Journal of Engineering Sciences
Vol 11, Issue 6,June/2020
ISSN NO:0377-9254

SMART WEATHER FORECASTING USING MACHINE LEARNING

A.B.S. Aravind[1], D. Lavanya[2] , B.T.S.S Nagarjuna[3] ,Ch. Vinay[4], G.V.S.N.R.V. Prasad[5]

[1] *Student, Department of Computer Science & Engineering, Gudlavalleru Engineering College, India*
[2] *Student, Department of Computer Science & Engineering, Gudlavalleru Engineering College, India*
[3] *Student, Department of Computer Science & Engineering, Gudlavalleru Engineering College, India*
[4] *Student, Department of Computer Science & Engineering, Gudlavalleru Engineering College, India*
[5]*Professor, Department of Computer Science & Engineering, Gudlavalleru Engineering College, India*
aravind.akula1223@gmail.com, dvlavanya2015@gmail.com, chakka.vinay2427@gmail.com, sainagarjuna184@gmail.com,guttaprasad1@gmail.com

ABSTRACT-Traditionally, weather predictions are performed with the help of large complex models of physics, which utilize different atmospheric conditions over a long period of time. These conditions are often unstable because of perturbations of the weather system, causing the models to provide inaccurate forecasts. In this paper, we present a weather prediction technique that utilizes historical data from multiple weather stations to train simple machine learning models, which can provide usable forecasts about certain weather conditions for the near future within a very short period of time. This paper proposes a simulated system which was developed to predict various weather conditions across Indian subcontinent using Data Analysis and Machine learning techniques such as regression and classification. The main source of data used for supervised learning is collected from data.gov.in, ncdc.noaa.gov and UCI machine learning data repository. The existing weather condition parameters ex. rainfall etc are used to fit a model and further using machine learning techniques and extrapolating the information, the future variations in the parameters are analysed.

Keywords: weather forecast, prediction, regression, classification, machine learning.

1.INTRODUCTION

Weather conditions around the world change rapidly and continuously. Correct forecasts are essential in today's daily life. From agriculture to industry, from traveling to daily commuting, we are dependent on weather forecasts heavily. As the entire world is suffering from the continuous climate change and its side effects, it is very important to predict the weather conditions without any error to ensure easy and seamless mobility, as well as safe day to day operations. The current weather prediction models heavily depend on complex physical models and need to be run on large computer systems involving hundreds of HPC nodes. The computational power of these large systems is required to solve the models that describe the atmosphere. Despite using these costly and complex devices, there are often inaccurate forecasts because of incorrect initial measurements of the conditions or an incomplete understanding of atmospheric processes. Moreover, it generally takes a long time to solve complex models like these.

Inspite of using those complex techniques, Machine learning comes with a better solution. As weather systems can travel a long way over time in all directions, the weather of one place depends on that of others considerably. In this work, we propose a method to utilize surrounding city's historical weather data along with a particular city's data to predict its weather condition. We combine these data and use it to train simple machine learning models, which in turn, can predict correct weather conditions for the next few days. These simple models can be run on low cost and less resource-intensive computing systems, yet can provide quick and accurate enough forecasts to be used in our day-to-day life.

2. PROBLEM STATEMENT

Climate is an important aspect of human life. So, the Prediction should accurate as much as possible. In this paper we try to deal with the prediction of the rainfall which is also a major aspect of human life and which provide the major resource of human life which is Fresh Water. Fresh water is always a crucial resource of human survival – not only for the drinking purposes but also for farming and many other purposes. Making a good prediction of climate is always a major task now a day because of the climate change. An incorrect rainfall prediction can affect the agriculture mostly framers as their whole crop is depend on the rainfall and agriculture is always an important part of every economy. So, making an accurate prediction of the rainfall is necessary. There are number of techniques available and used of machine learning but accuracy is always a matter of concern in prediction made in rainfall. There are number of causes made by rainfall affecting the world ex. Drought, Flood and intense summer heat etc. And it will also affect water resources around the world. Based on past data we predict average rainfall of next year, then farmers may choose which crop will benefit them according to that average rainfall. We also predict tomorow's rainfall based on some factors.(Did it rain the next day? Yes or No.)

3. REVIEW OF LITERATURE

There are different techniques used for the prediction of rainfall such as Regression analysis, clustering and Artificial Neural Networks (ANN). Fundamentally, two approaches are used for predicting rainfall. One is Empirical approach and the other is Dynamical approach. The empirical approach is based on analysis of historical data of the rainfall and its relationship to a variety of atmospheric and oceanic variables over different parts of the world. The most widely used empirical approaches, which are used for climate prediction, are regression, artificial neural network, fuzzy logic and group method of data handling. On the other hand in dynamical approach, predictions are generated by physical models based on systems of equations that predict the evolution of the global climate system in response to initial atmospheric conditions [1]. The different rainfall estimation models were developed by Ozlem Terzi [2] by using the monthly rainfall data of Isparta, Senirkent, Uluborlu, Egirdir, and Yalvac stations of Turki . Rainfall estimation models were built using Decision Table, KStar, Multilinear Regression, M5'Rules, Multilayer Perceptron, RBF Network, Random Subspace, and Simple Linear Regression algorithms and quality of these models were tested using chosen coefficient of determination (R^2) and root mean-squared error (RMSE) which are the most well known and the commonly used performance criteria. Using different combinations of Input given to above developed Models, he has generated the MLR model that gives the best results to estimate rainfall over Isparta region. J.M. Spate et al [3] has prepared a model to measure a stream flow from the measured and estimated /interpolated rainfall. K-medoid algorithm on clustering has been discussed to clustering shapes/peaks. The paper has discussed the various classification and association rule extraction methods. Instead, they have selected all those catchments in their region of interest where high-intensity rainfall data does exist for at least some temporal interval. Then they applied some simple criteria to the high-intensity data; for example so much rain must fall in such a small time interval on a given day for that fall to be flagged as an intense event. Having generated a Boolean series with 1's on every day with an intense event and 0's elsewhere, they used data mining to automatically extract those combinations of daily data characteristics which tend to occur on a day with 1 in the Boolean series.

Pratap Singh Solanki et al [4] reviewed the studies related to use of data mining techniques in the field of water resource sector for Water Management. Presently, the Water Resource Management has become most challenging, interesting and fascinating domain around the world since many years. Scientist tries to predict the Rainfall, Flood Warning, Water Inflow, Water Availability and Requirements etc. based on huge available metadata using various methods. In this article, they tried to search the use of data mining techniques for predicting the inflow, drought possibility, weather report, rainfall, evaporation, temperature, wind speed etc. This paper provides the survey of some literature and work done by the researchers using various algorithms and modeling method viz. Associations rules, Classification, Clustering, Decision Tree, and

Artificial Neural Network etc.

Pinky Saikia Dutta [5] in her Project, Rainfall prediction is implemented with the use of empirical statistical technique. She used 6 years (2007-2012) datasets 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). Regression is a statistical empirical technique that utilizes the relation between two or more quantitative variables on observational database so that outcome variable can be predicted from the others. One of the purposes of a regression model is to find out to what extent the outcome (dependent variable) can be predicted by the independent variables. Predictors selected for the model are minimum temperature, maximum temperature, mean sea level pressure, wind speed and rainfall.

Jyothis Joseph [6] described empirical method technique belonging to clustering and classification approach. ANNs are used to implement these techniques. He used Relative Humidity, Pressure, Temperature, Precipitable Water, Wind Speed. In this paper subtractive clustering is used. Subtractive clustering is a fast, onepass algorithm for estimating the number of clusters and the cluster centers in a set of data. Applying subtractive clustering, the optimum numbers of clusters are obtained .The rainfall values are categorized as low, medium & heavy. The classifier model has been evaluated against a confusion matrix and the results have been obtained. This paper applies neural network for rainfall prediction. In this paper two methods such as classification and clustering are implemented. The neural network Bayesian regularization has been applied in the implementation.

4. PROPOSED METHOD

METHODOLOGY:

The Basic process evolved in the entire program is represented with a simple flowchart:

Fig: Architecture

In this work we used two algorithms one is regression and another one is classification for two datasets. Regression is used to predict average rainfall and classification is used for next day's rainfall.

Step 1: Importing Libraries

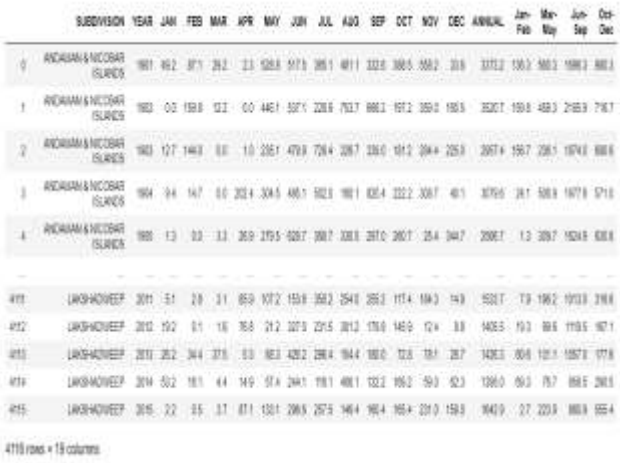
We used pandas, numpy and Multilinear Regression classifier for prediction. Pandas will be used for performing operations on data frames. Further, more using numpy, we will perform necessary mathematical operations.

```
In [1]:
import pandas as pd
import numpy as np
import math
import sklearn
from sklearn.model_selection import train_test_split
from sklearn.metrics import mean_squared_error,r2_score,mean_absolute_error
import matplotlib.pyplot as plt
import seaborn as sb
from sklearn import preprocessing
```

Step 2: Reading the dataset

To pick the right variables, you have got to have a basic understanding of your dataset, enough to know that your data is relevant, high quality, and of adequate volume. As part of your model building

efforts, you will be working to select the best predictor variables for your model.



	SUBDIVISION	YEAR	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP	OCT	NOV	DEC	ANNUAL	Jan-Feb	Mar-May	Jun-Sep	Oct-Dec
0	ANDAMAN & NICOBAR ISLANDS	1901	49.2	87.1	29.2	2.3	528.8	517.5	365.1	481.1	332.6	388.5	558.2	33.6	3373.2	136.3	560.3	1696.3	980.3
1	ANDAMAN & NICOBAR ISLANDS	1902	0.0	159.8	12.2	0.0	446.1	537.1	228.9	753.7	666.2	197.2	359.8	160.5	3520.7	159.8	458.3	2185.9	716.7
2	ANDAMAN & NICOBAR ISLANDS	1903	12.7	144.0	0.0	1.0	235.1	479.9	728.4	326.7	339.0	181.2	284.4	225.0	2957.4	156.7	236.1	1974.0	690.6
3	ANDAMAN & NICOBAR ISLANDS	1904	9.4	14.7	0.0	202.4	304.5	495.1	502.0	160.1	820.4	222.2	308.7	40.1	3079.6	24.1	506.9	1977.8	571.0
4	ANDAMAN & NICOBAR ISLANDS	1905	1.3	0.0	3.3	26.9	279.5	628.7	368.7	330.5	297.0	260.7	25.4	344.7	2566.7	1.3	309.7	1624.9	630.8
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
4111	LAKSHADWEEP	2011	5.1	2.9	3.1	65.9	107.2	153.8	358.2	254.0	235.3	117.4	184.3	14.9	1532.7	7.9	196.2	1013.9	314.6
4112	LAKSHADWEEP	2012	10.2	0.1	1.6	16.6	21.2	327.9	231.5	261.2	170.6	145.9	12.4	8.8	1408.5	10.3	39.4	1119.5	167.1
4113	LAKSHADWEEP	2013	35.2	34.4	37.5	0.0	98.3	429.2	286.4	164.4	180.0	72.8	78.1	28.7	1438.5	69.6	137.1	1057.8	177.8
4114	LAKSHADWEEP	2014	53.2	18.1	4.4	14.9	57.4	244.1	118.1	468.1	132.2	169.2	59.0	62.3	1396.0	69.3	75.7	868.5	382.5
4115	LAKSHADWEEP	2015	2.2	0.5	3.7	87.1	130.1	286.5	257.5	146.4	160.4	165.4	231.0	159.8	1640.9	2.7	220.9	880.8	535.4

4116 rows × 19 columns

Fig1: dataset for regression

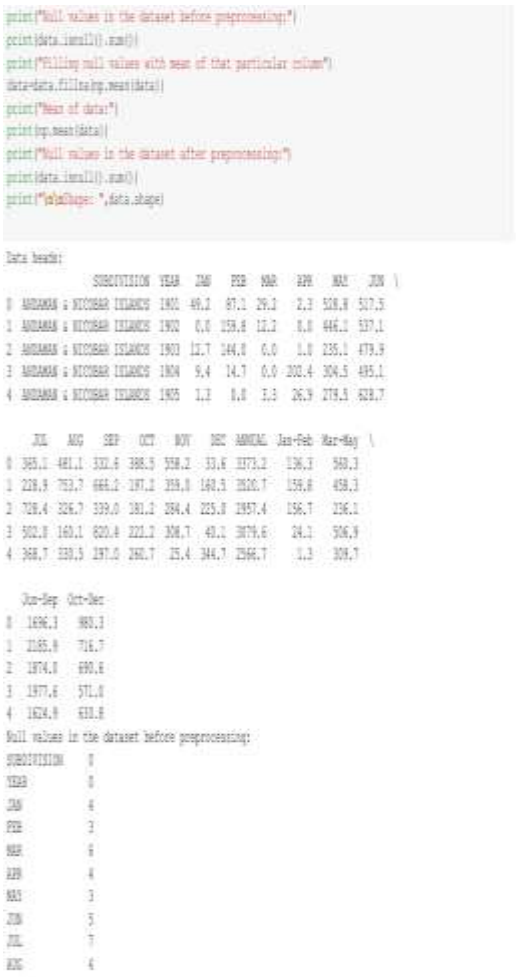
	MinTemp	MaxTemp	Rainfall	Evaporation	Sunshine	WindGustSpeed	WindSpeed9am	WindSpeed3pm	Humidity9am	Humidity3pm	Pressure9am
0	13.4	22.9	0.6	0.3	0.0	44	20	24	71	22	1007.7
1	7.4	25.1	0.0	0.3	0.0	44	4	22	44	25	1010.8
2	12.9	25.7	0.0	0.3	0.0	46	19	26	38	30	1007.6
3	9.2	28.0	0.0	0.3	0.0	24	11	9	45	16	1017.6
4	17.5	32.3	1.0	0.3	0.0	41	7	20	82	33	1010.8
...	...	...	...	...	...	...	...	...	...	...	...
142186	3.5	21.8	0.0	0.3	0.0	31	15	13	59	27	1024.7
142187	2.8	23.4	0.0	0.3	0.0	31	13	11	51	24	1024.6
142188	3.6	25.3	0.0	0.3	0.0	35	13	9	56	24	1023.5
142189	5.4	26.9	0.0	0.3	0.0	37	9	9	53	24	1021.0
142192	7.8	27.0	0.0	0.3	0.0	28	13	7	51	24	1019.4

142193 rows × 15 columns

Fig2: datasetfor classification

Step3:Checking for null values:

Sometimes you may find some data are missing in the dataset. We need to be equipped to handle the problem when we come across them. Obviously you could remove the entire line of data but what if you are unknowingly removing crucial information? Of course we would not want to do that. One of the most common idea to handle the problem is to take a mean of all the values of the same column and have it to replace the missing data.



```
print("Null values in the dataset before preprocessing:")
print(data.isnull().sum())
print("Filling null values with mean of that particular column")
data=data.fillna(np.mean(data))
print("Mean of data:")
print(np.mean(data))
print("Null values in the dataset after preprocessing:")
print(data.isnull().sum())
print("shape: ",data.shape)
```

```
Data heads:
       SUBDIVISION  YEAR   JAN   FEB   MAR   APR   MAY   JUN \
0  ANDAMAN & NICOBAR ISLANDS  1901  49.2  87.1  29.2   2.3 528.8 517.5
1  ANDAMAN & NICOBAR ISLANDS  1902   0.0 159.8  12.2   0.0 446.1 537.1
2  ANDAMAN & NICOBAR ISLANDS  1903  12.7 144.0   0.0   1.0 235.1 479.9
3  ANDAMAN & NICOBAR ISLANDS  1904   9.4  14.7   0.0 202.4 304.5 495.1
4  ANDAMAN & NICOBAR ISLANDS  1905   1.3   0.0   3.3  26.9 279.5 628.7

    JUL   AUG   SEP   OCT   NOV   DEC  ANNUAL  Jan-Feb  Mar-May \
0 365.1 481.1 332.6 388.5 558.2  33.6  3373.2    136.3    560.3
1 228.9 753.7 666.2 197.2 359.8 160.5  3520.7    159.8    458.3
2 728.4 326.7 339.0 181.2 284.4 225.0  2957.4    156.7    236.1
3 502.0 160.1 820.4 222.2 308.7  40.1  3079.6     24.1    506.9
4 368.7 330.5 297.0 260.7  25.4 344.7  2566.7      1.3    309.7

   Jun-Sep  Oct-Dec
0   1696.3    980.3
1   2185.9    716.7
2   1974.0    690.6
3   1977.6    571.0
4   1624.9    630.8
Null values in the dataset before preprocessing:
SUBDIVISION     0
YEAR            0
JAN             4
FEB             3
MAR             0
APR             4
MAY             3
JUN             5
JUL             7
AUG             4
```

First dataset null values

Feature	% of Null values
Sunshine	43%
Evaporation	48%
Cloud3pm	40%
Cloud9am	38%

Null values for second data set.

Step4: Data Preprocessing

Data preprocessing is a data mining technique that involves transforming raw data into an understandable format. Real-world data is often incomplete, inconsistent, and/or lacking in certain behaviors or trends, and is likely to contain many errors. We have carried below preprocessing steps.

Missing Values:

we learned that we have few instances with null values. Hence, this becomes one of the important

step. To impute the missing values, we will group our instances thereby replace the null values by their respective mean values.

Step5:Finding correlation coefficent for all column pairs

We can already see some potentially interesting relationships between the target variable (the number of fatal accidents) and the feature variables (the remaining three columns).

To quantify the pairwise relationships that we observed in the scatter plots, we can compute the Pearson correlation coefficient matrix. The Pearson correlation coefficient is one of the most common methods to quantify correlation between variables, and by convention, the following thresholds are usually used:

- 0.2 = weak
- 0.5 = medium
- 0.8 = strong
- 0.9 = very strong

Step6: Training and Test Sets: Splitting Data

The module introduced the idea of dividing your data set into two subsets:

training set—a subset to train a model.

Test set—a subset to test the trained model.

You could imagine slicing the single data set as Slicing a single data set into a training set and test set.

Make sure that your test set meets the following two conditions:

Is large enough to yield statistically meaningful results.

Is representative of the data set as a whole. In other words, don't pick a test set with different characteristics than the training set.

Assuming that your test set meets the preceding two conditions, your goal is to create a model that generalizes well to new data. Our test set serves as a proxy for new data. For example, consider the following figure. Notice that the model learned for the training data is very simple. This model doesn't do a perfect job—a few predictions are wrong. However, this model does about as well on the test data as it does on the training data. In other words, this simple model does not overfit the training data.

Validating the trained model against test data. **Never train on test data.** If you are seeing surprisingly good results on your evaluation metrics, it might be a sign that you are accidentally training on the test set. For example, high accuracy might indicate that test data has leaked into the training set.

For example, consider a model that predicts whether an email is spam, using the subject line, email body, and sender's email address as features. We apportion the data into training and test sets, with an 80-20 split. After training, the model achieves 99% precision on both the training set and the test set. We'd expect a lower precision on the test set, so we take another look at the data and discover that many of the examples in the test set are duplicates of examples in the training set (we neglected to scrub duplicate entries for the same spam email from our input database before splitting the data). We've inadvertently trained on some of our test data, and as a result, we're no longer accurately measuring how well our model generalizes to new data.

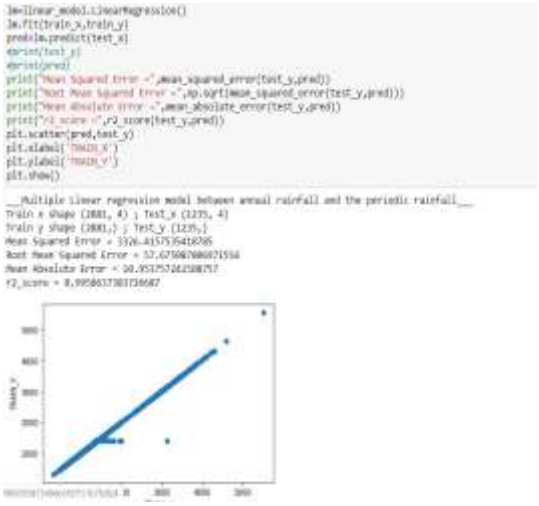
```
X = df.iloc[:, :-1].values

y = df.iloc[:, 8].values
```

Scikit-learn provides a range of supervised and unsupervised learning algorithms via a consistent interface in Python. It is licensed under a permissive simplified BSD license and is distributed under many Linux distributions, encouraging academic and commercial use. The library is built upon the SciPy (Scientific Python) that must be installed before you can use scikit-learn.

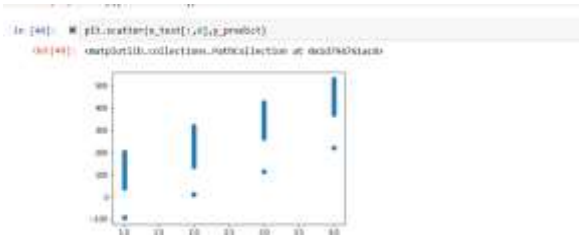
```
From sklearn.cross_validation import train_test_split

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.2, random_state = 0)
```

The algorithm can be used for predicting an output vector y given an input matrix X. In the first step a tree ensemble is generated with gradient boosting. The trees are then used to form rules, where the paths to each node in each tree form one rule. A rule is a binary decision if an observation is in a given node, which is dependent on the input features that were used in the splits. The ensemble of rules together with the original input features are then being input in a L1-regularized linear model, also called Lasso, which estimates the effects of each rule on the output target but at the same time estimating many of those effects to zero.

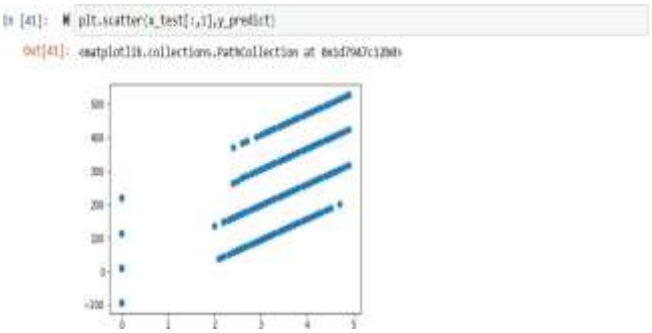
You can use rule fit for predicting a numeric response (categorial not yet implemented). The input has to be a numpy matrix with only numeric values.



Step7:Visualization

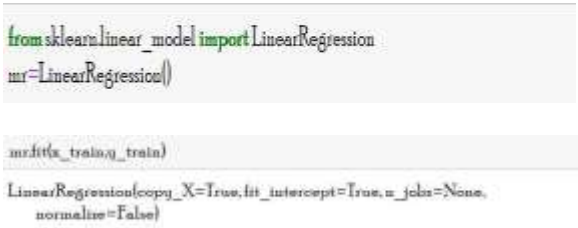


Step 8: Model Fitting



from sklearn.linear_model import LinearRegression

LinearRegression fits a linear model with coefficients w = (w1, …, wp) to minimize the residual sum of squares between the observed targets in the dataset, and the targets predicted by the linear approximation.



Step9:Predicting Results

Confusion Matrix

It gives us a matrix as output and describes the complete performance of the model. It focuses on True Positives - the cases in which we predicted YES and the actual output was also YES; True Negatives - the cases in which we predicted NO and the actual output was NO; False Positives - the cases in which we predicted YES and the actual output was NO; False Negatives - the cases in which we predicted NO and the actual output was YES.

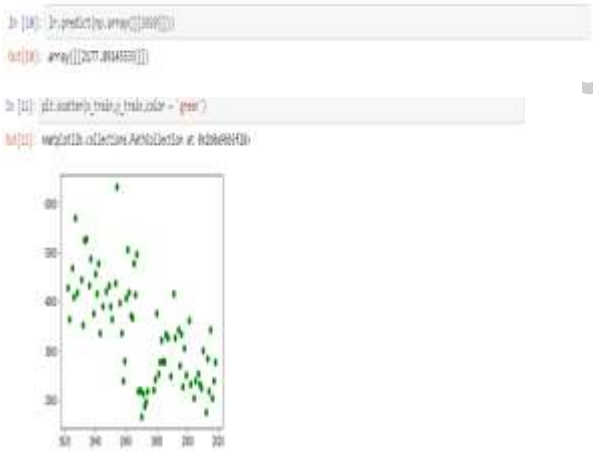
6. RESULTS AND DISCUSSION

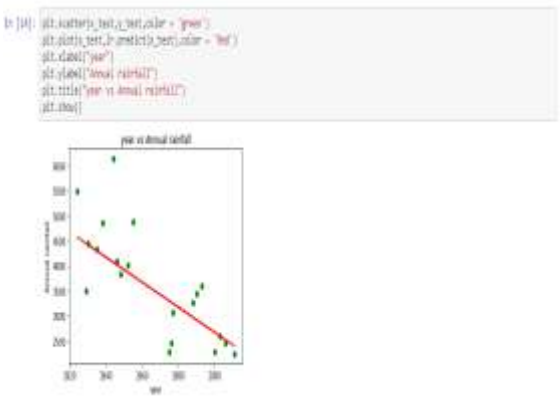
Experimental Analysis:

In this work we used two algorithms, multiple linear regression for predicting annual rainfall and classification for next day's rainfall prediction. Here we will have comparison between them.

Using multilinear regression:

In the dataset the attributes are subdivisions, years and months from January to December. Here we have taken months as x-attributes and predicted the annual rainfall(y-attribute).





Using Classification:

By using classification we predict next day's rainfall by taking some factors like temperature, humidity sunshine, evaporation. It gives output as "1", if there is chance of rainfall else it gives "0".

With the issues with our original dataset, we learned many things considering all the preprocessing steps that we carried to rectify them. The first important thing we learned is the importance of knowing your data. While imputing the missing value, we grouped two other features and calculated the mean instead of directly calculating the mean for all the instances. This way our imputed values were closer to the correct information. Another thing we learned is about the leaky features. While exploring our data, we came to know that one of our feature (RiskMM) was used for generating the target variable and hence it made no sense to use this feature for predictions.

7. CONCLUSION

In this work, we presented a technology to utilize machine learning techniques to provide weather forecasts. Machine learning technology can provide

intelligent models, which are much simpler than traditional physical models. They are less resource-hungry and can easily be run on almost any computer including mobile devices. Our evaluation results show that these machine learning models can predict weather features accurately enough to compete with traditional models.

A look into multilinear regression which includes the investigation of the subject, information mining methods, information mining forms, information mining calculations and its usage bitterly to make it more intelligent to the clients. Presenting the crude information subsequent to preparing and executing the information mining procedures in intuitive way to the clients for better understanding.

We explored and applied several preprocessing steps and learned there impact on the overall performance of our classifiers. We also carried a comparative study of all the classifiers with different input data and observed how the input data can affect the model predictions

8.REFERENCES

[1] Nikhil Sethi, Dr.Kanwal Garg," Exploiting Data Mining Technique for Rainfall Prediction",(IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (3) , 2014, 3982-3984.

[2] Ozlem Terzi,"Monthly Rainfall Estimation Using Data-Mining Process", Hindawi Publishing Corporation Applied Computational Intelligence and Soft Computing ,Volume 2012,Article ID 698071, 6 pages doi:10.1155/2012/698071.

[3] J.M. Spate, B.F.W. Croke, A.J. Jakeman, "Data Mining in Hydrology" ,Department of Mathematics, The Australian National University, Canberra ACT 0200, Australia .

[4] Pratap Singh Solanki , R. S. Thakur '"A Review of Literature on Water Resource Management Using Data Mining Techniques",International Journal of Science and Research (IJSR) ISSN (Online): 2319-7064 Index Copernicus Value (2013): 6.14 | Impact Factor (2015): 6.391 Volume 5 Issue 7, July 2016 www.ijsr.net Licensed Under Creative Commons Attribution CC BY

[5] Pinky Saikia Dutta et.al,"Prediction Of Rainfall Using Datamining Technique Over Assam", Indian Journal of Computer Science and Engineering (IJCSE).

[6] Jyothis Joseph, Ratheesh T K ," Rainfall Prediction using Data Mining Techniques ", International Journal of Computer Applications (0975 – 8887) Volume 83 – No 8, December 2013- 11 .

[7] K Poorani ,K Brindha ," Data Mining Based on Principal Component Analysis for Rainfall Forecasting in India ",International Journal of Advanced Research in Computer Science and Software Engineering" Volume 3, Issue 9, September 2013 ISSN: 2277 128X , Research Paper Available online at: www.ijarcsse.com

[8] E. Sreehari, J. Velmurugan, Dr. M. Venkatesan," A Survey Paper on Climate Changes Prediction Using Data mining ",International Journal of Advanced Research in Computer and Communication Engineering Vol. 5, Issue 2, February 2016 Copyright to IJARCCE DOI 10.17148/IJARCCE.2016.5261 294 .

[9] Narasimha Prasad LV, Naidu MM ," An Efficient Decision Tree Classifier to Predict Precipitation Using Gain Ratio", The International Journal of Soft Computing and Software Engineering [JSCSE], Vol. 3, No. 3, Special Issue: The Proceeding of International Conference on Soft Computing and Software Engineering 2013.

[10] A Geetha,G. M. Nasira," Artificial Neural Networks' Application in Weather Forecasting – Using RapidMiner ",International Journal of Computational Intelligence and Informatics, Vol. 4: No. 3, October - December 2014 ISSN: 2349-6363 177.

[11] Neha Khandelwal, Ruchi Davey,"Climatic Assessment Of Rajasthan's Region For Drought With Concern Of Data Mining Techniques", International Journal Of Engineering Research and Applications (IJERA) ISSN: 2248-9622 www.ijera.com Vol. 2, Issue 5, September- October 2012, pp.1695-1697, 1695.