Rainfall Prediction Using Machine Learning

1st Vineetha Gourishetty 2nd Naveen Kumar Mahendarker 3rd Janaki Rama Satyanarayana Murthy Toleti

4th Kumar Baddula

***Abstract*—Massive rainfall forecast is a significant problem for the meteorological department. This paper investigates the performance of the various Machine Learning (ML) models, namely Lasso regression, ridge regression, elastic net regression, random forest, gradient boosting, and decision tree regressor. To predict rainfall, several types of research have been conducted using data mining and machine learning techniques of different countries' environmental datasets. An erratic rainfall distribution in the country affects the agriculture on which the economy of the country depends on. Correct usage of rainfall water should be desinged and practiced in the country to minimize the problem of the drought and flood occurred in the country. The main objective of this study is to identify the relevant atmospheric features that cause rainfall and predict the intensity of daily rainfall using machine learning techniques.Root mean squared error and Mean absolute Error methods were used to measure the performance of the machine learning model. The result of the study revealed that the Extreme Gradient Boosting machine learning algorithm performed better than others.**

***Index Terms*—Machine Learning, ridge regression, Elasticnet, Random forest, Mean Absolute Error, Root mean square**

I. Introduction

In the hydrological study, the main problem is accurately predicting the rainfall. Due to natural hazards and storm, farmers will lose and destroy their crops. To avoid these problems, accurately and timely predict the rainfall prediction earlier and give caution more first to farmers. Rainfall is said to be an environmental aspect which affects the human activities such as farming production, construction, energygeneration, forestry and tourism, etc. The rainfall prediction is more essential as it is concerned with the maximum association with other factors like landslides, flooding, avalanches, earthquakes, and so on [1]. The rainfall prediction is more required as it is concerned with the maximum association with other factors like landslides, flooding, avalanches, earthquakes, and so on. Such disasters affect the public severely for many decades [2]. Hence, developing effective model to predict the rainfall helps to prevent the natural disaster to the limited extent [3]. We applied different regression techniques of machine learning algorithms to build the ML models to make accurate and timely predictions. Machine learning is used to study and develop the system behavior model. Machine learning modeling techniques used to design models which can be further predicted vital system parameters with regards to Indian panther ecosystem [4]. This article aims to deliver end to end machine learning life cycle right from Data acquisition to evaluating the models. For evaluation metrics of regress or is R2 , Mean Absolute Error (MAE), Mean Square Error (MSE) and Root Mean Square (RMSE)

The rainfall prediction is more required as it is concerned with the maximum association with other factors like landslides, flooding, avalanches, earthquakes, and so on. Such disasters have affected the public severely for many decades. Hence, developing an effective model to predict rainfall helps to prevent natural disasters to a limited extent.

Several environmental factors affect the existence of rainfall and its intensity. The temperature, relative humidity, sunshine, pressure, evaporation, etc. are some of the factors that affect the existence of rainfall and its intensity directly or indirectly. The study conducted by Chaudhari and Choudhari [12] indicated that temperature, wind, and cyclone were important features of the atmosphere over the Indian region to predict rainfall, however, the study did not measure the correlations of each feature to determine the strength of the independent features on the rainfall. On the other hand, a correlation study by Thirumalai et al. [13] identified the most important features like solar radiation, perceptible water vapor, and diurnal features for rainfall prediction using a linear regression model. Whereas, scholars (for example, [10, 11, 14]) used atmospheric features of temperature, relative humidity, pressure, and wind speed as an important feature to predict rainfall accurately using machine learning Random forest, and multiple linear regression model respectively. Hence, important atmospheric features that have a direct or indirect impact on rainfall should be studied to predict the existence and the intensity of rainfall.

II. Related work

Sawale G.J et al., presented an ANN-based technique to predict atmospheric conditions. They used the dataset that consisted of various weather attributes, e.g., humidity, temperature, and wind speed. The proposed technique integrated the Back Propagation Network and Hopfield Network in such a way that the output of the back propagation network is the input to the Hopfiled network.

https://github.com/Vineetha-gourishetty/ProjectRainfallPrediction
https://github.com/nxm19440/ProjectRainfallPrediction
https://github.com/MurthyToleti/ProjectRainfallPrediction
https://github.com/KumarBaddula/ProjectRainfallPrediction

Abhishek K. et al. used Artifivcial neural networks to predict the monthly average rainfall of monsoon weather in India. A dataset covering a period of 8 months each year was used for prediction. Feed Forward Back Propagation, Layer Recurrent, and Cascaded Feed Forward Back Propagation. According to the results, Feed Forward Back Propagation outperformed the others.

Liu J.N.K. et al., presented a framework with deep neural networks to predict weather changes over the next 24 h. For prediction, they used a dataset covering 30 years, from 1983 to 2012, obtained from Hong Kong Observatory (HKO). The dataset consisted of four weather attributes: temperature, dew point, mean sea level pressure, and wind speed.

Cramer S. et al., compared "Markov Chain extended with rainfall prediction" with other widely used data mining techniques, including Radial Basis, Neural Networks, Genetic Programming, Support Vector Regression, M5 Rules, k-Nearest Neighbors, and M5 Model trees. A dataset obtained from 42 cities was used for the experiment. The results showed that the Markov Chain technique can be outperformed by machine learning techniques.

Wang L et al., developed the technique of modular-based Support Vector Machine (SVM) to predict and simulate rainfall prediction. It includes generation of training sets with the bagging sampling technique, training of SVM kernel function, selection of SVM combination members with the PLS (Partial Least Square) technique, and production of –SVM.

Chen C. et al., conducted an advanced statistical technique for solar power forecasting based on an artificial intelligence approach. With several features as input, such as past power measurements and meteorologically related forecasts. The required metrological data included solar irradiance, relative humidity, and temperature. A SOM (Self organized map) was trained to classify the local weather 24 h in advance with the help of online meteorological services. The proposed method was considered to be suitable for the forecasting of 24 h ahead power output of a PV (photovoltaic) system, as well as for trading in electricity markets of PV power system operators.

III. Motivation

Rainfall prediction is much needed for the current updated society inorder to prevent the consequences of the public to be facing after massive storms without prior safety measures. Rain is the biggest natural source for the feild of agriculture, so the prediction helps for the agricultarists to reep their crop and also to get the good yeild for the crops. ML Algorithms are capable of learning a large amount of malicious and benign requests of different patterns and can predict them effectively in production.

IV. Objectives

This paper is focused on evaluating different Machine Learning Algorithms for the prediction of rainfall using the supplied dataset.

- To prepare the data effectively to train the machine learning model, cleaning data without missing values or null values, results in the effective performance of the model.
- To implement and evaluate the regression models available in machine learning for the considered rainfall dataset.
- To develop a UI with ease of use for the users, using the programming components and modules available in Python.

For this, we have evaluated the following algorithms for prediction:

- Lasso regression
- Ridge regression
- Elasticnet
- Lassolar regression
- Decision Tree Classifier
- Gradient Boosting
- Random forest

V. Proposed frame work

The dataset is loaded and as a first step, the data is preprocessed. After data preprocessing, feature extraction is done. This data is trained to different machine learning models such as Lasso regression, ridge regression, Elastic net regression, Lassolar regression, Decision tree algorithm, Gradient boosting alorithm. After training, testing is done. Finally the result analysis is done. The whole methodology is depicted in Fig. 1

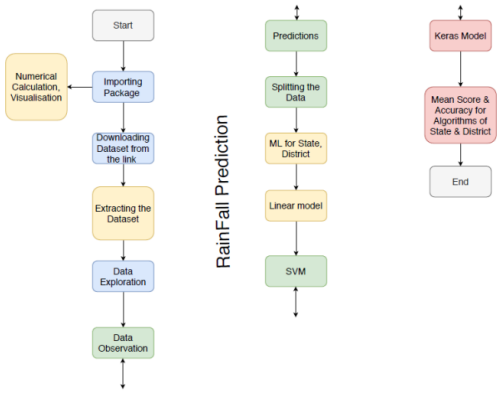

Fig. 1. Process Flow Diagram

A. Dataset Collection

Here we have collected thet dataset with rainfall measures for the states of India for 116 years month wise. Our data consists of 4188 rows and 19 feature columns. This dataset consists of parameters such as (['SUBDIVISION', 'YEAR', 'JAN', 'FEB', 'MAR', 'APR', 'MAY', 'JUN', 'JUL', 'AUG', 'SEP', 'OCT', 'NOV', 'DEC', 'ANNUAL', 'JF', 'MAM', 'JJAS', 'OND']. Here in the below screenshot you can see the first 5 rows of the data that is pulled out from the original dataset.

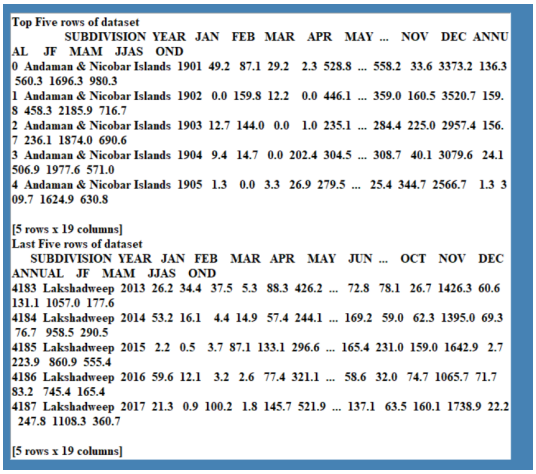

Fig. 2. Process Flow Diagram

B. Analyzing the Data

In this section, we are plotting the data points on the graph grouping them as months and years cumulatively. Here in data analysis we figured out the correlation between the features present in the dataset and also the month wise rainfall over the years and found out the months with highest rainfall also the trend of the rainfall over the year quarterly.

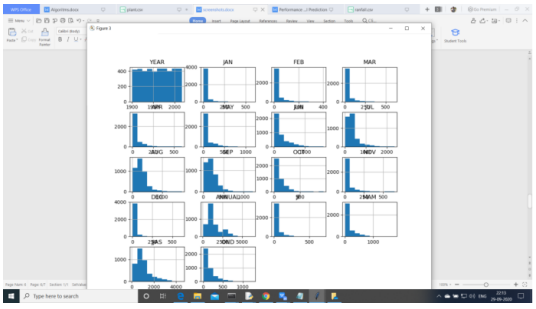

Fig. 3. Month wise rainfall over the years 1901 - 2017

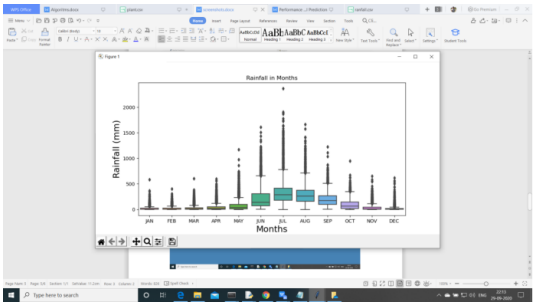

Fig. 4. Cummiliative rainfall over months

C. Splitting data for training and validation

Here in this section we used the sklearn module in python to split our numerical tabular dataset into 2 with 80 percent as training data and remaining as validation data.

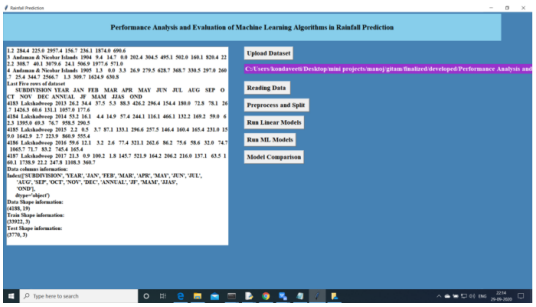

Fig. 5. Test and train split of data

D. Model building

Here in this section we used sklearn module from python to implement the machine learning algorithms.

1) Lasso regression: This is a regularization technique. It is used over regression methods for a more accurate prediction. This model uses shrinkage. Shrinkage is where data values are shrunk towards a central point as the mean. The lasso procedure encourages simple, sparse models (i.e. models with fewer parameters). This particular type of regression is well-suited for models showing high levels of multicollinearity or when you want to automate certain parts of model selection, like variable selection/parameter elimination.

```
# linear model
lassoreg = linear_model.Lasso(alpha=0.9)
lassoreg.fit(X_train, y_train)
y_pred = lassoreg.predict(X_test)
lasso_mae = mean_absolute_error(y_test, y_pred)
text.insert(END,"MAE value for Lasso: "+str(lasso_mae)+"\n")
```

Fig. 6. Lasso regression

2) Ridge regression: Ridge regression is a model tuning method that is used to analyse any data that suffers from multicollinearity. This method performs L2 regularization. When the issue of multicollinearity occurs, least-squares are unbiased, and variances are large, this results in predicted values being far away from the actual values.

```
# linear model
ridgereg = linear_model.Ridge(alpha=0.01)
ridgereg.fit(X_train, y_train)
y_pred = ridgereg.predict(X_test)
ridge_mae = mean_absolute_error(y_test, y_pred)
text.insert(END,"MAE value for Ridge: "+str(ridge_mae)+"\n")
```

Fig. 7. Ridge regression

3) Elasticnet regression: Elastic net linear regression uses the penalties from both the lasso and ridge techniques to regularize regression models. The technique combines both the lasso and ridge regression methods by learning from their shortcomings to improve the regularization of statistical models.

4) Decisiontree regressor: Decision tree builds regression or classification models in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets while

at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes.

```
dtreg = DecisionTreeRegressor(random_state=0)
dtreg.fit(X_train, y_train)
y_pred = dtreg.predict(X_test)
dt_mae = mean_absolute_error(y_test, y_pred)
text.insert(END,"MAE value for Lasso: "+str(dt_mae)+"\n")
```

Fig. 8. Decisiontree regressor regression

5) Random forest regressor: Random Forest Regression is a supervised learning algorithm that uses ensemble learning method for regression. Ensemble learning method is a technique that combines predictions from multiple machine learning algorithms to make a more accurate prediction than a single model.

```
rfreg = RandomForestRegressor(random_state=0)
rfreg.fit(X_train, y_train)
y_pred = rfreg.predict(X_test)
rf_mae = mean_absolute_error(y_test, y_pred)
text.insert(END,"MAE value for Decisiontree: "+str(rf_mae)+"\n")
```

Fig. 9. Random forest regression

6) Gradient boosting regressor: Gradient boosting is one of the most popular machine learning algorithms for tabular datasets. It is powerful enough to find any nonlinear relationship between your model target and features and has great usability that can deal with missing values, outliers, and high cardinality categorical values on your features without any special treatment.

```
gbreg = GradientBoostingRegressor(random_state=0)
gbreg.fit(X_train, y_train)
y_pred = gbreg.predict(X_test)
gb_mae = mean_absolute_error(y_test, y_pred)
text.insert(END,"MAE value for GradientBoosting: "+str(gb_mae)+"\n")
```

Fig. 10. Gradient boosting regression

VI. Results and Analysis

The main goal of this article is to predict the rainfall using the machine learning linear models. We evaluated seven linear machine learning algorithms, they are Lasso regression, Ridge ression, Lassolar regression, Elasticnet Regression, Decision tree regressor, Random forest regressor, and Gradient boosting regressor.

A. Interface

We developed an interface using Tkinter and binded the best model for the prediction of rainfall using the trained machine learning models. The main page or home page of the interface is as below.

Next to upload the dataset:

Then, reading the data from the file and generating the graph for data

Then, data analysis yearly for every state and also quarterly from the given data.

Then, we executed the linear models for the cleaned data.

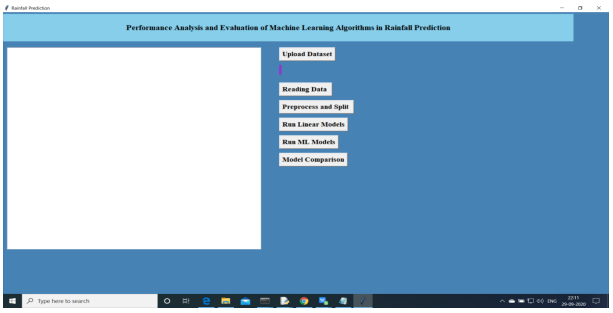

Fig. 11. Home Page in the Interface

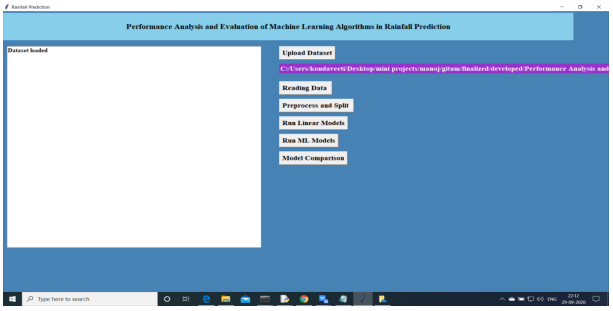

Fig. 12. Load the dataset

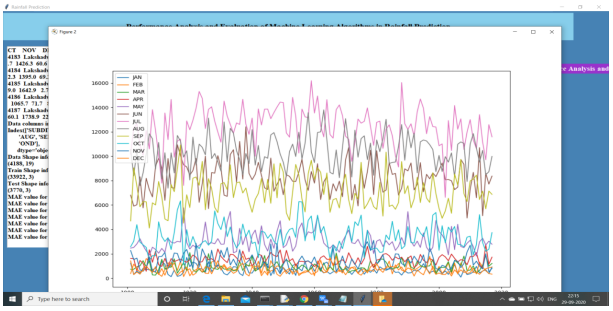

Fig. 13. Reading dataset

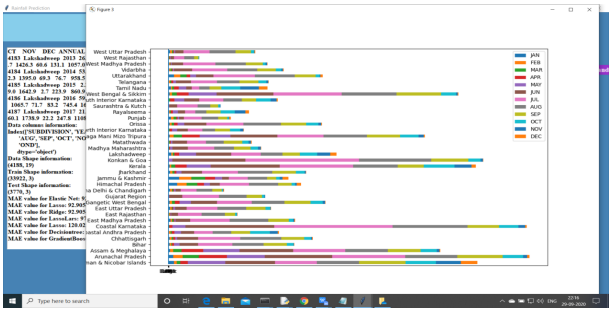

Fig. 14. yearly data analysis state wise

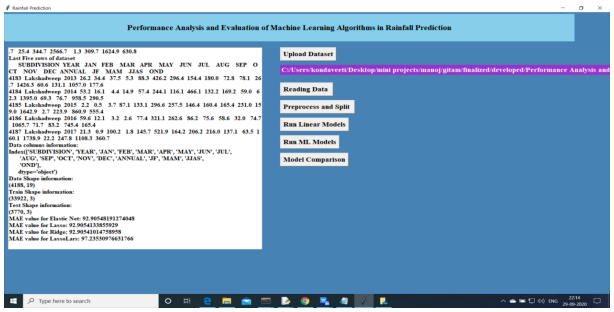

Fig. 15. Linear models

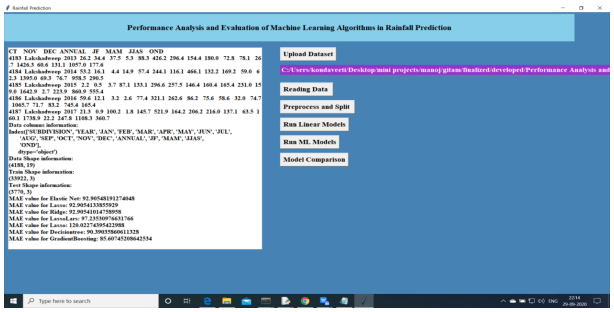

Fig. 16. ML models

CONCLUSION

Many ML algorithms have been successfully applied for the automatic regression of rainfall. This research paper summarizes and exemplifies the working logic of the six ML algorithms and empirically evaluates the regression performance of all the ML algorithms to the benchmark rainfall dataset. Among the six algorithms, lasso regression got the highest R2 score of 99.21 at 80-20 of training and validation dataset. Apart from this, the performance of all ML algorithms is evaluated and compared to the actual target values with predicted values.

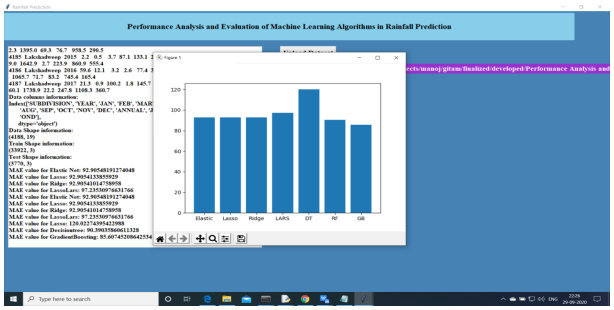

Fig. 17. Model comparison

REFERENCES

[1] N. Gnana Sankaran, E. Ramaraj, "A Multiple Linear Regression Model to Predict Rainfall Using Indian Meteorological Data", International Journal of Advanced Science and Technology (IJAST) Vol. 29, No. 8s, (2020), pp. 746-758.

[2] Neville Nicholls. Atmospheric and Climatic Hazards: Improved Monitoring and Prediction for Disaster Mitigation. Natural Hazards, 23(2-3):137–155

[3] Puneet Sharma and Nadim Chishty, "Machine Learning-Based Modelling of Human Panther Interactions in Aravalli Hills of Southern Rajasthan", Indian Journal of Ecology 46(1): 126-131.

[4] A.El-shafie, M.Mukhlisin, Ali A. Najah and M.R. Taha, " Performance of artificial neural network and regression techniques for rainfall-runoff prediction", International Journal of the Physical Science vol

[5] N. K. A. Appiah-Badu, Y. M. Missah, L. K. Amekudzi, N. Ussiph, T. Frimpong and E. Ahene, "Rainfall Prediction Using Machine Learning Algorithms for the Various Ecological Zones of Ghana," in IEEE Access Journal in 2022, vol. 10, pp. 5069-5082, 2022, doi: 10.1109/ACCESS.2021.3139312.

[6] Ms Ashwini Mandale, Mrs Jadhawar B.A, "Weather Forecast Prediction: A ata Mining Application", International Journal of Engineering Research and General Science Volume 3, Issue 2, March, April 2015, ISSN 2091-2730.

[7] J.N.K. Liu, B. N. L. Li, and T. S. Dillon. An improved naive Bayesian classifier technique coupled with a novel input solution method [rainfall prediction]. Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions, (2):249 –256, 2001.

[8] https://data.gov.in/resources/sub-divisional-monthly-rainfall-1901-2017.

[9] Shen Rong, Zhang Bao-wen, The research of regression model in machine learning field MATEC Web of Conferences 176, 01033 (2018)

[10] S. Chowdhury and M. P. Schoen, "Research Paper Classification using Supervised Machine Learning Techniques," 2020 Intermountain Engineering, Technology and Computing (IETC), Orem, UT, USA, 2020, pp. 1-6, doi: 10.1109/IETC47856.2020.9249211.

[11] Ahmad M., Aftab S., Salman M., Hameed N., Ali I., Nawaz Z. SVM Optimization for Sentiment Analysis. Int. J. Adv. Comput. Sci. Appl. 2018;9:393–398. doi: 14569/IJACSA.2018.090455. [CrossRef] [Google Scholar]

[12] Ahmad M., Aftab S., Salman M., Hameed N. Sentiment Analysis using SVM: A Systematic Literature Review. Int. J. Adv. Comput. Sci. Appl. 2018;9:182–188. doi: 10.14569/IJACSA.2018.090226. [CrossRef] [Google Scholar].

[13] Ahmad M., Aftab S., Ali I. Sentiment Analysis of Tweets using SVM. Int. J. Comput. Appl. 2017;177:25–29. doi: 10.5120/ijca2017915758. [CrossRef] [Google Scholar]

[14] Ahmad M., Aftab S. Analyzing the Performance of SVM for Polarity Detection with Different Datasets. Int. J. Mod. Educ. Comput. Sci. 2017;9:29–36. doi: 10.5815/ijmecs.2017.10.04. [CrossRef] [Google Scholar]

[15] Ahmad M., Aftab S., Muhammad S.S. Machine Learning Techniques for Sentiment Analysis: A Review. Int. J. Multidiscip. Sci. Eng. 2017;8:27. [Google Scholar]

[16] Available online: http://ru8.rp5.ru/

[17] Sivapragasam C., Liong S.-Y., Pasha M.F.K. Rainfall and runoff forecasting with SSA–SVM approach. J. Hydroinformatics. 2001;3:141–152. doi: 10.2166/hydro.2001.0014. [CrossRef] [Google Scholar]

[18] Isa D., Hong L.L., Kallimani V.P., Rajkumar R. Text Document Pre-Processing Using the Bayes Formula for Classification Based on the Vector Space Model. Comput. Inf. Sci. 2008;20:79–90. doi: 10.5539/cis.v1n4p79. [CrossRef] [Google Scholar]

[19] Sawale G.J., Gupta S.R. Use of Artificial Neural Network in Data Mining For Weather Forecasting. Int. J. Comput. Sci. Appl. 2013;6:383–387. [Google Scholar]

[20] Abhishek K., Kumar A., Ranjan R., Kumar S. A rainfall prediction model using artificial neural network; Proceedings of the 2012 IEEE Control and System Graduate Research Colloquium, ICSGRC 2012, no. Icsgrc; Selangor, Malaysia. 16–17 July 2012; pp. 82–87. [Google Scholar]