International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 08 Issue: 08 | Aug 2021 www.irjet.net p-ISSN: 2395-0072

Weather Forecasting Using Data Mining

Harsha Dessai[1], Siddhi Naik[2]

[1]Student, Department of Information Technology and Engineering, Goa Engineering College, Ponda Goa, Goa, 403401, India
[2]Professor, Department of Information Technology and Engineering, Goa Engineering College, Ponda Goa, Goa, 403401, India

---

Abstract - *India is an agricultural country and therefore, a large number of Indians depend on weather and other climatic conditions. Weather forecasts include rain forecast, wind forecast, cloud, fog, and lightening. It is a challenging task to predict the weather due to the powerful and unpredictable nature of weather data. Climate forecasting has become increasingly important as more and more lives on earth are subject to climate change. In this study, different classification methods were applied to predict the rainfall data in Panjim Goa, India. The predicting model uses four different data mining algorithms namely Naïve Bayes, K-Nearest Neighbor, Classification and Regression Tree (CART), and Random Forest Classifier. The main motive of the project is to implement a method to forecast future rainfall in the city of Goa for the next few days using the patterns of past dataset from a single weather station measurement. The method used to carry out this forecasting is Linear Regression.*

Key Words: Rain forecasting, Naïve Bayes algorithm, K-nearest neighbor, Random Forest, CART, weather patterns, Linear Regression.

1. INTRODUCTION

Predicting future weather patterns based on historical data is called Weather Prediction. Forecasting weather is an application of science technology to predict the atmospheric state for a given location and for future time. Climate speculation has been the subject of much research as climate change directly affects people. The process of weather forecasting is complex and challenging because it depends on a variety of factors. Knowing the weather conditions in advance helps us in many ways to minimize losses. Traditional methods of weather forecasting using satellite imagery and weather channels are expensive because they involve more expensive and sophisticated methods. Weather forecasting using a data mine is less expensive, less time consuming, easier and more realistic, more accurate in nature. Forecasting weather patterns using a machine learning process involves using a lot of past weather data. It is therefore very important that any machine learning model is trained with the most accurate details. Information obtained from various sources is not always reliable and therefore data should be processed in advance. Preliminary data processing includes removing unnecessary columns from the model prediction, removing zero values, merging the same columns following many other pre-processing steps. The Department of Environment around the world is putting a lot of effort into climate research sites. The historical data used in this project is collected from NOAA (National Ocean Atmospheric Administration) [21]. This data has 15 indicators, from which a few attributes are used to predict rainfall. To facilitate the operation of the K-NN, Naïve Bayes, Random Forest, and CART models of rain forecast, pre-data processing and transformation is done using factor analysis and regression analysis techniques.

The rest of the paper is organized as follows: Section II provides a survey on several related work done in this field. The proposed methodology is described in Section III. Section IV gives the details of the dataset and the experimental results obtained while evaluating the proposed approach. Finally, Section V concludes the paper and provides a brief outline of the possible future work.

2. LITERATURE SURVEY

Most of the approaches for weather forecasting use data mining for future prediction. The author in paper [1] worked on data collected from various stations of Malaysia and performed a comparative analysis using different data mining algorithms. In [2] several algorithms such as Naive Bayes, K- Nearest Neighbor algorithm, Decision Tree, Neural Network and fuzzy logic algorithms are compared. The research in [3] performed rainfall prediction in Lahore city using five data mining techniques with 12 years of past weather data from December 1, 2005 to November 31, 2017, is used for prediction in this research. Three accuracy measures such as precision, recall and f-measure are used to perform analysis of data mining techniques. In paper [4] Predictive Decision Tree model, Artificial Neural Network model and Naïve Bayes model are developed for rainfall prediction and comparison and the results show that the decision tree model performed better compared to other predictive models. The author in [5] employed a deep learning architecture to predict the accumulated rainfall for the next day. Two networks namely an auto-encoder

network and a multilayer perceptron network is used in the architecture. The study in paper [6] aims to provide a comparative analysis of the multiple machine learning classifiers for rainfall prediction based on Malaysian data. Several classifiers were explored which are Naïve Bayes (NB), Support Vector Machine (SVM), Decision Tree (DT), Neural Network (NN) and Random Forest (RF).The most effective classifier was the neural network as shown by the analysis. In [7] an improved KNN algorithm has first been proposed. The author in [8] represents a new prediction model for rainfall-induced shallow landslides with application for the tropical hilly area of the Quy Hop area, central of Vietnam. DE optimization and fuzzy K-NN algorithm is combined to from the proposed model. Performances of three artificial intelligence techniques such as K-nearest neighbor (KNN), artificial neural network (ANN), and extreme learning machine (ELM) were evaluated in [9] for the prediction of summer monsoon and post-monsoon rainfall in Kerala India. By using different statistical tests the performance of the aforementioned approaches has been gauzed.

A deep learning based model is designed in paper [10] in order to predict the rainfall using several machine algorithms like SVM, random forest and decision tree. Several algorithms such as Naïve Bayes and the C4.5 (J48) Decision Tree algorithm are simultaneously developed with a database containing weather data collected of 2 years in paper [14] and was found that the performance of the C4.5 (J48) algorithm of the decision tree was significantly better than that of the Naïve Bayes.

In [15] the author focused on historical weather data collected locally from the city of Faisalabad and carried analysis and using data mining techniques by assessing climate change patterns including high temperatures, low temperatures, wind speed, rainfall. In [16] the proposed data model includes the Hidden Markov Model for predicting and issuing weather forecasts using a collection of K methods. Predicting new or future conditions the system needs to adapt to current weather conditions.

The different data mining techniques used for weather predictions are studied and explained in [17] by the authors. Some of the methodologies includes Artificial neural network(ANN), supervised and unsupervised machine learning algorithms, Support Vector Machine, FP Growth Algorithm , K-medoids algorithm, Naive Bayes algorithm and decision tree classification algorithm.In [18] the author describes nearly 5 data mining algorithms namely random forest, neural network, classification and regression tree, support vector machine and the k-nearest neighbor algorithm. The author in [20] researched on finding out several patterns of the average temperature of Bangladesh per year as well as the average temperature per season. Different algorithms such as Linear Regression, Polynomial Regression, Isotonic Regression, and Support Vector Regressor are used. The results showed that Isotonic Regression algorithm predicts the training dataset most accurately, but Polynomial Regressor and Support Vector Regressor predicts the future average temperature most accurately.

3. METHODOLOGY

This section presents the methods used for preprocessing raw data attributes and the algorithms used for predicting rainfall data and forecasting future rainfall for the next few days in the city of Panjim Goa based on the past dataset. As seen in Fig. 1, the raw dataset collected from year 2010 to 2021 is cleaned using the data preprocessing approaches such as factor analysis and regression analysis using the IBM SPSS software and separated the attributes such as temperature, dew point, visibility, and wind speed listed in Table 1 and created linear regression models of each feature against the dates. We also trained the classification models to predict if it will rain or not with the same dataset. Feature set taken to train these classifiers is temperature, dew point, visibility, and wind speed. Now to predict if it will rain or not:

1. For a single day

We provided a date and with that date based on regression models it will try to get temperature, dew point, visibility, and wind speed for that particular day based on linear regression models and use those values and pass it to the classifier for it to predict if it will rain or not.

2. For a date range

We created all dates in the range and then took each date and repeated the same procedure as in step one.

For the classification of rainfall, we used four different classification approaches namely (1) K-nearest neighbor and (2) Naïve bayes classifier (3) Random forest classifier (4) Classification and regression tree (CART) model. Each of the approaches is described below:

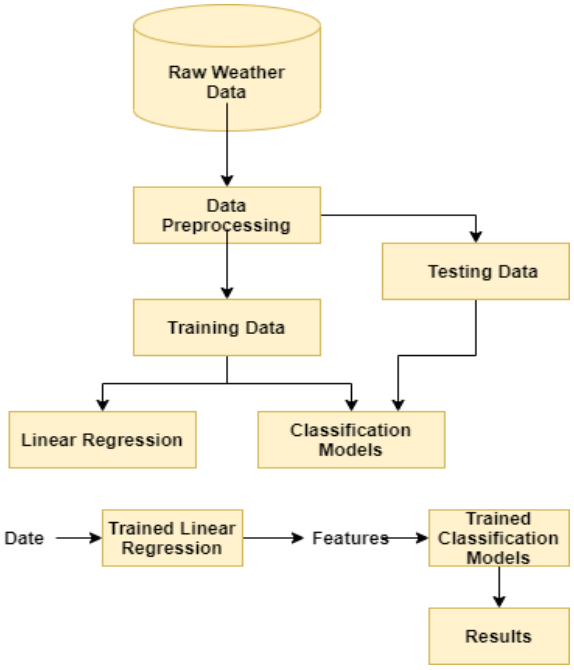

Fig -1: The proposed framework for rainfall forecasting.

(1) K-nearest neighbour: Each of the test entry is compared with K of its closest training data. The class (rainfall: yes/no) to which majority of these K neighbours is determined and the test entry is assigned to this most popular class

(2) Naïve Bayes classifier: Naive Bayes classifier is a classification algorithm based on Baye's Theorem. It is a family of algorithms where all of them share a common principle that is, every pair of features being classified is independent of each other.

(3) Classification and Regression tree (CART): One of the data mining algorithms for the retrieval of decision tree is the CART algorithm. It elucidates how a target variable's values can be predicted based on other values and allows for rapid classification of new observations.

(4) Random Forest classifier: A random forest classifier forms decision trees using the given data points and then predicts the classes using these decision trees by taking a majority vote among the predicted classes. It averages the results thus reducing over-fitting in a tree.

4. PERFORMANCE EVALUATION

In this section, we have described the details of the dataset that we have used for our experiments. We have also analyzed the results and compared the approaches used for prediction of rainfall.

4.1 Dataset

We have used the data collected from NOAA (National Oceanic Atmospheric Administration) of 11 years for a particular station of state Goa, India which is spilt into 2 parts training and testing data set. The data set of the year 2010 to 2019 is used for training the models and one year data of 2020 is used for testing whereas to observe the changing weather patterns we have also trained the models using four years data of year 2011 to 2014 and testing data is of year 2015. The raw data had 15 parameters (Station, Date, mean temperature, dew point, pressure, Mean sea level pressure, visibility, mean Wind speed, maximum sustained wind speed, maximum wind gust, maximum temperature, Minimum temperature, precipitation amount, Snow Depth, Rain) out of which 5 parameters (Temperature, Dew point pressure, Wind Speed, Visibility, Rain) are considered for the implementation of the four algorithms. These 5 attributes are considered to be the most suitable for predicting rainfall by performing factor analysis and regression analysis of raw data and obtaining results. Features with low value and duplicate values are ignored. Regression analysis results have shown that Mean Temp, Visibility, dew point, wind speed are the most appropriate predictions for Rainfall and therefore only these parameters are included in the forecast.

The pre-processed attributes used are listed in the table 1 below. Symbol values are represented by numerical values.

Table -1: Preprocessed attributes used for rain forecasting.

ATTRIBUTE	TYPE	DESCRIPTION
Visibility	Numerical	Kmph
Wind speed	Numerical	Kmph
Dew point	Numerical	Fahrenheit
Mean temperature	Numerical	Fahrenheit
Rainfall	String	Yes/No

4.2 Experimental Results

The formulas used for calculating accuracy and error are listed below:

Accuracy: Total number of relevant results

Total collection number

Error: Total number of incorrect results

Total collection number

```
predicting for day: 2021-12-10 00:00:00
temp(F) [82]
Dewpoint [78]
Visibility [2]
Windspeed [2]
prediction using NB model
[0]
prediction using categorical NB model
[0]
prediction using KN model
no of neighbours  5
[1]
no of neighbours  25
[1]
no of neighbours  125
[1]
no of neighbours  625
[1]
prediction using decision tree model
[1]
prediction using random forest classifier model
[1]
```

Fig -2. Linear regression result for a single day.

Figure 2 and 3 shows the output of linear regression for a single day and for the range of days. The 1s represents that there will be rainfall in the coming days and the 0s shows that there will be no rain.

```
predicting for date range from 2021-08-01 to 2021-09-01
prediction using NB model
[0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0]
prediction using categorical NB model
[0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0]
prediction using KN model
no of neighbours  5
[1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1]
no of neighbours  25
[1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1]
no of neighbours  125
[1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1]
no of neighbours  625
[1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1]
prediction using decision tree model
[1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1]
prediction using random forest classifier model
[1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1]
```

Fig -3. Linear regression result for a range of days.

Table -2: Prediction accuracy for rainfall of year 2020.

METHOD	ACCURACY	ERROR
Naïve Bayes Classifier	99.18%	0.82%
K-Nearest Neighbor	65.02%	34.98%
Random Forest classifier	60.10%	39.90%
Classification and Regression Tree	58.46%	41.54%

Table 2 shows the prediction accuracy obtained for the classification of rainfall for the test data of year 2020 using the four classification algorithms which are trained with ten years data and table 3 shows the prediction accuracy obtained for the test data of year 2015 using classification methods trained on four years data. As can be seen clearly, the naïve bayes method has provided the highest prediction accuracy of 99.18% with more training data whereas with less training data the accuracy obtained is 66.94%. The k-nearest neighbor performed better when trained with less data and accuracy provided is 82.36% whereas with more training data the accuracy is 65.02%, The CART algorithm has given the least accuracy as compared to random forest method when trained with more data whereas random forest performed well when trained with less amount of data.

Table -3: Prediction accuracy for rainfall of year 2015.

METHOD	ACCURACY	ERROR
K-Nearest Neighbor	82.36%	17.64%
Random Forest classifier	79.66%	20.34%
Classification and Regression Tree	75.82%	24.18%
Naïve Bayes Classifier	66.94%	33.06%

Tables 4, 5, 6, and 7 shows the confusion matrix obtained for rainfall prediction of year 2020 using naïve bayes classifier, K-nearest neighbor, random forest, and classification and regression tree method. The naïve bayes classifier could correctly classify 363 entries and 3 are misclassified whereas CART could classify only 214 entries correctly and 152 test entries are misclassified.

Table -4: Confusion matrix obtained using Naïve Bayes algorithm.

		Actual	
		No	***Yes***
Predicted	***No***	363	3
	Yes	0	0

Table -5: Confusion matrix obtained using K-nearest Neighbor algorithm.

		Actual	
		No	***Yes***
Predicted	***No***	238	128
	Yes	0	0

Table -6: Confusion matrix obtained using Random Forest algorithm.

		Actual	
		No	***Yes***
Predicted	***No***	220	146
	Yes	0	0

Table -7: Confusion matrix obtained using Classification and Regression Tree algorithm.

		Actual	
		No	***Yes***
Predicted	***No***	214	152
	Yes	0	0

Fig. 4 shows the percentage of test examples of the year 2020 correctly classified by all four of the classification methods on monthly basis. In this we can see that for most of the months, Naïve Bayes classifier outperforms the other three algorithms.

5. CONCLUSIONS

In the proposed work, four data mining algorithms namely naïve bayes, random forest, decision tree and k-nearest neighbor algorithms were trained using four years and ten years data separately for predicting the classification accuracies of rainfall for the year 2015 and 2020. The change in weather patterns could clearly be noticed when trained with less and more number of dataset. Experiments on the dataset used were carried out using the IBM SPSS software. The naïve bayes classifier gave highest accuracy compared to the other 3 algorithms whereas the least accuracy is obtained using classification and regression tree method. To measure the accuracy we will have to predict for the next day and create statistics for each of the date. The accuracy for the rain forecast was not measured since there is no way to give accuracy for future dates and we do not know if it actually rained or not in the future. The proposed model can be enhanced by finding an approach to give accuracy for future dates and improve the accuracy.

REFERENCES

[1] Zainudin, Suhaila, Dalia Sami Jasim, and Azuraliza Abu Bakar. "Comparative analysis of data mining techniques for malaysian rainfall prediction." International Journal on Advanced Science, Engineering and Information Technology 6, no. 6 (2016): 1148-1153.

[2] Kumar, R. Senthil, and C. Ramesh. "A study on prediction of rainfall using datamining technique." In 2016 International Conference on Inventive Computation Technologies (ICICT), vol. 3, pp. 1-9. IEEE, 2016.

[3] Shabib Aftab, Munir Ahmad, Noureen Hameed, Muhammad Salman Bashir, Iftikhar Ali, and Zahid Nawaz. "Rainfall Prediction in Lahore City using Data Mining Techniques." International journal of advanced computer science and applications 9, no. 4 (2018).

[4] Zamani, Nabila Wardah, and Siti Shaliza Mohd Khairi. "A comparative study on data mining techniques for rainfall prediction in Subang." In AIP Conference Proceedings, vol. 2013, no. 1, p. 020042. AIP Publishing LLC, 2018.

[5] Hernández, Emilcy, Victor Sanchez-Anguix, Vicente Julian, Javier Palanca, and Néstor Duque. "Rainfall prediction: A deep learning approach." In International Conference on Hybrid Artificial Intelligence Systems, pp. 151-162. Springer, Cham, 2016.

[6] SamsiahSani, Nor, Israa Shlash, Mohammed Hassan, Abdul Hadi, and Mohd Aliff. "Enhancing Malaysia Rainfall Prediction Using Classification Techniques." J. Appl. Environ. Biol. Sci 7, no. 2S (2017): 20-29.

[7] Huang, Mingming, Runsheng Lin, Shuai Huang, and Tengfei Xing. "A novel approach for precipitation forecast via improved K-nearest neighbor algorithm." Advanced Engineering Informatics 33 (2017): 89-95.

[8] Bui, Dieu Tien, Quoc Phi Nguyen, Nhat-Duc Hoang, and Harald Klempe. "A novel fuzzy K-nearest neighbor inference model with differential evolution for spatial prediction of rainfall-induced shallow landslides in a tropical hilly area using GIS." Landslides 14, no. 1 (2017): 1-17.

[9] Dash, Yajnaseni, Saroj K. Mishra, and Bijaya K. Panigrahi. "Rainfall prediction for the Kerala state of India using artificial intelligence approaches." Computers & Electrical Engineering 70 (2018): 66-73.

[10] Naik, Akshay R., A. V. Deorankar, and Premchand B. Ambhore. "Rainfall Prediction based on Deep Neural Network: A Review." In 2020 2nd International Conference on Innovative Mechanisms for Industry Applications (ICIMIA), pp. 98-101. IEEE, 2020.

[11] Anwar, Muchamad Taufiq, Wiwien Hadikurniawati, Edy Winarno, and Wahyu Widiyatmoko. "Performance Comparison of Data Mining Techniques for Rain Prediction Models in Indonesia." In 2020 3rd International Seminar on Research of Information Technology and Intelligent Systems (ISRITI), pp. 83-88. IEEE, 2020.

[12] Zaman, Yousuf. "Machine Learning Model on Rainfall-A Predicted Approach for Bangladesh." PhD diss., United International University, 2018.

[13] Swain, S., S. Nandi, and P. Patel. "Development of an ARIMA model for monthly rainfall forecasting over Khordha district, Odisha, India." In Recent Findings in Intelligent Computing Techniques, pp. 325-331. Springer, Singapore, 2018.

[14] Hanif, and Nafees Ayub. "Application of data mining techniques in weather data analysis." International journal of computer science and network security 17, no. 6 (2017): 22-28.

[15] Refonaa, J., and M. Lakshmi. "Cognitive computing techniques based rainfall prediction—A study." In 2017 International Conference on Computation of Power, Energy Information and Commuincation (ICCPEIC), pp. 142-144. IEEE, 2017.

[16] Sheikh, Fahad, S. Karthick, D. Malathi, J. S. Sudarsan, and C. Arun. "Analysis of data mining techniques for weather prediction." Indian Journal of Science and Technology 9, no. 38 (2016).

[17] Kunjumon, Christy, Sreelekshmi S. Nair, Padma Suresh, and S. L. Preetha. "Survey on weather forecasting using data mining." In 2018 Conference on Emerging Devices and Smart Systems (ICEDSS), pp. 262-264. IEEE, 2018.

[18] Taksande, Amruta A., and P. S. Mohod. "Applications of data mining in weather forecasting using frequent pattern growth algorithm." Int. J. Sci. Res 4, no. 6 (2015): 3048-3051.

[19] Shivaranjani, M. P., and K. Karthikeyan. "A review of weather forecasting using Data Mining Techniques." International Journal of Engineering and Computer Science 5, no. 12 (2016): 19784-19788.

[20] Shafin, Ashfaq Ali. "Machine learning approach to forecast average weather temperature of Bangladesh." Global Journal of Computer Science and Technology (2019).

[21] https://www7.ncdc.noaa.gov/

[22] https://en.wikipedia.org/wiki/Weather_Forecasting/

[23] https://codefying.com/2015/03/03/k-nearest-neighbor-classifier/

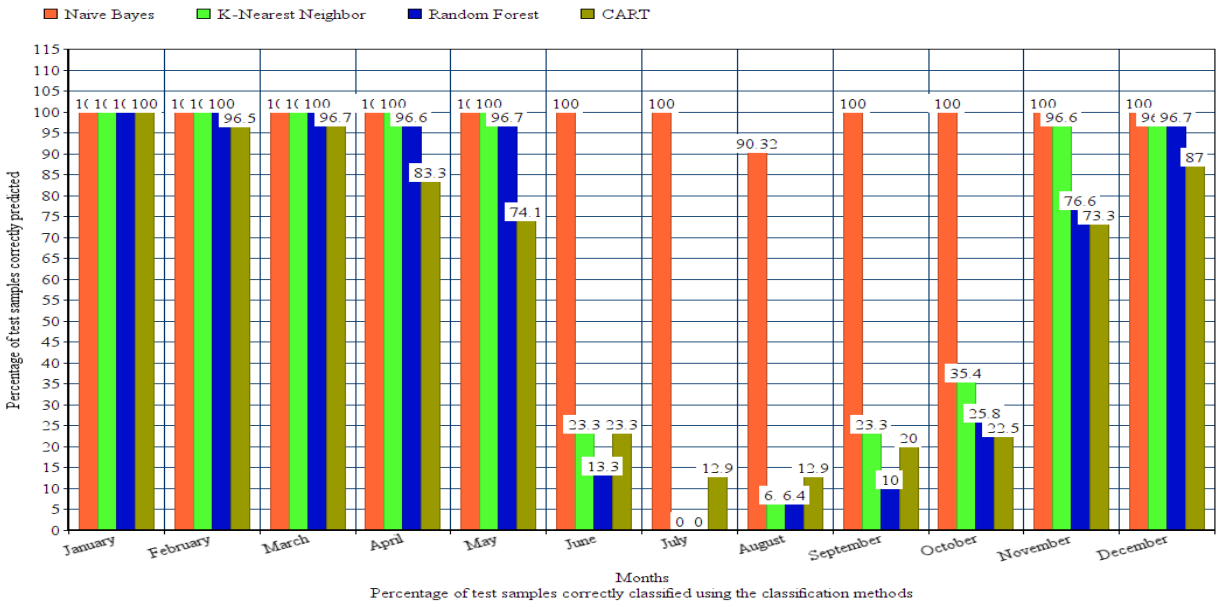

Fig -4: Percentage of test samples correctly classified using the classification methods