Annual Rainfall Prediction of Various States in India using Linear Regression

Deepa.G, Dinesh.B, Naveen Kumar.K

Abstract: ***Rainfall prediction is a significant part in agriculture, so prediction of rainfall is essential for the best financial development of our nation. In this paper, we represent the linear regression method to predict the yearly rainfall in different states of India. To predict the estimate of yearly rainfall, the linear regression is implemented on the data set and the coefficients are used to predict the yearly rainfall based on the corresponding parameter values. Finally an estimate value of what the rainfall might be at a given values and places can be establish easily. In this paper, we demonstrate how to predict the yearly rainfall in all the states from the year 1901 to 2015 by using simple multi linear regression concepts. Then we train the model using train _test_ split and analyze various performance measures like Mean squared error, Root mean squared error, R^2 and we visualize the data using scatter plots, box plots, expected and predicted values.***

Keywords: Rainfall Prediction, Linear Regression, Learning Process, Machine Learning.

I. INTRODUCTION

Rainfall is a climatic factor that affects many human activities like agricultural production, power generation, construction, forestry and tourism, between others. At this range, rainfall prediction is essential as this variable is the one with the maximum correlation with opposing natural events such as flooding, landslides and avalanches. These incidents have affected the world for many years. To solve this uncertainty, we used various machine learning techniques and models such as linear regression, multiple linear regressions, mean absolute error and root mean squared error to make exact and suitable predictions. It plays a fundamental role in preventing causalities caused due to natural disasters. It additionally encourages us to keep up our water resources properly. Exact estimation of rainfall is helpful in case of heavy rainfall which may cause flood and no rainfall which may cause drought to retain our water resources. In our nation, yields are collected on the monsoon season, so having more information in predication is significant. A few states are caused by drought while some with floods. To solve this issue, there are different types of rainfall prediction methods are proposed.

II. LITERATURE SURVEY

In [1] authors discussed RNN(Recurrent Neural Network), TDNN(Time Delay Recurrent Neural Network) features and composite models. Multilayer Feed Forward Neural Network (MLFNN) is used for predicting Indian summer monsoon rainfall in [2]. Error Back Propagation (EBP) method is used for predicting rainfall in [3]. These network models are analyzed with two, three and ten input parameters. Also they used statistical models to compare their output result. In [4] Adaptive Basis Function Neural Network (ABF NN) is used for annual rainfall prediction in kerala region. In [5] authors used Artificial Neural Network (ANN) for prediction of rainfall in Hyderabad. In [6] authors proposed Conjugate Gradient Decent (CGD) and Levenberg–Marquardt (LM) learning algorithm for training. In [7] authors implemented Modular Artificial Neural Network (MANN) to predict rainfall in India and China. In [8] authors used monthly, quarterly, half-yearly, yearly rainfall data for rainfall forecasting with the method Focused Time Delay Neural Networks (FTDNN). In [9] authors predicted short term rainfall. In [10] authors discussed monthly rainfall of Chennai region. In this paper we discussed simple multi linear regression concepts to predict the yearly rainfall in all the states from the year 1901 to 2015.

III. PROPOSED SYSTEM

a. Problem statement:

Our proposed "Annual Rainfall Prediction" represents a numerical method called linear regression to forecast the rainfall in different states of India. The linear regression technique is improved so as to acquire the most ideal error level by emphasizing and adding some level of error to the input values. This strategy gives a prediction of yearly rainfall using distinct environmental parameters like normal temperature and cloud cover to forecast the rainfall. Linear regression is used on the data information and the coefficients are utilized to forecast the rainfall dependent on the corresponding values of the parameter. In this way, yearly rainfall estimated values is predicted exactly.

b. Objective:

The objectives of the study is to implement rainfall prediction with machine learning techniques such as linear regression, multi linear regression ,mean squared error, root mean squared error, mean absolute error concepts.

c. System architecture:

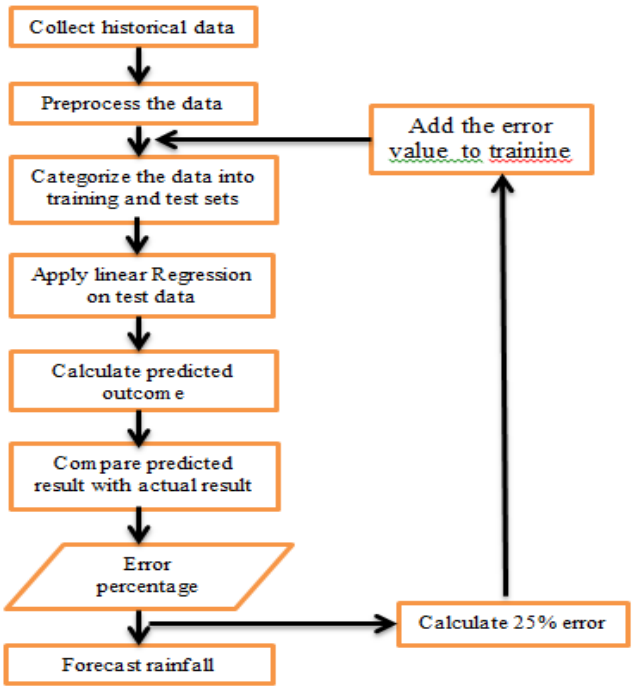
Revised Manuscript Received on June 22, 2020.

* Correspondence Author

Dr. G. DEEPA, Assistant Professor, Department of Mathematics, School of Advanced Sciences, Vellore Institute of Technology, Tamil Nadu, India.

Dinesh.B, Student, M.Sc Data Science, Department of Mathematics, School of Advanced Sciences, Vellore Institute of Technology, Tamil Nadu, India.

Naveen Kumar.K, Student, M.Sc Data Science, Department of Mathematics, School of Advanced Sciences, Vellore Institute of Technology, Tamil Nadu, India.



Data collection:
The data set we used in this paper was from Kaggle lnc which is an open source dataset and consists of 4000 records with 19 parameters. Out of these 19 parameters only 4 were chosen which are bound to affect the yearly rainfall. Parameters such as Jan-Feb, Mar-May, are independent variables. Annual is a dependent variable on several other independent variables.

Preprocessing:
It is a technique by changing raw data into a reasonable format. Real-world data is frequently inconsistent, incomplete or lacking in certain behaviors and is possible to cover several errors. It includes the procedure of finding out missing and redundant data in the data set. Here entire data set is checked for NaN and whichever observation contains of NaN will be erased. Therefore, this gets consistency in the dataset. Anyway in our dataset, there were no missing values qualities discovered implying that each record was established by comparing its feature values.

Data classification:
It spreads raw data into a suitable number of groups according to the values of the variable.

Data regression:
Regression is a statistical approach to seek out the relationship between the variables.

Linear Regression (LR):
LR makes the task to predict a dependent variable value (y) based on a given independent variable value (x). So, this regression method finds out a linear relationship between x (input) and y (output). Hence, it is called LR. It is of the form $y= \alpha + \beta.x$ where α (intercept) and β (slope) are called regression coefficients.
In this paper, LR used to predict the amount of rainfall (that is how many inches of rainfall we can expect). It is significant to closely determine the rainfall for current use of water resources, crop productivity and pre-planning of water structures.

Multiple linear regressions (MLR):
MLR is a model to predict a dependent variable (y) from two or more independent variables (xi).
It is of the form:

$y= \alpha 0+ \beta 1x1 + \beta 2x2 + \beta 3x3+ \beta 4x4+..+ \beta txt+e$

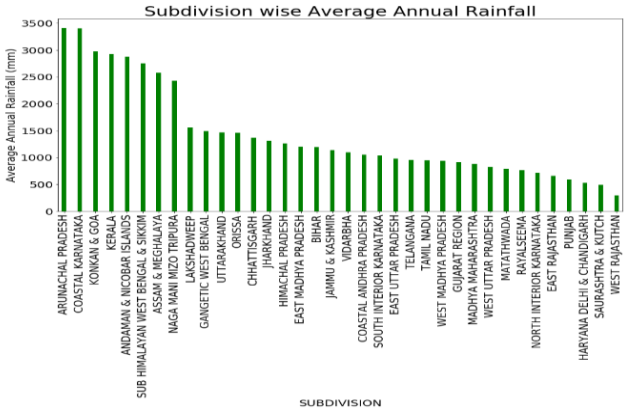
- ✓ Where $\alpha 0$, $\beta 1$, $\beta 2$, $\beta 3$, $\beta 4$… βt are regression coefficients
- ✓ e is the model's error (residuals).

Mean squared error (MSE):
The MSE conveys us how close a Regression Line is to a set of points. It ensures this by taking the distances from the points to the regression line (these distances are the "errors") and squaring them. The squaring is essential to eliminate any negative signs. It also provides extra weight to larger differences. It's called the MSE.

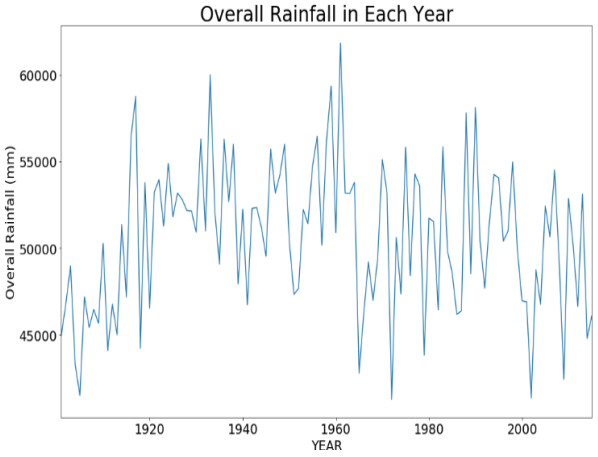
Root mean square error (RMSE):
It is the standard deviations of the residuals (prediction errors). Residuals are a measure of how far from the regression line data points are; RMSE is a measure of how spread out these residuals are. Further, it states that how focused the data is around the line of best fit. RMSE is generally used in forecasting, regression analysis and climatology to confirm experimental results.

Mean absolute error (MAE):
The MAE is the average of all absolute errors.

IV. RESULTS AND DISCUSSION

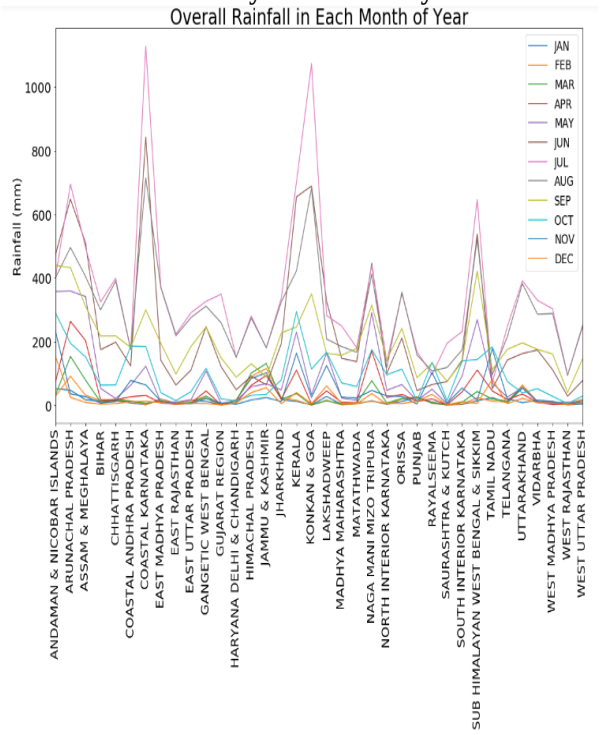


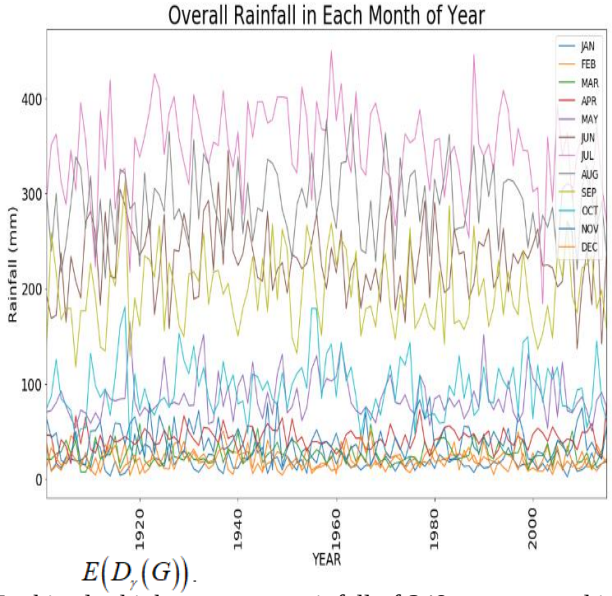
- Subdivisions with highest annual Rainfall are "Arunachal Pradesh", "Coastal Karnataka" and "Konkan & Goa" with approximate annual rainfall of 3418mm, 3408mm and 2977mm respectively.
- Subdivisions with lowest annual rainfall are "West Rajasthan", "Saurashtra & Kutch" and "Haryana Delhi & Chandigarh" with approximate annual rainfall of 292mm, 495mm and 530mm respectively



- Maximum overall rainfall (sum of all 36 subdivision) of 61815mm occurred in the year 1961.
- Minimum overall rainfall (sum of all 36 subdivision) of 41273mm occurred in the year 1972.
- Average (of all 36 subdivision) overall rainfall (sum of all 36 subdivision) is 50182mm.

Overall rainfall in every month of the year:



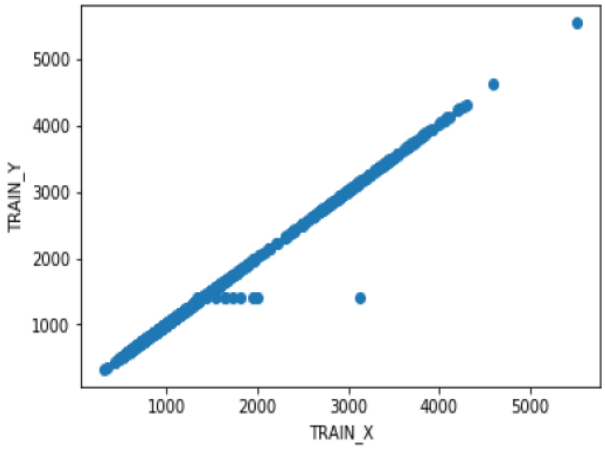


$E(D_\gamma(G))$.

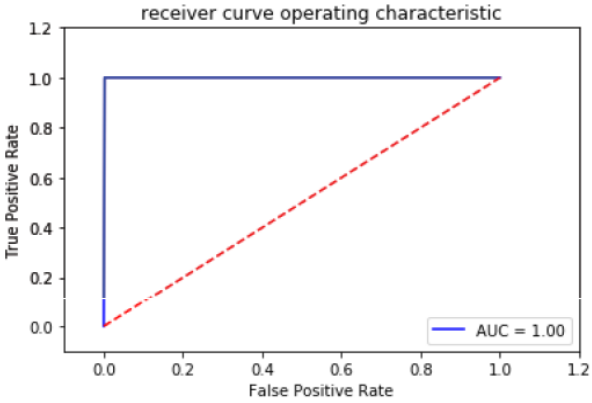
- In this plot highest average rainfall of 348mm occurred in the month of July.
- In this plot lowest average rainfall of 19mm occurred in the month of January.

Multiple Linear regression models between annual rainfall and the periodic rainfall.
Train x shape (2881, 4) ; Test_x (1235, 4)
Train y shape (2881,) ; Test_y (1235,)
Mean Squared Error = 3326.4157535418863
Root Mean Squared Error = 57.67508780697162
Mean Absolute Error = 10.953757241508946
r2_score = 0.9958637383726687.

Prediction:



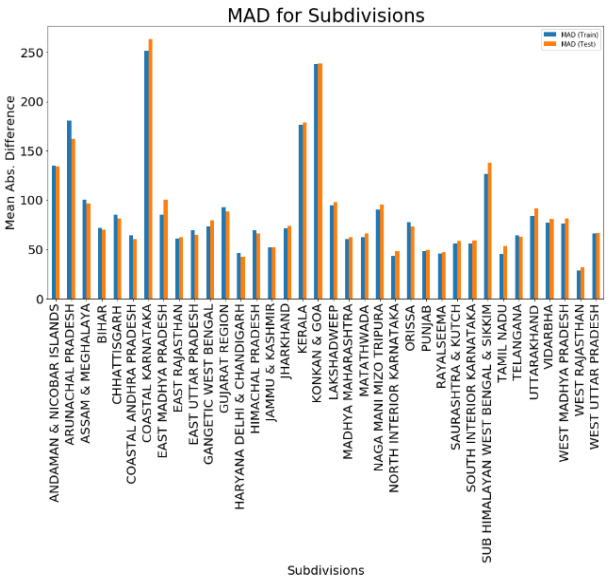

```
accuracy
0.9983805668016195


classification
              precision    recall  f1-score   support

        high       1.00      0.99      1.00       237
         low       1.00      1.00      1.00       998

avg / total       1.00      1.00      1.00      1235
```



Using linear regression: $\sum_{i=1}^{n} |\delta_i|$

To begin with, we shall predict the rainfall for the current month with predi $\sum_{i=1}^{n} |\delta_i|$ ariables as the rainfall in previous three months. Da arranged into 36810 rows and 4 columns with first three columns as the predictor variables and the last column is dependent variable. For each month from April to December, four columns are appended at the bottom of the data to create a data of 36810 rows and 4 columns. Training/Testing is split in the ratio 80:20 randomly.

```
MAD(Training Data) :93.9377544009
MAD (Test Data): 93.1975199159
```




```
Overall MAD (Training): 85.954531315
Overall MAD (Testing): 88.67399873
```

V. RESULT:

Since the Mean Absolute Error and Root Mean Squared Error are relatively less values and scatter plot is pretty straight line. There will 36 different linear models for each category corresponding to each subdivision. First Data is extracted for each subdivision. For each month from April to December, data is appended to end of the data from. There will be four columns (one dependent variable and three independent variables). And for each subdivision the total number rows in the data come out to be 115*9 (approximately. 9 months (april to dec) appended at the bottom). Training/Testing is split in the ratio 80:20 randomly.

VI. CONCLUSION:

It is one the best important natural phenomenon it's not only for the human beings but also for the all other living organisms. Our study pointed a structure of expecting system using linear regression that can predict annual rainfall exactly and efficiently with minimum error. Finally using linear regression algorithms we analyzed and predicted the annual rainfall.

REFERENCES:

1. Goswami.P and Srividya, "A novel Neural network design for long range prediction of rainfall pattern," *Current Sci.(Bangalore)*, vol. 70, no. 6, pp. 447-457, 1996.
2. Venkatesanet.C, S. D. Raskar , S. S. Tambe , B. D. Kulkarni , and R.N. Keshavamurty , "Prediction of all India summer monsoon rainfall using Error Back-Propagation Neural Networks," *Meteorology and Atmospheric Physics*, pp. 225-240, 1997.
3. Sahai.A. K., M. K. Soman, and V. Satyan, "All India summer monsoon rainfall prediction using an Artificial Neural Network," *Climate dynamics*, vol. 16, no. 4, pp. 291-302, 2000.
4. Philip.N.S. and K. B. Joseph, "On the predictability of rainfall in Kerala-An application of ABF neural network," *Computational Science- ICCS*, Springer Berlin Heidelberg, pp. 1-12, 2001.
5. Somvanshi.V. K., O. P. Pandey, P. K. Agrawal, N.V.Kalanker1, M.Ravi Prakash, and Ramesh Chand, "Modeling and prediction of rainfall using Artificial neural Network and ARIMA techniques," *J. Ind. Geophys. Union*, vol.10, no.2, pp. 141-151, 2006.
6. Chattopadhyay.S. and G. Chattopadhyay, "Comparative study among different neural net learning algorithms applied to rainfall time series", *Meteorological applicat*., vol. 15, no. 2, pp. 273-280, 2008.
7. Wu. C. L, K. W. Chau, and C. Fan, "Prediction of rainfall time series using Modular Artificial Neural Networks coupled with data preprocessing techniques," *J. of hydrology*, vol. 389, no. 1, pp. 146-167, 2010.
8. Htike. K. K and O. O. Khalifa, "Rainfall forecasting models using Focused Time-Delay Neural Networks," *Comput. and Commun. Eng.(ICCCE)*, Int. Conf. on IEEE, 2010.
9. Kannan.M., S.Prabhakaran, P.Ramachandran, "Rainfall Forecasting Using Data Mining Technique", *International Journal of Engineering and Technology*, Vol.2 (6), 397-401, 2010.
10. G. Geetha and R. S. Selvaraj, "Prediction of monthly rainfall in Chennai using Back Propagation Neural Network model," *Int. J. of Eng. Sci. and Technology*, vol. 3, no. 1, pp. 211 213, 2011.

AUTHORS PROFILE

Dr. G. DEEPA, Assistant Professor, Department of Mathematics, School of Advanced Sciences, Vellore Institute of Technology, Tamil Nadu, India.

Dinesh B, Student, M.Sc Data Science, Department of Mathematics, School of Advanced Sciences, Vellore Institute of Technology, Tamil Nadu,India

Naveen Kumar.K, Student, M.Sc Data Science, Department of Mathematics, School of Advanced Sciences, Vellore Institute of Technology, Tamil Nadu,India