Analysing the performance of Machine Learning Algorithms on Rainfall Prediction

[1]S Eshita kamalavalli, [2] Battula S S L Sravya, [3] I Srikar, [4] U Praveen Kumar, [5] Dr Poosapati Padmaja
[1,2,3,4] Student
[5]Professor,
[1]Department of Information Technology,
[1]Anil Neerukonda Institute of Technology and Sciences, Visakhapatnam, India

[1]eshitakamalavalli77@gmail.com, [2]sravyabattula@gmail.com, [3]srikarcool1@gmail.com,
[4]Upkumar2k18@gmail.com, [5]poosapatipadmaja@gmail.com

ABSTRACT

The prediction of rainfall has become a vital part of weather forecast. With the advancements in communication and networking, this prediction has taken over another major step. The benefits are numerous, be it for predicting the crop predictivity by predicting the rainfall every year and planning accordingly or pre-planning the water structures in an area by predicting the rainfall in the region in that period of time or efficient usage of water resources rainfall prediction plays a pivotal role in it. One of the approaches to achieve this is by developing a machine learning model using classification algorithms on weather data set to predict rainfall. However, rainfall prediction being a random phenomenon, due to unpredictable and dynamic climate data sets, which can frequently change according to global climatic changes. there is no one perfect algorithm to use hence, we are evaluating the performance of kNN (K nearest neighbours), Naïve Bayes Classification, Decision Tree Algorithm on the periodic weather dataset of a region extracted from Kaggle.com to predict rainfall. The performance measures being accuracy, precision and recall. We have deduced how the change in the training size can affect the accuracy of the algorithm as the test size is kept constant.

***Keywords*:** kNN, Naïve Bayes, Decision Tree, Accuracy, Precision, Recall.

1.INTRODUCTION

Prediction of rainfall has become a vital part of weather forecast. This includes the benefits estimating the crop productivity in a region over a period based on rainfall prediction in that area, pre planning of building water structures as dams or reservoirs based on the volume of rainfall in that area, efficient usage of water resources i.e. water conservation or precautionary measures in case of estimation of low rainfall. This calls for a high need for the necessity of predicting rainfall over a region. One of the approaches of achieving this is by building a machine learning model using classification algorithms.However, rainfall prediction being a random phenomenon places a challenge of choosing the suitable algorithm for training the model. Moreover, there are numerous classification algorithms suitable for the data sets.As there is no one algorithm perfect for any dataset, we are analysing the performance of three classification algorithms namely, kNN, Naïve Bayes Classification and Decision Tree for the same dataset. We chose the dataset of periodic weather dataset of a region consisting of atmospheric factors such as temperature, apparent temperature, humidity, pressure, wind bearing etc. from Kaggle.com for this project. The performance measures of analysis would be accuracy, precision and recall.

2. RELATED WORKS

With the advancements in technology, obtaining weather data at various meteorological data has become easier however developing a suitable or efficient model has become a challenge. Moreover, when factors such as wind bearing or pressure keep fluctuating with their floating-point values, it places a challenge in accommodating these huge amounts of data with good precision and accuracy measures. Several works are being carried out trying to solve this menace. Few works related to our work are:

PinkySaikia Dutta, Hitesh Tahbilder [1] "Prediction of Rainfall Using Data Mining Technique over Assam". In this paper, they have described data mining technique in forecasting monthly Rainfall and traditional statistical technique Multiple Linear Regression. They included Six years data from 2007 to 2012 this was collected locally from Regional Meteorological Center, Guwahati, Assam, India.

Improved Nave Bayesian classification is considered by James-N.K.Liu by varying the weather state data. The Bayesian classification is used for identifying the weather prediction. The algorithm is compared with different models based on genetic algorithms and results showed that the considerable amount of accuracy is obtained.

F.Dell presented a more general approach for identifying the varying the wind speed. The metrological data is divided into two groups and performed the classification on these groups to identify the hurricanes and non-hurricanes. The above authors analysed and predicted the rainfall.

Valmik.B et al. proposed a model for predicting the weather data based on classification technique and considered several important attributes such as wind pressure, humidity, vapor, wind speed and pessimistic results obtained from the experimental result. The results showcased good accuracy by correlating the above parameters.

E. G. Petre[10] proposed a model for Weather prediction using Decision tree CART based on parameters such as Pressure, clouds quantity, humidity, precipitation, temperature with time period of 4 years which resulted with a 83% accuracy.

Z Jan et al.[38] Inter annual climate prediction using kNN considered parameters such as Wind speed, dew point, seal level, snow depth, rain considering data over 10 years with 40000 instances gave 96% Long term accurate results with large set of attributes.

3. EVALUATION METRICS

3.1 Accuracy

This metric is calculate by finding the total number of instances that are correctly predicted as positive cases to the total number of data that is present, the instances are classified into positive cases or negative cases by calculating the data that are divided into True positive (TP), True negative (TN), False positive (FP), False positive (FN).

Accuracy = (TP+TN)/(TP+FP+FN+TN)

3.2 Precision

It refers to the total data which is correctly predicted to be positive over the total number of data that are predicted to be positive, by observing the false positive and true positive instances, it can be calculated as:

Precision = TP/(TP+FP)

3.3 Recall

It is also known as the true positive rate (TPR), Sensitivity (SN) or detection rate. It indicates the total number of instances that are correctly predicted as positive over the total number of positive instances present. While detecting the overall positive data in the dataset the recall serves best as the main evaluation metric or the best performance indicator of positive data, it is calculated as follows:

Recall = TP/(TP+FN)

4. DATASET USED

For this project, we have worked on a periodic weather dataset of a region for the past 10 yearsfrom Kaggle.com. The dataset contains attribute classes namely temperature, apparent temperature, atmospheric pressure, wind speed, humidity, wind bearing 0and the targetclass attribute as precipitation i.e. if there has been rainfall or not. The data set contains 96,454 rows of labelled periodic data of the same city over a period. Though we measure against the performance measures, cleaned or pre-processed dataresults in better results and performance of the algorithms thereby increasing their efficiency.

One deficiency in this data set is that it is unscaled i.e. values vary largely in terms of magnitude. Unscaled data puts forth the problem of unevenness especially while calculating the Euclidean distance in terms of kNN between two points. Hence the data had to be scaled to avoid this discrepancy.

Another discrepancy is missing values. Missing values result in issues like Zero frequency in case of Naïve Bayes algorithm thereby affecting the model. One effective way to overcome this issue is by mean imputation wherein every missing value is replaced with the mean value of the respective column.

5. IMPLEMENTATION

Initially import the data set in the form of a csv file. Then perform the required pre processing required for cleaning the data. After its extraction divide the data into training data and testing data samples. The training data aids in building the model while the testing data is used to evaluate model's performance.

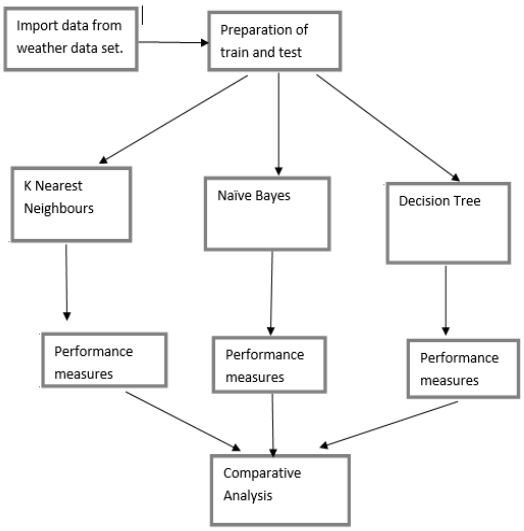

Fig 5.1: Flow Process of System

Train the algorithm taken with the training data set. Now take a detection model, input the test data and the trained algorithmic input to it. This starts predicting the class to which the output belongs to i.e., either it has precipitated or not. According to the output predicted and the actual output it should predict, evaluate the metrics like accuracy, precision, recall. Continue the same process for all algorithms under consideration. Finally, rank them all based on their performance and find the best performing algorithm in the given context.

6 RESULTS

Consider if it rains as a positive class and if it doesn't as a negative class evaluation measures precision and recall are based on the factors in a confusion matrixnamely:

- **True Positive** is an outcome where the model correctly predicts the positive class.
- **True negative** is an outcome where the model correctly predicts the negative class.
- **False Positive** is an outcome where the model incorrectly predicts the positive class.
- **False Negative** is an outcome where the model incorrectly predicts the negative class.

For conducting this, we have chosen a weather data set of 96,454 rows with 6 columns and one class column. This data is split randomly into training and testing data in the ratio 7:3 respectively. Thereby training data consisting 67,517 as training data tuples and 28,937 as testing data tuples.

6.1 K Nearest Neighbours

Train size: 7000 Test size : 3000 Accuracy : 0. 9815

Train size: 14000 Test size : 6000 Accuracy : 0. 9830

Train size: 21000 Test size : 9000 Accuracy : 0. 9833

Train size: 28000 Test size : 12000 Accuracy : 0. 9829

Train size: 35000 Test size : 15000 Accuracy : 0. 9816

Train size: 42000 Test size : 18000 Accuracy : 0. 9807

Train size: 49000 Test size : 21000 Accuracy : 0. 9819

Train size: 56000 Test size : 24000 Accuracy : 0. 9818

Train size: 63000 Test size : 27000 Accuracy : 0. 9832

Table 6.1 : Results of KNN algorithm

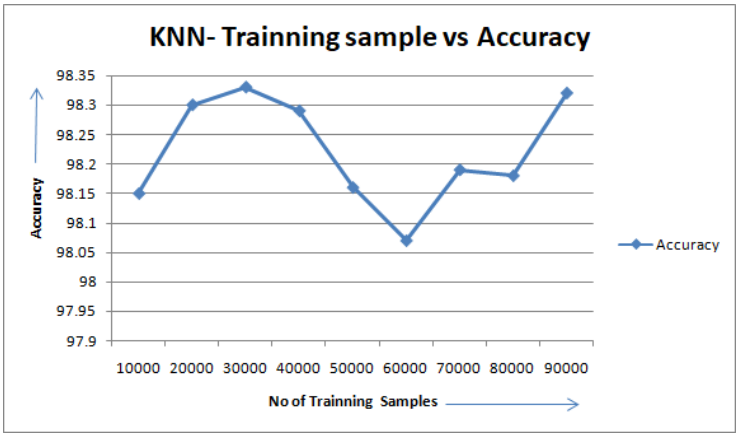

Fig 6.1: Graph plotting training size vs KNN accuracy

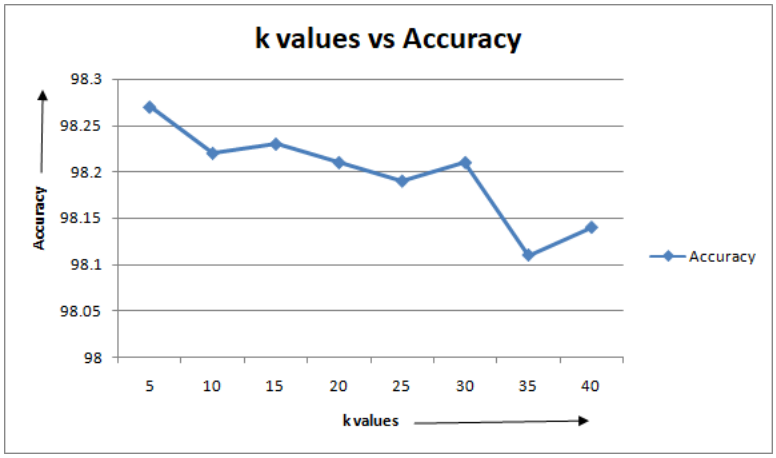

Fig 6.1.1: Graph plotting k values vs accuracy

6.2 Naïve Bayes

Train size: 7000 Test size : 3000 Accuracy : 0. 9235

Train size: 14000 Test size : 6000 Accuracy : 0. 9407

Train size: 21000 Test size : 9000 Accuracy : 0. 9442

Train size: 28000 Test size : 12000 Accuracy : 0. 9454

Train size: 35000 Test size : 15000 Accuracy : 0. 9465

Train size: 42000 Test size : 18000 Accuracy : 0. 9223

Train size: 49000 Test size : 21000 Accuracy : 0. 9227

Train size: 56000 Test size : 24000 Accuracy : 0. 9270

Train size: 63000 Test size : 27000 Accuracy : 0. 9223

Table 6.2: Naïve Bayes Results

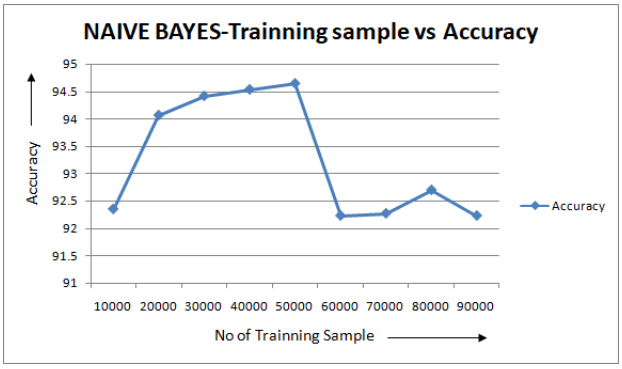

Fig 6.2: Graph denoting the accuracy vs training sample set

6.3 Decision Tree

Train size: 7000 Test size : 3000 Accuracy : 0. 1

Train size: 14000 Test size : 6000 Accuracy : 0. 1

Train size: 21000 Test size : 9000 Accuracy : 0. 1

Train size: 28000 Test size : 12000 Accuracy : 0. 1

Train size: 35000 Test size : 15000 Accuracy : 0. 9955

Train size: 42000 Test size : 18000 Accuracy : 0. 9965

Train size: 49000 Test size : 21000 Accuracy : 0. 9968

Train size: 56000 Test size : 24000 Accuracy : 0. 9969

Train size: 63000 Test size : 27000 Accuracy : 0. 9975

Table 6.3: Decision Tree results

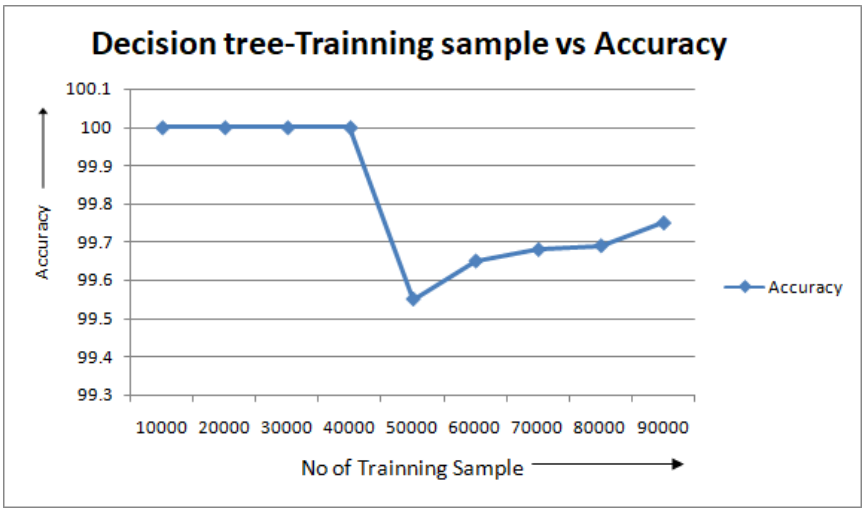

Fig 6.3: Graph denoting accuracy vs Training sample in decision tree

6.4 Overal Results:

Table 6.4 shows the results when trained with entire dataset and then evaluated with the test set. Accuracy is being calculated based on the number of right predictions.

Based on the confusion matrix i.e. the number of true positives, true negatives, false positives and false negatives the performance measures as precision and recall are generated.

Algorithms	Accuracy	Precision	Recall
kNN	98.25	90.34	94.39
Naïve Bayes	90.62	93.68	99.29
Decision Tree	99.28	99.63	99.55

Table 6.4. Overall performance measures

The Fig 6.4 indicates the performance measures of each algorithm in terms of accuracy, precision, recall while the training data set keeps increasing.

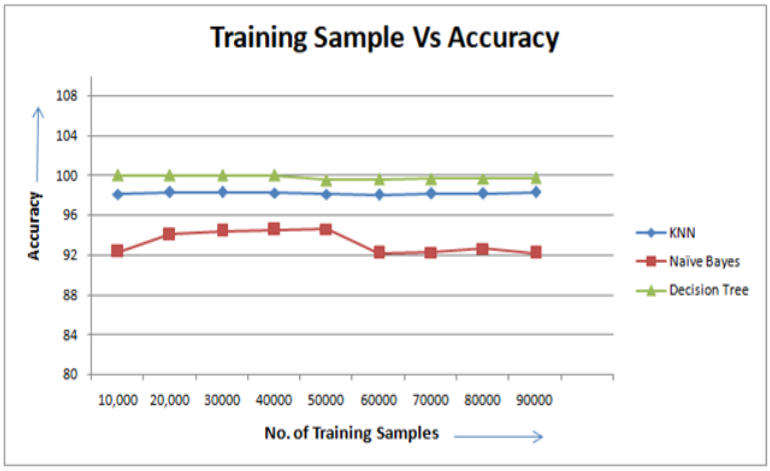

Fig 6.4 Comparision of accuracy vs training size

7. FUTURE SCOPE

In addition to Delhi the model will collect data from other regions as well for evaluation and prediction. Moreover, other related features of data will be collected and the existing model will be enhanced using more features for making the model more robust for weather prediction.

8. CONCLUSION

The idea behind this paper is to predict rainfall with the most possible accuracy and help farmers in better determination of their crop productivity. However, rainfall prediction being a random phenomenon cannot be concluded with just one algorithm.

From the three algorithms namely K Nearest Neighbours, Naïve Bayes and Decision Tree algorithm we have evaluated that Decision tree algorithm performs better in terms of accuracy, precision and recall.

9. REFERENCES

[1] PinkySaikia Dutta, Hitesh Tahbilder"Prediction of Rainfall Using Data Mining Technique over Assam"Indian Journal of Computer Science and Engineering (IJCSE) Vol -5, 2014.

[2] Valmik B Nikam and B.B. Meshram, “Modeling Rainfall Prediction using Data Mining Method”, Fifth International Conference on Computational Intelligence, Modeling and Simulation, Issue No: 2166-8531, PP:132-136, 2013.

[3] Divya Chauhan and Jawahar Thakur, “Weather Prediction Based on Decision Tree Algorithm Using Data Mining Techniques”, International Journal on Recent Innovation Trends in Computing and CommunicationVol. 5, Issue 5, May 2016.

[4]Zahoor Jan, M.Abrar, Shariq Bashir, Anwar M.Mirza, “Seasonal to Inter-Annual Climate Prediction using Data Mining kNN Technique”, IMTIC 2008, CCIS 20,PP:40-51, 2008.

[5]F.DellAcqua and P.Gamba, "A simple Model Approach to the problem of meteorological object Tracking", Volume: 1, Issue No: 7803-6359, PP: 2152-2154, Italy, 2000.

[6]Badhiye S.S, Dr.Chatur P.N, Wakode B.V, "Temperature and Humidity Data Analysis for Future Value Prediction using clustering Technique: An Approach", International Journal of Emerging Technology and Advanced Engineering, ISSN:2250-2459, volume-2, Issue-1, PP- 88-91, January 2012. [11].

[7]G.Vamsi Krishna ,Prediction of Rainfall Using K-Means algorithm,International Journal of Mathematical Sciences and Computing (IJMSC),2015

[8]Siddharth S Bhatkande Roopa G Hubbli "Weather Prediction Based on Decision tree Algorithm Using Data Mining Techniques",International Journal of Advanced Research in Computer and Communicational Engineering , Vol.5 , Issue 5 ,May 2016

[9]Ayisha Sddiqua L , Senthil Kumar N C,"Heavy Rainfall Predition Using Gini Index in Decision Tree",International Journal of Recent Technology and Engineering(IJRTE),ISSN :2277-3878,Vol 8,Issue 4,Nov 2019

[10]Shahista Navaz1 , Huma Khan2 , Dr. S. M. Ghosh3, A Survey on Ensemble Computing Method for Rainfall Prediction in Different Regions of Chhattisgarh, International Journal of Science and Research (IJSR), Volume 6 Issue 6, June 2017

[11]L. Maria Michael Visuwasam1 , Dr.P.Geetha2 , G.Gayathri3 , K. Divya4 , S. Hariprasath, Prognostication Of Rainfall Using Data Mining Techniques, International Journal of Engineering, Applied and Management Sciences Paradigms (IJEAM),Volume 54 Issue 1 April 2019.