RAINFALL PREDICTION USING THE XGBOOST MODEL

K.N.S.P.Anjana[1] ,Dr.K.Srinivasa Rao[2] ,A.Prashanthi[3] ,V.V.S.S Varshini[4] ,G.Prashanth Kumar[5]

[2]Assistant Professor, Dept of CSE, Sri Vasavi Engineering College (Affiliated by JNTUK), Tadepalligudem, AP, INDIA

[1,3,4,5]Students, Dept of CSE , Sri Vasavi Engineering College (Affiliated by JNTUK), Tadepalligudem, AP, INDIA

***Corresponding Author(s)-Email-ID(s):** srinivasarao.cse@srivasaviengg.ac.in
anjanakaruturi3@gmail.com
prashanthiattili43@gmail.com
2004varshi@gmail.com
prasanthgeddam99088@gmail.com

Abstract

Rainfall has a wide range of effects on human life, affecting industries such as agriculture and transportation while also causing natural disasters such as droughts and floods. Accurate rainfall prediction is crucial for weather forecasting and disaster preparedness, notably influencing agriculture, water resource management, and disaster risk reduction. The main aim is to decrease the effects of rainfall. Past research on rainfall prediction models has shown limitations and suboptimal performance in their results. This study endeavors to construct an advanced rainfall prediction model using the highly efficient XGBoost algorithm, known for its outstanding predictive performance due to its adeptness at managing complex data relationships. The research focuses on selecting optimal environmental variables as inputs for this machine-learning algorithm, leveraging historical weather data with features like temperature, humidity, wind speed, and atmospheric pressure to train the XGBoost model.

Keywords: Rainfall Prediction, Weather Forecasting, Machine Learning, XGBoost Algorithm.

I. Introduction

Rainfall affects human life in various sectors, like agriculture and transportation, and causes natural disasters like drought and floods. Rainfall remains one of the meteorological parameters in many aspects of our daily lives. Rainfall prediction, also known as rainfall forecasting, uses various theoretical and sample methods to estimate the amount of rainfall that is expected to occur in a specific regional area during a defined period in the future. Predicting rainfall is essential for various sectors, including agriculture, water resource management, and disaster preparedness, as it enables better planning and risk mitigation. The aim is to provide necessary information about the chances of weather-related precipitation, which is key for a range of applications and sectors. Rainfall prediction involves studying various climatic factors and old weather data to make familiar forecasts about future rainfall. These factors can include atmospheric pressure, temperature, humidity, wind patterns, and environmental features. Machine learning, a facet of artificial intelligence, centres on the creation of algorithms and models. It stands as a robust instrument for tackling intricate problems, employing techniques and algorithms implemented through computer programs. Machine learning involves three types of predicting rainfall.

They are supervised, unsupervised, and reinforcement learning.

Supervised learning:

Supervised machine learning is a model that involves the collection of data relations from input and output data. It uses labelled datasets to train the model. It helps the model solve real-world problems. In this algorithm, it learns how to make predictions. The capacity of supervised learning to use annotated training data is one of its key features[1].

Unsupervised learning:

Unsupervised machine learning is a model that analyses and clusters unlabelled data. Since unsupervised learning is probably far more prevalent in the brain than supervised learning, it is significant[2]. It tries to find hidden patterns without any predefined labelled data. It helps to extract information from raw data. This model is helpful to train for large datasets.

Reinforcement learning:

Reinforcement learning is a technique that works on feedback. Reinforcement learning (RL) is a subfield of machine learning as well as a learning issue[3]. It is a small area in machine learning that consists of agents. In this technique, the agents are responsible for making decisions. In this technique, The agent engages with the environment. and learns about optimal behaviour. It helps in multiple areas.

XGBoost is a brilliant-boosting supervised learning algorithm. Machine learning, especially the XG Boost algorithm, has become increasingly popular for rainfall prediction because of its effectiveness in handling complex data relationships. XG Boost is a combination-learning method that uses decision tree performance to make accurate predictions. In this study, we will study the application of XG Boost to predicting rainfall. We will use old weather data, including temperature, humidity, wind speed, and other pertinent features, to train the XG Boost model. The main aim of this model is to create effective predictions of rainfall based on features.

II. Literature Survey

AriYairBarrera-Animas[4] proposed a research article “Applied machine learning” A comparative analysis of modern machine learning algorithms for time-series forecasting This study compares rainfall forecasting models, including LSTM, Stacked-LSTM, Bidirectional-LSTM, XGBoost, and an ensemble model, using two decades of UK city weather data. Bidirectional- LSTM views similar performance to Stacked-LSTM in guessing hourly rainfall. Stacked-LSTM with 10 unseen layers performs poorly. Models based on LSTM are likely to overfit training data. Future work involves parameter tuning, handling missing data, and exploring alternative forecasting methods. Enhancing feature importance analysis and considering more weather factors are also important for improving rainfall forecasting models.

K.Sarvani[5] proposed a research article “Predictive Analytics for Rainfall” This project focuses on choosing the best apt algorithm for predicting rainfall, considering various affecting factors. Random Forest Regression is recognized as an adjustable strategy for exact predictions. Machine Learning offers smart models with less computational requirements and manual work. The focus is on regression algorithms, alongside one classification algorithm. The study covers important preprocessing techniques to ensure dataset quality, addressing noise and overfitting. Through a comparative analysis, the project increases the understanding of how unique algorithms react to input data disturbances.

KOPPOLLA KRISHNABABU[6]

Proposed a research article “Forecasting Rainfall” This study explores the application of data mining techniques, including classification, clustering, neural networks, and decision trees, to upgrade historical weather forecasting models. The goal is to improve classification and prediction performance. However, a few limitations are noted in the model, which requires another review before implementation. Additionally, challenges exist in implementing data mining techniques in weather forecasting, highlighting the need for better integration in this field So, this paper proposes HSMS (Hospital Services Management System) which aims at improving quality of services, identifying cost reduction areas, analyses and evaluate/rate healthcare services.

Liyew[7] proposed a research article ”unveiling Nature’s secretes” This projectexplores the application of data science and machine learning in forecasting rainfall, a critical element for efficient water resource management and crop production, and diminishing weather-related risks. Three machine learning algorithms—MLR, RF, and XGBoost—were judged using a meteorological report from Bahir Dar City, Ethiopia. Related environmental featureswere identified using Pearson correlation, serving as inputs for the models. Among the algorithms, XGBoost exhibits higherperformance in daily rainfall prediction. The potential for further improvement lies in incorporating sensor and meteorological datasets, suggesting a future avenue for big data analysis in rainfall prediction.

B.Meena Preethi[8] proposed a research article “Machine learning Meteorology” Rainfall prediction is necessary for agricultural success, as it directly affects crop growth. To support farmers, this method predicts rainfall for the Indian dataset using multiple linear regression (MLR). The results indicate that this MLR-based approach excels existing algorithms in terms of accuracy, mean squared error (MSE), and correlation. This means it provides more efficient rainfall forecasts, which are vital for effective agricultural planning and production.

D.Sirisha[9] proposed a research article “unleashing Predictive potential” This project employs three boosting techniquesand an artificial neural network with forwardand backpropagation to predict rainfall. It relies on a meaningful dataset ofapproximately 145,000 data points from Australia in Excel alignment. After training the models, they were assessed on hidden data, yielding correct temperature predictions. Notably, the artificial neural network, utilizing forward and backward propagation, outperformed the boosting algorithms, making it a more suitable possibility for forecasting rainfall conditions for the following day.

Chalachew Muluken Liyew[10] proposed a research article “Machine Learning Expedition” Rainfall prediction, a key application of data science and machine learning, assists forecast rainfall potency for efficient water resource management, crop production, and flood prevention. In this study conducted in Bahir Dar City, Ethiopia, three machine learning algorithms—MLR, RF, and XGBoost — are evaluated using meteorological station data. To recognize the most relevant environmental aspects for prediction, Pearson correlation coefficients were employed. Results indicated that XGBoost outperformed MLR and RF in daily rainfall amount prediction. However, the study would not consider sensor data, which could further enhance accuracy. Future work should explore sensor and meteorological datasets to refine rainfall predictions. R **Vijayan[11]** proposed a research article “Precise Rainfall Prophecy” This study aimed to predict rainfall in the city using five data mining methods, including support vector machines, random forests, and multilayer perceptron’s. It used 12 years of historical weather data from December 2005 to November 2017. The data mining techniques are estimated, and the results were presented in tables and graphs. Results showed that while the methods performed well for predicting no-rain conditions, they struggled with rain predictions, possibly due to missing data, inadequate climatic information, and low overall precipitation in the area. Future work suggests testing different classification techniques and climate attributes on various dates to improve accuracy, with a focus on random forest and logistic regression.

Gowtham Sethupathi[12] proposed a research article ”Streamlining rainfall prediction” Rainfall forecasting is challenging, especially in complex situations due to the helplessness to predict unseen patterns accurately. To raise accuracy two methods, random forest and logistic regression, are being compared for rainfall prediction. Data mining techniques are employed, and the results are presented visually. A classification approach is used, involving data cleaning and normalization, to make accurate predictions. While these methods work well for predicting no rain, they are less victorious for predicting rain, possibly due to missing data and insufficient climate information. Future efforts should inspect additional classification methods and climate attributes to improve accuracy. Currently, a combination of random forest and logistic regression proves to be highly efficient for accurate rainfall predictions.

Antonio Sarasa-Cabezuelo[13] proposed a research article “Anticipating Rainfall” This study focused on using machine learning for rainfall prediction in Australia, specifically in the regions of Victoria and Sydney. They collected meteorological data from 49 cities, including information like whether it rained (Rain Today) and various weather properties. Different machine learning models, such as Knn, Decision Tree, Random Forest, and Neural Networks, were tested. The results showed that Neural Networks were the most effective for this type of prediction. The study also looked at how well the models worked in different cities and explored the impact of varying training data, but no significant improvement was observed. This research highlights the potential of machine learning in rainfall prediction, particularly Neural Networks, and its applicability in different locations.

III. Proposed System:

The Rainfall Predicting System relies on crucial meteorological variables like temperature, humidity, atmospheric pressure, wind speed, and wind direction at a specific geographical location to make accurate rainfall forecasts. What sets this system apart is its inclusion of both maximum and minimum temperatures as pivotal factors in the prediction process. The dataset is diligently curated, with data points being collected every six hours, ensuring its high quality and reliability for precise forecasting. At the heart of this predictive mechanism lies the XGBoost model, chosen for its well-regarded proficiency in making accurate predictions.

IV. Methodology

XG boost which refers to extreme gradient boosting. It is the most popular mission learning algorithm Tianqi Chen stamped it. It is a supervised Machine learning technique. It supports both classification and regression problems XG boost which refers to extreme gradient boosting. It is the most popular machine learning algorithm it was stamped by Tianqi Chen[13]. It is a supervised Machine learning technique. It supports both classification and regression problems

Classification:

In machine learning, the data is typically divided into several classes using a supervised learning approach. The results of classification problems are categorical. By default, the square root of the total number of features is used to pick the features.A decision tree's majority vote is used to determine the outcome.

Regression:

The result of regression problems is continuous or real. There are several metrics involved in regression like root-mean-squared- error(RMSE), and mean-squared-error(MSE)(It is an absolute sum of actual and predicted differences, but it is lacked mathematically). In XGBoost the model is based on Decision trees. It builds the decision trees in sequential form. Similar to a road map, a decision tree represents a test of a certain characteristic at each decision point.

In machine learning, decision tree learning employs a decision tree as a predictive model that maps observations regarding an item to judgments regarding the intended worth of the item[14]. The journey's branches show results depending on these checks as it go along, leading to the terminal nodes with separate class labels. The tree grows by repeatedly splitting the starting set, assisted by attribute tests, until subsets converge with consistent target variable values or more splits produce insignificant predictive value. Weights play a crucial role in this algorithm. Based on the weights only the result is predicted.

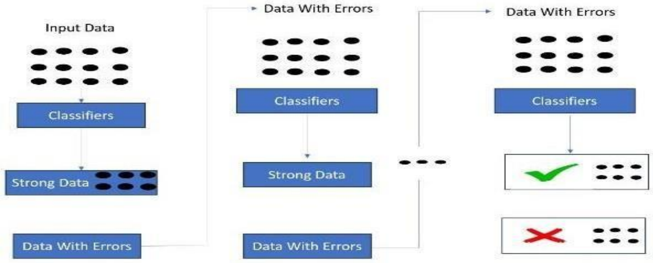
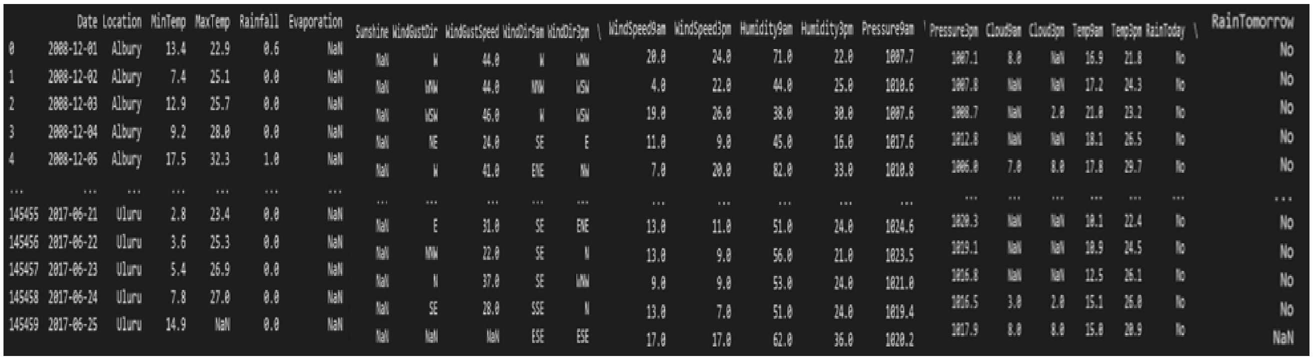


Fig1: Represents the Decision tree

Data Collection:

Data is collected from Kaggle. We have taken a CSV file that is related to Australia. It has attributes like max temp, wind gust speed, pressure, humidity, etc. The following is the dataset:



```
              Date Location MinTemp MaxTemp Rainfall Evaporation  Sunshine WindGustDir WindGustSpeed WindDir9am WindDir3pm \  WindSpeed9am WindSpeed3pm Humidity9am Humidity3pm Pressure9am \ Pressure3pm Cloud9am Cloud3pm Temp9am Temp3pm RainToday \  RainTomorrow
0       2008-12-01   Albury    13.4    22.9      0.6         NaN       NaN           W          44.0          W        WNW            20.0         24.0        71.0        22.0      1007.7        1007.1      8.0      NaN    16.9    21.8        No             No
1       2008-12-02   Albury     7.4    25.1      0.0         NaN       NaN         WNW          44.0        NNW        WSW             4.0         22.0        44.0        25.0      1010.6        1007.8      NaN      NaN    17.2    24.3        No             No
2       2008-12-03   Albury    12.9    25.7      0.0         NaN       NaN         WSW          46.0          W        WSW            19.0         26.0        38.0        30.0      1007.6        1008.7      NaN      2.0    21.0    23.2        No             No
3       2008-12-04   Albury     9.2    28.0      0.0         NaN       NaN          NE          24.0         SE          E            11.0          9.0        45.0        16.0      1017.6        1012.8      NaN      NaN    18.1    26.5        No             No
4       2008-12-05   Albury    17.5    32.3      1.0         NaN       NaN           W          41.0        ENE         NW             7.0         20.0        82.0        33.0      1010.8        1006.0      7.0      8.0    17.8    29.7        No             No
...            ...      ...     ...     ...      ...         ...       ...         ...           ...        ...        ...             ...          ...         ...         ...         ...           ...      ...      ...     ...     ...       ...            ...
145455  2017-06-21    Uluru     2.8    23.4      0.0         NaN       NaN           E          31.0         SE        ENE            13.0         11.0        51.0        24.0      1024.6        1020.3      NaN      NaN    10.1    22.4        No             No
145456  2017-06-22    Uluru     3.6    25.3      0.0         NaN       NaN         NNW          22.0         SE          N            13.0          9.0        56.0        21.0      1023.5        1019.1      NaN      NaN    10.9    24.5        No             No
145457  2017-06-23    Uluru     5.4    26.9      0.0         NaN       NaN           N          37.0         SE        WNW             9.0          9.0        53.0        24.0      1021.0        1016.8      NaN      NaN    12.5    26.1        No             No
145458  2017-06-24    Uluru     7.8    27.0      0.0         NaN       NaN          SE          28.0        SSE          N            13.0          7.0        51.0        24.0      1019.4        1016.5      3.0      2.0    15.1    26.0        No             No
145459  2017-06-25    Uluru    14.9     NaN      0.0         NaN       NaN         NaN           NaN        ESE        ESE            17.0         17.0        62.0        36.0      1020.2        1017.9      8.0      8.0    15.0    20.9        No            NaN
```

Fig2: Denotes the Dataset

Data processing:

An essential phase in the pipeline for data analysis and machine learning is data pretreatment. It entails preparing unprocessed data for analysis or machine learning model training by cleaning and converting it into a format that works well. Key elements of data preparation are listed below:
Handling Missing Data:
Identify and manage missing values: Use methods like as mean or median imputation, or apply more sophisticated approaches, to remove or impute missing values.
Data cleaning:
Eliminate duplicates: Find and eliminate records that are duplicates in the dataset.
Identifying and managing outliers: Outliers that could skew the analysis or model training should be identified and dealt with.
Data Transformation:
Scaling: To make sure numerical features are on the same scale, standardize or normalize them. Z-score normalization and Min-Max scaling are popular methods.
Coding categorical variables: Use methods
such as label encoding or one-hot encoding to translate category variables into a numerical format.
Engineering Features:
Make fresh features: To give the model additional information, extract new features from the ones that already exist.
Binding: To transform continuous variables into categorical ones, group numerical values into bins.

Features of polynomials: In order to identify non-linear relationships in the data, create polynomial features.

Handling Unbalanced Information:

If there is a class imbalance, address it in the target variable. Methods include using fake data, under sampling, and oversampling.

Managing Data on Dates and Times:

From date and time variables, such as day of the week, month, or hour, extract pertinent information.

Normalization and Scaling of Data:

Numerical features are guaranteed to be on the same scale by scaling. Z-score normalization and Min-Max scaling are popular techniques.

Managing Text Information:

Tokenization: Segment text into discrete words or units of measurement.

Eliminating stop words: Get rid of terms that are frequently used but don't add much information.

Lemmatization and stemming reduce words to their most basic form.

Data division:

To assess the performance of the model, divide the dataset into training, validation, and testing sets.

Information Visualization:

Analyze the feature distribution visually to spot trends, anomalies, and connections.

Use correlation matrices, scatter plots, or histograms to examine relationships between variables.

Standardization and Normalization:

To make sure that the features have comparable scales, normalize or standardize them.

Managing Types of Data:

Utilize methods such as label encoding or one-hot encoding to transform categorical values into a numerical representation.

Managing Uncertain Data:

Eliminate any noisy data that can skew the analysis or cause inaccuracies.

Integration of Data:

If necessary, combine data from several sources to guarantee consistency and compatibility.

The following are steps for working of an extreme gradient boosting algorithm:

The ensemble-based approach includes the boosting technique known as extreme gradient boosting (XGBoost)[15].

Algorithm: Rainfall Prediction Using XGBoost

```
Input:
-Training dataset D_train = {(X_train_i, y_train_i)},
i = 1 to N_train
-Testing dataset D_test = {(X_test_i, y_test_i)},
i = 1 to N_test
-Hyperparameters: max_depth, learning rate,   n_estimators, objective, …
Output:
-Predicted rainfall values for testing dataset: y_pred
Initialization:
-Initialize an empty ensemble of decision trees:
G = 0
Training:
-For t = 1 to n_estimators:
-Calculate gradients G_t and Hessians H_t for each sample (X_train_i, y_train_i)
-Train a decision tree T_t to fit the gradients and Hessians using features from D_train
-Update the ensemble: G = G + learning_rate * T_t
```

Testing:
-For each sample (X_test_i, y_test_i) in D_test:
-Predict the rainfall value y_pred_i using the ensemble of decision trees G
Evaluation:
-Calculate the mean squared error (MSE) between y_pred and the actual rainfall values y_test
Future Rainfall Prediction:
-Given new feature values for future periods, use the trained ensemble G to predict future rainfall values
End

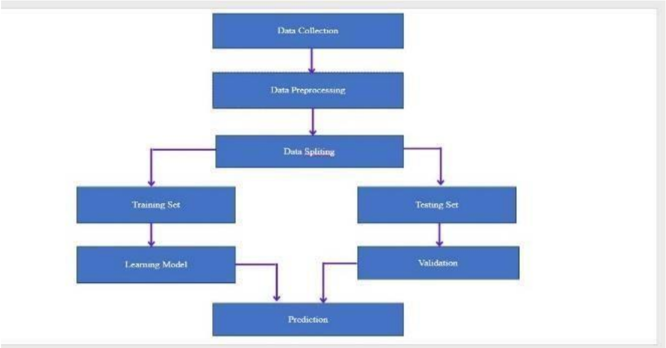


Fig 3: Denotes a Flowchart

Benefits of XGBOOST:

1.XGBoost frequently creates models that are very precise. Decision trees serve as the base learners in its ensemble technique, which helps capture complicated data correlations.

2.XGBoost already has a method in place to deal with missing values in the dataset, which eliminates the need for laborious data preprocessing.

3.Both regression and classification issues can be solved with XGBoost. Due to its versatility, it can be used for a variety of activities.

4.XGBoost frequently performs quicker than implementations of conventional gradient boosting. For rapid model training, it uses parallelization and optimization approaches

V.RESUL AND DISCUSSION

At first, the dataset was collected from the Kaggle site and it was downloaded as a CSV file and then it was processed using Python. The collected dataset consists of 145460 rows and 23 columns. For the processing of algorithms, various Python libraries including matplotlib, NumPy, pandas, and Sklearn are used. 80% of the dataset is split into a training dataset and the remaining 20% is split into a test dataset. The purpose of training data is to train the model whereas testing data is used to check the performance of the model. The algorithm explored in this proposed paper is Extreme Gradient Boosting. It can be used for large Datasets and to get accurate predictions.

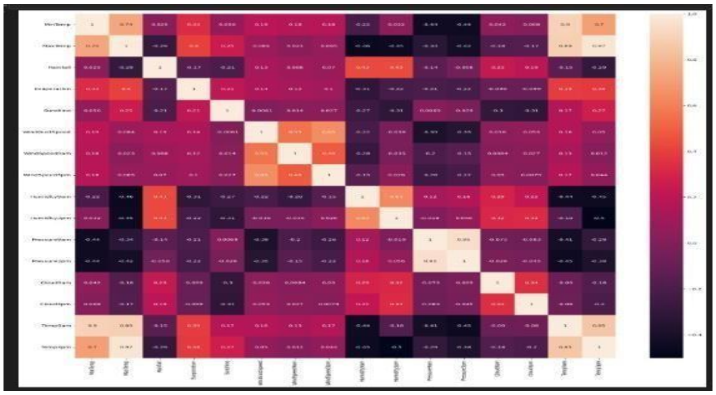


Fig4: Denotes the Heat Map

Explanation: Heat map shows the correlation between the data in matrices such as correlation matrices. The intensity of the color represents magnitude of the data at a specific data.

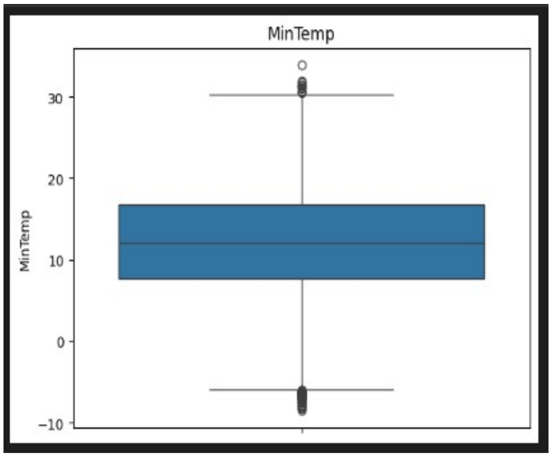


Fig5: Represents the Boxplot of MinTemp

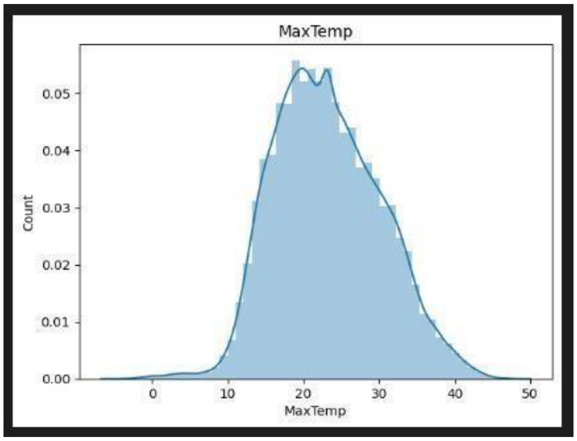


Fig6: Denotes the Max Temp

VI. Conclusion

The accurate prediction of rainfall is crucial for effective management of water resources, safeguarding human life, and preserving the environment. However, this can be a challenging task due to the variations in geography and region. In this paper, we have reviewed and concluded that the use of the XGBOOST algorithm for rainfall prediction is a powerful and effective approach. This algorithm has proven its ability to identify complex relationships within the data and provide accurate forecasts in sectors such as agriculture, disaster management, and water resources management. Its agility and predictive accuracy make it the preferred approach over all other available methods.

VII. Future Scope

The future scope of rainfall prediction using machine learning involves the collection of current atmospheric data through advanced sensors and the application of sophisticated algorithms. Future advancements in rainfall prediction will likely involve the deployment of highly sophisticated sensors that can collect real-time data on various atmospheric parameters.

Reference:

[1]Vladimir Nakeski, A summary of techniques for supervised machine learning.

[2]Peter Dayan ,unsupervised learning.

[3]Morgam & clay pool Publisher, Csaba Szepesvari.

[4] Ari Yair Barrera-Animas, Lukumon Akinosho, Juan Manuel Davila Delgado, Lukman Adewale Akanbi. Machine Learning with Applications 7 (2022) 100204.

[5] K.Sarvani , Y. Sai Priya , Ch. Teja , T. Lokesh and E.Bala Bhaskara Rao. RAINFALL ANALYSIS AND PREDICTION USING MACHINE LEARNING TECHNIQUES.

[6] KOPPOLLAKRISHNABABU(38110263),KASU PURUSHOTHAMREDDY (38110244).RAINFALL PREDICTION USING MACHINELEARNING TECNIQUES

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

[8] B.Meena Preethi, R.Gowtham, Aishwarya, S.Karthick, D.G.Sabareesh. Rainfall Prediction using Machine Learning and Deep Learning Algorithms.

[9] D Sirisha , P. Srijani, R. Dharani , K. Durga Vinusha , K. Pavan Kalyan Predicting Rainfall using Machine Learning Techniques.

[10] Chalachew Muluken Liyew* and Haileyesus Amsaya Melese. Machine learning techniques to predict daily rainfall amount.

[11] R Vijayan, V Mareeswari, P Mohankumar, G Gunasekaran K Srikar. Estimating Rainfall Prediction using Machine Learning Techniques on a Dataset.

[12] Gowtham Sethupathi.Ma , Yenugudhati Sai Ganesh b , Mohammad Mansoor Ali. Efficient Rainfall Prediction and Analysis using Machine Learning Techniques.

[13] Antonio Sarasa-Cabezuelo. Prediction of Rainfall in Australia Using Machine Learning

[14] Tianqi Chen, Carlos Guestrin, 16, August 13- 17, 2016, San Francisco, CA, USA c 2016 ACM.ISBN978-1-4503-4232-2/16/08

[15] Bhoopesh Singh Bhati1 Garvit Chugh2 Fadi Al-Turjman3 Nitesh Singh Bhati2, A better ensemble-based intrusion detection method that makes use of XGBoost,Trans Emerging Tel Tech. 2020;4076

[16] 7.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.

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

[18] Aswin S, Geetha P, Vinayakumar R. Deep learning models for the prediction of rainfall. In 2018 International Conference on Communication and Signal Processing (ICCSP). IEEE: New York. 2018; pp. 0657– 0661.

[19] Zeelan BCMAK, Bhavana N, Bhavya P, Sowmya V. Rainfall prediction using machine learning & deep learning techniques. Proceedings of the International Conference on Electronics and Sustainable Communication Systems (ICESC 2020). Middlesex University: IEEE Xplore. 2020; pp. 92–97.

[20] Vijayan R, Mareeswari V, Mohankumar P, Gunasekaran G, Srikar K, (JUNE,. Estimating rainfall prediction using machine learning techniques on a dataset. Int J Sci Technol Res. 2020;9(06):440–5.

[21] Chaudhari MM, Choudhari DN. Study of various rainfall estimation & prediction techniques using data mining. Am J Eng Res. 2017;6(7):137–9.