Machine Learning Techniques to Predict Daily Rainfall Amount

Chalachew Muluken Liyew (chalachewsweet@gmail.com)
Bahir Dar University Institute of Technology https://orcid.org/0000-0003-4031-8032
Haileyesus Amsaya Melese
Bahir Dar University Institute of Technology

Research

Keywords: Machine learning, MLR, RF, XGBoost, Rainfall Prediction

Posted Date: September 20th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-801241/v1

Version of Record: A version of this preprint was published at Journal of Big Data on December 1st, 2021.
See the published version at https://doi.org/10.1186/s40537-021-00545-4.

Abstract

It is crucial to predict the amount of daily rainfall to improve agricultural productivities to secure food, and water quality supply to keep the citizen healthy. To predict rainfall, various researches are conducted using data mining and machine learning techniques of different countries' environmental datasets. The Pearson correlation technique is used to select relevant environmental variables which are used as an input for the machine learning model of this study. 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. The dataset is collected from the local meteorological office to measure the performance of three machine learning techniques as Multivariate Linear Regression, Random Forest and Extreme Gradient Boost. Root mean squared error and Mean absolute Error are used to measure the performance of the machine learning model for this study. The result of the study shows that the Extreme Gradient Boost gradient descent machine learning algorithm performs better than others.

1. Introduction

Rainfall prediction is crucial for increasing agricultural productivities to secure food and quality water supply for the citizen. Rainfall shortage has negative influence on the aquatic ecosystem, quality water supply and agricultural productivities. The agriculture and water quality depend on the rainfall water (Namitha et al., 2015 ; Andrew et al., 2013 ; Chowdari et al., 2015) amount in daily and annual bases. Therefore, accurate prediction of daily rainfall is a challenging task to manage the rainfall water for agriculture and water supply.

Various researchers conduct studies to improve the prediction of daily, monthly and annual rainfall amount using different countries meteorology data. Researchers applied data mining techniques (Andrew et al., 2013; Chowdari et al., 2015; Zainudin et al., 2016; Tharun et al., 2018), Big Data analysis (Namitha et al., 2015; Manandhar et al., 2019) and different machine learning algorithms (Vijayan et al., 2020; Zeelan et al., 2020; Arnav Garg and Kanchipuram, 2019; Aswin et al., 2018) to improve the accuracy of daily, monthly and annual rainfall prediction. According to the results of the studies, the prediction process is now shifted from the data mining techniques to the machine learning techniques. Namitha et al. (2015) confirmed to predict weather, machine learning algorithms are proved to be better replacing the traditional deterministic method so that the machine learning method performs better than the traditional data mining techniques to predict rainfall. This paper analyses different machine learning algorithms to identify the better machine learning algorithms for accurate rainfall prediction.

There are several environmental factors that 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 research conducted by Chaudhari M.M. & Choudhari D.N. (2017) studies important features of the atmosphere lie temperature, wind and cyclone over the Indian region to predict rainfall, however, the study do not measure the correlation of each features to determine the strength of the independent features on the rainfall. The studies Thirumalai et

al. (2017) conduct the correlation study to identify the most important features like solar radiation, perceptible water vapor and diurnal features stand out for rainfall prediction using a linear regression model. And (Zeelan et al., 2020 ; Vijayan et al., 2020; Gnanasankaran and Ramaraj, 2020) use atmospheric features of temperature, relative humidity, pressure and wind speed as an important features to predict rainfall accurately using machine learning such as Artificial Neural Network, Random forest and multiple linear regression model respectively. Hence, important atmospheric features that have direct or indirect impact for rainfall should be studied to predict the existence and the intensity of rainfall.

Therefore, 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. To achieve this objective, related literatures are reviewed, data sets are collected from meteorology stations, experiments are conducted and result of the experiments are analyzed and finally concluding the work.

2. Related Work

The machine learning algorithm called linear regression is used for predicting the rainfall using important atmospheric features by describing the relationship between atmospheric variables that affect the rainfall (Thirumalai et al., 2017; Prabakaran et al., 2017). The correlation study is conducted (Manandhar et al 2019), and identified solar radiation, perceptible water vapor and diurnal features are important variables for daily rainfall prediction using data driven machine learning algorithm. The future work identified by Manandhar et al. (2019) is studying the impact of using different atmospheric features using a larger data set. The researches address the relationship between independent and dependent features to identify which features impact the rainfall to rain or not to rain. The intensity of the daily rainfall is not addressed.

Tharun et al. (2018) perform the accuracy measure of the comparative study of statistical modelling and regression techniques (SVM, RF and DT) for rainfall prediction using environmental features. According to the result of the study, the regression techniques of rainfall prediction outperforms the statistical modeling. The experimental result show that the RF model performs and predicts accurately than the SVM and DT. Hence, rainfall prediction is accurate and show high performance in machine learning models than the traditional models.

The study by Arnav Garg and Kanchipuram, (2019) shows three machine learning algorithm experiments such as support vector machine (SVM), support vector regression (SVR) and K-nearest neighbor (KNN) using the patterns of rainfall in the year. The SVM algorithm performs best among the three machine learning algorithms. This research did not show the experiment result that which environmental features impact the intensity of rainfall. This paper shows the environmental features that has positive and negative impact for rainfall and predict the daily rainfall amount using those features.

(Balan et al., 2019; Gnanasankaran and Ramaraj, 2020) confirm that the multiple linear regression machine learning algorithm outperforms well to predict rainfall using dependent weather variables of temperature, humidity, moisture, wind speed, and finally the study showed the performance of the rainfall

prediction can be improved using deep learning models as a future work. The researchers (Zeelan et al., 2020 ; Aswin et al., 2018) study the deep learning algorithm for the rainfall prediction by using different dependent weather variables. To provide accurate prediction of rainfall, prediction models have been developed and experimented using machine learning techniques.

Therefore, this paper selects the appropriate environmental features that correlate with rainfall positively or negatively to examine the performance of the daily rainfall amount prediction machine learning algorithms using MAE and RMSE.

3. Machine Learning Algoriths

To choose the better machine learning algorithms to study the daily rainfall amount prediction, various papers have been reviewed concerning the rainfall prediction. To predict the daily rainfall intensity using the real time environmental data, three algorithms such as MLP, RF and XGBoost gradient decent are chosen for experiment. Hence, the three machine learning algorithms are experimented and compared to report the better algorithms to predict the daily rainfall amount.

A. Multivariate Linear Regression (MLR)

Linear regression can be multivariate which has multiple independent variables used as an input features and simple linear regression which has only one independent or input feature. Both linear regressions have one dependent variable which can be forecasted or predicted based on the input features. This paper presents the multivariate linear regression because multiple environmental variables or features are used to predict the dependent variable called daily rainfall amount. The linear regression is a supervised machine learning technique used to predict the unknown daily rainfall amount using the known environmental variables. The multivariate linear regression uses multiple explanatory or independent variable (X) and single dependent or output variable denoted by Y. Hence, the general equation of the multiple linear regression is given as:

$$Y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \cdots + \beta_p x_{ip} + \varepsilon_i = x_i^T \beta + \varepsilon_i \qquad \text{i= 1, 2, 3 ... n}$$

Where x_i^T is transpose of x_i the input or independent variable, β is regression coefficient, ε_i is error term or noise, Y_i is a dependent variable.

The general multivariate linear regression equation of this paper is given as

*Daily rainfall = (year * β_1) + (month * β_2) + (day * β_3) + (MaxTemp * β_4) + (MinTemp * β_5) + (Humidity * β_6) + (Evaporation * β_7) + (sunshine* β_8) + (windspeed * β_9) +*

B. Random Forest (RF)

A Random Forest Regression model is powerful and accurate. It usually performs great on many problems, including features with non-linear relationships. Random forest regression is a supervised machine learning algorithm that uses ensemble learning method for regression. A Random Forest works by building several decision trees during training time and outputting the mean of the classes as the prediction of all the trees. The Random Forest algorithm works on the following steps:

a. Take at random p data points from the training set
b. Build a decision tree associated to these p data points
c. Take the number N of trees to build and repeat a and b steps
d. For a new data point, make each one of the N tree trees predict the value of y for the data point and assign the new data point to the average all of the predicted y values.

Random forest algorithm is one of the supervised machine learning algorithm that is selected as the predictive model for daily rainfall prediction using environmental input variables or features. Random forest regression is operated by constructing a multitude of decision tree at the training time and outputting the class that is the mode of mean prediction or regression of the individual trees. According to (Andrew Kusiak, Anoop Prakash Verma and Evan Roz, 2013) the RF algorithm is efficient for large datasets and good experimental result is obtained using large datasets having a large proportion of the data is missing.

C. XGBoost Gradient Descent

XGBoost stands for e**X**treme **G**radient **B**oosting; it is a specific implementation of the Gradient Boosting method which uses more accurate approximations to find the best tree model. XGBoost is implemented for the supervised machine learning problem that has data with multiple features of x_i to predict a target variable y_i. Most authors use XGBoost for different regression and classification problems due to the speed and prediction accuracy of the algorithm.

Extreme Gradient Boosting (XGBoost) is one of the efficient (Srinivas et al., 2020) algorithm in the gradient descant that has linear model algorithm and tree learning algorithm. It is faster than other gradient descent algorithms because of the parallel computation on a single machine. This paper chooses XGBoosting algorithm for experiments to predict the target variable daily rainfall intensity using various input or dependent environmental variables. XGBoost is a powerful algorithm which is fast learning through parallel and distributed computing and offers efficient memory usage that produces robust solution.

4. Methodology

4.1. DATA COLLECTION

To study and analyze the selected machine learning algorithms, row data is collected from the regional meteorological station at bahir dar city, Ethiopia. The data is collected form the meteorology station includes 10 features which are year, month, date, evaporation, sunshine, max-temp, min-temp, humidity, wind speed, rainfall.

The data is collected from Bahir Dar city Meteorology station for a period of 20 years from 1999–2018. Since the data is row data which is collected from the measuring devices in the station, the data contains missing values and it is not arranged in the appropriate format for the experiment.

4.2. DATA PREPROCESSING

The data preprocessing step includes the data conversion, manage missing values, categorical encoding and splitting dataset for training and testing dataset. A total of 20 years data is collected from the metrology office. Data conversion is the process of converting the data into the appropriate data format for the experiment. The row data is collected from meteorology station arranged in year based and the attributes in rows that need to combine and rearrange the features in column. Finally the data is converted from excel data to CSV data.

The dataset collected at the meteorology office is with missing values so that the missing value of the target variable is removed and the other features are filled using mean of the data. Encoding the dataset is performed and then the dataset is prepared for experiment. The important features for rainfall prediction are selected and the dataset is splitting as 80% for training and 20% for testing considered as an input for the model.

4.3. MODEL

In this paper, the rainfall is predicted using machine learning technique. Three machine learning algorithms such as Multivariate Linear Regression (MLR), Random Forest (RF) and gradient descent XGBoost are analyzed which took input variables having moderately and strongly related environmental variables with rainfall. The better machine learning algorithm is identified and reported based on the performance measure using RMSE and MAE.

4.4 MEASURING PERFORMANCE

Pearson correlation is used to measure the strength of the relationship between two variables. The two variables can be positively or negatively correlated and no relationship between two variables if the Pearson correlation coefficient is zero. The Pearson correlation coefficient model is mathematically described as:

$$r_{xy} = \frac{\sum_{i=n}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\ \sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}}$$

Where r_{xy} is the Pearson correlation coefficient, $\{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$ are paired data consisting of n pairs and are mean of x and y respectively.

Various researchers interpret the Pearson coefficient values differently as the relation is weak, moderate and strong positively or negatively, and two variables are not related. To show the relevant features of the environmental variables to predict daily rainfall intensity, the following Pearson coefficient ranges and interpretations are used as shown in table 1.

Table 1 Pearson coefficient ranges and Interpretations

Pearson coefficient |r|	Interpretation
0.00 < 0.10	negligible
0.10 < 0.20	Weak
0.20 < 0.40	Moderate
0.40 < 0.60	Relatively Strong
0.60 < 0.80	strong
0.80 < 1.00	Very strong

The machine learning algorithms take the input data features which are selected using the Pearson correlation coefficient as a relevant features.

The rainfall prediction performance of each machine learning algorithms that are used in this study should be measured using RMSE and MAE to compare which machine learning algorithm outperforms better than others. Root mean squared error (RMSE) and Mean Absolute Error (MAE) are two of the most common metrics used to measure accuracy for continuous variables. The MAE measures the average magnitude of the errors in a set of forecasts and the corresponding observation, without considering their direction.

$$MAE = \frac{1}{n}\sum_{j=1}^{n} |y_j - \hat{y}_j|$$

The RMSE is a quadratic scoring rule which measures the average magnitude of the error. It's the square root of the average of squared differences between prediction and actual observation.

$$RMSE = \sqrt{\frac{1}{n}\sum_{j=1}^{n} (y_j - \hat{y}_j)^2}$$

RMSE gives a relatively high weight to large errors. This means the RMSE is most useful when large errors are particularly undesirable. The MAE and the RMSE can be used together to diagnose the variation in the errors in a set of forecasts. The RMSE will always be larger or equal to the MAE; the greater difference between them, the greater the variance in the individual errors in the sample. If the RMSE = MAE, then all the errors are of the same magnitude.

5. Findings

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. The research findings are summarized here.

To choose the environmental variables that correlate with the rainfall, the Pearson correlation is analyzed on the environmental variables and the result of the correlation table is presented in Table 1. Since the dataset is large, the variables that correlates greater than 0.20 with rainfall are considered as the participant environmental features to the experiment for rainfall prediction. Hence, to predict the amount of daily rainfall, the environmental attributes relevant to daily rainfall prediction are Evaporation, Relative Humidity, Sunshine, Maximum Daily Temperature, and Minimum Daily Temperature as shown in Table 2.

Table 2
Environmental features and their Pearson coefficient value

Features	|r|
Year	0.012
Month	0.101
Day	0.017
Evaporation	0.279
Relative Humidity	0.401
Max daily Temperature	0.296
Min daily Temperature	0.204
Sunshine	0.351
Wind Speed	0.046
Daily rainfall	1.000

The Pearson Correlation coefficient experimental results on the given data show that the attributes such as year, month, day and wind speed have no significant impact on the prediction of rainfall. This paper took environmental values which have correlation coefficient greater than 0.2 and analyze the rainfall

prediction. The highly correlated environmental features for rainfall prediction is relative humidity which measure the Pearson coefficient of 0.401 and then daily sunshine of 0.351.

The machine learning model used the selected environmental features as an input for the algorithms. The regression models are implemented in python and the performance of the MLR, RF, and XGBoost are measured using MAE and RMSE.

Table 3
Performance Measurements

Algorithms	MAE	RMSE
Random Forest	4.49	8.82
MLR	4.97	8.61
XGBoost	3.58	7.85

As shown in Table 3, the comparison result of the three algorithms such as the MLR, RF, and XGBoost, the XGBoost Gradient descent outperforms than MLR and RF. The MAE and RMSE values of the XGBoost gradient descent algorithms are 3.58 and 7.85 respectively so that The XGBoost algorithm predicts the rainfall using relevant selected environmental features better than the RF and the MLR.

6. Conclusion

Rainfall Prediction is the application area of data science and machine learning to predict the state of the atmosphere. It is important to predict the rainfall intensity for effective use of water resources, crop productivity, and reduce mortality due to flood and any disease caused by rain. This paper analyses various machine learning algorithms for rainfall prediction. Three machine learning algorithms such as MLR, FR and XGBoost are presented and tested using the data collected from metrological station at Bahir Dar, Ethiopia.

The relevant environmental features for rainfall prediction are selected using Pearson correlation coefficient. The selected features are used as the input variables for the machine learning model used in this paper. The recent algorithms are analyzed in this work are MLR, RF and XGBoost and compared the results of the study. A comparison is made and showed the XGBoost is better suited machine learning algorithm for daily rainfall amount prediction using selected environmental features. The accuracy of the rainfall amount prediction my increase if the sensor data is incorporated for the study. The sensor data is not considered in this study.

The Rainfall prediction accuracy can be improved using sensor and meteorological datasets with additional different environmental features. Hence, as a future work, big data analysis can be used for rainfall prediction if the sensor and meteorological datasets are used for the daily rainfall amount prediction study.

Abbreviations

XGBoost- Extreme Gradient descent; MLR- Multivariate Linear Regression; RF – Random Forest RMSE- Root Mean Squared Error; MAE- Mean Absolute Error; SVM- Support Vector Machine DT- Decision Tree

Declarations

Competing Interest

The authors declare that they have no competing interests.

Ethics Approval and Consent to Participate

Not Applicable

Funding

There are no funding organization or individuals

Author's Contribution

CML designed, coordinated this research, draft the manuscript and conduct the experiment. CML and HAM carried out the data collection and data analysis. The authors read and approved the final manuscript.

Acknowledgments

We gratefully acknowledge North West of Ethiopia Meteorology agency for providing meteorological data, kind help and giving valuable information for the completion of this study.

Availability of Data and Material

The row data collected from the North West of Ethiopia Meteorology agency is available by researchers if it is requested and the materials that the authors used are available at the authors.

Consent for publication

Not Applicable

References

1. Andrew Kusiak, Anoop Prakash Verma and Evan Roz. (2013, April). Modeling and Prediction of Rainfall Using Radar Reflectivity Data: A Data-Mining Approach. *IEEE Transactions on Geoscience and Remote Sensing, 51*, 2337-2342.

2. Arnav Garg, and Kanchipuram Tamil Nadu. (2019). Rainfall Prediction Using Machine Learning. *International Journal of Innovative Science and Research Technology*, 56-58.
3. Aswin, S., Geetha, P., & Vinayakumar, R. (2018, April). Deep learning models for the prediction of rainfall. *In 2018 International Conference on Communication and Signal Processing (ICCSP)* (pp. 0657-0661). IEEE.
4. Balan, M. S., Selvan, J. P., Bisht, H. R., Gadgil, Y. A., Khaladkar, I. R., & Lomte, V. M. (2019). Rainfall Prediction using Deep Learning on Highly Non-Linear Data. *International Journal of Research in Engineering, Science and Management, 2*(3), 590-592.
5. Chaudhari, M. M., & Choudhari, D. N. (2017). Study of Various Rainfall Estimation & Prediction Techniques Using Data Mining. *American Journal of Engineering Research (AJER), 6*(7), 137-139.
6. Chowdari, K. K., Girisha, R., & Gouda, K. C. (2015). A study of rainfall over India using data mining. *In 2015 International Conference on Emerging Research in Electronics, Computer Science and Technology (ICERECT)* (pp. 44-47). IEEE.
7. CMAK Zeelan Basha, Nagulla Bhavana, Ponduru Bhavya, and Sowmya V. (2020). Rainfall Prediction Using Machine Learning & Deep Learning Techniques. *Proceedings of the International Conference on Electronics and Sustainable Communication Systems (ICESC 2020)* (pp. 92-97). Middlesex University: IEEE Xplore.
8. Manandhar, S., Dev, S., Lee, Y. H., Meng, Y. S., & Winkler, S. (2019). A data-driven approach for accurate rainfall prediction. *IEEE Transactions on Geoscience and Remote Sensing, 5*(11), 9323-9331.
9. N. Gnanasankaran, and E. Ramaraj. (2020). A Multiple Linear Regression Model To Predict Rainfall Using Indian Meteorological Data. *International Journal of Advanced Science and Technology, 29*(8), 746-758.
10. Namitha K, Jayapriya A and G Santhosh Kumar. (2015). Rainfall Prediction using Artificial Neural Network on Map-Reduce Framework. *ACM*, 492-495. doi:http://dx.doi.org/10.1145/2791405.2791468
11. Prabakaran, S., Kumar, P. N., & Tarun, P. S. M. (2017). Rainfall prediction using modified linear regression. *ARPN Journal of Engineering and Applied Sciences, 12*(12), 3715-3718.
12. R Vijayan, V Mareeswari, P Mohankumar, and G Gunasekaran K Srikar. (JUNE 2020). Estimating Rainfall Prediction using Machine Learning Techniques on a Dataset. *INTERNATIONAL JOURNAL OF SCIENTIFIC & TECHNOLOGY RESEARCH, 9*(06), 440-445.
13. Srinivas, A. S. T., Somula, R., Govinda, K., Saxena, A., & Reddy, P. A. (2020). Estimating rainfall using machine learning strategies based on weather radar data. *International Journal of Communication Systems, 33*(13), 1-11.
14. Tharun, V. P., Prakash, R., & Devi, S. R. (2018). Prediction of Rainfall Using Data Mining Techniques. *In 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT)* (pp. 1507-1512). IEEE Xplore.

15. Thirumalai, C., Harsha, K. S., Deepak, M. L., & Krishna, K. C. (2017, May). Heuristic prediction of rainfall using machine learning techniques. *In 2017 International Conference on Trends in Electronics and Informatics (ICEI)* (pp. 1114-1117). IEEE.
16. Zainudin, S., Jasim, D. S., & Bakar, A. A. (2016). Comparative analysis of data mining techniques for malaysian rainfall prediction. *International Journal on Advanced Science, Engineering and Information Technology, 6*(6), 1148-1153.

Figures

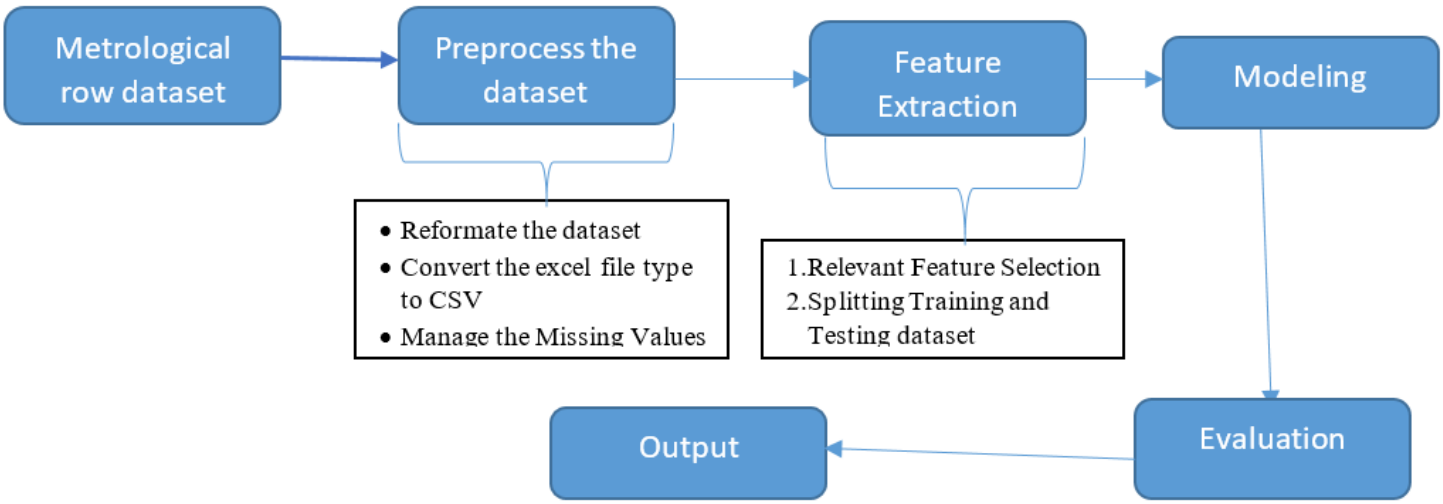

Figure 1

Machine Learning Model