JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

REGRESSION TECHNIQUES BASED ON WHETHER FORECASTING PREDICTION

Gunda Nithin[1], M Sai Trilochan[2], G Sangamesh[3], Mr. Vigneshwara Reddy[4], Dr. Subhani Shaik[5]

[1,2,3]IV Year Students, [4]Assistant Professor,[5]Associate Professor

Dept. of Information Technology, Sreenidhi Institute of Science and Technology(Autonomous), Hyderabad, India.

Abstract: Weather forecasting is the most important thing for observing the atmosphere observations by location and time. Regression techniques are a mainstay of information and have been chosenfornumericalmachine-learning concepts. This may be perplexing because we can use regression to refer to the class of problem and algorithm. This paper provides temperature predictions in weather forecasting from time to time using regression-based techniques. We have to use four techniques Ordinary Least Squares Regression, Logistic Regression, Multivariate Adaptive Regression Splines, and Locally Estimated Scatterplot Smoothing. We focus on the mean square error rate of each technique. Different types of metrics are tested in this research. Low MSE is indicated for better weather forecasting results. The statistical results consider the lowerMSE rateas optimal.

Keywords: Regression techniques, weather forecasting, Prediction, MSE

I. INTRODUCTION

Weather forecasting is most important scientific technique to predict the status of atmosphere at certain time and locations. In past days weather forecasting carried out by manually alters barometric pressure and current weather conditions.But now it measures on computer-based models that consider many factors for predict [1].The researchers connect linear relationship between input attribute to corresponding target attribute. This procedure is different in non linear prediction of the weather with multiple attribute relations. Weather forecasting make by gathering quantitative data about the present status and past trend of the atmosphere by using technology to predict the conditions of atmosphere analysis.Weather conditions most important for farmers, business and normal public for protection of life and properties [2].Actually, machine learning based weather forecasting started from 2018. Initially creating medium range weather forecasting for machine learning. For benchmark problems machine learning play key role in many research areas. Due to this accessibility researchers work on wide variety of backgrounds [3].

Machine learning based weather forecasting is an initial technology for predict the weather conditions in high quality [4]. These machine learning models take less time and resources in single attempt. Machine learning based techniques heavily used for weather forecasting throughout the world [6]. The procedure train and validation of the data continuously generate the accurate results for weather conditions. Quality forecasting is most important for daily life of public [7].

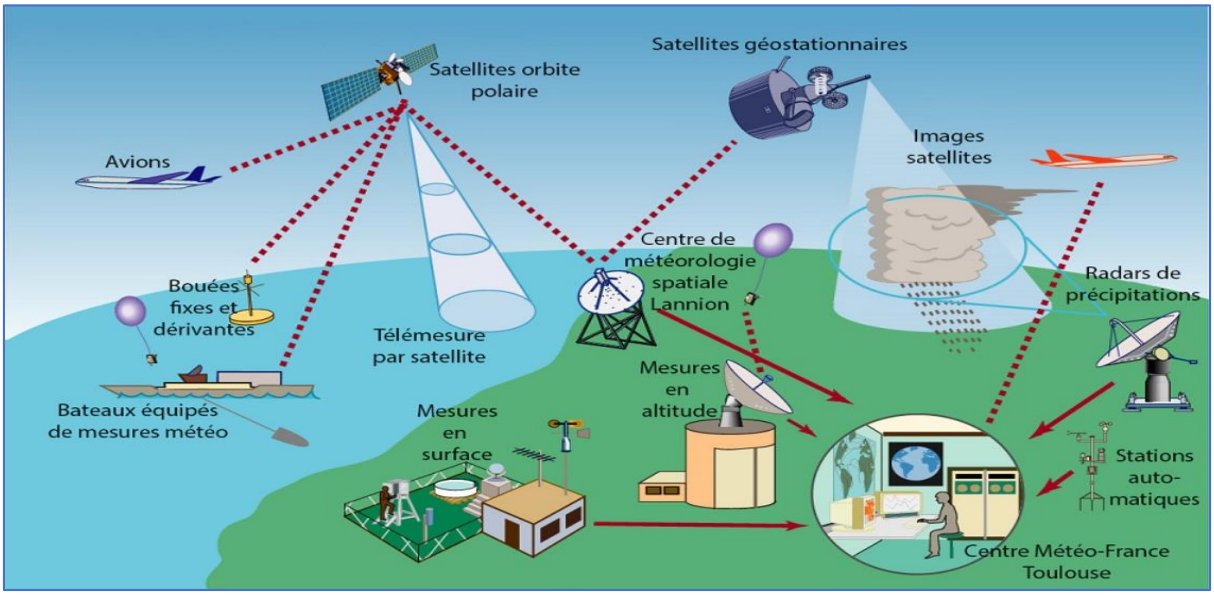

Figure 1: system in operational methodology [5]

The figure 1 shows the system for weather forecasting analysis of data from outside environment. Different fields connected with weather forecasting for successfully execute their work in proper time without any loss.

The following paper discuss the proposed method and architecture in section 2. Section 3 state the results and analysis. Final section concludes the paper.

II. PROPOSED METHODS AND ARCHITECTURE

Regression techniques are a mainstay of information and have been chosen into numerical machine learning concepts. This may be perplexing because we can use regression to refer to the class of problem and algorithm [8]. This paper provides the temperature predictions in weather forecasting time to time using regression-based techniques. We have to use four techniques like Ordinary Least Squares Regression, Logistic Regression, Multivariate Adaptive Regression Splines and Locally Estimated Scatterplot Smoothing. We focus on mean square error rate of each technique.The following figure 2 is a proposed architecture of our research. Different phases included in this architecture for weather forecasting.

1. Environment (outside world)
2. Weather sensors (perceive the information for environment)
3. Model Implementation (suitable model prepare based on algorithms)
4. Classification (classify the data based on certain conditions)
5. Weather conditions (predict the results)

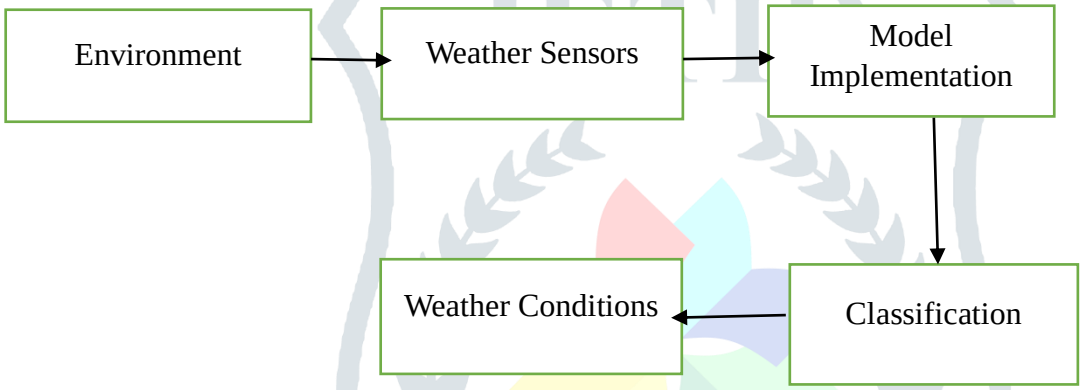

Figure 2: block diagram of proposed system[9]

III. RESULTS AND ANALYSIS

The statistical results generated using python programming environment with Jupiter notebook tool. Different types of regression techniques used for predicting the result of weather conditions.

3.1 Dataset description

Upload the datasetinto system for next preparation of data for classification. It consists of six attributes and 1461 records.

Table 1: Dataset [10]

	date	precipitation	temp_max	temp_min	wind	weather
0	2012-01-01	0.0	12.8	5.0	4.7	drizzle
1	2012-01-02	10.9	10.6	2.8	4.5	rain
2	2012-01-03	0.8	11.7	7.2	2.3	rain
3	2012-01-04	20.3	12.2	5.6	4.7	rain
4	2012-01-05	1.3	8.9	2.8	6.1	rain
...	...	...	...	...	...	...
1456	2015-12-27	8.6	4.4	1.7	2.9	rain
1457	2015-12-28	1.5	5.0	1.7	1.3	rain
1458	2015-12-29	0.0	7.2	0.6	2.6	fog
1459	2015-12-30	0.0	5.6	-1.0	3.4	sun
1460	2015-12-31	0.0	5.6	-2.1	3.5	sun

1461 rows × 6 columns

3.2 Data Preprocessing

Data preprocessing is a procedure of making the pure data from given raw data. Prepare data for suitable machine learning models. It is the initial and vital step while generating a machine learning model [11].

3.3 Data Visualization

The following figures 3 and 4 shows the data visualization of temperature in minimum and maximum. [12]

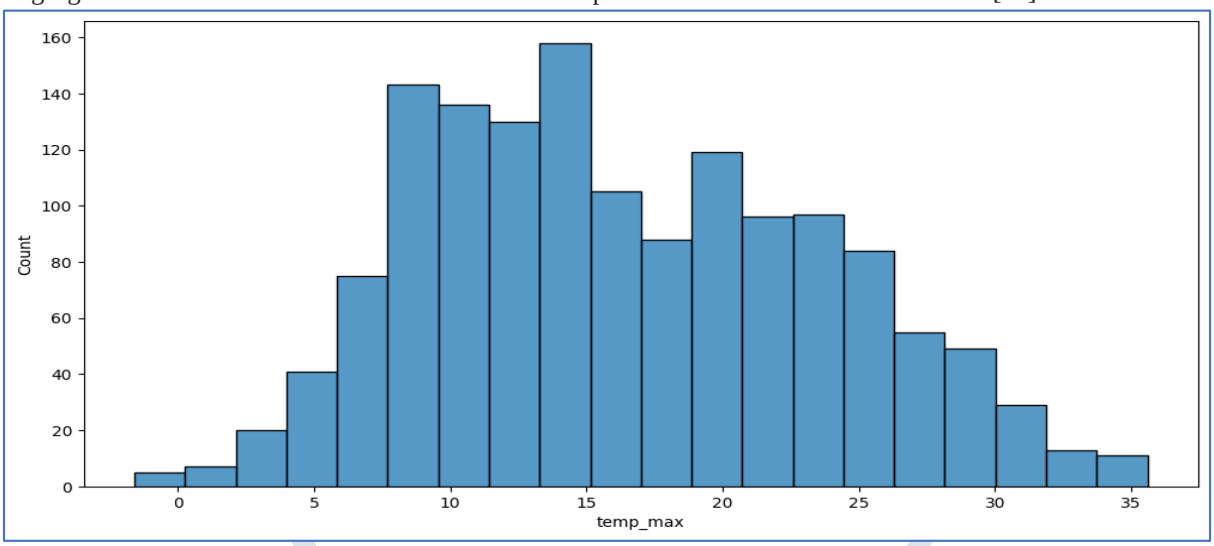

Figure 3:graph for data visualization in maximum temperature

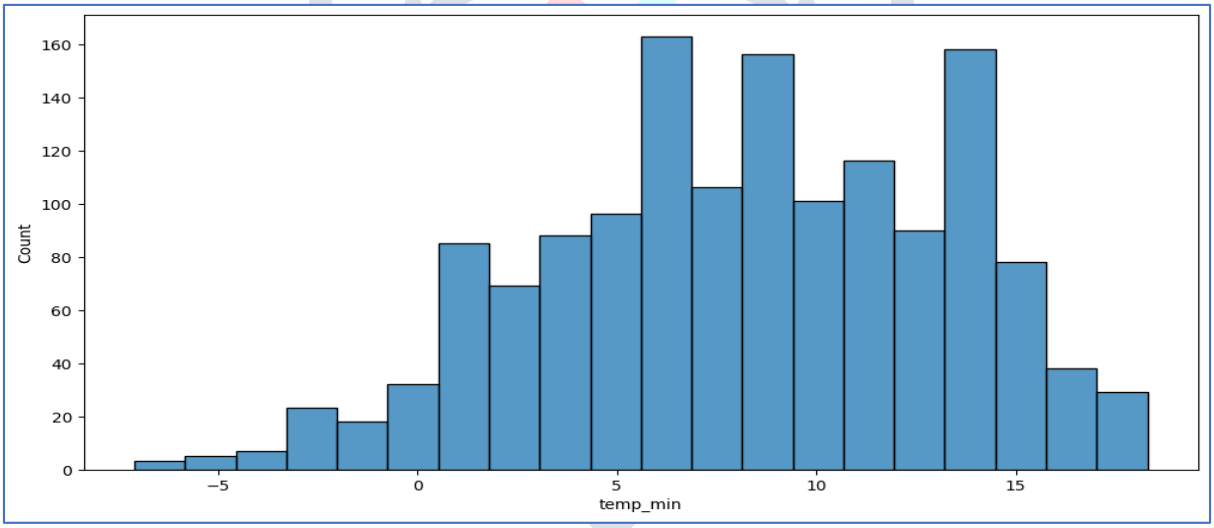

Figure 4:graph for data visualization in minimum temperature

The following figure 5 represents the maximum temperature in each month in each yearfrom 2019 to 2023.

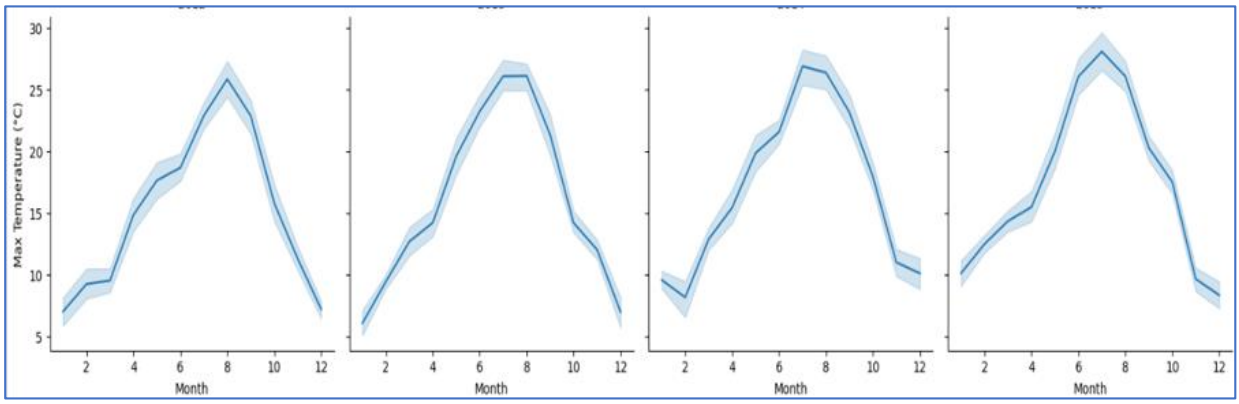

Figure 5:data visualization max temperature in each month in each year

The following figure 6 represents the minimum temperature in each month in each year from 2019 to 2023.

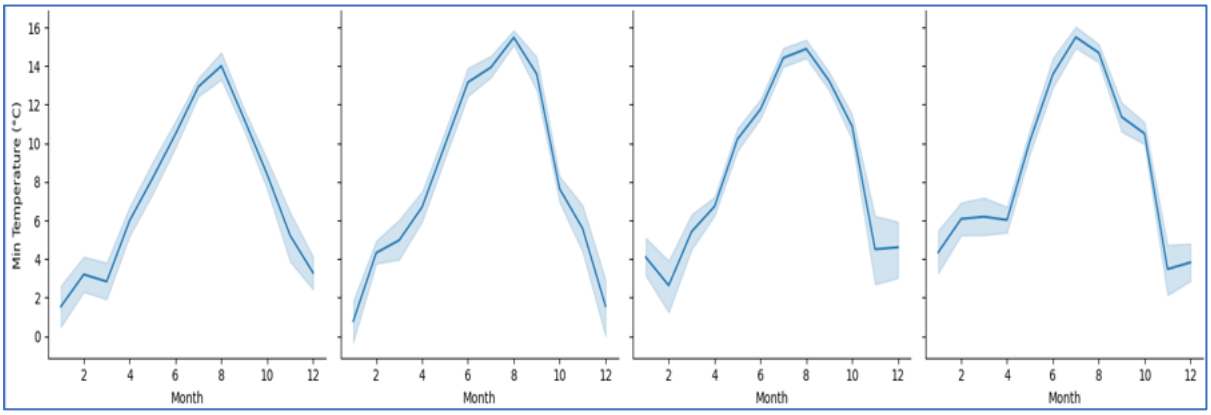

Figure 6:data visualization min temperature in each month in each year

A form of water, such as rain, snow, or sleet, that condenses from the atmosphere, becomes too heavy to remain suspended, and falls to the Earth's surface.The following figure 5.6shows the overall precipitation in each month in each yearand 7 shows the Precipitation in wind speed in each month in each year [13].

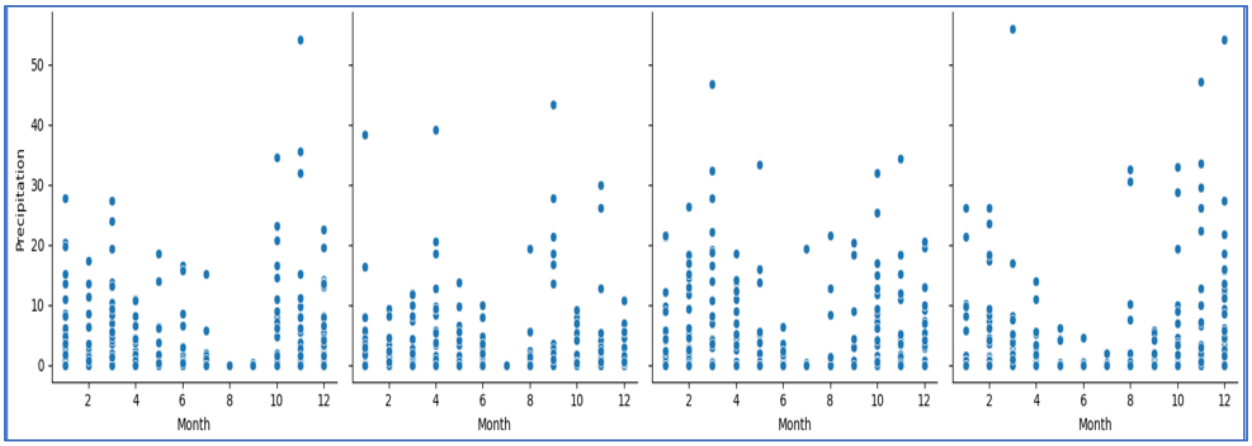

Figure 7: Precipitation in each month in each year

```
g = sns.FacetGrid(df, col='year', col_wrap=4, height=5)
g.map(sns.scatterplot, 'month', 'wind')
g.set_axis_labels('Month', 'Wind speed')
g.set_titles(col_template="{col_name}")
plt.show()
```

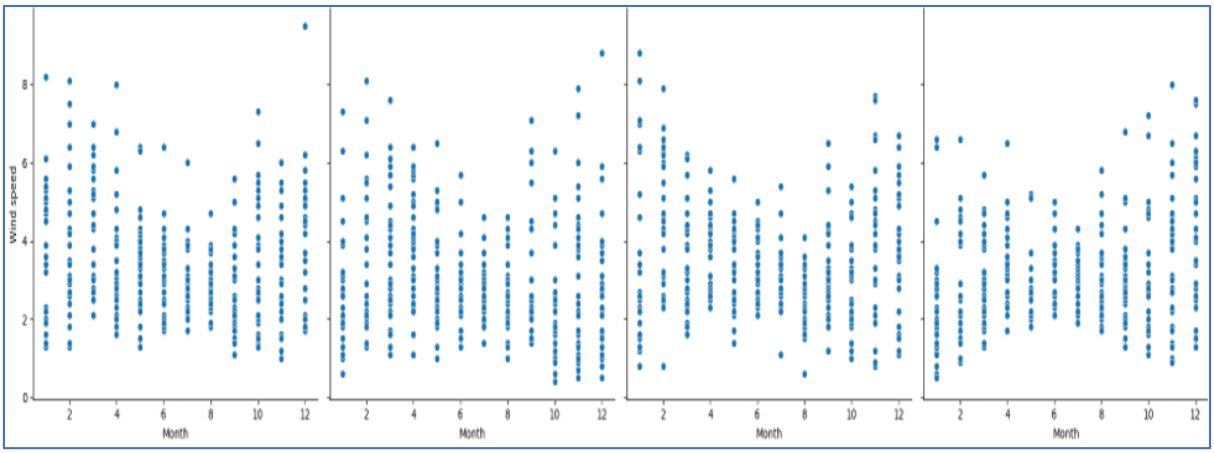

Figure 8: Precipitation in each month in each year (wind speed)

The following figure 8shows thedata visualization of different weather conditions like fog, snow, rain, sun and dazzle. Figure 9shows thedata visualizationfor distribution of weather types in percentage.

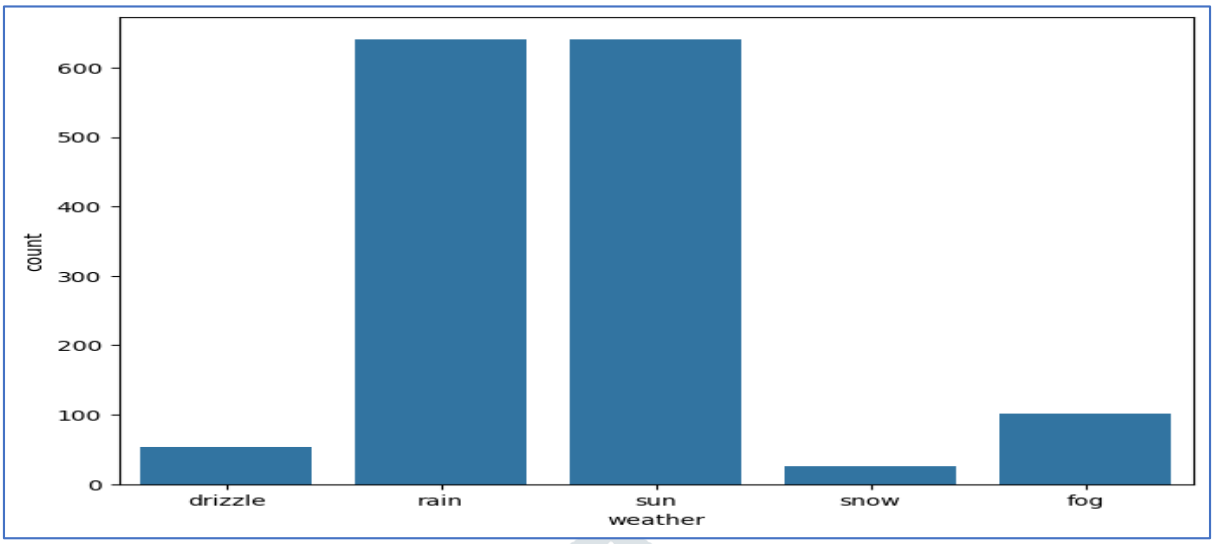

Figure 9:data visualization of different weather conditions

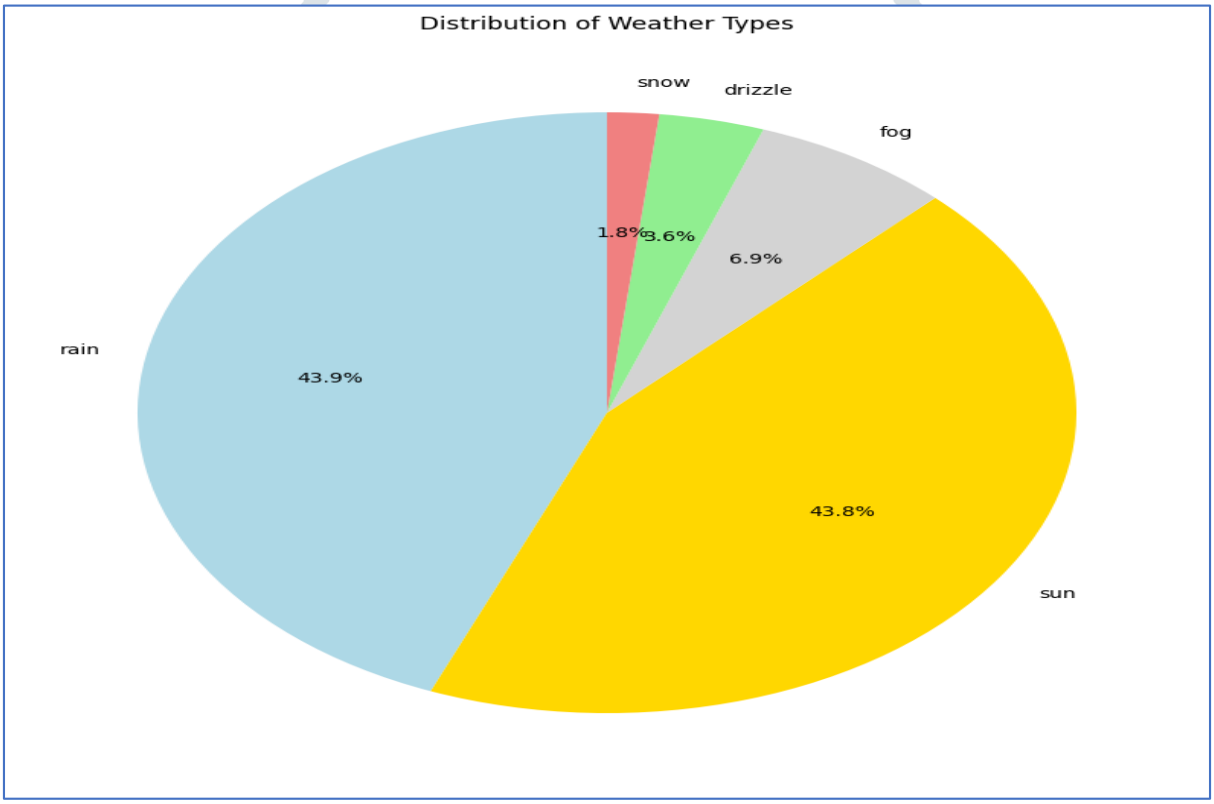

Figure 10:data visualization of different weather conditions

3.4 Regressor Analysis of dataset

Regression analysis is a set of numerical methods used for the assessment of associations between a dependent variable and independent variables.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1461 entries, 0 to 1460
Data columns (total 6 columns):
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   date           1461 non-null   datetime64[ns]
 1   precipitation  1461 non-null   float64
 2   temp_max       1461 non-null   float64
 3   temp_min       1461 non-null   float64
 4   wind           1461 non-null   float64
 5   weather        1461 non-null   object
dtypes: datetime64[ns](1), float64(4), object(1)
memory usage: 68.6+ KB
```

```
def huber_loss(y_true, y_pred, delta=1.0):
    error = y_true - y_pred
    is_small_error = np.abs(error) <= delta
    squared_loss = 0.5 * error**2
    linear_loss = delta * (np.abs(error) - 0.5 * delta)
    return np.where(is_small_error, squared_loss, linear_loss)
```

```
from sklearn.preprocessing import LabelEncoder
le=LabelEncoder()
df['weather']=le.fit_transform(df['weather'])
x = df[['temp_min', 'temp_max', 'precipitation', 'wind']]
y=df['weather']
```

3.5 Ordinary Least Squares Regression (OLSR)

This method OLSR generate the MSE is 1.07, MAE is 0.83, MSLE is 0.24, maximum error is 3.28, median absolute error is 0.67 and huber loss is 0.44.

```
Mean Squared Error (OLSR): 1.072145890750182
Mean Squared Error: 1.072145890750182
Mean Absolute Error: 0.8337729527271713
R-squared: 0.24691005911461006
Mean Squared Logarithmic Error: 0.12272273074166824
Explained Variance Score: 0.24691085731054307
Maximum Error: 3.285459408477359
Median Absolute Error: 0.6748184043980787
Huber Loss: 0.4432257028587715
```

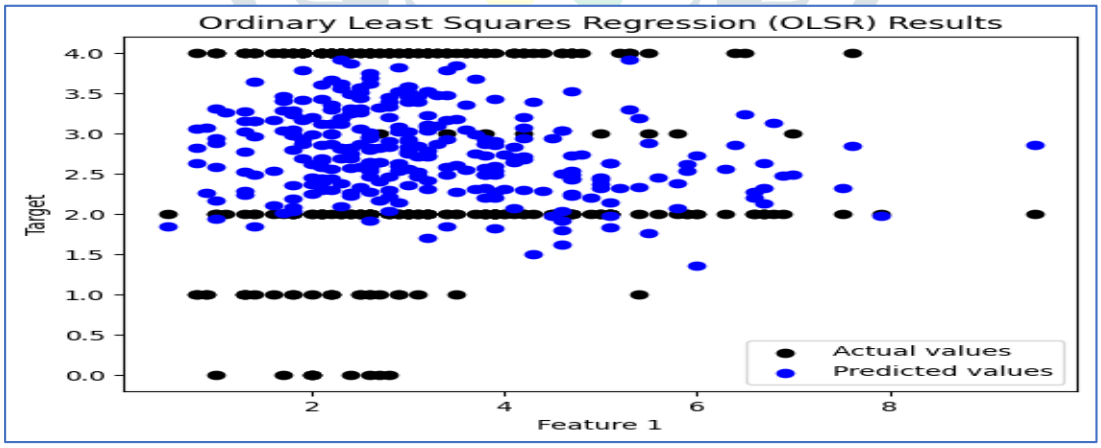

Figure 11:OLSR results for actual vs target values

3.6 Logistic Regression

This method Logistic regression generate the MSE is 1.39, MAE is 0.45, MSLE is 0.16, maximum error is 4, median absolute error is 0.0 and huber loss is 0.37.

```
Mean Squared Error (Logistic Regression): 1.3924914675767919
Mean Squared Error: 1.3924914675767919
Mean Absolute Error: 0.45733788395904434
R-squared: 0.021894943544428114
Mean Squared Logarithmic Error: 0.16030443872226058
Explained Variance Score: 0.13971526755031893
Maximum Error: 4
Median Absolute Error: 0.0
Huber Loss: 0.37372013651877134
```

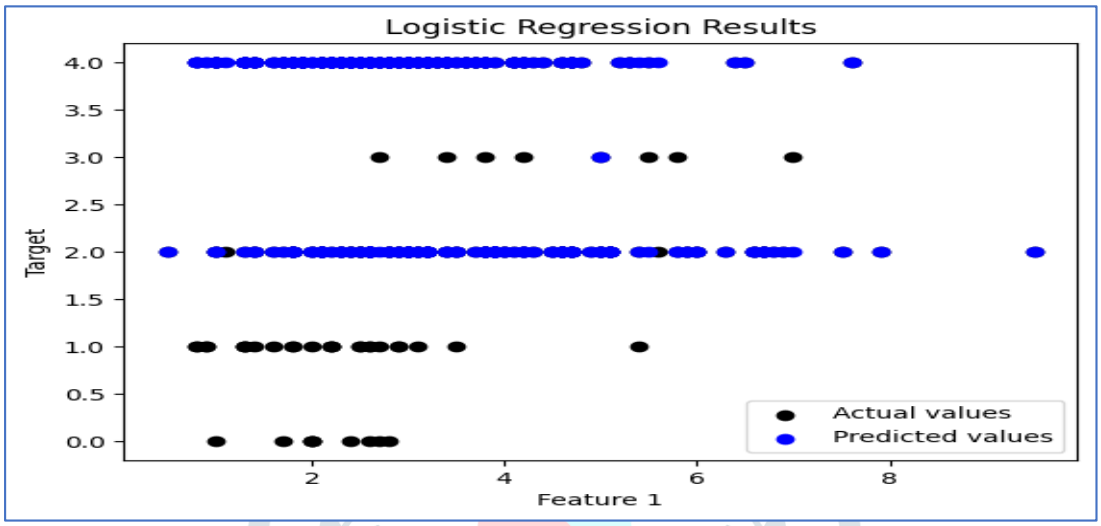

Figure 12: Logistic Regression results for actual vs target values

3.7 Multivariate Adaptive Regression Splines (MARS)

This method MARS generate the MSE is 1.99, MAE is 0.75, MSLE is 0.12, maximum error is 3.27, median absolute error is 0.60 and huber loss is 0.39.

```
Mean Squared Error (MARS): 0.9963239517346019
Mean Squared Error: 0.9963239517346019
Mean Absolute Error: 0.7507737873755044
R-squared: 0.3001684263421385
Mean Squared Logarithmic Error: 0.12092198810950003
Explained Variance Score: 0.30034662229333653
Maximum Error: 3.2706111740060844
Median Absolute Error: 0.6072167513929974
Huber Loss: 0.39462514443330554
```

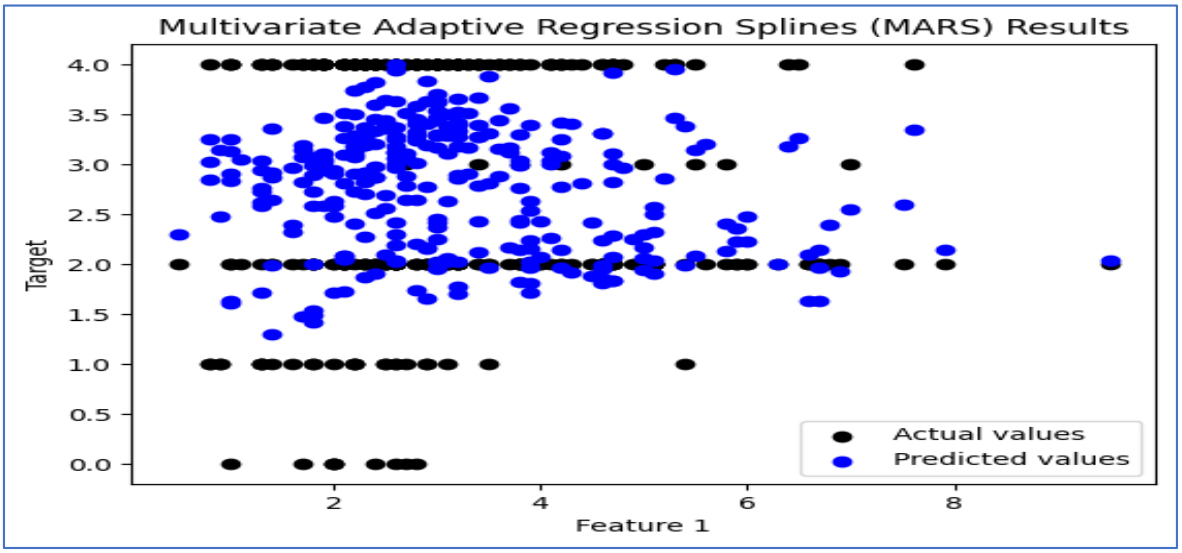

Figure 13:MARS results for actual vs target values

3.8 Locally Estimated Scatterplot Smoothing (LOESS)

This method LOESSgenerates the MSE is 1.26, MAE is 0.97, MSLE is 0.13, R squared is 0.11,Variance 0.11, Maximum error is 3.01, Median absolute error is 0.94 and Huber loss is 0.54.

```
Mean Squared Error (LOESS): 1.2670347358605656
Mean Squared Error: 1.2670347358605656
Mean Absolute Error: 0.9776203055885682
R-squared: 0.11001746818120028
Mean Squared Logarithmic Error: 0.13348889019611102
Explained Variance Score: 0.11048769758025989
Maximum Error: 3.015671863461137
Median Absolute Error: 0.9429770367667869
Huber Loss: 0.5442133209933444
```

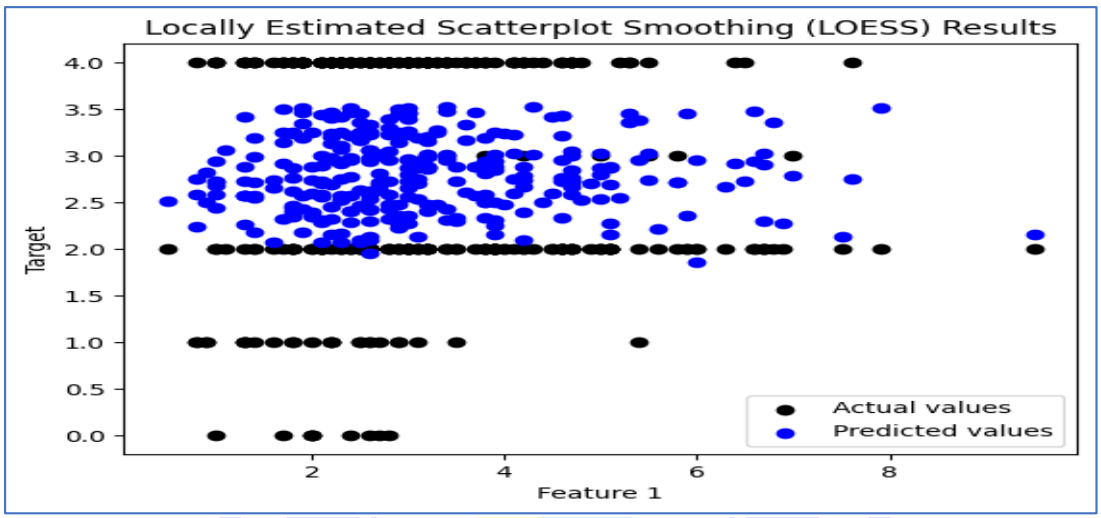

Figure 14:LOESS results for actual vs target values

IV. COMPARATIVE STUDY

The following table 2demonstrations the comparative study of dissimilar values of four models for detect the finest model for weather temperature prediction.

4.1 Metrics for Error Rate

4.1.1Mean square error

It is calculated by taking the average, specifically the mean, of errors squared from data as it relates to a function. Among four models MARS and OLSR get less MSE values. Less value prefers in MSE context for good accuracy.

4.1.2 MeanAbsolute error

MAE takes the average of absolute errors for a group of predictions and observations as a measurement of the magnitude of errors for the entire group.

The MAE is higher than the MSE, which suggests that your model has some small errors that accumulate in the MAE. Less than 5% is most acceptable MAE in research.

Four models get less than 1 MAE. It shows considered all models. Among four models' logistic regression get less amount of MAE. It prefers more than other models.Lower values prefer mostly than higher. In outlier point of view, we prefer MAE than MSE because MAE is more relevant to error rate prediction.

4.1.3 R Squared Value

R-squared is a statistical measure that represents the goodness of fit of a regression model. The value of R-square lies between 0 to 1. All models are accepted in the sense of R squared value.

4.1.4 Mean Square Logarithmic Error (MSLE)

It is similar to mean square error metric. It is called the Mean Squared Logarithmic Error. It can also be interpreted as a measure of the ratio between the actual and predicted values.Among four models get less MSLE values. Less value prefers in MSLE context for good accuracy.

4.1.5 Variance

Variance refers to the changes in the model when using different portions of the training data set. So, it is required to make a balance between bias and variance errors. For an accurate prediction of the model, algorithms need a low variance and low bias. Among four models Logistic Regression and LOESS preferable in variance context. Because these models predict less values.

4.1.6 Max Error

The Max-Error metric is the worst-case error between the predicted value and the true value. Among four models OLSR, MARS and LOESS get less error value in max error metric. LOESS more less error value among other two. Logistic regression gets highest error value.

4.1.7 Median Absolute Error

The Median Absolute Error is the median difference between the observations (true values) and model output (predictions). Lower values are preferable to accurate result. Here Logistic regression model consider for better to generate perfect accuracy.

4.1.8 Huber Loss

Huber Loss is a popular loss function used in machine learning for regression tasks. Huber Loss is a hybrid between MSE and MAE and is designed to provide the benefits of both loss functions. Huber Loss has the advantage of being less sensitive to outliers than MSE while still providing a more balanced approach to evaluating the performance of a regression model compared to MAE.

Huber loss metric also prefer low value for best model in machine learning. Four models are generatingless than 1 value. Logistic regression generates 0.37 less value compares to others.

Table 1: Comparative analysis of four models

Name of the Regressor	MSE	MAE	R squared Value	MSLE	Variance	Max Error	Median AE	Huber Loss
OLSR	1.07	0.83	0.24	0.12	0.24	3.28	0.67	0.44
Logistic Regression	1.39	0.45	0.02	0.16	0.13	4	0	0.37
MARS	0.99	0.75	0.30	0.12	0.30	3.27	0.60	0.39
LOESS	1.26	0.97	0.11	0.13	0.11	3.01	0.94	0.54

In our research used four models in different metrics like MSE,MAE, R Squared value, MSLE, Variance, Median Absolute Error, Max Error and Huber loss. Among all metrics most preferable metrics are MSE, MAE, and Huber loss. Among these three metricsMAE most preferable for outliers. Logistic regression get 0.45 for MAE value. The is less compare to other models. But one thing is remembering all these metrics values based on dataset size, dataset type, training data, test data and pre-processing of dataset.

V. CONCLUSION

Regression techniques are a mainstay of information and have been chosen into numerical machine learning concepts. This may be perplexing because we can use regression to refer to the class of problem and algorithm. This paper provides the temperature predictions in weather forecasting time to time using regression-based techniques. We have to use four techniques like Ordinary Least Squares Regression, Logistic Regression, Multivariate Adaptive Regression Splines and Locally Estimated Scatterplot Smoothing. We focus on mean square error rate of each technique. Different types of metrics are test in this research. Low MSE is indicate for better weather forecasting results. The statistical results consider the less MSE rate for optimal.

REFERENCES

[1] ImranMaqsood, Muhammad Riaz Khan, and AjithAbraham, “Anensembleofneuralnetwork for weather forecasting”, Neural Computing& Application (2004) 13: 112–122.

[2]AmanpreetKaur,JKSharma,andSunilAgrawal, “Artificial neural networks in forecasting maximum and minimum relative humidity”. International Journal of Computer Science and Network Security, 11(5):197-199, May 2011.

[3] https://rp5.ru/aRussianwebsitewithweatherdata of airport weather station fromall around the world

[4] https://colab.research.google.com/ Google’s free cloudserviceforAIdevelopersforeducationalor research purposes.

[5] https://www.encyclopedie-environnement.org/en/air-en/introduction-weather-forecasting/

[6] Culclasure, Andrew, "Using Neural Networks to Provide Local Weather Forecasts" (2013). Electronic Theses and Dissertations.32. https://digitalcommons.georgiasouthern.edu/etd/32

[7] Trebing, K., Stańczyk, T. and Mehrkanoon, S., 2021. Smaat-unet: Precipitation nowcasting using a small attention-unet architecture. Pattern Recognition Letters, 145, pp.178-186.

[8] Mr. Sujan Reddy, Ms. Renu Sri andSubhani Shaik,” Sentimental Analysis using Logistic Regression”, International Journal of Engineering Research and Applications (IJERA), Vol.11, Series-2, July-2021.

[9] Ms. Mamatha, Srinivasa Datta and Subhani Shaik,” Fake Profile Identification using Machine Learning Algorithms”, International Journal of Engineering Research and Applications (IJERA), Vol.11, Series-2, July-2021.

[10] R. Vijaya Kumar Reddy, Subhani Shaik,B. Srinivasa Rao, “Machine learning based outlier detection for medical data” Indonesian Journal of Electrical Engineering and Computer Science, Vol. 24, No. 1, October 2021.

[11] M. Jagan Chowhaan, D. Nitish, G. Akash, Nelli Sreevidya, Dr. Subhani Shaik, "Machine Learning Approach for House Price Prediction" Asian Journal of Research in Computer Science , Volume 16, Issue 2, Page 54-61, June- 2023.
[12] Neeraja, Anupam, Sriram, Subhani Shaik and V. Kakulapati," Fraud Detection of AD Clicks Using Machine Learning Techniques", Journal of Scientific Research and Reports, Volume 29, Issue 7, Page 84-89, June-2023.
[13] P. Pranathi, V. Revathi, P. Varshitha, Subhani Shaik and Sunil Bhutada,"Logistic Regression Based Cyber Harassment Identification", Journal of Advances in Mathematics and Computer Science, Volume 38, Issue 8, Page 76-85, June-2023.