Climatic Temperature Trends: Analyzing Past Data and Predicting Future Temperature

DR. S. K. SINGH[1], HARSH PANCHAL[2], RISHIKA RAO[3]

[1] *H.O.D (IT), Department of Information Technology, Thakur College of Science and Commerce, Thakur Village, Kandivali (East), Mumbai Maharashtra, India*

[2, 3] *PG Student, Department of Information Technology, Thakur College of Science and Commerce, Thakur Village, Kandivali (East), Mumbai Maharashtra, India*

Abstract— Machine Learning (ML) techniques for time series prediction are becoming increasingly accurate and helpful, particularly in considering climate change, such as weather forecasting, climate research, and environmental planning, accurate temperature prediction is essential. This work utilizes a 40-year historical dataset and three machine learning algorithms, Linear Regression (LR), Random Forest (RF) and Polynomial Regression to give a thorough method to temperature prediction. To increase the accuracy of temperature forecasts, it is important to examine how well various algorithms work on a large dataset. A large historical dataset covering four decades is first gathered and pre-processed for the study. This dataset includes temperature records as well as data on a wide range of meteorological factors, including humidity, wind speed, and precipitation. The dataset is optimized for modeling using feature selection and engineering methods. The first step in predicting temperature is to use linear regression, polynomial Regression - Single modeling approach and use random forest algorithm – Ensemble modeling and for the same and then compare these algorithms. The data has been collected via NASA climate datasets. It consists of 40 years data from 1980 to 2020 in month wise format.

Indexed Terms— Machine learning, Temperature prediction, Ensemble Models, Single Models.

I. INTRODUCTION

Weather forecasting, climate research, agriculture, and energy management are just a few of the scientific, industrial, and societal pursuits that depend critically on temperature prediction. Making wise decisions in these domains requires the capacity to precisely predict temperature variations. Machine learning algorithms have become effective tools for improving temperature prediction accuracy in recent years, providing new insights and capabilities.

The objective of this project is to create and assess temperature prediction models using Linear Regression (LR), Random Forest (RF) and Polynomial Regression (PR) three different machine learning algorithms. The use of a thorough 40-year historical dataset, which offers a rich source of temperature and meteorological data spanning several decades, distinguishes this research from others. This expanded dataset enables a more thorough investigation of temperature trends and patterns as well as a thorough evaluation of the effectiveness of the chosen algorithms over a longer time period.

Linear regression for temperature prediction is a statistical modelling technique used to forecast temperature values based on historical or measured data. It assumes that there is a linear relationship between one or more independent variables (predictors) and the temperature, which is the dependent variable. This method can be employed for short-term or long-term temperature predictions, depending on the available data and the specific application. Random Forest is a powerful machine learning algorithm that can be applied to temperature prediction tasks. It is particularly well-suited for complex and nonlinear relationships between input variables (predictors) and the target variable (temperature). In polynomial regression, you create higher-degree polynomial features by raising the original features to various powers and use these new features to model non-linear relationships between variables. Temperature often exhibits non-linear behaviour, with daily and seasonal fluctuations. Polynomial regression addresses this by introducing polynomial terms of the independent variable(s) into the model.

The goal of the project is to show how machine learning may improve temperature forecast by giving stakeholders in many fields more precise and useful data. We may more effectively comprehend, foresee, and reduce the effects of temperature changes on our environment and society by utilizing the power of machine learning. Additionally, this research supports continuing initiatives to use cutting- edge technology to address the urgent problems brought on by climate change and extreme weather.

II. LITERATURE REVIEW

This study Evaluates of Machine Learning Methods Application in Temperature Prediction in which six machine learning (ML) methods, including LR, kNN, SVM, ANN, Random Forest, and AdB, were used to predict temperature. Data from 1/1/1980 to 12/31/2014 were split into training (60%) and test (40%) sets. Test results revealed that all methods except SVM performed well, with SVM having the lowest R2 value of 0.5560. Different performance indicators, including R.S.R., NSE, M.A.E., Index of Agreement, PBIAS (%), and RMSE, led to varying conclusions on the best method. However, three out of six indicators identified ANN as the most suitable method[Babak Azari 1*, Khairul Hassan1, Joel Pierce1, Saman Ebrahimi2][1]. This paper predicts Atmospheric Temperature using Support Vector Machines(SVM).The study compared the performance of two machine learning models, MLP and SVM, using Mean Square Error (MSE) as the evaluation metric.Here are the MSE values: MLP: MSE ranged from 8.07 to 10.2, depending on the order,SVM: MSE ranged from 7.07 to 7.56, with no significant improvement beyond order [Y.Radhika and M.Shashi][2]. The focus of this model is to Forecast daily mean, maximum and minimum temperature time series by three artificial neural network methods. The results of the method are summarized, while the FFBP method performed quite well, compared with the other ANN methods and the MLR method, providing good agreement with the forecasts in the daily mean temperature(tmean) forecasting in terms of the selected performance criteria, RMSE and IA, the RBF method provided superior performance in predicting the temperature[B. Ustaoglu,H.K.Cigizoglu and M.Karacaa[3]. This study is performed for Accurate long-term air temperature prediction with Machine Learning models and data reduction techniques.Average daily mean August air temperature in Córdoba ranges from 30 degrees to 25 degrees with a mean 27.66 °C and a variance 1.02 °C. Two extreme temperature events are spotted in the test period considered, one in the famous 2003 summer, the other in years 2017– 2018[Fister , J. Pérez-Aracil , C. Peláez-Rodríguez , J. Del Ser , S. Salcedo-Sanz][4]. This paper explains the Weather - Temperature Pattern Prediction and Anomaly Identification using Artificial neural network. In this case, ANN achieves minimum MSE of around -0.0705 degree Celsius for 2500 test cases, 0.8941 degree Celsius for 2250 test cases and -1.035 degree Celsius for 2000 test cases out of 13908 test cases. Using the same it can be induced that approximately 48.53% test cases have MSE +/- 0.95 degree Celsius[Vishwajeet Pattanaik, Shweta Suran, Himani Tyagi][5]. The goal of this research is to Predict global patterns of long-term climate change from short-term simulations using machine learning algorithms.This evaluates the performance of the two different machine learning methods (Ridge, GPR) by benchmarking them against a traditional pattern scaling approach, often used for estimatingfuture patterns of climate change. The latter relies on multiplying the long-term response pattern for the 2xCO2 scenario by the relative magnitude of global mean response for eachindividual climate forcer[L. A. Mansfield, P. J. Nowack, M. Kasoar, R. G. Everitt, W. J. Collins][6]. This research induces on an Analysis of Climate Change Based on Machine Learning and an Endoreversible Model.They monitored various machine learning models to check the climate factors and to predicts the climate factors using climate models extensively[Sebastián Vázquez-Ramírez, Miguel Torres-Ruiz,Rolando Quintero, Kwok Tai Chui Carlos Guzmán Sánchez-Mejorada][7]

III. METHODOLOGY

1. Single Model ML Methods

Single models contain only one method in their procedure, unlike ensemble models, which contain multiple methods combined. In this study, two different single models were used. This section discusses the single models used in this study.

Linear regression: Linear regression utilizes a linear curve based on the best fit criterion to estimate the trend of a dataset. Linear regression can be used with single or multiple variable datasets. With a single variable, the method minimizes a single objective function based on that variable. The training set of data is used to create a linear regression model. The input features and the goal temperature will be related linearly by the model.

The general form of a linear regression equation is:
$y = \beta_0 + \beta_1 * x + \varepsilon$

Use metrics like Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared (R2) to measure the models performance on the test dataset and determine its correctness.

Polynomial regression: Polynomial regression is a powerful technique in statistics and machine learning used to model relationships between variables that are not adequately explained by simple linear models. In this method, the relationship between a dependent variable and one or more independent variables is expressed as a polynomial equation of a specified degree.
The general form of a polynomial regression equation is:
$y = \beta_0 + \beta_1 * x + \beta_2 * (x^2) + \beta_3 * (x^3) + ... + \beta_k * (x^k) + \varepsilon$

2. Ensemble Model ML Methods

Ensemble methods are a class of machine learning techniques that combine the predictions of multiple individual models to produce a more accurate and robust overall prediction. The idea behind ensemble methods is that by aggregating the results of multiple models, the weaknesses of individual models can be mitigated, and their strengths can be amplified, leading to better predictive performance.

Random Forest: Random Forest is a popular ensemble learning technique used in machine learning for both classification and regression tasks. It is a powerful and versatile algorithm that improves the accuracy and robustness of individual decision trees by combining the results of multiple trees.

3. Performance metrics

To compare the predictive capability of different ML algorithms, the following evaluation metrics are used, which have been used in many studies to evaluate the results in various fields:
(1) Mean Absolute Error (MAE),
(2) Root Mean Absolute Error (RMAE) and
(3) Root Mean Square Error (RMSE)

The acronym MAE means "Mean Absolute Error." It is a frequently used statistic for assessing the effectiveness of regression models, particularly when you want to comprehend the average size of errors the model makes. The average absolute difference between the expected and actual values is measured by MAE.

$$MAE = \frac{\sum_{i=1}^{n}\left|Y_i^{obs} - Y_i^{sim}\right|}{n}$$

RMSE is a common evaluation metric used in regression tasks to measure the average magnitude of errors between the predicted values and the actual values. It is calculated as the square root of the mean of the squared differences between the predicted and actual values.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}\left(Y_i^{obs} - Y_i^{sim}\right)^2}{n}}$$

RMAE is not a standard or recognized abbreviation in machine learning or statistics. However, it's possible that you are referring to "RMAE" as a variation or abbreviation for "Root Mean Absolute Error (RMAE)," which would be the square root of the Mean Absolute Error (MAE).

$$RMAE = \sqrt{MAE}$$

IV. RESULT & DISCUSSION

In this study, meteorological data were extracted for the city of Mumbai based on multiple meteorological features, including air temperature, relative humidity, wind speed, wind direction. Data covered almost 21 years, from 1/1/1980 to 31/12/2021. We performed a comparative analysis of machine learning models, Linear Regression, Random Forest Classifier and polynomial Regression to predict temperature.

Figure.1 shows the schematic of the measures pertaining to the highest and lowest temperatures ever noted or seen in a given setting, like a certain area, time frame, or dataset. The greatest and lowest temperatures are often represented by numerical values in this data, which can be utilized for a variety of tasks like statistical analysis, weather forecasting, and climate study. It makes it possible for analysts, meteorologists, and academics to comprehend trends and fluctuations in temperature over time or between various sites.

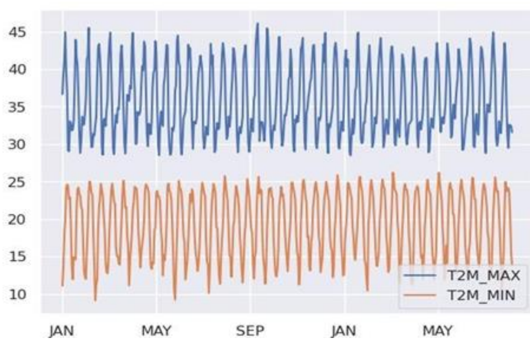

Figure 1: Data concerning the maximum and minimum temperature

PERFORMANCE METRICS	ML MODELS		
	LINEAR REGRESSION	POLYNOMIAL REGRESSION	RANDOM FOREST
RSME	2.33	0.43	0.66
MAE	1.78	0.37	0.48
RMAE	1.34	1.40	0.69
ACC	93.50	98.35	98.15

Table 1: Assessments of the ML model's performance on the test dataset in terms of their ratings.

In Table.1 we have compared the models' performance based on the respective metrics. For regression tasks, lower values of RMSE, MAE, and RMAE are desirable, while for classification tasks, you want higher accuracy values. The goal was to evaluate the performance of these models in terms of accuracy and error metrics to determine which model provides better temperature predictions.

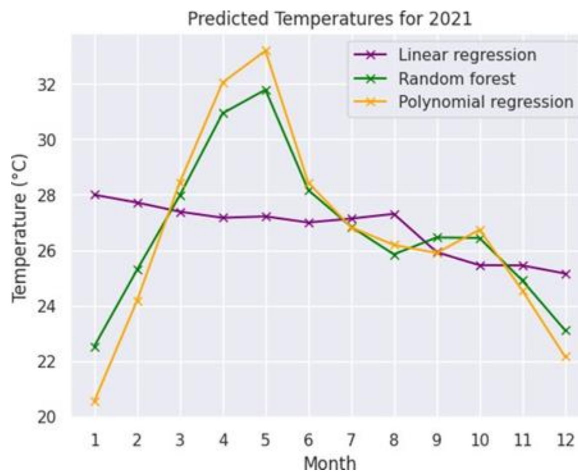

Figure 2: Graphical representation to compare the temperature predictions produced by three distinct models.

The above Figure.2 provides a comprehensive comparison of predicted temperature values generated by three different models. Each line on the graph corresponds to the temperature forecasts produced by a specific model. This visual representation allows us to assess how each model performs in predicting temperature, enabling a clear and detailed examination of their respective strengths and weaknesses. By examining the trends and variations among the lines, we can gain insights into which model offers the most accurate and reliable temperature predictions, aiding in the selection of the best forecasting approach for the given context or task.

CONCLUSION

The meteorological data for the city of Mumbai, several features were taken into consideration. These features were air temperature, relative humidity, wind direction, and speed. The dataset is collected from NASA. Data covered almost 21 years, from 1980 to

2020. This study was used to compare the ability of various machine learning techniques to forecast air temperature in order to investigate climate change. The choice of the best model depends on the specific characteristics of the temperature prediction problem.

Linear Regression can serve as a baseline, while Random Forest and Polynomial Regression offer more flexibility in capturing non-linear patterns. The effectiveness of each model would depend on the dataset. Therefore, these ML methods were used to predict the temperature. To help decide the best trained method, the three other performance indicators were included: M.A.E., R.M.A.E and RMSE.

REFERENCES

[1] Azari B, Hassan K, Pierce J, Ebrahimi S. Evaluation of machine learning methods application in temperature prediction. Environ Eng. 2022;8:1-2.

[2] Radhika Y, Shashi M. Atmospheric temperature prediction using support vector machines. International journal of computer theory and engineering. 2009 Apr 1;1(1):55.

[3] Smith BA, McClendon RW, Hoogenboom G. Improving air temperature prediction with artificial neuralnetworks. International Journal of Computational Intelligence. 2006;3(3):179-86.

[4] Salcedo-Sanz S, Deo RC, Carro-Calvo L, Saavedra-Moreno B. Monthly prediction of air temperature in Australia and New Zealand with machine learning algorithms. Theoretical and applied climatology. 2016 Jul;125:13-25.

[5] Nury AH, Hasan K, Alam MJ. Comparative study of wavelet-ARIMA and wavelet-ANN models for temperature time series data in northeastern Bangladesh. Journal of King Saud University-Science. 2017Jan 1;29(1):47-61.

[6] Shikoun N, El-Bolok H, Ismail MA. Climate change prediction using data mining. International Journal of Intelligent and Cooperative Information Systems. 2005;5(1):365-79.

[7] Steinbach M, Tan PN, Kumar V, Potter C, Klooster S, Torregrosa A. Data mining for the discovery of ocean climate indices. InProc of the Fifth Workshop on Scientific Data Mining 2002 Apr.

[8] Mansfield LA, Nowack PJ, Kasoar M, Everitt RG, Collins WJ, Voulgarakis A. Predicting global patternsof long-term climate change from short-term simulations using machine learning. npj Climate and Atmospheric Science. 2020 Nov 19;3(1):44.

[9] Anjali T, Chandini K, Anoop K, Lajish VL. Temperature prediction using machine learning approaches.In2019 2nd International Conference on Intelligent Computing, Instrumentation and Control Technologies (ICICICT) 2019 Jul 5 (Vol. 1, pp. 1264-1268). IEEE.

[10] Tyagi H, Suran S, Pattanaik V. Weather-temperature pattern prediction and anomaly identification using artificial neural network.

[11] International Journal of Computer applications 2016 Apr:975:8887.

[12] Evaluation Abhishek K, Singh MP, Ghosh S, Anand A. Weather forecasting model using artificial neural network. Procedia Technology. 2012 Jan 1:4:311-8.

[13] Buszta A, Mazurkiewicz J. Climate changes prediction system based on weather big data visualization, International Conference on Dependability and Complex Systems 2015 jun 29 (pp.75-86). Cham: Springer International Publishing.

[14] Doblas-Reyes FJ, Garcia-Serrano J, Lienert F,Biescas AP, Rodrigues LR. Seasonal climate Predictability and forecasting: status and prospects. Wiley Interdisciplinary Reviews: Climate Change. 2013 Jul;4(4):245-68.

[15] Fister D, Pérez Aracil J, Peláez-Rodríguez C, Del Ser J, Salcedo-Sanz S. Accurate long-term air temperature prediction with Machine Learning models and data reduction techniques. Applied Soft Computing. 2023 Mar 1;136:110118.

[16] Ustaoglu B, Cigizoglu HK, Karaca M. Forecast of daily mean, maximum and minimum temperature time series by three artificial neural network methods. Meteorological Applications: A journal of forecasting, practical applications, training techniques and modelling. 2008 Dec;15(4):431-45.