International Journal of Emerging Technology and Advanced Engineering
Website: www.ijetae.com (ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 7, Issue 8, August 2017)

Conglomerate Predictive Indicators for Rainfall Scalable Clustering and Classification Models

Smita Pallavi[1], Pratyaksha Sinha[2], Priyanka Dalmia[3]
[1]*Research Fellow, Dept. of CSE, Birla Institute of Technology Mesra, Patna Campus, India*
[2,3]*Research Student., Dept. of CSE, Birla Institute of Technology Mesra, Patna Campus, India*

***Abstract*—Indicators and techniques involved in prediction of weather conditions or rainfall estimation has been widely researched and classified by attribute selection and class labels in the dataset. The paper describes an integrated Analytic system that pre-processes large data through density estimation of clusters by pillar k-means and further classifies them by cluster label. The label of the data point is predicted by the help of two multiclass classification methods. First is by using decision tree or classification tree and second is through random forests which are ensembles of decision trees. These categories of supervised learning algorithms used to search interesting data patterns are further validated for accuracy. The knowledge predicted by this schema (KDD) empowers the pre-disaster management to take precautionary measures for rainy season. The data used Machine Learning library of Apache Spark, Mllib has been implemented in this paper.**

***Keywords*— Pillar K-means, Minkowski distance, Rainfall prediction, Decision trees, Random Forest**

I. INTRODUCTION

The estimation of amount of rainfall plays a vital role in flood forecasts considering the uncertainty in the input, output and model parameters. In this paper cascaded uncertainties are clustered and further classified to provide robust month ahead predictions. Data mining or knowledge discovery in databases (KDD) is the automatic extraction of implicit and interesting patterns from large data collections [8]. The multidisciplinary area of data mining contains convergence of computing paradigms like decision tree construction, random forests, artificial neural networks, bayesian learning, statistical algorithms, etc. and is mined using tools visualization, clustering, classification, association rule mining, sequential and text mining etc.

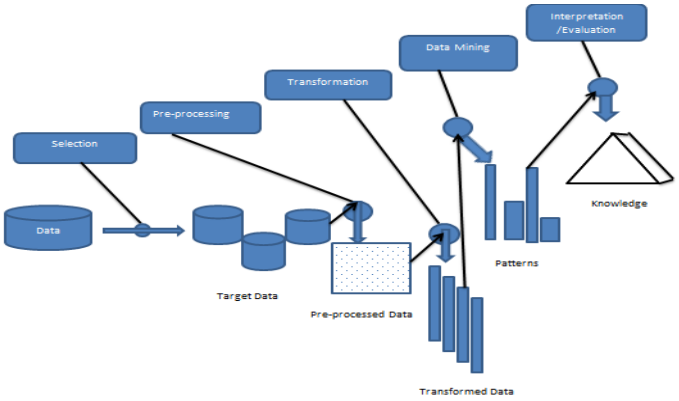

Fig 1 : The Data-Knowledge Interpretation Model

Source: Data Mining- Recommender Systems 0.7.5 documentation

Rainfall forecasts are limited by atmospheric dynamics parameterisation and are sensitive to the initial condition patterns. This paper proposes a methodology for predicting rainfall through pillar K-means clustering and classification techniques. Weather data of Patna District, Bihar has been collected from www.indiawaterportal.org as a case study. The data contains monthly mean for each of 6 parameters which include- Average temperature(in Celsius), minimum temperature(in Celsius),maximum temperature(in Celsius), precipitation(in mm), cloud cover(in %) and vapor pressure(in hPa). The rest of the paper is organized as follows: a brief literature review of the existing techniques of weather predictions is presented in section II, the model adopted is traversed in section III, findings and results are analytically discussed in section IV followed by conclusion in section V.

International Journal of Emerging Technology and Advanced Engineering
Website: www.ijetae.com (ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 7, Issue 8, August 2017)

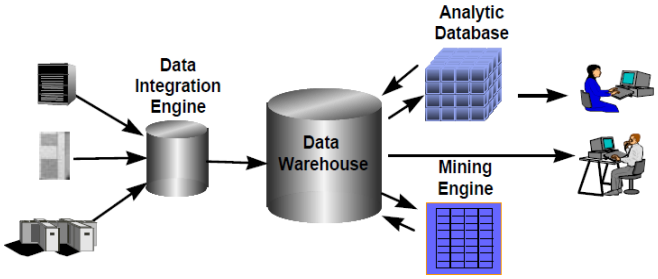

Fig 2 : Database infrastructure for analytical applications

II. Related Literature Review

Rainfall Prediction is very important for an agricultural country like India. In recent past, related work with ability to predict the future by studying past patterns are listed as :-

Arit Thammano et. al.(2013)[1] employed entropy concept to adapt the traditional K-means by employing a scheme to select the initial cluster centers. The models were tested on seven benchmark data sets from the UCI machine learning repository. Experimental results were shown to outperform the learning vector quantization network in most of the tested data sets.

Chakraborty, S et.al(2012) [2] used incremental K-Means clustering to develop list of weather categories based on Air Pollution database. For meteorological data clustering simple K-means and DBSCAN are simulated on real time air pollution data of the city. Performance analysis of the two algorithms has been done. To achieve better clustering simple K-means and DBSCAN algorithms were combined. Then hybrid DBK-means algorithm was proposed to predict future weather condition. The accuracy of the proposed method was 83.3%.

Chau et.al(2010) [3] integrated ANN-SVR artificial neural networks and support vector regression for daily rainfall prediction by Fuzzy C-means clustering to split the training set into three crisp subsets to be associated with low-, medium- and high-intensity rainfall. Two local artificial neural network models enumeration and correlation analysis were involved in training and predicting low- and medium-intensity subsets whereas a local support vector regression model was applied to the high-intensity subset.

A.Kumar et.al(2012)[4] used the numerical results generated through the Probability Density Function algorithm as the basis of recommendations in favor of the K-means clustering for weather-related predictions.

They proposed a model for predicting the probability of the outcome of the Play class as YES or NO through K-means clustering on weather data.

Pappenberger F., (2015)[6] implemented rainfall forecasts based on the European Centre for Medium Range Weather Forecasting (ECMWF) control and ensemble forecasts (known as the Ensemble Prediction System or EPS). 50 additional realisations for the next 10 days with a grid scale of 80 km were proposed. In addition, a deterministic forecast was provided, which was the control forecast at higher resolution (40 km) and thus claim to provide better precipitation predictions.

Random forest model has also been used for monthly temperature forecasting for historical time series data [10].

The novelty of our work is to adopt pillar K-means for clustering and usage of Apache Spark provides caching capabilities, hence results in faster decision making.

III. Methodology And Model Adopted

The objective was to analyse the 102 years (1901-2002) monthly means of the parameters and categorize the data according to the 4 important parameters responsible for rainfall. In this paper, the analysis is based on the weather data of Patna district, Bihar state which was collected from the website www.indiawaterportal.org. The context of case study undertaken is shown as :-

Fig 3 : Heavy Rainfall Zones of Bihar

Source :- BMTPC, India

It contains monthly means of each year of attributes that affect rainfall. Hence the total number of data points computed was 102x12=1224. The feature set is as shown:-

International Journal of Emerging Technology and Advanced Engineering
Website: www.ijetae.com (ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 7, Issue 8, August 2017)

TABLE I
FEATURES NOTED FOR AMOUNT & CLASS OF RAINFALL PREDICTION

Feature #	Particulars of the features	
	Description	Unit
1	Minimum Temperature	^{O}C
2	Maximum Temperature	^{O}C
3	Average Temperature	^{O}C
4	Precipitation	mm
5	Cloud Cover	%
6	Vapour Pressure	hpa
7	Relative Humidity	%

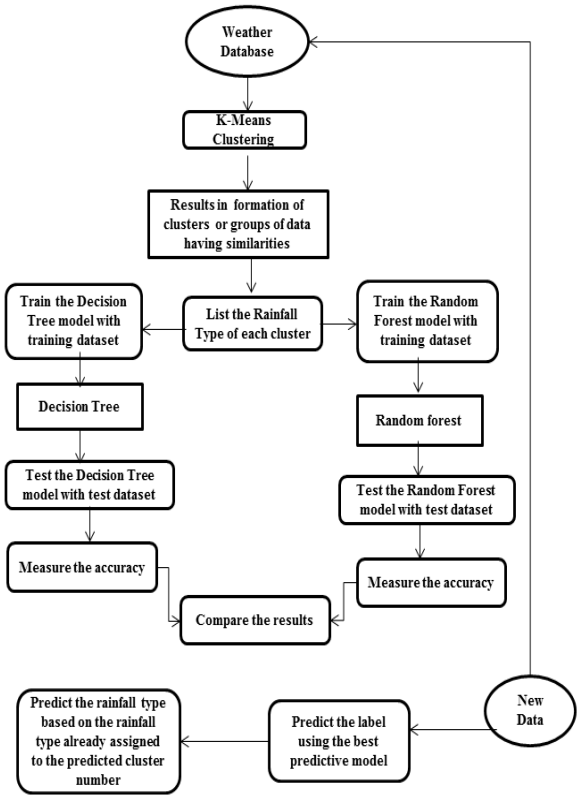

Fig 4 : Our Proposed Work Flow Diagram

The process flow of our work involves application of Pillar K-Means clustering on the weather database, which results in the formation of clusters of data. We assign the rainfall type to each cluster and then train the decision tree model and the random forest model. After this we test the two models for accuracy and compare the respective accuracies. We choose the model with higher accuracy as our predictive model. When new data enters the database, we predict the rainfall type of that data by predicting the cluster number or label of the data point with the help of our predictive model.

We then conclude the rainfall type by looking at the pre assigned rainfall type of the predicted cluster number. Apache Spark mllib used in the proposed work provides caching capabilities, hence results in faster decision making. It's in-memory clustering capability increases the speed of applications. It has been made on top of Resilient Distributed Datasets(RDDs), which makes it fault tolerant. Clustering is an unsupervised machine learning technique which divides the dataset into subsets, called clusters. The data is divided into four categories by applying pillar K-Means clustering algorithm on the dataset and each data point is assigned a cluster number. The data points in the same cluster share some similarities. We have used Apache Spark, since it provides in-memory clustering which increases processing speed of applications. The rainfall categories developed are scanty, light, moderate and heavy.

Pillar K-means Algorithm :-

Let X ={ x_i }, i = 1 ,2,….n be the set of n d-dimensional points to be clustered into a set of M clusters, C = {c_m; m= 1, 2, . . . M}. K-means algorithm finds a partition such that the squared error between the empirical mean of a cluster and the points in the cluster is minimized.

Following formula is to compute minimum distance between centroid and object (feature) for form clusters,

$$j = \sum_{j=1}^{k} \sum_{i=1}^{n} \left\| x_i^{(j)} - c_j \right\|^{\lambda} \quad \text{.....(i)}$$

where $\left\| x_i^{(j)} - c_j \right\|^{\lambda}$ is chosen distance measure between a data point $x_i^{(j)}$ and the cluster centre c_j ,which is at an distance of the n data points from their k respective cluster centres. For any $\lambda > 0$, it is seldom used for values further than 1, 2 and ∞. Where, j is k-means objective function to be reduced, c_j is centroid of the j^{th} cluster, k is number of clusters, n is total data points and $x_i^{(j)}$ is j^{th} data point. [9]

The Minkowski distance hence calculated is :-

$$\partial_k = \frac{1}{2}\left(Y_n - O_n\right)^2 \quad \text{.....(ii)}$$

International Journal of Emerging Technology and Advanced Engineering
Website: www.ijetae.com (ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 7, Issue 8, August 2017)

where ∂_k is the k^{th} learning error , Y_n is the desired/predicted output and O_n is the actual output.

Classification Algorithms :-

1) Decision Trees :-

Decision tree is a classifier in the form of a tree structure which consists of the following :- [7]

Decision node: specifies a test on a single attribute.

Leaf node: indicates the value of the target attribute.

Edge: split of one attribute

Path: a disjunction of test to make the final decision. Decision trees classify instances or examples by starting at the root of the tree and moving through it until a leaf node.

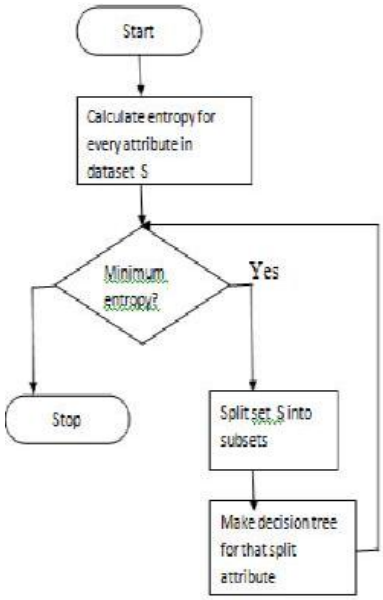

Fig 4 : Flow Diagram of Decision Tree Algorithm

On the chosen real dataset, multiclass classification is performed using Decision Tree which aids in predicting the label of the unseen data points with the help of a tree whose leaves denote the labels and branches represent the decision based on the input features. The model hence formed is tested against the test data to categorize the rainfall corresponding to the data point into one of the four categories. The same classification is then performed with the help of random forests that are ensembles of decision trees which use more than one decision trees to predict the target.

2) Random Forest Algorithm

Random forests are an ensemble method used for classification. Random forests are used to rank the importance of variables in a classification problem.

To measure the importance of a variable in a data set $S_n=\{(X_i,Y_i)\}$ i=1….n we fit a random forest to the data. During the fitting process the error for each data point is calculated and averaged over the forest.

Algorithm_Random_Forest

i] Consider **n**- #of training cases; **v**- # of classifier variables

ii] **p**- # of decision variables at the node of a tree **(p<v)**

iii] Create training set for **k** times with replacement from **n** training cases by bootstrapping computation. The *k* set of overlap between 1^{st} ,2^{nd} block = k overlap 2^{nd} ,3^{rd} and so on. The remaining cases estimate the error of the random tree.

iv] For each node :- Random variables are selected on which to search for the best split. Class of new data can be predicted by considering the majority votes in the tree.

v] Calculate the best split with least deviation based on chosen variables in the training set. The whole tree is retained without pruning.

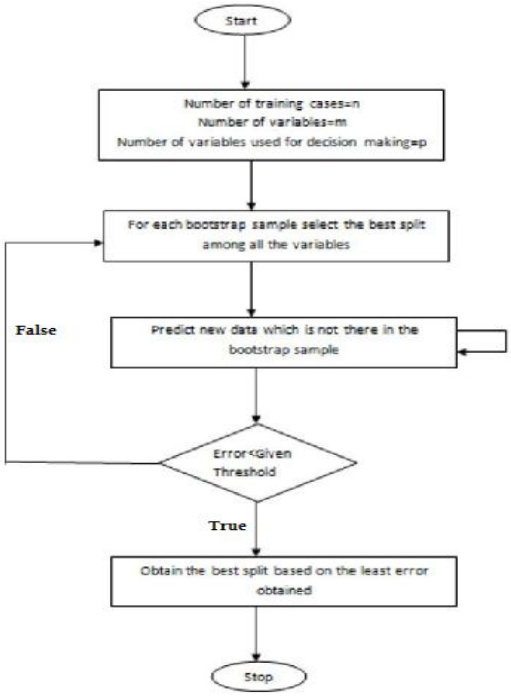

Fig 5 : Flow Diagram of Random Forest Algorithm

International Journal of Emerging Technology and Advanced Engineering
Website: www.ijetae.com (ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 7, Issue 8, August 2017)

Parameter Evaluation Metrics

i) Gini Index :-

The Gini Index is a measure of the impurity of each node t.

For binary classification, it can be computed as:

$$\text{Gini (t)} = 1 - \sum_{i=1}^{m} \left\| p_i^2 \right\| \quad \text{.....(iii)}$$

where
—t is referred as node,
—i is the class label of class C
—p(i|t) is the conditional probability that subject fell in the node t belongs to the class i

ii) Confusion Matrix:-

The confusion matrix is a visualization tool commonly used to present performances of classification tasks. It shows the relationships between real class attributes and that of predicted classes. The level of effectiveness of the classification model is calculated with the number of correct and incorrect classifications.

iii) Root Mean Square Error :-

A comprehensive assessment of model performance comprises the Root Mean Square Error (RMSE) and the Nash–Sutcliffe Coefficient of Efficiency (NCE). They are respectively formulated as :-

$$\text{RMSE} = \sqrt{\frac{1}{m}\sum_{i=1}^{m}\left(f_i - O_i\right)2} \quad \text{.....(iv)}$$

$$\text{NCE} = 1 - \frac{\sum_{i=1}^{m}\left(f_i - O_i\right)2}{\sum_{i=1}^{m}\left(f_i - M_i\right)2} \quad \text{.....(v)}$$

where m is the number of observations; f_i stands for the forecasted rainfall; O_i is the observed rainfall; M_i denotes the average observed rainfall. NCE =1 is a perfect fit.

iv)Accuracy :-

The predictive accuracy of the classifier measures the proportion of correctly classified instances

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

v) True Positive Rate (TPR/ Recall/ Sensitivity): It measures the rate of occurrence of true positives. i.e. percent of actual positives that are correctly classified

TPR = TP/(TP + FN).....(vii)

vi)True Negative Rate (TNR or Specificity):

It measures the percent of actual negative examples that are correctly classified = TN/(TN + FP).....(viii)

vii) Positive Predictive Value (PPV/ Precision):

It measures the percentage of the examples predicted to be positive that were correct = TP/(TP + FP).....(ix)

viii) False Negative Rate (FNR):

It measures the percentage of positive examples that were incorrectly classified.

FNR = FN/(TP + FN) = 1 – TPR .....(x)

ix) False Positive Rate (FPR):

It measures the percentage of negative example that were incorrectly classified.

FPR = FP/(TN + FP) = 1 – TNR .....(xi)

IV. Dataset Description

The total number of data points considered for rainfall categorization was 102x12=1224 stored in csv format. A part of the database is shown below:

TABLE II
Sample Data Synthesised For Model Implementation

#	Year	Mon	Avg	Min	Max	Preci	Clcove	RH	Press
1	X01	Jan	15.26	8.06	22.47	61.81	22.672	44.4	12.10
2	X01	Feb	18.87	11.4	26.38	23.44	22.602	37.9	13.06
3	X01	Mar	24.20	16.0	32.39	4.177	29.132	26.7	13.01
4	X01	Apr	30.77	22.7	38.81	8.506	25.191	22.9	15.89
5	X01	May	32.93	25.7	40.16	8.874	34.144	31.4	23.38
6	X01	Jun	34.20	28.8	39.67	102.6	58.715	45.6	33.06
7	X01	Jul	30.36	26.6	34.10	171.3	73.110	62.1	33.25
8	X01	Aug	29.12	25.7	32.50	266.5	68.575	68.7	33.64
9	X01	Sep	28.64	24.8	32.52	208.2	57.506	65.0	31.87
10	X01	Oct	27.21	21.9	32.52	23.64	34.717	48.9	23.98
11	X01	Nov	21.41	14.5	28.36	5.664	20.209	42.7	16.50
12	X01	Dec	16.71	9.42	24.04	0.000	23.054	46.2	13.82
13	X02	Jan	17.32	10.1	24.55	4.640	22.672	43.3	13.39
14	X02	Feb	19.26	11.8	26.77	3.255	22.602	38.4	13.51
15	X02	Mar	26.40	18.2	34.60	6.051	29.132	25.2	13.91

International Journal of Emerging Technology and Advanced Engineering
Website: www.ijetae.com (ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 7, Issue 8, August 2017)

An additional parameter *month of the year* was added for clustering. The month attribute was converted as a categorical parameter field containing values from 0 to 11 representing January to December respectively. The relevance of Year column was not much observed hence removed from the final dataset.

Analysis of clustering showed that the clusters were more homogenous if the precipitation attribute was excluded. Also, rainfall is a type of precipitation, i.e., liquid precipitation. Hence, amount of rainfall is the same thing as amount of liquid precipitation. So the final model was formed by the pillar k means clustering including the 5 parameters namely, Month, Avg Temperature, Relative Humidity , Cloud Cover and Vapor Pressure.

TABLE III
FINAL DATA SET AFTER FEATURE REDUCTION

#	Mon	AvgTemp	cloud.cov	RH	pressure
1	0	15.268	22.672	44.4674	12.102
2	1	18.879	22.602	37.9655	13.061
3	2	24.209	29.132	26.7751	13.019
4	3	30.774	25.191	22.9536	15.894
5	4	32.936	34.144	31.4118	23.382
6	5	34.206	58.715	45.6138	33.069
7	6	30.369	73.110	62.1360	33.252
8	7	29.121	68.575	68.7481	33.645
9	8	28.644	57.506	65.0569	31.878
10	9	27.212	34.717	48.9406	23.981
11	10	21.416	20.209	42.7187	16.500
12	11	16.710	23.054	46.2076	13.829
13	0	17.322	22.672	43.3995	13.393
14	1	19.265	22.602	38.4022	13.514
15	2	26.400	29.132	25.2931	13.919

For selecting the appropriate number of clusters, the sum of squared errors for all the clusters was computed and the elbow method was opted to analyse the graph of wssse versus number of clusters. K-Means was performed for k= 2 to 20. For each iteration the within set sum of squared errors were calculated and a plot of SSE vs. Number of clusters was analysed. The appropriate number of clusters was found as 4. Thus, four rainfall categories were developed. After this the cluster number column is appended to the data points.

The data in the form of comma separated values was parsed into dense vector. The clustering parameters (four) were initialized and number of iterations was initialized as 50. The Pillar K-means model was then trained and allocation of clusters was found.

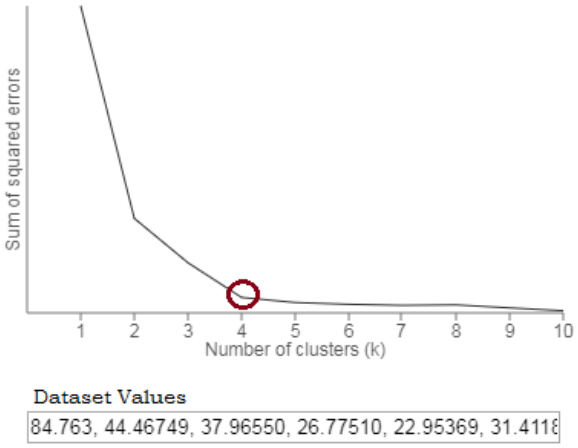

Fig 6 : Elbow Method depicting Four Clusters for Dataset

V. SIMULATION RESULTS AND DISCUSSIONS

The allocation variable was an RDD, in the form of (Key,Value) pairs, where key is the cluster number and value is the month label (0 – 11). The cluster distribution was analysed by grouping the RDD by the key i.e., by the cluster number.

The following table shows the cluster number and the corresponding month. This shows that the months in the same cluster share similar rainfall patterns. The months of June and October generally have moderate rainfall whereas July, August and September witness heavy rainfall. January, February, November and December share the same kind of rainfall pattern, that is, Scanty rainfall or almost no rainfall. Light rainfall is seen in the months of March, April and May.

TABLE IV
CLUSTER FORMATION WITH MONTHLY LABELLED SETS

CLUSTER #	MONTH LABEL	MONTHS
0	5,9	June, October
1	2,3,4	March, April, May
2	0,1,10,11	January, February, November, December
3	6,7,8	July, August, September

International Journal of Emerging Technology and Advanced Engineering
Website: www.ijetae.com (ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 7, Issue 8, August 2017)

Hence, Rainfall patterns can be assigned as follows:-

TABLE V
CLUSTER CLASSIFICATION WITH RAINFALL TYPE

CLUSTER NUMBER	RAINFALL TYPE
0	Moderate
1	Light
2	Scanty
3	Heavy

The cluster number for each data point was then appended in the data set and the month label was removed, in order to perform classification using Decision Tree and Random Forest Algorithm of Spark Mllib.

The resultant data set obtained was parsed and is divided into training and test sets in the ratio of 70:30. Thus, 857 data points were used for fitting the model (training set) and the rest 369 were used to test the accuracy of prediction (test set) .

```
If (feature 1 <= 23.577)
 If (feature 3 <= 27.68910997)
  Predict: 1.0
 Else (feature 3 > 27.68910997)
  If (feature 2 <= 34.144)
   Predict: 2.0
  Else (feature 2 > 34.144)
   If (feature 0 in {10.0})
    Predict: 0.0
   Else (feature 0 not in {10.0})
    Predict: 2.0
Else (feature 1 > 23.577)
 If (feature 3 <= 36.21301231)
  If (feature 2 <= 45.187)
   Predict: 1.0
  Else (feature 2 > 45.187)
   Predict: 0.0
 Else (feature 3 > 36.21301231)
  If (feature 3 <= 59.84651143)
   If (feature 2 <= 30.015)
    Predict: 2.0
   Else (feature 2 > 30.015)
    If (feature 2 <= 66.972)
     Predict: 0.0
    Else (feature 2 > 66.972)
     Predict: 3.0
  Else (feature 3 > 59.84651143)
   If (feature 2 <= 55.626)
    If (feature 3 <= 64.74179682)
     Predict: 3.0
    Else (feature 3 > 64.74179682)
     Predict: 3.0
   Else (feature 2 > 55.626)
    Predict: 3.0
```

Fig 7 : Classification Decision Tree Model (DT)

The DT Classification Model predicts the cluster label, given a certain combination of Avg temperature, Relative humidity, Cloud cover and Vapor pressure. Thus a rainfall category for each data point was predicted thus indicating disasters like Drought and Flood for rainfall categories of Scanty and Heavy Rainfall respectively.

Thus, for D as set of examples over Feature Space X and a set of classes C= $\{c_1,c_2,c_3,c_4\}$

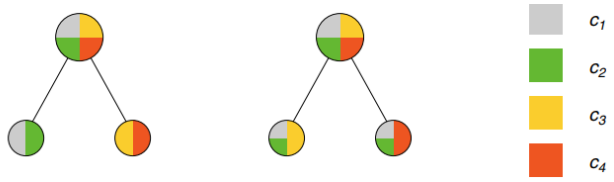

Fig 8 : Cluster Split w.r.t impurity identified by Gini Index

TABLE VI
CLUSTER NUMBER APPENDED DATASET FOR CLASSIFICATION

#	Cluster	Avg	Cloud.Co	RH	Pressure
1	2	15.268	22.672	44.46749	12.102
2	2	18.879	22.602	37.96550	13.061
3	1	24.209	29.132	26.77510	13.019
4	1	30.774	25.191	22.95369	15.894
5	1	32.936	34.144	31.41186	23.382
6	0	34.206	58.715	45.61382	33.069
7	3	30.369	73.110	62.13608	33.252
8	3	29.121	68.575	68.74818	33.645
9	3	28.644	57.506	65.05698	31.878
10	0	27.212	34.717	48.94069	23.981
11	2	21.416	20.209	42.71877	16.500
12	2	16.710	23.054	46.20760	13.829
13	2	17.322	22.672	43.39955	13.393
14	2	19.265	22.602	38.40222	13.514
15	1	26.400	29.132	25.29317	13.919

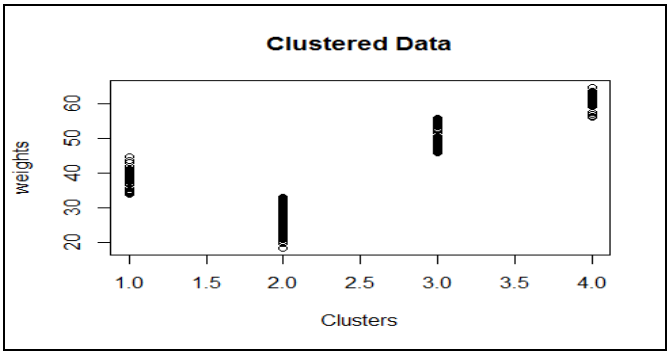

Fig 9 : Cluster Formed on the Experimental Rainfall Dataset

The time taken to construct the Random Forest Model with the Training Dataset was 0.768 seconds which was greater as compared to Decision tree Model of 0.115 seconds. However it was observed that the Re-substitution Error Rate (RER) which is calculated as Product of Relative error and Root Node Error, of random forest is less than decision tree since random forest is ensemble of trees.

International Journal of Emerging Technology and Advanced Engineering
Website: www.ijetae.com (ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 7, Issue 8, August 2017)

RER (DT) = 0.7155 * 0.11491 = *0.082218105*
RER (RF) = 0.5607 * 0.76756 = *0.04303709*

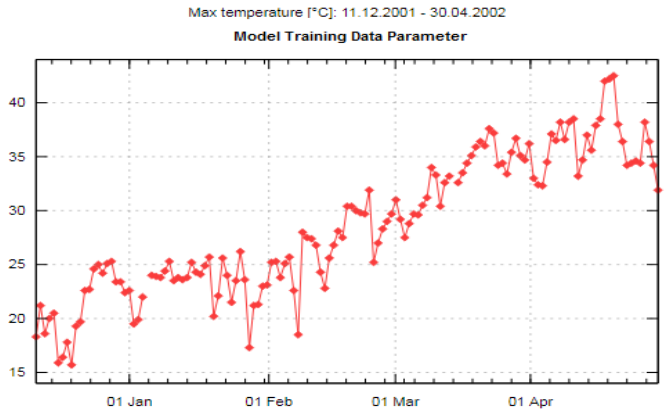

Fig 10 : Average Temperature for Trained Model Dataset

TABLE VII
Confusion Matrix For Testing Dataset For Both Classifications

Decision tree Model	OBSERVED			
PREDICTED	Confusion Matrix	YES	NO	TOTAL
	YES	TP = 208	FP = 5	213
	NO	FN = 6	TN = 150	156
	TOTAL	214	155	369
Random Forest Model	OBSERVED			
PREDICTED	Confusion Matrix	YES	NO	TOTAL
	YES	TP = 201	FP = 2	203
	NO	FN = 4	TN = 162	166
	TOTAL	305	164	369

The number of instances used for testing the DT and RF classification models were 369. The comparative confusion matrix shows the following performance measure metrics.

TABLE VIII
COMPARATIVE PERFORMANCE MEASURES FOR CLASSIFICATION MODELS

Model Type	Accuracy	TPR	TNR	PPV	FNR	FPR
Decision Tree	0.97	0.972	0.9677	0.976	0.028	0.0122
Random Forest	0.9837	0.98	0.9878	0.99	0.02	0.0323

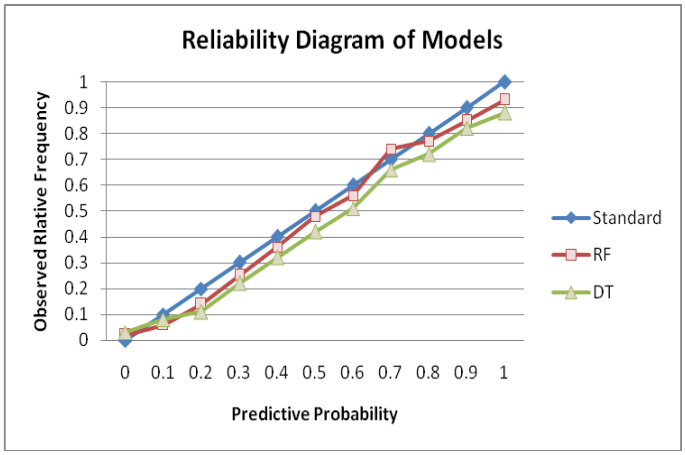

Fig 11 : Graphical Reliability of Prediction of Rainfall

VI. CONCLUSION

In this paper, a hybrid technique is established to predict monthly rainfall by pillar Kmeans and twin classifiers.

The results of both training and testing sets showed that the Random Forest in general performs better than decision tree in the classification. Monthwise Prediction model of clustered classes with Avg Temperature, Relative Humidity, Cloud Cover and Vapor Pressure as parameters performed with accuracy of 98.37 and 99% precision in case of Random Forest classification model and that of accuracy of 97% and 97.6% precision with Decision Tree model.

In future, other variants of modified K-means clustering algorithms can be used to predict online weather conditions with extended feature selection models. Optimization for big analytics can also be improvised.

REFERENCES

[1] Arit Thammano, Pannee Kesisung, "Enhancing K-means algorithm for solving classification problems", IEEE International Conference on Mechatronics and Automation (ICMA) 2013, pp. 1652-1656, 2013.

[2] Chakraborty, S. , Nagwani, N.K., Dey, L. , "Weather Forecasting using Incremental K-means Clustering", CiiT International Journal of Data Mining & Knowledge Engineering, May 2012.

[3] K. W. Chau and C. L. Wu , " A hybrid model coupled with singular spectrum analysis for daily rainfall prediction", Journal of Hydroinformatics | 12.4 | 2010, pp 458-472

[4] Kumar Abhay, Ramnish Sinha, Daya Shankar Verma, Vandana Bhattacherjee, Satendra Singh "Modeling using K-Means Clustering Algorithm" , 1st Int'l Conf. on Recent Advances in Information Technology | RAIT-2012 |

[5] Ordonez, C. and Omiecinski E.: "An Efficient Disk- Based K-Means Clustering for Relational Databases, IEEE transaction on knowledge and Data Engineering,Vol.16,2004.

[6] Pappenberger F., K. J. Beven, N. M. Hunter, P. D. Bates, B. T. Gouweleeuw, et al.. Cascading model uncertainty from medium range weather forecasts (10 days) through a rainfall-runoff model to flood inundation predictions within the European Flood Forecasting System (EFFS). Hydrology and Earth System Sciences Discussions, European Geosciences Union, 2015, 9 (4), pp.381-393.

[7] Purva Sewaiwar,K.K. Verma : "Comparative Study of Various Decision Tree Classification Algorithm Using WEKA"‖International Journal of Emerging Research in Management and Technology(IJERMT) ISSN: 2278-9359, Vol. 4, Issue 10,Oct 2015

[8] Romero, C., & Ventura, S. (2010). Educational data mining: a review of the state of the art. Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on, 40(6), 601-618.

[9] Smita Pallavi, K.Lal, S.P. Lal (2017) " Analyse the Academic Performance of Students Using ANN Classification with Modified Pillar K-means and IWFA", Wireless Personal Communications, Springer Professional , ISSN: 0929-6212 ,E ISSN: 1572-834X

[10] Vaibhavi Mistry , Vibha Patel "Weather Condition Prediction Using Semi-Supervised Data Mining Technique", International Journal of Engineering Trends and Technology (IJETT), V20(4),179-183 Feb 2015. ISSN:2231-5381. www.ijettjournal.org. published by seventh sense research group

[11] Veronica J. BERROCAL, Adrian E. RAFTERY, Tilmann GNEITING, and Richard C. STEED "Probabilistic Weather Forecasting for Winter Road Maintenance" American Statistical Association Journal of the American Statistical Association June 2010, Vol. 105, No. 490

[12] G.Shrivastav, "Application of ANN for weather Forecasting", International Journal of Computer Applications, Volume 51 No.18, August 2012.

[13] Wikipedia contributors, ―C4.5_algorithm,‖ Wikipedia, The Free Encyclopedia. Wikimedia Foundation, Accessed 28- March-2017.

[14] Wikipedia contributors, ―Random_tree,‖ Wikipedia, The Free Encyclopedia. Wikimedia Foundation, Accessed 13-Jun- 2017.