Prediction of Rainfall based on Statistical and Computational Approach

Writaparna Mukherjee
Institute of Engineering & Management, Kolkata. India
writaparnam@iemcal.com

Angshuman Ray
Institute of Engineering & Management, Kolkata. India
angshuman@iemcal.com

Bimal Datta
The Budge Budge Institute of Technology, Budge Budge. India
bkd_hetc@yahoo.co.in

Pintu Pal
Asansol Engineering College, Asansol. India
p5_pal@yahoo.co.in

***Abstract*— A comparative study is done in this paper in the prediction of rainfall at ground level multiple linear regression and, feature selection and k-means clustering method. Based on the past observations of the last three days atmospheric parameters like minimum and maximum Temperature, minimum and maximum relative humidity, minimum and maximum air pressure, minimum and maximum vapour pressure and minimum and maximum radiation the model is developed. In this paper it is observed that considering the seasonality effect better results can be achieved. It has also been observed that the selection of appropriate features can also improve the performance of the prediction.**

Keywords*— *Multiple linear regression, feature selection, k-means clustering, rainfall.

I. INTRODUCTION

The turbulent idea of the environment, the gigantic computational force needed to settle the conditions that depict the climate including different boundaries like most extreme temperature, least temperature, relative mugginess, fume pressure, wind speed and bearing, precipitation fall, and so forth which are difficult to count and quantify precisely. Many important initiatives to solve the challenge of weather forecasting using statistical modelling, including machine learning techniques, have been published in the last decade.

In temperature gauging one needs to recognize the occasions the gauge goes on, for instance temperature one hour ahead or least and greatest temperature of a given day. A few works has been done and diverse statistical and computational models have been tried. Hiyam Abobaker Yousif Ahmed and Sondos W. A. Mohamed portray a model of rainfall prediction using temperature, wind speed, and dew point for Khartoum state. The efficiency of the model has been measured by comparing the average value of the mean square error of the training data with the test data. Barrera-Animas and Ari Yair depicted a model which predict rainfall an hourly basis using time series data. Isabelle Roesch and Tobias Günther designed a comprehensive and interactive system that allows users to study the output of recurrent neural networks on both the complete training data and testing data. They follow a coarse-to-fine strategy, providing overviews of annual, monthly and daily patterns in the time series and directly support a comparison of different hyper-parameter settings. They applied our method to a recurrent convolutional neural network that was trained and tested on 25 years of climate data to forecast meteorological attributes, such as temperature, pressure and wind velocity.

II. DATA PREPARATION

The current investigation built up a conglomerate model dependent on the past perceptions of a few meteorological boundaries like temperature, humidity, vapour, radiation etc. as a contribution for preparing the model. The information was gathered day by day by the meteorological division of Dumdum Airport. The parameters of the data acquisition are:

i. Minimum Temperature (Min. Temp(t))
ii. Maximum Temperature (Max. Temp(t))
iii. Minimum Relative Humidity (Min. RH(t))
iv. Maximum Relative Humidity (Max. RH(t))
v. Minimum Air Pressure (Min. Press.(t))
vi. Maximum Air Pressure (Max. Press.(t))
vii. Minimum Vapour Pressure (Min. VP(t))
viii. Maximum Vapour Pressure (Max. VP(t))
ix. Minimum Radiation (Min. Rad.(rd))
x. Maximum Radiation (Max. Rad.(rd))
xi. Rainfall (Rain(r))

The prediction was on rainfall on the basis of past three days atmospheric parameters. This information is stored in an input file. The file contains data of seven years. So, there is an observation of 10 variables three consecutive days, say *t*. In the model the rainfall for the next (t_{th}) day is determined by the atmospheric parameters for the current day i.e. day *(t-3)*. To empower the determination of the best model, the preparation informational collection should cover precipitation at various seasons. So the information for whole years were picked as the preparation informational collections. The data is pre-processed before training. Our model has the following functional relations.

Rainfall(r) = f(Temp_Min(t-1), Temp_Max(t-1), Hum_Mint(t-1), Hum_Max(t-1), Pressure_Min(t-1), Pressure_Max(t-1), Vapor_Min(t-1), Vapor_Max(t-1), Rad_Min(t-1), Rad_Max(t-1), Rainfall(t-1), Temp_Min(t-2), Temp_Max(t-2), Hum_Mint(t-2), Hum_Max(t-2), Pressure_Min(t-2), Pressure_Max(t-2), Vapor_Min(t-2), Vapor_Max(t-2), Rad_Min(t-2), Rad_Max(t-2), Rainfall(t-2), Temp_Min(t-3), Temp_Max(t-3), Hum_Mint(t-3), Hum_Max(t-3), Pressure_Min(t-3), Pressure_Max(t-3), Vapor_Min(t-3), Vapor_Max(t-3), Rad_Min(t-3), Rad_Max(t-3), Rainfall(t-3)) (1)

III. METHODOLOGY

The target of regression analysis is to choose the assessments of limits for a limit that cause the ability to best fit a lot of data insights that you give. In linear regression, the capacity is a linear (straight-line) equation. Likewise with connection, regression is utilized to analyze the connection between two continuous (scale) factors. Nonetheless, regression is more qualified for studying practical conditions between factors. The term useful dependency infers that X partially determines the level of Y.

Furthermore, regression is more qualified than relationship for studying tests in which the specialist fixes the appropriation of X. Regression analysis is utilized to anticipate a consistent ward variable from various autonomous factors. On the off chance that the reliant variable is dichotomous, calculated regression ought to be utilized. (If the split between the two levels of the dependent variable is almost half, then, at that point both determined and direct regression will end up giving relative results). The free factors used in regression can be either constant or dichotomous. Autonomous factors with beyond what two levels can likewise be utilized in regression analyses, however they initially should be changed over into factors that have only two levels. This is called dummy coding and will be talked about later. As a rule, regression analysis is used with normally happening factors, as opposed to tentatively controlled variables, regardless of the way that you can use regression with tentatively controlled components. One highlight remember with regression analysis is that causal connections among the factors can't be resolved. While the terminology is to such an extent that we say that X "predicts" Y, we can't say that X "causes" Y.

In this paper the different independent variables we have considered as

i. Minimum Temperature (Temp_Min(t))
ii. Maximum Temperature (Temp_Max(t))
iii. Minimum Relative Humidity (Hum_Mint(t))
iv. Maximum Relative Humidity (Hum_Max(t))
v. Minimum Air Pressure (Pressure_Min(t))
vi. Maximum Air Pressure (Pressure_Max(t))
vii. Minimum Vapour Pressure (Vapor_Min(t))
viii. Maximum Vapour Pressure (Vapor_Max(t))
ix. Minimum Radiation (Rad_Min(t))
x. Maximum Radiation (Rad_Max(t))

And the dependent variable is Rainfall (Rain(r)).

After the initial regression were performed in the normalized data set, the feature selection algorithm was applied in it. The appropriate features were only kept and the data set were cleaned on the basis of the same. After applying feature selection k-means clustering was applied in the data set. As expected there were 4 significant clusters were the output of the application as the data was collected season wise. Then again the multiple linear regression was performed in each cluster and a significant improved results were found.

IV. RESULT AND OBSERVATIONS

After the input file is prepared, the training is done taking into consideration all the parameters. After the training process is over, multiple linear regression was performed on the collected data set. The result is shown below.

TABLE I. OUTPUT OF MULTIPLE LINEAR REGRESSION ON THE GIVEN DATA SET

Coefficients	Estimate	Std. Error
Regression Constant	675.674930	112.833767
Pressure_Max	0.083326	0.304744
Pressure_Min	0.302526	0.288641
Vapor_Max	-0.05599	0.200713
Vapor_Min	-0.14914	0.185311
Hum_Max	-0.01152	0.051431
Hum_Min	0.057332	0.061827
Temp_Max	0.085435	0.233911
Temp_Min	0.449337	0.243745
Rad_Max	0.320937	0.698987
Rad_Min	0.281948	0.981226
Rainfall_1	0.012865	0.026139
Pressure_Max_2	0.07553	0.315234
Pressure_Min_2	-0.14694	0.32006
Vapor_Max_2	0.557232	0.199978
Vapor_Min_2	0.106149	0.185662
Hum_Max_2	-0.1728	0.051868
Hum_Min_2	-0.06442	0.06022
Temp_Max_2	0.068554	0.252792
Temp_Min_2	-0.35544	0.274286
Rad_Max_2	0.380665	0.720005
Rad_Min_2	0.913103	0.987394
Rainfall_2	-0.00706	0.026926
Pressure_Max_3	-0.97165	0.276571
Pressure_Min_3	-0.01137	0.30947
Vapor_Max_3	-0.5605	0.199342
Vapor_Min_3	-0.36748	0.186541
Hum_Max_3	0.266316	0.053075
Hum_Min_3	0.158933	0.060562
Temp_Max_3	-0.20494	0.253389
Temp_Min_3	0.258418	0.254571
Rad_Max_3	-2.84136	0.673489
Rad_Min_3	-2.04699	0.987444
Rainfall_3	0.058032	0.025154

The output of multiple linear regression gives a linear equati on, which is mentioned below:

Rainfall = (Pressure_Max) * (0.083326) + (Pressure_Min) * (0.302526) + (Vapor_Max) * (-0.05599) + (Vapor_Min) * (-0.14914) + (Hum_Max) * (0.01152) + (Hum_Min) * (0.057332)+Temp_Max) * (0.085435) + (Temp_Min) * (0.449337) + (Rad_Max) * (0.320937) + (Rad_Min) * (0.281948) + (Rainfall_1) * (0.012865) + (Pressure_Max_2) * (0.07553) + (Pressure_Min_2) * (0.14694) +

(Vapor_Max_2) * (0.557232) + (Vapor_Min_2) * (0.106149) + (Hum_Max_2) * (-0.1728) + (Hum_Min_2)* (0.06442) + (Temp_Max_2) * (0.068554) + (Temp_Min_2) * (-0.35544) + (Rad_Max_2) * (0.380665) + (Rad_Min_2) * (0.913103) + (Rainfall_2) * (-0.00706) + (Pressure_Max_3) * (-0.97165) + (Pressure_Min_3) * (-0.01137) + (Vapor_Max_3) *(-0.5605) + (Vapor_Min_3) * (-0.36748)+ (Hum_Max_3)* (0.266316) + (Hum_Min_3)* (0.158933) + (Temp_Max_3) * (-0.20494) + (Temp_Min_3)* (0.258418) + (Rad_Max_3) * (-2.84136) + (Rad_Min_3) * (-2.04699) + (Rainfall_3) * (0.058032) +(675.67493) (2)

It was recorded that Residual standard error is 11.54 on 140 0 degrees of freedom and Multiple R-squared value is 0.231 9, Adjusted R-squared value is 0.2138.The root mean square d error is 15.83106, which is significantly high.

To achieve more accuracy of the model, feature selection algorithm was applied in the data set to reduce the number of variable of the data set to reduce the computational cost of modelling. Here the " Boruta" package were used to perform feature selection in R programming and 500 iterations were performed.

Boruta is a feature selection algorithm. Precisely, it works as a wrapper algorithm around Random Forest. This package derive its name from a demon in Slavic mythology who dwelled in pine forests. This technique achieves supreme importance when a data set comprised of several variables is given for model building. Below is the step wise working of boruta algorithm:

1. Firstly, it adds randomness to the given data set by creating shuffled copies of all features (which are called shadow features).
2. Then, it trains a random forest classifier on the extended data set and applies a feature importance measure (the default is Mean Decrease Accuracy) to evaluate the importance of each feature where higher means more important.
3. At every iteration, it checks whether a real feature has a higher importance than the best of its shadow features (i.e. whether the feature has a higher Z score than the maximum Z score of its shadow features) and constantly removes features which are deemed highly unimportant.
4. Finally, the algorithm stops either when all features gets confirmed or rejected or it reaches a specified limit of random forest runs.

Out of 33 parameters 28 attributes were confirmed important, 3 attributes confirmed unimportant: Rad_Max, Rad_Max_2, Rainfall_1 and 2 tentative attributes left: Rad_Min_2, Rainfall_2.The results are shown below.

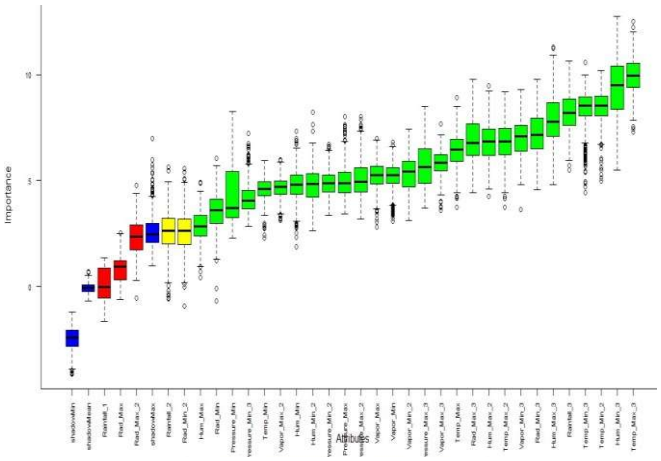

Fig. 1. Correlation of the features of the data set

TABLE II. BORUTA FINAL DECISION

Features	Boruta final decision
Pressure_Max	confirmed
Pressure_Min	confirmed
Vapor_Max	confirmed
Vapor_Min	confirmed
Hum_Max	confirmed
Hum_Min	confirmed
Temp_Max	confirmed
Temp_Min	confirmed
Rad_Max	unimportant
Rad_Min	confirmed
Rainfall_1	unimportant
Pressure_Max_2	confirmed
Pressure_Min_2	confirmed
Vapor_Max_2	confirmed
Vapor_Min_2	confirmed
Hum_Max_2	confirmed
Hum_Min_2	confirmed
Temp_Max_2	confirmed
Temp_Min_2	confirmed
Rad_Max_2	unimportant
Rad_Min_2	tentative
Rainfall_2	tentative
Pressure_Max_3	confirmed
Pressure_Min_3	confirmed
Vapor_Max_3	confirmed
Vapor_Min_3	confirmed
Hum_Max_3	confirmed
Hum_Min_3	confirmed
Temp_Max_3	confirmed
Temp_Min_3	confirmed
Rad_Max_3	confirmed
Rad_Min_3	confirmed
Rainfall_3	confirmed

After the feature selection the unimportant and tentative features were discarded and a new data set was created and linear regression also performed.

TABLE III. THE OUTPUT OF LINEAR REGRESSION AFTER THE FEATURE SELECTION

Coefficients	Estimate	Std. Error
Regression Constant	-0.00329	0.023221
Pressure_Max	0.134813	0.146964
Pressure_Min	0.059284	0.136995
Vapor_Max	-0.14409	0.114682
Vapor_Min	0.05452	0.111223
Hum_Max	0.051937	0.046166
Hum_Min	0.003889	0.070611
Temp_Max	-0.00853	0.066304
Temp_Min	0.135283	0.093188
Rad_Min	0.080893	0.03135
Pressure_Max_2	0.095757	0.158709
Pressure_Min_2	-0.16336	0.154165
Vapor_Max_2	0.12838	0.118011
Vapor_Min_2	0.292234	0.112391
Hum_Max_2	-0.11169	0.047715
Hum_Min_2	-0.15908	0.068584
Temp_Max_2	-0.00408	0.071758
Temp_Min_2	-0.11303	0.10533
Pressure_Max_3	-0.65068	0.148693
Pressure_Min_3	0.257571	0.157747
Vapor_Max_3	-0.16939	0.11424
Vapor_Min_3	-0.21864	0.110457
Hum_Max_3	0.18574	0.047743
Hum_Min_3	0.085519	0.068532
Temp_Max_3	-0.05347	0.070447
Temp_Min_3	0.078956	0.099122
Rad_Max_3	-0.1709	0.031131
Rad_Min_3	-0.00455	0.031815
Rainfall_3	0.032378	0.027234

This output of multiple linear regression also gives a linear e quation, which is mentioned below:

Rainfall = (Pressure_Max) * (0.134813) + Pressure_Min * (0.059284) + Vapor_Max * (-0.144087) + Vapor_Min * (0.05452) + Hum_Max * (0.051937) + Hum_Min * (0.003889) + Temp_Max * (-0.008531) + Temp_Min * (0.135283) + Rad_Min * (0.080893) + Pressure_Max_2 * (0.095757) + Pressure_Min_2 * (-0.16336) + Vapor_Max_2 * (0.12838) + Vapor_Min_2 * (0.292234) + Hum_Max_2 * (-0.111692) + Hum_Min_2 * (-0.159078) + Temp_Max_2 * (-0.004077) + Temp_Min_2 * (-0.113032) + Pressure_Max_3 * (-0.650682) + Pressure_Min_3 * (0.257571) + Vapor_Max_3 * (-0.169391) + Vapor_Min_3 * (-0.218637) + Hum_Max_3 * (0.18574) + Hum_Min_3 * (0.085519) + Temp_Max_3 * (-0.053472) + Temp_Min_3 * (0.078956) + Rad_Max_3 * (-0.170903) + Rad_Min_3 * (-0.00455) + Rainfall_3 * (0.032378) + (-0.003291) (3)

It was recorded that Residual standard error is 0.86 on 1405 degrees of freedom and Multiple R-squared value is 0.227, Adjusted R-squared value is 0.2115.The root mean squared error is 11.7406, which is also very high.

Then k-means clustering were performed on the above said data set. First the total within sum of squared calculated and wss plot was made by using elbow method. This method determined that the size of " k" will be 4. The method was also validated by gap_stat method in R language to determine the number of cluster.

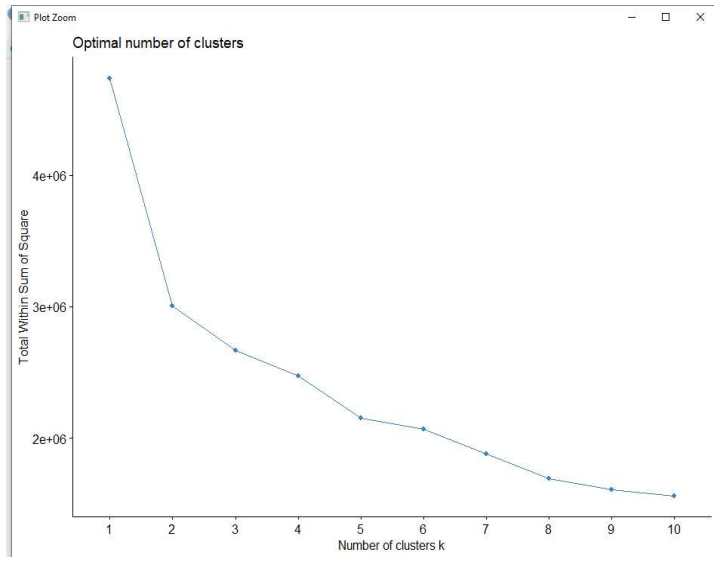

Fig. 2. Wss plot showing the optimum number of clusters

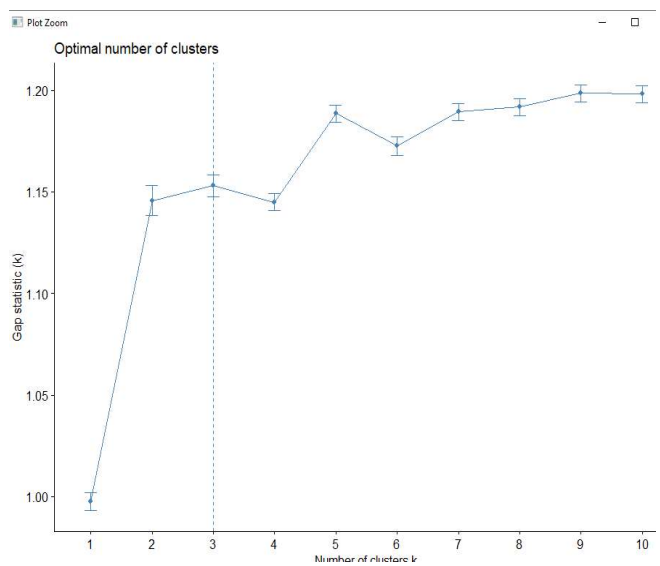

Fig. 3. Gap_stat method showing the optimum number of clusters

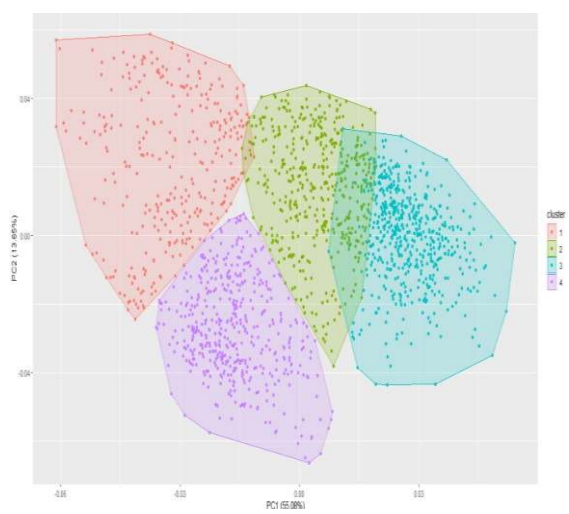

Fig. 4. Clusters of the data set (k-means where k=4)

Each cluster was fed into the model and multiple linear regression were applied again to achieve the result. A significant amount of accuracy were recorded in the output.

TABLE IV. THE OUTPUT OF LINEAR REGRESSION AT CLUSTER II

Coefficients	Estimate	Std. Error
Regression Constant	63.234355	159.459190
Pressure_Max	0.083326	0.304744
Pressure_Min	0.302526	0.288641
Vapor_Max	-0.05599	0.200713
Vapor_Min	-0.14914	0.185311
Hum_Max	-0.01152	0.051431
Hum_Min	0.057332	0.061827
Temp_Max	0.085435	0.233911
Temp_Min	0.449337	0.243745
Rad_Max	0.320937	0.698987
Rad_Min	0.281948	0.981226
Rainfall_1	0.012865	0.026139
Pressure_Max_2	0.07553	0.315234
Pressure_Min_2	-0.14694	0.32006
Vapor_Max_2	0.557232	0.199978
Vapor_Min_2	0.106149	0.185662
Hum_Max_2	-0.1728	0.051868
Hum_Min_2	-0.06442	0.06022
Temp_Max_2	0.068554	0.252792
Temp_Min_2	-0.35544	0.274286
Rad_Max_2	0.380665	0.720005
Rad_Min_2	0.913103	0.987394
Rainfall_2	-0.00706	0.026926
Pressure_Max_3	-0.97165	0.276571
Pressure_Min_3	-0.01137	0.30947
Vapor_Max_3	-0.5605	0.199342
Vapor_Min_3	-0.36748	0.186541
Hum_Max_3	0.266316	0.053075
Hum_Min_3	0.158933	0.060562
Temp_Max_3	-0.20494	0.253389
Temp_Min_3	0.258418	0.254571
Rad_Max_3	-2.84136	0.673489
Rad_Min_3	-2.04699	0.987444
Rainfall_3	0.058032	0.025154

The output of multiple linear regression gives a linear equation, which is mentioned below:

Rainfall = (Pressure_Max) * (-0.170394) + (Pressure_Min) * (0.261837) + (Vapor_Max) * (0.146748) + (Vapor_Min) * (-0.094494) + (Hum_Max) * (-0.052744) + (Hum_Min) * (0.001777) + (Temp_Max) * (-0.015656) + (Temp_Min) * (0.01272) + (Rad_Max) * (0.501028) + (Rad_Min) * (0.162517) + (Rainfall_1) * (0.04587) + (Pressure_Max_2) * (0.063313) + (Pressure_Min_2) * (0.039806) + (Vapor_Max_2) * (-0.034743) + (Vapor_Min_2) * (0.002312) + (Hum_Max_2) * (0.304489) + (Hum_Min_2) * (0.151876) + (Temp_Max_2) * (-0.540959) + (Temp_Min_2) * (0.16598) + (Rad_Max_2) * (-0.139332) + (Rad_Min_2) * (-0.185507) + (Rainfall_2) * (0.059424) + (Pressure_Max_3) * (0.122321) + (Pressure_Min_3) * (0.100949) + (Vapor_Max_3) * (-0.391693) + (Vapor_Min_3) * (1.257434) + (Hum_Max_3) * (2.082708) + (Hum_Min_3) * (-0.020157) + (Temp_Max_3) * (-0.20494) + (Temp_Min_3) * (0.258418) + (Rad_Max_3) * (-2.84136) + (Rad_Min_3) * (-2.04699) + (Rainfall_3) * (0.058032) +(63.234355) (4)

It was recorded that Residual standard error is 3.831 and Multiple R-squared value is 0.1505, Adjusted R-squared value is 0.04529. The root mean squared error is 2.320498, which is significantly lower than the previous experiment.

TABLE V. A COMPARISON BETWEEN THE OUTPUT BEFORE AND AFTER FEEDING THE MODEL

	Output of standard multiple linear regression	Output after the feature selection	Output after the data set was fed into the model
Residual standard error	11.54	0.86	3.831
Multiple R-squared	0. 2319	0.227	0.1505
Adjusted R-squared	0. 2138	0.2115	0. 04529
Root mean squared error	15. 83106	11.7406	2. 320498

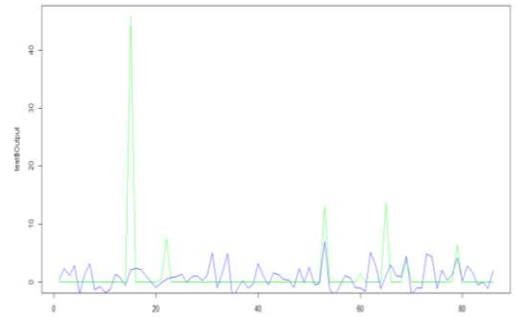

Fig. 5. Comparison between the actual and predicted data before fed into the model

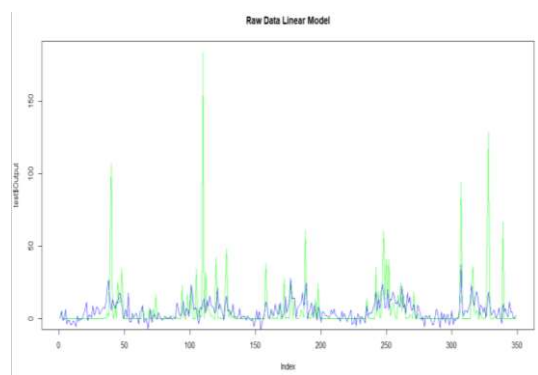

Fig. 6. Comparison between the actual and predicted data after fed into the model

V. CONCLUSIONS

This comparison talked about here has been created to foresee rainfall for a specific day dependent on the information of past three days. The meteorological information were gathered from Kolkata Meteorological center and utilized for investigation of the proposed model. A conglomeration of statistical and computational models were tried to figure the yield and this processed yield was contrasted and the objective yield for example rainfall. After testing these models, the following conclusions are made.

i. The clustering approach turns out to be an excellent tool that can predict the rainfall accurately by overcoming the seasonality effect on air pressure.
ii. The comparative model proposed here can be good alternatives for traditional meteorological approaches for weather forecasting. In the future works, the combined use of Feature selection and Artificial neural network may result in an excellent paradigm for prediction of air pressure. Moreover, we can also incorporate time series analysis in our data set to get more accurate result.
iii. In this paper the sensitivity analysis can also be incorporated to determine how different values of an independent variable affect a particular dependent variable under a given set of assumptions.

REFERENCES

Ahmed, H. A. Y., & Mohamed, S. W. (2021, February). Rainfall Prediction using Multiple Linear Regressions Model. In *2020 International Conference on Computer, Control, Electrical, and Electronics Engineering (ICCCEEE)* (pp. 1-5). IEEE.

Fayaz, S. A., Zaman, M., & Butt, M. A. (2022). Knowledge discovery in geographical sciences—a systematic survey of various machine learning algorithms for rainfall prediction. In *International Conference on Innovative Computing and Communications* (pp. 593-608). Springer, Singapore.

Barrera-Animas, A. Y., Oyedele, L. O., Bilal, M., Akinosho, T. D., Delgado, J. M. D., & Akanbi, L. A. (2022). Rainfall prediction: A comparative analysis of modern machine learning algorithms for time-series forecasting. *Machine Learning with Applications*, *7*, 100204.

Mohd, R., Butt, M. A., & Baba, M. Z. (2022). Grey wolf-based linear regression model for rainfall prediction. *International Journal of Information Technologies and Systems Approach (IJITSA)*, *15*(1), 1-18.

Duan, T., Anand, A., Ding, D. Y., Thai, K. K., Basu, S., Ng, A., & Schuler, A. (2020, November). Ngboost: Natural gradient boosting for probabilistic prediction. In *International Conference on Machine Learning* (pp. 2690-2700). PMLR.

Roesch, I., & Günther, T. (2019, February). Visualization of neural network predictions for weather forecasting. In *Computer graphics forum* (Vol. 38, No. 1, pp. 209-220).

Das, R., Mishra, J., Mishra, S., & Pattnaik, P. K. (2022). Design of mathematical model for the prediction of rainfall. *Journal of Interdisciplinary Mathematics*, *25*(3), 587-613.

Gad, I., & Hosahalli, D. (2022). A comparative study of prediction and classification models on NCDC weather data. *International Journal of Computers and Applications*, *44*(5), 414-425.