Asian Journal of Computer Science and Technology
ISSN: 2249-0701 (P) Vol.7 No.3, 2018, pp.13-19
© The Research Publication, www.trp.org.in
DOI: https://doi.org/10.51983/ajcst-2018.7.3.1901

Comparative Study of Rainfall Prediction Modeling Techniques (A Case Study on Srinagar, J&K, India)

Razeef Mohd[1], Muheet Ahmed Butt[2] and MajidZaman Baba[3]
[1]Student, [2]Scientist-D, Department of Computer Science, [3]Scientist-D, Directorate of IT&SS,
University of Kashmir, Jammu and Kashmir, India
E-Mail: m.razeef@gmail.com, ermuheet@gmail.com, zamanmajid@gmail.com

(Received 18 August 2018; Revised 27 August 2018; Accepted 26 September 2018; Available online 4 October 2018)

***Abstract- Prediction* of rainfall is one of the most essential and demanding tasks for the weather forecasters since ages. Rainfall prediction plays an important role in the field of farming and industries. Precise rainfall prediction is vital for detecting the heavy rainfall and to provide the information of warnings regarding the natural calamities. Rainfall prediction involves recording the various parameters of weather like wind direction, wind speed, humidity, rainfall, temperature etc. From last few decades, it has been seen that data mining techniques have achieved good performance and accuracy in weather prediction than traditional statistical methods. This research work aims to compare the performance of few data mining algorithms for predicting rainfall using historical weather data of Srinagar, India, which is collected from http://www.wundergrounds.com website. From the collected weather data which comprises of 9 attributes, only 5 attributes which are most relevant to rainfall prediction are considered. Data mining process model is followed to obtain accurate and correct prediction results. In this paper, various data mining algorithms were explored which include decision tree based J48, Random forest, Naive Bayes, Bayes Net, Logistic Regression, IBk, PART and bagging. The experimental results show that J48 algorithm has good level of accuracy than other algorithms.**

***Keywords:*Rainfall Prediction, Data Mining, J48, Random Forest, IBk, Naive Bayesian, Bagging**

I. INTRODUCTION

In today’s information technology era, weather forecasting has become the most challenging and important technique which helps us to predict the atmosphere of a location. Weather prediction is an important application in meteorology and has become one of the most scientifically and technologically challenging problem for meteorologists around the world [1]. From the last few decades the advancement and development in science and technology enable scientists to make better and precise weather prediction. Rainfall prediction has a vital role to play in the field of agriculture and horticulture. Rainfall prediction help in water resource management, crop production plan and other things that are of greater concern for the mankind. More advance techniques and technologies are used by the scientists to make more accurate rainfall predictions. Rainfall forecasting has been most interesting and alluring field since the beginning of time and still is one of the most challenging and appealing domain. Number of methods and techniques are used by the scientists to forecast rain; some of these techniques are more accurate than others. Weather forecasting is a process of collecting data on atmospheric conditions, which records the temperature, humidity, rainfall, wind speed and its direction etc. High speed computers, wired and wireless sensors, meteorological satellites and weather radars are the tools used to collect the weather data for weather forecasting [2]. Weather forecasting helps to prevent climatic hazards and in climate monitoring, drought detection, severe weather prediction, agriculture and production, planning in energy industry, aviation industry, communication, pollution dispersal etc. There is huge amount of weather data available which is rich in information and can be used for weather prediction. Various data mining techniques are applied to the weather data to predict atmospheric parameters like temperature, wind speed, rainfall, meteorological pollution etc. which tend to change from time to time and weather calculation varies with the geographical location along with its atmospheric parameters. Some commonly used data mining techniques for weather prediction are Decision Trees, Artificial Neural Networks (ANN), Naive Bayes Networks, Support Vector Machines (SVM), Fuzzy Logic, Rule-based Techniques which includes Memory based reasoning Techniques and Genetic Algorithms.

The prediction of correct weather condition especially prediction of rainfall is very important. Rainfall is important for crop production, water resource management, humidifying the atmosphere, producing streams and rivers, replenishing the water table and redistribution of fresh water in the water cycle. The occurrence of prolonged dry period or heavy rain at the critical stages of the crop growth and development may lead to significant reduce crop yield [3]. The prolonged and incessant rainfall can also lead to floods which cause destruction of crops, livestock, orchards etc and can also decimate housing and business infrastructure. Thus rainfall prediction becomes a significant factor in agricultural countries like India.

In this work, we have employed various data mining techniques for the rainfall prediction based on various atmospheric variables. The weather data used in this work is taken from Srinagar, India, from November 2015 to November 2016[1]. Srinagar is the largest city and the summer capital of the Indian state of Jammu and Kashmir. It lies on the banks of Jhelum River, a tributary of the Indus River, Dal and Anchar Lakes. The average annual rainfall is around 720 millimeters (28 in). Spring is the wettest season while autumn is the driest. The highest temperature reliably recorded is 38.3 °C (100.9 °F) and the lowest is −20.0 °C (−4.0 °F) [4].

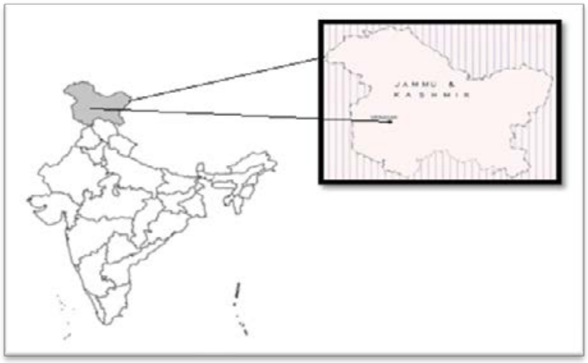

Fig. 1 Study Area

II. RELATED WORK

Data mining techniques have significantly changed the weather prediction process [29]. In the past decades lot of weather prediction models were designed and implemented using data mining techniques [31]. These predictive models showed appreciable accuracy in weather forecasting. In this section we will give a brief literature review of work based on rainfall prediction models using various data mining techniques.

In [5], Kannan*et al,* used regression method to forecast rainfall. In this work, the authors used five years of historical weather data from Statistical department of Tamil Nadu, Chennai, India. The authors have computed values for rainfall fall in the ground level by using Karl Pearson correlation which is then used to predicted rainfall for future years in ground level by using multiple linear regression technique.In [6], Valmik*et al,*proposed a rainfall prediction model based on Bayesian approach. In this work the authors used historical weather data from Indian Meteorological Department (IMD) Pune, India. Data preprocessing and data transformation is performed on raw weather data set, so that it shall be possible to work on Bayesian model. Posterior probabilities were used to calculate likelihood of each class label for input data instance and the one with maximum likelihood is considered resulting output. The proposed model can be deployed on commodity hardware without the need of high-performance computers. The results show that the model has good accuracy and takes moderate compute resources to predict the rainfall. The Bayesian approach used in this research has proved that the proposed rainfall prediction model works well with appreciable accuracy.

In [7], Nhita*et al,* proposed rainfall forecasting system using fuzzy system based on genetic algorithm (GA). The weather data for this research is taken Indonesian Agency for Meteorology, Climatology and Geophysics (BMKG) for Kemayoran area, Jakarta. The experimental results showed that the combination of GA and Fuzzy System for Kemayoran weather data can produce prediction model with more than 90% accuracy and can better predict rainfall.In [8], Mahajan*et al,* proposed a rainfall prediction model for 30 Indian sub divisions using on artificial neural network based on frequency analysis approach using Fast Fourier Transform. Weather data for 30 sub divisions was collected from IITM, Pune, India. The proposed prediction model is able to predict the quantity of rainfall before 1 year which is quite helpful for crop planning and management.

In [9], Geetha*et al,* proposed rainfall prediction model using decision trees. In this work the author highlights the prediction model using decision tree to predict weather phenomena like fog, rainfall, cyclones and thunderstorms. The results revealed 100%accuracy in performance vector. The success rate for the year 2014 is 80.67% when compared with the actual and target data. The proposed model is very promising, encouraging and further opens the door to extend with other soft computing techniques like fuzzy, genetic algorithms and artificial neural networks.In [10], Dutta*et al,* proposed rainfall prediction model using Multiple Linear Regression data mining technique which can predict monthly rainfall. For this research weather data of six years from 2007-2012 was collected from Regional Meteorological Center, Guwahati, Assam, India. The performance of the model was measure in adjusted R-squared. The prediction model shows acceptable accuracy and acceptability.

In [11], Sharma *et al,* proposed Bayesian network model for mean monthly rainfall prediction of 21 stations in Assam, India. This work can be useful for better management of water resources. Monthly data of 20 years from 1981 to 2000 for all the atmospheric parameters is used for this study which was taken from different sources. Rainfall at a station is taken as a variable for this model and dependencies between rainfalls at different station is shown by Bayesian network. In this work, the author used K2 algorithm and conditional probability is found using maximum likelihood approximations. Five different atmospheric parameters viz. Temperature, Cloud cover, Relative humidity, Wind speed and Southern Oscillation Index (SOI) are used. The results revealed that temperature is found most efficient and wind speed least. SOI is also found important in improving the results. Some station got efficiency above 95% whereas other station got satisfactory results.

In [12], Akash D Dubey, proposed a rainfall prediction model using artificial neural networks (ANN). In this work the author has used the weather data of Pondicherry, India. Three different training algorithms viz. feed-forward back propagation algorithm, layer recurrent algorithm and feed-forward distributed time delay algorithm were used to create ANN models and keeping number of neurons for all the models to 20. Of all the algorithms, the results showed that feed-forward distributed time delay algorithm has best accuracy and MSE value as low as 0.0083.

III. DATA COLLECTION ANDPREPROCESSING

Following data mining process steps have been applied to pre-process and clean the collected raw weather data set as shown in figure 2. The most time consuming and essential part of data mining process is data collection and preparation. Understanding how the data is collected,stored, transformed, reported, and used is essential for the data mining process [20].

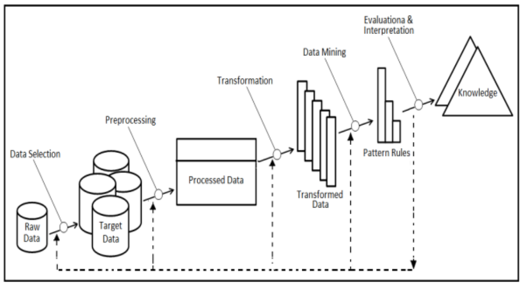

Fig. 2 Data Mining Process Model

A. Data collection

For this work we have collected weather data for one year from http:///www.wundergrounds.com website. For the prediction model, we used weather data of Srinagar, India from November 2015 to November 2016 [13]. The raw weather data collected consists of nine measured attributes which are date, temperature (high, low, average) in ^{0}c , Dew point (high, low, average) in ^{0}c , Humidity (high, low, average) in %age, sea level pressure (high, low, average) in hPa, visibility (high, low, average) in Km, wind (high, low, average) in Km/h, precipitation (high, low, average) in mm, Events (Rainfall snow, thunderstorm, fog). For this work out of these 9 features we have used the Average temperature, Average Humidity, Average sea level pressure, Average wind and Events features as shown in table I. We have ignored less relevant features in the dataset for better model computation and prediction.

TABLE I WEATHER DATA DESCRIPTION

Attribute	Type	Description
Temperature	Numerical	Temp is in deg. C
Humidity	Numerical	Humidity in Percentage
Sea Level Pressure	Numerical	Sea Level Pressure in hpa
Windy	Numerical	Wind Speed in Kmph
Events	Numerical	Rainfall in mm

B. Data Preprocessing and Data Cleaning

The main challenge in weather prediction is the poor data quality and selection. For this reason we try to preprocess data carefully to obtain accurate and correct prediction results. In this phase unwanted data or noise is removed from the collected data set which is done by removing the unwanted attributes and keeping the most relevant attributes that help in better prediction. Another major issue that is to be rectified is the missing values in the collected data set. Missing values in the data set is filled by using various techniques. In this work, the missing values for attributes in the dataset are replaced with the modes and means based on existing data. Adding the missing values provides a more complete dataset for the classifiers to be trained on [14].Data mining is the process of extracting the useful information from a large collection of data which was previously unknown [15]. For extracting useful information we need to follow data mining process model that will give us clean valuable dataset for model computation and better prediction. Very rarely data are available in the form required by the data mining algorithms. Most of the data mining algorithms would require data to be structured in a tabular format with records in rows and attributes in columns. The methodological discovery of useful relationships and patterns in data is enabled by a set of iterative activities known as data mining process. Not all discovered patterns leads to knowledge. It is up to the practitioner to invalidate the irrelevant patterns and identify meaningful information [16].

IV. RESEARCH METHODOLOGY

There are two main types of data mining approaches; supervised learning and unsupervised learning. In this work we have carried out research on supervised learning. Classification is a supervised learning approach which is based on training sample set. WEKA [31] machine learning tool is used to build predictive models. We have implemented eight classifiers which represent five categories of classifiers (i.e., trees, functions, Bayesian classifiers, lazy classifiers, and rules). These Classification algorithms are bagging, bayesNet, IBk, J48, Logistic Regression, NaiveBayes, PART and RandomForest which are experimentally implemented and compared against each other.

A. Bagging

Bagging (bag) stands for bootstrap aggregating is a machine learning algorithm that relies on an ensemble of different models. Bagging predicts an outcome multiple times from different training sets that are combined together either by uniform averaging or with voting [17]. The training data is re-sampled from the original data set. According to Witten and Frank [18], bagging typically performs better than single method models and almost never significantly worse. Bagging is a “bootstrap” ensemble method that creates individuals for its ensemble by training each classifier on a random redistribution of the training set. The bagging algorithm creates an ensemble of models (classifiers or predictors) for a learning scheme where each model gives an equally weighted prediction.

B. Decision Tree (J48 Algorithm)

J48 is Weka’s implementation of the C4.5 decision tree learning. C4.5 constructs a classifier in the form of decision tree. A set of data representing things e.g software fault data that are already classified is fed to the C4.5 algorithms to produce a decision tree. A classifier is a tool in data mining that takes a group of data representing things we want toclassify and attempts to predict which class the new data belongs to. Research on C4.5 algorithm was funded by the Australian Research Council for many years. Decision trees have internal nodes, branches, and terminal nodes. Internal nodes represent the attributes, terminal nodes show the classification result, and branches represent the possible values that the attributes can have. C4.5 is a well-known machine learning algorithm.

C. Naive Bayes

The Naive Bayesian classifier was first described in [19] in 1973 and then in [20] in 1992.Bayesian classifiers are statistical classifiers. Naïve Bayes algorithm is one of the most robust machine learning algorithms for rainfall prediction [11]. The Naïve Bayes classifier [21] is based on Bayes rule of conditional probability. It analysis each attribute individually and assumes that all of them are independent and important. Naive Bayes classifiers have been used extensively in fault-proneness prediction, for example in [22]. An advantage of the naive Bayes classifier is that it requires a small amount of training data to estimate the parameters necessary for classification.

D. Random Forest

Random Forest [23] is also another approach under ensemble classifier. Random Forest is a classifier based on decision trees which exhibits great performance in computer engineering studies by Guo*et al*., [24]. Random forest has one important advantage that it is fast and is able to handle large number of input attributes. It includes tens or hundreds of trees. In the construction of decision tree a random choice of attributes is involved. The trees are created using the following strategy [25]:

1. Each tree's root node has a sample bootstrap data which is equal to the actual data. There is a different bootstrap sample for each tree.
2. Using best split method subset of variables is randomly selected from input variables.
3. Each tree is then grown to the maximum extent possible without pruning.
4. When all trees are built in the forest, new instances are attached to all the trees then voting process takes place to select the classification with maximum votes as the new instance(s) prediction.

E. Logistic Regression (LR)

Logistic regression is a classification scheme which uses mathematical logistic regression functions. The most popular models are generalized linear models. Logistic regression measures the relationship between the categorical dependent variable and one or more independent variables by estimating probabilities using a logistic function, which is the cumulative logistic distribution [30]. Thus, it treats the same set of problems as probit regression using similar techniques, with the latter using a cumulative normal distribution curve instead. Equivalently, in the latent variable interpretations of these two methods, the first assumes a standard logistic distribution of errors and the second a standard normal distribution of errors [26]. The regression coefficients are usually estimated using maximum likelihood estimation. Researchers have applied statistical methods such as uni-variate or multivariate binary logistic regression to predict rainfall [10].

F. IBk

IBk is the Weka tool implementation of k-nearest-neighbor Classifier [27]. The number of nearest neighbors (k) can be set manually, or determined automatically using cross-validation. With k=1 the default value, IBk is in fact IB1. This is the basic nearest-neighbor instance based learner that searches for the training instance closest in Euclidean distance to the given test instance and uses the result of the search for classification (Witten and Frank 2005).

V. EXPERIMENTAL STUDY

Experiments are conducted on weather data of Srinagar, J&K, India from November 2015 to November 2016 [13] which is first pre-possessed and cleaned by implementing the data mining process model. The experiments are conducted in order to compare various data mining algorithm for rainfall prediction. In our collected weather data set, EVENT is predicted variable which tells whether it will rain on a particular day or not. WEKA tool [28] is used for the implementation of experiments. The 10- fold cross validation test is chosen for the experiments which randomly split the data into training and test data. By applying various algorithms on the cleaned data set models are generated which are also known as classifiers. The percentage of correctly classified instances by the classifier (model) known as classification accuracy gives us the performance measure of the classifier (model). There are total 540 records in dataset. Each record has 5 attributes including the last attribute defines the class label of the record, whether it will rain or not.

A. Confusion Matrix

Prediction results are usually explained using confusion matrix and related performance measures. Confusion matrix is the matrix visualization of outcome of machine learning prediction model as shown in table II.

TABLE II A SAMPLE CONFUSION MATRIX

Predicted by Model	**Actual Labels**		
		YES	**NO**
	YES	True-Positive**(TP)**	False-Positive**(FP)**
	NO	False-Negative **(FN)**	True-Negative**(TN)**

As shown in table II, confusion matrix consists of two rows and two columns that consist of True Negatives, True Positives, False Positive and False Negative.

1. True-Positive (TP), are the number of instances which are actually positive and are also predicted positive by the model.
2. True-Negative (TN), are the number of instances which are actually negative and are also predicted negative by the model.
3. False-Positive (FP), are the number of instances which are actually negative and are predicted positive by the model. False-Negative (FN), are the number of instances which are actually positive and are predicted negative by the model.

B. Performance Measures

There are many performance measures for classification algorithms. In this work we have implemented following performance measures: Accuracy, Precision, Recall, F-measure, Receiver Operating Characteristic (ROC), root mean square error (RMSE), and mean absolute error (MAE).

1. Accuracy: Accuracy is the percentage of correctly classified modules [41]. It is one the most widely used classification performance metrics.

$$\text{Overall Accuracy} = \frac{TN+TP}{TP+FP+FN+TN}$$

2. True Positive (TP): TP is the number of correctly classified fault-prone modules. TP rate measures how well a classifier can recognize fault-prone modules. It is also called sensitivity measure.

$$\text{True Positive rate/Sensitivity} = \frac{\text{TP}}{\text{TP}+\text{FN}}$$

3. False Positive (FP): FP is the number of non-fault-prone modules that is misclassified as fault-prone class. FP rate measures the percentage of non-fault-prone modules that were incorrectly classified.

$$\text{False Positive rate} = \frac{\text{FP}}{\text{FP}+\text{TN}}$$

4. True Negative (TN): TN is the number of correctly classified non-fault-prone modules. TN rate measures how well a classifier can recognize non-fault-prone modules. It is also called specificity measure.

$$\text{True Negative rate/Specificity} = \frac{\text{TN}}{\text{TN}+\text{FP}}$$

5. False Negative (FN): FN is the number of fault-prone modules that is misclassified as non-fault-prone class. FN rate measures the percentage of fault-prone modules that were incorrectly classified.

$$\text{False Negative rate} = \frac{\text{FN}}{\text{FN}+\text{TP}}$$

6. Precision: This is the number of classified fault-prone modules that actually are fault-prone modules.

$$\text{Precision} = \frac{\text{TP}}{\text{TP}+\text{FP}}$$

7. Recall: This is the percentage of fault-prone modules that are correctly classified.

$$\text{Recall} = \frac{\text{TP}}{\text{TP}+\text{FN}}$$

8. F-measure: It is the harmonic mean of precision and recall. F-measure has been widely used in information retrieval [42].

$$\text{F-measure} = \frac{2 \text{ x Precision x Recall}}{\text{Precision} + \text{Recall}}$$

9. ROC: It is tool for comparing capabilities of classification model. It plots true positive rate on Y-axis and false positive rate on X-axis.

10. Mean Absolute error (MAE): Mean Absolute Error is the average of difference between actual and predicted value in all test cases.

11. Root Mean Square Error (RMSE): Root Mean Square Error is a measure of differences between values that are actually observed from thing which is being modeled or estimated and values predicted by a model or estimator.

12. Relative Absolute Error (RAE): It takes the total absolute error and normalizes it by dividing by the total absolute error of the simple predictor.

TABLE III PERFORMANCE MEASURE OF ALGORITHMS USING SRINAGAR WEATHER DATA

Algorithms	**Precision**	**Recall**	**F-Measure**	**ROC**	**MAE**	**RMSE**	**RAE**	**Accuracy**
Naïve Bayes	0.841	0.85	0.842	0.873	0.2002	0.3353	58.69%	85.01%
Simple Logistic	0.865	0.872	0.864	0.871	0.201	0.3151	58.92%	87.15%
IBK	0.844	0.847	0.845	0.777	0.1552	0.3897	45.51%	84.70%
Bagging	0.862	0.869	0.863	0.847	0.2094	0.325	61.38%	86.85%
PART	0.815	0.826	0.818	0.804	0.2175	0.3705	63.76%	82.56%
j48	0.825	0.832	0.828	0.739	0.2161	0.3811	63.37%	83.18%
Random Forest	0.874	0.878	0.875	0.878	0.1938	0.3138	56.82%	87.76%

The results of various machine learning algorithms are compared on the basis Accuracy, Precision, Recall, F-measure, ROC, RMSE, and mean absolute error (MAE), and weighted average. Prediction accuracy and performance measures of applied prediction models based on weatherdataset is shown in table 3 and is also graphically observed in figure- 3. Each value presented in the table 3 is the result of 10-fold cross-validation run.

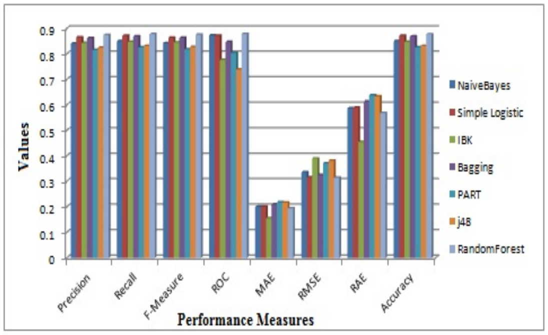

Fig. 3 Different Performance Measures

It can be seen from the results that RandomForest has the best prediction model as compared to other algorithms and is shown with the boldfaced letters in the table 3 above. The graph given in figure-3 shows the result of various classification algorithms and their performance measures.

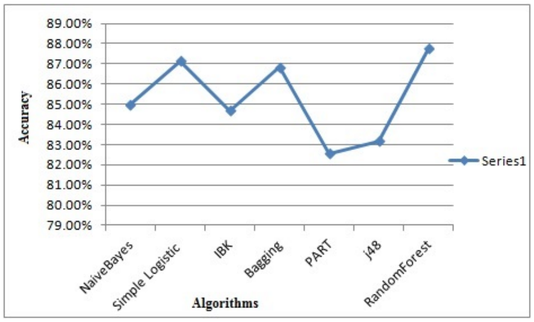

Fig. 4 Accuracy of Various Algorithms

From figure 4, we can observe that accuracy value of RandomForest is highest as compared to other data mining algorithms. The different values of Precision, Recall, F-Measure, ROC, MAE, RMS, RAE and Accuracy for given weather dataset is shown in table 3. It can be observed that out of seven classification algorithms, RandomForest exhibits highest values of Precision, Recall, F-measure, ROC and Accuracy. RandomForest also produces minimum amount of Root Mean Square Error (RMSE) among all the eight algorithms used.

VI. CONCLUSION

In this work we carried out an experimental work to compare popular data mining algorithms for rainfall prediction using various performance measures over weather data of Srinagar, J&K, India. The different measuring attributes play a pivotal role in giving precise rainfall prediction. We have observed that RandomForestproduces best rainfall prediction results with an accuracy of 87.76% and also exhibits highest values in Recall, F-Measure and ROC as compared to other classification algorithms. In our case, RandomForest approach proves to be an efficient and acceptable method for rainfall prediction. The level of accuracy and prediction highly depends on the data being used as input for classification and prediction. Every algorithm has its advantages and limitations; it is difficult to choose the best algorithm. The prediction accuracy of the model can be increased by developing a hybrid prediction model where multiple machine learning algorithms are put to work together. For our weather dataset, it was concluded after analyzing various models of supervised learning that the RandomForest classification algorithm has appreciable level of accuracy and acceptance.

REFERENCES

[1] Olaiya, Folorunsho and Adesesan Barnabas Adeyemo, “Application of data mining techniques in weather prediction and climate change studies”, *International Journal of Information Engineering and Electronic Business,* Vol. 4, No. 1, pp. 51, 2012.

[2] Sawaitul, D. Sanjay, K. P. Wagh and P. N. Chatur, “Classification and prediction of future weather by using back propagation algorithm-an approach”, *International Journal of Emerging Technology and Advanced Engineering,* Vol. 2, No. 1, pp. 110-113, 2012.

[3] M. Kannan, S. Prabhakaran and P. Ramachandran, “Rainfall forecasting using data mining technique”, 2010.

[4] Retrieved from https://en.wikipedia.org/wiki/Srinagar.

[5] M. Kannan, S. Prabhakaran and P. Ramachandran, “Rainfall forecasting using data mining technique”, 2010.

[6] Nikam, B. Valmik and B. B. Meshram, “Modeling rainfall prediction using data mining method: A Bayesian approach”, Computational Intelligence, Modelling and Simulation (CIMSim), *2013 Fifth International Conference on. IEEE,* 2013.

[7] Nhita, Fhira, “A rainfall forecasting using fuzzy system based on genetic algorithm”, *Information and Communication Technology (ICoICT), 2013 International Conference of IEEE,* 2013.

[8] Mahajan, Seema and HimanshuMazumdar, “Rainfall Prediction using Neural Net based Frequency Analysis Approach”, *International Journal of Computer Applications,* Vol. 84, No. 9, 2013.

[9] A. Geetha, and G. M. Nasira, “Data mining for meteorological applications: Decision trees for modeling rainfall prediction”, *Computational Intelligence and Computing Research (ICCIC), 2014 IEEE International Conference on. IEEE,* 2014.

[10] Dutta, Pinky Saikia and Hitesh Tahbilder, “Prediction of rainfall using data mining technique over Assam”, *IJCSE,* Vol. 5, No. 2, 2014, pp. 85-90.

[11] Sharma, Ashutosh and Manish Kumar Goyal, “Bayesian network model for monthly rainfall forecast”, *Research in Computational Intelligence and Communication Networks (ICRCICN), 2015 IEEE International Conference on IEEE,* 2015.

[12] Dubey and D. Akash, “Artificial neural network models for rainfall prediction in Pondicherry”, International Journal of Computer Applications, Vol. 120, No. 3, 2015.

[13] Retrieved from https://www.wunderground.com/history/airport/ VISR/2015/11/6/CustomHistory.html?dayend=6&monthend=11&yearend=2016&req_city=&req_state=&req_statename=&reqdb.zip=&reqdb.magic=&reqdb.wmo=

[14] Ahmed, Bilal, “Predictive capacity of meteorological data: Will it rain tomorrow?”,*Science and Information Conference (SAI), 2015, IEEE,* 2015.

[15] D. Hand, H. Mannila and P. Smyth, “Principles of data mining”, MIT, 2001.

[16] Kotu, Vijay and BalaDeshpande, Predictive analytics and data mining: concepts and practice with rapidminer, Morgan Kaufmann, 2014.

[17] T. Wang, W. Li, H. Shi and Z. Liu, “Software defect prediction based on classifiers ensemble”, *Journal of Information & Computational Science*, Vol. 8, No. 16, pp. 4241-4254, 2011.

[18] I. H. Witten and E. Frank. Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann, Los Altos, US, 2005.

[19] R. O. Duda and P. E. Hart, Pattern classification and scene analysis, John Wiley and Sons, 1973.

[20] P. Langley, W. Iba and K. Thompson, “An analysis of Bayesian Classifiers”, *in Proceedings of the Tenth National Conference on Artificial Intelligence,* San Jose, CA, 1992.

[21] A. Mccallum and K. Nigam, “A Comparison of Event Models for Naive Bayes Text Classification”, Proceedings of the 15th National Conference on Artificial Intelligence (AAAI-98)-Workshop on Learning for Text Categorization, pp. 41-48, 1998.

[22] T. Menzies, J. Greenwald and A. Frank, “Data Mining Static Code Attributes to Learn Defect Predictors”, *IEEE Transactions on Software Engineering,* Vol. 33, No. 1, 2-13, 2007.

[23] L. Breiman, “Random forests”, Machine Learning, Vol. 45, No. 1, pp. 5-32, 2001.

[24] L. Guo, Y. Ma, B. Cukic and H. Singh. Robust prediction of fault-proneness by random forests, *In Proc. of the 15th International Symposium on Software Relaibility Engineering ISSRE’04,* pp. 417-428, 2004.

[25] Y. Jiang, B. Cukic, T. Menzies and N. Bartlow, “Comparing design and code metrics for software quality prediction”, *Proc. Fourth Int. Workshop on Predictor Models in Software Engineering,* PROMISE’08, New York, USA, 2008, pp. 11-18.

[26] Retrieved from https://en.wikipedia.org/wiki/Logistic_regression.

[27] D. Aha, “Tolerating noisy, irrelevant and novel attributes in instance-based learning algorithms”, *Int. J. Man-Machine Studies,* Vol. 36, 267-287, 1992.

[28] R. Mohammad, M. Butt Ahmed and M. Baba Zaman, “Tools for Predictive Analytics : An Overview”, *International Journal of Scientific Research Engineering & Technology (IJSRET),* Vol. 6, No. 7, pp. 748-750, 2017.

[29] R. M.Shah, M. A. Butt and M. Z. Baba, “Predictive Analytics Modeling: A Walkthrough*”, Int. J. Adv. Res. Comput. Sci. Softw. Eng.,* Vol.7, No.6, pp. 421-426, June 2017.

[30] R. M. Shah, M. A. Butt and M. Z. Baba, “ Review of Predictive Analytic Modeling techniques”, *International Journal of Emerging Trends & Technology in Computer Science (IJETTCS),* Vol. 6, No. 4, pp. 58-62, 2017.

[31] R. Mohammad, M. Butt Ahmed and M. Baba Zaman, “Predictive Analytics: An Application Perspective”, *International Journal of Computer Engineering and Applications,* Vol. 9, No. 8, Aug. 2017.