e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
Volume:03/Issue:06/June-2021 Impact Factor- 5.354 www.irjmets.com

THE LITERATURE SURVEY ON RAINFALL PREDICTION USING MACHINE LEARNING TECHNIQUES

Ishwarya G[*1], Santhrupthi M B[*2], Shanthi B[*3], Varsha N[*4]

[*1,2,3]UG Student, Department Of Information Science & Engineering, VVIET, Mysore, Karnataka, India.

[*4]Faculty, Department Of Information Science & Engineering, VVIET, Mysore, Karnataka, India.

ABSTRACT

In this model mainly focused on predicting the rainfall in future using different machine learning algorithms. As India's economy significantly depends on horticulture, precipitation plays on important part. This model is very helpful for agriculture. In order to predict precipitation, an attempt is formed to a few of factual procedures and machine learning techniques to forecast and estimate meteorological parameters. For experimentation purpose daily observations were considered. We are using machine learning algorithms such as SVM , Linear regression methods for predicting the rainfall and also used to achieve crop yield prediction using rainfall, various machine learning algorithms such as Multiple linear regression, Decision tree, (ANN)Artificial neural networks, Support vector machine (SVM) have been used. In proposed system we are using SVM approach to analyze the datasets and also used to testing the dataset. Linear regression algorithm are used to trained the datasets for future prediction.

Keywords: Keywords: Precipitation, ARIMA , SVM, Linear Regression, Decision Tree, Holt Winter, Rainfall , Machine Learning, Random Forest.

I. INTRODUCTION

In India, where the bulk of agriculture business depends on precipitation as its standard wellspring of water , the time and measure of precipitation hold high importance and may impact the whole economy of the state. meteorology is one among the foremost challenge's issues seen by the planet, during a most up-to-date few century within the field of science and technology. As India's economy significantly depends on horticulture, precipitation plays on important part .The monthly climatic changes using spatiotemporal mining is being analyzed and therefore the variability in seasonal rainfall using the IMD data with many rain gage station information is completed by K. Chowdary. Cluster analysis technique is additionally performed using number of rainy days and rainfall because the input variable. Ingraining in has done a comparative study for rainfall prediction using different machine learning techniques on the north-eastern a part of Thailand. The paper shows that, how the feature selection are often wont to find the correlation between other weather parameter and therefore the rainfall, the paper also shows an equivalent day, Thai meteorological department (TMD) data is employed for experimental purpose. Attributes like temperature, humidity, pressure, wind, rain occurrence are used as input to the model. S.N Kohail . we has used daily historical data of the Gaza city and outlier analysis, prediction, classification, and clustering is completed for temperature prediction. Onset monsoon for the Indian sub-continent is predicted supported features extracted from the satellite image using data processing methods. KNN with Euclidean distance is employed for sea surface temperature (SST), cloud top temperature (CTT), cloud density, water vapour attributes were used. It predicts the onset monsoon beforehand 10-30 days is proposed. Production also plays an important role in it. At times the productivity of crops is moderate. As a result, the demand of food is increasing.

Modern technological advancement in the field of yield prediction using rainfall may also aid farmers in cost prediction using rainfall based upon the production. Mainly there are two categorizations in yield prediction using rainfall, classification and prediction using rainfall phase. In this paper we are presenting an outline on classification and prediction using rainfalls techniques. Prediction using rainfalls can be obtained by examining large sets of pre-

existent databases in order to generate new knowledge. Thus, resulting in an automatic and approximate prediction using rainfalls. Various steps that can be involved in crop yield prediction using rainfall are data acquisition, data pre-processing, feature selection, classification and prediction using rainfall.

II. RELATED WORK

"Nishchala C Barde and Mrunalinee Patole [1]. Classification and forecasting of weather using ann, k-nn and naive bayes algorithms ". In This paper, here some algorithms are used to predict the rainfall. K-NN, ANN, and Naïve bayes algorithms are used to predict the weather.

SML Venkata Narasimhamurthy & et al [2]. "Rice Crop Yield Forecasting Using Random Forest Algorithm". Highest accuracy of 85.89% is obtained by using this method. In this paper gives good accuracy for crop prediction using rainfall prediction model.

Dr. Bharat Mishra & et al [3]. "Soybean Productivity Modelling using Decision Tree Algorithms". Decision tree is being converted to classification rules using IF-THEN- ELSE. In this model tries to give good prediction for crops.

Here data mining techniques are used to building the model.

K Chowdary, R Girisha, and KC Gouda [4] . "A study of rainfall over India using data mining". Some datamining techniques are used to predict the output in future.

Pinky Saikia Dutta and Hitesh Tahbilder [5]. "Prediction of rainfall using data mining technique over assam ".

This paper done based on datamining techniques are using to trained the model for rainfall. and some techniques are used such as statistical techniques , multilinear regression . this model predicts rainfall based on month in assam.

Arun Kumar & et al [6]. "Efficient Crop Yield Prediction using rainfall Using Machine Learning Algorithms". in this paper includes rainfall prediction And the range of productivity will be defined.

Kawsar Akhand & et al [7]. "Yield Prediction using rainfall in Bangladesh Using Satellite Remote Sensing Data and Artificial Neural Network". In this machine learning algorithms are used to test the datasets and also used for predicting the system.

P. Priya*1 & et al [8]. "Predicting the Crop Yield Using Machine Learning Algorithm". in this paper having dataset cleaning, testing are done using machine learning algorithms and Decision tree is used for classification purposes.

III. PROPOSED SYSTEM

In the primary a part of the proposed model retrieved weather data is cleaned and reordered, then the rainfall data is categorized into different categories consistent with IMD guidelines. the info is partitioned into two parts 70% for training and 30% for testing. From the study, it's found that each one four parameters have significant importance with the rainfall. All past year's maximum temperature and minimum temperature were retrieved except last year. supported the past data six different forecasting methods (Holt winter method, ARIMA model, Simple Moving Average model, Neural Network method, Seasonal Naive method) were applied and therefore the best-fitted model output was taken into consideration. this method compared to an immediate forecast of the individual. In the fusion part, four forecast parameters are given as input to the trained data (1979 to 2013). Based on this input parameters next year and next monsoon season rainfall is forecasted. The individual accuracy of the model was also analyzed with confusion matrix. For the experimental purpose we have taken only Jun to Dec data because in most of regions of India rainfall occurs in this period. Considering the forecast for whole year gives higher accuracy as there are more no. of non rainy days which gets correctly classified but our focus is to predict the rainfall for those months who have chances of rainfall.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
Volume:03/Issue:06/June-2021 Impact Factor- 5.354 www.irjmets.com

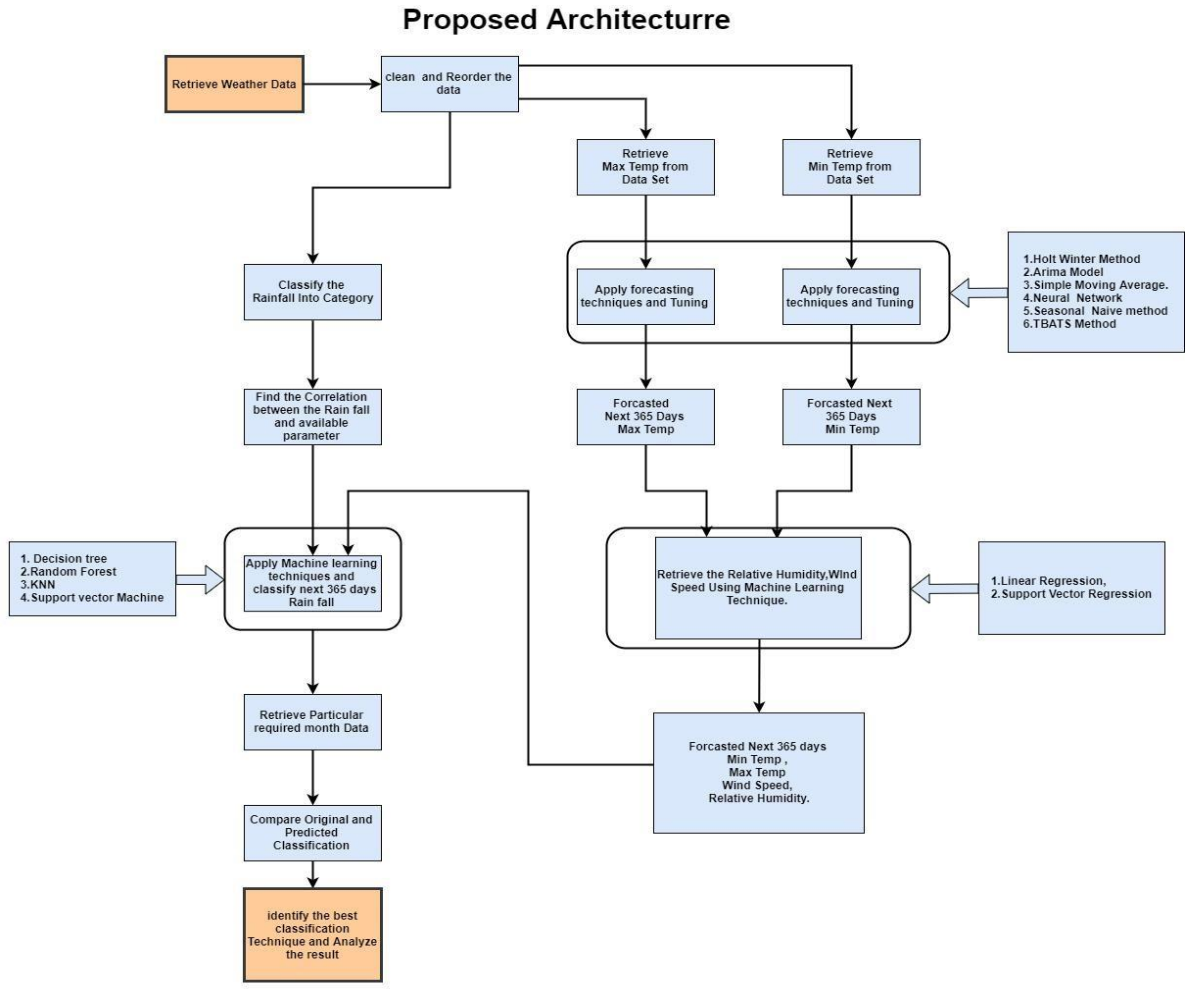

Figure 1: proposed architecture.

IV. RESULTS AND DISCUSSION

This sections includes the information about the data which is used for the experimentation along with the results of the forecasting and machine learning methods with the detailed explanation of tuned parameters. The detailed analysis of the best-fitted model and comparison of all methods based on performance is done. The data for the experimental purpose is retrieved from global weather site and it is provided by National Centers for Environmental Prediction (NCEP). For experimentation, daily data from 1/1/1979 to 7/31/2014 is collected from five locations. Data also contains parameters like minimum temperature, maximum temperature, relative humidity, wind speed, and precipitation. The rainfall is classified into seven categories according to the forecast manual provided by the Indian Meteorological Department (IMD).Save As command, and use the naming convention prescribed by your conference for the name of your paper.

A. Forecasting Parameters

Forecasting Maximum Temperature: As the temperature is comparatively easy to forecast compare to other meteorological parameters. We have forecasted maximum temperature using different forecasting techniques and RMSE (Root Mean Square Error) were compared with the original data set. Table 1 shows, the 365 days forecasted maximum temperature error. The result shows that in the case of maximum temperature ARIMA model performs better than the other model. Arima (3,0,4) is the best-fitted model. Forecasting Minimum Temperature: Various forecasting method for fore-casting the minimum temperature were analyzed, neural network show significant low RMSE compared to the other model. NNR(30,1,16)[365] performs the best fit model. Average of 20 networks, each of which is a 31-16-1 network with 529 weights options were -linear output units. Estimated sigma2 =0.01786

Table 1: Forecasted Maximum, Minimum Temperature RMSE

Method	RMSE(C)
ARIMA	3.45
TBATS Model	3.53
Naïve Method	4.33
Moving Average	6.93
Neural Network	8.66
Holt Winters Additive	17.47
Holt Winters Multiplicative	13.57

Forecasting Relative Humidity: As the correlation between relative humidity and rainfall is significant 0.303. We have also forecasted relative humidity. We have used minimum temperature and maximum temperature as the input to the model and predicted relative humidity. Forecasted minimum and maximum temperature were given as the input instead of the measured temperature to get the final model accuracy. The result shows that support vector regression which is a combination of linear regression and support vector machine works best. Forecasting Wind Speed: Wind speed is one of the important parameters for predicting the rainfall as its correlation with the rainfall is 0.49. It is also important to forecast the wind speed(m/s). We have also applied the two regression techniques for predicting the wind speed giving two input parameters minimum temperature and maximum temperature, as a result support vector regression gives less RMSE compare to simple linear regression. For papers with more than six authors: Add author names horizontally, moving to a third row if needed for more than 8 authors.

Table 2: Forecasted Relative Humidity and Wind Speed RMSE

Forecasted Relative Humidity		Forecasted Wind Speed	
Method	RMSE(Fraction)	Method	RMSE(m/s)
Linear Regression	0.75	Linear Regression	0.1345
Support Vector Regression	0.68	Support Vector Regression	0.1116

B. Machine Learning Model

KNN Method: To identify the best k nearest neighbor, we have tried with different values of K. The study reveals that k=15 gives best classification accuracy for the 1-year forecast, and k=9 gives best classification accuracy for June to December month forecast. Confusion matrix shows that very heavy rain classic to none. Results also show the considerable accuracy for the no rain, very light rain, moderate rain. For the very heavy rain, heavy rain and rather heavy rain results were not impressive. Decision Tree: In this method, we have used Gini index algorithm for the selection of the most homogeneous node. Higher the value of Gini higher the homogeneity and based on that decision tree is generated.

Method	RMSE(C)
ARIMA	3.05
TBATS Model	3.57
Naïve Method	3.38
Moving Average	7.92
Neural Network	2.55
Holt Winters Additive	7.19
Holt Winters Multiplicative	7.14

The process of pruning is also done in order to limit the level of the tree. To ensure that tree is not overfitted or underfitted we have also tuned tree. For level 5, it shows the best result, to avoid overfitting, we have taken only up to 5 level.10-fold cross validation is done on this data set for measuring the accuracy of the model. Results were also analyzed by confusion matrix. It is found that unlike the KNN, this method has classified very heavy rain. But same as the case in KNN it only shows the considerable accuracy for the no rain, moderate rain and for very light rain. Support Vector Machine: In order to give best classification accuracy different combination of kernels, gamma, C values were tried for the tuning purpose. Radial base function kernel, linear kernel, sigmoid kernel were given for kernel parameter, different gamma values and C values were also given. It is found that linear kernel with gamma value 0.1 and C value 1 gives best accuracy compared to others. From the confusion matrix, it is found that SVM is unable to classify Heavy Rain and Very Heavy Rain. For even light rain and for very light rain results were poor. In the experimentation we have taken more number of classes to classify the rainfall, but as SVM works best with optimal margin, there may be the case that multiple category overlap each other and because of which SVM performs worst compare to others.

Random Forest: Random forest is a tree based model, it is a collection of many tree models. We have applied different tuning parameters for tuning it. As in random forest case, one of the parameters is how many trees should be used to get the more accurate results. It works well with high variance low bias models. It is noticed that after 250 number trees error rate is constant. So, we will restrict number of trees to 250 in the forest. From the confusion matrix, it is found that for very light rain Random forest method gives the best accuracy. It also performs well for the no rain, moderate rain, and for light rain.

C. Accuracy Measurements and Analysis

Table 3 is the result of 70% training and 30% training data set. It shows that random forest out performs compared to another method. For the experimental purpose, we have given actual Real time values of Maximum Temperature, Min temperature, Relative Humidity, Wind Speed as an input to the trained model and analysis is done on 30% testing data set. But as we want to forecast the rainfall it is necessary to give forecasted maximum temperature, minimum temperature, relative humidity and wind speed values as an input parameter to the trained model. This forecasted parameters also have their own error so if we put the forecasted value as input parameter to this classification technique there are chances to decrease the final accuracy of the model. As the random forest gives the best accuracy we have shown the final con-fusion matrix for the random forest only (for Jun to Dec). Table 5 shows the confusion matrix for the random forest. Diagonal shows the correctly classified category. It is found that it shows good accuracy for No rain, moderate rain, very light rain, light rain. Figure 2 shows the ROC (Receiver Operator Characteristics) Curve analyzed through comparing results of the different methods, category wise.

Table 3: Accuracy on 30% test data.

Method	AUC (Area under curve)	Classification Accuracy	Precision	Recall
KNN	0.873	0.721	0.691	0.721
Tree	0.755	0.721	0.716	0.721
SVM	0.684	0.539	0.659	0.539
Random Forest	0.914	0.762	0.744	0.76

Table 4: Final Accuracy Comparison on Forecast

Method	Final Accuracy (1 Year-365 days)	Final Accuracy (Fr June To Dec)
Decision Tree	69.58	61.21
Random Forest	70.50	70.09
KNN	69.31(K=15)	66.35(K=9)
SVM	67.05	69.15
Neural Network	68.49	68.69

Random forest uses their own sample of training data i.e. there are some observations which might appear several times in the sample. The final prediction is based on voting by each tree in the forest. Random Forest is characterized by their efficiency to deal with large data set, relatively robustness for outliers and noise and ability to deal with highly correlated predictor variables. axis against False Positive Rate (1-Speci city) on x-axis for each categories .

Table 5 shows the final classification accuracy of each method with forecasted parameters as an input to the trained model. In a country like India, where rainfall occurs in only limited no. of the month. So for that, we have also analyzed our accuracy for monsoon season and it is noticed that it gives considerable classification accuracy.

Table 5: Confusion Matrix for the Random Forest (for Jun to Dec)

	Actual(days)						
Predicted(days)	Heavy Rain	Light Rain	Moderate Rain	No Rain	Rather Heavy	Very Heavy Rain	Very Light Rain
Heavy Rain	0	0	0	0	0	0	0
Light Rain	0	12	4	0	1	0	8
Moderate Rain	0	0	33	0	1	1	9

No Rain	0	0	0	74	0	0	5
Rather Heavy	1	0	0	0	1	0	0
Very Heavy Rain	0	0	0	0	0	0	0
Very Light Rain	0	11	2	12	0	0	30

V. CONCLUSION

The proposed work is an attempt to forecast rainfall using a fusion of different machine learning and forecasting techniques. Even though the rainfall is dependent on many parameters, we are able to get impressive classification accuracy using limited parameters. It is also found that even after classifying rainfall into eight different categories, we are getting acceptable accuracy. Validations for forecasted parameters are done using RMSE measure. Empirical results show ARIMA for maximum temperature, Neural Network for minimum temperature and SVR for relative humidity and wind speed works best. Validation of classification is measured through accuracy, precision and recall. ROC curve for all classifiers shows random forest works best for rainfall classification.

As rainfall is dependent on the various parameters it is also required to study how other meteorological parameters affect the Rainfall prediction. We can also perform the same exercise on hourly data using various parameters to forecast next hour rainfall. A study can also be done using more observations for particular region or area, and design this kind of model on big data framework so that computation can be faster with higher accuracy.

VI. REFERENCES

[1] Nishchala C Barde and Mrunalinee Patole. Classification and forecasting of weather using ann, k-nn and na•ve bayes algorithms.

[2] Debasish Basak, Srimanta Pal, and Dipak Chandra Patranabis. Support vector regression. Neural Information Processing-Letters and Reviews, 11(10):203{224, 2007.

[3] Leo Breiman. Random forests. Machine learning, 45(1):5{32, 2001.

[4] KK Chowdari, R Girisha, and KC Gouda. A study of rainfall over india using data mining. In Emerging Research in Electronics, Computer Science and Technology (ICERECT), 2015 International Conference on, pages 44{47. IEEE, 2015.

[5] Pinky Saikia Dutta and Hitesh Tahbilder. Prediction of rainfall using data mining technique over assam. IJCSE, 5(2):85{90, 2014.

[6] G Gregoire. Multiple linear regression. European Astronomical Society Publications Series, 66:45{72, 2014.

[7] Rob J Hyndman. Moving averages. In International Encyclopedia of Statistical Science, pages 866{869. Springer, 2011.

[8] Elia Georgiana Petre. A decision tree for weather prediction. BULETINUL UniversitaNii Petrol{Gaze din Ploiesti, pages 77{82, 2009.

[9] Narasimha Prasad, Prudhvi Kumar, and Naidu Mm. An approach to prediction of precipitation using gini index in sliq decision tree. In Intelligent Systems Modelling & Simulation (ISMS), 2013 4th International Conference on, pages 56{60. IEEE, 2013.

[10] Sirajum Monira Sumi, MFaisal Zaman, and Hideo Hirose. A rainfall forecasting method using machine learning models and its application to the fukuoka city case.