PERFORMANCE ANALYSIS OF RAINFALL PREDICTION USING MACHINE LEARNING ALGORITHMS

Karumuri Sai Krishna Shamita, Gandham Sri Ganga Prudvi Sai Chiru, Kandregula Hindu Harshini,Vindula Basanthi, Mrs.R. Uma Maheswari (HOD of Dept) Department Of Electronics And Computer Engineering, Vignan's Institute of Information Technology, Visakhapatnam-India

Abstract

Predicting rainfall is critical because heavy rains are capable of causing many disasters. Agricultural damage, landslides, and flooding are all hazards associated with heavy rainfall. Predictions aid in preventing disasters and even more importantly, are more accurate if they are accurate.The rain can also cause infrastructure damage and human fatalities. Weather conditions threaten farmers severely. The climatic variability for an area is appertained to the long term change in rain fall, temperature, moisture, evaporation, wind speed and other meteorological parameters. Machine Learning algorithms are also used in predicting rainfall. This paper comprises of algorithms like Logistic Regression, Decision tree and Random forest classifier. The dataset consists of various meteorological parameters.

Keywords: Rainfall, Random Forest, Decision Tree,Logistic Regression

Introduction

India's wealth is husbandry. Heavy rains cause damage to crops, property damage, and many other losses due to their intense impact. It is important to forecast rainfall. It is therefore crucial to develop a better prediction model to provide an early warning system that reduces risk of human life and property risks. It also helps with water coffers. Rainfall information in the history will help to build the model which is further helpful to cultivators in managing their crops. It has farther lead to the substance of determining whether the trend is adding or dwindling. The changes in the most important climatological parameter i.e rainfall, may be responsible for the natural disasters like flooding and landslides. Timely and accurate prognostications can helps us to reduce mortal and fiscal loss. Agriculture is that the crucial point for survival. For husbandry, rainfall is most important. Vaticination of rainfall gives mindfulness for the people and know in advance and take certain preventives to cover their crop from heavy rainfall.

Review of Literature

In Literature , many authors have contributed for this study. S.Poornima et.al[1] proposed a technique to predict rainfall with the help of Feed forward neural network. It is simplest form of Artificial Neural Network (ANN).The proposed work is based on various parameters which are used in prediction. CMAK Zeelan Basha et.al[2] used a downfall vaticination system using algorithms of machine learning. Major algorithms are ARIMA Model, Support Vector Machine (SVM), ANN, Logistic Regression and Self Organizing Map. In this architecture the accuracy is measured using Root Mean Square Error(RMSE) and Mean Square Error(MSE). R.Kingsy Grace et.al[3] introduced a rainfall vaticination system using Multiple linear regression. To validate the proposed model, Mean Square Error (MSE), delicacy, and correlation are used, plus the input data includes various meteorological factors to provide more precise predictions. Urmay Shah et.al[4] used four algorithms i.e Decision Tree, Random Forest, K-Nearest Neighbour on the partitioned data and observed that Random forest characterized the efficiency to deal with larger data sets and ARIMA model shows empirical results for maximum temperature. B.Vasantha et.al [5] proposed an approach to use convolution neural network the system. Sunil Kaushik et.al[6] proposed three techniques of machine learning for the vaticination. The techniques were compared on the parameters that measure the performance. The techniques used are Extreme Learning Machines, KNN,SVM. Suhaila Zainudin et.al[7] analysed different classifiers like Naïve Bayes, Support Vector Machine, Decision Tree, Neural Network and Random Forest to predict the rainfall. Chandrasegar Thirumalai et.al[8] introduced an approach to predict the rainfall using linear regression. This paper calculates various categories of data by the mentioned approach for better understanding purpose.Moulana Mohammed et.al[9] used three prediction models Lasso Regression,Support Vector Regression and Multiple Linear Regression.

Copyright @ 2022 Authors

Pengcheng Zhang et.al[10] proposed a model to improve forecasting accuracy.The approach is proposed using a multi-layer perceptron i.e a dynamic regional combined approach.

Research Methodology

The proposed system model is depicted in the Fig 1. It is the block diagram of the system model.

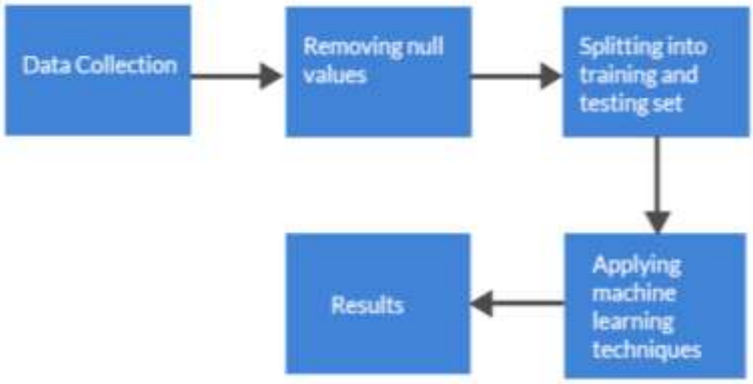
Data Collection
Removing null values
Splitting into training and testing set
Applying machine learning techniques
Results

Fig.1 System model

According to this system model it is known that the first step is data collection i.e the factors required to predict rainfall are collected into a dataset. There will be inconsistenices and null values in the dataset such values are dropped that influence the accuracy of the model.The data is splitted further for the specified purpose.

A. Data Collection

In this research, a dataset is taken which contains the data records. The dataset consists of 23 columns out of which 2 columns are date and location , remaining columns are meteorological parameters.

1. Date
2. Location
3. MinTemp
4. MaxTemp
5. Rainfall
6. Evaporation
7. Sunshine
8. WindGustDir
9. WindGustSpeed
10. WindDir9am
11. WindDir3pm
12. WindSpeed9am
13. Windspeed3pm
14. Humidity9am
15. Humidity3pm
16. Pressure9am
17. Pressure3pm
18. Cloud9am
19. Cloud3pm
20. Temp9am
21. Temp3pm
22. RainToday
23. RainTomorrow

RainTomorrow is the target variable in the collected data.

B.Removing Null Values

Removal of null values is a necessary step. A machine learning algorithm that uses null values will perform poorly and produce inaccurate results. In other words, the dataset should be cleaned up of null values before applying any machine learning algorithm. Null values in the dataset influence the results and accuracy of the model. The columns consisting of more number of null values are dropped . In the dataset six columns have more number of null values hence they are dropped. The six columns are Evaporation, Sunshine, Cloud9am , Cloud3pm, Location and date.

C.Training and Testing

Training is one of the most important aspect in machine learning and and helps to make accurate predictions or perform a desired task. The model is trained using the training data and helps in understanding various hidden features of the data. The model is trained using independent variables.The model's performance is validated using validate set.The Model's performance is evaluated. The test set is different from train set where the model is tested using the test set. Data is splitted into training dataset and testing dataset.

RainTomorrow is a testing dataset variable that is dependent on the independent variables. Other than RainTomorrow rest all factors are independent variables.

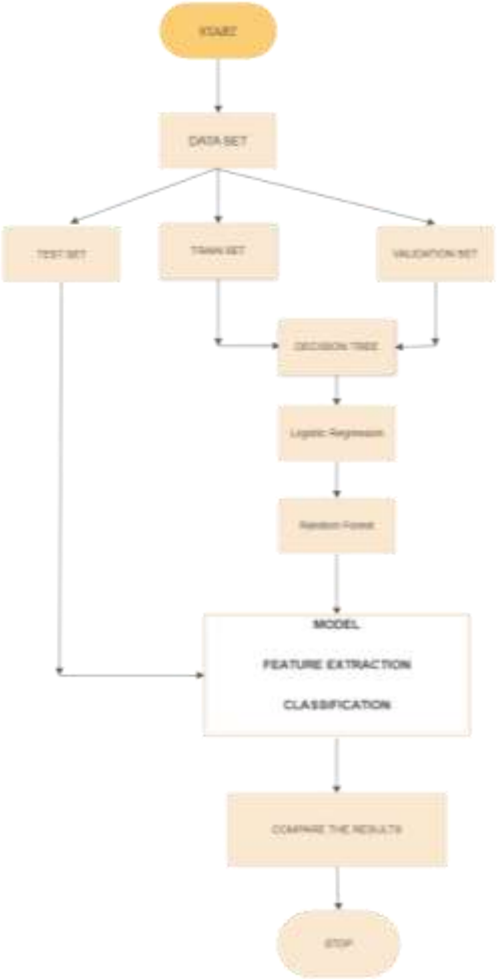

Fig.2 Flow diagram of proposed model

Fig2 depicts the flow diagram of the proposed model used in this research.

D.Machine Learning Algorithms used

The three machine learning algorithms used in this research work are Decision Tree , Logistic Regression, Random Forest Classifier. The three techniques are used in predicting the rainfall with the help of parameters present in the dataset which is further cleaned from the null values.

Results and Discussion

Three supervised machine learning algorithms are used to prognosticate the delicacy of model. The three algorithms are

a.Decision Tree
b.Logistic regression
c.Random Forest Classifier

Decision Tree:

In supervised learning, Decision Trees represent, based on given conditions, all possible solutions to a problem/decision.A decision tree can contain categorical data (yes/ no) as well as numeric data.By using Decision Tree ,the accuracy predicted result is 78.8%.

Logistic regression:

The log-regression algorithm utilizes a set of independent variables to predict a categorical dependent variable.By using Logistic Regression , the accuracy predicted result is 84.5%

Random Forest:

Classification and regression problems are often solved with Random Forests. Based on the majority vote of different samples, it builds decision trees for classification and regression, respectively. Nonetheless, this method is supervised machine learning. By using Random Forest Classifier, the accuracy predicted result is 85.4%

•Scatter plots

Scatter plot is used to determine whether or not two variables have a relationship or correlation. For visualization a scatter plot is drawn using seaborn. The library named seaborn is used for making graphs.

The below two scatter plots graphically represent the relationship between the parameters such as a.MaxTemp vs MinTemp and b.Humidity9am vsTemp9am and

Two paramters namely Humidity9am and Temp9am are taken as xlabel and ylabel respectively. MaxTemp and MinTemp are taken as xlabel and ylabel respectively.

TABLE 1. Few Data Considered For Scatter Plot Fig.3

S.NO	MIN TEMP	MAX TEMP	RAIN FALL TOMORROW
1.	13.4	22.9	NO
2.	9.7	31.9	YES
3.	20.1	32.7	NO
4.	19.7	27.2	YES
5.	20.8	30.6	NO
6.	16.4	27	YES
7.	9.3	28	NO
8.	5.9	11.1	YES
9.	0.9	16.4	NO
10.	6.9	11.2	YES

Copyright @ 2022 Authors

The data taken in the TABLE 1 is a set of few values from the dataset which are used in drawing the scatter plot Fig3

a.Scatter plot for MaxTemp vs MinTemp

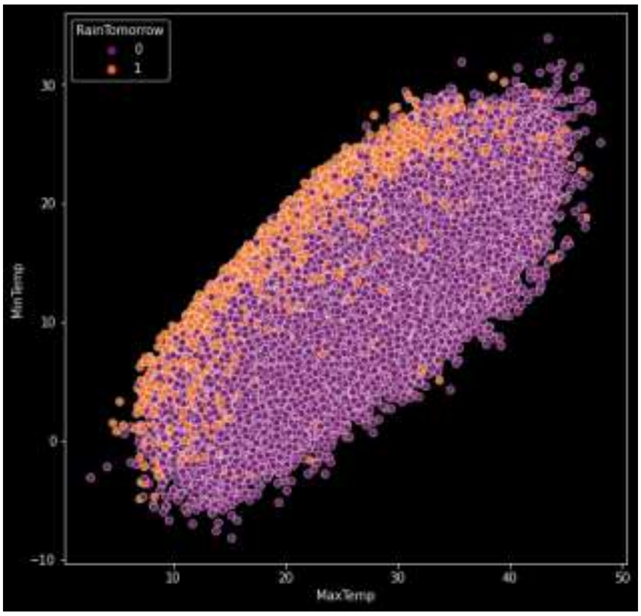

Fig.3 Scatter plot for MaxTemp vs MinTemp

In this x data , maximum temperature is considered while in the y data minimum temperature is considered. Hue is rain tomorrow while the palette is inferno and the data is dataset which is considered. From the plot it is observed that as the minimum temperature is increasing, maximum temperature is also increasing. It is a cluster but it follows linear relationship.

b.Scatter plot for Humidity9am vs Temp9am

TABLE 2. Few Data Considered For Scatter Plot Fig.4

S.NO	Humidity9am	Temp9am	RAIN FALL TOMORROW
1.	38	21	NO
2.	42	18.3	YES
3.	58	20.1	NO
4.	48	20.4	YES
5.	65	15.8	NO
6.	69	18.1	YES
7.	47	15.5	NO
8.	49	21.6	YES
9.	78	12.5	NO
10.	60	26.1	YES

The data taken in TABLE 2 is a set of few values taken from the dataset which are used in drawing the scatter plot Fig4. Another plot is drawn for Humidity9am and Temp9am i.e x data is Humidity9am and y data is Temp9am. Hue is Rain Tomorrow and palette is inferno while data is the dataset which is considered. From the graph is observed that as the humidity9am increases, the temp9am increases the probability of rain tomorrow is also increasing.

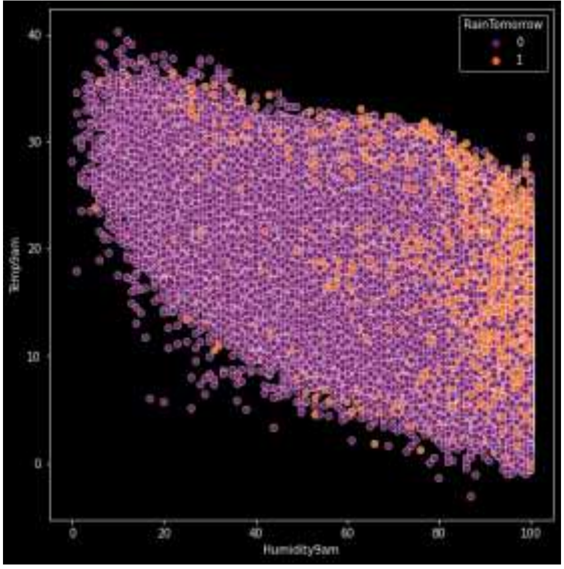

Fig.4 Scatter plot for Humidity9am vs Temp9am

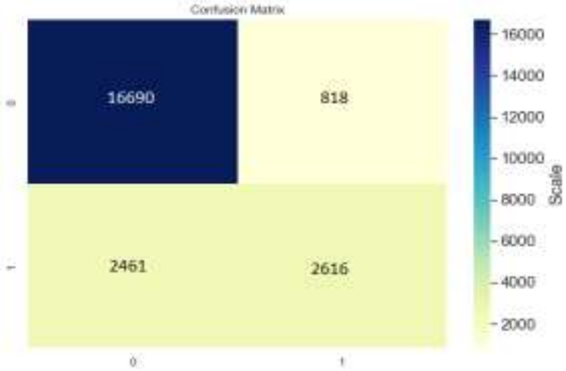

Fig.5 Confusion Matrix for Random Forest Classifier

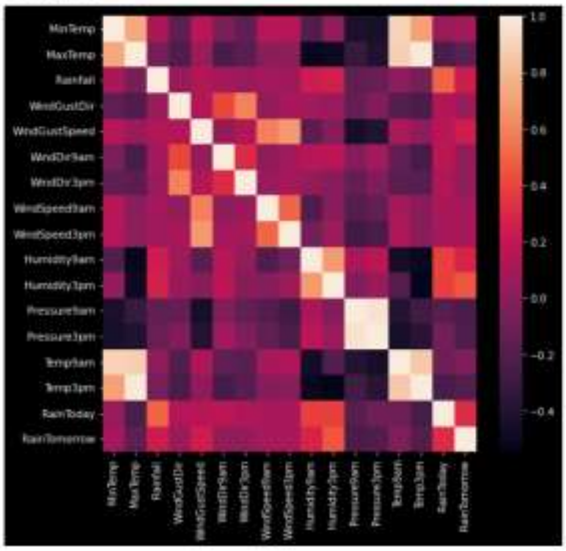

Fig.6 Heatmap

Copyright @ 2022 Authors

From Fig.5 the confusion matrix of random forest classifier is observed. As the name implies, a confusion matrix is a matrix that is helpful in measuring the performance of classification model on dataset i.e set of test data.

From Fig.6 the heatmap is observed. Heatmap helps in drawing correlation among each and every column. In the heatmap lighter the color , better the relationship between columns/parameters/factors. It is observed that Rain Tomorrow has good relationship with Humidity9am , Humidity 3pm , Rain Today and few other parameters.

TABLE3. Accuracy Of Three Algorithms

ALGORITHMS	ACCURACY
Random Forest	85.4%
Logistic Regression	84.5%
Decision Tree	78.8%

Table3 depicts the comparison of accuracies of the three algorithms used in the proposed model in this paper

Conclusion

In this Research paper three of Machine Learning algorithms such as Decision tree, Logistic regression and Random Forest Classifiers are used to predict the rainfall based on the parameters. In spite of having many inconsistencies and null values in datasets, various feature transforms have been used to improve the consistencies of datasets . By using Random Forest Classifier, the accuracy obtained is 85.4% . The accuracy predicted for model using logistic regression is 84.5%. The accuracy i.e 78.8% is obtained on Decision tree algorithm .The results impersonate that the Random forest classifier is achieving the highest accuracy score and the least accuracy score is obtained on Decision tree. Scatter plots are included to represent the relationship between the meteorological parameters.

References

[1] Poornima, S., Devi, S., Oviya, D., Taj, A. S., & Elakkiya, G. T. (2020)”. Prediction of Rainfall using Machine Learning.” In International Journal of Recent Technology and Engineering (IJRTE) (Vol. 9, Issue 1, pp. 1374–1377)

[2] Basha, Cmak Zeelan, Nagulla Bhavana, Ponduru Bhavya, and V. Sowmya. "Rainfall prediction using machine learning & deep learning techniques." In 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC), pp. 92-97. IEEE, 2020.

[3] Grace, R. Kingsy, and B. Suganya. "Machine learning based rainfall prediction." In 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS), pp. 227-229. IEEE, 2020.

[4] Shah, Urmay, Sanjay Garg, Neha Sisodiya, Nitant Dube, and Shashikant Sharma. "Rainfall prediction: Accuracy enhancement using machine learning and forecasting techniques." In 2018 Fifth International Conference on Parallel, Distributed and Grid Computing (PDGC), pp. 776-782. IEEE, 2018.

[5] Vasantha, B., and R. Tamilkodi. "Rainfall pattern prediction using real time global climate parameters through machine learning." In 2019 International Conference on Vision Towards Emerging Trends in Communication and Networking (ViTECoN), pp. 1-4. IEEE, 2019.

[6] Kaushik, Sunil, Akashdeep Bhardwaj, and Luxmi Sapra. "Predicting Annual Rainfall for the Indian State of Punjab Using Machine Learning Techniques." In 2020 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN), pp. 151-156. IEEE, 2020.

[7] Zainudin, Suhaila, Dalia Sami Jasim, and Azuraliza Abu Bakar. "Comparative analysis of data mining techniques for Malaysian rainfall prediction." Int. J. Adv. Sci. Eng. Inf. Technol 6, no. 6 (2016): 1148-1153.

[8] Thirumalai, Chandrasegar, K. Sri Harsha, M. Lakshmi Deepak, and K. Chaitanya Krishna. "Heuristic prediction of rainfall using machine learning techniques." In 2017 International Conference on Trends in Electronics and Informatics (ICEI), pp. 1114-1117. IEEE, 2017.

[9] Mohammed, Moulana, Roshitha Kolapalli, Niharika Golla, and Siva Sai Maturi. "Prediction of rainfall using machine learning techniques." International Journal of Scientific and Technology Research 9, no. 01 (2020): 3236-3240.

[10] Zhang, Pengcheng, Yangyang Jia, Jerry Gao, Wei Song, and Hareton Leung. "Short-term rainfall forecasting using multi-layer perceptron." IEEE Transactions on Big Data 6, no. 1 (2018): 93-106.

[11] Khan, Mohd Imran, and Rajib Maity. "Hybrid deep learning approach for multi-step-ahead daily rainfall prediction using GCM simulations." IEEE Access 8 (2020): 52774-52784.

[12] Appiah-Badu, Nana KA, Yaw M. Missah, Leonard K. Amekudzi, Najim Ussiph, Twum Frimpong, and Emmanuel Ahene. "Rainfall Prediction Using Machine Learning Algorithms for the Various Ecological Zones of Ghana." IEEE Access (2021).

[13] Liu, James NK, Bavy NL Li, and Tharam S. Dillon. "An improved naive Bayesian classifier technique coupled with a novel input solution method [rainfall prediction]." IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews) 31, no. 2 (2001): 249-256.

[14] Ahmed, Zanyar Rzgar. "RAINFALL PREDICTION USING MACHINE LEARNING TECHNIQUES." PhD diss., NEAR EAST UNIVERSITY, 2018.

[15] Sai Tarun, G. B., J. V. Sriram, K. Sairam, K. T. Sreenivas, and M. V. B. T. Santhi. "Rainfall prediction using machine learning techniques." Int J Innovative Technol Exploring Eng 8, no. 7 (2019): 957-963.