PAPER 62 – DESCRIPTIVE ANALYSIS OF NASA WEATHER DATA USING MACHINE LEARNING ALGORITHM

Aneke Chikezie Samuel[1*], Philip Michael Asuquo[2], Simeon Ozuomba[3], Okechukwu Collins Eze[4], Chetachukwu P Ogbu[5]

[1,2,3]Department of Computer Engineering, University of Uyo, Akwa Ibom State, Nigeria
[4]Washington University in St Louis, Missouri, USA
[5]Verizon Business Group, Richardson, Texas, USA
*Email: anekechikezies@uniuyo.edu.ng

ABSTRACT
Descriptive Analysis of National Aeronautics and Space Administration (NASA) weather data using a machine learning algorithm is presented in this paper. NASA has provided a range of weather data capable of aiding in the prediction of various weather conditions. The paper looked at the prediction analysis of NASA weather data using 70128 entries from 2010 to 2017. Nine weather data were introduced and later scaled down to three namely Relative humidity, Temperature and Pressure. For temperature, the correlation coefficient between T2M and Next T2M, T2MWET and Next T2M, and RH2M and Next T2M is 0.96, 0.82 and -0.72 respectively. For Relative Humidity, the correlation coefficient between R2M and Next RH2M, Next RH2M and T2M is 0.96 and -0,7 respectively. For Pressure, the correlation coefficient between R2M and Next RH2M, Next RH2M and T2M is 0.96 and -0.7 respectively. This paper presented some machine learning algorithms to analyze NASA weather data in determining which algorithm is better suited for weather forecasting or prediction. Ridge Regression, Lasso Regression, Decision Tree Regression, and XGboost algorithm were used to validate their performances. The result showed that the XGboost algorithm performed better than the other algorithms with a root mean square error value (RMSE) of 0.26901and coefficient determination (R^2) value of 0.93385

KEYWORDS: Weather Forecasting, Machine learning, Algorithm, Performance, Temperature Prediction

1. INTRODUCTION

The study of weather is as old as creation itself, and it has always had a significant influence on people's environments and behaviour. The effects of disasters caused by weather and climate change can be lowered with prior knowledge of these factors. In this age of global warming, getting the most recent weather information and taking the necessary measures in light of recent environmental catastrophes brought on by climate change have become top priorities. The model that can forecast various weather data for any place on Earth, including those without weather stations, weather satellite coverage, or other weather measuring equipment, has been created. A lot of devices have been developed to aid in weather acquisition [1], [2].
Researchers have improved the accuracy of forecasts for a variety of important meteorological occurrences using machine-learning techniques. [3], [4], [5], [6]. By combining (i.e., ensemble) the results of different machine learning techniques in various domains, some researchers have recently tried to increase the performance of their predicting methods. [8], [9], [10].

The need for accurate meteorological forecasts is on the rise because some sectors need accurate weather prediction like the agriculture sector, maritime sector and the aviation sector. Correct weather prediction is very important in forecasting dangerous weather conditions to put up preventive measures. With regard to the measurement of atmospheric weather parameters, this research evaluates the forecasting accuracy of NASA data and also looked at the prediction analysis of NASA weather data.

To give clear guidance for predictive and prescriptive analyses, NASA weather data were examined through exploratory analysis and four machine learning algorithms were used for the analysis

2. THEORETICAL ANALYSIS

2.1 Data Cleaning and Pre-Processing

Although errors are frequently unavoidable, data cleaning might help to reduce them. An erroneous study conclusion might occur if these flaws are not corrected. Errors and inconsistencies are examples of dirty data. These data may result from any step in the study process, such as a defective research design, use of the wrong measurement tools, or incorrect data entry. While filthy data are faulty in one or more ways, clean data meet some criteria for high quality.

The NASA dataset was obtained in.csv format for analysis. The dataset is made up of nine (9) features which are Temperature, Dew Temperature, Wet Temperature, Specific Humidity, Relative Humidity, Precipitation, Surface Pressure, Wind Speed, and Wind direction. The dataset obtained from NASA does not require much pre-processing. The Year, Month, and Day columns were combined into a single column called Date.

	YEAR	MO	DY	HR	T2M	T2MDEW	T2MWET	QV2M	RH2M	PRECTOTCORR	PS	WS10M	WD10M
0	2010	1	1	0:00:00	24.37	21.81	23.09	16.30	85.56	0.0	100.66	1.79	240.97
1	2010	1	1	1:00:00	24.14	21.77	22.95	16.30	86.56	0.0	100.62	1.65	243.92
2	2010	1	1	2:00:00	23.99	21.75	22.87	16.24	87.19	0.0	100.58	1.47	250.12
3	2010	1	1	3:00:00	23.87	21.80	22.83	16.30	88.12	0.0	100.58	1.31	255.88
4	2010	1	1	4:00:00	23.74	21.85	22.80	16.36	89.12	0.0	100.60	1.19	263.58
...	...	...	...	...	...	...	...	...	...	...	...	...	...
70123	2017	12	31	19:00:00	25.40	20.30	22.85	14.83	73.25	0.0	100.65	1.75	298.86
70124	2017	12	31	20:00:00	25.11	19.84	22.48	14.40	72.56	0.0	100.73	1.73	295.75
70125	2017	12	31	21:00:00	24.87	19.41	22.14	14.04	71.62	0.0	100.78	1.66	296.93
70126	2017	12	31	22:00:00	24.74	18.89	21.82	13.55	69.88	0.0	100.78	1.48	303.61
70127	2017	12	31	23:00:00	24.62	18.39	21.51	13.18	68.25	0.0	100.75	1.28	317.23

70128 rows × 13 columns

Figure 1: Preview of the first five and last five of the NASA dataset

The table below shows the weather parameters used for the research, their description, and the international system of units.

Table1: Dataset Features Used for the Research

Weather Parameter	Feature Description	SI Unit
Date	This is the date each data was logged	second
Temperature	Normal Temperature value	^oC
Dew Temperature	Dew Temperature value	^oC
Wet Temperature	Wet Temperature value	^oC
Specific Humidity	Specific Humidity value	g/kg
Relative Humidity	Relative Humidity value	%
Precipitation	Precipitation value	m/Hr
Surface Pressure	Surface Pressure value	kPa
Wind Speed	Wind Speed value	m/s
Wind Direction	Wind Direction value	^{o}

3. METHODOLOGY

3.1 Descriptive Analysis

This part looks at the data visualization using the available dataset. Python libraries like Pandas, Matplotlib, Plotly, Seaborn, and NumPy were used for this analysis. Visualization is used to show the time series plots of the weather parameters for the NASA dataset.

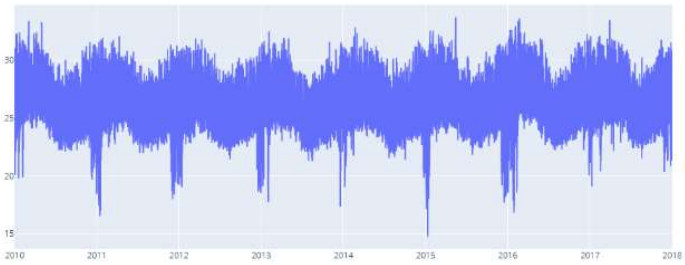

Figure 2: Time series plot of the T2M variable

A time series chart sometimes referred to as a times series graph or time series plot, is a data visualization tool that shows data points at progressively shorter time intervals. A measurement of a quantity and a time is represented by each point on the graph. This was done for all the variables used in this report with indications of signs of seasonality. Usually, there are annual changes in seasons precipitated by changes in the beginning, middle, and end of the season.

For temperature, there is an increase at the beginning of the year, a drop in the middle of the year, an increase from midyear to the third quarter, and a slight drop at the end of the year.

3.2 Predictive Analysis

Intelligent algorithms are used in predictive analysis to produce predictions based on patterns, trends, and correlations in historical data. Because predictive analytics is probabilistic, it cannot predict what will happen in the future but rather predictions of what might be.

A target variable *Next T2M* is created by shifting the value of the *T2M* variable by one step backwards

The Seaborn library is used to show the Pearson Correlation heatmap of the features of the dataset. Pearson Correlation is given as:

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[n\Sigma x^2 - (\Sigma x)^2][n\Sigma y^2 - (\Sigma y)^2]}}$$

Where,
n = total number of observations,
x = first variable
y = second variable
r = Pearson correlation value.

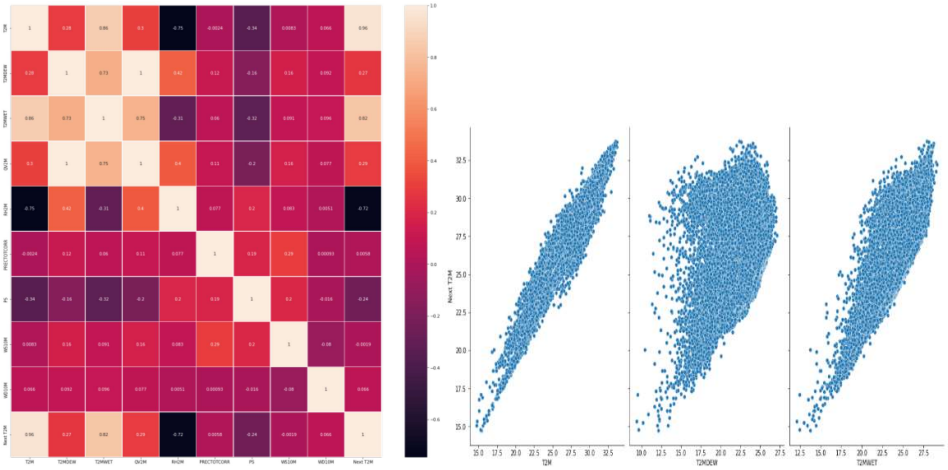

Figure 3: (a) Pearson Correlation Heatmap of Features in the Dataset (b) Scatter Plots of the Next T2M against T2M, T2MDEW, and T2MWET

Exploratory Data analysis shows that T2M and T2MWET are highly positively correlated with the Next T2M variable. Also, RH2M is highly negatively correlated with the Next T2M variable. The Pearson correlation coefficient between T2M and Next T2M is 0.96. The Pearson correlation coefficient between T2MWET and Next T2M is 0.82. The Pearson correlation coefficient between RH2M and Next T2M is -0.72.

A target variable Next RH2M is created by shifting the value of the RH2M variable by one step backwards.

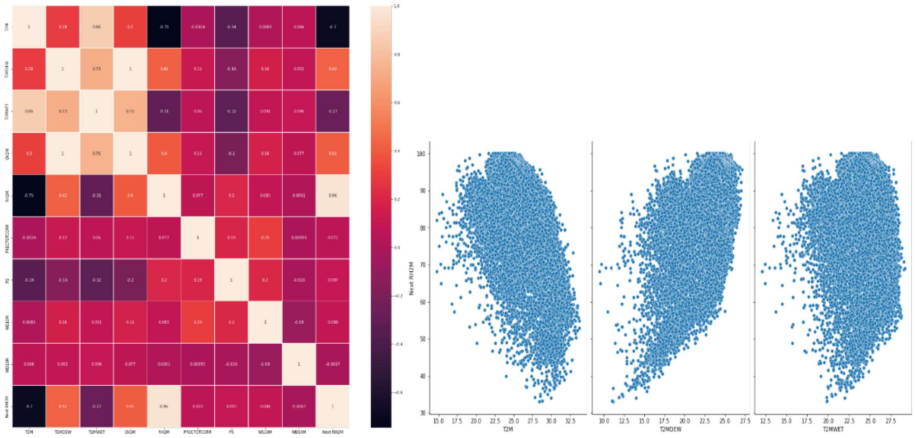

Figure 4: (a) Pearson Correlation Heatmap of Features in the Dataset (b)Scatter Plots of the Next RH2M against T2M, T2MDEW, T2MWET

Exploratory Data analysis shows that RH2M is highly positively correlated with the Next RH2M variable. Also, T2M is highly negatively correlated with the Next RH2M variable.
The Pearson correlation coefficient between R2M and Next RH2M is 0.96. The Pearson correlation coefficient between Next RH2M and T2M is -0.7.

The Scatter plot below was used for visualization to see which features correlate with the future temperature value.

3.3 Splitting Data into Training Set and Test Set

Before splitting the dataset, standardization is applied to it. This is to ensure all the variables are of the same variance.

The formula for standardization is:

$$x = x - \text{mean}(x)/\text{standard deviation}(x)$$

The dataset is split in the ratio of 75:25. The dataset is not shuffled because we are dealing with time-series data. We use 75% of the data for training, while we use 25% of the data for testing.

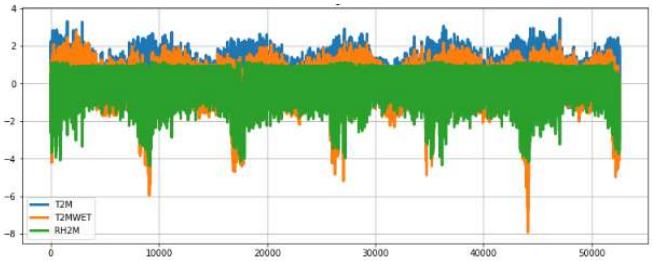

Figure 5: Line Plot of the T2M, T2MWET & RH2M features after applying standardization in the training dataset

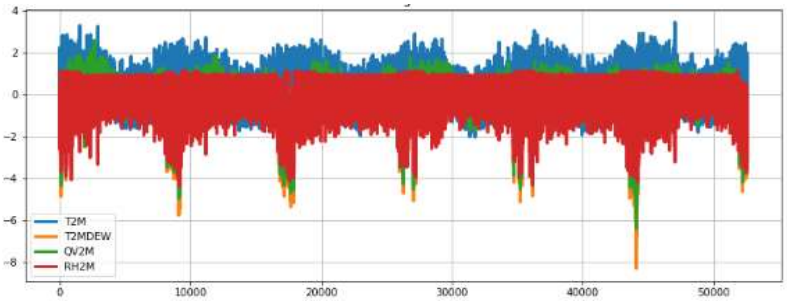

Figure 6: Line Plot of the T2M, T2MDEW, QV2M & RH2M features after applying standardization in the training dataset.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(P_i - O_i)^2}{n}} \qquad R^2 = 1 - \frac{\sum(y_i - \hat{y})^2}{\sum(y_i - \bar{y})^2}$$

Where, $\check{Y}$ - Predictive value of Y
$\bar{Y}$ - Mean value of Y

To model the training set for the prediction of future temperature, the following algorithms were used for the training; Ridge Regression, Lasso Regression, Decision Tree Regression, and XGBoostRegressor. This was also done for Relative Humidity and pressure.

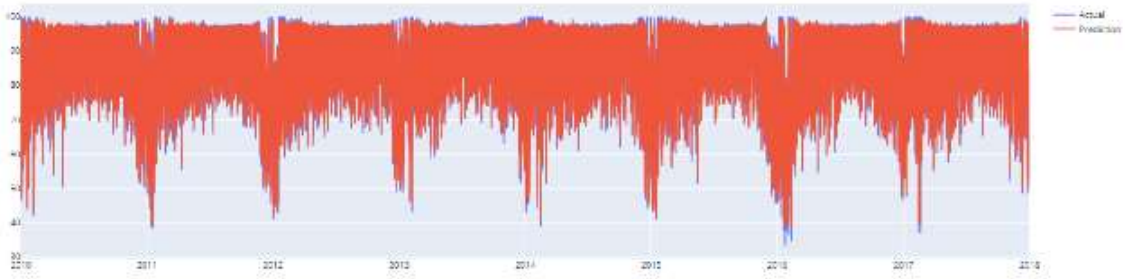

Figure 7: Line plot of actual result vs predicted result of the XGBoost regression model.

The table below shows the R Squared score and the Root Mean Squared Error on the test set of each algorithm.

Table 2: R Squared score and the Root Mean Squared Error of each Algorithm

Model	**R Squared**	**RMSE**
Ridge Regression	0.92599	0.28455
Lasso Regression	0.92580	0.28491
Decision Tree Regression	0.92519	0.28609
XGBoost Regression	0.93385	0.26901

4. RESULTS AND DISCUSSION

For temperature, the correlation coefficient between T2M and Next T2M, T2MWET and Next T2M, RH2M, and Next T2M is 0.96, 0.82, and -0.72 respectively. The analysis shows that T2M and T2MWET are highly positively correlated with the Next T2M variable. Also, RH2M is highly negatively correlated with the Next T2M variable. For Relative Humidity, the correlation coefficient between R2M and Next RH2M, Next RH2M, and T2M is 0.96 and -0.7 respectively. The analysis shows that RH2M is highly positively correlated with the Next RH2M variable. Also, T2M is highly negatively correlated with the Next RH2M variable.

For Surface pressure, the correlation coefficient between R2M and Next RH2M, Next RH2M, and T2M is 0.96 and -0.7 respectively. The analysis shows that RH2M is highly positively correlated with the Next RH2M variable. Also, T2M is highly negatively correlated with the Next RH2M variable.
The Scatter plot was used for visualization to see which features correlate with the future temperature value. From the exploratory data analysis, the three features used to predict the next hourly temperature are T2M, T2MWET, and RH2M.
The dataset was split and, standardization was applied to it. This is to ensure all the variables are of the same variance. The dataset is split in the ratio of 75:25. The dataset is not shuffled because we are dealing with time-series data. We use 75% of the data for training, while we use 25% of the data for testing. This was done for temperature, Relative humidity, and pressure.

Table 3: Predictive Algorithm deployed and their performance evaluation

Feature	Algorithm	R Squared	RMSE
Temperature	Ridge Regression	0.92321	0.27913
	Lasso Regression	0.92325	0.27908
	Decision Tree Regression	0.92149	0.28226
	XGBoost Regression	0.92308	0.27938
Relative Humidity	Ridge Regression	0.92599	0.28455
	Lasso Regression	0.92580	0.28491
	Decision Tree Regression	0.92519	0.28609
	XGBoost Regression	0.93385	0.26901
Pressure	Ridge Regression	0.92599	0.28455
	Lasso Regression	0.92580	0.28491
	Decision Tree Regression	0.92519	0.28609
	XGBoost Regression	0.93385	0.26901

5. CONCLUSION

Most previous studies that used machine learning approaches only applied single machine learning methods [5], [6]. In contrast, in this study, descriptive and predictive analysis of the NASA dataset was performed using four prediction models (i.e. Ridge Regression, Lasso Regression, Decision Tree Regression, and XGBoost regression) for performance analysis of NASA data. It reported a root mean square error (RMSE) value of between 0.27913 and 0.28455, 0.27908 and 0.28491, 0.28226 and 0.28609, 0.27938 and 0.26901 for Ridge Regression, Lasso Regression, Decision Tree Regression, and XGBoost regression respectively. The correlation analysis reported a coefficient of determination (R2) value of between 0.92321 and 0.92599, 0.92325 and 0.92580, 0.92149 and 0.92519, 0.92308 and 0.93385 for Ridge Regression, Lasso Regression, Decision Tree Regression, and XGBoost regression respectively.
XGBoost regression performed better than other algorithms with a root mean square error (RMSE) value of 0.26901 and a coefficient of determination (R2) value of 0.93385. For further improvements, a more predictive algorithm can be used to check for performance.

ACKNOWLEDGEMENT

I want to sincerely acknowledge the contributions of my co-authors in making this research a success. I want to thank the National Aeronautics and Space Administration for the dataset used in this research.

REFERENCES

[1] Aneke C.S, Asuquo P.M, and Ozuomba S. (2023)" Design and Implementation of a Microcontroller-based Weather Acquisition Device, "Journal of Engineering Research and Reports, Vol. 24, no. 6, Page 35-46, DOI: 10.9734/JERR/2023/v24i6823
[2] K. E. Ukhurebor, I. C. Abiodun, S. O. Azi, I. Otete and L. E. Obogai (2017) "A Cost Effective Weather Monitoring Device" Archives of Current Research International, Vol 7 no.4: page 1-9, DOI: 10.9734/ACRI/2017/32885

[3] Dongjin Cho, CheolheeYoo, JunghoIm, Dong-Hyun Cha. "Comparative Assessment of Various Machine Learning-Based Bias Correction Methods for Numerical Weather Prediction Model Forecasts of Extreme Air Temperatures in Urban Areas " Earth and Space Science
Vol 7, no. 4:2020 https://doi.org/10.1029/2019EA000740
[4] Kim, M., Im, J., Park, H., Park, S., Lee, M.-I., &Ahn, M.-H. (2017). "Detection of tropical overshooting cloud tops using himawari-8 imagery" Remote Sensing, Vol 9 no.7, Page 685.
[5] Lee, S., Han, H., Im, J., Jang, E., & Lee, M.-I. (2017). Detection of deterministic and probabilistic convection initiation using Himawari-8 advanced Himawari imager data. Atmospheric Measurement Techniques, 10 (5), 1859–1874
[6] Sim, S., Im, J., Park, S., Park, H., Ahn, M., & Chan, P. W. (2018). Icing detection over East Asia from geostationary satellite data using machine learning approaches. Remote Sensing,10 (4), 631
[7] Yoo, C., Im, J., Park, S., &Quackenbush, L. J. (2018). Estimation of daily maximum and minimum air temperatures in urban landscapes using MODIS time series satellite data. ISPRS Journal of Photogrammetry and Remote Sensing, 137, 149–162.
[8] Chou, J.-S, & Pham, A.-D. (2013). Enhanced artificial intelligence for ensemble approach to predicting high-performance concrete compressive strength. Construction and Building Materials, 49, 554–563.
[9] Healey, S. P., Cohen, W. B., Yang, Z., Kenneth Brewer, C., Brooks, E. B., Gorelick, N., et al. (2018). Mapping forest change using stacked generalization: An ensemble approach. Remote Sensing of Environment, 204, 717–728. https://doi.org/10.1016/j.rse.2017.09.02
[10] Ren, Y., Zhang, L., &Suganthan, P. N. (2016). Ensemble classification and regression-recent developments, applications, and future directions. IEEE Computational Intelligence Magazine,11 (1), 41–53