52.6	1010.3	2.1	3.2	6	78.3	49.1	0.12
42131099999	22-01-1988	60	52.7	1015	1.9	2.6	6	72.3	49.5	0
42131099999	25-01-1988	54.1	43.3	1016	2.1	1.8	2.9	69.8	40.8	0
42131099999	26-01-1988	57	41.4	1017.6	2	1.8	5.1	72	40.8	0
42131099999	27-01-1988	55.7	41.7	1017.2	2.1	1.7	4.1	74.8	39.6	0
42131099999	29-01-1988	60.8	45.1	1016.3	1.9	2.8	6	73.2	46.2	0.03
42131099999	30-01-1988	64.6	45.9	1016.2	2.2	1.5	1.9	77.2	46.8	0

Figure 4: Data set of 1988 of Hisar after pre-processing

IV. RESULT AND DISCUSSION

Preprocessing of data set is indispensable part of the knowledge discovery process and hance should be done with full precaution. [15] Before preprocessing the total number of instances in the data set are 9500 and after removal of rows having missing values and outliers the number of instances remains 8695. Also, there are total 11 attributes in the original data set but two attribute station code and date are found as redundant attributes and hance removed from the data set which makes the total number of attributes as 9. [16],[17] In this paper, preprocessing of weather data of having 9500 instances approximately collected from secondary data source i.e., National climate data center has been showcased. [18] The weather data contains daily values of 11 atmospheric weather attributes for the district Hissar of Haryana. [19], [20] The various steps taken for the analysis and preprocessing of the collected data for rainfall prediction has been discussed in this paper.

V. CONCLUSION AND FUTURE SCOPE

Rainfall is an interrelated and diverse phenomenon which depends upon many different atmospheric attributes.

© Copyright: All rights reserved.

It is a very challenging task to predict rainfall. Many government and private organizations are working on predicting rainfall with accuracy by sharing various atmospheric attributes. In this work, supervised data mining techniques are used to predict rainfall by analyzing historical weather data. In data mining "data set" is the foundation stone and should be consistent and error free.

Analysis of data helps in order to find out the relationship of different attribute with rainfall. In this work, the historical weather has been collected from national climate data center (NCDC) repository. According to world weather watch program of the world meteorological organization (WMO), the national climate data center has been storing the daily weather data for more than 8000 stations. When data is stored in such large scale there is a possibility of outliers and noisy data. It is always recommended to do preprocessing of the data before using it. After preprocessing, the data becomes error free and give good accuracy in predicting the outcome.

DECLARATION STATEMENT

Funding/ Grants/ Financial Support	No, I did not receive.
Conflicts of Interest/ Competing Interests	The article bears 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.

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 12 years. Her area of research includes Data mining, Big data, Software Engineering, Machine Learning.

© Copyright: All rights reserved.

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.

© Copyright: All rights reserved.