Website : www.asthafoundation.in
Progressive Research : An International Journal
Vol 18 (2) : 145-149 (July-2023), Print ISSN : 0973-6417, Online ISSN : 2454-6003
Society for Scientific Development in Agriculture and Technology, Meerut (U.P.) India

Assessment of Rainfall by Observing Historical Trends Using Supervised Data Mining Technique

Deepak Sharma and Priti Sharma
DCSA, Maharshi Dayanand University, Rohtak, Haryana, India
Email : erdeepaksharmabwn@gmail.com

Abstract

Rainfall being a complete weather phenomenon, depends directly or indirectly upon many different weather aspects. It plays a key role in the process of agriculture. Also, in some places the cultivation of crop is completely dependent upon the rainfall during that period. Early estimation and prediction of rainfall helps in dealing with the uncertainties of nature. Data mining is a process of finding out meaningful patterns after analyzing large amount of data. In this work, thirty-six years of weather data has been acquired which is collected by the world meteorological department. The collected data contains a total of 11 weather attributes for which daily values has been collected and maintained. In this work, a lazy learning approach called as k-nearest neighbor is used to analyze the data in order to predict the rainfall. The prediction model proposed in this work has been validated using cross validation technique.

Key words : *Data mining, rainfall prediction, k- nearest neighbor, weather forecasting, cross validation, rapid miner, machine learning.*

Introduction

India's growth rate is heavily dependent on agriculture sector. A good harvest simply results in more money in the hands of government to provide its welfare schemes also a depression in agriculture sector reciprocates the above scenario. Indian economy is also described as Agro-economy. Agriculture sector is also identified and marked as the sector of livelihood for more than half of the Indian population and the count becomes even more in the rural areas.

To increase the efficacy and productivity in the field of agriculture, more scientific research should be encouraged in this area. This work is dedicated towards the prognosis and estimation of rainfall by observing historical trends in the weather data using data mining techniques.

Data mining is a constructive process that involves study of data in order to find the relationship between various attributes of data set and trends that are often repetitive in nature. K- nearest neighbor is a lazy learning approach that is used to analyze thirty-six years of historical weather data. This paper includes details regarding the rainfall prediction model and its implementation using rapid miner.

K-NN stands for k-nearest neighbor where value of 'K' varies as per requirement of model. It is a supervised learning algorithm in which data has been supplied to the prediction model and based on that data, predictions will take place. It is also categories as a non-parametric algorithm because in this machine learning approach no assumption has been made on the supplied data and used the training data at the time of actual predictions.

Some machine learning algorithm such as decision tree and artificial neural network learn from the supplied data but on the other hand K-NN does not learn from the supplied data and performs prediction and analysis at the time of actual prediction this approach is called lazy leaning and such algorithms are categories as lazy learning algorithms. K-NN can be used for both classification and regression problems.

Tuple A_1 : (a_{11}, a_{12}, a_{13}…………., a_{1n})

Tuple A_2 : (a_{21}, a_{22}, a_{23}…………., a_{2n})

$$\text{dis}\,(A_1, A_2) = \sqrt{\sum_{i=1}^{n} (a_{1i} - a_{2i})^2}$$

In order to understand the working of K-nearest neighbor, Let's consider two tuples A_1 and A_2. The difference between value of each corresponding attribute in tuple A_1 and A_2 is taken and square this difference. After the addition of all the squared difference of the corresponding attribute, square root is applied. The values of attributes are normalized wherever required before calculation of Euclidean distance.

Received : June-2023; Revised : June-2023; Accepted : July-2023

The unknown tuple is classified based on the inference made from its nearest made from its nearest neighbor. For example, if the value of k=1 then the unknown tuple has assigned the class of its closest neighbor in the pattern space.

Materials and Methods

Data collection and preprocessing : Data collection for any research work can be done in two ways. One is to collect the data specifically for that research work. This type of collected data is known as Primary data. Other is to gather data which is already been collected by some other entity. This type of collected data is known as secondary data.

Table-1 : Represents data utilisation for model.

S. No.	Data Set	Description	Instances
1.	Training Data Set	(1988-2016) 31 years	7000
2.	Testing Data Set	(2017-2022) 6 years	2000

Dedicated resources are required to collect primary data which are costly and it is a very time-consuming process. In this research work, thirty-seven year of historical weather data has been used which is collected under the "World weather watch program" by WMO i.e. World meteorological organization (WMO). The data used in the work has been collected by NCEI (National centre environment Information). The details of attribute present in this data have been given in table-2.

Table-2 : Description of attributes present in the data set.

S. No.	Attribute	Type
1.	Station Code (STN)	Integer
2.	Date (DATE)	Integer
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

Implementation : The collected data has been divided into two parts i.e., training data and testing data. As already discussed in the introduction section, K-nearest neighbor is a supervised machine learning and data from year 1988 to 2016 that contains approximately 7000 instances is used as training set and data from 2017 to 2022 that contains approximately 2000 instances is used as testing set. The proposed rainfall prediction model based on K-nearest neighbor has been implemented using Rapid Miner. The implemented version of the model is being represented in figure-1.

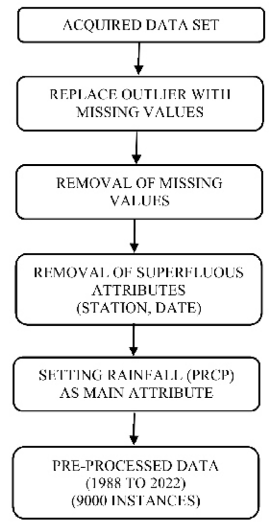


Figure-1 : Pre-processing of Acquired Data Set.

Results and Discussion

In this research work, a K-NN based rainfall prediction model has been proposed and implemented. The validation of the model has been done using cross validation technique shown in figure 4 and 5 respectively. The model has been tested for different values of 'K' and compared on the basis of Accuracy, Precision, recall, Root mean square error (RMSE). The results are compiled and shown in table-3.

Conclusion and Future Scope

K-Nearest neighbor data mining technique can be used for both classification and regression problem. In this work a rainfall prediction model has been proposed to predict rainfall using historical weather data having 9000 instances for 11 weather attributes.

The supervised machine learning technique used to design the model is k nearest neighbor. The data from the year 1988 to 2016 is used to train the proposed model having 7000 instances and data from year 2017 to 2022 is used for testing purpose having 2000 instances.

The validation of the proposed models has been carried out using cross validation method. The model has been tested out and analyzed for different value of 'K'. The detail comparative analysis has been discussed in the result section of this paper.

The K-NN based model performs best in case of K=8 with accuracy of 88.08 %, Precision of 64.78%, Recall of 78.94% and RMSE i.e. Root mean square error of 0.303.

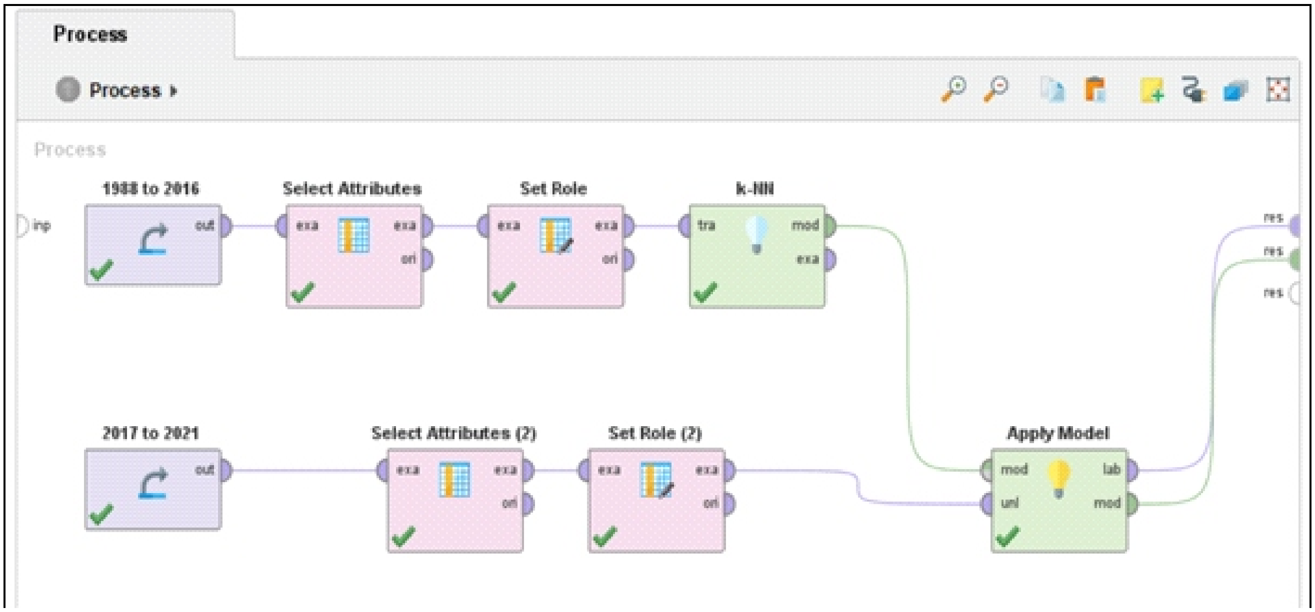
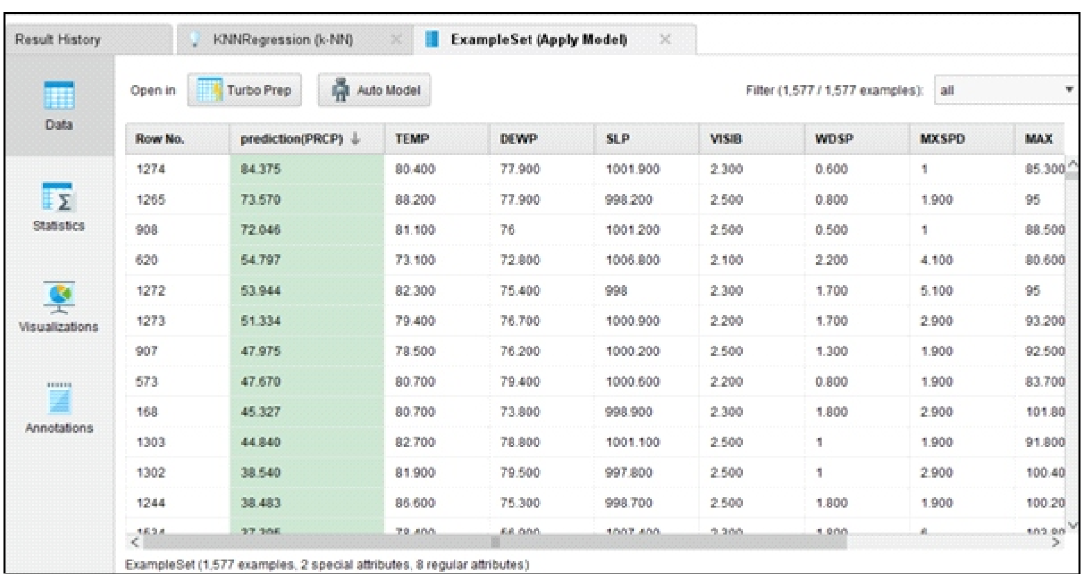


Figure-2 : Rainfall Prediction model using K-Nearest Neighbor (RPM_KNN).

Figure-3 : Results generated using rapid miner (RPM_KNN).



Row No.	prediction(PRCP)	TEMP	DEWP	SLP	VISIB	WDSP	MXSPD	MAX
1274	84.375	80.400	77.900	1001.900	2.300	0.600	1	85.300
1265	73.570	88.200	77.900	998.200	2.500	0.800	1.900	95
908	72.046	81.100	76	1001.200	2.500	0.500	1	88.500
620	54.797	73.100	72.800	1006.800	2.100	2.200	4.100	80.600
1272	53.944	82.300	75.400	998	2.300	1.700	5.100	95
1273	51.334	79.400	76.700	1000.900	2.200	1.700	2.900	93.200
907	47.975	78.500	76.200	1000.200	2.500	1.300	1.900	92.500
573	47.670	80.700	79.400	1000.600	2.200	0.800	1.900	83.700
168	45.327	80.700	73.800	998.900	2.300	1.800	2.900	101.80
1303	44.840	82.700	78.800	1001.100	2.500	1	1.900	91.800
1302	38.540	81.900	79.500	997.800	2.500	1	2.900	100.40
1244	38.483	86.600	75.300	998.700	2.500	1.800	1.900	100.20

Figure-4 : Predicted value of PRCP (Rainfall) in Numerical form.

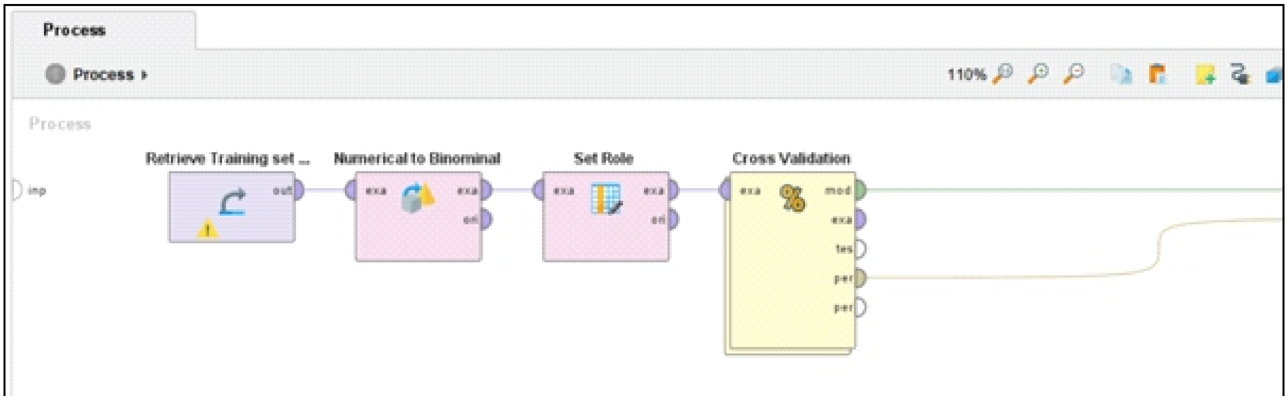


Figure-4 : Cross validation on RPM_KNN.

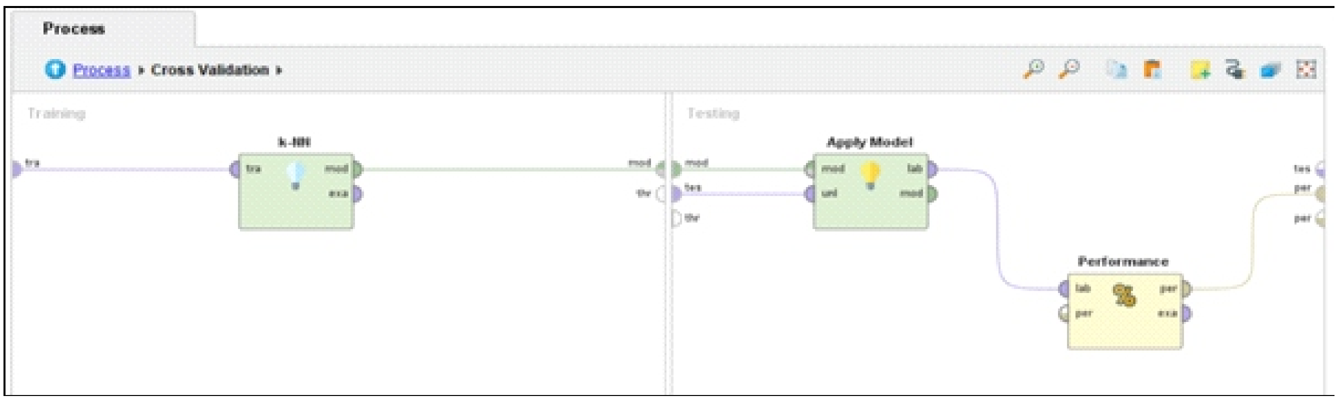


Figure-5 : Training and Testing module of Cross validation on RPM_KNN.

Table-3 : Results and analysis.

Model	Accuracy	Precision	Recall	RMSE
K-NN (K=2)	86.51	60.18	72.98	0.350
K-NN (K=4)	87.20	62.44	75.61	0.318
K-NN (K=6)	87.80	64.16	77.83	0.308
K-NN (K=8)	88.08	64.78	78.94	0.303
K-NN (K=10)	87.84	64.60	77.89	0.300

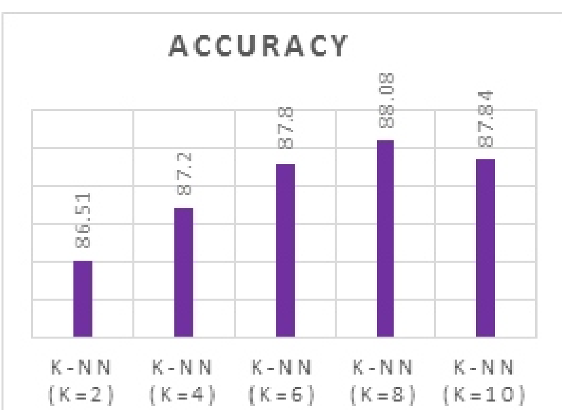


Figure-6 : Comparison : Accuracy.

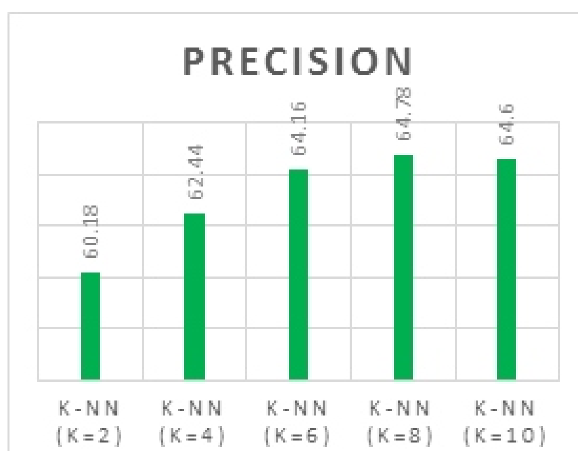


Figure-7 : Comparison: Precision.

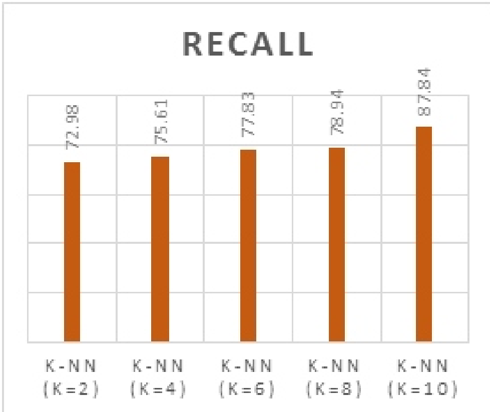


Figure-8 : Comparison: Recall.

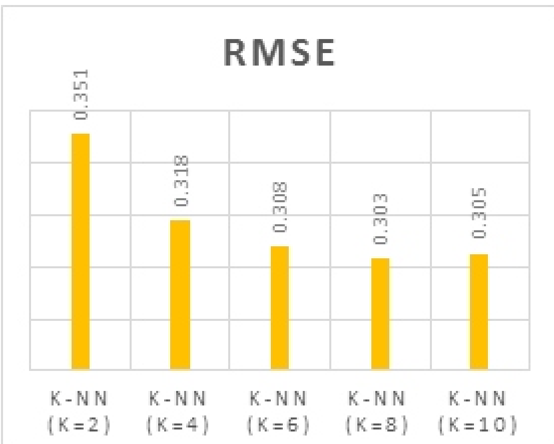


Figure-9 : Comparison : RMSE.

References

1. Tharun V.P, Ramya Prakash S. Renuga Devi (2018). "Prediction of Rainfall Using Data Mining Techniques", *2nd International Conference on Inventive Communication and Computational Technologies, IEEE*-2018.
2. Abishek B., R. Priyatharshini, Akash Eswar M., P. Deepika (2017). "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.
3. Fahad Sheikh S. Karthick D. Malathi J.S. Sudarsan C. Arun (2016). "Analysis of Data Mining Techniques for Weather Prediction", *Indian Journal of Science and Technology,* 9(38): ISSN: 0974-6846.
4. Nikhil Sethi, Kanwal Garg (2014). Exploiting Data Mining Technique for Rainfall Prediction, (IJCSIT) *International Journal of Computer Science and Information Technologies,* 5(3): 3982-3984.
5. Sapara G.K., Parmar R.S., Barad H.R. and Patel J.B. (2022). Combining ability studies in F_2 generation of sesame (*Sesamum indicum* L.) over environments. *Frontiers in crop improvement,* 10(2): 134-140.
6. B. Renuka Devi, K. Nageswara Rao, S. Pallam Setty, M. Nagabhushana Rao (2013). Disaster Prediction System Using IBM SPSS Data Mining Tool", *International Journal of Engineering Trends and Technology*, 4(8): 126-130.
7. Gao S. and L.S. Chiu (2012). Development of statistical typhoon intensity prediction: Application to satellite-observed surface evaporation and rain rate, *Weather Forecasting*, 27: 240–250.
8. Deepak Sharma, Priti Sharma (2019). Rain Fall Prediction using Data Mining Techniques with Modernistic Schemes and Well-formed Ideas, *International Journal of Innovative Technology and Exploring Engineering,* 9(1): 258-263.