IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Real Time Rainfall Prediction For Indian States Using XGBoost And Random Forest Approach

Varun Kumar P, Muduganti Manish Reddy, Tandra Sushanth, Raparthy Sandeep Kumar, Dr.T.Sathish Kumar

Student, Student, Student, Student, Associate Professor

Department of computer Science and Engineering,

Hyderabad Institute of Technology And Management, Hyderabad, India

Abstract: Predicting daily rainfall boosts agricultural output and ensures a consistent supply of food and water to maintain healthy populations. There have been a variety of studies to predict rainfall using data mining and machine learning approaches in various nations. The agriculture, which is the foundation of the nation's economy, is impacted by the country's uneven rainfall distribution. Planning and implementing the judicious use of rainfall water is necessary for the country in order to lessen the issue of the country's drought and flooding. This study's primary goal is to determine the pertinent atmospheric factors that contribute to precipitation and utilize artificial intelligence to forecast the daily amount of rain that will fall. The machine learning model's input variables were chosen using the Pearson correlation technique to be pertinent environmental factors. The root mean squared error and mean absolute error methodologies are used to assess the performance of the machine learning model. The Extreme Gradient Boosting machine learning technique outperforms competitors, according to the study's results. We accomplish this using the machine learning methods of Extreme Gradient Boost, Random Forest, and Multivariate Linear Regression.

Index Terms - Rainfall Prediction, Machine Learning, Multivariate Linear Regression, Extreme Gradient Boost, Random forest.

I. INTRODUCTION

The development of the fauna and flora of the natural world is significantly influenced by rainfall. Its significance must be taken into account by all living things, including humans, animals, and plants. Water is without a doubt one of the world's greatest natural resources and plays a key part in farming and agriculture. The planet Earth and people are finding it more challenging to experience the essential amount of rainfall needed to meet human demands and its uninterrupted use in daily life due to changing climatic conditions and rising greenhouse gas emissions.

Although rain is beneficial to agriculture, it can harm crops when it rains heavily. Flooding, which poses a threat to human life and causes infrastructure and building damage. Transport and communications are hampered by landslides, which pose a hazard to human life. The environment, business and industry, agriculture, and human activities can all be negatively impacted by an excessive amount of precipitation. One of the difficult aspects in the weather forecasting process is predicting the amount of rain. Due to the significant climate changes, it is now harder than ever to estimate rainfall accurately.

The ability to predict rainfall will aid not just in understanding the shifting patterns of precipitation but also planning for emergency preparedness and catastrophe management. The forecast of precipitation might be helpful in developing policies and methods to address the growing global problem of ozone depletion. Changes in Rainfall patterns and weather are connected to global warming, which is the rise in Earth's temperature brought on by increased emissions of chlorofluorocarbons from everyday items like refrigerators, air conditioners, deodorants, printers, etc.

In fact, the climate is being significantly impacted by the rising temperature. Similar to this, rainfall forecasts and weather updates assist in controlling micro-level issues like flood control and agricultural challenges brought on by little or excessive rainfall. By informing the public and tracking rainfall trends, the ability to predict rainfall could potentially improve the welfare and comfort of the populace.

Predictions of rains assist individuals in coping with the hot, humid weather. The space for innovation and revolution has increased as a result of technical advancement in the modern world. One must take into account the variety of options and possibilities that this technical progress has given up to people, even though the issues at hand are likely related to these technological developments. Another factor that poses a challenge in managing the water reserve is the incorrect or subpar rainfall forecast.

The study will be important for flood management agencies as well since a more exact and accurate forecast for high monsoon rains will keep the agencies ready and focused for an impending catastrophe, the destruction of which might be

reduced by adopting preventative steps. Water is a limited resource that must be conserved for the benefit of humans since it is a rare resource. The rainfall prediction will significantly aid in addressing this growing problem. Additionally, it will assist people in managing and scheduling their social activities appropriately.

II. LITERATURE SURVEY

The goal of Kunverji et al. was to create a working, flood-decisive prototype. The forecast model was created by the authors using DTs, RFs, and gradient boost algorithms, three supervised ML techniques. The Indian Water Portal for Bihar and Orissa provided the data set that was utilized in their research. With a 94.4% accuracy rate, it was found that DTs outperformed all other algorithms used. Temperature and rainfall intensity parameters have been incorporated in ML algorithms for flood prediction. KNNs, naïve Bayes, and SVMs—supervised learning techniques—were contrasted with deep learning models.

Therefore, to attain improved accuracy, a fresh data collection is necessary. The India Water Portal provided the set of information used in this study. Deep neural networks were found to perform better than the other algorithms, obtaining 91.18% accuracy. In their study, the scientists utilized both an SVM and a convolutional neural network (CNN), and they discovered that the CNN performed significantly better in terms of spatial resolution imaging while the SVM can make predictions based on linear data.

A robust flood map is also produced when an SVM and CNN are combined. The analysts confirmed that a probabilistic approach to examining the possibility of coastal flooding in the future would successfully aid in the creation of decisions for a coordinated beachside zone of the board . To forecast floods, they used a variety of machine learning algorithms on a set of flood data from southern Korea. The official website of the Korean Government served as the source of the data set for this research work. KNNs outperformed the other algorithms, according to the results, obtaining 94.6% accuracy.

The goal of Rani et al. was to create a reliable technique that can identify local floods and warn residents. Mean absolute error (MAE) and standard deviation are used to assess the efficacy of various flood detection techniques, including neural networks, SVMs, linear regression (LR), logistic regression, and linear regression (LR). The Indian Meteorological Department provided the data set that the researchers used. In the results, it can be shown that neural networks outperformed the other machine learning methods, reaching an MAE of 21.809 as opposed to the SVM's 90.606 and logistic regression's 40.246.

In a quicker region-based CNN-based multilayer perceptron (R-CNN) is suggested for detecting coastal rubbish.

If you want to get a generally faster R-CNN performance, it is challenging to synchronize a number of settings. To find small items, high-resolution features from a low-resolution image could be combined with high-dimensional ones. Automated coastal garbage identification could perform better if region of interest (RoI) align was used in place of RoI pooling to address position offset. Investigated for river-flood prediction accuracy were models using the radial basis function-firefly algorithm (RBF-FA) and support vector machine-firefly algorithm (SVM-FA) . RBF-FA and SVM-FA models were developed by combining an FA with an RBF and an SVM. The statistical measures used to analyze the error of the approaches show reduced root-mean-square error and higher R2 compared to normal SVM and RBF models that take into account all nodes. The assessment's findings also showed that the SVM-FA and SVM models fared better in predicting river floods than the RBF-FA and RBF models.

Finding the ideal supervised learning model configuration was the aim . According to the investigation, water levels at both nearby stations and in the control station's past are the most important prediction factors. It has been proven that rainfall amounts are a poor indicator of when floods may occur. This article offered a challenge because there was a lack of experimental data and unknown significant variables.

Using imperviousness maps and data from social media, emergency warning response management can identify crucial locations during pluvial flooding disasters. Eleven different flood model combinations were used along with the top model. The findings allow for the following inferences: In comparison to the other models, the LR model has a higher prediction rate (area under the receiver operating characteristic (AUROC) of 86.8%). Researchers used KNNs and extreme gradient boosting (XGB) supervised learning models to study flash flood-prediction mapping . The ROC area under the curve accuracy for the XGB and KNN algorithms was 90.2 and 80.7%, respectively, with XGBoost's capacity to offer more outputs permitting higher accuracy. It is hoped that employing different optimization strategies may improve the model's performance in next studies. In addition, it was discovered that the topographic wetness index (TWI) conditioning factors, slope, topography, and distance from the stream network were the parameters that had the most influence on the modelling processes.

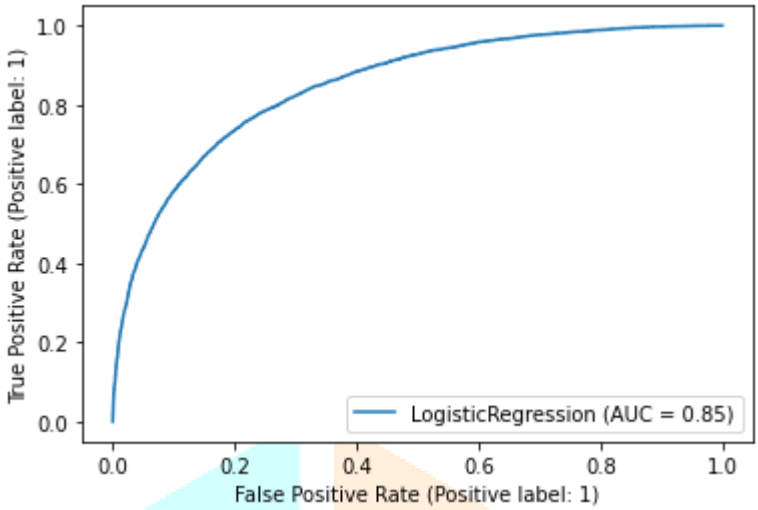
According to study by Tehrany et al. adding condition-based factors to the parameters obtained from lidar does not necessarily increase the accuracy of the results. Some academicians claim that height is a significant effect in flooding. This creates additional chances for enhancing and fortifying the Malaysian flood monitoring model. The computational components were efficiently combined to create the model. The device can automatically send out a warning message if the water level rises above a certain threshold. The tool can also be used to keep an eye on flash floods. But the system hasn't yet been put to the test in a real-world scenario in a flood-affected area. Future work should consider the following elements to be more effective: 1) Data transmission (such as wireless communication), 2) multipurpose messaging, and 3) the Android app. Given Malaysia's enormous developments in information and communication technology, the development of these three is highly anticipated. A technique that will outperform the probability mapping system and produce flood susceptibility was put forth by Mosavi et al. By utilizing CNN's social media data, emergency warning response management can employ imperviousness maps to pinpoint crucial regions during pluvial flooding calamities.

Working with high-dimensional images and items of interest results in a strong spatial structure and aids in other research' conclusions, which is advantageous. Nonlinear data can also be predicted using an SVM. These two distinct network designs when combined will produce a flood map that is both accurate and dependable.

III. EXISTING METHODS

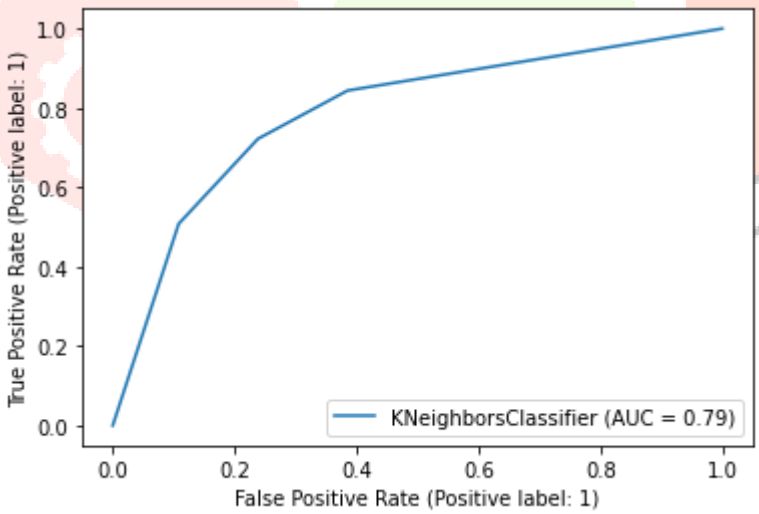
Logistic Regression :

A classification process known as logistic regression is employed to calculate the likelihood of an event occurring based on logistic function. When the dependent variable has only two possible values, such as 0 and 1 or True and False, it is described as a binary or dichotomous dependent variable. For each independent variable, the logistic regression model generates a coefficient that estimates the relative influence of the independent variable on the dependent variable.



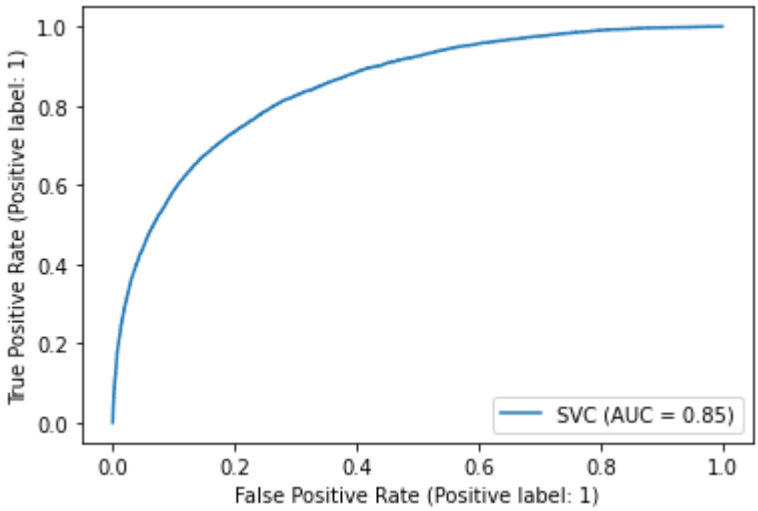
K -Nearest Neighbor :

Instance-based learning, often known as lazy learning, is the category in which the k-Nearest Neighbors (K-NN) method belongs. Although It is workable for regression, classification is where it is most frequently utilized. Many measures, consisting of Euclidean distance, Manhattan distance, Minkowski distance, etc., can be employed to establish the distance between the test and training data. The K-NN technique has a number of advantages, including being straightforward, simple to use, and easy to comprehend. It also offers good accuracy for smaller datasets. Yet it can be expensive to compute, especially for larger datasets., and it might not work well if the dataset is unbalanced.



Support Vector Machine (SVM) :

Machine learning uses the support vector machine (SVM) method as a classification tool. It is a binary linear classifier that distinguishes between the various classes of data by identifying the ideal border, or hyperplane. The SVM's goal is to maximize the distance between the nearest datapoints in each class and the hyperplane, which separates the two classes' data points. SVM is a potent algorithm that is frequently employed when high accuracy classification in complex datasets is required. The SVM technique has a number of benefits, including the capacity to handle big datasets, low-dimensional data, and non-linear data by applying kernel approaches.



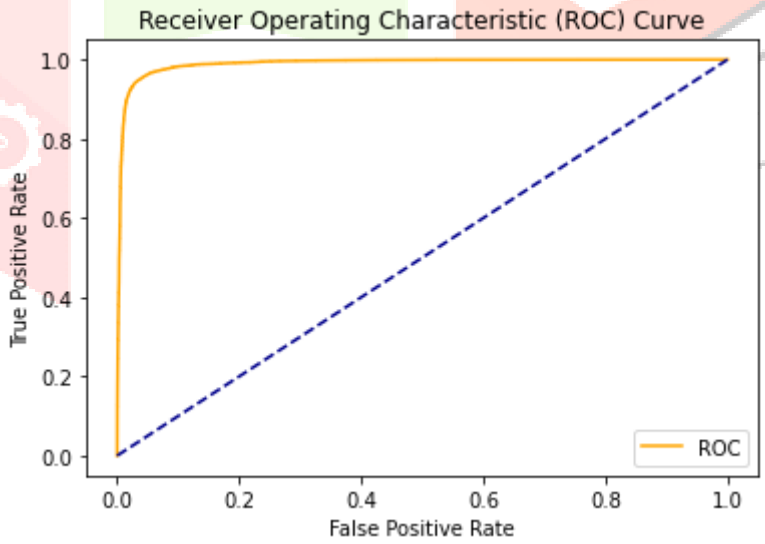
IV. PROPOSED METHODS

India's climate can be categorized into four categories. Winter is the first, lasting from December to February; summer, also known as pre-monsoon, is the second, lasting from March to May; monsoon, also known as rainy season, is the third, lasting from June to September; and the fourth is autumn, also known as post monsoon, lasting from October to November. The wettest months in India are June through September. Due to a rapid rise in rainfall, this time of year might result in flooding in some locations. Rainfall has the greatest impact on how often there are floods in any given place.

Extreme Gradient Boosting :

XGBoost is a supervised machine learning algorithm that belongs to the boosting algorithm family. It is an optimized version of the gradient boosting algorithm, whose goal is to make the model faster and more accurate by shortening the time required to calculate the gradient.

Due to its prowess in managing big datasets, high accuracy, and handling missing values, XGBoost has become quite well-liked in the data science field. Decision trees, which are weak learners, are systematically added to the model as part of the process, with each new tree being trained to rectify the mistakes produced by its predecessors. To put it another way, XGBoost creates a collection of decision trees, where each tree aims to identify the patterns in the data that were missed by the previous trees.

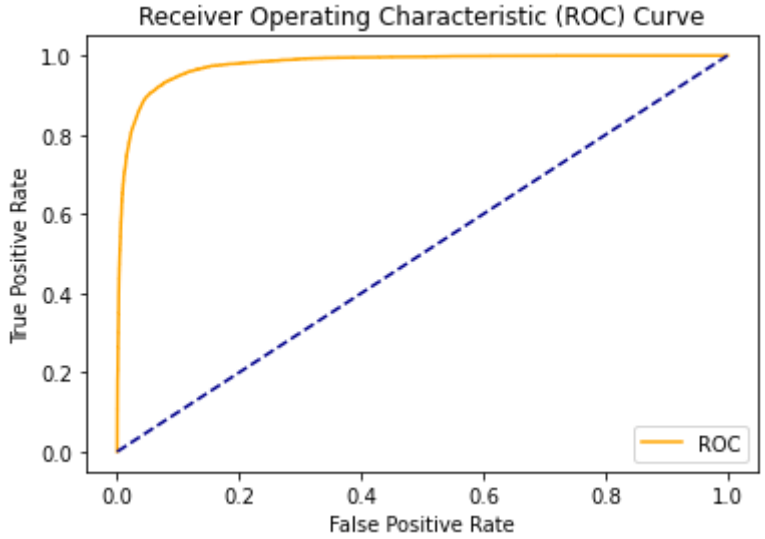


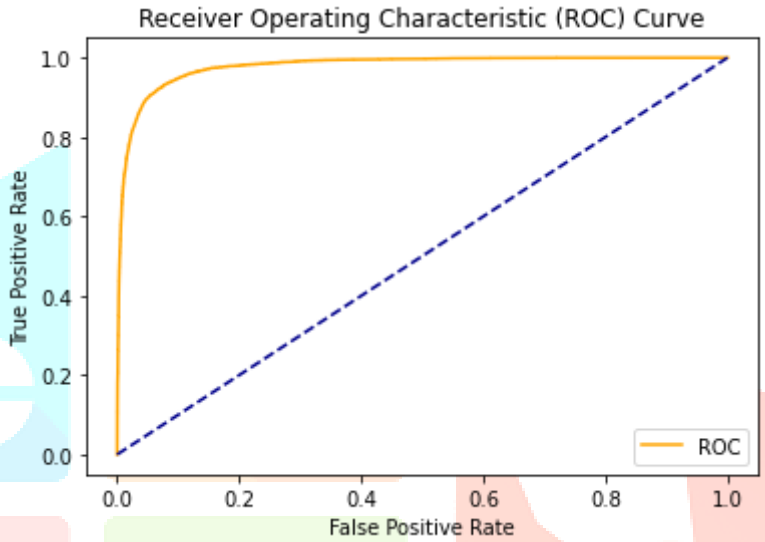
Random Forest :

An approach for ensemble learning called RF is utilized to solve classification and regression issues. A final prediction is made by combining the results of many decision trees created by this supervised learning technique.

A forest of decision trees is created using Random Forest, and each tree is trained using a random subset of the training data and the features. This lessens overfitting and enhances generalization ability. The user can alter two hyperparameters: the size of the subgroups and the number of trees in the forest.

Each decision tree in the forest makes a separate prediction about the class or value of the new data point during prediction. The final forecast is then made by averaging or taking the majority vote (for classification) from all of the decision trees' outputs. (in the case of regression).



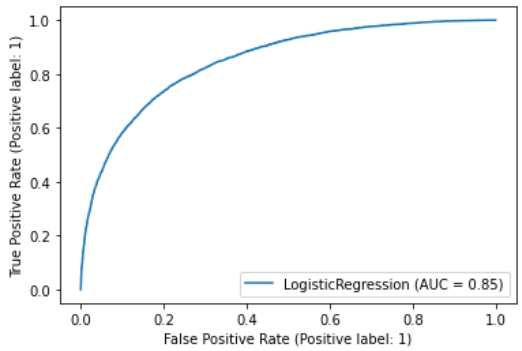
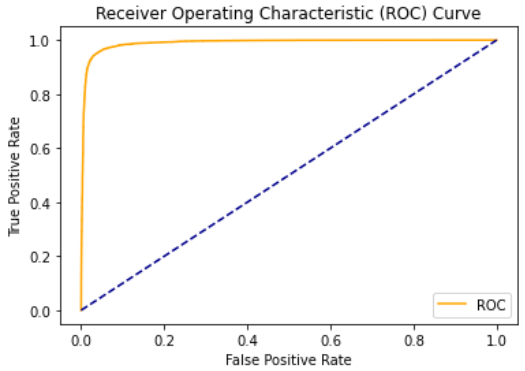
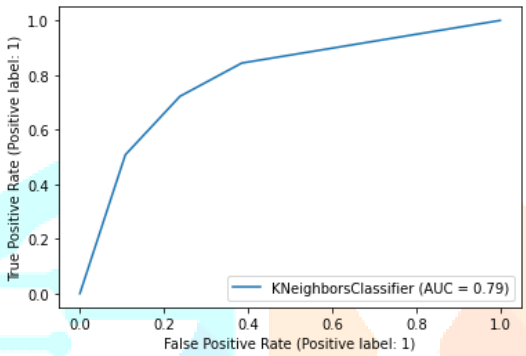
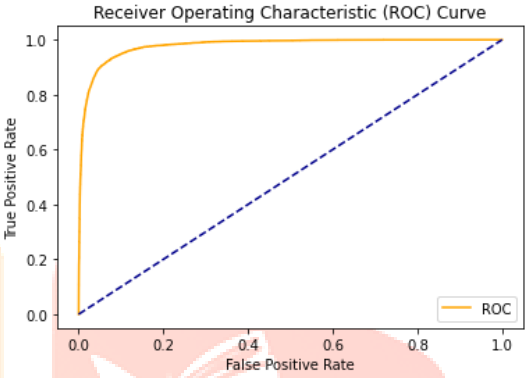
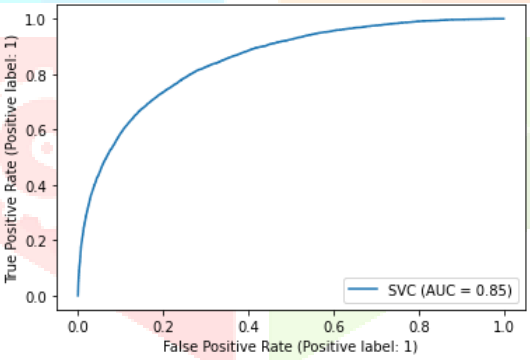


V. EVALUATION METRICS

Algorithms	Recall	Precision	F1-Score	Accuracy
Logistic Regression	78	92	84	85
K-Nearest Neighbour	76	91	83	79
Support Vector Machine	78	92	85	85
Random Forest	93	94	94	92
XGBoost	95	97	96	95

VI. COMPARISON

Based on the predicted day of rainfall, the accuracy of the rainfall prediction is higher than in the previous paper

EXISTING METHODS	PROPOSED METHODS
	
	
	

VII. IMPLEMENTATION

An method known as machine learning enables software programs to improve their accuracy and predictability without having to be explicitly designed. A subset of AI promotes a system's capacity to learn from data, recognize patterns, and make judgement calls to push for best solutions with the least amount of human involvement. Algorithms for machine learning (ML) can be divided into two categories: unsupervised and supervised.

Data Collection :

The regional weather station in Bahir Dar City, Ethiopia, provided the raw data for this study. Ten various forms of data were included, including year, month, date, evaporation, daylight, maximum temperature, minimum temperature, humidity, wind speed, and rainfall. The equipment inside the meteorological station immediately record the environmental variable's values every day for every year. The data were entered into a tabular format using Microsoft Excel. In the row of tables relating to environmental factors in the table's column, the year and the days of the month were ordered.

Data preprocessing :

The data conversion, missing value management, categorical encoding, and separation of the dataset into training and testing datasets were all included in the data preprocessing step. Data from the meteorology office were collected for a total of 20 years . The target variable's missing values were eliminated, and the other features were filled using the data's mean because the raw data had missing values and incorrectly encoded values.

In the meteorology office, the raw data were likewise arranged by year, with attributes in rows that needed to be combined and features in columns that needed to be rearranged. However, data were transformed from excel to CSV.

The dataset was first encoded, and then it was ready for the experiment. The dataset was split into 80% for training and 20% for testing, with the key features for rainfall prediction being regarded as an input for the model.

Model :

In this work, rainfall amounts were predicted using machine learning. To analyze two machine learning techniques that used input variables with moderately and strongly associated environmental factors with rainfall, Random Forest (RF) and gradient descent XGBoost were used. Based on the performance metric utilizing RMSE and MAE, there is a machine learning technique. and published.

Findings :

The major goal of this work was to employ machine learning approaches to identify the pertinent atmospheric variables that generate rainfall and predict the severity of daily rainfall. The machine learning model fed its algorithms with the chosen environmental features. The RF and XGBoost performances were evaluated using MAE and RMSE, and the regression models were constructed in Python. The XGBoost algorithm outperformed the RF at forecasting rainfall using pertinent, carefully chosen environmental parameters.

VIII. CONCLUSION

Machine learning and data science are employed in the application area of rainfall prediction to foretell atmospheric conditions. Predicting rainfall intensity is crucial for efficient water usage, crop production, and the reduction of rain-related disease and food-related death. This article examined different machine learning algorithms for predicting rainfall. Using the data gathered from the meteorological station in Bahir Dar City, Ethiopia, two machine learning algorithms, FR and XGBoost, were presented and tested.

The input variables for the machine learning model utilized in this paper were the chosen features. The XGBoost algorithm was discovered to be a more effective machine learning method for daily rainfall amount prediction utilizing specified environmental parameters when results from the two algorithms (RF and XGBoost) were compared. If the sensor Data is employed in the study, The precision of the rainfall amount prediction may improve. However, this study did not take into account the sensor data.

Using sensor and meteorological datasets with extra different environmental parameters, the accuracy of the rainfall prediction can be increased. Therefore, if sensor and meteorological information are combined to predict daily rainfall amounts, big data analysis can be employed for rainfall prediction in future studies.

IX. ACKNOWLEDGEMENT

I am grateful for the opportunity to acknowledge the guidance and support of Dr. T. Sathish Kumar during the course of our project. His mentorship has been instrumental in helping us navigate through the challenges of the project and delivering the desired outcomes. His insightful feedback, constructive criticism, and unwavering support have been invaluable to us, and we could not have achieved this level of success without his guidance. We greatly appreciate his dedication and commitment to the project and for being a source of inspiration for our team. Thank you, Dr. T. Sathish Kumar, for your invaluable contribution to the project.

X. REFERENCES

[1.] Ehsan MA. Seasonal predictability of Ethiopian Kiremt rainfall and forecast skill of ECMWF's SEAS5 model. Climate Dynamics. 2021; 1–17.

[2.] Kusiak A, Verma AP, Roz E. Modeling and prediction of rainfall using radar refectivity data: a data-mining approach. IEEE Trans Geosci Remote Sens. 2013;51:2337–42.

[3.] Namitha K, Jayapriya A, SanthoshKumar G. Rainfall prediction using artifcial neural network on map-reduce framework. ACM. 2015. https://doi.org/10.1145/2791405.2791468.

[4.] Tharun VP, Prakash R, Devi SR. Prediction of Rainfall Using Data Mining Techniques. In 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT). IEEE Xplore. 2018; pp. 1507–1512.

[5.] Zainudin S, Jasim DS, Bakar AA. Comparative analysis of data mining techniques for malaysian rainfall prediction. Int J Adv Sci Eng Inform Technol. 2016;6(6):1148–53.

[6.] Manandhar S, Dev S, Lee YH, Meng YS, Winkler S. A data-driven approach for accurate rainfall prediction. IEEE Trans Geosci Remote Sens. 2019;5(11):9323–31.

[7.] Arnav G, Kanchipuram Tamil Nadu. Rainfall prediction using machine learning. Int J Innovative Sci Res Technol. 2019. 56–58.

[8.] Aswin S, Geetha P, Vinayakumar R. Deep learning models for the prediction of rainfall.. IEEE: New York. 2018; pp. 0657–0661.

[9.] Zeelan BCMAK, Bhavana N, Bhavya P, Sowmya V. Proceedings of the International Conference on Electronics and Sustainable Communication Systems (ICESC 2020). Middlesex University: IEEE Xplore. 2020; pp. 92–97.