Forecast of Rainfall Quantity and its Variation using Environmental Features

Preetham Ganesh
Department of Computer Science and Engineering
Amrita School of Engineering,
Amrita Vishwa Vidyapeetham
Coimbatore, India
preetham.ganesh2015@gmail.com

Harsha Vardhini Vasu
Department of Computer Science and Engineering
Amrita School of Engineering,
Amrita Vishwa Vidyapeetham
Coimbatore, India
harshavardhini2019@gmail.com

Dayanand Vinod*
Department of Computer Science and Engineering
Amrita School of Engineering,
Amrita Vishwa Vidyapeetham
Coimbatore, India
v_dayanand@cb.amrita.edu

***Abstract*—Rainfall plays a crucial role in the lives of an ordinary man. Developing a prediction model that captures sudden fluctuations in rainfall has always been a challenging task. The paper aims at developing three models which predict monthly rainfall for all districts in Tamil Nadu, India and also drawing a district-wise comparison among them to find the best model for prediction. The models developed are District-Specific Model, Cluster-Based Model and Generic-Regression Model. The District-Specific Model trains on data from a particular district, the Cluster-Based Model groups districts based on the climatic conditions and trains on data from a particular cluster and the Generic-Regression Model trains on combined data from all the districts. The paper also aims at finding the monthly variation of rainfall across geographical regions.**

***Index Terms*—Rainfall Prediction, Tamil Nadu, Regression, Clustering.**

I. Introduction

Agriculture is the backbone of India's economy. According to the survey conducted by the World Travel and Tourism Council (WTTC) [1], agriculture contributed approximately 500 billion US Dollars to the Indian economy in the year 2016, which is roughly 24% of India's GDP (Gross Domestic Product) and engages 59% of India's human resources. Indian agriculture is sundry, ranging from poor farm villages to evolved farms using present-day agricultural technologies. Rainfall is the central source of water for the country's agricultural land. It is a boon if the rainfall quantity is in the right amount and a bane if the rainfall is too low or too high where the crops get destroyed. The knowledge about the rainfall quantity and its variation can help the farmers to plan their crops, thus saving time, effort and resources. Predicting rainfall can also help the general public and the government, as they can take precautionary measures in the case of heavy rains which may lead to floods. These preventive measures can not only save human lives but can also minimise the recovery and reconstruction costs for the state.

Predicting rainfall using Machine Learning can be done using various methods and the most commonly used method is Regression. Regression analysis is widespread in various domains. Jeyakumar et al. in [2] used Support Vector Regression for identification of symbols in huge Multiple Input and Multiple Output (MIMO) systems and Jackson Isaac et al. in [3] used Logistic Regression in DBMS to forecast the class of the given query.

Generally, for rainfall prediction, a district's or location's climatic data is trained using a few regression algorithms and evaluated using a few error measures. In this paper, three models are used to predict rainfall in a particular district namely District-Specific Model, Generic-Regression Model and Cluster-Based Model. Section IV-B, IV-C and IV-D discuss in detail about the models. The regression algorithms used for developing the models mentioned above are Multiple Linear Regression (MLR), Support Vector Regression (SVR), Polynomial Regression (PR) and Decision Tree Regression (DTR).

The anatomy of the proposed work is as follows: Section II lists the previous works related to the rainfall prediction; Section III explains the machine learning algorithms and the error measures used in the paper; Section IV explains the process flow for the proposed solution; Section V discusses in detail about the derived results; Section VI concludes the paper based on the derived results.

II. Literature Survey

This section reviews in detail about the previous researches conducted in the same territory. The papers are grouped and discussed based on the methods used in them.

Niu et al. in [4] proposed the use of classification algorithms such as Naive Bayes (NB), Support Vector Machine (SVM) and Back Propagation Neural Network (BPNN) on the open dataset from the China Meteorological Administration. The various features used for forecasting include latitude, longitude, altitude and average temperature. The performance measure used for comparing the models is Accuracy. Based on accuracy, it was concluded that BPNN outperforms SVM and NB.

Tharun et al. in [5] predicted rainfall in the Nilgiris District, Tamil Nadu, India using various regression methods such as SVR, Random Forest Regression (RFR) and DTR. The performance measures used to evaluate the regression models are R^2 and Adjusted R^2. Adjusted R^2 is the customised version

978-1-5386-8189-3/19/$31.00

of R^2 that takes into account the effect of adding an influential weekly predictor. Based on the performance measures RFR outperforms SVR and DTR. Kusiak et al. in [6] used a data mining approach to predict rainfall in Oxford and Iowa. The machine learning models used for prediction are Multiple Layer Perceptron (MLP), RFR, Classification and Regression Tree (C&RT), SVR and K-Nearest Neighbors Regression (KNN). The error measures used are Mean Absolute Error (MAE), Mean Squared Error (MSE) and Standard Deviation (SD). Smaller values of MAE, MSE and SD indicate that a particular model has an excellent fit to the data. According to the analysis, MLP outperforms the other models.

Lu et al. in [7] investigated the performance of various regression methods to predict average monthly rainfall in Guangxi, China using data from January 1965 to December 2009. The methods used are Simple Averaging Ensemble, Mean Squared Error Ensemble, Variance Weighed Ensemble and SVR. The error measures used are Normalised Mean Squared Error (NMSE), Mean Absolute Percentage Error (MAPE) and Pearson Relative Coefficient (PRC). The outcome of the analysis is that SVR performs best. Mohapatra et al. in [8] used MLR to model the rainfall data of Bangalore obtained from the India Water Portal for the years 1901 to 2002 and compare the performance of the validation techniques such as Holdout method and K-Fold Cross-Validation method. The prediction was season-wise (Rainy, Summer and Winter) and the features used are Precipitation and Wet Day Frequency. In all the seasons K-Fold Cross Validation method outperforms the Holdout Method.

Chatterjee et al. in [9] used a combination of clustering and Hybrid Neural Network (HNN) to predict rainfall in the Southern part of West Bengal, India. It is a two-step process where the first step is using the Greedy Forward Selection algorithm to reduce the feature set and find the best possible feature set and then K-Means clustering is applied. The second step is to train each cluster with the Neural Network discreetly. The performance measures such as Accuracy, Precision and Recall are used to compare the HNN and MLP Feed Forward Network (MLP-FFN). The HNN outperformed MLP-FFN in both feature selected and non-feature selected methods.

R. Venkata Ramana et al. in [10] predicted rainfall in Darjeeling Rain Gauge Station, West Bengal, India using a combination of Wavelet Neural Network (WNN) and Artificial Neural Network (ANN). The dataset consisted of average monthly rainfall for 74 years. The performance measures used are Root Mean Squared Error (RMSE), Correlation Coefficient (R) and Coefficient of Efficiency (COE). Using 44 years of data as the training set and the rest of the years as the test set, based on performance measures WNN performed better than ANN. Mislan et al. in [11] proposed two different architectures of Neural Networks, which are 2-50-10-1 and 2-50-20-1. The first digit is the number of neurons in the input layer, the second and the third digits are the number of neurons in the hidden layer, and the last number indicates the number of neurons in the output layer. Architecture 2-50-20-1 outperformed the other. Manek et al. in [12] compared BPNN, Generalised Regression Neural Network (GRNN) and Radial Basis Function Neural Network (RBFNN) to predict the rainfall in Thanjavur district of Tamil Nadu, India using the data obtained from the India Water Portal - Met Data Repository. The features used for prediction are Precipitation, Cloud Cover, Vapor Pressure and Average Temperature. RBFNN outperformed GRNN and BPNN. Dash et al. in [13] used Single Layer Feed Forward Neural Network (SLFN) and Extreme Machine Learning (ELM) to predict the rainfall season-wise in the years 1871 to 2014, where the networks were trained with the years 1871 to 2004 and for testing set 2005 to 2014. The performance measures used for evaluating the models are MAE and RMSE. On analysis, SLFN outperformed ELM.

Most of the papers mentioned above use a particular location's data to predict rainfall, but in this paper, the collective knowledge of all the 29 districts data in Tamil Nadu, India is used to predicting rainfall in a particular district. Also, to optimise the result, different parameters for each regression algorithm across all the models are tested. The primary focus is on finding the best model among the District-Specific Model, Generic-Regression Model and the Cluster-Based Model along with the best regression algorithm and the corresponding parameter for each district. Furthermore, Section V-E discusses the variation of rainfall across the geographic regions in a detailed manner.

III. Methodology

This section describes the dataset used for investigation and defines all the regression algorithms, clustering algorithms and performance measures used in this paper.

A. Dataset Description

The India Water Portal - Met Data Repository is used to collect the data. The data collected for a particular district comprises of the dependent attribute 'Rainfall' and eight independent attributes namely 'Average Temperature', 'Cloud Cover', 'Maximum Temperature', 'Minimum Temperature', 'Crop Evapotranspiration', 'Potential Evapotranspiration', 'Vapor Pressure' and 'Wet Day Frequency'. The dataset of each feature contains 102 records and 12 columns where each row contains data of a particular year, and each column contains data of a particular month across the years.

B. Regression Algorithms

1) Multiple Linear Regression (MLR): It is a straight line approach to model the correlation between the dependent variable and multiple independent variables using single-dimensional predictor functions. The model parameters depend on the dataset and is not standard for all the datasets.

2) Support Vector Regression (SVR): It is a supervised learning method which builds hyper-plane(s) in a dimensional space used for regression and classification examination or detecting outliers. The various kernels used to transform the data for prediction are Linear, Non-Linear, Polynomial, Sigmoid and Radial Basis Function (RBF).

3) Polynomial Regression (PR): It is similar to MLR where the relationship between the dependent and the independent variable modelled as n^{th} order polynomial on the independent variables.

4) Decision Tree Regression (DTR): Decision tree uses supervised learning to build regression or classification models in the form of a tree structure. The tree has three different nodes, namely the root node, decision nodes and leaf nodes. The root node is the primary node, decision node has branches (two or more), and the leaf node is a node at the end of the tree.

C. Clustering Model

1) K-Means Clustering: It focuses on dividing N points into K clusters with the closest mean, serving as the centre of the cluster. The Euclidean Distance is used to allocate a data point to a particular cluster centre.

2) Elbow Method: This method focuses on finding the optimal number of clusters. Sum Square Error (SSE) is the sum of the mean Euclidean Distance of all the points against the centroid. SSE is computed for every increment in the number of clusters (K). When the SSE starts dropping by decidedly smaller angles, then that K value is the optimal number of clusters.

D. Evaluation Measures

1) Mean Squared Error (MSE): It measures the quality of a model where the value is the mean squared difference between the actual and predicted value as given in (1).

$$MSE = \frac{\sum_{i=1}^{N}(Y_t - Y_p)^2}{N} \tag{1}$$

Where Y_t is the actual value, Y_p is the predicted value and N is the number of observations in the dataset.

2) Root Mean Squared Error (RMSE): It is the square root of the mean squared error as given in (2).

$$RMSE = \sqrt[2]{MSE} \tag{2}$$

3) Mean Absolute Error (MAE): It measures the mean absolute difference between the actual and the predicted value in a set of predictions as given in (3).

$$MAE = \frac{\sum_{i=1}^{N}|Y_t - Y_p|}{N} \tag{3}$$

4) Median Absolute Error (MDAE): It measures the variance of a uni-variate sample of quantitative data. It is defined as the median of the absolute residuals between the original and the predicted value as given in (4).

$$MDAE = median(|Y_t - Y_p|) \tag{4}$$

5) Explained Variance Score (EVS): It measures the ratio to which a regression model is capturing the dispersion in the dataset. It is the mean squared difference between the predicted value and the mean of the actual values in the dataset as given in (5).

$$EVS = \frac{\sum_{i=1}^{N}(Y_p - Y_m)^2}{N} \tag{5}$$

Where Y_m is the mean of Y_t in the dataset.

6) R^2 Score (R^2): It is commonly known as the coefficient of determination which is the ratio of dispersion in the predictive variable from independent variables as given in (6).

$$R^2 = \frac{EVS}{TV} \tag{6}$$

Where EVS is the Explained Variance Score and TV is Total Variation. If the value of R^2 is 0%, then none of the variability of the response data is around the mean, and if it is 100%, then all the variability of the response data is around the mean.

IV. Process Flow

The process flow for the proposed architecture is given in Fig. 1.

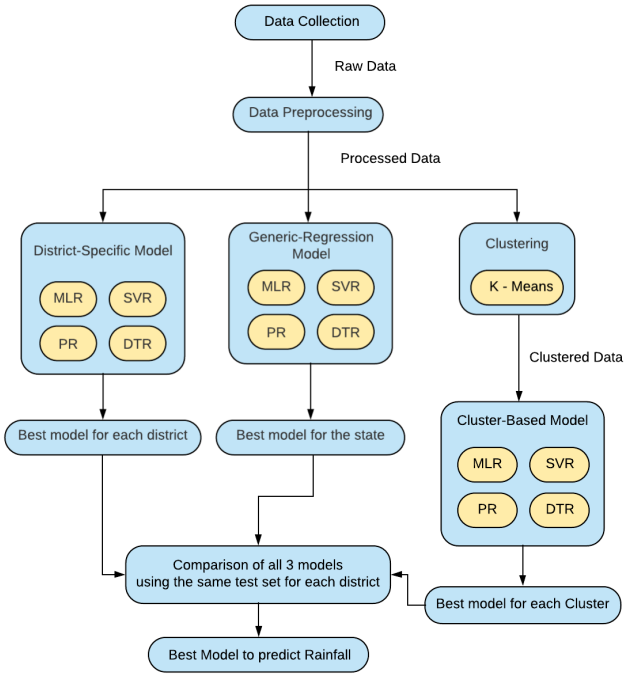

Fig. 1. Process Flow

A. Data Pre-processing

1) Data Transformation: The dataset obtained from the source had separate files for each feature in all the districts. Combining the datasets of features into a single dataset makes computation far easier where each column contains data of a particular feature and is arranged sequentially from 1901 January to 2002 December.

2) Data Normalisation: All the attributes used are numerical and have different ranges. For the regression algorithm to work with high efficiency and accuracy, the attributes have to be normalised. Using Min-Max Normalisation for this purpose brings the range of all the features from 0 to 1, thereby reducing the chance of having different weights for the features. The formula for the Min-Max Normalisation is given in (7).

$$X'_{\mathrm{i}} = \frac{X_{\mathrm{i}} - X_{\mathrm{min}}}{X_{\mathrm{max}} - X_{\mathrm{min}}} \tag{7}$$

Where X_i is the i^{th} element in the feature, X_{min} is the minimum value of the feature, X_{max} is the maximum value of the feature and X'_i is the normalised value of the i^{th} element in the feature.

B. District-Specific Model

For each district, the rainfall has been predicted using four regression algorithms with different parameter values. To predict rainfall for a particular district, only the data collected from that district is used to train the model. Repeated K-Fold Cross Validation method has been used to validate the model with ten splits and ten repetitions, and the evaluation measures have been used to find the best model for each district.

C. Generic-Regression Model

The data of all the districts have been combined into one single dataset (Generic Dataset) to build the Generic-Regression Model. The generic dataset has 35496 tuples (29 districts * 1224 tuples per district) to which the same process as in the District-Specific Model has been used for prediction.

D. Cluster-Based Model

K-Means clustering has been used to find the districts with similar climatic conditions. The datasets are combined based on the clusters formed, and rainfall is predicted for a district by training the model with data of the cluster to which the district belongs. The rest remains the same as the District-Specific Model.

V. Results and Discussion

This section discusses in detail the results obtained on using the machine learning regression algorithms to model the data for all the districts in Tamil Nadu, India.

A. Correlation between the attributes

The Pearson Correlation Coefficient finds the linear relationship between any two continuous variables. It helps in finding the right set of features for predicting the target variable. The formula for calculating the correlation between any two attributes is given in (8).

$$\rho_{\mathrm{x,\,y}} = \frac{\sum(X_{\mathrm{i}} - X_{\mathrm{m}})(Y_{\mathrm{i}} - Y_{\mathrm{m}})}{\sqrt{\sum(X_{\mathrm{i}} - X_{\mathrm{m}})^2 \sum(Y_{\mathrm{i}} - Y_{\mathrm{m}})^2}} \tag{8}$$

Where X and Y are the continuous variables. X_i and Y_i represents the i^{th} element in the vectors and X_m and Y_m are the mean values of the corresponding vectors. In this process, the actual values are re-scaled, and the Standard Deviation is computed. If ρ is closer to 1 then the variables are positively correlated, if ρ is closer to -1, then the variables are negatively correlated, and if the variables are independent of each other, then ρ is closer to 0. The correlation heat-map of the attributes is given in Fig. 2.

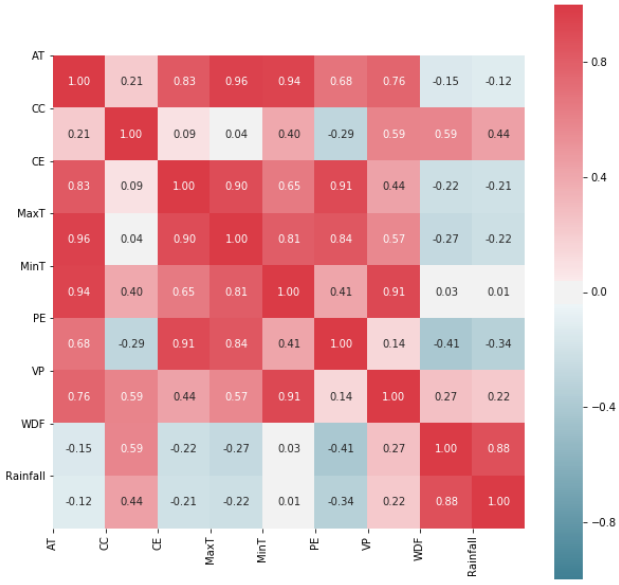

Fig. 2. Correlation Heat Map

Cloud Cover, Vapour Pressure and Wet Day Frequency are positively correlated with rainfall, and Average Temperature, Crop Evapotranspiration, Maximum Temperature and Potential Evapotranspiration are negatively correlated with rainfall as shown in Fig. 2. Also, Minimum Temperature has only a slight impact on the amount of rainfall, so it is excluded from the prediction process.

B. Parameter Selection for the Regression Algorithms

For the regression algorithms to predict more accurately, their corresponding parameter values have to be tuned. For SVR, different kernels like Linear, Polynomial (Degree = 3), Sigmoid and RBF are tested. Similarly for DTR, maximum depths ranging from two to seven and for PR, degrees ranging from two to five are tested.

C. Performance Analysis of the Models

1) District-Specific Model: The regression algorithms and the parameter required to build the best model for a district is chosen based on the models' performance measures. A good model should have low MSE, RMSE, MAE and MDAE and high EVS and R^2 values. The performance of all the regression algorithms with different parameters for the Chennai District is given in Table I.

On observing the values in Table I, for PR, degree two is an excellent choice, as the degree rises the MSE, RMSE, MAE and MDAE values tend to increase, and the EVS and

R^2 values tend to decrease. For degree four and five the EVS and R^2, scores are negative which indicates that the models are unstable and do not capture the variation well. Also, DTR with a maximum depth of five outperforms the others, and SVR with RBF kernel outperforms SVR with other kernels.

Extending the analysis done in Table I to the other districts, it was found that MLR performs better for the districts Dharmapuri, Dindigul, Madurai, Ramanathapuram, Theni, Tirunelveli and Virudhunagar. Likewise, SVR with RBF kernel for Kancheepuram, Tiruvannamalai and Vellore and PR with degree two performs better for the other districts.

2) Generic-Regression Model: The generic data is used for training the model, where the performance measures of all the regression algorithms along with their parameters are given in Table II.

From Table II, it can be inferred that PR with degree four fits the data better and outperforms the other degrees. DTR with a maximum depths of six outperforms the other depth values, and SVR with RBF kernel outperforms the other kernels.

3) Cluster-Based Model: For each district, the median of all the features across 102 years has been considered as input for clustering using K-Means, and Elbow Method has been used to find the optimal number of clusters. The graph of the number of cluster centres versus the sum of squared distances is shown in Fig. 3.

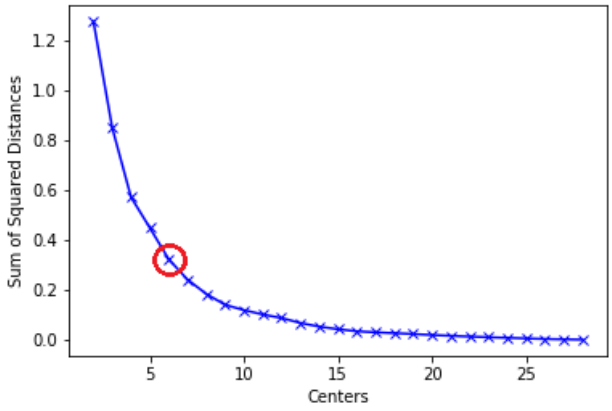

Fig. 3. Elbow Method

It can be clearly observed in Fig. 3 that six is the optimal number of clusters as the sum of squared distances drops by minimal angles from that point. The clusters formed as a result of K-Means clustering with K as six is shown in Fig. 4 and that is cross-verified with works done by Palanisami et al. in Diversification of Agriculture in Coastal Districts of Tamil Nadu- a Spatio- Temporal Analysis [14].

Based on the results obtained after performing clustering, the districts were grouped, and all the chosen regression algorithms with different parameter values have been applied to each of the grouped data. The performance measures of all the regression algorithms along with different parameter values for Cluster 1 are shown in Table III.

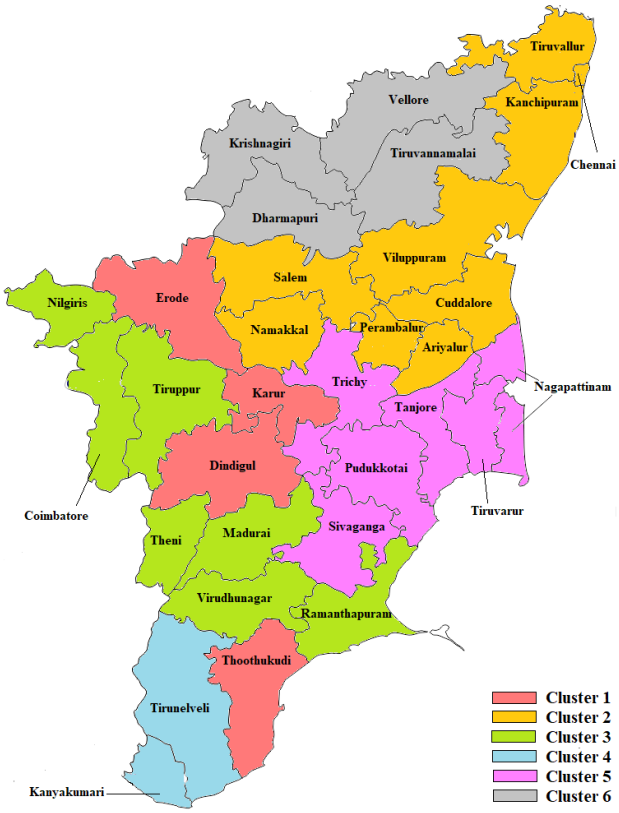

Fig. 4. Graphical Representation of formed Clusters

As shown in Table III, PR with degree three, DTR with a maximum depth of five and SVR with linear kernel outperforms the other regression models for the cluster 1. The same has been extended to the other clusters, and it was found that PR is the best regression algorithm where the best degree for cluster 4 is two, cluster 2 is four and for the other clusters is three.

D. Comparison on performance of District Specific Model, Cluster-Based Model and Generic Regression Model

A comparison was drawn between the performance of the District-Specific Model, Cluster-Based Model and the Generic-Regression Model by testing them on the same test data. At a time only one district is considered for comparison. Repeated K-Fold Cross-Validation with ten folds and ten repeats, has been applied a district's data, where the test set obtained in each iteration has been removed from the respective clustered data and the generic data using a customised index. The same set of record has been removed from the generic dataset and the clustered dataset that contains that district, for testing. Then the remaining records have been used for training the respective models. The comparison between the performance of the three models is shown in Table IV and Table V.

Based on all six performance measures used, Cluster-Based Model performs better than the District-Specific Model across all the districts as shown in Table IV and Table V. The

TABLE I
COMPARISON ON PERFORMANCE OF THE REGRESSION ALGORITHMS FOR THE CHENNAI DISTRICT

Method	Parameter	MSE	RMSE	MAE	MDAE	EVS	R^2
Multiple Linear Regression	-	0.004	0.0635	0.0442	0.0324	0.8257	0.8241
Polynomial Regression	**Degree = 2**	**0.0039**	**0.0615**	0.0423	0.0282	**0.8305**	**0.8289**
	Degree = 3	0.004	0.0629	**0.0402**	**0.0239**	0.8273	0.8253
	Degree = 4	0.0558	0.1933	0.0846	0.0408	-1.5956	-1.6172
	Degree = 5	4149.7	50.1	15.3	2.7	-187415	-188520
Decision Tree Regression	Max Depth = 2	0.0057	0.0747	0.0484	0.0252	0.7569	0.7552
	Max Depth = 3	0.0043	0.0646	0.0393	0.0198	0.817	0.8156
	Max Depth = 4	0.004	0.0628	0.0371	0.0183	0.8278	0.8264
	Max Depth = 5	**0.0039**	**0.0616**	**0.036**	**0.0181**	**0.8342**	**0.833**
	Max Depth = 6	0.0042	0.0639	0.0368	0.0184	0.8211	0.8199
	Max Depth = 7	0.0044	0.0653	0.0374	0.0184	0.8132	0.812
Support Vector Regression	Kernel = Linear	0.0046	0.0674	0.05	0.0406	0.8053	0.8002
	Kernel = Poly	0.0103	0.1004	0.0727	0.0592	0.5758	0.5637
	Kernel = RBF	**0.0038**	**0.0609**	**0.0424**	**0.031**	**0.8395**	**0.8372**
	Kernel = Sigmoid	0.2638	0.5119	0.3532	0.2414	-10.41	-10.93

TABLE II
COMPARISON ON PERFORMANCE OF THE REGRESSION ALGORITHMS FOR THE GENERIC-REGRESSION MODEL

Method	Parameter	MSE	RMSE	MAE	MDAE	EVS	R^2
Multiple Linear Regression	-	0.0006	0.0254	0.0156	0.0101	0.7845	0.7844
Polynomial Regression	Degree = 2	0.00057	0.0239	0.0145	0.0089	0.8081	0.8081
	Degree = 3	0.00054	0.0231	0.0137	0.0079	0.8207	0.8206
	Degree = 4	**0.00052**	**0.0227**	**0.0134**	**0.0076**	**0.8268**	**0.8267**
	Degree = 5	0.00053	0.0229	0.0135	0.0078	0.8236	0.8235
Decision Tree Regression	Max Depth = 2	0.00098	0.0313	0.0192	0.0111	0.6731	0.673
	Max Depth = 3	0.00074	0.0272	0.016	0.0091	0.7518	0.7518
	Max Depth = 4	0.00067	0.0259	0.0149	0.0083	0.7759	0.7758
	Max Depth = 5	0.00064	0.0253	0.0145	0.0081	0.7862	0.7862
	Max Depth = 6	**0.00063**	**0.0252**	**0.0142**	**0.0079**	**0.7878**	**0.7877**
	Max Depth = 7	0.00065	0.0254	**0.0142**	**0.0079**	0.7833	0.7833
Support Vector Regression	**Kernel = Linear**	**0.0015**	**0.0388**	**0.0311**	**0.0279**	0.6963	**0.494**
	Kernel = Poly	0.0041	0.0641	0.0574	0.0577	0.5824	-0.3829
	Kernel = RBF	0.0027	0.0523	0.0463	0.0466	0.6845	0.0814
	Kernel = Sigmoid	0.0016	0.0394	0.033	0.0318	**0.7475**	0.4795

TABLE III
COMPARISON ON PERFORMANCE OF THE REGRESSION ALGORITHMS FOR THE CLUSTER 1

Method	Parameter	MSE	RMSE	MAE	MDAE	EVS	R^2
Multiple Linear Regression	-	0.0044	0.0663	0.0462	0.0309	0.7412	0.7406
Polynomial Regression	Degree = 2	0.0043	0.0654	0.0453	0.0307	0.748	0.7475
	Degree = 3	**0.0041**	**0.0635**	**0.0437**	**0.029**	**0.7624**	**0.7619**
	Degree = 4	0.0044	0.0659	0.0454	0.0304	0.7444	0.7438
	Degree = 5	0.0106	0.1005	0.0604	0.0.0384	0.379	0.3777
Decision Tree Regression	Max Depth = 2	0.0056	0.0745	0.0534	0.0363	0.6729	0.6722
	Max Depth = 3	0.0048	0.069	0.0478	0.0325	0.7193	0.7187
	Max Depth = 4	0.0047	0.0682	0.0463	0.0309	0.7257	0.7251
	Max Depth = 5	**0.0046**	**0.0678**	**0.0454**	0.0299	**0.7287**	**0.7282**
	Max Depth = 6	0.0048	0.0688	0.0456	**0.0295**	0.7204	0.7198
	Max Depth = 7	0.005	0.0705	0.0463	0.0296	0.7062	0.7056
Support Vector Regression	**Kernel = Linear**	**0.0048**	**0.0694**	**0.0526**	**0.0419**	**0.7343**	**0.7162**
	Kernel = Poly	0.0068	0.0824	0.0675	0.0641	0.6359	0.6001
	Kernel = RBF	0.005	0.071	0.0557	0.0469	0.7292	0.7031
	Kernel = Sigmoid	0.7071	0.8392	0.5115	0.3084	-38.35	-40.84

Generic-Regression Model has the least MSE, RMSE, MAE and MDAE values for all the districts and the Cluster-Based Model has the highest EVS and R^2 scores for a maximum number of districts, which implies that Cluster-Based Model captures the variation well compared to the other models. However, the difference in value between the Cluster-Based Model and the Generic-Regression Model is negligible.

E. Variation in Rainfall Distribution across the Geographical Regions and Time

To visualise the variation of rainfall across months, the median of the rainfall values recorded for a particular month across years for all the districts in a cluster has been calculated. The continuous lines plots in Fig. 5 and 6 is the line connecting the median rainfall across months for each cluster and dotted

TABLE IV
COMPARISON BETWEEN THE MODELS USING THE PERFORMANCE MEASURES MSE, RMSE AND MAE

Cluster	District Name	MSE			RMSE			MAE		
		District	Cluster	Generic	District	Cluster	Generic	District	Cluster	Generic
Cluster 1	Dindigul	0.0064	0.0055	**0.0006**	0.0796	0.0734	**0.0245**	0.0559	0.0505	**0.0166**
	Erode	0.0042	0.0026	**0.0003**	0.064	0.0503	**0.0165**	0.0435	0.0338	**0.0109**
	Karur	0.0114	0.0031	**0.0003**	0.1064	0.0555	**0.0184**	0.0778	0.0398	**0.0131**
	Thoothukkudi	0.0074	0.0029	**0.0003**	0.0857	0.0533	**0.0177**	0.0593	0.0369	**0.0121**
Cluster 2	Ariyalur	0.003	0.0008	**0.0001**	0.0539	0.0275	**0.0107**	0.0356	0.0179	**0.0069**
	Chennai	0.0031	0.0022	**0.0003**	0.055	0.0466	**0.0183**	0.0353	0.0287	**0.0112**
	Cuddalore	0.002	0.0005	**0.0001**	0.0441	0.0231	**0.0093**	0.0286	0.0146	**0.0059**
	Kancheepuram	0.0041	0.0017	**0.0003**	0.0634	0.0413	**0.0165**	0.0456	0.0261	**0.0103**
	Namakkal	0.008	0.0012	**0.0002**	0.0891	0.0346	**0.0132**	0.0646	0.0247	**0.0094**
	Perambalur	0.0045	0.0013	**0.0002**	0.0669	0.0352	**0.0134**	0.0453	0.0239	**0.0091**
	Salem	0.0071	0.0011	**0.0002**	0.084	0.0333	**0.0127**	0.0593	0.0229	**0.0087**
	Thiruvallur	0.0046	0.0016	**0.0002**	0.0674	0.0396	**0.0155**	0.0486	0.0253	**0.0098**
	Viluppuram	0.002	0.0005	**0.0001**	0.0441	0.0223	**0.0086**	0.0298	0.0148	**0.0057**
Cluster 3	Coimbatore	0.0035	**0.0009**	**0.0009**	0.059	0.0301	**0.0295**	0.0385	0.0193	**0.0187**
	Madurai	0.007	**0.0008**	**0.0008**	0.0827	**0.028**	**0.028**	0.0546	0.0183	**0.018**
	Ramanathapuram	0.0061	**0.0003**	**0.0003**	0.0776	**0.018**	**0.018**	0.0542	0.0122	**0.012**
	Theni	0.0051	0.0018	**0.0017**	0.0703	0.0412	**0.0411**	0.0433	0.0248	**0.0245**
	The Nilgiris	0.0025	0.002	**0.0019**	0.0486	0.0436	**0.0428**	0.0253	0.0224	**0.0214**
	Virudhunagar	0.0088	**0.0007**	**0.0007**	0.0927	**0.0266**	0.0269	0.0621	0.0176	**0.0175**
Cluster 4	Tirunelveli	0.0082	0.0075	**0.0005**	0.09	0.0856	**0.0228**	0.0611	0.0577	**0.0154**
Cluster 5	Nagapattinam	0.0032	0.0024	**0.0002**	0.0563	0.0489	**0.0151**	0.0378	0.0324	**0.0101**
	Pudukkottai	0.004	0.0019	**0.0002**	0.0625	0.0434	**0.0137**	0.0425	0.0292	**0.0092**
	Sivaganga	0.0048	0.0024	**0.0002**	0.0686	0.0485	**0.0152**	0.0483	0.0337	**0.0105**
	Thanjavur	0.0044	0.0024	**0.0002**	0.0656	0.0483	**0.0149**	0.044	0.032	**0.0098**
	Thiruvarur	0.0054	0.0041	**0.0004**	0.0733	0.0636	**0.0198**	0.0514	0.0426	**0.0132**
	Tiruchirapalli	0.0063	0.0021	**0.0002**	0.0787	0.0458	**0.0143**	0.0576	0.0326	**0.0101**
Cluster 6	Dharmapuri	0.0056	0.0032	**0.0002**	0.0744	0.0566	**0.0125**	0.053	0.0382	**0.0084**
	Tiruvannamalai	0.0054	0.004	**0.0002**	0.0735	0.0628	**0.014**	0.0523	0.0416	**0.0091**
	Vellore	0.0074	0.0049	**0.0002**	0.0854	0.0696	**0.0153**	0.0612	0.0467	**0.0102**

TABLE V
COMPARISON BETWEEN THE MODELS USING THE PERFORMANCE MEASURES MDAE, EVS AND R^2

Cluster	District Name	MDAE			EVS			R^2		
		District	Cluster	Generic	District	Cluster	Generic	District	Cluster	Generic
Cluster 1	Dindigul	0.0379	0.0336	**0.0112**	0.7174	**0.7589**	0.7527	0.7146	**0.7571**	0.7508
	Erode	0.0279	0.0212	**0.0069**	0.8407	**0.8654**	**0.8654**	0.8395	0.8643	**0.8646**
	Karur	0.057	0.0268	**0.009**	0.6979	**0.7445**	0.7383	0.6945	**0.7424**	0.7355
	Thoothukkudi	0.0421	0.0251	**0.0081**	0.7018	**0.75**	0.7446	0.6987	**0.7478**	0.7423
Cluster 2	Ariyalur	0.0237	0.0109	**0.0042**	0.8882	**0.9187**	0.9069	0.8873	**0.9181**	0.9062
	Chennai	0.0206	0.0155	**0.0058**	0.8663	**0.9033**	0.8894	0.865	**0.9024**	0.8884
	Cuddalore	0.0186	0.0087	**0.0035**	0.9374	**0.9573**	0.948	0.9369	**0.9569**	0.9475
	Kancheepuram	0.0351	0.0146	**0.0055**	0.8517	**0.916**	0.9016	0.8495	**0.9153**	0.9006
	Namakkal	0.0448	0.0166	**0.0062**	0.7828	**0.8286**	0.8135	0.7803	**0.8273**	0.8117
	Perambalur	0.0283	0.0151	**0.0058**	0.8059	**0.8409**	0.8273	0.8045	**0.8398**	0.8258
	Salem	0.0394	0.0149	**0.0055**	0.7982	**0.8406**	0.8289	0.7963	**0.8391**	0.8274
	Thiruvallur	0.0357	0.0145	**0.0056**	0.8243	**0.9043**	0.8909	0.8227	**0.9035**	0.8901
	Viluppuram	0.0193	0.009	**0.0034**	0.9386	**0.9568**	0.9521	0.9381	**0.9565**	0.9517
Cluster 3	Coimbatore	0.0232	0.0112	**0.0107**	0.8636	0.8847	**0.8888**	0.8625	0.8839	**0.888**
	Madurai	0.0348	0.0115	**0.011**	0.6421	**0.6847**	0.6822	0.639	**0.6814**	0.6797
	Ramanathapuram	0.038	0.0081	**0.0076**	0.7631	**0.8117**	0.8103	0.761	**0.81**	0.8092
	Theni	0.026	0.0145	**0.0138**	0.7353	0.7668	**0.7679**	0.7329	0.7648	**0.7664**
	The Nilgiris	0.0116	0.0099	**0.0085**	0.8494	0.8785	**0.8817**	0.8483	0.8774	**0.8809**
	Virudhunagar	0.0417	0.0115	**0.0112**	0.6459	0.6991	**0.6934**	0.6423	**0.6968**	0.6907
Cluster 4	Tirunelveli	0.0407	0.0388	**0.01**	0.6777	0.7094	**0.7374**	0.6736	0.7069	**0.7348**
Cluster 5	Nagapattinam	0.0247	0.0204	**0.0065**	0.8671	0.8862	**0.8875**	0.8658	0.8853	**0.8866**
	Pudukkottai	0.0275	0.018	**0.0057**	0.8381	**0.8579**	0.8532	0.8368	**0.8567**	0.8523
	Sivaganga	0.0333	0.022	**0.0069**	0.8038	**0.8321**	0.8254	0.8021	**0.8306**	0.8239
	Thanjavur	0.028	0.0198	**0.006**	0.8167	0.8446	**0.8453**	0.8152	0.8434	**0.8442**
	Thiruvarur	0.0359	0.0266	**0.0081**	0.7617	**0.8242**	0.8196	0.7598	**0.8224**	0.8181
	Tiruchirapalli	0.0416	0.0227	**0.0069**	0.7559	**0.8034**	0.8022	0.7528	**0.8019**	0.8
Cluster 6	Dharmapuri	0.0363	0.023	**0.005**	0.7909	**0.8504**	0.8448	0.7894	**0.8493**	0.8439
	Tiruvannamalai	0.0379	0.0251	**0.0053**	0.8366	**0.879**	0.873	0.8352	**0.878**	0.872
	Vellore	0.0448	0.0272	**0.0059**	0.7833	**0.8396**	0.8343	0.7817	**0.8385**	0.8332

lines are the average rainfall for each cluster.

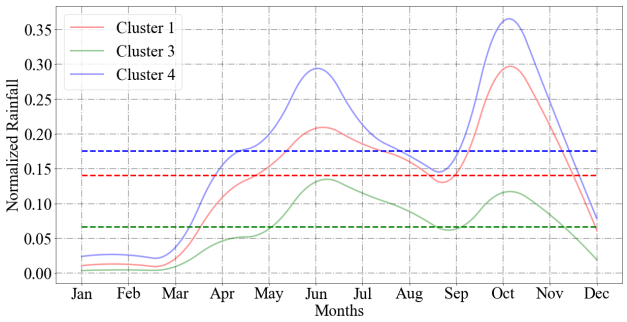

Fig. 5. Variation of Rainfall across months for Clusters 1, 3 and 4

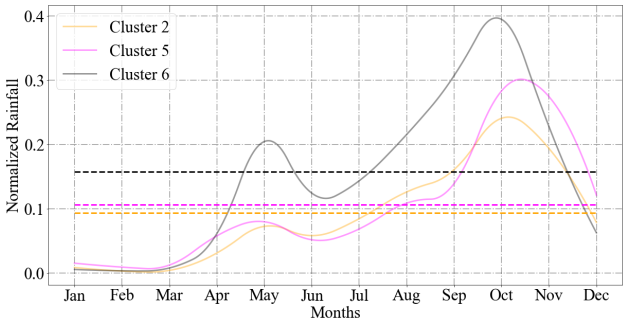

Fig. 6. Variation of Rainfall across months for Clusters 2, 5 and 6

For every cluster, the rainfall is deficient in the first three months of the year and is maximum in October. All the districts in Tamil Nadu, India receives a high amount of rainfall twice in a year. The first time it is caused by South-West Monsoon, and the second time it is caused by North-East Monsoon. Two patterns are observed in the variation of rainfall among the clusters which is displayed in Fig. 5 and Fig. 6. Clusters 1, 3 and 4 receive high rainfall in June and October, these are clusters of districts which lies on the western half of the state whose rainfall is influenced by the Western Ghats whereas clusters 2, 5 and 6 receives high rainfall in May and October, these are clusters of districts which lies on the eastern half of the state near the coastal regions. The dotted lines in Fig. 5 and Fig. 6 shows that cluster 4 has the maximum rainfall followed by clusters 6, 1, 5, 2 and 3 across the months in the respective order.

VI. Conclusion

In this paper, we have developed a regression model that predicts rainfall with minimum error and captures sudden fluctuations in it. Based on the analysis, it was observed that the Generic-Regression Model using Polynomial Regression with degree 4 outperforms all the other models and predicts the rainfall in all the districts with comparatively low error rates. However, the Cluster-Based Model using Polynomial Regression captures variation in most of the districts and performs better than the Generic-Regression Model only by a fractional value. Hence, it can be concluded that Generic-Regression Model is the best model to predict rainfall for the state of Tamil Nadu, India. Also, on an analysis of variation of rainfall among the formed clusters, it was concluded that the districts in the eastern half and western half of the state have distinct patterns of rainfall across the months.

Acknowledgment

We would like to express our sincere gratitude to Dr. C Shunmuga Velayutham, (Associate Professor, Department of Computer Science and Engineering, Amrita School of Engineering, Amrita Vishwa Vidyapeetham, Coimbatore) for the guidance and useful critiques of this research work.

References

[1] WTTC. Country Reports 2017 - India. Technical report, 2017.

[2] R. Ramanathan and M. Jayakumar. A support vector regression approach to detection in large-MIMO systems. *Telecommunication Systems: Modelling, Analysis, Design and Management*, 64(4):709–717, April 2017.

[3] J. Isaac and S. Harikumar. Logistic regression within dbms. In *2016 2nd International Conference on Contemporary Computing and Informatics (IC3I)*, pages 661–666, Dec 2016.

[4] Jinghao Niu and Wei Zhang. Comparative analysis of statistical models in rainfall prediction. In *2015 IEEE International Conference on Information and Automation*, pages 2187–2190. IEEE, aug 2015.

[5] V.P Tharun, Ramya Prakash, and S. Renuga Devi. Prediction of Rainfall Using Data Mining Techniques. In *2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT)*, pages 1507–1512. IEEE, apr 2018.

[6] Andrew Kusiak, Xiupeng Wei, Anoop Prakash Verma, and Evan Roz. Modeling and Prediction of Rainfall Using Radar Reflectivity Data: A Data-Mining Approach. *IEEE Transactions on Geoscience and Remote Sensing*, 51(4):2337–2342, apr 2013.

[7] Kesheng Lu and Lingzhi Wang. A Novel Nonlinear Combination Model Based on Support Vector Machine for Rainfall Prediction. In *2011 Fourth International Joint Conference on Computational Sciences and Optimization*, pages 1343–1346. IEEE, apr 2011.

[8] Sandeep Kumar Mohapatra, Anamika Upadhyay, and Channabasava Gola. Rainfall prediction based on 100 years of meteorological data. In *2017 International Conference on Computing and Communication Technologies for Smart Nation (IC3TSN)*, pages 162–166. IEEE, oct 2017.

[9] Sankhadeep Chatterjee, Bimal Datta, Soumya Sen, Nilanjan Dey, and Narayan C. Debnath. Rainfall prediction using hybrid neural network approach. In *2018 2nd International Conference on Recent Advances in Signal Processing, Telecommunications & Computing (SigTelCom)*, pages 67–72. IEEE, jan 2018.

[10] R. Venkata Ramana, B. Krishna, S. R. Kumar, and N. G. Pandey. Monthly Rainfall Prediction Using Wavelet Neural Network Analysis. *Water Resources Management*, 27(10):3697–3711, aug 2013.

[11] Mislan, Haviluddin, Sigit Hardwinarto, Sumaryono, and Marlon Aipassa. Rainfall Monthly Prediction Based on Artificial Neural Network: A Case Study in Tenggarong Station, East Kalimantan - Indonesia. *Procedia Computer Science*, 59:142–151, jan 2015.

[12] Aishwarya Himanshu Manek and Parikshit Kishor Singh. Comparative study of neural network architectures for rainfall prediction. In *2016 IEEE Technological Innovations in ICT for Agriculture and Rural Development (TIAR)*, pages 171–174. IEEE, jul 2016.

[13] Yajnaseni Dash, S.K. Mishra, and B.K. Panigrahi. Rainfall prediction of a maritime state (Kerala), India using SLFN and ELM techniques. In *2017 International Conference on Intelligent Computing, Instrumentation and Control Technologies (ICICICT)*, pages 1714–1718. IEEE, jul 2017.

[14] Chieko Palanisami, Kuppannan and R. Ranganathan, C and Senthilnathan, S and UMETSU. Diversification of Agriculture in Coastal Districts of Tamil Nadu- a Spatio- Temporal Analysis. page 673, 2009.