Analysis of machine learning models in predicting weather conditions and its severity

R.Devi[1], D.Siva[2]

[1]*Assistant professor, Department of ComputerScience and Engineering,Sree Sastha Institute of Engineering and Technology , Chennai ,India*

[2] *Assistant professor ,Department of ComputerScience,SRM Istitute of Science and Technology,Ramapuram campus,Chennai-600018 ,Tamilnadu,India*

***Abstract*— Quantifying the climate estimation can lead to insight by considering the environment as a liquid.The changes in the condition of the air and the future condition of the environment can be analysed with parameters of thermodynamics and liquid elements.The more conventional methods of analyzing the climatic conditions are sometimes unreliable.So it leads to an inadequate strategy of weather prediction. But Machine learning is moderately warm to most barometric unsettling influences when contrasted with customary techniques. Another favorable position of machine learning is that it isn't reliant on the physical laws of environmental procedures. The present study examines using machine learning models and regression to predict the weather conditions by several additional parameters and identifying the reliability of each model.**

Keywords*— *machinelearning;regression,python,confusionmatrix;

I. INTRODUCTION

Weather prediction is the task of prediction of the atmosphere at a future time and a given area.

In early days, this has been done through physical equations in which the atmosphere is consider as fluid. The current state of the environment is inspected, and the future state is predicted by solving those equations numerically, but we cannot determine a very accurate weather for more than 10 days and this can be improved with the help pf science and technology. There are numerous kinds of machine learning calculations, which are Linear Regression, Polynomial Regression, Random Forest Regression, Artificial Neural Network and Recurrent Neural Network.[8] These models are prepared dependent on the authentic information gave of any area. Contribution to these models are given, for example, in the event that anticipating temperature, least temperature, mean air weight, greatest temperature, mean dampness, and order for 2 days. In light of this Minimum Temperature and Maximum Temperature of 7 days will be accomplished.

II. PROBLEM STATEMENT

Heavy rainfall can lead to numerous hazards, for instance:flooding, including danger to human life, harm to structures and framework, and loss of products and domesticated animals. avalanches, which can compromise human life, upset transport and interchanges, and cause harm to structures and foundation. Where overwhelming precipitation happens with high breezes, hazard to ranger service crops is high.

For example if we consider an area affected by tropical cyclone the fundamental impacts of tropical cyclone incorporate heavy rain, strong wind, huge tempest floods close landfall, and tornadoes. The devastation from a tropical cyclone, for example, a sea tempest or hurricane, depends for the most part on its power, its size, and its area. Tropical tornados act to evacuate woods shade and additionally change the scene close beach front zones, by moving and reshaping sand ridges and causing broad disintegration along the drift. Indeed, even well inland, overwhelming precipitation can prompt mudslides and avalanches in rugged regions. Their belongings can be detected after some time by concentrate the convergence of the Oxygen-18 isotope inside caverns inside the region of typhoons' ways. So we are providing a better way to get accurate predictions.[1]

As mentioned above, the benefits of identifying important features of mechanical learning, complex data sets, play an important role in forecasting of weather. Since the best results can be achieved with engineering learning algorithms,we should use these techniques to aware people from natural disasters. This is because learning engineering algorithms can provide more accurate results. Apart from this, the results are achieved at a short time and people get enough time to do preparations or to escape from that place .

A. Methodology

The dataset utilized in this arrangement will be gathered from Weather Underground's complementary plan API web benefit. I will utilize the solicitations library to collaborate with the API to pull in climate information since 2015 for the city of Lincoln, Nebraska. When gathered, the information should be process and collected into an organization that is appropriate for information examination, and afterward cleaned.

Then we will concentrate on examining the patterns in the information with the objective of choosing fitting highlights for building a Linear Regression, Polynomial Regression. We will examine the significance of understanding the suppositions vital for utilizing a Linear and Polynomial Regression show and exhibit how to assess the highlights to fabricate a hearty model. This will finish up with a discourse of Linear and Polynomial Regression show testing and approval.

At last we will concentrate on utilizing Neural Networks. I will look at the way toward building a Neural Network show, deciphering the outcomes and, by and large precision between the Linear and Polynomial Regression demonstrate worked earlier and the Neural Network display.

We have a problem statement which comes under the category of Classification. It is a multiclass classification in which the classes given to us are

1. Rain, tempest, and snow into precipitation
2. For the most part shady, foggy, and cloudy into exceptionally shady
3. Scattered mists and somewhat shady into modestly shady
4. Clear as clear

Our aim is to classify the given data into the above given classes. In order to do so, we have to first analyze the data given to us. For analyzing the features, we are using different techniques. The training of model can be done in many ways. It depends on how the data is prepared for further processing. The data can be used directly depending on the situation or the data can be used to form a histogram. After these modifications, we choose a particular model on which we will train our data. This model can be: Linear regression, Logistic Regression, SVM, Neural NetworksDecision Tress, K-Nearest Neighbours etc. Parameter tuning can also be done in order to increase our accuracy.

Once the model is trained, we can test our data by applying our algorithms on the Test Data. With the help of this we can find the learning ability of our algorithm

B .System Design

The record has just been separated into train set and test set. Each information has just been labelled. First we take the trainset organizer.

We will train our model with the help of histograms. The feature so extracted is stored in a histogram. This process is done for every data in the train set. Now we will build the model of our classifiers. The classifiers which we will take into account are Linear Regression, Polynomial Regression, Random Forest and Neural Networks. With the help of our histogram, we will train our model. The most important thing to in this process is to tune thee parameters the accordingly, such that we get the most accurate results.

Once the training is complete, we will take the test set. Now for each data variable of test set, we will extract the features using feature extraction techniques and then compare its values with the values present in the histogram formed by train set. The output is then predicted for each test day. Now in order to calculate accuracy, we will compare the predicted value with the labelled value. The different metrics that we will use are confusion matrix, accuracy score, f1 score etc.

C.Model Development

Our strategy for model improvement is exploratory. The objective of our undertaking is to ensure the conclusion of malignancy with greatest accuracy. This must be accomplished by exploring different avenues regarding distinctive systems from a specific field. We have considered the programmed learning descriptors and algorithms The Machine Learning Algorithms that we are using are:

- Linear Regression
- Polynomial Regression
- Random Forest
- Neural Networks

Subsequently our point is to locate the best mix which will furnish us with greatest precision. Along these lines this task is absolutely test based. In addition parameter tuning is a noteworthy piece of any Machine Learning Algorithm. Regardless of whether the calculation works exceptionally solid in specific conditions, at that point too because of terrible determination of parameters, the precision could be low. In this manner we likewise need to center around the right arrangement of parameters. Hence parameter tuning must be done in whichever show we pick Parameter tuning should either be possible physically or by utilizing the lattice seek technique. Network looking is the procedure in which information is checked with the end goal to discover ideal parameters for some random model. Contingent upon the kind of model that we are utilizing, tuning of specific parameters is vital. Framework seeking applies to a solitary model sort as well as number of models. Network looking can be connected in machine learning with the end goal to ascertain the best parameters for its utilization in some random model. It very well may be computationally greatly costly and may set aside a long opportunity to keep running on the machine. Matrix Search constructs a model on every conceivable parameter mix. At that point it repeats through every parameter blend lastly stores a model for each mix.

III. DATASET

The dataset being used for our prediction models comprises of weather records of the city in focus collected over a period of time using various different parameters like temperature, humidity, atmospheric pressure, and so on as given in figure1. Till date it consists of a record of weather over a period of 20 years (1997-2016). Temperature is a measure of the degree of hotness or coldness of the surroundings. It, like all weather conditions, varies from instance to instance. Similarly, atmospheric pressure and humidity, that plays a vital role in predicting whether an area will receive precipitation or not, is also included in the dataset. Details about fog and dew point are included in the dataset as well, as they only contribute to improving the accuracy of the predictions made by the prediction models. All the data gathered in the dataset was collected from Wunderground that has an easy to use API, which makes data collection all the more simpler.

Date and time	Precipitation	Atmospheric pressure	Humidity	Fog	Temperature
18-11-1996 11:00	0	934	2	0	18
18-11-1996	0	936	3	0	19
12:00					
18-11-1996 1:00	0	932	4	0	20

Fig 1.Sample Dataset with parameters

A.Test Metrics

Scikit-Learn library in Python is a free machine learning library for Python. It highlights different algorithms like support vector machine, random forests, and k-neighbors, and it likewise underpins Python numerical and scientific libraries like NumPy and SciPy.The library has functions like accuracy_score(), RandomForestRegressor() and many other very useful regression functions that enable us to make accurate predictions.

B.Test Setup

The test process is already in-built in our system.

The testing process taking place just after the model is trained. After the completion of the training process, we analyze each data entry in the test set. In order to analyze each entry, we use descriptors to extract features.

Now we compare these feature values with the feature values which were initially retained using the train set. The comparison is done according to the Machine Learning model used and finally the output for each entry is received. Since each data entry is already labeled, we can compute accuracy by comparing the predicted value with the received values.

IV. RESULT

The results of the implementation of the project are demonstrated below.

A.Multiple Linear Regression

This regression model has high variance, hence turned out to be the least accurate model. Given below is a snapshot of the actual result from the project implementation of multiple linear regression.

TABLE I: ACTUAL VS PREDICTED VALUES

S.No	Actual Value	Predicted Value
1.	0	0.0459157
2.	0	0.0423579
3.	0	0.0474239
4.	1	0.8654278
5.	0	0.0325468
6.	0	0.0023542
7.	0	0.1236582

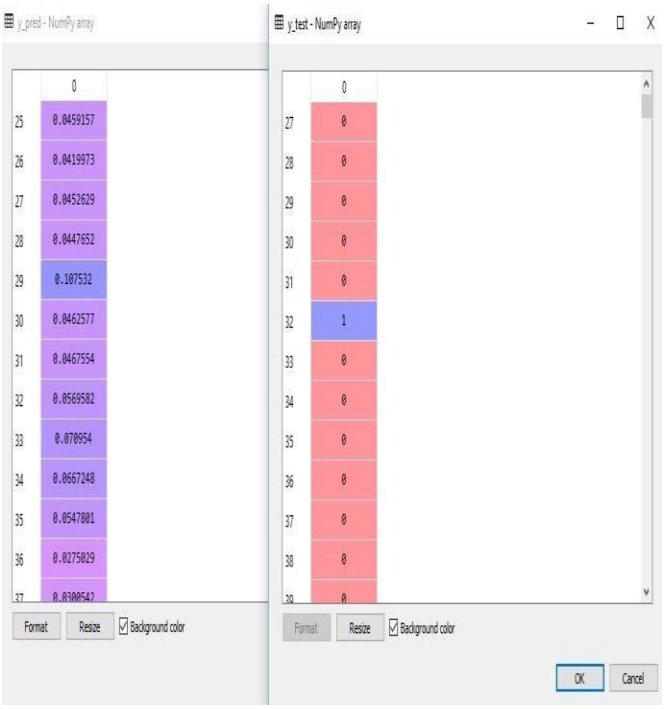

Fig 2. Predicted and actual values using Multiple linear regression.

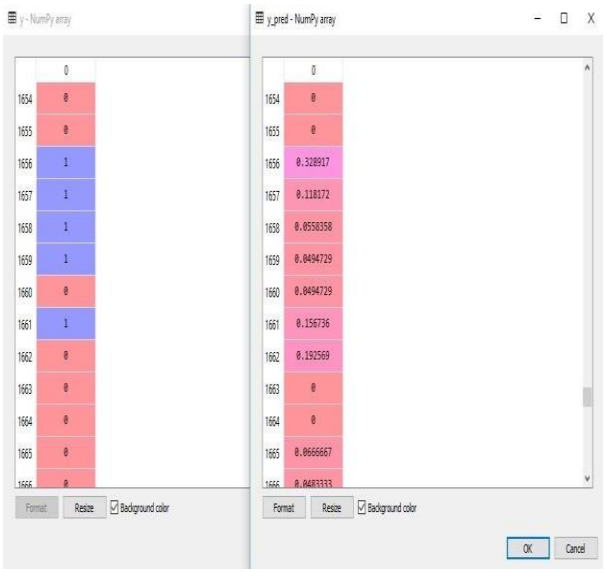

Fig 3. Actual and predicted values using Random Forest Regression

TABLE II: ACTUAL VS PREDICTED VALUES FROM RANDOM FOREST REGRESSION

S.No	Actual Values	Predicted Values
1	0	0
2	0	0
3	1	0.324546
4	1	0.121647
5	1	0.564642
6	1	0.195487
7	0	0

B.Polynomial Linear Regression

This regression model is much more accurate than the multiple linear regression model, hence it made predictions that were more closer to the actual result that linear regression.

Below is a snapshot of its implementation in the code, and the result it displayed.

TABLE III: ACTUAL VS PREDICTED VALUES FROM POLYNOMIAL REGRESSION

S.no	Actual Value	Predicted Value
1	0	0.0214568
2	0	0.2669756
3	1	0.8165476
4	0	0.0165959
5	1	0.6326548
6	1	0.7656548
7	0	0.0436597

C.Logistic regression

This regression technique is used to classify the predictions. Here, I used binary logistic regression. The result of this regression technique was justified using the confusion matrix. The accuracy was 97%, as per the confusion matrix. Below is a snapshot of the same.

TABLE IV: CONFUSION MATRIX FOR LOGISTIC REGRESSION

	Precision	Recall	F1-score	Support
0	0.97	1.00	0.99	24613
1	0.00	0.00	0.00	635
Avg/total	0.95	0.97	0.96	25248

V. CONCLUSION

All the machine learning models: linear regression, various linear regression, polynomial linear regression, logistic regression, random forest regression and Artificial neural systems were beaten by expert climate determining apparatuses, in spite of the fact that the error in their execution reduced significantly for later days, demonstrating that over longer timeframes, our models may beat genius professional ones.

Linear regression demonstrated to be a low predisposition, high fluctuation model though polynomial regression demonstrated to be a high predisposition, low difference model. Linear regression is naturally a high difference model as it is unsteady to outliers, so one approach to improve the linear regression model is by gathering of more information. Practical regression, however, was high predisposition, demonstrating that the decision of model was poor, and that its predictions can't be improved by further accumulation of information. This predisposition could be expected to the structure decision to estimate climate dependent on the climate of the previous two days, which might be too short to even think about capturing slants in climate that practical regression requires. On the off chance that the figure were rather founded on the climate of the past four or five days, the predisposition of the practical regression model could probably be decreased. In any case, this would require significantly more calculation time alongside retraining of the weight vector w, so this will be conceded to future work.

Coming to the Logistic Regression, it proved vital to classify whether a day would be rainy or not. Its significance was proven by the accuracy of the results, where it predicted the classification right, more often than not.

Talking about Random Forest Regression, it proves to be the most accurate regression model. Likely so, it is the most popular regression model used, since it is highly accurate and versatile

ANN with backpropagation utilizes an iterative procedure of preparing where, it more than once contrasts the watched yield and focused on yield and computes the mistake.[7] This blunder is utilized to rearrange the estimations of loads and predisposition to show signs of improvement yield. Subsequently this technique attempts to limit the blunder. In this manner, Artificial Neural system with Backpropagation algorithm is by all accounts most fitting strategy for estimating climate precisely.

ACKNOWLEDGMENT

I hereby thank my husband and all my friends who helped me in publishing this paper.

REFERENCES

[1] Mohammad Wahiduzzaman, Eric C. J. Oliver, Simon J Wotherspoon, Neil J. Holbrook, "A climatological model of North Indian Ocean tropical cyclone genesis, tracks and landfall".

[2] Jinglin Du, Yayun Liu , Yanan Yu and Weilan Yan, "A Prediction of Precipitation Data Based on Support Vector Machine and Particle Swarm Optimization (PSO-SVM) Algorithms"

[3] Prashant Kumar, Atul K. Varma, " Atmospheric and Oceanic Sciences Group,EPSA, Space Applications Centre (ISRO), Ahmedabad, IndiaAssimilation of INSAT-3D hydro- estimator method retrieved rainfall for short-range weather prediction"

[4] Prashant Kumar, C. M. Kishtawal, P. K. Pal, "Impact of ECMWF, NCEP, and NCMRWF global model analysis on the WRF model forecast over IndianRegion"

[5] H. Vathsala, Shashidhar G. KoolagudiPrediction, "Model for peninsular Indian summer monsoon rainfall using data mining and statistical approaches"

[6] Mark Holmstrom, Dylan Liu, Christopher Vo, "Machine Learning applied to weather forecasting", Stanford University, 2016.

[7] Ayman M. Abdalla, Iyad H. Ghaith, Abdelfatah A. Tamimi, "Deep Learning Weather Forecasting Techniques: Literature Survey", Information Technology (ICIT) 2021 International Conference on, pp. 622-626, 2021

[8] Nitin Singh, Saurabh Chaturvedi, Shamim Akhter, "Weather Forecasting Using Machine Learning Algorithm**",** IEEE, International Conference on Signal Processing and Communication (ICSC)2019**.**

AUTHORS

First Author – R.Devi, ME(CSE), Sree Sastha Institute of Engineering and Technology,devi.cse@ssiet.in.
Second Author – D.Siva, MTech(CSE), SRM Institute of Engineering andTechnology,d.siva885@gmail.com