Weather Forecasting Analysis using Linear and Logistic Regression Algorithm

Tanvi Patil [1], Dr. Kamal Shah[2]

[1,2]Thakur College of Engineering and Technology, Kandivali (E), Mumbai, Maharashtra 400101

I. ABSTRACT

Weather forecasting is the process of technology and science which used to predict the atmospheric conditions for the given location. The great scientist have also attempted to predict the weather formally and informally from many centuries. Weather forecasts are made by collecting the data about the current state with different attributes of the climate at a given place and using meteorology to project how the attributes affect the atmosphere.

The objective of the system is to predict the weather over a given period. The weather condition at any instance is described by using different kinds of variables. Out of these attributes, only significant attributes are used in the process of weather prediction. The selection of such attributes depends strongly on the location you have selected. The existing weather condition attributes are used to fit a model and by using the machine learning techniques and extrapolating the information, the future variations in the attributes are analysed.

Keywords: - Data Mining, Linear regression, weather forecasting, logistic regression, classification, data pre-processing, prediction.

II. INTRODUCTION:

The most essential application in the meteorology is weather forecasting and also the most scientific challenging problems around the world. Weather prediction is mainly concerned with the prediction of weather condition for a given environment. The essential purposes of weather prediction are climate monitoring, drought detection, agriculture and production, planning in aviation industry, communication, pollution dispersal, and many more. There is a considerable historical record of instances where weather conditions have altered the course of battles in the military operations. Accurate prediction of weather conditions is a difficult task since weather is non-linear and dynamic process i.e., it varies from day to day and even from minute to minute. The accuracy of the prediction depends on the knowledge of previous weather conditions over large areas and over large period. The critical information about the future weather is provided by the forecasting department. There are many approaches available in weather forecasting, from relatively simple observation of the sky to the highly complex computerized mathematical models.

III. RELATED THEORY:

Linear Regression:

Linear Regression is a machine learning algorithm used for the prediction of parameter which is in continuous nature. In this project, linear regression has been used for forecasting the minimum and maximum temperature and wind speed.

The major objectives of Linear Regression:

Linear regression has been used for the following two objectives:

- In order to find the relationship among variables (here maximum temperature rainfall and minimum temperature, etc.).
- In order to estimate the values of some attributes so that new observations are entertained.

Logistic Regression:

Logistic regression is used when the output are in categorical form. This paper consist the use of logistic regression for forecasting the weather and check the probability of rainfall which in turn decides whether it will rain or not.

The major objective of Logistic Function:

For this project, logistic regression is being for the categorizing the data and to predict the probability of rainfall.

Major formulae used in the project:

The fundamental equation of generalized linear model is:

$g(E(y)) = \theta 0 + x1\theta 1 + x2\theta 2 + x3\theta 3 + x4\theta 4 + x5\theta 5 + \cdots$ (1)

Where

g () is the function of the link,

E(y) is the linear predictor i.e. minimum and maximum temperature, relative humidity and wind speed.

The role of link function is to link the expectation of y to linear predictor. The cost function is use to predict the optimum value of θ0, θ1, θ2, θ3, θ4, θ5.

For those values cost function has minimum value and the predicted line is best fit.

IV. METHODOLOGY:

System architecture

The project is divided into is following phases. Each phase will require different technical fields which the project encompasses.

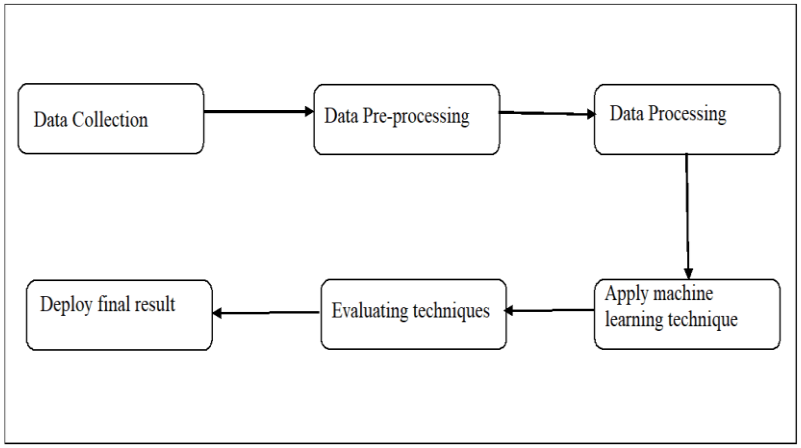

Figure 1: System Architecture

Data Collection:

The data is collected from Aurangabad district from year 1969 to 2015 having various attributes like Relative humidity, wind speed, max temp and min temp.

```
In [4]: df.head()
```

Out[4]:

	INDEX	YEAR	MN	HR	DT	.RH	DD	FFF	max	min	R/F
0	42920	1969	6	12	3	77	9	26	37.4	20.4	9.0
1	42920	1969	6	12	4	46	23	16	35.3	21.2	14.0
2	42920	1969	6	12	5	57	23	32	34.2	22.0	3.9
3	42920	1969	6	12	6	81	23	24	35.0	23.2	1.0
4	42920	1969	6	12	7	93	27	34	30.9	22.5	10.6

Fig 2: Reading the data

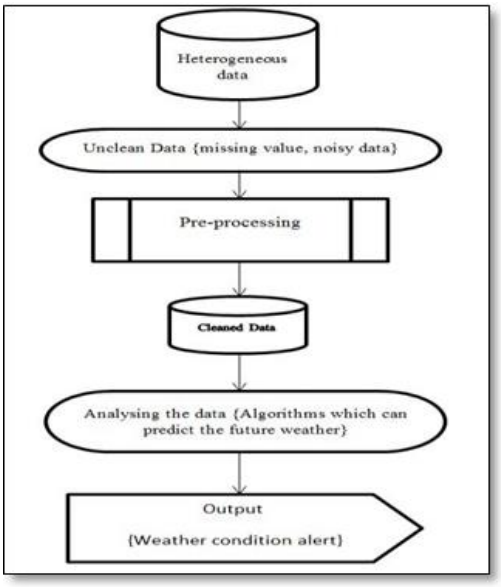

Fig 3: Weather Forecasting Model

The data collected from sensors isheterogeneous in nature which gets stored in server database. The data obtained may contain some noisy information or missing values. These missing values are removed using pre-processing of raw data. Pre-processing can be attained at the sensor node itself. It involves statistical techniques to smooth the data by applying methods.

Once the cleaned data is acquired, various algorithms can be applied to perform prediction. To obtain future events according tocurrent and past trends which algorithms are to be used is a majorpart of research. This paper focuses on algorithmic techniques which can be used for weather forecasting. Further section will describe the algorithms used for prediction of weather data.

Data Pre-processing:

Three common data pre-processing phases are:

• Formatting: The information we have picked probably won't be in an arrangement that permits one to work with. The information can be in a social database, and you might want to be in a plain document, or the information may be in a restrictive record configuration, and you need it in a social database or content record.

• Cleaning: Information cleaning includes the disposal or fixing of missing information. Occasions of information might be fragmented and don't hold the information that you trust you have to address the issue. We might need to erase these

cases. Also, a portion of the qualities may contain delicate data and these ascribes may should be anonymized or expelled from the information.

• Sampling: The information accessible might be significantly more separated than you have to work with. For calculations and more prominent computational and memory prerequisites, more information can bring about any longer running occasions. While assessing the whole dataset, you can take a littler agent test of the chosen information, which might be a lot simpler to investigate and test arrangements.

```
df.describe()
```

	INDEX	YEAR	MN	HR	DT	.RH	DD	FFF	max	min	R/F
count	10187.000000	10187.000000	10187.000000	10187.0	10187.000000	10187.000000	10187.000000	10187.000000	10181.000000	10187.000000	10187.000000
mean	43069.394522	1988.475115	7.433101	12.0	16.358987	81.536763	26.832237	11.886326	28.451577	22.410513	15.280126
std	94.679154	12.147825	1.008951	0.0	8.685724	10.954393	5.665943	9.919163	2.464947	1.625181	27.353372
min	42920.000000	1969.000000	6.000000	12.0	1.000000	25.000000	2.000000	1.000000	22.300000	14.900000	0.100000
25%	42920.000000	1978.000000	7.000000	12.0	9.000000	76.000000	25.000000	4.000000	26.800000	21.300000	1.600000
50%	43110.000000	1988.000000	7.000000	12.0	17.000000	83.000000	27.000000	8.000000	28.400000	22.100000	5.200000
75%	43157.000000	1998.000000	8.000000	12.0	24.000000	90.000000	27.000000	16.000000	29.800000	23.600000	16.100000
max	43157.000000	2015.000000	9.000000	12.0	31.000000	100.000000	99.000000	88.000000	42.200000	28.200000	385.100000

Fig 4: Rainfall Dataset

This paper focuses on four such algorithms which are:

1. Linear Regression.
2. Logistic Regression

Each of these techniques is briefly described below:

Linear Regression:

Linear regression is a simple technique for supervised learning which works on structural as well as time series data. It is the assumption that the dependency of Y on X_1, X_2, X_p is linear. This relationship is denoted by following formula:

$$Y = \beta_0 + \beta_1 X + €$$

Logistic Regression:

Logistic Regression is one of the most simple and commonly used Machine Learning algorithms for two-class classification. It is easy to implement and can be used as the baseline for any binary classification problem. The basic fundamental concepts of Logistic Regression are also constructive in deep learning. Logistic regression describes and estimates the relationship between one dependent binary attribute and independent attribute.

V. RESULT AND ANALYSIS

INTRODUCTION

This chapter describes about the dataset and the results obtained from the different machine learning algorithms. Results obtained have been checked on the different types of parameters. The result is shown in the bar diagram form. In this we will more discuss about the data description, Train the model, Model Evaluation.

LINEAR REGRESSION

Fig 5: Training the model

Model Development and Prediction

First, import the Logistic Regression module and create a Linear Regression classifier object using Linear Regression () function.

Then, fit your model on the train set using fit () and perform prediction on the test set using predict ().

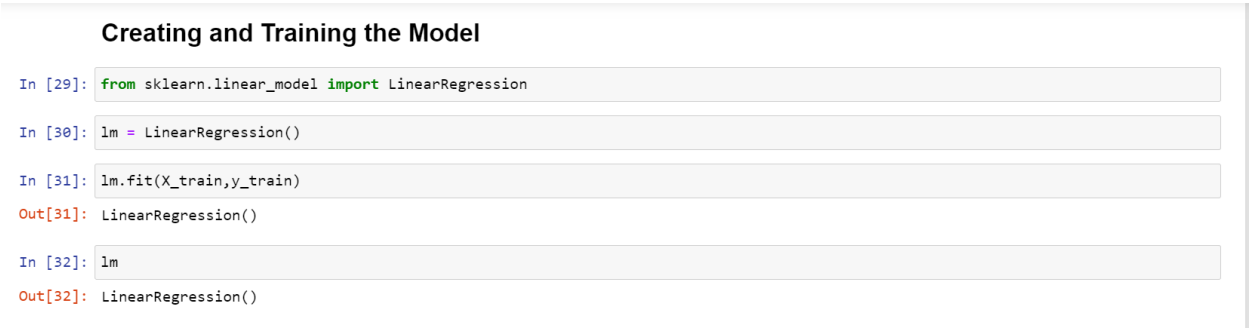

Fig 6: Training the model

Model Evaluation & Prediction

Let's evaluate the model by checking out it's coefficients and how we can interpret them.

```
In [33]: X.columns

Out[33]: Index(['HR', '.RH', 'DD', 'FFF', 'max', 'min'], dtype='object')

In [34]: coeff_df = pd.DataFrame(lm.coef_,X.columns,columns=['Coefficient'])
         coeff_df  #read linear regression
```

Out[34]:

	Coefficient
HR	0.000000
.RH	0.655399
DD	-0.151555
FFF	-0.169410
max	-0.240564
min	1.035796

```
In [35]: lm.score(X_train, y_train)  #R square

Out[35]: 0.08440399821390743
```

Fig 7: Model Evaluation

```
In [36]: predictions=lm.predict(X_test)

In [37]: predictions

Out[37]: array([ 0.28527308, 27.4507702 , 18.32683214, ..., 11.53626388,
                11.85850084, 16.57275932])

In [39]:
print('MAE:', metrics.mean_absolute_error(y_test, predictions))
print('MSE:', metrics.mean_squared_error(y_test, predictions))
print('RMSE:', np.sqrt(metrics.mean_squared_error(y_test, predictions)))  #Root mean square error,near to zero.

MAE: 15.212134082597709
MSE: 614.2127615155913
RMSE: 24.783316192866348
```

Fig 8: Prediction

Output:

```
print('MAE:', metrics.mean_absolute_error(y_test, predictions))
print('MSE:', metrics.mean_squared_error(y_test, predictions))
print('RMSE:', np.sqrt(metrics.mean_squared_error(y_test, predictions)))  #Root mean square error,near to zero.

MAE: 15.212134082597709
MSE: 614.2127615155913
RMSE: 24.783316192866348
```

Accuracy for linear regression:

```
In [212]: lm.score(X_test,y_test)

Out[212]: 0.07760990775483523
```

Scatter plot on various attributes:

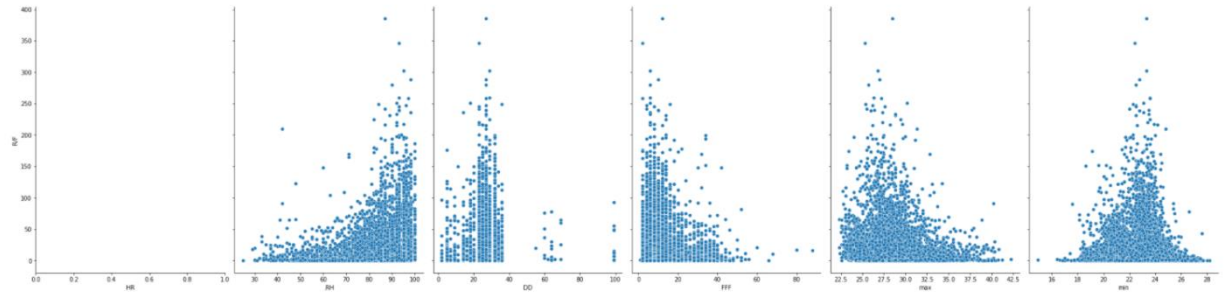

Fig 9: Scatter Plot

We have tested and evaluated the linear regression model on the rainfall dataset and the accuracy of the model is 0.084%.Further we will apply the logistic regression model to find the accuracy of the model on the dataset.

LOGISTIC REGRESSION:

Applying Logistic Regression:

```
In [42]: X = df1.drop(["T/F","R/F"],axis=1)
         y = df1['T/F']

In [43]: X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=101)

In [44]: from sklearn.linear_model import LogisticRegression

In [45]: lr= LogisticRegression()

In [46]: lr.fit(X_train,y_train)
Out[46]: LogisticRegression()

In [47]: y_pred=lr.predict(X_test)
```

Fig 10: Logistic regression

Accuracy for Logistic regression:

```
In [48]: from sklearn.metrics import accuracy_score

In [49]: accuracy_score(y_test, y_pred)
Out[49]: 0.6720667648502701
```

Confusion matrix:

A confusion matrix is a summary of prediction in the form of a table which is used to describe the performance of a model on a given dataset and give results on a classification problem. The correct and incorrect predictions are summarized with count values and broken down by each class. This is the key to the confusion matrix. The confusion matrix shows the ways in which your classification model is confused when it makes predictions. The Confusion matrix not only gives us the insight into the errors being made by a classifier but also gives more us the types of errors that are being made.

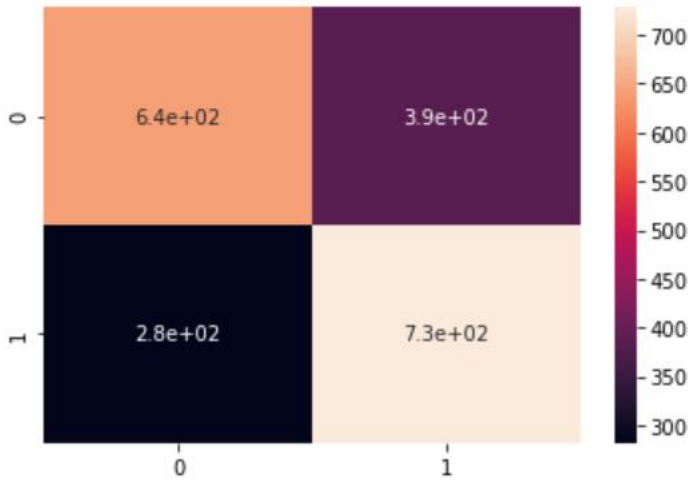

Fig 11: Confusion matrix of Logistic regression on weather data

VI. CONCLUSIONS AND FUTURE WORK

The most scientific and technical challenging problem around the world is forecasting the weather. Weather Prediction relies on two correct things 1) First the collection of the data from the meteorological department and 2) the appropriate selection of the data mining techniques for predicting the weather conditions. The major concerns of Weather prediction are the Accuracy of the model and its Timely output. The Problem domain of Weather Forecasting is very vast and therefore it is very feasible to use data mining techniques which can perform in a thorough manner with the complex problem domain of weather forecasting and give some accurate results. However more than one data mining technique is applied in parallel for better and accurate results for the weather prediction. The proposed work is an attempt to forecast different weather conditions using a fusion of different forecasting and data mining techniques. Even though the rainfall is dependent on many parameters, the proposed model was able to get an impressive classification accuracy using limited parameters.

VII. REFERENCES

[1] Shahi, R. Atan, M. N. Sulaiman, "An Effective Fuzzy C-Mean and Type-2 Logic for Weather Forecasting,". Journal of Theoretical and Applied Information, vol. 5 (5), pp. 556–567, 2009.

[2] Paras, Sanjay Mathur. "A Simple Weather Forecasting Model Using Mathematic Regression." *Indian Research Journal of Extension Education*12.2 (2016): 161-168.

[3] Beniwal, S. &Arora, J. (2012), —Classification and feature selection techniques in data mining‖, International Journal of Engineering Research & Technology (IJERT), 1(6)

[4] Rohit Kumar Yadav, Ravi Khatri, "A Weather Forecasting Model using the Data Mining Technique", International Journal of Computer Applications (0975 –8887)Volume 139 – No.14, April 2016

[5] Data Mining Concepts and Techniques Third EditionJiawei Han and MichelineKamber Morgan Kaufmann Publishers. Page 351

[6] A Quick Review of Machine Learning Algorithms, 2019, International Conference on Machine Learning, Big Data, Cloud and Parallel Computing.

[7] Vijayalakshimi, K. and Lakshmi, G.V.M. (2016) Real Time Weather Monitoring from Remote Location Using Raspberry Pi. International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering (IJARCCE)

[8] Ms.P.Shivaranjani, Dr.K.Karthikeyan. "A Review of Weather Forecasting Using Data Mining Techniques", International Journal of Engineering and Computer Science ISSN: 2319-7242 Volume 5 Issue 12 Dec. 2016.

[9] 2017 International Conference on Advanced Informatics, Concepts, Theory and Applications. Weather prediction based on fuzzy logic algorithm for supporting general farming automation system, 2017 5th International Conference on Instrumentation, Control, and Automation (ICA).