IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol. 10, Issue.2, February 2022 E-ISSN: 2347-2693

Data Mining Techniques for Rainfall Data Using WEKA

K. Anil Kumar[1], S. Venkatramana Reddy[2] , B. Sarojamma[3*]

[1]Assistant Professor, Department of Mathematics, School of Science, GITAM University, Hyderabad -502329,TS, India
[2]Associate Professor, Department of Physics, S.V.University, Tirupati – 517 502, AP, India
[3]Associate Professor, Department of Statistics, S.V.University, Tirupati – 517 502, AP, India

**Correspondence Author: saroja14397@gmail.com*

DOI: https://doi.org/10.26438/ijcse/v10i2.4548 | Available online at: www.ijcseonline.org

Received: 10/Feb/2022, Accepted: 20/Feb/2022, Published: 28/Feb/2022

Abstract**--**There are two types of monsoons or rainfall seasons in India: summer rainfall from October to March and winter rainfall from April to September. Rainfall plays a vital role in the cultivation, cropping, drinking and other purpose of human beings. Generally, in India, most of times the water source is from rain. In this paper, we are fitted isotonic regression model, linear regression, additive regression, Rep tree and simple linear regression by using machine learning models and are estimated using WEKA software for rainfall as dependent variable and time as an independent variable. The best model for the data is chosen using various accuracy measures like Absolute Mean Error, Root Mean Squared Error, Relative absolute error and Root Relative squared error.

Key words**-** Rainfall, Isotonic Regression, Rep tree, RMSE.

I. INTRODUCTION

Rainfall plays a vital role in agriculture,humankind, cultivating paddy, cattle feed etc., and needs plenty of water. In India, several people are depended upon rainfall or monsoons; they are summer rainfall and winter rainfall. Generally, summer rainfall is from April to September and winter rainfall is between October and April. Massive rain plays a prominent role in irrigation, drinking, and power generation.

P.E.NailHomani [1] in the paper entitled “Time series analysis models for Rainfall data in Jordan”. The author used Box and Jenkins model of Autoregressive Integrated Moving Average for forecasting the monthly rainfall data. MostataDastorani et al [2] discussed in their article “Comparative Study among different time series models applied to monthly rainfall forecasting in semi-arid climate condition”. In this paper they used Auto Regressive Integrated Moving Average and Seasonal Auto regressive Integrated Moving Average with different structures of trial and error and it was examined for North khorasan province from 1989 to 2012 using R software. Nasimul Hasan et al [3] published a paper “A support vector regression model for forecasting Rainfall”. In this paper they used support vector regression model and forecasted 7 days ahead results also. Kin C. Luk et al [4] in their article “An Application of Artificial Neural networks for rainfall forecasting”, used artificial neural networks like multilayer feed forward neural networks, partial recurrent neural networks and time delay neural networks. N.Q. Hung et al [5] explain an artificial neural network model for rainfall forecasting in Bangkok, Thailand. In this paper they used ANN for estimation of rainfall using meteorological parameters like relative humidity, air pressure, wet bulb temperature and cloudiness. M.P. Darji et al [6] explains rainfall forecasting using Artificial Neural Networks. In this paper they analyze crop productivity and use of water resources and different accuracy measures are used to test performance of ANN. Hari Mallikarjuna Reddy et al published a paper entitled “Data Mining Techniques for estimation of wind speed using WEKA”[7]. Damodaran et al. discussed various Quantile Regression Models for Rainfall Data [8].

II. METHODOLOGY

For fitting of data mining techniques of annual rainfall data from 2005 to 2017[9], we are using Waikato Environment for Knowledge Analysis (WEKA) software. For this data, time (years) is taken as independent and annual rainfall values as a dependent variable. Here we performed isotonic regression, Linear Regression Model, Additive Regression, Rep Tree, and simple linear regression.

Isotonic Regression: It is similar to linear regression. However, it is changed up and down with data. It is not in linear form. Generally, isotonic regression minimizes the mean square error. It is a technique of fitting free-form lines in a sequence of observations thatholds the line in a non-decreasing or non-increasing manner. But it lies close to the actual observations.

Linear regression model: If the data contains two variables: one is an independent and another is the dependent variable. The linear regression equation is $y = a + b x$
Where y is dependent variable
x is independent variable
a and b are the coefficients.

Additive Regression: It explains about unknown relationship of continuous output and a dimensional vector of inputs. Generally, regression predicts the outputs. In the additive regression model, unrestricted nonparametric multiple regression andthe conditional average value of y as a general smooth function of different x's. The general modeling is given below.

$E(Y/X_1, X_2, X_3 \ldots\ldots X_n) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k$

Rep Tree: The rep tree is a regression tree and is the fast decision tree. It deals with corresponding instances and automatically deletes the missing observations, and sorts the values for numeric attributes.

Simple linear regression: It is a model having one dependent and an independent variable. Here, the dependent variable can carry only continuous or real values, whereas an independent variable can carry either continuous or categorical values. The simple linear regression model is as follows.

$y = \beta_0 + \beta_1 x + \epsilon$

Where y is a dependent variable
x is the independent variable
$\beta_0, \beta_1$are the coefficients
$\epsilon$is the error term.

Various measures of error used to test the models are Mean absolute error, Root mean square error, Relative absolute error, and Root relative squared error.

Mean absolute error:
It is the average of sum of absolute deviations between predicted and actual values. The formula of MAE (Mean Absolute Error) = $\frac{\sum_{j=1}^{n}|a_j - b_j|}{n}$

$$Where\ a_j\ and\ b_j\ are\ the\ predicted\ and\ atual\ values\ for\ j^{th}\ pair\ of\ observation.$$

Root Mean Square Error:
It is calculated by using the following formula

Root Mean Squared Error (RMSE) = $\sqrt{\frac{\sum_{j=1}^{n}(a_j - b_j)^2}{n}}$

$$Where\ a_j\ and\ b_j\ are\ the\ predicted\ and\ atual\ values\ for\ j^{th}\ pair\ of\ observation.$$

Relative absolute error: It is used in machine learning and data mining techniques to check the accuracy of the model and is given by

$RAE = \frac{\sum|\hat{y}_i - y_i|}{\sum|y_i - \bar{y}_i|};\ here\ y_i\ is\ the\ actual\ and\ \hat{y}_i\ is\ predicted\ values.$

Root relative squared error:
It is one of the accuracy measure metrics and is calculated using the relative squared error takes the total squared error and normalizes it by dividing the total squared error of the simple predictor.

III. EMPIRICAL INVESTIGATIONS

We have fitted five data mining algorithms such as Isotonic regression, Linear regression, Additive regression, Rep Tree, and simple linear regression, using ten-fold WEKA software for annual rainfall data of India from 2005 to 2017. The actual values, predicted values, and error values of isotonic regression model are presented in the following table-1.

Table-1: The actual, predicted, and error values by using the isotonic regression model.

Year	Actual values	Predicted values	Error values
2006	11706	10574	-1132
2007	11091	11555.5	464.5
2008	9273	11477	2204
2009	11001	11477	476
2010	10162	9685	-477
2011	11602	11001	-601
2012	12291	9844	-2447
2013	11509	11346.5	-162.5
2014	9780	9876	96
2015	9943	11012	1069
2016	10789	10803.25	14.25
2017	9590	9971	381

Summary of the model like correlation coefficient, mean absolute error, Root mean square error, Relative absolute error and Root Relative squared error are given in Table-2.

Table-2: Model Summary for isotonic regression.

Model Summary	Isotonic model
Correlation coefficient	0.0857
Mean absolute error	793.6875
Root mean squared error	1098.5469
Relative absolute error	89.316 %
Root relative squared error	108.7063 %

Linear regression model using WEKA by taking rainfall as dependent and year wise time as an independent variable and fitted equation is Rainfall (Y) = 343707.7496-165.6339 * years.

The actual, predicted, and error values are shownin the table-3.

Table-3: The actual, predicted and error values by using the linear regression model.

Year	Actual values	Predicted values	Error values
2006	11706	10757.335	-948.665
2007	11091	11692.155	601.155
2008	9273	10965.786	1692.786
2009	11001	11133.792	132.792
2010	10162	9603.625	-558.375
2011	11602	11206.976	-395.024
2012	12291	10286.652	-2004.348
2013	11509	11425.78	-83.22
2014	9780	10016.188	236.188
2015	9943	10521.108	578.108
2016	10789	10598.343	-190.657
2017	9590	10250.822	660.822

Various model accuracy measures are produced to test the errors. Mean absolute error, Root mean squared error, Relative absolute error and Root relative squared error aregiven in the table-4.

Table-4: model accuracy metrics for the linear regression model.

Model Summary	**Linear regression**
Correlation coefficient	673.5118
Mean absolute error	889.7428
Root mean squared error	1098.5469
Relative absolute error	75.7923 %
Root relative squared error	88.0442 %

Actual values, estimated values using additive regression, Rep tree, and simple linear regression models and their corresponding error observations are tabulated in the table-5. These are calculated by taking rainfall as dependent and year wise time as an independent variable in WEKA software.

Table-5: Actual, estimated and error values of additive regression, Rep tree and simple linear regression models.

Year	**Actual values**	**Estimated values**			**Error values**		
		Additive regression model	***Rep tree model***	***Simple linear regression model***	***Additive regression model***	***Rep tree model***	***Simple linear regression model***
2006	11706	9858.074	917.544	10724.01	-1847.926	-788.456	-981.99
2007	11091	1507.434	917.544	314.965	416.434	-173.456	223.965
2008	9273	11580.29	805.051	1105.548	2307.29	1532.051	1832.548
2009	11001	11580.29	805.051	1241.994	579.29	-195.949	240.994
2010	10162	9881.398	783.889	10083.327	-280.602	621.889	-78.673
2011	11602	1124.825	744.657	1026.334	-477.175	-857.343	-575.666
2012	12291	9874.887	530.776	0176.995	-2416.113	-1760.224	-2114.005
2013	11509	1166.639	710.201	1157.582	-342.361	-798.799	-351.418
2014	9780	0072.398	771.423	347.101	292.398	991.423	567.101
2015	9943	1154.151	713.828	0626.256	1211.151	770.828	683.256
2016	10789	0895.981	428.639	0660.659	106.981	-360.361	-128.341
2017	9590	1704.336	848.198	0483.555	2114.336	1258.198	893.555

Various model accuracy measures like mean absolute error, Root mean squared error, Relative absolute error and Root relative squared error for additive regression, Rep tree and simple linear regression are presented in table-6.

Table-6: Model Summary for Additive, Rep tree and simple linear regression models.

Measures of accuracy	**Additive regression**	**Rep tree**	**Simple linear regression**
Mean Absolute Error	1032.6714	842.4147	722.6261
Root Mean Squared Error	1339.8664	965.4702	955.8689
Relative Absolute Error	116.2095 %	94.7994 %	81.3192 %
Root Relative Squared Error	132.586 %	95.5377 %	94.5876 %

IV. SUMMARY AND CONCLUSIONS

By considering rainfall is the dependent variable and year wise time is an independent variable, we perform five different data mining models: isotonic regression, linear regression, additive regression, Rep tree, and simple linear regression using WEKA software. The model accuracy was checked for the five models using measure of accuracy like Mean absolute error, Root mean squared error, Relative absolute error and Root relative squared error and are given in the following table.

Model	**MAE**	**RMSE**	**RAE**	**RRSE**
Isotonic Regression	793.6875	1098.5469	0.8932	1.0871
Linear regression	673.5118	889.7428	0.7579	0.8804
Additive Regression	1032.6714	1339.8664	1.1621	1.3259
Rep tree Regression	842.4147	965.4702	0.9480	0.9554
Simple linear Regression	722.6261	955.8689	0.8132	0.9459

From the observations of the above table we conclude that the best model among five models is thelinear regression model for rainfall data since the RMSE is minimum when compared to others.

REFERENCES

[1] P.E.NailHomani, “Time series Analysis model for Rainfall data in Jordan case study for using Time series Analysis”, American Journal of environmental sciences vol. 5 no5, pp. 599-604, 2009.

[2] MostataDastorani, Mohammad Mirzawad, Mohammad TaghiDastorani, and Syed Javad Sadatinejad,“Comparative study among different time series models applied to monthly rainfall forecasting in semi-arid climate condition”, *Natural Hazards*, vol. 81, pp.1811–1827, 2016.

[3] N. Hasan, N. C. Nath and R. I. Rasel, “A support vector regression model for forecasting rainfall”, 2015 2nd International Conference on Electrical Information and Communication Technologies (EICT), Khulna, Bangladesh, pp. 554-559, 2015. doi: 10.1109/EICT.2015.7392014.

[4] Kin C. Luk, J.E. Ball and A. Sharma,“An application of Artificial Neural Networks for rainfall forecasting”, mathematical and computer modelling, vol 33, pp. 683-693, 2001.

[5] N.Q.Hung, M.S. Babel, S. Weejakul and N.K. Tripathi,“An Artificial Neural Network model for rainfall forecasting in

Bangkok", Thailand, Hydrology and earth sciences, vol. 13, pp.1413-1425, 2009.

[6] M. P. Darji, V. K. Dabhi and H. B. Prajapati, "Rainfall forecasting using neural network: A survey, "Proceedings of the International Conference on Advances in Computer Engineering and Applications, Ghaziabad, India, pp.706-713, 2015, doi: 10.1109/ICACEA.2015.7164782.

[7] B. Hari Mallikarjuna Reddy, S. Venkatramana Reddy, and B. Sarojamma, "Data Mining Techniques for estimation of wind speed using WEKA", International Journal of Computer Sciences and Engineering(IJCSE), 9(9), 49-53. 2021.DOI: https://doi.org/10.26438/ijcse/v9i9.4851

[8] S. Damodharan, S. Venkatramana Reddy, B.Sarojamma, "Quantile Regression Models for Rainfall Data", International Journal of Computer Sciences and Engineering(IJCSE), 9(9), 83-85, 2021.DOI: https://doi.org/10.26438/ijcse/v9i9.8385

[9] Data website: URL:http://www.imd.gov.in/Welcome%20To%20IMD/Welcome.php.

AUTHOR PROFILE

Dr. K.Anil Kumar has done his Ph.D.in Statistics from Sri Venkateswara University, Tirupati, in the year 2009 and currently working as an Assistant Professor in the Department of Mathematics, GITAM (Deemed to be University), Hyderabad. He has ten years of Teaching and 13 years of research experience. He has published and presented papers in various national and internationally reputed Journals and National and International Conferences. Presently He is the coordinator for M.Sc. Data Science, Department of Mathematics, GITAM (Deemed to be University), Hyderabad, India.

Dr. S. Venkatramana Reddy is working as an Associate Professor in the Department of Physics, S.V. University, Tirupati. He has 21years' teaching experience and 27 years research experience. He has published 115 research papers in various internationally reputed Journals and presented more than 125 papers in various National and International Conferences. Received the Best Paper Award for presentation of the Research Paper entitled "Reducing the effect of ground clutter from wind profiler radar signal using wavelet transforms", in the National Conference. 8 Ph.D. and 4 M.Phil. degrees are awarded under his Supervision. He was Co-ordinator for 5^{yr} Integrated M.Sc. Course in Physics and presently Coordinator for M.Sc. Electronics, S.V. University, Tirupati.

Dr. B. Sarojamma is working as an Associate Professor in Dept. of Statistics, Sri Venkateswara University, Tirupati, A.P, India. She has 14 years of Teaching and 19 years Research Experience. She Published 92 Research papers and 2 Books. She Presented 134 Research Papers in Various National and International Conferences, Seminars and Symposiums. She attended 57 Workshops. She organized 4 National Seminars and workshops. 6 Ph.Ds and 2 M.Phil Degrees are awarded under her guidance.

