ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 2, February 2018

Weather Forecasting Using Data Mining

[1] Siddhant Revankar, [2] Shaba Desai
[1] Department of Information Technology, Padre Conciecao College of Engineering Verna, Goa
Goa University

Abstract - **Changing Climatic conditions are leading to alternate weather patterns. Accurately predicting weather patterns is important as they have wide social and economic impact. This paper proposes and analyzes a predictive model which analyse a wide range of data points with the aim of predicting likelihood and pattern of location-specific rainfall at a high degree of confidence.**

We extract knowledge from historical weather data collected from NOAA (National Oceanic Atmospheric Administration)[10]. From the collected weather data comprising of 15 attributes, only 5 attributes are most relevant to rainfall prediction. Data preprocessing and data transformation on raw weather data set is performed, so that it shall be possible to work on Bayesian and K-NN the data mining, prediction model used for rainfall prediction. The model is trained using training dataset and tested on test data for accuracy. We have used comparative approach of Bayesian and K-NN models and found Bayesian approach to be more accurate.

Keywords: --- **Rainfall prediction, Naive Bayes, K-Nearest Neighbor, Climate change, Weather patterns**

I. INTRODUCTION

Weather prediction is a challenging task and that too for weather is even more complex, dynamic and mind-boggling. Weather forecast postures right from the antiquated times as a major gigantic undertaking, because it depends on various parameters to predict the dependent variables like temperature, visibility, wind speed which are changing from weather calculation varies with the some specific location along with its atmospheric attributes. Accurate forecasts can help to identify possible floods in future and to plan for better water management. Weather forecasts can be categorized as: Forecasts. Which is forecasts up to few hours, Short term forecasts which is mainly Rainfall forecasts is 1 to 3 days forecasts, Forecasts for 4 to 10 days are Medium range forecasts and Long term forecasts are for more than 10 days. [1] Short range and Medium Range rainfall forecasts are important for flood forecasting and water resource management. There are many data mining techniques used for weather predictions. Naive Bayes approach and K-Nearest Neighbor has been used in this paper to forecast the Rainfall.

This paper utilize 4 years (2011-2014) data[10] from the month May to October as training dataset. Dataset contains 15 attributes out of which 5 relevant attributes i.e. (Temp, Visibility, Dewpoint, Speed, Rainfall) according to factor analysis and linear regression techniques are considered for rainfall prediction. The test dataset results of Naïve Bayes approach and K-NN(K-Nearest Neighbor) approach are compared for better results. This paper is organized in three sections, Introduction, Literature survey of the weather prediction models, proposed data mining model with the results of the implementations, and conclusion.

II. LITERATURE SURVEY

Literature survey provides the required knowledge about the project and its background. It also helps in following the best practices in project development. Literature survey also helps in understanding the risk and feasibility of the project.

In [1] the author uses Naïve Bayes classification method. Weather historical data is collected from Indian Meteorological Department (IMD) Pune. From the collected weather data attributes which are most relevant to rainfall prediction are chosen. Data pre-processing and data transformation on raw weather data set is performed, so that it shall be possible to work on Bayesian, The model is trained using the training data set and has been tested for accuracy on available test data. The meteorological centers use high performance computing and supercomputing power to run weather prediction model. To address the issue of compute intensive rainfall prediction model, Author proposed and implemented data intensive model using data mining technique. The model works with good accuracy and takes moderate compute resources to predict the rainfall. Prediction was found to be working well with good accuracy.

In [2] the author investigates the use of data mining techniques in forecasting attributes like maximum temperature, minimum temperature. This was carried out using Decision Tree algorithms and meteorological data collected between 2012 and 2015 from the different cities. On available datasets the Decision Tree Algorithm is applied for deleting the inappropriate data. On the

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 2, February 2018

percentage of these parameters they predict there is a full cold or full hot or snow fall.

In [3] Author's study carries historical weather data collected locally from Faisalabad city. Analysis and investigation was done using data mining techniques by examining changing patterns of weather parameters which includes maximum temperature, minimum temperature, wind speed and rainfall. After pre-processing of data and outlier analysis, K-means clustering algorithm and Decision Tree algorithm were applied. Two clusters were generated by using K-means Clustering algorithm with lowest and highest of mean parameters. Whereas in decision tree algorithm, a model was developed for modelling meteorological data and it was used to train an algorithm known as the classifier. The result obtained with smallest error (33%) was selected on test data set. While for the number of rules generated of the given tree was selected with minimum error of 25%. The results showed that for the given enough set data, these techniques can be used for weather analysis and climate change studies.

In [4] this paper describes empirical method technique belonging to clustering and classification and approach. ANNs are used to implement these techniques. The artificial neural networks analyse the data and learn from it for future predictions making them suitable for weather forecasting. Characteristics of neural networks can be used for the prediction of the weather processes. The input variables given are Temperature, Pressure, Relative humidity, Wind speed, perceptible water. The technique used for rainfall prediction is classification. In this technique, rainfall values are clustered using subtractive clustering and three classes or states are identified as low, medium and heavy.

In [5] the paper data used for the research work was obtained from meteorological tower of SRM University Chennai, India. Parameters like humidity, temperature, cloud cover, wind speed were used etc. Data Transformation and Data Pre-processing was performed. Different algorithms like Naïve Bayes and C4.5 (J48) Decision Tree algorithm was done simultaneously with dataset containing weather data collected over a period of 2 years. It was found that the performance of C4.5 (J48) decision tree algorithm was far better than that of Naïve Bayes.

In [6] the paper describes comparative approach. The training dataset used to train the classifier using Classification and Regression tree algorithm, Naive Bayes approach, K nearest Neighbor and 5-10-1 Pattern Recognition Neural Network and its accuracy is tested on a test dataset.

In [7] the proposed data model incorporates the Hidden Markov Model for prediction and for extraction of the weather condition observations the K-means clustering is used. For predicting the new or upcoming conditions the system need to accept the current scenarios of weather conditions.

In [8] the paper provides a survey of different data mining techniques being used in weather prediction or forecasting which helps the farmer for yield worthy productive and nourish the soil fertility such as artificial feed-forward neural networks (ANNs), fuzzy inference system, decision tree method, time series analysis, learning vector Quantization (LVQ) and biclustering technique.

III. PROPOSED MODEL

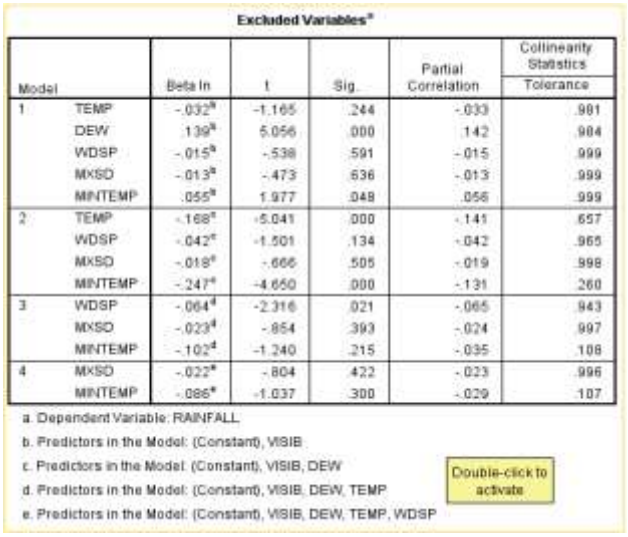
A. Data Collection and pre-processing

The data used for this work was collected from NOAA (National Oceanic Atmospheric Administration)[10]. The case data covered the period of 2011 to 2014 of the month May to October. Raw weather dataset contains 15 measured parameters like Station, Date, Mean temperature, Dew point, pressure, Mean sea level pressure, Mean station pressure Visibility, Wind speed, Maximum sustained wind speed, Maximum wind Gust, Max temperature, Min temperature Precipitation amount, Snow depth, Rainfall. Out of these 15 features we used only 5 most relevant attributes like Mean Temperature, Dew point pressure, Wind Speed, Visibility, Rainfall. We used factor analysis and linear regression techniques to find out the most relevant attribute needed for rainfall prediction. We ignored less relevant features in the dataset for model computation. We ignored variables like Station, Date as it had distinct value hence these attributes cannot be used for prediction. We also ignored Station pressure, Gust, Perception amount and Snow depth as these variables had similar duplicates values and factor reduction was not possible. Linear regression technique results shows that Mean Temp, Visibility, Dew point and Wind speed are the best predictors of Rainfall. Therefore we have included these attributes for prediction of rainfall. As shown in figure 1.

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 2, February 2018

Figure1 Linear Regression Results



Excluded Variables[a]

Model		Beta In	t	Sig.	Partial Correlation	Collinearity Statistics: Tolerance
1	TEMP	-.032[b]	-1.165	.244	-.033	.981
	DEW	.139[b]	5.056	.000	.142	.984
	WDSP	-.015[b]	-.538	.591	-.015	.999
	MXSD	-.013[b]	-.473	.636	-.013	.999
	MINTEMP	.055[b]	1.977	.048	.056	.999
2	TEMP	-.168[c]	-5.041	.000	-.141	.657
	WDSP	-.042[c]	-1.501	.134	-.042	.965
	MXSD	-.018[c]	-.666	.505	-.019	.998
	MINTEMP	-.247[c]	-4.650	.000	-.131	.260
3	WDSP	-.064[d]	-2.316	.021	-.065	.943
	MXSD	-.023[d]	-.854	.393	-.024	.997
	MINTEMP	-.102[d]	-1.240	.215	-.035	.108
4	MXSD	-.022[e]	-.804	.422	-.023	.996
	MINTEMP	-.086[e]	-1.037	.300	-.029	.107

a. Dependent Variable: RAINFALL
b. Predictors in the Model: (Constant), VISIB
c. Predictors in the Model: (Constant), VISIB, DEW
d. Predictors in the Model: (Constant), VISIB, DEW, TEMP
e. Predictors in the Model: (Constant), VISIB, DEW, TEMP, WDSP

Double-click to activate

The attribute values are numeric. The preprocessed attributes used are listed in Table 1.

Table 1. Data Description

Attribute	Type	Description
Mean Temperature	Numerical	Fahrenheit
Dewpoint Pressure	Numerical	Fahrenheit
Wind Speed	Numerical	Kmph
Visibility	Numerical	Kmph
Rainfall	String	Yes/No

B. Bayesian Rainfall prediction model

Bayesian classifiers[12] are statistical classifiers. They can predict class membership probabilities such as the probability that a given tuple belongs to a particular class.

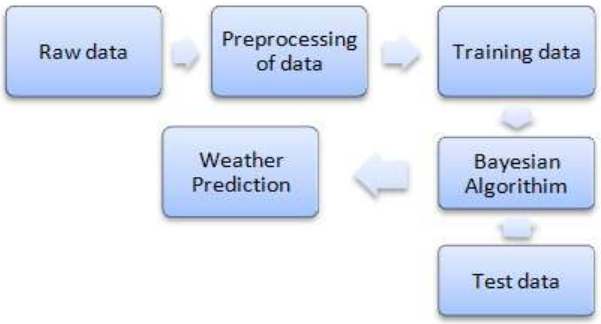


Figure 2 Bayesian Prediction Model

The Bayesian Classifier is capable of calculating the most probable output depending on the input. The flow of the model is shown in Fig 2. It is possible to add new raw data at runtime and have a better probabilistic classifier. A Naive Bayes classifier assumes that the presence (or absence) of a particular feature of a class is unrelated to the presence (or absence) of any other feature, given the class variable. The system consists of two functions Train classifier and Classify. Train classifier function will train the data set by calculating mean and variance of each variable as shown in Table 2.

Table-2. Mean and Variance Measurement

Mean Temp	Variance Temp	Mean Dew point	Variance Dewpoint	Mean Visibility	Variance Visibility	Mean Speed	Variance Speed
83.5	6.7	74.9	13.9	2.5	0.16	7.8	12.2
80.2	5.4	76.8	1.99	2.36	0.18	9.1	12.6

The classifier created from the training data set using a Gaussian distribution. The Classify function finds the probabilities using normal distribution. In order to get the probability of P(Temp/Yes) we use the formula.

$$F(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$$

Here x is the value of the temperature from the test data. μ and σ are mean and standard deviation of temperature calculated from the train dataset. Similar process is repeated for all the other attributes to get the individual probability.

1) Computational illustration of rainfall prediction:

For the classification as rainfall= Yes, the probability is given by:

P(Rainfall=Yes)=(P(Rain=yes)*P(Temp/Yes)*P(Dewpoint=Y es)* P(Visibility=Yes)* P(WindSpeed=yes))

For the classification as rainfall= No, the probability is given by:

P(Rainfall=No)=(P(Rain=No)*P(Temp/No)*P(Dewpoint=No)*P(Visibility=No)*P(Wind Speed =No))

If Value of P(Rainfall= Yes) > P(Rainfall=No) then Predict Rainfall= Yes.

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 2, February 2018

If Value of P(Rainfall= Yes) <P(Rainfall=No) then Predict Rainfall= No.

2) Result of rainfall prediction(Bayesian approach):
In this model, we have used Train datasets of Panjim city. We used the actual monsoon data of the year (2015) as test data to compare it with the model results. The dataset and the obtained results are shown below, in Table-2. The model observed to be accurate.

Table-3. Accuracy and Error Measurement

Dataset	Training Dataset	Test Dataset	Accuracy	Error
Panjim City	847	184	80.43%	19.56%

C. K-NN(K Nearest-Neighbor) Model
K Nearest Neighbor classifier[14] is based on learning by analogy, that is, by comparing a given test tuple with training tuples that are similar to it. The training tuples are described by n attributes. Each tuple represents a point in an n-dimensional space. In this way, all the training tuples are stored in an n-dimensional pattern space. When given an unknown tuple, a k-Nearest-Neighbor classifier searches the pattern space for the k training tuples that are closest to the unknown tuple.

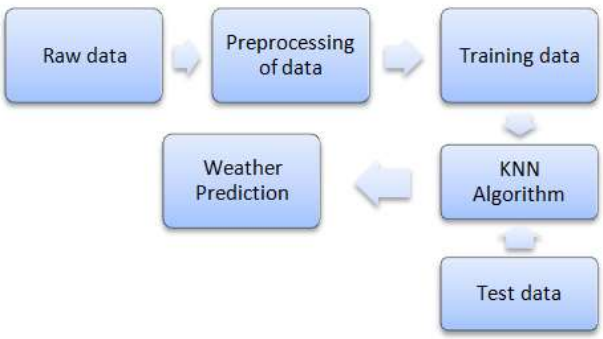


Figure 3 K-NN Prediction Model

The flow of the model is as shown in fig 3. Here K is the number of instances used to cast the vote when labelling previously unobserved instance. K-NN is a type of instance-based learning, or lazy learning, where the function is only approximated locally and all computation is deferred until classification Both for classification and regression, a useful technique can be to assign weight to the contributions of the neighbors, so that the nearer neighbors contribute more to the average than the more distant ones.

1) Computational illustration of rainfall prediction:
A common weighting scheme consists in giving each neighbor a weight of 1/d, where d is the distance to the neighbour. Given test dataset unobserved instance I={i0,…,in, class}, we calculate the Euclidean distance between I and each known instance in the dataset as follows:

$$D_i = \sqrt{\sum_{k=0}^{N}(Z_k^i - I_k^i)^2}$$

Here Z is a sequence of values from train dataset of some instance i in attribute k for which a classification is given and I is the unclassified test data instance. The distances were calculated on normalized data. We normalize each value according to:

$$\frac{Z_k^i - Min(Z_k)}{Max(Z_k) - Min(Z_k)}$$

Z is from the dataset. The instance that we need to classify is also normalized. Once the distances are calculated, we can proceed to vote on which class the instance I should belong to. To do this, we select K smallest distances and look at their corresponding classes. The value of K in this case was taken a 3.

2) Result of rainfall prediction(K-NN Approach):
In this model, we have used datasets of actual Panjim city. We used the actual monsoon data of the year (2015) as test data to compare it with the model results. The obtained results are shown below, in Table-4.

Table-4. Accuracy and Error Measurement

Dataset	Training Dataset	Test Dataset	Accuracy	Error
Panjim City	847	184	75.55	24.45

IV. CONCLUSIONS

In this paper, Bayesian and K-NN process were implement ted. The Bayesian prediction model can easily learn new classes. The accuracy will grow with the increase of learning data. The model returns good prediction results. The negative part of model is, when a predictor category is not present in the training data, the model assumes that a new record with that category has zero probability. This could be a major issue if this rare predictor value is important. On the hand KNN is effective when data is large but finding unknown patterns like forecasting the future trends. Value for K in K

Nearest Neighbor technique is difficult to determine. In this paper the Bayesian model proved to be more accurate then the K-NN Model.

ACKNOWLEDGMENT

The author of this paper thanks to authorities, and scientists, of, National Oceanic and Atmospheric Administration (NOAA) for providing the factual meteorological data; and helping the author to understand and interpret the data in the right direction. The understanding of data made authors convenient to find out the accuracy of the rainfall prediction model.

REFERENCES

[1] B.B Meshram, Valmik B Nikam, "Modelling Rainfall Prediction using Data Mining Method, A Bayesian Approach", Fifth International Conference on Computational Intelligence, Modelling and Simulation,2013.

[2] Siddhart S Bhatkande, Roopa.G Hubballi, "Weather Prediction Based on Decision Tree Algorithm Using Data Mining Techniques", International Journal of Advanced Research in Computer and Communication Engineering Vol.5, Issue 5, May 2016.

[3] M Ramzan Talib, Toseef Ullah, M Umer Sarwar, M Kashif Hanif and Nafees Ayub, "Application of Data Mining Techniques in Weather Data Analysis", IJCSNS International Journal of Computer Science and Network Security, June2017.

[4] Jyothis Joseph, Ratheesh T K, "Rainfall Prediction using Data Mining Techniques", International Journal of Computer Applications (0975 – 8887)Volume 83 – No 8, December 2013.

[5] Fahad Sheikh, S. Karthick, D. Malathi2, J. S. Sudarsan3 and C. Arun, "
Analysis of Data Mining Techniques for Weather Prediction",
Indian Journal of Science and Technology, October 2016.

[6] Deepti Gupta, Udayan Ghose, "A Comparative Study of Classification Algorithms for Forecasting Rainfall", 978-1-4673-7231-2/15/$31.00 ©2015 IEEE.

[7] Rohit Kumar Yadav, Ravi Khatri, "A Weather Forecasting Model using the Data Mining Technique", International Journal of Computer Applications (0975 – 8887)Volume 139 – No.14, April 2016.

[8] Ms.P.Shivaranjani, Dr.K.Karthikeyan. "A Review of Weather Forecasting Using Data Mining Techniques", International Journal of Engineering And Computer Science ISSN: 2319-7242 Volume 5 Issue 12 Dec. 2016.

[9] National Climatic Data Center, U.S. Department of Commerce.

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

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

[12] Data Mining Concepts and Techniques Third EditionJiawei Han and MichelineKamber Morgan Kaufmann Publishers. Page 351

[13] https://www.codeproject.com/Articles/318126/Naive-Bayes-Classifier/

[14] Data Mining Concepts and Techniques Third EditionJiawei Han and MichelineKamber Morgan Kaufmann Publishers. Page 423

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