Design of Data Mining Tools for Rainfall Forecasting

[1]M.S.Chaudhari, [2]Dr.N.K.Choudhari

[1]Research Scholar, [2]Principal,
[2]Priyadarshini Bhagwati College of Engineering,
Nagpur, India.

Abstract : The paper presents the classification and regression model developed for the rainfall forecasting of Nagpur, Pune, Mumbai and Delhi rainfall data set for the year 2008 to 2014. The research work used regression algorithms Regression Tree(RT), Multiple Linear Regression(MLR), Support Vector Machine(SVM) and Artificial Neural Network(ANN) for the development of regression model and found SVM as the best performing model using PCA as dimensionality reduction technique. The goodness of fit and predictive skill measures were used to compare the models. The classifiers Random Forest(RF),Decision Tree(DT), Naïve Bayes(NB) and K-nearest Neighbour(KNN) and found DT and NB models best classification models both before and after application of PCA using confusion matrix .

***IndexTerms* – Regression, Classification, Principal Component analysis, Data Mining, Machine Learning.**

I. INTRODUCTION:

Data mining is the process of extracting information or knowledge from the large databases also called as knowledge discovery from database (KDD). The environmental systems like rainfall consist of huge amount of data containing the large number of features or variables affecting the rainfall. Data mining along with machine learning algorithms proves to be the most useful for rainfall forecasting.

Environmental system like rainfall processes involves distributed, heterogeneous or decentralized data sources. Hence for today's resource driven world we are very much concerning with the best understanding and utilization of our natural environment and resources like the water received from rainfall . Hence these natural resources like rainfall requires significant applications of data mining. Data mining has wide variety of applications that will certainly make a high impact in the area of integrated data fusion and mining environmental applications.

India is an agricultural country where 60\% percent of population depends on the farming sector. Most of this farming sector is dependent on the rainfall that occurred during the four months of Monsoon that is from June to September. But the uncertainty in rainfall occurrences leads to the poor yield or no crop to the farmer that causes huge loss to farming. Hence the accurate prediction of rainfall with the lead time of years, months, weeks, days and hours can help water management to impressively control situation arising out of scarcity of water, floods and problems arising out of drinking water in the summer.

This research work has used data mining techniques in combination with the statistical approach and machine learning algorithms. The rainfall data is very large, voluminous , noisy and unstructured that requires preprocessing techniques of data mining, statistics to analyze and visualize the patterns and machine learning algorithms to forecast the future events using the past available information and various variables that contribute to the rainfall events and their correlations on the rainfall.

The tasks of classification and regression are concerned with predicting the value of one field from the values of other fields. The target field is called the class or dependent variable in statistical techniques. The other fields are called attributes or independent variables in statistical terminology. If the class is continuous or consists of numerical values, the task at hand is called regression. If the class is discrete or it has a finite set of nominal values, the task at hand is called classification. In both cases, a set of data is taken as input, and a model (a pattern or a set of patterns) is generated. This model can then be used to predict values of the class for new data.

II. RELATED WORK:

[1] has developed a data driven prediction model based on decision tree that maps climatic variable like temperature ,humidity, wind speed etc. The performance of the models is evaluated on three indicators like Naish Sutcliffe efficiency, RMSE,MSE etc. The performance of the model is compared with data driven ANN based model. First data preprocessing and data selection is done followed by extracting hidden relationship and developing decision tree from categorical and numerical variable followed by rainfall prediction. In this model correlation is applied among the variables of data set and a new data set with important variable has been generated. Then data set is prepared with preprocessing techniques and divided into 80 \% training and 20 \% test dataset . Three data mining algorithms are applied to map variable for rainfall prediction. They concluded that decision tree is more accurate than ANN with poorly correlated variable.

[2] has proposed ARMA time series techniques which is linear, model driven and nonparametric method. The data driven backpropogation ANN approach and KNN non parametric regression methods are used. The models were evaluated using RMSE, MAE, COP, Efficiency coefficient ,correlation coefficient and index of agreement.

[3] has investigated the ability of fuzzy rules/logic in modeling rainfall for South Western Nigeria. The developed Fuzzy Logic model is made up of two functional components; the knowledge base and the fuzzy reasoning or decision-making unit. Two operations fuzzification and defuzzification were performed on the Fuzzy Logic model. The input variables of data set temperature, pressure, humidity, dew point and wind speed are inputted for fuzzification model and fuzzy set is formed using membrane function.

[4] has proposed an improved accuracy rainfall prediction using preprocessing procedure like moving average (MA) and singular spectrum analysis (SSA) and used modeling method like local SVR and local ANN.

The fuzzy logic model adopted in this work composed of two functional components, the knowledge base, which contains a number of fuzzy if–then rules and a database to define the membership functions of the fuzzy sets used in the fuzzy rules.

In [5] rainfall prediction is implemented with the use of empirical statistical technique. She used 6 years (2007-2012) data set such as minimum temperature, maximum temperature ,pressure, wind direction ,relative humidity etc and performed prediction of Rainfall using Multiple Linear Regression(MLR) . This model forecasts monthly rainfall amount in summer monsoon season (in mm).

[6] has used principal component analysis for forecasting rainfall. The proposed PCA method is used when there is vital inter-correlation between the predictors. The PCA model avoids the inter-correlation and support to reduce the degrees of liberty by controlling the number of predictors. Their experiment studies therefore suggest that PCA has some more benefits over ANN in analyzing climatic time series such as rainfall, particularly with regards to the interpretability of the extracted signals.

III. SYSTEM DESIGN AND METHODOLOGY:

Data mining is the process of extraction of information or patterns from large databases. The quality of the information extracted by data mining algorithms largely depends on the quality of data sets used for the experimental purpose. These data sets are highly susceptible to noise, missing and inconsistent data due to their huge size because they are originated from multiple, heterogeneous sources. Low-quality data leads to low quality mining results.

Hence to obtain accurate prediction results the data must be of quality i.e. it should be accurate, complete and consistent. But the most of the data sets are inaccurate, incomplete and inconsistent. Specially the environmental data often includes measurement errors , uncertainty, imprecision, multi-scalarity, heterogeneous, non-linear, non stationary and non-normality. Following are the detail system design and methodology steps adopted in the design and development of DM Tools for rainfall forecasting.

1.To Develop and understand the domain, to capture relevant prior knowledge and the goals of the end-user:

In this research work the specific ES domain identified and understood is rainfall prediction. The rainfall process is very uncertain and a lot of factors are associated that affects the phenomena of rainfall. It's very difficult to accurately predict the rainfall considering variables affecting it.

2.To create the target data set by selecting a proper set of variables:

This research has identified four different data sets with number of relevant variables for 4 cities of India that is Nagpur, Pune, Mumbai and Delhi.

3. Data cleaning and preprocessing:

The various preprocessing techniques like removing outliers, replacing missing values by mean or median, normalization, Principal Component Analysis (PCA) and visualization techniques are identified and applied for the proposed work.

4.Data reduction and projection:

This work has implemented PCA as dimensionality reduction techniques for data reduction i.e. reducing the variables and Min-Max transformation for data projection.

5.Choosing the data mining task, with reference to the goal of the KDD process :

This work has identified classification and regression task for the development of the DM tool.

6.Selecting the data mining algorithms and Machine Learning algorithm :

In this research approach , the algorithms Artificial Neural Network (ANN), SVM(Support Vector Machine),Regression Tree(RT), and Multilinear Regression (MLR) are identified and applied for development of regression model. The algorithms Random Forest(RF),Naive Bayes (NB), Decision Tree (DT) and K Nearest Neighbour (KNN) are identified and applied for development of classification model .

7.Data Mining :

The results obtained for above task are analysed by using the goodness of fit and predictive accuracy criteria . The improvement in the results has been observed by using PCA as a preprocessing step for regression and classification model.

8.Interpreting mined patterns:

In this work,the results obtained are interpreted by visualizing the plots of both predicted and observed values of rainfall .

3.1Data Preprocessing:

Preprocessing is required to better understand the data set, to detect imperfections in data sets and manage them in the proper way and to correctly prepare data for the selected DM techniques. It is done to remove the noise that might be present in the data set due to error in observation. Hence to obtain quality results, these data sets needs preprocessing before applying them to various data mining algorithms to get correct patterns or result. The purpose of data preprocessing is to clean noisy data, extract and merge the data from different sources and then transform and convert the data into a proper format.

3.2. Principal Component Analysis:

Principal Components Analysis (PCA) is one of the several statistical tools available for reducing the dimensionality of data sets. The major goal of principal component analysis is to reveal hidden structure in a data set. PCA is able to identify how different variables work together to create the dynamics of the system, to reduce the dimensionality of the data, decrease redundancy in the data, filter some of the noise in the data ,compress the data, prepare the data for further analysis [7].

Thus the central idea of PCA is to reduce the dimensionality of a data set consisting of a large number of interrelated variables, while retaining as much as possible of the variation present in the data set. The PCA approach uses all of the original variables to obtain a smaller set of Principal Components (PCs) which can be used to approximate the original variables. PCs are uncorrelated and are ordered so that the first few components retain most of the variation present in the original set.

This research work is categorically divided into two main parts. That is regression and classification modeling. The main objective of this research work is to design regression model to forecast the rainfall as a response numeric variables, to design the classification model to predict the rainfall as response categorical variables .

3.3 Evaluation of Model:

The regression models are evaluated using goodness of fit for in sample data and cross validation for predictive skill using out of sample data. For goodness of fit, the metrics MSE(Mean Square Error),RMSE(Root Mean Square),MASE(Mean Absolute Scaled Error),MAE(Mean absolute error) and Correlation Coefficient(COR) are used. For cross validation, K- Fold cross validation(K-Fold) , Repeated K Fold(R-KFold) cross validation ,Leave one out cross validation(LOOCV) and validation set are used .

IV EXPERIMENTAL RESULTS AND DISCUSSION:

4.1 Data Collection:

The below mentioned data sets are used for the design and development of regression and classification model. All the below mentioned data sets consist of dependent numeric variable **rain in mm** and other dependent **nominal variable events** among other predictor or independent variables like temperature, sea level pressure, humidity, wind speed , visibility and dew point. All these predictor variables are given as their numeric values maximum , average and minimum. In this data set some of the variables have nonlinear effects and other variables have linear or constant or zero effects on the response variable.

The dependent numeric variable rain in mm is used as response variable for the development of regression model keeping other variables as independent and nominal variable events is used for the development of classification model having remaining variables as independent.

The data set I consist of rainfall data of following four city.

Data set for Nagpur for the period 2008 to 2014.

Data set for Pune for the period 2008 to 2014.

Data set for Mumbai for the period 2008 to 2014.

Data set for Delhi for the period 2010 to 2014.

4.2 Statistical Tool used for the experiment:

For this research work, R programming language is used. We used 32 bit R programming language version 1.1.383 with the help of 32 bit IDE RStudio version 3.4.3 .

4.2 Development of Regression Model:

Following steps are applied in the development of regression model.

1. For the data set I, apply preprocessing steps to remove outliers, missing values and replace missing values by mean values.
2. Apply Normalization to scale the data in the range [-1,+1] .
3. Apply algorithms RT, MLR, SVM and ANN on data set of Nagpur, Pune, Mumbai and Delhi .
4. Compare the models using goodness of fit (in-sample) by using the MAE, MSE, RMSE, MASE and Correlation coefficient (COR).
5. Compare the models for predictive accuracy using cross validation approach (out-of-sample) .
6. Using above criteria, select the best model.
7. Apply Principal Component Analysis (PCA) on data set I and decide the principal components (PC) as the predictors.
8. Apply RT, MLR, SVM and ANN using these selected principal components and check the improvement in the results for goodness of fit and cross validation approach for predictive accuracy.
9. Select the best model which has shown the best improvement using PCA.
10. This model is our regression model.

The regression model is developed for data set I consisting rainfall data with multiple variables affecting the rainfall for Nagpur, Pune, Mumbai and Delhi city. Four algorithms Multiple Linear Regression (MLR), Support Vector Machine (SVM), Regression Tree(RT) and Artificial Neural Network(ANN) are applied on these data sets. The data set I is divided into 70\% training set and 30\% testing set (validation set). The performance of these algorithms is tested against the performance metrics as mentioned above and accordingly best suitable model is selected. The Table I, Table II , Table III and Table IV shows the goodness of fit and cross validation results obtained without applying PCA and after applying PCA on data set I.

From the Table I it is observed that SVM has shown the best in sample accuracy parameters as compared to RT,MLR and ANN before applying the PCA.But RMSE and MSE values of MLR has outperformed the SVM for Nagpur and Pune . The SVM has given the highest correlation coefficient of 0.66 for Mumbai and which higher than all other correlation coefficient of RT,MLR and ANN.

Table II shows the results of predictive skill for out of sample-accuracy through cross validation R-FOLD,LOOCV, R-KFOLD and V-SET approach. The cross validation approach has shown MLR outperforming other models for Nagpur and Pune except SVM has better MAE in both city. The SVM has also shown much better values than other model in Mumbai but RT shown great improvements in its values in Delhi compared to other models. The SVM and MLR can be considered to give better predictive skill for the Nagpur, Pune and Mumbai and RT has good cross validation results for out of sample accuracy in Delhi.

The Table III shows results of RT,MLR, SVM and ANN algorithms with performance metrics. From the table after comparing the values of MAE,RMSE,MASE,MSE and correlation coefficient with Table IV, it is observed that all algorithms has improved their performance considerably after applying PCA.

The Nagpur results has shown SVM performing the best model in all the assessing parameters of in-sample-accuracy. But in Pune, Mumbai, Delhi SVM has outperformed the others but there is no improvement in correlation coefficient .It is decreased with respect to without PCA. In Pune and Delhi, RT has improved its correlation coefficient among other parameters. In Mumbai, all the models has shown the improvement except correlation coefficient where as ANN has not been improved at all due to overfitting or underfitting.ANN has improved performance in Delhi but has negative correlation coefficient.

Thus among the four model, SVM model has shown much improvement as compared to others. Hence we selected SVM as the final model according to goodness-of-fit for in-sample data.

Table 1: Regression model with goodness of fit values without PCA

Data set	Algoj Metrics	MAE	RMSE	MSE	MASE	COR
Nagpur	RT	6.94	19.46	378.84	0.85	0.38
	MLR	8.01	18.7	349	0.99	0.45
	SVM	5.92	19.6	384.37	0.78	0.42
	ANN	6.55	19.05	363.15	0.80	0.45
Pune	RT	7.71	20.89	436.47	1.02	0.45
	MLR	9.8	20.20	408.29	1.3	0.49
	SVM	7.4	20.49	418.67	0.98	0.53
	ANN	8.39	23.88	570.37	1.11	0.37
Mumbai	RT	8.45	18.34	336.38	0.80	0.62
	MLR	10.10	19.12	365.79	0.95	0.58
	SVM	7.43	18.96	359.8	0.70	0.66
	ANN	9.62	22.25	495.42	0.91	0.52
Delhi	RT	3.74	11.24	126.46	1.28	0.014
	MLR	4.64	9.52	90.7	1.59	0.15
	SVM	2.62	9.25	85.61	0.89	0.19
	ANN	3.76	11.91	141.98	1.29	0.08

Table 2: Regression model with goodness of fit values with PCA

Data Set	Algo/Metrics	MAE	RMSE	MSE	MASE	COR
Nagpur	RT	5.34	11.55	133.6	0.98	0.56
	MLR	7.10	11.86	140.77	1.3	0.48
	SVM	4.16	10.56	111.6	0.76	0.55
	ANN	6.33	14.96	223.93	1.16	-0.082
Pune	RT	6.34	15.32	234.88	1.03	0.51
	MLR	9.1	15.03	226.15	1.48	0.48
	SVM	5.91	15.03	226.04	0.96	0.42
	ANN	10.52	22.62	511.67	1.71	-0.07

Mumbai	RT	9.11	16.79	282.0	1.32	0.29
	MLR	8.95	13.93	194.28	1.3	0.45
	SVM	5.92	13.14	172.74	0.86	0.47
	ANN	12.09	21.04	706.56	1.13	-0.019
Delhi	RT	3.52	9.6	92.21	1.37	0.299
	MLR	4.38	8.12	65.95	1.8	0.12
	SVM	2.47	7.95	63.28	0.94	0.16
	ANN	1.98	11.76	138.34	0.77	-0.01

Table 3: Regression model with cross validation values

Data set	Algorithm	CV Method/Metrics	RMSE	Rsquared	MAE
Nagpur	RT	K-FOLD	18.21	0.21	7.68
		LOOCV	20.87	0.096	7.98
		R-KFOLD	18.99	0.17	7.72
		V-SET	19.46	0.14	6.94
		K-FOLD	17.78	0.26	8.7
	MLR	LOOCV	19.01	0.17	8.74
		R-KFOLD	18.05	0.23	8.74
		V-SET	18.7	0.20	8.01
		K-FOLD	18.4	0.19	6.47
		LOOCV	20.14	0.13	6.46
	SVM	R-KFOLD	19.07	0.20	6.43
		V-SET	19.59	0.18	5.92
		K-FOLD	20.28	0.17	6.83
	ANN	LOOCV	21.62	0.003	6.83
		R-KFOLD	20.72	0.11	6.83
		V-SET	26.06	0.10	7.21
Pune	RT	K-FOLD	21.84	0.14	9.42
		LOOCV	20	0.24	8.69
		R-KFOLD	19.68	0.26	8.76
		V-SET	20.89	0.20	7.71

		MLR	K-FOLD	19.77	0.24	10.75
			LOOCV	20.58	0.21	10.76
			R-KFOLD	19.36	0.26	10.82
			V-SET	20.20	0.24	9.8
		SVM	K-FOLD	21.44	0.25	7.43
			LOOCV	22.59	0.21	7.43
			R-KFOLD	20.98	0.27	7.43
			V-SET	20.35	0.29	7.38
		ANN	K-FOLD	22.99	0.16	7.17
			LOOCV	24.03	0.24	7.68
			R-KFOLD	22.53	0.12	7.53
			V-SET	21.18	0.22	7.69
	Mumbai	RT	K-FOLD	19.4	0.38	9.24
			LOOCV	19.86	0.29	9.94
			RKFOLD	18.79	0.37	8.89
			V-SET	18.2	0.39	8.25
		MLR	K-FOLD	18.43	0.37	9.77
			LOOCV	19.32	0.32	9.73
			R-KFOLD	18.68	0.36	9.79
			V-SET	18.95	0.35	9.96
		SVM	K-FOLD	18.86	0.36	7.83
			LOOCV	19.56	0.31	7.85
			R-KFOLD	19.04	0.37	7.9
			V-SET	18.71	0.46	7.2
		ANN	K-FOLD	23.52	0.06	8.50
			LOOCV	24.48	0.0004	8.5
			R-KFOLD	23.31	0.07	8.51
			V-SET	63.76	0.02	30.67
	Delhi	RT	K-FOLD	9.78	0.23	2.97
			LOOCV	12.89	0.64	3.18
			R-KFOLD	11.29	0.25	3.20
			V-SET	14.16	0.014	3.91

	MLR	K-FOLD	11.37	0.076	5.15
		LOOCV	13.7	0.037	5.12
		R-KFOLD	11.61	0.07	5.17
		V-SET	13.5	0.048	4.52
	SVM	K-FOLD	11.08	0.035	3.36
		LOOCV	13.88	0.79	3.37
		R-KFOLD	10.09	0.021	3.36
		V-SET	13.86	0.004	3.11
	ANN	K-FOLD	9.34	0.051	2.45
		LOOCV	13.89	0.06	3.05
		R-KFOLD	10.94	0.027	3.05
		V-SET	29.59	0.002	6.41

Table 4: Predictive Accuracy of data set I after PCA

Data set	Algorithm	CV	RMSE	R2	MAE
Nagpur	RT	K-FOLD	0.04	0.26	0.021
		LOOCV	0.53	0.11	0.023
		R-KFOLD	0.051	0.17	0.023
		V-SET	13.08	0.31	5.89
	MLR	K-FOLD	0.048	0.25	0.024
		LOOCV	0.049	0.20	0.024
		R-KOLD	0.048	0.24	0.024
		V-SET	11.65	0.23	7.10
	SVM	K-FOLD	0.051	0.21	0.018
		LOOCV	0.054	0.07	0.01
		RKFOLD	0.052	0.22	0.018
		V-SET	0.30	0.30	3.577
	ANN	K-FOLD	0.050	0.29	0.034
		LOOCV	0.050	0.23	0.029
		R-KFOLD	0.049	0.28	0.029
		V-SET	16.74	0.004	5.79
Pune	RT	K-FOLD	0.047	0.244	0.021
		LOOCV	0.051	0.15	0.023
		R-KFOLD	0.048	0.21	0.022

		V-SET	15.48	0.26	5.96
	MLR	K-FOLD	0.48	0.23	0.025
		LOOCV	0.049	0.21	0.025
		R-KFOLD	0.047	0.22	0.026
		V-SET	13.40	0.23	9.09
	SVM	K-FOLD	0.053	0.22	0.017
		LOOCV	0.053	0.22	0.017
		R-KFOLD	0.051	0.22	0.017
		V-SET	0.17	0.17	4.09
	ANN	K-FOLD	0.055	0.21	0.0197
		LOOCV	0.049	0.23	0.030
		R-KFOLD	0.048	0.25	0.0312
		V-SET	34.11	0.009	10.56
Mumbai	RT	K-FOLD	0.053	0.13	0.026
		LOOCV	0.055	0.09	0.028
		R-KFOLD	0.052	0.17	0.027
		V-SET	15.36	0.08	8.09
	MLR	K-FOLD	0.050	0.22	0.026
		LOOCV	0.050	0.21	0.025
		R-KFOLD	0.050	0.24	0.025
		V-SET	13.23	0.20	8.80
	SVM	K-FOLD	0.051	0.29	0.020
		LOOCV	0.052	0.18	0.020
		R-KFOLD	0.051	0.25	0.020
		V-SET	10.29	0.22	4.7
	ANN	K-FOLD	0.052	0.21	0.034
		LOOCV	0.052	0.19	0.034
		R-KFOLD	0.052	0.22	0.033
		V-SET	25.18	0.002	13.87
Delhi	RT	K-FOLD	0.032	0.024	0.0131
		LOOCV	0.040	0.005	0.015
		R-KFOLD	0.040	0.005	0.0153

		V-SET	8.511	0.0039	2.89
	MLR	K-FOLD	0.033	0.032	0.0134
		LOOCV	0.038	0.003	0.0134
		R-KFOLD	0.034	0.028	0.0135
		V-SET	4.51	0.0157	3.6
	SVM	K-FOLD	0.032	0.0144	0.010
		LOOCV	0.038	0.85	0.010
		R-KFOLD	0.033	0.043	0.010
		V-SET	1.31	0.0269	1.17
	ANN	K-FOLD	0.040	0.029	0.029
		LOOCV	0.038	1.36	0.008
		R-KFOLD	0.034	0.021	0.0085
		V-SET	32.82	2.18	6.35

4.3 Development of Classification Model:

The classification model is developed for the nominal variable events as dependent variable and other variables as independent or predictor variables. The dependent variable events has two classes as NoRain for the day no rainfall is there and other class is Rain for the day rainfall has occurred .

For the development of Classification model, the algorithms KNN, RF, DT[C5.0] and NB are used. The following steps have been performed on data set I with nominal attribute as dependent feature and temperature, wind speed, sea level pressure, humidity, visibility and dew point as independent features.

1. For the data set I, apply preprocessing steps to remove outliers, missing values and replace missing values by mean values.
2. Apply Normalization to scale the data in the range [-1,+1].
3. Then data set I is applied with all the above four classifier.
4. The best classification model is selected using the confusion matrix and accuracy rate criteria.
5. Then data set I is applied with PCA followed by these four classifiers.
6. The model that has shown the best improvement than those without PCA has been selected as the best classification model using the confusion matrix and accuracy rate criteria.

The following Table 5 shows the classification model with confusion matrix along with accuracy of the classifier before applying the PCA. In the confusion matrix horizontal row indicates the class NoRain (Predicted values) and Class Rain(Predicted values). The vertical column of confusion matrix indicates class NoRain (Actual Values) and Rain(Actual Values).

From the Table 5 it is clear that the classifier DT[C5.0] and NB both has shown the best performance as compared to the accuracy of KNN and RF that ranges from 72\% for Nagpur, 70\% for Pune, 71\% for Mumbai, and highest 90\% in Delhi. But in Delhi ,all the classifiers KNN,RF, DT and NB outperformed in accuracy as compared with Nagpur, Pune and Delhi models . Since the Delhi data set contains the maximum number NoRain class as compared to Rain class that helps the classifier to give the maximum accuracy. The DT[C5.0] has shown the highest number of True positive values of NoRain in confusion matrix indicating the highest number of no rainfall event in data set. So DT and NB are selected as the best classifiers before applying the PCA.

From the Table 6 it is evident that all the classifiers has shown improvement in the accuracy prediction. The RF has shown maximum improvement in the Mumbai data set. The DT and NB has outperformed RF and KNN in all four data set with improvement in the prediction accuracy. Hence DT and NB are selected as the best classifiers for the classification model.

Table 5: Classification Model without PCA

Data Set	Classifier	Accuracy	Confusion Matrix	NoRain	Rain
Nagpur	KNN	0.64	NoRain	459	105
			Rain	171	41
	RF	0.65	NoRain	473	91
			Rain	176	36
	DT[C5.0]	0.72	NoRain	564	0
			Rain	212	0

	NB	0.72	NoRain	540	24
			Rain	190	22
Pune	KNN	0.64	NoRain	462	81
			Rain	195	29
	RF	0.66	NoRain	465	78
			Rain	177	47
	DT[C5.0]	0.70	NoRain	543	0
			Rain	224	0
	NB	0.70	NoRain	522	21
			Rain	203	21
Mumbai	KNN	0.64	NoRain	436	107
			Rain	168	53
	RF	0.66	NoRain	460	83
			Rain	177	44
	DT[C5.0]	0.71	NoRain	543	0
			Rain	221	0
	NB	0.71	NoRain	537	4
			Rain	218	3
Delhi	KNN	0.89	NoRain	495	9
			Rain	52	1
	RF	0.90	NoRain	503	1
			Rain	53	0
	DT[C5.0]	0.90	NoRain	504	0
			Rain	53	0
	NB	0.90	NoRain	504	0
			Rain	53	0

Table 6: Classification model with PCA.

Dataset	Classifier	Accuracy	Confusion Matrix	NoRain	Rain
Nagpur	KNN	0.65	NoRain	470	94
			Rain	178	34

	RF	0.68	NoRain	483	81
			Rain	177	35
	DT[C5.0]	0.73	NoRain	565	0
			Rain	211	0
	NB	0.73	NoRain	565	2
			Rain	212	3
Pune	KNN	0.65	NoRain	463	80
			Rain	197	27
	RF	0.68	NoRain	464	79
			Rain	177	47
	DT[C5.0]	0.71	NoRain	544	0
			Rain	223	0
	NB	0.71	NoRain	524	19
			Rain	108	16
Mumbai	KNN	0.66	Norain	442	101
			Rain	170	51
	RF	0.70	NoRain	501	42
			Rain	187	0
	DT[C5.0]	0.72	NoRain	544	0
			Rain	220	0
	NB	0.71	NoRain	539	4
			Rain	216	5
Delhi	KNN	0.90	NoRain	499	5
			Rain	51	1
	RF	0.91	NoRain	505	0
			Rain	52	0
	DT[C5.0]	0.91	NoRain	505	0
			Rain	52	0
	NB	0.91	NoRain	505	0
			Rain	52	0

V. CONCLUSION:

The research work applied 08 different algorithms for the design and development of data mining tools for the regression and classification task on 4 different data sets of Nagpur, Pune, Mumbai and Delhi. This work has found that SVM model performed reasonably well outperforming other models in terms of predictive skill while providing the reasonably better goodness of fit. The DT and NB model has shown best results for the classification task and both are selected as classification models in terms of accuracy and confusion matrix.

The multicollinearity among the predictors is eliminated using PCA and all the regression and classification models has shown much improvement in the results. Also it is found that the machine learning algorithms such as SVM and ANN to rainfall prediction should be approached cautiously because these algorithms are complex and viewed as black boxes that can obscure interpretation of the results.

It is also critical to distinguish between measures of predictive skill RMSE, MAE and R2 and measures of goodness of fit MAE,MSE,RMSE,MASE and correlation coefficient while developing the rainfall prediction models since these values varies from the goodness of fit to predictive skill for the same model using same data set. Here SVM model provided better benefits in terms of both predictive skill and goodness of fit.

The classification models DT and NB proved their accuracy in the prediction of classes upto 91% and increased with application of PCA as well.

REFERENCES

[1] Ramsundram N. Sathya S and Karthikeyan S(2016).Comparison of decision tree based rainfall prediction model with data driven model considering climatic variables. Irrigation Drainage Systems Engineering, 5(3).

[2] Toth E. and Brath A 2000. Comparison of short-term rainfall prediction models for real-time flood forecasting. Journal of Hydrology" (239) :132–147.

[3] Agboola A. Gabriel A. J., Aliyu E.O. and Alese B.K. 2013. Development of a fuzzy logic based rainfall prediction model. International Journal of Engineering and Technology, 3(4)

[4] Wu C,L.,and Chau K.W.2013.Prediction of rainfall time series using modular soft computing methods.Engineering Applications of Artificial Intelligence, (26) :997–1007.

[5] Pinky D.2014.Prediction of rainfall using data mining technique over Assam, Indian Journal of Computer Science and Engineering ,5(2):85-90.

[6] Poorani K and Brindha K,2013. Data Mining Based on Principal Component Analysis for Rainfall Forecasting in India,International Journal of Advanced Research in Computer Science and Software Engineering" 3(9).

[7] Emily Mankin .Principal Components Analysis: A How-To Manual for R.