Weather Forecasting Using Decision Tree

DOI : 10.36909/jer.ICCEMME.15635

Rashmi Bhardwaj* and Varsha Duhoon

University School of Basic and Applied Sciences, Nonlinear Dynamics Research Lab, Guru Gobind Singh Indraprastha University, Sec-16C, Dwarka, New Delhi-10078

E-mail: rashmib22@gmail.com

Abstract

The present situation of climate change calls for the need to develop a system which can forecast temperature in advance so that it can be helpful for making polices by government. Decision Trees is a method used to analyse combination of Mathematical & Computational Techniques in order to make description, categorisation & generalisation of given set of data using machine learning to predict behaviour for future performance. In this study the Decision Tree techniques like Quinlan's M5 algorithm (M5P), Reduced Error Pruning Tree (REP Tree), Random Forest, Logit Boosting, Ada Boosting M1Tool are used to analyse the weather parameter Maximum Temperature and Minimum Temperature for Delhi region. Daily data set of 17 months has been used for analysing and forecasting. It is observed that among the techniques of decision tree; Random Forest is much effective than statistical methods as it gives better results in less time and less statistical errors.

Keywords: Ada Boosting M1; Decision Tree; Logit Boosting; M5P; Optimisation; REP Tree; Random Forest

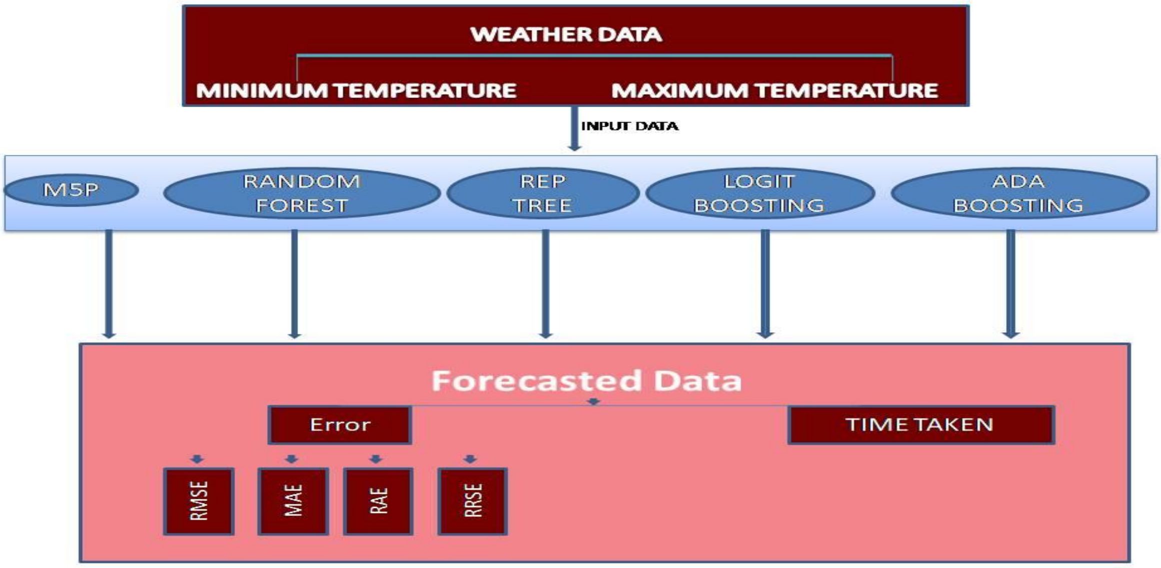
Introduction

In India, where the major share of economy is agriculture based as a result of which large share of the economic growth in terms of GDP depend on weather. An optimising model is an important characteristic to solve many problems which can be of the type: business oriented, society related or can be resource allocation problem. A mathematical modelling in which either objective function or constraints or both are in non linear form is called non linear mathematical modelling. Data mining is the process of analysing, processing and studying the patterns and relations among the huge data sets in order to forecast the future values. Data mining techniques involves different tools and techniques for the study of time series or data. In the present study the application of different tools of decision tree has been studied for weather parameters.

Decision Tree is an optimising technique used in order to make decisions on the basis of tree structure graphs or model by analysing possible events and hence possible chances of outcomes. Mathematical programming or Optimisation techniques are applied to select the best element from the available sets of alternatives. Optimisation technique is a method under Data Mining. Decision Trees is a combination of mathematical & computational techniques in machine learning. Ashwini et.al. (2018) forecasted rainfall on monthly basis using multiple linear regression technique [1]. Bhardwaj et.al. (2010) forecasted bias-free rainfall and temperature using T-170 model in monsoon season [2]. Among the initial studies Breiman (1984) using CART method studied trees, to increase size without pruning [3]. It was further discussed by Breiman et.al (1996, 2000, 2001, 2004) that gains in classification and regression can be collected using collection of trees such that every tree is the collection growing according to a studied random value. The average of the ensembles are useful in calculating the forecast. Collection includes tree structures as forecasters, all trees are formed by providing randomness, hence called “random forest” [4,5,6,7]. For monthly prediction Petre (2009) studied DT for representation and interpretation using CART[21]. New metric named information gain is applied for choosing the attributes for classifying set in study of Chandar (2013) on ID3 [8]. Since 1990’s decision tree has been studied in terms of the classification and regression approach by Svetnik et.al. (2003), Diaz et.al (2006), Genuer et.al. (2008,2010) [27, 9, 12, 13]. Adaptive Nearest Neighbour method and relation among random forest were studied by Lin et. al (2006) [18]. Weather Forecasting Using Machine Learning Algorithm was done by singh et. Al. (2019) [20]. Multi-Dimension-Model was used by Durai et.al. (2014) for rainfall prediction[10]. The prediction of weather was done by Dhamodaran et. Al. (2020) using Random Forest Algorithm and GIS Data Model [11]. The impact of weather on the human moods was studied by Howarth et.al. (1984) using the linear regression and canonical correlation analysis which included 10 mood variables which were related to the 8 weather parameters for multi-dimensional study [14]. The methods namely kNN, DT, Navie Bayes were compared by Khan et. al. (2014) in their work and it was showed that DT gave desirable results comparatively [16]. On similar lines Kaur (2012) compared and studied CART, Quest, C5.0, CHAID in forming a decision tree [15]. A different study of analysing the effect of using heat storage, aimed at the night shift operation was studied by Hardi et. al. in the form of absorber type optimization [30]. The optimisation technique applied by Paul et. al. for global pipeline design optimisation and reducing the cost for pipeline service has been studied [31]. In past along with weather parameters various other kinds of system parameters in other natural entities like air and water have been extensively studied using different tools for forecasting purpose. Lawrence et. al. (2005) studied humidity & Dew Point relation [17]. Manikandan et. al. (2017) studied weather and forecasted using C4.5[19]. Pasanen et.al. (1991) studied the growth simulation of 2 types of fungi in the presence of air temperature & relative humidity [19]. Using fractals water data was analysed by Parmar et. al. (2013) [23]. Rajesh (2013) studied weather parameters using optimising techniques [25]. Further Srivastava et. al. (2014) studied thunderstorm & cloudburst event [26]. Vathsala et.al. (2015) studied land& ocean variables using data mining techniques [28]. Xi et. al. (2012) studied genomic using random forest [29]. Quinlan (1986) studied different approaches to analyse decision tree which has been used in different areas and systems among which ID3 has been studied in detail to study ways to handle data with noise or which has missing values [24]. Earlier many attempts have been made to predict the temperature data in various regions using different mathematical models and optimising techniques. The present study contributes to selection and analysing efficient tool and less cost-effective tool for accurate prediction of weather paramters by analysing decision tree tools such as M5P tree, Random Forest, REP tree, Logit Boosting, Ada Boosting M1 and then choosing better tool among all on the basis of statistical errors and time taken.

Modelling and Analysis

Decision Tree: The algorithm which includes condition control statement is termed as decision tree. It helps to identify a strategy or method most likely to reach objective. Decision tree (DT) in the form of chart structure has internal nodes showing "test" on an attribute and branches shows outcomes of experiment and every node of leaf implies class label. Path from-roots-to-leaf shows classification rules. There are three nodes: Decision Nodes, Chances Nodes, End Nodes.



Flow Chart 1: The schematic diagram of weather data.

The following decision tree tools have been studied:

- ***M5P:*** M5P is the modulation of Quinlan’s M5 algorithm for inducing tree of regression models. M5P merges Convection DT with existence of Linear Regression function at the nodes. Multiple Regression models have been fitted for each node. M5P is many valued tree algo., which is preferred due to its following qualities: estimation of error, linear model’s simplification, pruning, Smoothening. Tree pruning begins from the base of the tree and implemented for each non-leaf node. M5’ is induced to generate “TREE”; Then after collapsing into smaller set of if-then rules by either adding or subtracting paths for roots to terminal nodes.
- ***REP Tree:*** REP Tree Classifier reduces Error Pruning (REP) tree-classifier is an algorithm which works fast on the principle of collecting information Gain by entropy and minimising disturbances arising from variance. REP applies Regression Tree (RT) Logic and generates Multiple Trees in altered iterations. REP Tree is a learner which is efficient in deciding Decision Tree or RT by collecting information gain as splitting criterion and cut back by applying reduced error pollard. It classifies values for numeric characteristic. Missing values are dealt by using C4.5’s method of using fractional instances.

The Algorithm works as follows:

- If all situations belong to same Class, tree is leaf and further leaf is again labelled Class.
- For every value, calculate required details obtained from the test parameters, further calculate “GAIN” of information obtained from test on parameters.
- Based on selection-criterion find appropriate-parameter on branch.

Gain, is a process including ENTROPY; which means existence of disordering of data. The entropy can be calculated as follows:

$$\frac{\sqrt{\frac{\sum abs(\beta-\alpha)^2}{N}}}{\sqrt{\frac{\sum abs(\gamma-\alpha)^2}{N}}} \qquad (1)$$

Iterating over All Possible value of $\vec{A}$. Conditional Entropy:

$$ENTROPY(K|\overrightarrow{A}) = \left|\frac{A_K}{\vec{A}}\right| \log \left|\frac{A_K}{\vec{A}}\right| \quad (2)$$

GAIN is given by: $GAIN(\overrightarrow{K,A}) = ENTROPY\left(\vec{A} - ENTROPY(K|\overrightarrow{A})\right)$ (3)

The objective is to maximise GAIN. Pruning is a step applied due to outliners. All data set includes few observations which are not well defined and are different than the others in the neighbourhood. Ones the tree has been built it must classify all observations in training set in which it is pruned. Objective is to reduce classification error, to make tree more general.

- ***Random Forest*:** Random forest is combination of randomise based regression trees such as $\left\{a_q(p, \alpha_n, \Delta_m), n \geq 1\right\}$ in which $\alpha_1, \alpha_2.....$ are results of randomising variable α . The trees combine to form average of regression estimate

$$\bar{a}_q(P,\alpha) = \left[\frac{\sum_{j=1}^{m} Z_i 1_{[P_i \in B_m(p,\alpha)]}}{\sum_{j=1}^{m} 1_{[P_i \in B_m(p,\alpha)]}} \bullet 1_{E_m(P,\alpha)}\right] \quad (4)$$

Such that E_α - expectation w.r.t random parameters; Δ_m - data set; α - randomising variable.

$B_m(P,\alpha)$ be rectangle cell of random partition having P; such that $P = (P^{(1)}, ..., P^{(i)})$.

$$a_m(P,\alpha) = \frac{\sum_{j=1}^{m} Z_i 1_{[p_i \in B_m(p,\alpha)]}}{\sum_{j=1}^{m} 1_{[p_i \in B_m(p,\alpha)]}} \bullet E_N(P,\alpha) \quad (5)$$

Further, random forest regression takes the form of: $\bar{a}_q(P) = E_\alpha[a_m(P,\alpha)] = E_\alpha\left[\frac{\sum_{j=1}^{m} Z_i 1_{[P_i \in B_m(p,\alpha)]}}{\sum_{j=1}^{m} 1_{[P_i \in B_m(p,\alpha)]}} \bullet 1_{E_m(P,\alpha)}\right]$ (6)

- ***Logit Boosting*:** Logit boosting is a boosting algorithm in which logistic regression is applied to the cost function, the formulation is: $f = \sum_t x_t q_t$.

The logistic loss can be reduced by: $\sum_{j} \log(1 + e^{-z_j f(p_i)})$ (7)

- ***Ada Boosting M1*:** It refers to additive boosting method, which is sensitive to noisy data. The booster is of the form: $X_A(y) = \sum_{n=1}^{A} f_n(y)$. Such that: f_n is a learner values from “y” as input and represents the class of the object. “A” is positive or negative based on the data input.

The Statistical calculations done are:

Mean Absolute Error (MAE): MAE calculates the amount of error among the forecasted values. MAE is used to measure efficiency of continuous variable. Average of difference among observed values and Actual Values is calculated. It calculates gap between model’s prediction and actual points. $MAE = \sum_{i=1}^{k} \frac{abs(X_i - \alpha(X))}{n}$. X_i = Predicted value; X = Actual value; N = No. of terms

Root Mean Squared Error (RMSE): Average magnitude of errors is measured using RMSE. Errors are squared before they are averaged. It can also be calculated by taking the square root of MSE. RMSE = $\sqrt{\frac{1}{N}\sum_{i=1}^{N}(X_i - x_i)^2}$ x_i =Predicted values; X_i = Actual value; N = No. of observations

Root Relative Squared Error (RRSE): It is used to calculate error which is normalised by taking the absolute of the difference and further square root. RRSE = $\frac{\sqrt{\frac{\sum abs(x_i - X_i)^2}{N}}}{\sqrt{\frac{\sum abs(\gamma - X_i)^2}{N}}}$

x_i =Predicted values; X_i = Actual value γ = Previous target value; N = No. of Observation

Mean Squared Error (MSE): Measures average squared error between actual value and predicted value. $MSE = \frac{1}{N}\sum_{i=1}^{N}(X_i - x_i)^2$. x_i =Predicted values; X_i = Actual value; N = No. of Observation

Mean Absolute Percentage Error (MAPE/MAPD): Measures percentage of difference between actual input & predicted output. Result is expressed in percentage. $MAPE = \frac{100\%}{N}\sum_{i=1}^{N}\frac{abs(X_i - x_i)}{abs(X_i)}$. x_i =Predicted values; X_i = Actual value; N = No. of Observation

Daily data of Temperature, rainfall, bright sun shine, evaporation, relative humidity, wind speed for Delhi with coordinates Longitude 77^0 09’ 27’’ Latitude 28^0 38 ‘23’’ N Altitude :228.61m has been taken from 1 January 2017 to 31May 2018.

The time series of Maximum and Minimum Temperature:

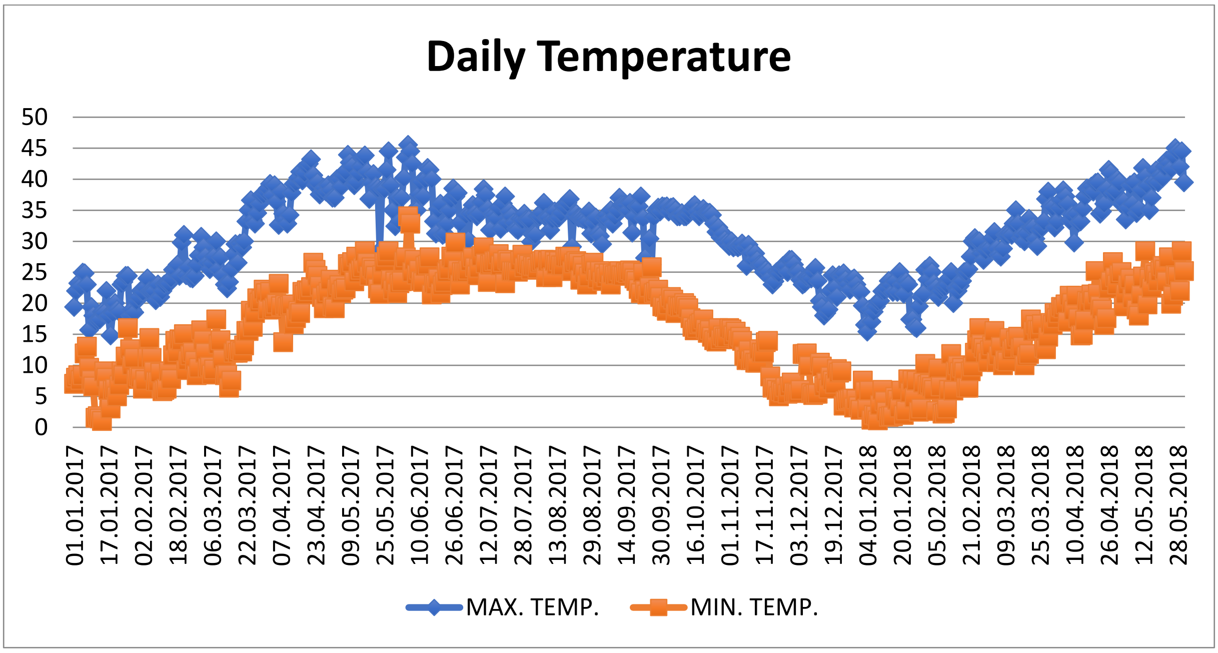


Figure 1: Time series plot of Maximum and Minimum Temperature 01.01.2017-31.05.2018

Table 1: Statistical Comparison for Minimum Temperature

MODEL	TIME TAKEN	MAE	RAE	RRSE	RMSE	MAPE
M5P	0.78	6.78	100.21	100.31	0.6938	0.6148
RANDOM FOREST	0.01	4.37	64.62	66.40	6.2964	0.3888
REP TREE	0.02	6.77	100	100.01	11.4142	0.6215
LOGIT BOOSTING	1.29	5.0104	99.7211	108.6237	11.0782	0.9788
ADABOOSTING M1	0.05	6.0103	100.2532	99.9797	9.0722	0.7887

From above table it can be analyzed that Random Forest shows least time for modeling and shows least error.

Table 2: Statistical Comparison for Maximum Temperature

MODEL	TIME TAKEN	MAE	RAE	RRSE	RMSE	MAPE
M5P	0.74	5.63	97.45	97.64	16.7595	0.1909
RANDOM FOREST	0.01	3.54	61.27	63.51	7.6853	0.1171
REP TREE	0.03	5.78	100.01	100.03	7.7854	0.1932
LOGIT BOOSTING	0.08	6.0121	99.9387	100.2526	12.078	0.1999
ADABOOSTING M1	1.54	5.012	99.3994	110.8701	8.0863	0.1989

From above table it can be analyzed that Random Forest shows least time for modeling and shows least error.

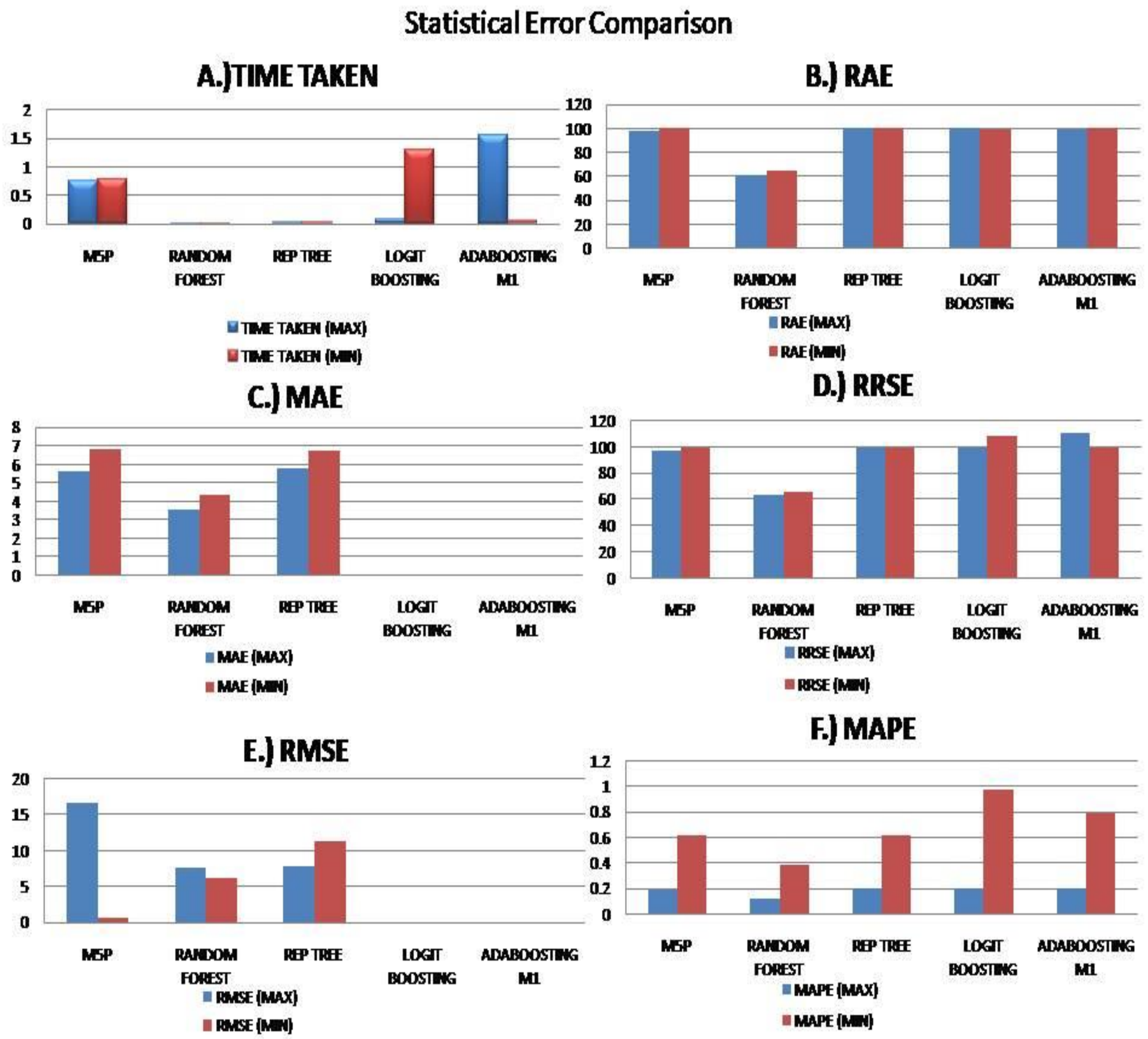


Figure 2: Statistical Error & time taken comparison by the tools of Decision Tree

Table 3: Error calculated among Predicted and Actual Maximum Temperature

DATE	Predicted Max. Temp. (M5P)	Error	Predicted Max. Temp. (Random Forest)	Error	Predicted Max. Temp. (REP)	Error	Actual Max. Temp.	Predicted Max. Temp. (Adaboost M1)	Error	Predicted Max. Temp. (Logit Boosting)	Error
01.06.2018	39.0	6.0	40.53	-4.47	44.2	0.8	45.0	48.8	-3.8	44.2	0.8
02.06.2018	36.3	6.9	40.93	-2.27	40.4	2.8	43.2	40.1	3.1	38.2	5
03.06.2018	44.2	-1.7	41.75	-0.75	40.6	1.9	42.5	38.2	4.3	36.2	6.3
04.06.2018	38.0	1.8	41.9845	2.1845	45.8	-6.0	39.8	35.6	4.2	39	0.8
05.06.2018	43.5	-2.5	42.332	1.332	44.6	-3.6	41.0	38.2	2.8	36.2	4.8

06.06.2018	45.4	-2.4	42.202	-0.798	48.2	-5.2	43.0	35.4	7.6	39.8	3.2
07.06.2018	47.6	-3.8	43.0093	-0.7907	46.8	-3.0	43.8	35.2	8.6	44.1	-0.3

Table 4: Error calculated among Predicted and Actual Minimum Temperature

DATE	Predicted Min. Temp. (M5P)	Error	Predicted Min. Temp. (Random Forest)	Error	Predicted Min. Temp. (REP)	Error	Actual Min. Temp.	Predicted Max. Temp. (AdaBoosting M1)	Error	Predicted Max. Temp. (Logit Boosting)	Error
01.06.2018	30.2	-3.4	24.9925	-1.8075	29.8	-3	26.8	29.8	-3	30.9	-4.1
02.06.2018	34.6	-2.9	25.2843	-6.4157	32.8	-1.1	31.7	30.9	0.8	24.9	6.8
03.06.2018	38.4	-8.5	25.57	-4.33	29.2	0.7	29.9	24.5	5.4	22.1	7.8
04.06.2018	39.2	-10.0	23.85	-5.35	20.0	9.2	29.2	23.3	5.9	34.2	-5
05.06.2018	36.1	-5.8	24.48	-5.82	34.3	-4	30.3	30.9	-0.6	32.1	-1.8
06.06.2018	29.2	-1.2	24.08	-3.92	39.6	-11.6	28	29.9	-1.9	30.3	-2.3
07.06.2018	29.0	-4	24.14	-0.86	30.2	-5.2	25	29.9	4.9	30.3	5.3

Conclusion

Decision Tree is the supervised form of learning as it has the fastest computation speed also it is cost effective and also decision tree is less complex. The study of analysing which tool in decision tree is fastest and accurate for weather prediction concludes that random forest is much efficient and takes least time for model formation. Results of minimum, maximum temperature were close to forecasted values using Random Forest. Hence it can be concluded that the efficient tool for forecasting is Random Forest; due to least prediction error in terms of MAE, RAE, RRSE, MSE, MAPE and time taken among others.

References

1. **Aswini R., Kamali D., Jayalakshmi S., Rajesh R., 2018.** Predicting Rainfall and Forecast Weather Sensitivity using Data Mining Techniques. *International Journal of Pure and Applied Mathematics*. 119(14).: 843-847.
2. **Bhardwaj R., Kumar A., Maini P., Kar S.C., Rathore L.S., 2010.** Bias-free rainfall forecast and temperature trend- based temperature forecast using T-170 model output during the monsoon season. *Meteorological Applications. Royal Meteorological Society*. 14(4).: 351-360.
3. **Breiman L., Friedman J. H., Olshen R.A., Stone C. J., 1984.** Classification and Regression Tree., *Chapman & Hall/CRC Taylor&Franci.*, New York.
4. **BreimanL. 1996.** Bagging predictors. *Machine Learning.* 24(2).: 123–140.
5. **Breiman L. 2000.** Some Infinity Theory for Predictor Ensembles. *Technical Report 577, UC Berkeley*. 1-30.
6. **Breiman L., 2001.** Random forests. *Machine Learning.* 45(1).: 5–32.
7. **Breiman L., 2004.** Consistency for a Simple Model of Random Forests. *Technical Report* 670.: 1-10.
8. **Chandar S. 2013.** An intelligent application of fuzzy id3 to forecast seasonal runoff. *International Journal on Cybernetics & Informatics*. 2(1).: 17-22.
9. **Diaz-Uriarte R., De Andres. S.A. 2006.** Gene selection and classification of microarray data using random forest. *BMC Bioinformatics.* 7.: 1471–2105.
10. **Durai V.R., Bhardwaj R., 2014.** Forecasting quantitative rainfall over India using multi-model ensemble technique. *Meteorology and Atmospheric Physics.* 126(1-2).: 31-48.
11. **Dhamodaran, Sridhar, Varma, Ch & Reddy, Chittepu., 2020.** Weather Prediction Model Using Random Forest Algorithm and GIS Data Model. 10.1007/978-3-030-38040-3_35.

12. **Genuer R., Poggi J.-M., Tuleau C., 2008.** Random Forests: Some Methodological Insights. arXiv:0811.3619.: 1-35.
13. **Genuer R., Poggi J. M., Tuleau-Malot C., 2010.** Variable selection using random forests. *Pattern Recognition Letters*. 31.: 2225–2236.
14. **Howarth, E., Hoffman M. S., 1984.** A multidimensional approach to the relationship between mood and weather. *British Journal of Psychology*. 75(1).: 15-23.
15. **Kaur G., 2012.** Meteorological data mining techniques: A survey. *International Journal of Emerging Technology and Advanced Engineering*. 2(8).: 325-327.
16. **Khan Z. U., Hayat M., 2014.** Hourly based climate prediction using data mining techniques by comprising entity demean algorithm. *Middle-East Journal of Scientific Research*. 21 (8).: 1295-1300.
17. **Lawrence, Mark G., 2005.** The relationship between relative humidity and the dewpoint temperature in moist air: A simple conversion and applications. *Bulletin of the American Meteorological Society*. 86(2).: 225-233.
18. **Lin Y., Jeon Y., 2006.** Random forests and adaptive nearest neighbours. *Journal of the American Statistical Association.* 101.: 578–590.
19. **Manikandan M., Mala R., 2017.** Optimal Prediction of Weather Condition Based on C4.5 Classification Technique. *International Journals of Advanced Research in Computer Science and Software Engineering*. 7(8).: 265-272.
20. **N. Singh, S. Chaturvedi and S. Akhter**, **2019.** Weather Forecasting Using Machine Learning Algorithm *2019 International Conference on Signal Processing and Communication (ICSC)*. doi: 10.1109/ICSC45622.2019.8938211.: 171-174.
21. **Pasanen, A.L.,** et al. **1991.** Laboratory studies on the relationship between fungal growth and atmospheric temperature and humidity. *Environment International*. 17(4).: 225-228.

22. **Petre E. G., 2009.** A decision tree for weather prediction. *SeriaMatematicǎ- Informaticǎ– Fizicǎ*. 11(1).: 77 – 82.

23. **Parmar K.S., Bhardwaj R., 2013.** Water quality index and fractal dimension analysis of water parameters. *International Journal of Environmental Science and Technology*. 10(1).: 151–164.

24. **Quinlan, J. R.**, **1986.** Induction of decision trees. Machine learning. *Kluwer Academic Publishers*. 1.: 81-106.

25. **Rajesh K., 2013.** Decision Tree for the Weather Forecasting. *International Journal of Computer Applications* (0975 – 8887). 76(2).: 31-34.

26. **Srivastava K., Bhardwaj R., 2014.** Real-time nowcast of a cloudburst and a thunderstorm event with assimilation of Doppler weather radar data. *Natural hazards*. 70(2).: 1357-1383.

27. **Svetnik V., Liaw A., Tong C., Culberson J., SheridanR., Feuston B. 2003.** Random forest: A classification and regression tool for compound classification and QSAR modelling. *Journal of Chemical Information and Computer Sciences*. 43.: 1947–1958.

28. **Vathsalaa H., Koolagudib S. G., 2015.** Rainfall A Data Mining Model with Land and Ocean Variables as Predictors. *Procedia Computer Science*. 54.: 271 – 280.

29. **Xi, C., Hemant I., 2012.** Random forests for genomic data analysis. *Genomics*. 99(6).: 323-329.

30. **Hardi A. M. R., Idress A. H., Ali M. I., 2021.** Absorber type Optimization for Night-Shift Operation of Solar air Heater. *Journal of Engineering Research*. 9.: https://doi.org/10.36909/jer.v9iICRIE.11673.

31. **Paul, Abdulrahman A., 2021.** Comprehensive Optimisation of project cost for long supply pipelines. *Journal of Engineering Research*. 9(3A).: 14-28.