# Prediction of Rainfall and its Affecting Factors using Machine Learning

Kajal Gangele[1], Priyanka Karande[2], Ashwini Thappa[3], Prof. Sujata Kullur[4]

[1,2,3,4]*B. Tech Information Technology Usha Mittal Institute of Technology,SNDT Women's University*

***Abstract*—Predicting rainfall is the most difficult aspect of weather forecasting. Accurate rainfall forecasting is now more challenging than ever because of significant climatic changes. So, it is essential to study how rain acts in connection to the variables namely, temperature, humidity, pressure, and wind speed. Then, and only then, will we be able to accurately predict the rain.Forecasting is the process of using past data to create assumptions about future variations in rainfall amounts. Furthermore, time forecasting is concerned with the creation of models and methods that lead to precise forecasts and predictions. This study, whichis a survey, used a thorough mapping study and a methodical literature review. The most popular linear time series models used in time series forecasting are ARIMA, Prophet, and Holt Winter's Exponential Smoothing, which have been around for a while due to their high predicting accuracy and will be employed in this study. The major objective of this project is to raise an understanding of time series forecasting and its techniques and to develop a system that not only forecasts rainfall but also the variables that influence it.**

***Index Terms*—Rainfall, Temperature, humidity, wind speed, pressure, Time series, forecasting, ARIMA, Holt Winter's Exponential Smoothing, and Prophet**

## I.INTRODUCTION

Water is one of the main contributors to its success. It is possible to employ a variety of water reservoirs, including rivers, canals, and bore wells, for agricultural purposes. Nonetheless, rainwater serves as these reservoirs' primary supply. Due to geographic and financial factors, every farmer cannot afford a bore well or other traditional technologies. Rainfall is thereforeagriculture's biggest asset and key component.

Accurately estimating rainfall is crucial for many industries,including agriculture, pest control, tourism, event planning, water conservation, flood and drought forecasting, etc. India is currently experiencing several disaster situations that are causing loss of lives and property as a result of inaccurate forecasting.

Rainfall prediction is essential for accurately managing water resources, and its patterns are affected by primary factors such as temperature, Humidity, Pressure, and Wind speed. Machine learning algorithms are used to forecast future weather conditions by using hidden patterns and relationships between weather data.

Machine learning is a classifier of Supervised, Unsupervised, and Ensemble Learning used to predict and find the accuracy of a given dataset. It can be implemented via two approaches Empirical and Dynamical. The empirical approach involves analyzing historical data on rainfall and its relationship to several atmospheric variables, while the Dynamical approach generates predictions based on systems of equations to predict the evolution of the global climate system. The most widely used empirical approaches for climate prediction are regression, Artificial Neural Networks (ANN), fuzzy logic, andgroup method of data handling.

Here, we'll apply one of the most well-liked machine-learning methods to forecast rainfall known as the Times seriesMethod. This technique has the 3 most used algorithms which we would use for analysis: ARIMA, Holt Winter's Exponential Smoothing, and Prophet. After comparing their results, we willuse the most precise algorithm for prediction.

We are employing the Time series model (empirical approach) for prediction because we will be working on daily data from all states. This model's (Time series) goal is to Forecast the values of some variables based on their historical values. Forecasting is the use of models and methods to generate a good forecast. It is used in numerous fields, such asproduction, marketing, economics, and finance. So, our motive is to generate a good forecast.

## II.LITERATURE SURVEY

In **Abhishek Kumar, Abhay Kumar, Rahul Ranjan & SarthakKumar in 2012 "A rainfall prediction model using artificial neural network"**. The average monthly rainfall of India's monsoon season was predicted by researchers using ANN.For the forecast, a dataset that covered 8 months per year was employed. It was determined that there was a substantial likelihood of rain throughout the chosen months. Feed Forward Back Propagation, Layer Recurrent, and Cascaded Feed Forward Back Propagation were the three different typesof networks that were used for the performance analysis. Feed Forward Back Propagation fared better than the others, according to the findings. [1]

**Jyothis Joseph& Ratheesh TK 2013 published "Rainfall Prediction using Data Mining Techniques"** This study dis-cusses a model based on the K-means clustering technique and a supervised data classification technique, namely the Classification and Regression Tree, which is used to generate rainfall states from large-scale atmospheric variables in a river basin. Anns are used to implement these techniques and analyze the four months of rainfall data from June to September of a particular region for nine years in India, whichis the monsoon season for the state. The accuracy of the predictions is 87% [2].

**Nikam V.B. & Meshram B.B. IN 2013 published "Modelingrainfall prediction using data mining method: A Bayesianapproach"**. Researchers used a Bayesian modeling approach to present a data-intensive model for predicting rainfall. TheIndian Meteorological Department supplied the dataset and the top 7 attributes were chosen. The proposed approach demonstrated good accuracy for rainfall prediction while utilizing moderate computing resources, reaching a 91% accuracy. [3]

**Deepti Gupta & Udayan Ghose 2015 published "A comparative study of classification algorithms for forecasting rainfall"**. This study has clearly explained Classification as partof data mining is very useful for finding unknown patterns likeforecasting future trends. The K-Nearest Neighbors method's value for K is difficult to calculate, hence this study presentsprediction results for two distinct values of K (accuracy of 80.7%). Decision trees require intensive calculations and time-consuming pruning (accuracy of 80.3%). Neural networks providebetter results than discussed algorithms with 10 neurons ina hidden layer (Accuracy of 82.1%). The most important idea is that it can handle noisy and continuous values better thanCART, but the number of neurons and training networks in multiple modeling is time-consuming and difficult. [4]

**Nasimul Hasan & Nayan Nath 2015 published "A Support Vector Regression Model for Forecasting Rainfall "**. Using current rainfall data for Bangladesh, this research illustrates a reliable rainfall forecast method using Support Vector Regression (SVR), a technique based on Support Vector Machines (SVM). The data was preprocessed to fit the algorithm and the SVR technique outperforms traditional frameworks in termsof accuracy and process running time. The approach gives an accuracy of 99.92% [5].

**Junaida Sulaiman and Siti Hajar Wahab 2018 published "Heavy Rainfall Forecasting Model Using Artificial Neural Network for Flood Prone Area"**. ANN model was used to predict heavy precipitation from 1965 to 2015 using various historical precipitation values from nearby meteorological stations. It was compared to ARIMA's auto-regression integratedmoving average (ARIMA) which was assessed using the correlation coefficient (R2) and root mean square error (RMSE). The results showed that the ANN model is dependable inpredicting the risky level of heavy precipitation events. [6].

**Shreekant Parashar and Tanveer Hurra published "A Study on Prediction of Rainfall Using Several Data Mining Techniques" in 2019**. Through this survey, we understood that there are majorly two approaches for predicting rainfall namely Empirical Method and Dynamic Method. The Empirical approachinvolves analyzing historical data, while Dynamic Method involves generating physical models to predict the evolution of the global climate system in response to initial atmospheric conditions. Bayes Theorem is used to find the probability ofan event that has already occurred. [7]

**Kaushal Kailas Sarda 2022 published "Rainfall Predictions using data visualization techniques"** which reviews rainfall prediction techniques using publicly available data. This paper explains that statistical techniques for rainfall forecasting can- not predict long-term rainfall forecasting due to the constant change of climate phenomena. It has the potential to providea precise classification of rainfall forecasting models and potential future research methods in this area for further discussion. [8]

## III.METHODOLOGY

Time series forecasting is a group of observations, which is being recorded at a particular time and the collection of such observations is known as time series data. Creating a time series model that helps to predict the future value of a time series after observing previous data. Firstly, data is analyzedto pull out statistical information, and characteristics of the data and to predict the output [9].

As the data tend to follow a pattern in time series data, the Machine Learning model finds it difficult to predict approx. hence time series analysis and its approaches have made it simpler for prediction. The time series data is analyzed and visualized to find out 3 important aspects: trend, seasonality, and heteroscedasticity.

- Trend: It is the observation of an increasing or decreasingpattern over a period.
- Seasonality: It is b cyclic happening of events. Statistical properties are the same over time, and its properties are constant means, constant variance, and no seasonality.
- Heteroscedasticity: It is the non-constant variance from the mean calculated at different periods and intervals.

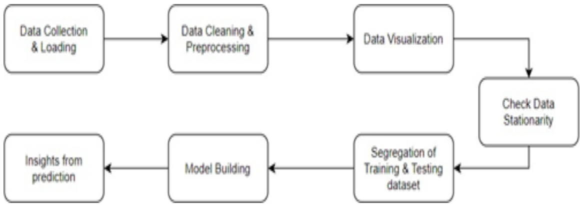

Data Collection & Loading
Data Cleaning & Preprocessing
Data Visualization
Check Data Stationarity
Segregation of Training & Testing dataset
Model Building
Insights from prediction

Fig. 1. Flowchart

Step:

Firstly, we must load the data & read it.

- Cleaning & Pre-processing data: All unwanted things areremoved or replaced. The input should be univariate.
- Check if data is stationary. (If not convert it to stationary)
- Fit the model in the univariate series.
- Predict values on validation set:
- Calculate The Mean Absolute Error, Root Mean Square Error for each forecasting model it is calculated for the testing of different forecasting models. The model having the smallest MAE (Mean Absolute Error) valueis preferred.

$$MAE = \frac{1}{N}\sum_{i=1}^{N}\left|y_i - \hat{y}_i\right|$$

Fig. 2. MAE

Where:

N = the number of errors

$\sum$= summation symbol (which means add them all up)

$|y_i - \hat{y}_i|$ = the absolute errors

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}\|y(i)-\hat{y}(i)\|^2}{N}},$$

Fig. 3. RMSE

Where:

N = number of
data points$y_i$ = i-th
measurement
$\hat{y}_i$ = $y_i$ corresponding prediction.

### *A. ARIMA*

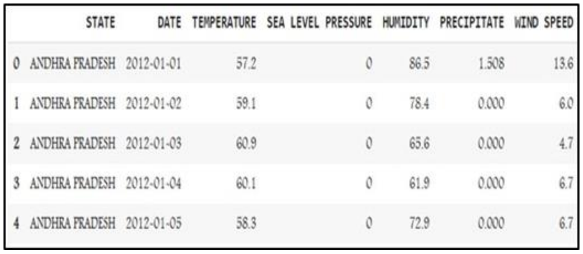
Auto-Regressive Integrated Moving Average is also known as ARIMA. ARIMA is a class of models that enables to capture of a suite of different standard temporal structures in time series data. ARIMA majorly has three components: AR i.e., Auto-Regressive; I i.e., the Differencing term (Integrated); and then MA which is the Moving Average term. Each component of ARIMA is explained below. The term AR Auto- Regressive particularly refers to the past year or historical values to be used to forecast the next future prediction value and is defined by the parameter 'p' in ARIMA. I term refers tothe difference in order and is defined by the parameter 'd' in ARIMA. MA is used to define the number of past forecast errors used to predict future values and is defined by the parameter ‘q’ in Arima.

### B. Holt Winter’s Exponential Smoothing (HWES)

Holt Winter’s Exponential Smoothing (HWES) also knownas the Triple Exponential Smoothing method models the next step as an exponentially weighted linear function of observations at prior time steps, taking trends and seasonality into account. It is a widely known smoothing model for forecasting data that has a trend. Holt Winter Exponential smoothing estimates the seasonal component called gamma (γ) For this method, the seasonal cycle should be specified example weekly (7), monthly (12), quarterly (4), and seasonality refers to a cycle of pattern that occurs in regular interval of time, which can be multiplicative or additive, Multiplicative seasonality has a pattern where the magnitude increases when the data increase. Additive seasonality reflects a seasonal pattern that has a constant scale even as the observations change, steps are similar to the simple exponential smoothing method.

### C. Prophet

Prophet is particularly helpful for the dataset that has a long history of detailed past observations (hourly, daily, or weekly) over an extended period (months or years), multiple strong seasonality, a history of significant but irregular events, large outliers or missing data points, and non-linear growth trends that are approaching a limit. An additive regression model called the prophet is distinguished by a piecewise linear or logistic growth curve trend. The Prophet model additionally includes seasonal components for the week and year that are represented by dummy variables and Fourierseries, respectively.



| | STATE | DATE | TEMPERATURE | SEA LEVEL PRESSURE | HUMIDITY | PRECIPITATE | WIND SPEED |
|---|---|---|---|---|---|---|---|
| 0 | ANDHRA PRADESH | 2012-01-01 | 57.2 | 0 | 86.5 | 1.508 | 13.6 |
| 1 | ANDHRA PRADESH | 2012-01-02 | 59.1 | 0 | 78.4 | 0.000 | 6.0 |
| 2 | ANDHRA PRADESH | 2012-01-03 | 60.9 | 0 | 65.6 | 0.000 | 4.7 |
| 3 | ANDHRA PRADESH | 2012-01-04 | 60.1 | 0 | 61.9 | 0.000 | 6.7 |
| 4 | ANDHRA PRADESH | 2012-01-05 | 58.3 | 0 | 72.9 | 0.000 | 6.7 |

Fig. 4. Dataset

In this project, we collected data from Visual Crossing & thedataset consists of attributes such as states, date, Temperature, Humidity, Sea Level Pressure, Wind speed, and Precipitate (rain) here we analyzed data and applied time series for the prediction of future annual rainfall.

## IV. RESULT

The suggested approach is applied to a dataset taken from Visual Crossing. The dataset being used spans 11 years, from 2012 to 2022 [10].

Considering the size of the dataset we segregated the data into input and targets as States and attributes namely Temperature, Humidity, Precipitate, Sea level pressure, and Wind speed.

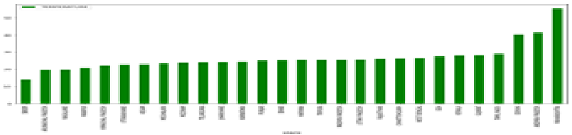


Fig. 5. Average of Temperature v/s State

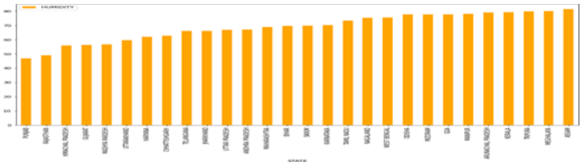


Fig. 6. Average of Humidity v/s State

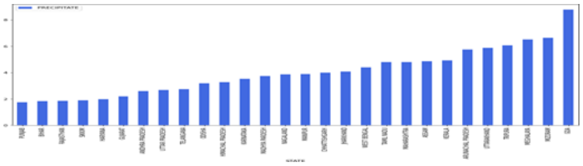


Fig. 7. Average of Precipitate v/s State

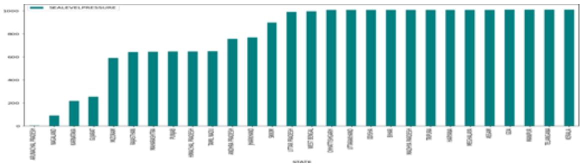


Fig. 8. Average of Sea Level Pressure v/s State

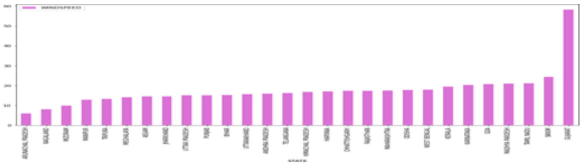


Fig. 9. Average of Wind Speed v/s State

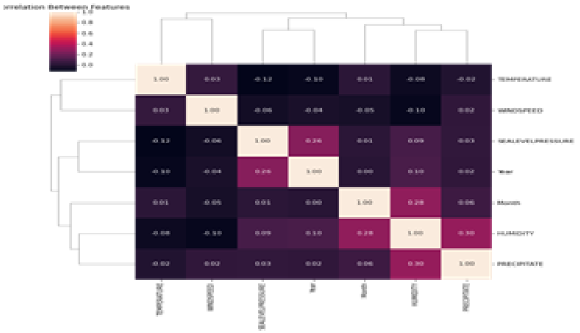


Fig. 10. Correlation Between Features

According to Time series, the data used for prediction should be stationary which means that its statistical properties are constant with time in order to determine this attribute there are two ways namely ADF and Kwiatkowski-Phillips-Schmidt-Shin (KPSS).

In this proposed system we have used ADF Test to determine whether the given data is stationary or not.

In ADF Test if the

p-value > 0.05 (Not stationary)

p-value > = 0.05 (Stationary)

And according to this approach applied, we have the calculated p-value which is less than 0.05, which determines thatthe data taken is stationary.

The dataset is then divided with respect to year i.e. Trainingset: 2012-2021 which is 10 years of data & Testing set: 2022, the last year of data.

TABLE I

| *Temperature* | | |
|---|---|---|
| *Algorithm* | *MAE* | *RMSE* |
| ARIMA | 0.86 | 1.1 |
| HOLT’S WINTER | 0.97 | 1.21 |
| PROPHET | 0.65 | 0.92 |
| *Humidity* | | |
| *Algorithm* | *MAE* | *RMSE* |
| ARIMA | 5.04 | 6.27 |
| HOLT’S WINTER | 6.78 | 8.15 |
| PROPHET | 3.89 | 4.73 |
| *Precipitate (Rain)* | | |
| *Algorithm* | *MAE* | *RMSE* |
| ARIMA | 7.57 | 11.17 |
| HOLT’S WINTER | 4.25 | 11.53 |
| PROPHET | 7.64 | 11.01 |
| *Sea Level Pressure* | | |
| *Algorithm* | *MAE* | *RMSE* |
| ARIMA | 2.83 | 3.18 |
| HOLT’S WINTER | 1.35 | 1.73 |
| PROPHET | 2.77 | 3.3 |
| *Wind Speed* | | |
| *Algorithm* | *MAE* | *RMSE* |
| ARIMA | 5.23 | 7.57 |
| HOLT’S WINTER | 5.43 | 7.86 |
| PROPHET | 3.29 | 5.3 |

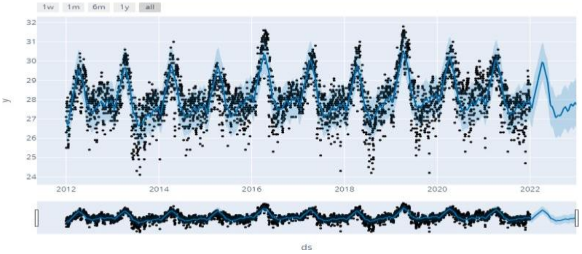


Fig. 11. Prophet-Temperature

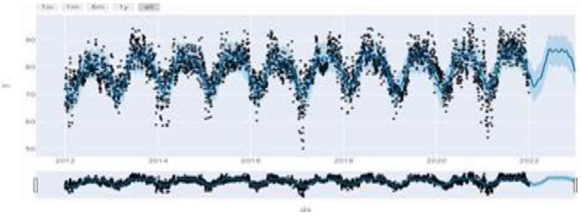


Fig. 12. Prophet-Humidity

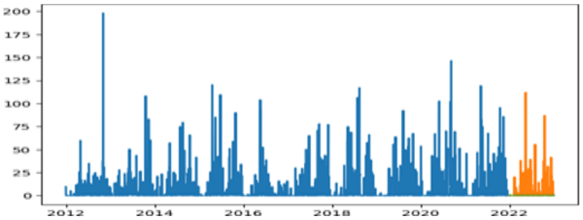


Fig. 13. HOLTS’S WINTER Exponential Smoothing Method – Precipitate(rain)

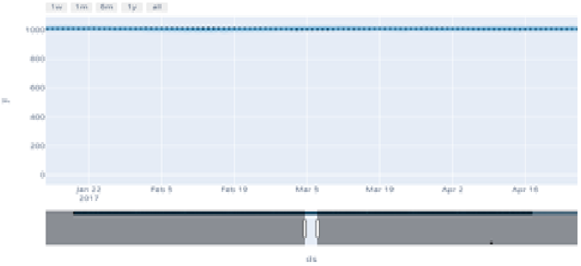


Fig. 14. Prophet-Sea Level Pressure

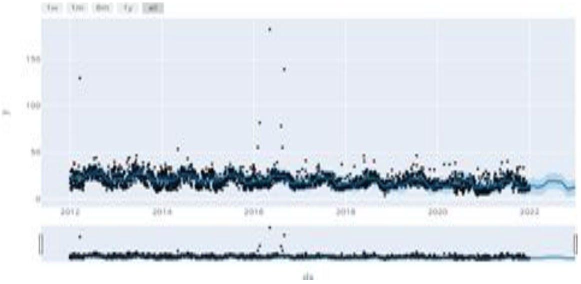


Fig. 15. Prophet-Wind Speed

After training and testing the model, Overall, we can say that Prophet is the algorithm that produces the best outcomes out of the three. Similar procedures were followed for other states.

## V. CONCLUSION

Everyday rain forecasting is essential since it has such a big impact on our daily life. The many varied atmospheric conditions influence rainfall. Because they are the ones that affect rain the most. So to create an accurate rain forecasting algorithm, this study analyzed daily data from India. To predict temperature, pressure, humidity, precipitate(rain), and wind speed, the time series algorithms ARIMA, Prophet, and Holt's winter exponential smoothing were utilized. The outcomes demonstrated that Prophet performed better than other algorithms and was the most accurate predictor of rain, which enhanced precipitation predictions. Our ability to forecast anything other than state-level statistics was limited by the size of the dataset. Additional features, like real-time updates and important alerts, can be provided on the website to better integrate the system with a river-specific flood prediction and alarm model.

## REFERENCE

[1] K. Abhishek, A. Kumar, R. Ranjan, and S. Kumar, "A rainfall prediction model using artificial neural network," in *2012 IEEE Control and System Graduate Research Colloquium*. IEEE, 2012, pp. 82–87.

[2] J. Joseph and R. T K, "Rainfall prediction using data mining techniques,"*International Journal of Computer Applications*, vol. 83, pp. 11–15, 12 2013.

[3] V. B. Nikam and B. Meshram, "Modeling rainfall prediction using data mining method: A bayesian approach," in *2013 Fifth International Conference on Computational Intelligence, Modelling and Simulation*. IEEE, 2013, pp. 132–136.

[4] D. Gupta and U. Ghose, "A comparative study of classification algorithms for forecasting rainfall," in *2015 4th International Conference on Reliability, Infocom Technologies and Optimization (ICRITO)(trends and future directions)*. IEEE, 2015, pp. 1–6.

[5] N. Hasan, N. C. Nath, and R. I. Rasel, "A support vector regression model for forecasting rainfall," in *2015 2nd international conference on electrical information and communication technologies (EICT)*. IEEE, 2015, pp. 554–559.

[6] J. Sulaiman and S. H. Wahab, "Heavy rainfall forecasting model using artificial neural network for the flood-prone area," in *IT Convergence and Security 2017: Volume 1*. Springer, 2018, pp. 68–76.

[7] S. Parashar and T. Hurra, "A study on prediction of rainfall using different data mining techniques," 05 2020.

[8] K. Sarda, "Rainfall predictions using data visualization techniques," 05 2022.

[9] K. Sarvani, Y. S. Priya, C. Teja, T. Lokesh, and E. B. B. Rao, "Rainfall analysis and prediction using machine learning techniques," 2021.

[10] V. C. Corporation, "Visual crossing weather (data query range)," https://www.visual crossing.com/, 2017-19.