Prediction of Rainfall Using Data Mining Techniques

Tharun V.P
School of Electronics Engineering
VIT University
Vellore, Tamil Nadu, India

Ramya Prakash
School of Electronics Engineering
VIT University
Vellore, Tamil Nadu, India

S. Renuga Devi
Associate Professor
School of Electronics Engineering
VIT University
Vellore, Tamil Nadu, India

***Abstract*- The occurrence of rainfall is an outcome of various natural factors such as temperature, humidity, cloudiness, wind speed, etc. Rainfall prediction is a major concern for meteorological department as it is closely associated with the economy and sustenance of human life. In this work, we use regression techniques and statistical modelling to predict the rainfall intensity of Coonoor in Nilgiris district, Tamil Nadu. It is a comparative study of various regression techniques based on the Relative error. The regression techniques used for prediction are Support Vector Regression (SVR), Random forest (RF) and Decision Tree (DT). The parameters considered for training the model includes the daily recorded temperature, humidity, cloud speed, wind speed and wind direction of Coonoor. The rainfall prediction model was made more efficient by including the rainfall intensities of nearby stations within an area of 7 km^2. The developed forecasting models were analysed on the basis of R-square and Adjusted R-square values. A statistical model was developed out of all the techniques by generating the regression equation used for prediction of rainfall by each of the model. The proposed models were implemented in Python platform. The prediction model developed by the RF regression technique was found out to be a better and efficient model compared to SVR and DT models.**

Index Terms—Prediction; Decision Tree; Random Forest; Regression; Statistical model; Support Vector Regression.

I. INTRODUCTION

Rainfall is a natural phenomenon defined as the outcome of interaction between several complex atmospheric processes. A large uncertainty involved in determining the contribution of the atmospheric processes is one of the biggest challenges to face in developing rainfall prediction models. Rainfall prediction is very difficult to model since the atmospheric processes involved follow a rather complex nonlinear pattern. The temperature, relative humidity, wind speed, wind direction, cloud coverage etc. are some of the factors that critically affect the occurrence of rainfall. The rainfall forecast is essential information to support crop, water and flood management. It also has a vital role to play in disaster management on occurrence of landslides caused due to heavy downpour. Rainfall prediction models can help in saving the lives and properties of the people, which indirectly support the economy of the country.

To provide a good and accurate prediction, prediction models have been developed and implemented using statistical modelling and regression techniques. This paper provides a comparative study of statistical modelling and regression techniques such as SVR, RF and DT on the basis of accuracy of prediction. Statistical modelling fails to provide good accuracy for rainfall prediction as compared to regression techniques due to dynamic nature of atmospheric composition. The objective of the study is to predict rainfall intensity of Coonoor in Nilgiris district, Tamil Nadu using statistical modelling and regression techniques.

The methodology of this study is as follows:

(a) Data Collection
(b) Data Pre-Processing
(c) Model Development
(d) Performance Measure

The collected data is pre-processed using data pre-processing techniques like Imputation, Feature Scaling, Categorical Encoding and Normalisation. R-square and Adjusted R-square values are used as parameters to evaluate the model performance.

II. BACKGROUND STUDY

Regression technique is a form of predictive modelling which investigates the relationship between the target and predictors. It’s extensively used for forecasting, time series modelling and to find causal effect between the variables. The regression techniques used in this study are as follows:

A. SUPPORT VECTOR REGRESSION

Support Vector Regression uses the same principle as Support Vector Machine which is a supervised machine learning algorithm used for classification and regression problems. It decides on the right hyper-plane to maximize the margin and minimize the error.

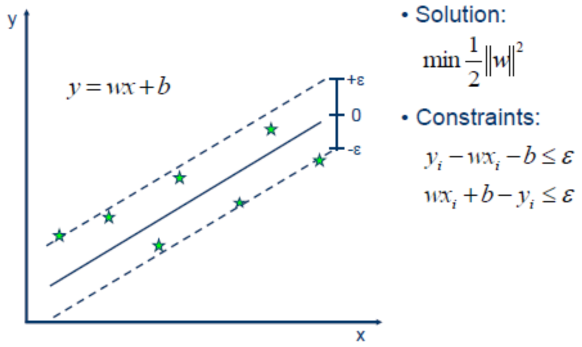
y
x
$y = wx + b$
+ε
0
-ε
• Solution:
$\min \frac{1}{2}\|w\|^2$
• Constraints:
$y_i - wx_i - b \le \varepsilon$
$wx_i + b - y_i \le \varepsilon$

Figure 1. *Support Vector Regression Technique*

B. DECISION TREE

Decision tree learning algorithm uses decision tree as a predictive model by breaking the dataset into smaller subsets so that each internal node holds an attribute. It contains a root node, branches and leaf nodes [3].

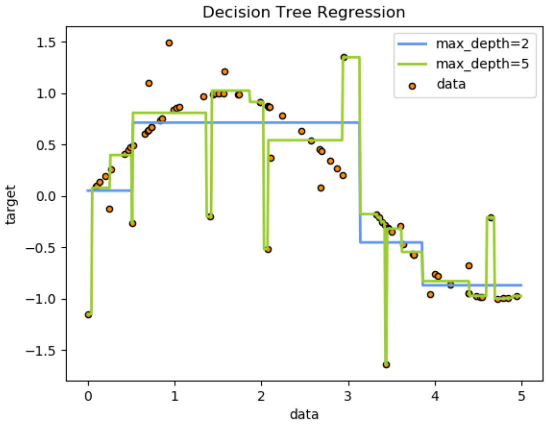

Figure 2. *Decision Tree Regression Technique*

C. RANDOM FOREST

Random decision forests are an ensemble learning method for regression that uses the mean prediction of the individual trees for predictive modelling. Random forests are constructed by large number of decision trees at the time of training [2].

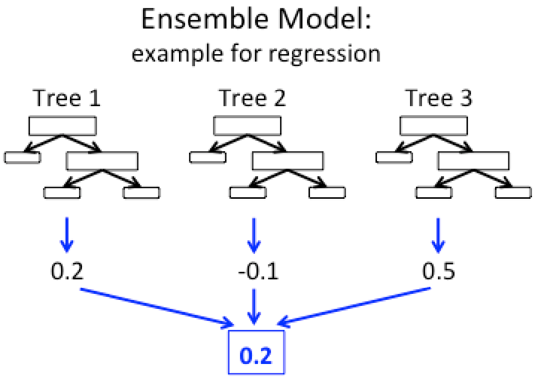

Figure 3. *Random Tree Regression Technique*

III. METHODOLOGY

A. DATASET

The area of study is Coonoor in Nilgiris district, Tamil Nadu, India. Coonoor is a Taluk and a municipality in the Nilgiris district. It is located at an altitude of 1850 m above sea level and is the second largest hill station in the Nilgiri hills after Ooty. There are 14 rain gauge stations in and around Coonoor, which keep a record of the amount of rainfall received.

The data have been collected from the India Meteorological Department (IMD), Chennai and the Public Works Department of Coonoor for a period of 9 years from 2005-2014 [8]. The rainfall intensity of 3 nearby stations such as Coonoor rain gauge station, Coonoor railway station and Runnymedu are also considered as the input parameters.

B. DATA PRE-PROCESSING

The pre-processing of data includes imputing data, categorical encoding, feature scaling, normalisation and splitting dataset. Imputing data is a measure of filling out the missing data using mean of the column, median or the most frequently used value. Imputing data is an important step as the dataset with missing information may wrongly train the model.

Categorical encoding is a way of converting strings into binary forms followed by one hot encoding which converts the strings into 0s and 1s is performed to ensure that no information is lost. In the dataset considered, wind direction has been categorically encoded using one hot encoding.

Feature scaling and normalisation scales the data into the range of -1 to 1 or 0 to 1. If all the features have largely varying values then feature with higher values dominates other features. Feature scaling results in better convergence and less computational time as compared to the normalisation.

C. MODEL DEVELOPMENT

The statistical equation is developed using each of the regression techniques such as SVR, RF and DT. All the regression techniques were implemented and analysed in Python. The overall dataset was divided into 80% for training data and 20% for testing data. After the pre-processing of data, the computed number of inputs was 28. The inputs included 17 wind directions (E, ENE, ESE, N, NE, NNE, NNW, NW, S, SE, SSE, SSW, SW, VRB, W, WNW, and WSW), temperature (max), temperature (min), relative humidity (08:30 IST), relative humidity (17:30 IST), wind speed (08:30 IST), wind speed (17:30 IST), cloud coverage (08:30 IST), cloud coverage, (17:30 IST), rainfall intensities of Coonoor rain gauge station, Coonoor Railway Station and Runnymedu. The inputs were labelled from [x1, x2, .., x28] and the output forecasted is rainfall (in mm) [4]. Three forecasting models were implemented using SVR, RF and DT (Model-1,2 and 3)[6]. Further, a comparative study was made within SVR (Model-1S & 2S), RF (Model-1R & 2R) and DT (Model-1D & 2D) based on various criteria. Later, the actual amount of rainfall recorded and the predicted amount of rainfall were compared using scatter plots.

D. PERFORMANCE MEASURES

The performance measures used to compare the performance of the regression model are R-square and adjusted R-square. R-square is a statistical measure of how close the output/regression line is fitted to the mean line. Adjusted R-square is a modified version of R-square that takes into account the effect of adding a weekly influential predictor with the help of a penalising factor.

IV. RESULTS AND DISCUSSIONS

The regression models were implemented in python and the results obtained are discussed below. Table I shows the comparison of the performance measures of the various regression techniques.

Table I. Comparison of SVR, RF and DT techniques

Forecasting Model	R-Square	Adjusted R-square
SVR	0.814	0.806
DT	0.904	0.900
RF	0.981	0.980

The tabulated results show that the RF model gives a better prediction and performance in comparison to SVR and DT with an improvement of 7.85% and 17.21% respectively. The forecasted model demonstrates the RF models can be chosen for accurate prediction models compared to SVR and DT. The scatter plots for three models are shown in Figure 4, 5, 6:

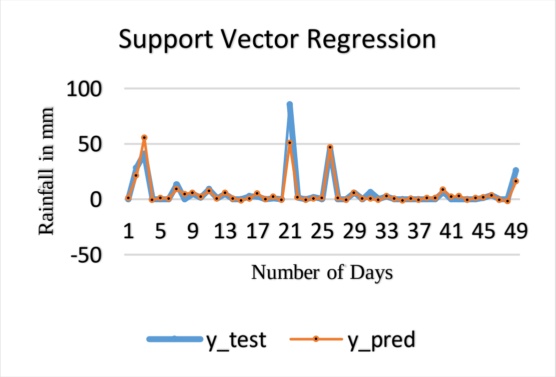

Figure 4. *SVR scatter plot between y_test (actual rainfall) and y_pred (predicted rainfall)*

A sample of 20 values of actual amount of rainfall and predicted amount of rainfall (in mm) using SVR with RBF kernel function is given in Table II:

Table II. Actual vs Predicted using SVR

	Actual Rainfall	Predicted Rainfall
Day 1	66.7	50.9822
Day 2	0	1.10645
Day 3	28.4	21.1975
Day 4	41.1	55.5796
Day 5	0	-0.48903
Day 6	0	1.07324
Day 7	0	0.5702
Day 8	13.4	8.8994
Day 9	0	4.7177
Day 10	4.6	5.9811
Day 11	1.6	2.5014
Day 12	9	7.1762
Day 13	1	0.6504
Day 14	4.8	5.8424
Day 15	0	0.6174
Day 16	0	-1.1598
Day 17	2.8	0.69156
Day 18	2.5	5.32732
Day 19	0	-0.32794
Day 20	1	2.07851

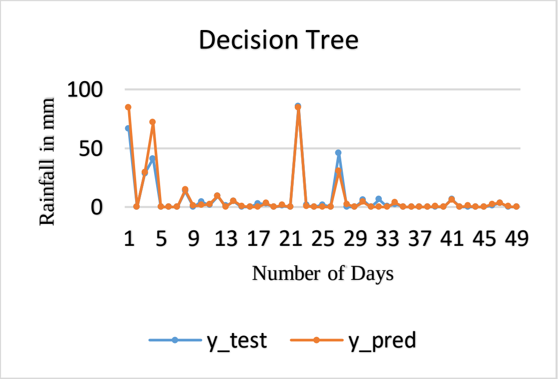

Figure 5. *DT scatter plot between actual and predicted rainfall*

A sample of 20 values of actual amount of rainfall and predicted amount of rainfall (in mm) using DT with MSE as the evaluation criterion is given in Table III:

Table III. Actual vs Predicted using DT

	Actual Rainfall	Predicted Rainfall
Day 1	66.7	84.3
Day 2	0	0
Day 3	28.4	29.3
Day 4	41.1	71.8
Day 5	0	0
Day 6	0	0
Day 7	0	0
Day 8	13.4	14.6
Day 9	0	0.8
Day 10	4.6	1.5
Day 11	1.6	2.3

Day 12	9	9.4
Day 13	1	0
Day 14	4.8	5
Day 15	0	0.7
Day 16	0	0
Day 17	2.8	0
Day 18	2.5	3.2
Day 19	0	0
Day 20	1	1.4

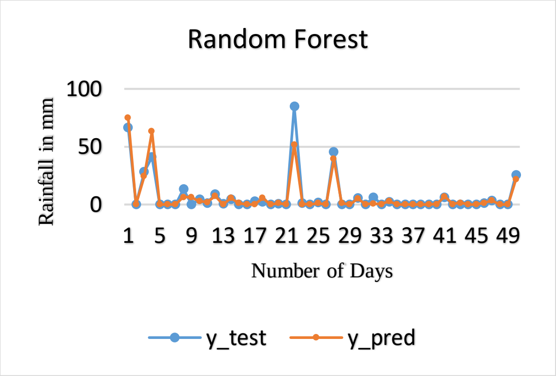

Figure 6. *RT scatter plot between actual and predicted rainfall*

A sample of 20 values of actual amount of rainfall and predicted amount of rainfall (in mm) using RF with the number of estimators as 300 is given in Table IV:

Table IV. Actual vs Predicted using RF

	Actual Rainfall	Predicted Rainfall
Day 1	66.7	75.136
Day 2	0	0.0026
Day 3	28.4	24.463
Day 4	41.1	63.354
Day 5	0	0.031
Day 6	0	0
Day 7	0	0.007
Day 8	13.4	6.514
Day 9	0	6.246
Day 10	4.6	2.94
Day 11	1.6	2.14467
Day 12	9	7.338
Day 13	1	0.1373
Day 14	4.8	5.725
Day 15	0	1.148
Day 16	0	0
Day 17	2.8	0.1253
Day 18	2.5	5.6403
Day 19	0	0.2346
Day 20	1	1.2413

Table V compares the SVR based on the kernel function [5]. Table VI depicts the change in R and adjusted R-square values in DT forecasted model due to change in error evaluation criteria [6]. Table VII compares the change in performance measures of RF regression based on the number of estimators.

Table V. Comparison of SVR based on kernel function

Forecasting Model	Kernel Function	R-square	Adjusted R-square
Model-1S	Polynomial function	0.649	0.634
Model-2S	Radial Basis function	0.814	0.806

The above results show that radial basis function should be used as the preferred kernel function for better performance.

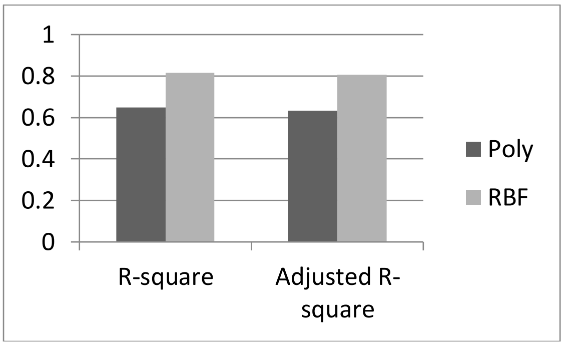

Figure 7. *SVR Comparative Analysis based on kernel function*

Table VI. Comparison of DT based on MSE and MAE

Forecasting Model	Criteria	R-square	Adjusted R-square
Model-1D	Mean abso error	0.899	0.895
Model-2D	Mean squa error	0.904	0.900

Using mean square error as performance measure criteria gives a slight improvement in the forecasted model [3].

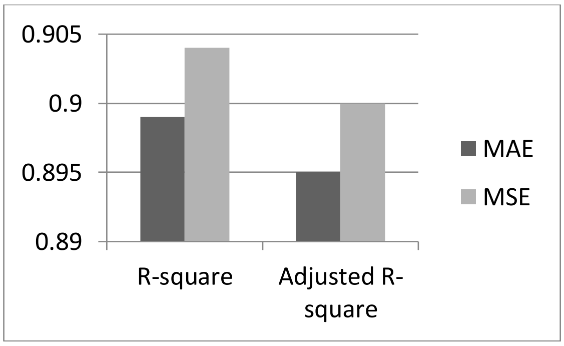

Figure 8. *DT comparative analysis based on evaluation criteria*

Table VII. Comparison of RT based on number of estimators

Forecasting Models	Number estimators	R-square	Adjusted square
Model-1R	10	0.965	0.964
Model-2R	100	0.979	0.980

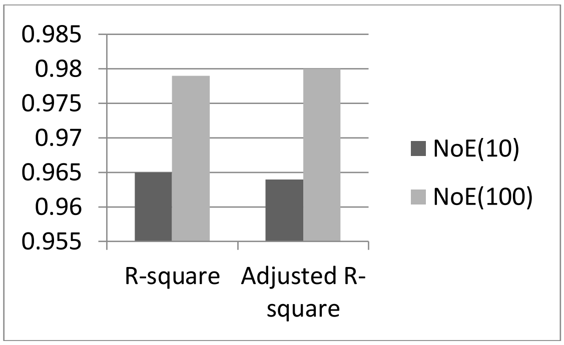

Figure 9. *RF comparative analysis based on number of estimators*

The overall relative and root mean square error of the regression techniques for actual vs predicted amount of rainfall can be tabulated as in Table VIII.

Table VIII. Overall Relative and Root Mean Square Error

	Relative Error	RMSE
Model 1D	0.0375	6.410
Model 2D	0.040	6.915
Model 1R	0.032	5.556
Model 2R	0.030	5.228
Model 3R	0.029	5.085
Model 1S	0.048	8.290
Model 2S	0.046	7.422

The statistical modelling for each of the regression techniques can be tabulated as below: (Table IX)

The statistical equation from the above table can be written as [7]:

$$Y = \sum_{n=1}^{28} coeff_n \times x_n$$

Y= forecasted rainfall in mm
n=1, 2, 3, …, 28

The statistical equation can be used if all the input parameters are given in hand to forecast the rainfall. It can be clearly seen that the standard deviation is less in RF regression method in comparison to DT and SVR.

Table VIII. Statistical Equation Coefficient Table

Statistical equation coefficient and standard error comparison table						
	SVR		DT		RF	
I/p	Coeff	Std err	Coeff	Std err	Coeff	Std err
x1	1.502	1.490	1.7880	1.514	0.2097	0.638
x2	4.7300	1.202	1.2732	1.221	1.5643	0.515
x3	2.0559	0.920	0.6484	0.934	0.6506	0.394
x4	-0.896	4.894	-0.293	4.973	0.0374	2.096
x5	-9.614	2.323	-3.502	2.360	-0.736	0.995
x6	-5.508	1.601	-0.551	1.627	0.5354	0.686
x7	2.3812	2.280	-3.330	2.316	2.2150	0.976
x8	0.6653	2.111	-2.086	2.145	-1.002	0.904
x9	1.2216	1.168	-0.891	1.187	-0.398	0.500
x10	0.6287	1.063	0.2669	1.080	0.6734	0.455
x11	0.4912	0.788	-0.126	0.800	0.1399	0.337
x12	0.5907	0.842	-0.110	0.855	0.0253	0.361
x13	0.1609	1.039	0.1809	1.056	0.1695	0.445
x14	-4.912	2.152	5.3375	2.187	5.1933	0.922
x15	-1.350	1.892	-1.094	1.923	-0.407	0.810
x16	0.8197	1.790	0.2139	1.819	0.3543	0.767
x17	-0.090	1.162	-0.619	1.181	0.0855	0.498
x18	-0.102	0.073	-0.130	0.075	-0.076	0.031
x19	0.0398	0.125	-0.004	0.127	0.0595	0.054
x20	0.0323	0.017	0.0527	0.017	0.0253	0.007
x21	-0.013	0.016	-0.011	0.016	-0.018	0.007
x22	-0.040	0.090	0.0539	0.091	0.0368	0.039
x23	0.0700	0.092	0.0452	0.093	0.0279	0.039
x24	0.1032	0.085	-0.077	0.087	-0.023	0.036
x25	0.0891	0.094	0.0509	0.095	0.0507	0.040
x26	0.2989	0.025	0.3953	0.026	0.5249	0.011
x27	0.1386	0.022	0.3454	0.023	0.2723	0.010
x28	0.2524	0.021	0.2199	0.022	0.1718	0.009

V. CONCLUSION

The study shows that the best regression technique among SVR, DT and RF is RF with 0.981 and 0.980 r-square and adjusted r-square values [1]. Further, it was seen that Radial Basis kernel function tends to give a better performance than Polynomial function with 0.814 and 0.806 r-square and adjusted r-square values. Similarly, in DT and RF models change in evaluation criteria and number of estimators provided a much better forecasted model. Based on the Relative error values Model 1D performs better amongst the DT models, Model 3R performs the best amongst the RF models and finally the Model 2S performs the best amongst the SVR models. Statistical Modelling fails to provide good accuracy of prediction due to the complexity involved in the input parameters. Rainfall forecasting has got great implication in the district of Coonoor for it being a landslide prone area. Knowing the rainfall in advance, would help us deal with landslides, floods and crops and water management effectively.

VI. FUTURE WORK

Considering the dataset is following a non-linear pattern it will be logically efficient to adopt a non-linear machine learning algorithm like Neural Networks to predict rainfall [9]. Sensitivity analysis can be done to decide upon the optimum hyper parameters to improve efficiency of the model [10].

ACKNOWLEDGMENT

The authors wish to express their sincere thanks to India Meteorological Department (IMD), Chennai and Public Works Department of Coonoor for providing the data to carry out the research.

REFERENCES

[1] Wai Yan Nyein Naing and ZawHtike. Forecasting of Monthly Temperature Variations Using Random Forests. ARPN Journal of Engineering and Applied Sciences, 2015.

[2] Ramsundram N, Sathya S and Karthikeyan S. Comparison of Decision Tree Based Rainfall Prediction Model with Data Driven Model Considering Climatic Variables. Irrigation and Drainage Systems Engineering, 2016.

[3] Nasimul Hasan, Nayan Chandra Nath and Risul Islam Rasel. A Support Vector Regression Model for Forecasting Rainfall. EICT, 2015.

[4] Siddharth S. Bhatkande and Roopa G. Hubballi. Weather Prediction Based on Decision Tree Algorithm Using Data Mining Techniques. International Journal of Advanced Research in Computer and Communication Engineering, 2016.

[5] Ashok Kumar, D.S. Pai, J. V. Singh, Ranjeet Singh and D. R. Sikka. Statistical Models for Long-range Forecasting of Southwest Monsoon Rainfall over India Using Stepwise Regression and Neural Network. Atmospheric and Climatic Sciences, 2012.

[6] WintThidaZaw and Thinn Thu Naing. Empirical Statistical Modelling of rainfall prediction over Myanmar. International Journal of Computer and Science engineering, 2008.

[7] Nikhil Sethi, Dr. Kanwal Garg. Exploiting Data Mining Technique for Rainfall Prediction. International Journal of Computer Science and Information Technologies, 2014.

[8] India Meteorological Department, Chennai, India and Public Works Department, Coonoor- Data Source Collection.

[9] Amruta A.Taksande, Dr. S. P. Khandait, Prof. Manish Katkar. A Data Mining Approach Using Artificial Neural Network to Predict Indian Monsoon Rainfall. International Journal of Research in Advent Technology, 2014.

[10] Bhaskar Pratap Singh, Pravendra Kumar, Tripti Srivastava, Vijay Kumar Singh. Estimation of Monsoon Season Rainfall and Sensitivity Analysis Using Artificial Neural Networks. Indian Journal of Ecology 44, 2017.