iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 11 **Issue:** I **Month of publication:** January 2023

DOI: https://doi.org/10.22214/ijraset.2023.48291

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue I Jan 2023- Available at www.ijraset.com

Machine Learning Methods to Weather Forecasting to Predict Apparent Temperature

Saurabh Nayak[1], Kamal Shrivastav[2]
[1]*HFCL Pvt. Ltd. Gurgaun*
[2]*SKU, Chhatarpur*

Abstract: ***Due to the impact that it has on human life across the globe, weather forecasting has attracted the attention of researchers from a wide variety of research communities. Many researchers have been motivated to investigate hidden hierarchical patterns in large volumes of weather datasets by the recent emergence of deep learning techniques over the past decade. This motivation has been influenced by these methods, as well as by the ubiquitous accessibility of enormous amounts of weather observing data and the development of information and technology. In this study, we explore use of the machine learning techniques to apparent temperature weather forecasting. We use the dark sky apis data set for forecasting and assessment, and this study uses a random forests, decision tree, linear regression, or polynomial regression. Finally, performance is evaluated using metrics like mean square error & r-square.***
Keywords: Weather Forecasting, temperature prediction, machine learning, mean square error, r-square.

I. INTRODUCTION

Predicting the atmosphere's state for a given time and place is the goal of weather forecasting. Just a few of the atmospheric components that can fluctuate throughout the day and produce what we refer to as "the weather" include temperature, wind speed and direction humidity, sunlight, cloud cover, and precipitation. Through the Internet of Things, the weather can now be predicted almost instantly (IoT)[1]. Low-cost deployment of several weather sensors allows for widespread collection of weather data from a variety of sites. Next, highly accurate predictions can be made using machine learning algorithms applied to the amassed data. The Infrastructure as a Service (IaaS) paradigm of commercial cloud platforms works well not only for Internet of Things (IoT) but also for applications like weather forecasting that only need access to computational resources on an as-needed basis[2]. It may be more cost-effective to rent high-end servers rather than buy and maintain specialised hardware in order to run applications with unpredictable compute demands. Users of cloud services only have to pay for the time that their resources are actually put to use. Numerous studies have focused on IoT, machine learning or cloud computing to develop cutting-edge weather forecasting systems in recent years. Following that, a succinct overview of a couple of such plans is provided [3].

Massive, humongous amounts of data in structured, semi-structured, and unstructured forms make up Big Data. That's why it's so challenging to handle, monitor, and store such data. As of now astonishing sorts of mechanisms, approaches or procedures are there to cope with Big Data. In this study, we use data mining [4]with machine learning as one of these methods to keep track of weather-related data and make predictions about the weather's trajectory and potential outcomes. In this framework, we propose making use of data mining and the systematic retrieval of data by machine learning in advance of climatological and meteorological predictions[5]. There has been a recent uptick in extreme weather events, pollution, and their corresponding reactions in India. Ranchers frequently face challenges in the horticulture field due to erratic weather patterns. Forecasting the weather is based solely on the distribution of naturally occurring particles in the atmosphere, such as ozone, nitrogen dioxide, carbon dioxide, sulfur dioxide, and others. In [6]study, we zero in on Delhi as our focus. To lessen these responses up to a certain degree there are several ways and computations through that we can forecast the climate on the basis of given information. Data mining employing a machine learning approach is applied as a component of Weather prediction process. Throughout each stage of human existence, favorable weather conditions are an absolute necessity[7]. Because of this, climate prediction is becoming increasingly important in many industries, including agriculture, science, and the management of food security crises. There was never a reliable source of information about the weather back in the old days. Many challenges existed in the food supply chain, industrial sector, and agricultural business sector at the time. But there are many ways we can find out about the weather now that we live in a highly advanced era. To determine the current state of the weather, scientists are using machine learning techniques like support vector machines and linear regression as well as Big Data and its Eco-System.

The basic objective of meteorology is to make predictions about the state of the earth's atmosphere at a given time and location. This study shows how real-time data from a sensor network can be used to forecast precipitation. During the last 30 years, numerical hydrodynamic approaches have progressed both in terms of research and development and in terms of their operational application in weather prediction[8]. The growth of computational power, the refinement of applicable mathematical models, as well as the extension of a foundational body of meteorological data all contributed to this breakthrough. Both the present and the future of the atmosphere can be predicted by numerically resolving the equations of fluid dynamics or thermodynamics[9]. However, the system of normal differential equations governing this physical model is unsteady under fluctuations, and uncertainties in initial earth's atmosphere measurements and an incomplete capacity to adapt to new atmospheric processes severely restrict the reliability of weather forecasts beyond the first ten days[10].

II. LITERATURE REVIEW

This study compares 24 machine learning methods for probabilistic afternoon power forecast using NWP with those examined by Markovics et al. over a two-year period using datasets with a 15-min resolution for 16 Pv in Hungary. We also evaluate the importance of hyperparameter adjustments and the significance of picking the appropriate predictors. The least accurate models were found to be multilayer perceptrons or kernel ridge regression, with a forecast skills score of up to 44.6% under persistence. The 13.1% reduction in the root mean square errors (RMSE) obtained by adding Solar position angle and statistically processed irradiation value system as the input of a learning models to the basic NWP data makes it evident how important predictor selection is. Hyperparameter tweaking is essential for maximising the capability of the models, especially for much less robust models that will be vulnerable to under and generalisation with adequate tuning. The RMSE of the best forecasts is, on average, 13.9% lower than that of the baseline approach, which was linear regression. The effectiveness of computer vision and with scant data is further demonstrated by the fact that the RMSE of power forecasts based just on daily average irradiance projections and the Sun positioning angles is only 1.5% larger than the best-case scenario. The results of this work can be used by researchers and practitioners to create more precise NWP-based Pv power forecasting techniques [6].

Birkelund et al. gives real-time measurements of wind speed and direction in a numerical weather forecasting a novel machine learning-based method for measuring and forecasting predicted errors, sometimes known as residuals. An Arctic wind turbines verifies the effectiveness of the framework. The residuals still contain significant meteorological information that can be accurately predicted with machine learning, as shown by a comparison of the four prediction learning algorithms. It is also demonstrated that the linear some of works well enough for multi-timestep forecasts of summary, East-West, East-West, or North-South North-South wind speeds residuals. The National Weather Service's wind model can be improved using the forecasts of wind power forecasting systems [8].

Fowdur et al. A practical collaborative machine learning-based weather forecasting system that incorporates information from several sources has been proposed. Across four different locations in Mauritius, we use five distinct methods of machine learning in this study for forecast temperatures, wind velocity, prevailing winds, pressures, humidity, and cloudiness. Data was retrieved from a mobile edge device and a desktop edge device using the OpenWeather API. The JSON file containing the data were stored in two databases: an IBM Cloudant database and a MySQL database that was hosted locally. The data from edge device was processed by both an on-premises servers and a servlet housed on IBM's cloud platform. Five machine learning algorithms—Multiple Linear Regression (MLP), Multi Polynomials Regression (MPR), K-Nearest Neighbors (KNN), Multi - layer perceptron (MLP), & Convolutional Neural Network—were assessed using collaborative and non-collaborative methods (CNN). The experiments' findings demonstrated that the Mean Absolute Percentage Error (MAPE) was reduced by 5% for cooperative versus non-collaborative regression schemes, and that Multiple Polynomial Regression (MLR) automated system outperformed them all with errors ranging from 0.009% per 9% for the different weather parameters. Overall, the findings demonstrated that using several predictor sites in a collaborative weather forecasting setting has the potential to improve the accuracy of forecasts made by machine learning systems[9].

Pavuluri et al. The decision tree, K-NN, and random forest method calculations by demonstrating the greatest accuracy result of the these three algorithms present a straightforward method for weather prediction in future years by mining historical data. Predicting the weather is a vital skill in many fields, and in this study, we accomplish it by looking at how a region's average temperature fluctuates over time. The algorithms compute standard deviations, medians, confidence intervals, probabilities, and visualise the variation between the three techniques, among other things. In the end, we may use the algorithms used in this study to foretell whether the temperature will rise or fall, and whether or not rain will fall. The weather in a certain region is the sole focus of this data set, which consists of a few objects (year, month, temperature, forecasted values, and so on)[11].

Siddula et al. Utilizing the restricted Boltzmann machine (RBM), create the best configuration of solar-wind energy systems (RBM). The RBM scales the best placement technique by taking into consideration a wide range of factors, enabling adequate size and positioning to maximise power generation from wind and solar systems. The MATLAB simulation results show better accuracy coupled with lowered mean average error & calculation time for the multi-objective criteria from both solar and wind farms. The results show that the RBM outperforms other approaches in terms of enhanced rate for identifying the appropriate placement at such a lower cost & calculation duration of the less than 2 ms [12].

III. PROPOSED METHODOLOGY

This study proposed weather temperature forecasting as well as prediction based on Szeged city data from 2006 to 2016 data, applied preprocessing tasks to improve better feature extraction and removed missing values, removed outliers and performed exploratory data analysis, found skewness for temperatures and humidity and plot probability stats, performed scaling, and finally implemented a machine learning model, evaluated its performance, and predicted weather conditions[13].

A. Data Collection

The implementation of this work makes use of the Darksky API for data extraction. The data contain the dates 01-01-2006 to 01-01-2017, possess 96500 samples to zero mismatch as well as missing values, and the csv contains 11 features. These features are as follows: time; summary; precipitation type; temperature; apprarent temperature; humidity; windspeed; wind bearing; visibility; loud cover; and pressure.

B. Data Preprocessing

Data enhancement involves the utilization of three primary variables, namely the likely temperature, the apparent temperature, and the humidity content of 96429, the discovery of statistical analyses such as count, mean, standard deviation, minimums, and maximums, the search for missing values, which resulted in the discovery of no missing values, and the final step, the detection of an outlier and its subsequent removal[12].

C. Splitting Data

Data splitting is the division of a set of data into two in subsets. When data is divided into two parts, usually one of these parts is used to analyse and test the data, the other is utilised to train a model. Dataset splitting is a crucial step in the data mining process, especially when creating models that are based on data. We divided the data in our study into a 70:30 ratio. There are 91882 train samples and 64317 test samples.

D. Detect Skewness

Data is considered skewed when it produces an uneven or skewed curve when plotted on a graph. A data set that has a normal distribution will have a graph that is symmetrical & shaped like a bell when it is plotted. Skewed data, on the other hand, will always have a "tail" along either of the graph's sides. In our study we didn’t find skewness in temperature but skewness present in humidity.

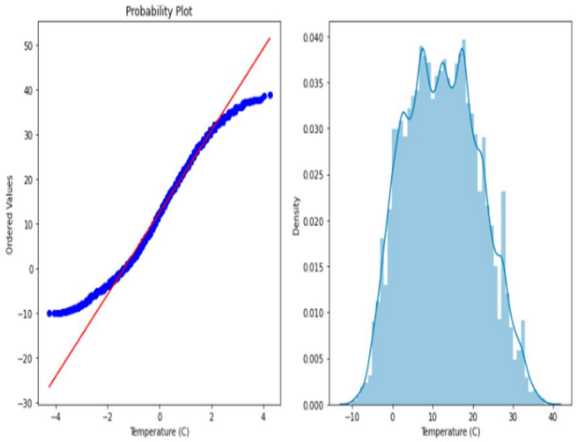


Figure 1. Probability and Histogram of Temperature

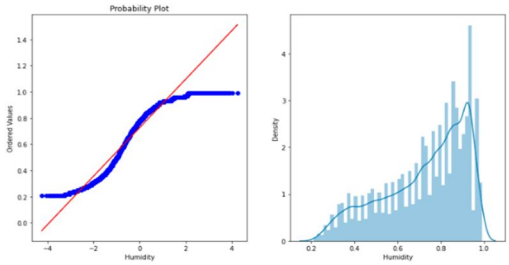


Figure 2. Probability and Histogram of Humidity

E. Standard Scaling

Data splitting is the division of a set of data into two in subsets. When data is divided into two parts, usually one of these parts is used to analyse and test the data, the other is utilised to train a model. Dataset splitting is a crucial step in the data mining process, especially when creating models that are based on data. We divided the data in our study into a 70:30 ratio. There are 91882 train samples and 64317 test samples.

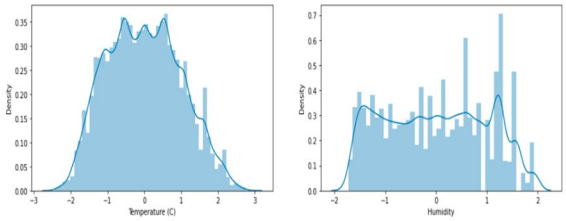


Figure 3. After Scaling of Temperature and Humidity

F. Machine Learning Modeling

The likely linear regression, polynomials regression, decision tree, & random forest regressors are just a few of the four machine learning models we constructed for this mode.

1) *Linear Regressor:* Linear regression is an approach that models the relationship between such a scalar response and one or many explanatory factors in a linear manner. The process is known is simple linear regression when there is only one explanatory variable, and as linear regression model when there are many. Contrary to multivariate linear regression, that predicts several dependent variables which are correlated with one another as opposed to a single scalar variable, this phrase should not be mistaken with that phrase.

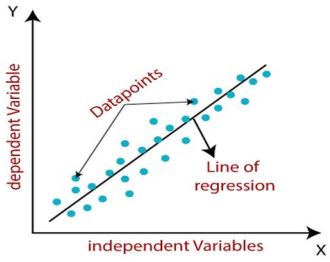


Figure 4. Linear Regression Plan

2) *Polynomial Regression:* A type is linear regression known as "polynomial regression" models the relationship between an independent (x) or a dependent variable (y) as the highest degree polynomial in x. The term "multivariate regression" is occasionally used to describe this kind of regression study. The polynomial regression technique can be used to describe a nonlinear relationship between the value in x and the corresponding conditional means of y, denoted by the symbols E(y |x). From a statistical standpoint, polynomial regression is a linear estimate issue even though it may fits a nonlinear to a data. This is thus because the inferred uncertain variables from the data have constant values for the regress functions E(y | x). Polynomial regression is therefore seen as a subset to multiple linear regression.

3) *Decision Tree:* In decision tree regression, properties of an item are analyzed, and a model is trained to fit inside the structure of a tree. This model is then used to forecast data that will occur in the future and produce meaningful output continuously. Continuous output signifies that the outcome or result is not discrete, which means that it is not represented only by a discrete, defined set of values or values. In other words, continuous output signifies that the outcome or result is not discrete.

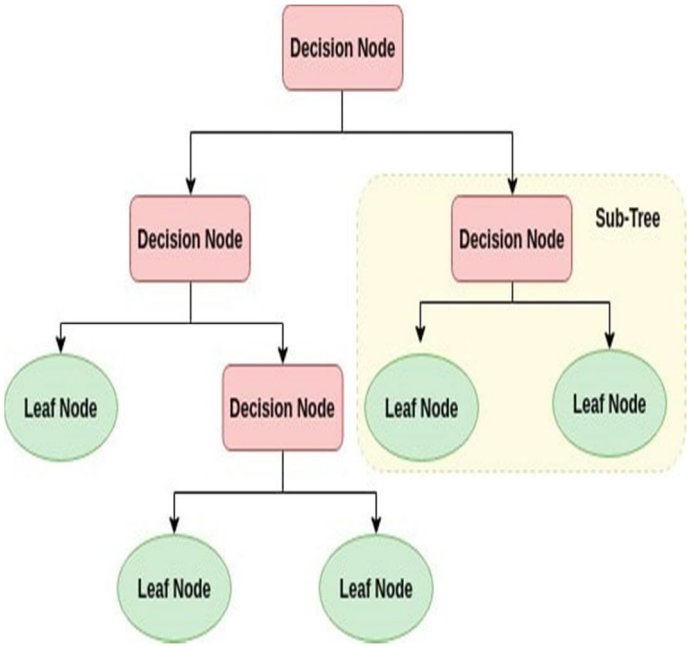


Figure 5. Decision Tree

4) *Random Forest:* A predictor from a random forest. A random forest is a particular kind of meta estimation that makes use of averaging to improve prediction accuracy or reduce overfitting. It accomplishes it by fitting a variety of classifying decision trees to various subsamples of a dataset.

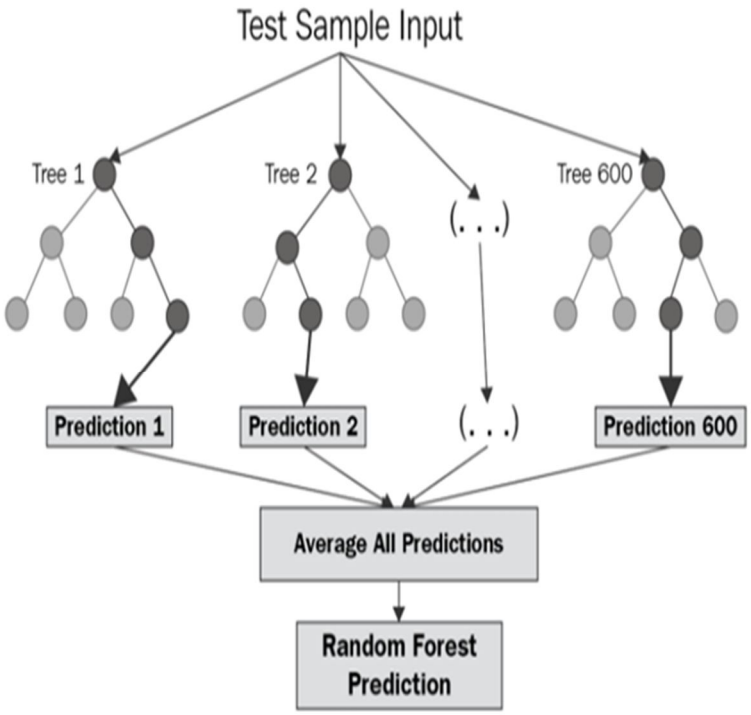


Figure 6. Random Forest

G. Flowchart of Proposed Work

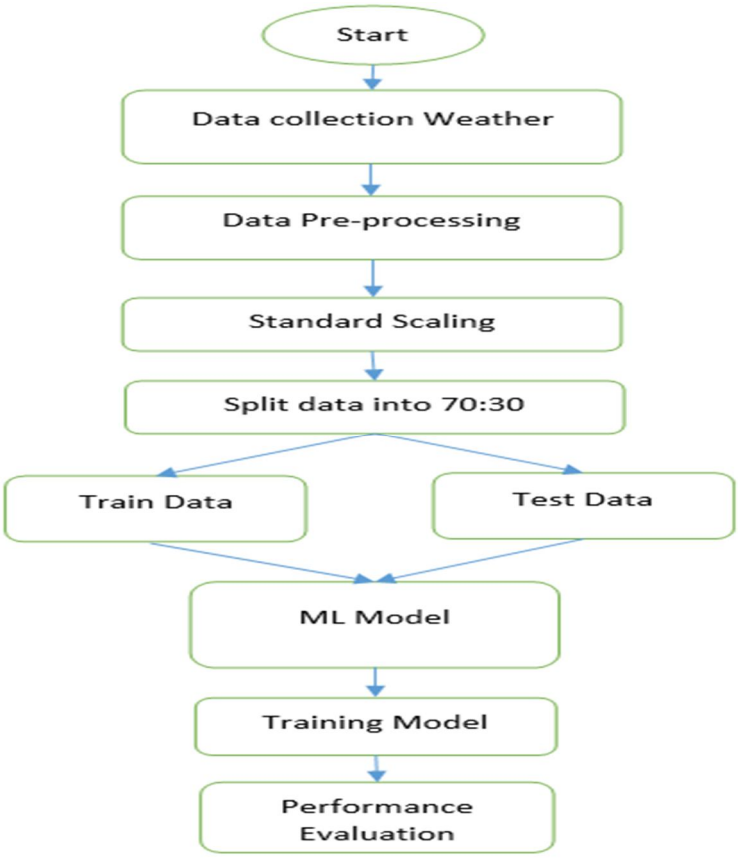


Figure 7. Flowchart of Proposed method

IV. RESULTS AND DISCUSSION

Figure 7 depicts the proposed method's workflow, which starts with the shortlisting of information through the use of an API, continues with the preprocessing of that data, which involves removing null, missing, & nan value systems from the data frame, then visualises the data though the exploratory data analysis, applies standard scaling, looks for skewness, eliminates outliers, and, finally, uses machine learning. Performance can be assessed using r-square and even the mean square error (MSE). The following sentence in this section describes the mse and r-square formulas.

A. Exploratory Data Analysis

To analysis of data using visualization design python function based on matplotlib and seaborn for data frame analysis in which plot histogram, bar plot, correlation matrix shown in below figure and outlier boxplot before and after removing and finally skewness graph of data[14].

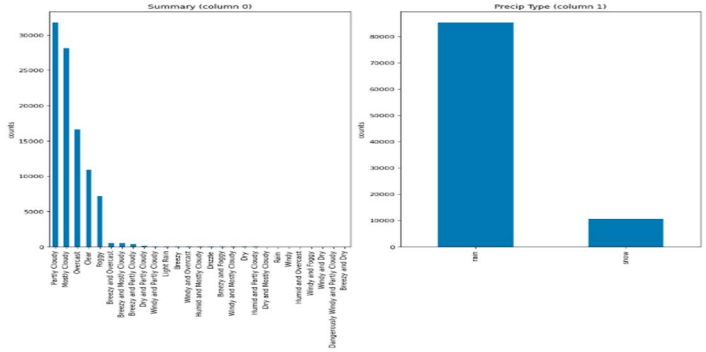


Figure 8. Histogram of features with frequency

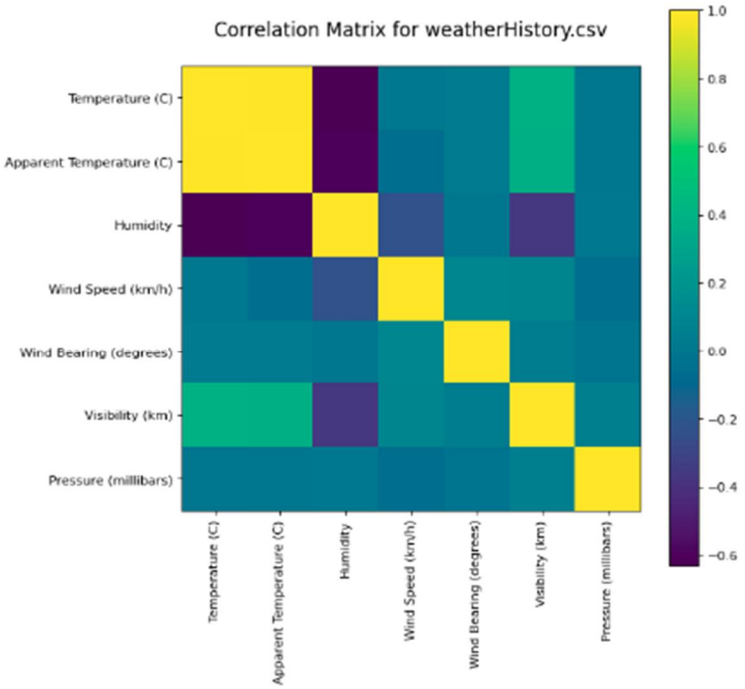


Figure 9. Correlation Matrix of features

B. Performance Analysis

After training the machine learning model to evaluate its performance use two distinct approaches as follows:

1) R-Square: The percentage of the dependent variable's overall variance that can be attributed to the independent variables is expressed by the coefficient of determination (R2) for a regression model. Correlation measures the strength of a connection between two variables, whereas R-squared measures the amount in which the variance for one variable explains for variance in another. Is if R2 is 0.50, the model's inputs may account for half of the observed variance [15].

$$R^2 = 1 - \frac{Unexplained\ Variation}{Total\ Varation} \quad (1)$$

2) MSE: The average of a squares of the errors, and the average squared differences between the estimated or the actual value, is quantified by the mean squared error (MSE) & mean square deviation (MSD) of an estimator (or a method for estimating a unobserved variable). A risk indicator it is proportional to squared error loss is the mean - square error (MSE). MSE is typically non-negative, whether by chance or because an estimator disregards data that might contribute to a more accurate estimate [16].

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y}_i) \quad (2)$$

Table 1. Performance Evaluation

ML Models	R-Square	MSE
Linear Regression	98.06	0.019
Polynomial Regression	99.89	0.3473
Decision Tree	100	0.000084
Random Forest	100	0.024

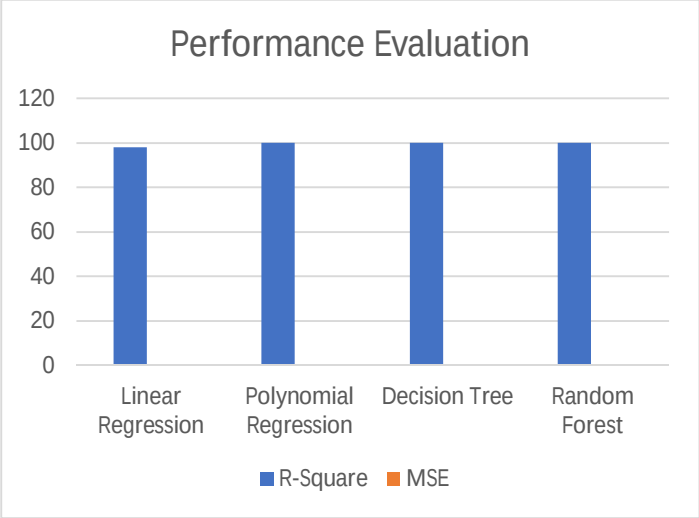


Figure 10. R-Square and MSE of ML Models

The achievement of various algorithms for machine learning is represented in the table above. The decision tree as well as the random forest both achieve the highest r-square, while the linear regression algorithm achieves the lowest mean square and the lowest r-square. Polynomial regression achieves the highest mse while linear regression achieves the lowest r-square[17].

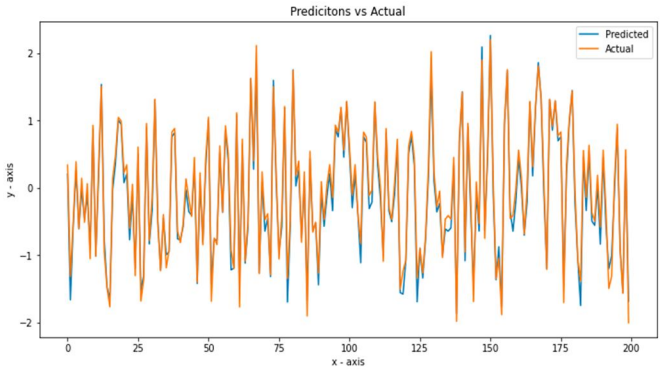


Figure 11. Predicted vs Actual values

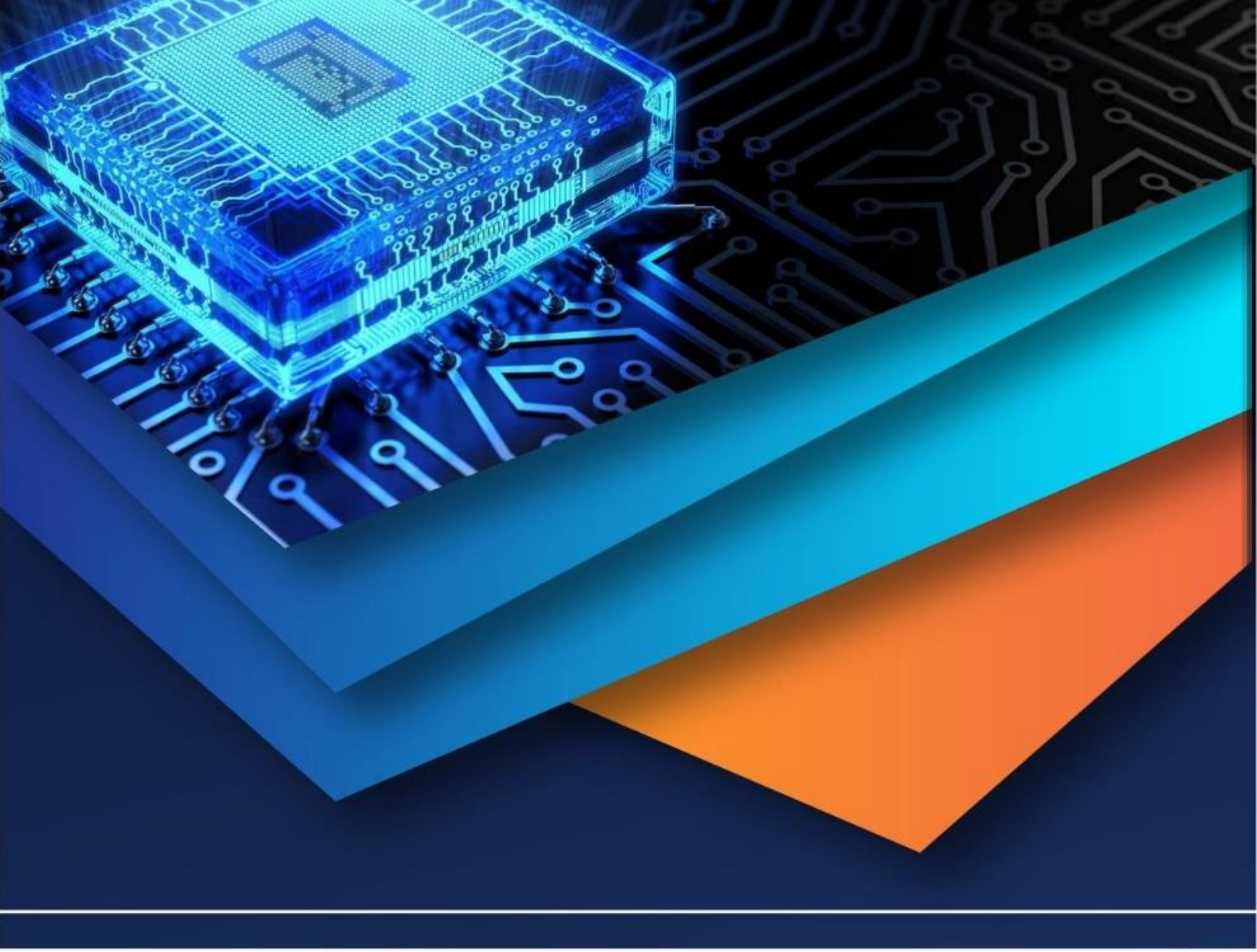
V. CONCLUSION

In recent years, scientific efforts have been focused on improving our ability to anticipate the weather. The reappearance of traditional risks as a result of erratic weather conditions is an event that poses a threat to the evolution of humankind. As a result of this, a variety of information gathering strategies for forecasting the weather have been studied throughout this research. In addition, the ways in which weather forecasting can be used to provide a favorable influence in a variety of industries, most notably agriculture, water systems, the determination of solar-based power sources, and so on, were further broken down. The examination is carried out with regard to a variety of performance measurements. In this contest after train models get the decision tree outperform with high r square like 100 percent and lowest mean square error like 0.000084.

REFERENCES

[1] C. Vennila et al., "Forecasting Solar Energy Production Using Machine Learning," Int. J. Photoenergy, vol. 2022, 2022, doi: 10.1155/2022/7797488.
[2] C. N. Obiora, A. Ali, and A. N. Hasan, "Estimation of Hourly Global Solar Radiation Using Deep Learning Algorithms," 11th Int. Renew. Energy Congr. IREC 2020, vol. 2020, 2020, doi: 10.1109/IREC48820.2020.9310381.
[3] H. Ali-Ou-Salah, B. Oukarfi, K. Bahani, and M. Moujabbir, "A New Hybrid Model for Hourly Solar Radiation Forecasting Using Daily Classification Technique and Machine Learning Algorithms," Math. Probl. Eng., vol. 2021, 2021, doi: 10.1155/2021/6692626.
[4] H. Zhou, Q. Liu, K. Yan, and Y. Du, "Deep Learning Enhanced Solar Energy Forecasting with AI-Driven IoT," Wirel. Commun. Mob. Comput., vol. 2021, 2021, doi: 10.1155/2021/9249387.
[5] N. Singh, S. Chaturvedi, and S. Akhter, "Weather Forecasting Using Machine Learning Algorithm," 2019 Int. Conf. Signal Process. Commun. ICSC 2019, pp. 171–174, 2019, doi: 10.1109/ICSC45622.2019.8938211.

[6] D. Markovics and M. J. Mayer, “Comparison of machine learning methods for photovoltaic power forecasting based on numerical weather prediction,” Renew. Sustain. Energy Rev., vol. 161, no. December 2021, 2022, doi: 10.1016/j.rser.2022.112364.
[7] K. Purwandari, J. W. C. Sigalingging, T. W. Cenggoro, and B. Pardamean, “Multi-class Weather Forecasting from Twitter Using Machine Learning Aprroaches,” Procedia Comput. Sci., vol. 179, no. 2019, pp. 47–54, 2021, doi: 10.1016/j.procs.2020.12.006.
[8] H. Chen, Q. Zhang, and Y. Birkelund, “Machine learning forecasts of Scandinavian numerical weather prediction wind model residuals with control theory for wind energy,” Energy Reports, vol. 8, pp. 661–668, 2022, doi: 10.1016/j.egyr.2022.08.105.
[9] T. P. Fowdur and R. M. Nassir-Ud-Diin Ibn Nazir, “A real-time collaborative machine learning based weather forecasting system with multiple predictor locations,” Array, vol. 14, no. April, p. 100153, 2022, doi: 10.1016/j.array.2022.100153.
[10] Y. Uchôa da Silva, G. B. França, H. M. Ruivo, and H. Fraga de Campos Velho, “Forecast of convective events via hybrid model: WRF and machine learning algorithms,” Appl. Comput. Geosci., vol. 16, no. May, 2022, doi: 10.1016/j.acags.2022.100099.
[11] B. L. Pavuluri, R. S. Vejendla, P. Jithendra, T. Deepika, and S. Bano, “Forecasting Meteorological Analysis using Machine Learning Algorithms,” Proc. - Int. Conf. Smart Electron. Commun. ICOSEC 2020, no. Icosec, pp. 456–461, 2020, doi: 10.1109/ICOSEC49089.2020.9215440.
[12] S. Siddula et al., “Optimal Placement of Hybrid Wind-Solar System Using Deep Learning Model,” Int. J. Photoenergy, vol. 2022, pp. 1–7, 2022, doi: 10.1155/2022/2881603.
[13] J. Mu, F. Wu, and A. Zhang, “Housing Value Forecasting Based on Machine Learning Methods,” Abstr. Appl. Anal., vol. 2014, 2014, doi: 10.1155/2014/648047.
[14] C. Srivastava, S. Singh, and A. P. Singh, “Estimation of air pollution in Delhi using machine learning techniques,” 2018 Int. Conf. Comput. Power Commun. Technol. GUCON 2018, no. August, pp. 304–309, 2019, doi: 10.1109/GUCON.2018.8675022.
[15] M. R. Taufik, E. Rosanti, T. A. Eka Prasetya, and T. Wijayanti Septiarini, “Prediction algorithms to forecast air pollution in Delhi India on a decade,” J. Phys. Conf. Ser., vol. 1511, no. 1, 2020, doi: 10.1088/1742-6596/1511/1/012052.
[16] D. Pruthi and Y. Liu, “Low-cost nature-inspired deep learning system for PM2.5 forecast over Delhi, India,” Environ. Int., vol. 166, no. June, p. 107373, 2022, doi: 10.1016/j.envint.2022.107373.
[17] A. Masood and K. Ahmad, “A model for particulate matter (PM2.5) prediction for Delhi based on machine learning approaches,” Procedia Comput. Sci., vol. 167, no. 2019, pp. 2101–2110, 2020, doi: 10.1016/j.procs.2020.03.258.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY