ZKG INTERNATIONAL ISSN: 2366-1313

Forecasting of Rainfall using Machine learning based Multiple Linear Regressions Model

[1]MUVVALA VENKATA NARESH, [2] M. NARESH

[1]PG Scholar, Dept. of MCA, Newton's Institute of Engineering, Guntur, (A.P)

[2]Associate Professor, Dept. of CSE, Newton's Institute of Engineering, Guntur, (A.P)

Abstract: *Scientists in the field of meteorology are always on the lookout for new techniques to learn about Earth's atmosphere and create reliable forecasting tools. Weather forecasting techniques have evolved throughout time. As an alternative to traditional approaches, machine learning strategies have become more popular in recent years for predicting the weather. One of the most important aspects of the weather system that has an obvious impact on the agricultural and biological industries is the rate of precipitation. In order to forecast the precipitation rate (PRCP), or rainfall rate, in Khartoum state, the authors of this work will design a multiple linear regression model. Conditions including wind speed, humidity, and dew point are taken into account. The information for this study came from the National Climatic Data Centre's online database. The Python code for this model, which makes use of ANNs, was created using the Pytorch package. Mean square error was calculated between the training and test data to determine the model's performance. When the same quantity of data is utilised in both the training and testing phases, the findings reveal that the average of the mean square error has decreased by 85% throughout test time. When more data is available during the test phase than during the training phase, this percentage lowers to 59%.*

Keywords: *Weather Prediction, rainfall, Linear Regression, Machine Learning. Artificial Neural Networks.*

I. INTRODUCTION

Weather forecasting is the science and art of determining what the weather will be like at a given place at a future time [1]. From the beginning of recorded history, people have been curious about the future by predicting the weather. There are a number of techniques used to predict the weather, and they all have their advantages and disadvantages. When attempting to predict the future state of the atmosphere, there are three crucial steps that must be taken first: collecting as much atmospheric data as possible; understanding the data and its inter-relation to determine the behaviour of the atmosphere; and using it

ISSN: 2366-1313

in numerical models. In recent years, scientists have favoured using machine learning algorithms for weather prediction since they do not need an in-depth knowledge of the atmospheric process. The goal of machine learning (ML) is to train a computer to execute a specified activity with little human input, with performance improvements attributable only to further training. There are three distinct categories of learning strategies: supervised learning (which relies on labelled data), unsupervised learning, and reinforcement. Feature extraction and utilisation are important processes in all machine learning approaches.

for use in classification and regression [3]; derived characteristics for other methods. It is possible to compensate for the complexity of the meteorological physics model by using machine learning methods applied to weather forecasting. They were advised to use multiple linear regression as their supervised learning approach [1] rather than unsupervised learning or reinforcement learning due to the availability of a metrological data set. Machine learning makes use of several kinds of regression, including linear regression, logistic regression, and polynomial regression. For prediction, linear regression is the most common and straightforward technique [4].The goal of this study is to create a multiple linear regression model to forecast the rainfall rate in Khartoum state, which is influenced by a wide range of factors. The remaining sections of this work will be structured as follows. In Section II, we offer a high-level overview of the relevant literature; in Section III, we detail our resources and methodology; and in Section IV, we provide our findings. The essay finishes with Section V.

II. LITERATURE SURVEY

Gupta (1981) detailed the computational procedures needed in the reconstruction of lost flows in addition to describing synthesis methods and a complex model of monthly streamflow simulation. One of the largest drainage systems in Liberia, located on the west coast of Africa, is the John River, and this model was used to produce the probable sequences of monthly flows at the two locations along this river.

The Thomas-Firing model was found to best preserve the mean, standard deviation, and lag-one correlation of historic streamflow's by Vedula and Reddy (1981), who compared the applicability of different streamflow generating models using historical data of monthly streamflow's into Hemavathy and Krishnasagar reservoirs in the upper Cauvery River Basin.

ISSN: 2366-1313

For ARMA model identification of hydrologic time series, Salas and Obeysekera (1982) showed how to employ the GPAF together with Grey et al.'s (1978) R- and S-functions. The suggested identification method was shown with many instances. Annual streamflow statistics from the Saint Lawrence River and the Nile River were used in practical examples.

Some streamflow sequences in West Bengal were examined by Gorantiwar and Majumdar (1988), who then simulated the stream flows using a first-order autoregressive model.

The ARMA model was created by Mujumdar and Nagesh Kumar (1990) to predict monthly and ten-day stream flows for Indian rivers. The most accurate prediction and data-representation models were chosen. According to the results, the AR (1) model is superior for predicting future streamflow's.

By carefully selecting variables to keep a high degree of similarity in the monthly equations, Garen (1992) was able to achieve consistency of projections from month to month. Regression approaches offer large gains in prediction accuracy over previous procedures, as shown by results for the South Fork Boise River at Anderson Ranch dam and other basins in the West, without compromising month-to-month forecast consistency.

Daily and hourly streamflow predictions in Korea's Pyung Chang River basin were made using ANNs and ARMA models by Kang et al. (1993). The authors found that ANNs are effective tools for predicting stream flows after looking at various three-layered designs.

Using data from stream gauging stations 30 km upstream and 20 km downstream of the gauging site, Kurunanithi et al. (1994) approximated stream flows at an ungauged location on the Huron River in Michigan. The authors compared the efficiency of ANNs to an empirical power law relationship calculated from the log-transformed values of the streams' flows. The ANN inputs and cascade correlation method were just the raw data. The greatest deviations from the empirical regression equations occurred at the highest stream flows. While both techniques performed well in predicting low flows, ANNs were shown to be superior at predicting large flow occurrences. It was said that ANNs may adjust their level of complexity to account for shifts in time series data, such as those seen in archives of past streamflow measurements. It was also discovered that the performance of the regression

ISSN: 2366-1313

approach deteriorated, but the performance of the ANN was unaffected, when an additional ganging station was included that had little or no influence on stream flows at the gauging site. The authors argued that ANNs are more likely to succeed when dealing with noisy data at the inputs.

III. RELATED WORK

This section summarises the various studies that used machine learning on weather records to make forecasts. These studies made forecasts of meteorological conditions including temperature and precipitation using neural networks or linear regression. Information on each publication is provided below.

For the purpose of studying the weather and making temperature predictions, E. B. Abrahamsen and O. M. Brastein [1] created a Python API to read meteorological data and ANN models in Tensor Flow. The research relied on two meteorological variables to account for rain and heat. The maximum and lowest temperatures for the next seven days were predicted using data from the previous two days using a linear regression model and a variant of a functional regression model by Mark Holmstrom, Dylan Liu, and Christopher [2]. Machine learning methods like the linear regression model and the normal equation optimisation approach were used to forecast the weather based on a small number of inputs by Sanyam Gupta, Indhumathi K, and Govind Singhal [4]. In order to predict weather variables (highest temperature, rainfall, and wind speed), Folorunsho Olaiya [7] employed Artificial Neural Network and Decision Tree algorithms with meteorological data (2000-2009) for the city of Ibadan, Nigeria. Rainfall was predicted with fewer error % by S. Prabakaran and colleagues [8] using a modified linear regression model by adding percentage to the input data.

Maximum and lowest temperatures, relative humidity, and rain type were the four meteorological variables that Paras and Sanjay Mathur [9] predicted using the Multiple Linear Regression (MLR) model. Two approaches, Autocorrelation Function (ACF) and projected error, have been used to predict rainfall by Wanie M. Ridwana, b, Michelle Sapitang et al. [10]. Bayesian Linear Regression, Boosted Decision Tree Regression, Decision Forest Regression, and Neural Network Regression were all used in the two approaches, along with daily, weekly, ten-day, and monthly time frames. Boosted Decision Tree Regression was found to have the highest coefficient of determination and thus be the best regression developed for M1. However, in M2, the overall model performance gave a

ISSN: 2366-1313

good result of each category except for 10-days when using Boosted Decision Tree Regression or Decision Forest Regression.

In this study, the authors forecast rainfall rate using a set of meteorological variables with a high degree of correlation, and they do it using a method of multiple linear regression rather than a single one.

IV. METHODOLOGY

A. Sampling and Data Collection

The weather information for this research came from the National Climatic Data Centre's online database.

Since the data came via an exchange conducted by the World Metrology Organisation (WMO) [11], it may be used without cost in any academic setting.

The data used in this study was collected from the meteorological station in Khartoum, Republic of Sudan, and split into two parts: one set was used to train the model (from 1990 to 2005), while the other set was used to evaluate its performance (from 2006 to 2020). Picked data set with TMP for 10 characteristics.

As indicated in Table I, the dependent variable rainfall (precipitation) PRCP rate is correlated with the maximum temperature MX, minimum temperature MN, dew point WP, sea level pressure SLP, station pressure STP, mean visibility VS, and wind speed WSP. Below

TABLE I. METEOROLOGICAL DATA USED AS INDEPENDENT VARIABLES OF THIS MODEL

Predictor Variable	Abbreviations
mean temperature	TMP
maximum temperature	MX
minimum temperature	MN
Dew point	WP
sea level pressure	SLP
station pressure	STP
mean visibility	VS
wind speed	WSP

B. Scrubbing and Reformatting the Data

The Excel programme was used for the manual data cleansing procedure. The process consisted of four stages: gaining familiarity with the data and its relationships, eliminating irrelevant variables, addressing gaps and outliers, and cleaning the data for easier manipulation.

The data in Table II (a, b) below are a subset of what will be utilised during the training phase. Table II(a) displays the model's first five independent parameters. and the remaining parameters are shown in table II (b).

TABLE II(A). SAMPLES OF METEOROLOGICAL DATA USED AT TRAINING PHASE

	TMP (x1)	WP (x2)	SLP (x3)	STP (x4)	VS (x5)
0	69.8	39.8	1012.8	967.4	2.1
1	69.3	36.4	1014.7	970.0	7.6
2	70.1	33.4	1012.8	968.1	10.3
3	73.5	38.0	1012.6	968.3	9.8
...	...	..	...	...	...

V. RESULT

ISSN: 2366-1313

Table IV. SHOWS THE Actual AND PREDICTED PRCP VALUES DURING TRAINING PHASE, THE TABLE SHOWS THAT THE DIFFERENCE BETWEEN ACTUAL AND PREDICTED VALUES WAS LARGE ESPECIALLY AT THE BEGINNING OF THE TRAINING

	Actual	predicted
0	0.0	-202.818268
1	0.0	-184.136261
2	0.0	-161.219086
3	0.0	-137.441727
4	0.0	-118.243645
..	...	...
95	0.0	-139.289917
96	0.0	-134.579208
97	0.0	-131.520737
98	0.0	-129.182846
99	0.0	-125.826492

Fig. 2 shows the learning curve of the model, in which the orange line and the blue line represents the actual and predicted values of the PRCP, respectively.

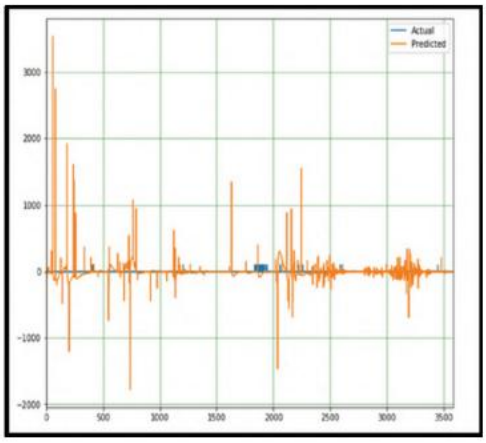

Fig. 2 Learning Curve of the Model the orange line and the blue line represents the actual and predicted values of the PRCP

VI. Conclusion

Precipitation rates (rainfall rates) for Khartoum state were predicted using a multivariate linear regression model using a selection of meteorological factors as independents.

The average, highest, and lowest temperatures, as well as the dew point, sea level pressure, station pressure, average visibility, and average wind speed are all included. Mean squared error was estimated as the average of the errors seen throughout the training and testing phases.

During testing, it was discovered that the acquired findings demonstrate a considerable reduction in the mean square error between the actual and anticipated values of the rainy precipitation rate (PRCP). When the test data is proportional to the training data, it is 85%, but drops to 59% when additional test data is utilised.

More study is required to explain this decrease.

It might mean, for instance, that the used model requires more data during its training phase.

REFERENCES

1 E. Abrahamsen, O. M. Bastien, and B. Lie, “Machine Learning in Python for Weather Forecast based on Freely Available Weather Data,” Proceedings of the 59th Conference on immolation and Modelling (SIMS 59), 26-28 September 2018, Oslo Metropolitan University, Norway, 2018.

ISSN: 2366-1313

2 M. Holmstrom, D. Liu, and C. Vo, "Machine Learning Applied to Weather Forecasting," Dec. 2016.

3 J. Remona, M. Lakshmi, R. Abbas, and M. Raziullha, "Rainfall Prediction using Regression Model," International Journal of Recent Technology and Engineering (IJRTE), vol. 8, no. 2S3, Jul. 2019.

4 S. Gupta, I. K, and G. Singhal, "Weather Prediction Using Normal Equation Method and Linear regression Techniques," International Journal Computer Science and Information Technologies, vol. 7, no. 3, pp. 1490-1493, 2016.

5 S. Gupta, I. K, and G. Singhal, "Weather Prediction Using Normal Equation Method and Linear regression Techniques," International Journal of Computer Science and Information Technologies, vol. 7, no. 3, pp. 1490-1493, 2016.

6 C. Bishop, Pattern recognition and machine learning. Springer Verlag, 2006.

7 F. Olaiya and A. B. Adeyemo, "Application of Data Mining Techniques in Weather Prediction and Climate Change Studies," International Journal of Information Engineering and Electronic Business, vol. 4, no. 1, pp. 51-59, 2012.

8 S. Prabakara, P. N. Kumar, and P. S. M. Tarun, "RAINFALL PREDICTION USING MODIFIED LINEAR REGRESSION," ARPN Journal of Engineering and Applied Sciences, vol. 12, no. 12, Jun. 2017.

9 S. M. Paras, "A Simple Weather Forecasting Model Using Mathematical Regression," Indian Research Journal of Extension Education, vol. 12, pp. 161-168, 2016.

10 W. M. Ridwan, M. Sapitang, A. Aziz, K. F. Kushiar, A. N. Ahmed, and A. El-Shafie, "Rainfall forecasting model using machine learning methods: Case study Terengganu, Malaysia," Ain Shams Engineering Journal, 2020

11 Climate Data Online - Select Area. [Online]. Available: https://www7.ncdc.noaa.gov/CDO/cdoselect.cmd. [Accessed: 21-Jan-2021]

12 Prasadu Peddi (2022), A Hybrid-Method Neighbor-Node DetectionArchitecture for Wireless Sensor Networks, ADVANCED INFORMATION TECHNOLOGY JOURNAL ISSN 1879-8136, volume XV, issue II.