Comparative Study of Rainfall Prediction Modeling Techniques: Case Study on Karapur, India

Sushma Iliger, Mario Pinto

Goa College of Engineering, Farmagudi, Goa, India

Corresponding author: Sushma Iliger, Email: siliger98@yahoo.in

India's economy is highly dependent on agriculture which is in turn dependent on rainfall. Every region receives varying amounts of rainfall and based on that the crops are grown to suit the geographical conditions. Due to the lack of proper predictive technology in making use of rainfall, the agricultural practices have put high pressure on the underground water tables leading to depletion of water resources in local areas. Considering these factors, we have tested the suitability of the different existing machine learning model for the data collected from Karapur weather station of Goa state, India to predict rainfall locally. The different models being used are Multiple Linear regression (MLR), Decision tree Regressor, Random forest Regressor, XGBoost and Artificial Neural Network algorithm(ANN). After analysis we obtained 85% accuracy and the algorithm performance compared in terms of Mean absolute error and R2 Score. . The result of the study revealed that ANN outperformed the others by delivering an average R2 score close to 1.

Keywords: Machine learning algorithms, Rainfall prediction, Multiple Linear Regression, Random forest regressor, Artificial neural network , Xgboost.

2023. In Saroj Hiranwal & Garima Mathur (eds.), *Artificial Intelligence and Communication Technologies*, 159–166. Computing & Intelligent Systems, SCRS, India. https://doi.org/10.52458/978-81-955020-5-9-16

1 Introduction

Rainfall is a climatic phenomenon that results due to the interactions of various environmental cycles. One of the most difficult issues in constructing rainfall prediction models is the high uncertainty involved in identifying the role of different atmospheric variables. Temperatures, evaporation rate, relative humidity, wind speed, wind direction, and other attributes all have a role in deciding the frequency of rain. However, rainfall forecasts are critical information for agriculture practices since it allows the farmers of the region to make use of natural resources rather than over exploiting the underground water tables.

In this experiment the features used in our studies are not commonly used for prediction purposes in previous work. We have carried out a comparative study of existing regression techniques such as MLR, Decision tree regression, Random Forest Regression, XGBoost and ANN on the basis of accuracy of prediction. The objective of the study is to predict rainfall in the region surrounding Karapur weather station, Goa, India and find the best model for the given set of features in the dataset. The methodologies followed in this study are:

- Data Collection
- Data Pre-Processing
- Model Development
- Performance Measure

The data records are pre-processed using filling Nans, elimination of outliers and Normalisation. R-square and Mean Absolute Error values are used as evaluation metrics to analyze the model performance.

2 Related Work

Mohapatra et al. [1] investigated the data mining techniques to create a rainfall prediction system for the historical meteorological data of 100 years of Bangalore, India. For linear regression technique there used a fixed sampling size for testing and training phase but in K fold technique, an ensemble technique, the sampling size of the data is random. It was proved that ensemble technique provides better results because the final output is the average of all the iterations. Mohini P et al. [2] discussed the survey of various artificial Neural network models used for prediction of rainfall in India as since in other researches that Statistical models such as ARIMA may not be suitable for long term weather prediction. They concluded that Feed forward, Recurrent and Time delay neural network models are suitable for yearly rainfall prediction but for daily and monthly data Neural networks delivered poor performance.

Geetha and Nasira [3] used Rapidminer to implement a model of Chi-Square Automatic Interaction Detector decision tree for predictions of different weather phenomenas. The historical data used for training and testing purposes was collected from Trivandrum weather station for two years. The model had delivered performance with accuracy of 85%. Surajit Chattopadhyay [4] discussed the superior performance of the ANN predictive model over persistence forecasting and MLR forecasting. They have used the historical data of 45 years to predict the average summer monsoon rainfall of India. Using 9 predictors and after 500 epochs ANN outperformed the other two models by giving 0.150 of prediction error.

Mekanik F et al. [5] discussed the performance of ANN and MR to predict long-term seasonal spring rainfall in Victoria, Australia. Both models were evaluated using statistical metrics such as mean square error, mean absolute error, Pearson correlation and Willmott index of agreement. MR. P. P. Sengar et al. [10] studied the breast cancer detection problem by carrying out early prediction by using the Wisconsin (Diagnostic) Data Set and feeding to 2 machine learning models including Logistic Regression and Decision Tree algorithm. The dataset consisted of 570 data entries of which 75% were used to train the models and 32 attributes. The decision tree classifier gave more accurate results with test accuracy 0.95.

Kesavulu Poola et al. [12] discussed the performance of XGBoost algorithm used to predict rainfall on monthly scales in areas of Visakhapatnam using historical data of 30 years. XGBoost uses the boosting technique where the previous level trees reduced the errors generated by older trees and the algorithm is highly dependent on the data accuracy. The results were consistent with an accuracy of 95%. Liyew et al. [13] discussed the performance of three machine learning techniques namely MLR, RF and XGBoost by applying them to predict rainfall daily, monthly and annually for the data obtained from the meteorological office at Bahir Dar City, Ethiopia. The evaluation metrics used root mean squared error and mean absolute error. XGBoost algorithm suited the best to carry out daily rainfall forecasting.

3 Background Study

The problem of rainfall prediction is recognised as a regression problem. Therefore in our work we have used existing regression algorithms such as Multiple Linear regression, Decision tree regressor, Random Forest Regressor, XGBoost and Artificial Neural Network.

(i) Multiple Linear Regression: It is a supervised machine learning algorithm used to predict numerical values using a set of predictors. A constant bias is applied to the summation of input feature values in order to perform prediction. Unlike simple regression which derives dependent variables from single predictors, MLR gives dependent variables from multiple predictors. Main objective of the algorithm is to produce a line that best fits the dataset and try to get minimum error.

(ii) Decision Tree Regressor: It is a supervised machine learning algorithm used to solve regression and classification problems. It uses a single decision tree as the predictive model where the dataset is broken into smaller subsets so that each internal node is labeled with a feature. The tree contains a single root node, branches and leaf nodes. Leaf nodes hold the decision to be made.

(iii) Random Forest Regressor: It is an ensemble machine learning algorithm which constructs random forests using a large number of decision trees that are randomly split during the training phase. The mean prediction value of individual decision trees is used for predictive modeling. The decision boundary gets stable and accurate as more single trees are added.

(iv) XGBoost: XGBoost is a scalable machine learning system for tree boosting. Features such as parallel and distributed computing and the procedure of handling sparse data makes learning faster. It allows faster model exploration than other learning algorithms. It uses regularization techniques to avoid overfitting by smoothening the final trained weights.

(v) Artificial Neural Network: The inspiration for development of this algorithm is the biological neurons of the human brain. The neural network consists of several layers consisting of artificial neurons called nodes. The input to the nodes is the summation of weights received from the previous layer which is combined with the activation function that dominates the orientation of the output. There are different kinds of activation functions such as linear,

exponential, binary etc. The accuracy of the algorithm is tested by calculating the error which is expected to be minimum.

In order to assess the different models two evaluation metrics were used Mean Absolute Error and R2 score. R2 score value varies between 0 to 1 and can also be negative. However a model giving value closer to 1 is considered a good model for prediction. Meanwhile MAE is the error calculated by finding the difference between actual and obtained output resulted by the model.

4 Methodology

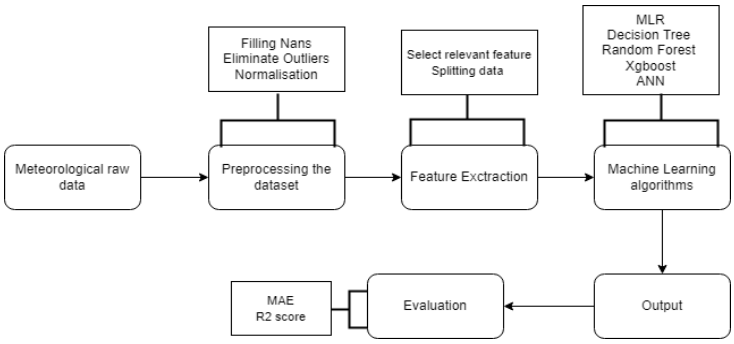

Fig. 1. Overall Architecture of Machine learning algorithm modeling

The problem of rainfall prediction is recognized as a regression problem. Therefore in our work we have used existing regression algorithms such as Multiple Linear regression, Decision tree regressor, Random Forest regressor, XGBoost and Artificial Neural Network. Figure 1 shows the overall flow of the experiment.

A. Data Collection

Table 1. Literature survey of different approaches

Algorithm	Type	Description
Temperature Dry Bulb	Numerical	Degree Celsius
Temperature Wet Bulb	Numerical	Degree Celsius
Relative Humidity	Numerical	Percentage
Instant Wind Speed	Numerical	Kilometers per hour
Average Wind Speed	Numerical	Degree Celsius
Pan Evaporation	Numerical	Degree Celsius
Temperature Pan Water	Numerical	Degree Celsius
Precipitation Value	Numerical	Millimeters

For our study, the data were collected from the regional weather station at Karapur in the state of Goa, India. Eight attributes such as Temperature Dry Bulb, Temperature Wet Bulb, Relative Humidity, Instant Wind Speed, Average Wind Speed, Pan Evaporation, Temperature Pan Water, Precipitation value were included. The data were recorded in the Microsoft Excel tabular format. The data ranging

over 4 years (2010–2014) was used for the study. Table 1 shows the selected attributes in the dataset and their types and description.

B. Data Preprocessing

In India many regions receive heavy rainfall during the months of June, July, August, September, October and affect the agricultural practices of the farmers. Therefore techniques used for data preprocessing for training and testing phase:

(i) Filling NANs: Since the data is gathered from the year 2010 to 2014 it has some missing data in different columns. Missing data values were replaced by the average value of the respective month in the individual column.

(ii) Eliminate Outliers: Local data is highly affected by global climatic changes. The presence of numerical values that are anomaly in nature can give inaccurate trained models. Therefore we decided to drop those rows.

(iii) Normalisation: As our data contains several irregularities, and the rainfall values have large variation from January to December. So it was decided for normalising the dataset which resulted in better results.

C. Model Development

All the machine learning algorithms were implemented and analysed using the Scikit Learn package in Python. The overall dataset was divided into 70% for training data and 30% for testing data for all the algorithms. The number of hidden layers in the ANN model was set to 3 and Sigmoid was chosen as the activation function. The epoch was set 1000 by trial and error.

5 Experimental Results

This section deals with the results obtained from after training and testing data with the 5 machine learning algorithms for rainfall prediction problems. The total number of data in the training and testing data set is maintained the same for testing all the algorithms. To evaluate the algorithm we have applied evaluation metrics such as mean absolute error and r2 score. The comparison is shown in Table 2.

Table 2. Results obtained from Machine Learning Algorithms

Algorithm	Mean Absolute Error	R2 Score
Multiple Linear Regression	13.50	0.18
Decision Tree Regression	8.42	0.19
Random Tree Regression	8.11	0.34
XGBoost	8.56	0.39
Artificial Neural Network	1.41	0.84

Figures 2-6 depict the graph of tested data for each algorithm. The actual rainfall value was compared with the predicted rainfall value on each day. The size of the testing data remained the same for the algorithms.

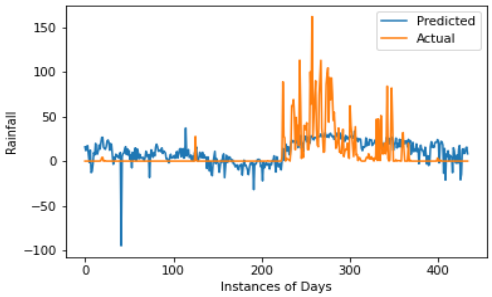

Fig. 2. Precipitation value predicted using Multiple Linear Regression

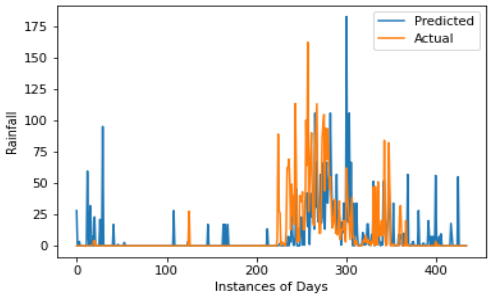

Fig. 3. Precipitation value predicted using Decision Tree Regression

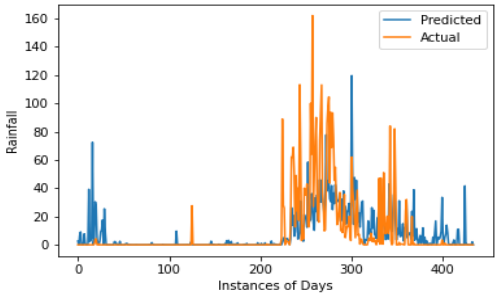

Fig. 4. Precipitation value predicted using Random Tree Regression

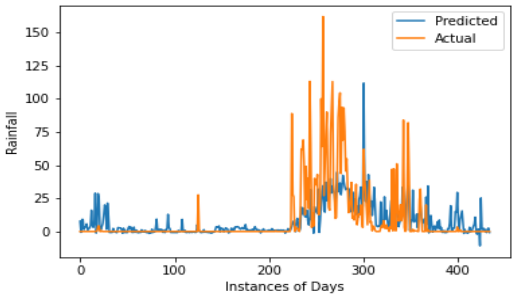

Fig. 5. Precipitation value predicted using XGBoost

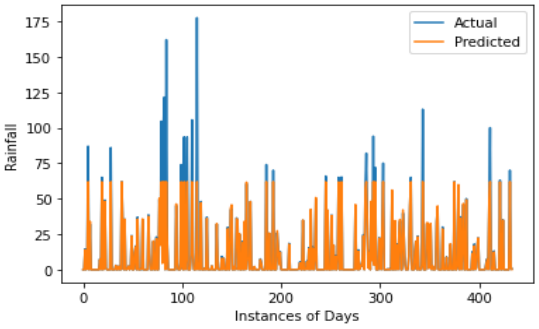

Fig. 6. Precipitation value predicted using Artificial Neural Network algorithms

6 Conclusion

In this study we have analyzed various state of the art machine learning algorithms which includes Multiple linear regression, Decision Tree regressor, Random forest regressor, XGBoost, Neural Networks and evaluated them using two metrics i.e. MAE and r2 score. Multiple linear Regression and decision tree regressor models turned out to be the least accurate model showing high variance. However, Random Forest and XGBoost algorithms showed significantly better performance and both obtained similar results for the given data. Artificial neural network model outperformed all the five machined learning algorithms by delivering a very high R2 score which indicated that the model is best suited for the given data. This study of prediction using machine learning algorithms will vary for other geographical regions however it helps the farmers to perform their agricultural practices by maximum utilisation of environmental resources thereby reducing the pressure on underground water table.

References

[1] Sandeep Kumar Mohapatra, Anamika Upadhyay, Channabasava Gola, "Rainfall Prediction based on 100 years of Meterological Data",2017 International Conference on Computing and Communication Technologies for smart Nation, pp.162 – 166.

[2] Mohini P. Darji, Vipul K. Dabhi, Harshadkumar B.Prajapati, "Rainfall Forecasting Using Neural Network: A Survey", 2015 International Conference on Advances in Computer Engineering and Applications (ICACEA) IMS Engineering College, Ghaziabad, India, pp.706 – 713

[3] Geetha, A. and G.M. Nasira, 2014, data mining for meteorological applications: Decision trees for modeling rainfall prediction, Proceedings of the IEEE International Conference on Computational Intelligence and Computing Research, Dec. 18-20, IEEE Xplore press, India. DOI: 10.1109/ICCIC.2014.7238481

[4] Surajit Chattopadhyay (2007). Feed forward Artificial Neural Network model to predict the average summer-monsoon rainfall in India. , 55(3), 369–382. doi:10.2478/s11600-007-0020-8

[5] Mekanik, F.; Imteaz, M.A.; Gato-Trinidad, S.; Elmahdi, A. (2013). Multiple regression and Artificial Neural Network for long-term rainfall forecasting using large scale climate modes. Journal of Hydrology, 503(), 11–21. doi:10.1016/j.jhydrol.2013.08.035

[6] Jui J.J., Imran Molla M.M., Bari B.S., Rashid M., Hasan M.J. (2020) Flat Price Prediction Using Linear and Random Forest Regression Based on Machine Learning Techniques. In: Mohd Razman M., Mat Jizat J., Mat Yahya N., Myung H., Zainal Abidin A., Abdul Karim M. (eds) Embracing Industry 4.0. Lecture Notes in Electrical Engineering, vol 678. Springer, Singapore. https://doi.org/10.1007/978-981-15-6025-5_19

[7] Kumar Abhishek, Abhay Kumar, Rajeev Ranjan, and Sarthak Kumar. A rainfall prediction model using artificial neural network. In 2012 IEEE Control and System Graduate Research Colloquium, pages 82–87. IEEE, 2012.

[8] Valmik B Nikam and BB Meshram. Modeling rainfall prediction using data mining method: A bayesian approach. In 2013 Fifth International Conference on Computational Intelligence, Modelling and Simulation, pages 132–136. IEEE, 2013.

[9] I. Salehin, I. M. Talha, N. N. Moon, M. Saifuzzaman, F. N. Nur and M. Akter, "Predicting the Depression Level of Excessive Use of Mobile Phone: Decision Tree and Linear Regression Algorithm," 2020 International Conference on Sustainable Engineering and Creative Computing (ICSECC), 16-17 December 2020, President University, Indonesia., in press.

[10] P. P. Sengar, M. J. Gaikwad and A. S. Nagdive, "Comparative Study of Machine Learning Algorithms for Breast Cancer Prediction," 2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT), 2020, pp. 796-801, doi: 10.1109/ICSSIT48917.2020.9214267.

[11]S. K. J. and G. S., "Prediction of Heart Disease Using Machine Learning Algorithms.," 2019 1st International Conference on Innovations in Information and Communication Technology (ICIICT), 2019, pp. 1-5, doi: 10.1109/ICIICT1.2019.8741465.

[12] Kesavulu Poola,P Hema Sekhar; Prediction of rainfall by using extreme gradient boost (XG boost) in Vishakapattanam area, Andhra Pradesh;International Journal of Statistics and Applied Mathematics 2021; 6(3): 83-86

[13] Liyew, C.M., Melese, H.A. Machine learning techniques to predict daily rainfall amount. J Big Data 8, 153 (2021). https://doi.org/10.1186/s40537-021-00545-4