Title: Predictive Review Modeling for Weather Prediction: A Machine Learning Approach

Abstract: Weather prediction plays a crucial role in various sectors, including agriculture, transportation, and disaster management. This research article explores the application of machine learning techniques to improve the accuracy of weather forecasting. By analyzing historical weather data, meteorological variables, and geographical factors, predictive models are developed to forecast future weather conditions. The study evaluates the performance of machine learning algorithms, such as decision trees, random forests, support vector machines, and neural networks, using metrics like accuracy, precision, recall, and F1-score. Additionally, feature importance analysis is conducted to identify the key factors influencing weather patterns.

Keywords: Weather prediction, Predictive modelling , Machine learning, Decision trees, Random forests, Support vector machines, Neural networks, Feature importance analysis, Meteorological variables.

# INTRODUCTION

Weather prediction is the task of predicting the atmosphere at a future time and a given area. This has been done through physical equations in the early days in which the atmosphere is considered fluid. The current state of the environment is inspected, and the future state is predicted by solving those equations numerically, but we cannot determine very accurate weather for more than 10 days and this can be improved with the help of science and technology. Machine learning can be used to process immediate comparisons between historical weather forecasts and observations. With the use of machine learning, weather models can better account for prediction inaccuracies, such as overestimated rainfall, and produce more accurate predictions. Temperature prediction is of major importance in a large number of applications, including climate-related studies, energy, agricultural, medical, or etc. There are numerous kinds of machine learning calculations, which are Linear Regression, Polynomial Regression, Random Forest Regression, Artificial Neural Network, and Recurrent Neural Network. These models are prepared dependent on the authentic information gave of any area. Contribution to these models is given, for example, if 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.

# Machine Learning

Machine learning is relatively robust to perturbations and does not need any other physical variables for prediction. Therefore, machine learning is a much better opportunity in the evolution of weather forecasting. Before the advancement of Technology, weather forecasting was a hard nut to crack. Weather forecasters relied upon satellites, data model's atmospheric conditions with less accuracy. Weather prediction and analysis have vastly increased in terms of accuracy and predictability with the use of the Internet of Things, for the last 40 years. With the advancement of Data Science, Artificial Intelligence, Scientists now do weather forecasting with high accuracy and predictability

## USE OF ALGORITHMS:

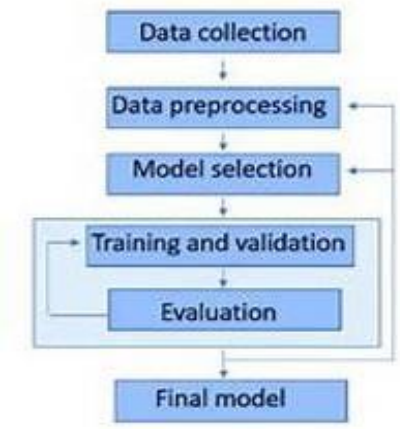
There are different methods of foreseeing temperature utilizing Regression and a variety of Functional Regression, in which datasets are utilized to play out the counts and investigation. To Train, the calculations 80% size of information is utilized and 20% size of information is named as a Test set. For Example, if we need to anticipate the temperature of Kanpur, India utilizing these Machine Learning calculations, we will utilize 8 Years of information to prepare the calculations and 2 years of information as a Test dataset. The as opposed to Weather Forecasting utilizing Machine Learning Algorithms which depends essentially on reenactment dependent on Physics and Differential Equations, Artificial Intelligence is additionally utilized for foreseeing temperature: which incorporates models, for example, Linear regression, Decision tree regression, Random forest regression. To finish up, Machine Learning has enormously changed the worldview of Weather estimating with high precision and predictivity. What's more, in the following couple of years greater progression will be made utilizing these advances to precisely foresee the climate to avoid catastrophes like typhoons, Tornados, and Thunderstorms.

## METHODOLOGY

The dataset utilized in this arrangement has been gathered from Kaggle which is "Historical Weather Data for Indian Cities" from which we have chosen the data for "Kanpur City". The dataset was created by keeping in mind the necessity of such historical weather data in the community. The datasets for the top 8 Indian cities as per the population. The dataset was used with the help of the worldweatheronline.com API and the wwo_hist package. The datasets contain hourly weather data from 01-01-2009 to 01-01-2020. The data of each city is for more than 10 years. This data can be used to visualize the change in data due to global warming or can be used to predict the weather for upcoming days, weeks, months, seasons, etc. Note: The data was extracted with the help of worldweatheronline.com API and we cannot guarantee the accuracy of the data. The main target of this dataset can be used to predict the weather for the next day or week with huge amounts of data provided in the dataset. Furthermore, this data can also be used to make visualization which would help to understand the impact of global warming over the various aspects of the weather like precipitation, humidity, temperature, etc. In this project, we are concentrating on the temperature prediction of Kanpur city with the help of various machine learning algorithms and various regressions. By applying various regressions on the historical weather dataset of Kanpur city we are predicting the temperature like first we are applying Multiple Linear regression, then Decision Tree regression, and after that, we are applying Random Forest Regression.

Table 2.1: Historical Weather Dataset of Kanpur City

| date_time | maxtempC | mintempC | cloudcover | humidity | tempC | sunHour | precipMM | pressure | windspeedKmph |
|---|---|---|---|---|---|---|---|---|---|
| 2009-01-01 00:00:00 | 24 | 10 | 17 | 50 | 11 | 8.7 | 0.0 | 1015 | 10 |
| 2009-01-01 01:00:00 | 24 | 10 | 11 | 52 | 11 | 8.7 | 0.0 | 1015 | 11 |
| 2009-01-01 02:00:00 | 24 | 10 | 6 | 55 | 11 | 8.7 | 0.0 | 1015 | 11 |
| 2009-01-01 03:00:00 | 24 | 10 | 0 | 57 | 10 | 8.7 | 0.0 | 1015 | 12 |
| 2009-01-01 04:00:00 | 24 | 10 | 0 | 54 | 11 | 8.7 | 0.0 | 1016 | 11 |



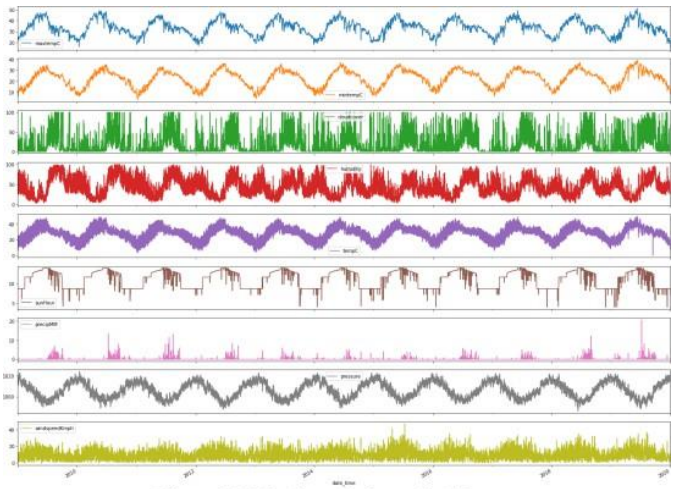


Figure 2.1: Plot for each factor for 10 years

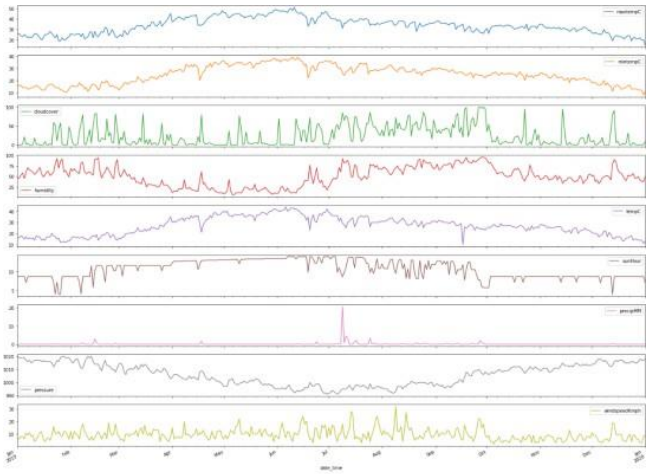
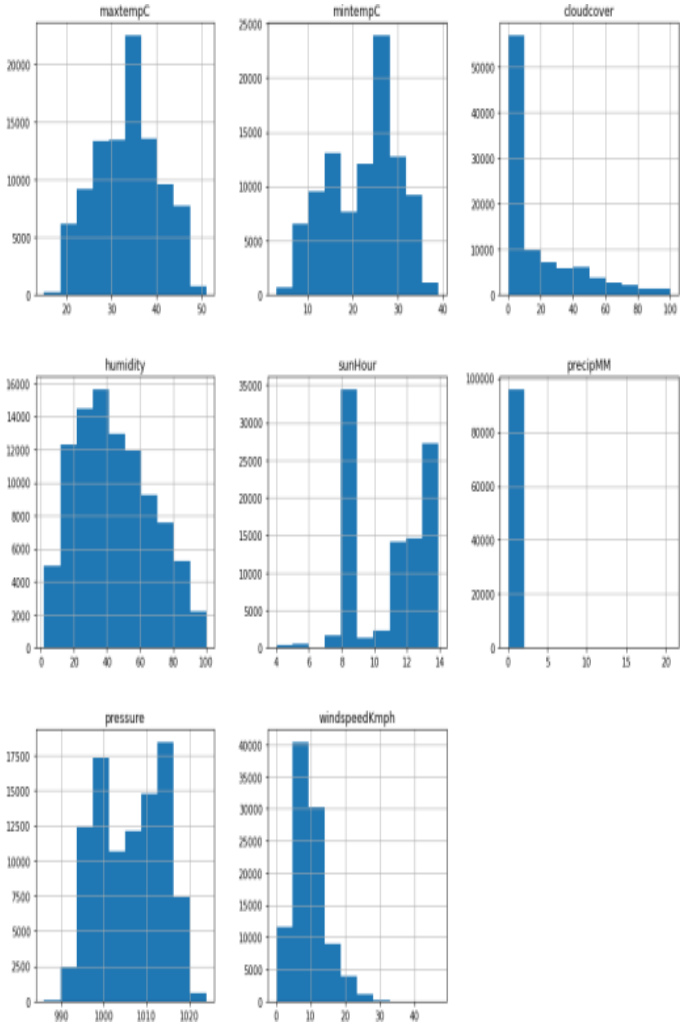


Figure 2.2: Plot for each factor for 1 year

# EXPERIMENTATION

The record has just been separated into a train set and a test set. Each information has just been labeled. First, we take the trainset organizer. We will train our model with the help of histograms and plots. 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 Random Forest Regression. With the help of our histogram, we will train our model. The most important thing in this process is to tune these parameters 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 the 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 the 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 labeled value. The different metrics that we will use confusion matrix, R2 score, etc.



# RESULT AND DISCUSSION

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

## Multiple Linear Regression:

Multiple linear regression is a statistical technique used to analyze the relationship between two or more independent variables and a dependent variable. Unlike simple linear regression, which involves only one independent variable, multiple linear regression incorporates several predictors.

This regression model has high mean absolute error, 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.

| date_time | Actual | Prediction | diff |
|---|---|---|---|
| 2013-07-10 08:00:00 | 34 | 33.209030 | 0.790970 |
| 2015-11-04 20:00:00 | 25 | 25.275755 | -0.275755 |
| 2015-09-21 09:00:00 | 34 | 31.975338 | 2.024662 |
| 2017-02-16 11:00:00 | 28 | 20.496727 | 7.503273 |
| 2012-07-21 01:00:00 | 28 | 28.401085 | -0.401085 |
| ... | ... | ... | ... |
| 2019-03-30 09:00:00 | 37 | 33.187428 | 3.812572 |
| 2015-11-12 12:00:00 | 32 | 28.483724 | 3.516276 |
| 2019-12-31 05:00:00 | 8 | 15.177361 | -7.177361 |
| 2019-08-02 17:00:00 | 35 | 35.363251 | -0.363251 |
| 2019-10-22 08:00:00 | 26 | 27.890691 | -1.890691 |

19287 rows × 3 columns

## Decision Tree Regression:

This regression model has medium mean absolute error, hence turned out to be the little accurate model. Given below is a snapshot of the actual result from the project implementation of multiple linear regression.

## Random Forest Regression:

Random forest regression is a machine learning technique that extends the concept of random forests to the task of regression. It belongs to the ensemble learning methods, which combine multiple models to improve prediction accuracy and robustness. Random forest regression is particularly useful when dealing with non-linear relationships between variables and handling datasets with a large number of features.

```
RandomForestRegressor
RandomForestRegressor(max_depth=90, random_state=0)
```

This regression model has low mean absolute error, hence turned out to be the more accurate model. Given below is a snapshot of the actual result from the project implementation of multiple linear regression.

| date_time | Actual | Prediction | diff |
|---|---|---|---|
| 2013-07-10 08:00:00 | 34 | 34.0 | 0.0 |
| 2015-11-04 20:00:00 | 25 | 25.0 | 0.0 |
| 2015-09-21 09:00:00 | 34 | 34.0 | 0.0 |
| 2017-02-16 11:00:00 | 28 | 28.0 | 0.0 |
| 2012-07-21 01:00:00 | 28 | 28.0 | 0.0 |
| ... | ... | ... | ... |
| 2019-03-30 09:00:00 | 37 | 39.0 | -2.0 |
| 2015-11-12 12:00:00 | 32 | 32.0 | 0.0 |
| 2019-12-31 05:00:00 | 8 | 9.0 | -1.0 |
| 2019-08-02 17:00:00 | 35 | 36.0 | -1.0 |
| 2019-10-22 08:00:00 | 26 | 27.0 | -1.0 |

19287 rows × 3 columns

# Correlation matrix

A correlation matrix is a tabular representation of the correlation coefficients between variables in a dataset. It's a square matrix where each cell contains the correlation coefficient between two variables. The correlation coefficient quantifies the strength and direction of the linear relationship between two variables.

Accuracy = True Purchase + True Not Purchase

_____

Total No of sample

Precision = True Positive

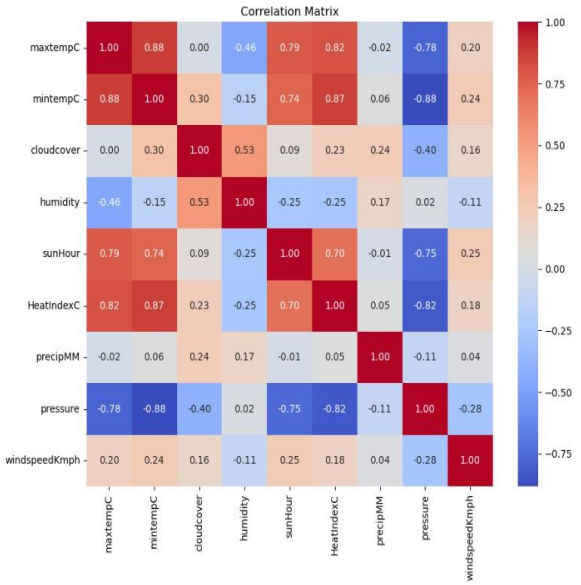
True Positive + False positive

Recall = True Positive

True Positive + False negative

F1 Score = 2*Precision*recall

Precision + recall




```
Mean absolute error: 1.20
Residual sum of squares (MSE): 2.51
R2-score: 0.96
```

```
Mean absolute error: 0.56
Residual sum of squares (MSE): 1.12
R2-score: 0.98
```

```
Mean absolute error: 0.47
Residual sum of squares (MSE): 0.63
R2-score: 0.99
```

```
Variance score: 0.99
```

```
              maxtempC  mintempC  cloudcover  humidity   sunHour  HeatIndexC  \
maxtempC      1.000000  0.881495    0.001678 -0.456879  0.785929    0.816059
mintempC      0.881495  1.000000    0.302925 -0.148172  0.744942    0.872525
cloudcover    0.001678  0.302925    1.000000  0.528443  0.092732    0.225983
humidity     -0.456879 -0.148172    0.528443  1.000000 -0.245537   -0.253712
sunHour       0.785929  0.744942    0.092732 -0.245537  1.000000    0.704134
HeatIndexC    0.816059  0.872525    0.225983 -0.253712  0.704134    1.000000
precipMM     -0.020540  0.060246    0.237962  0.171289 -0.006989    0.052894
pressure     -0.776002 -0.882492   -0.400311  0.023871 -0.753522   -0.823792
windspeedKmph  0.200734  0.244530    0.160360 -0.105510  0.245425    0.183965

              precipMM  pressure  windspeedKmph
maxtempC     -0.020540 -0.776002       0.200734
mintempC      0.060246 -0.882492       0.244530
cloudcover    0.237962 -0.400311       0.160360
humidity      0.171289  0.023871      -0.105510
sunHour      -0.006989 -0.753522       0.245425
HeatIndexC    0.052894 -0.823792       0.183965
precipMM      1.000000 -0.108859       0.038953
pressure     -0.108859  1.000000      -0.277053
windspeedKmph  0.038953 -0.277053       1.000000
```

Precision Score :-

```
0.4749165453503998
```

# CONCLUSION

All the machine learning models: linear regression, various linear regression, decision tree regression, random forest regression were beaten by expert climate determining apparatuses, even though 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 more information. Practical regression, however, was high predisposition, demonstrating that the decision of the model was poor and that its predictions can't be improved by the further accumulation of information. This predisposition could be expected to the structure decision to estimate temperature dependent on the climate of the previous two days, which might be too short to even think about capturing slants in a climate that practical regression requires. On the off chance that the figure was 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. 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. Below is a snapshot of the implementation of Random Forest in the project. Weather Forecasting has a major test of foreseeing the precise outcomes which are utilized in numerous ongoing frameworks like power offices, air terminals, the travel industry focuses, and so forth. The trouble of this determining is the mind-boggling nature of parameters. Every parameter has an alternate arrangement of scopes of qualities

# Reference


- Jaseena, K. U., and Binsu C. Kovoor. "Deterministic weather forecasting models based on intelligent predictors: A survey." *Journal of King Saud University-Computer and Information Sciences* 34.6 (2022): 3393-3412.
- Jaseena, K. U., & Kovoor, B. C. (2022). Deterministic weather forecasting models based on intelligent predictors: A survey. *Journal of King Saud University-Computer and Information Sciences*, *34*(6), 3393-3412.
- Salman, Afan Galih, Bayu Kanigoro, and Yaya Heryadi. "Weather forecasting using deep learning techniques." *2015 international conference on advanced computer science and information systems (ICACSIS)*. Ieee, 2015.
- Salman, A. G., Kanigoro, B., & Heryadi, Y. (2015, October). Weather forecasting using deep learning techniques. In *2015 international conference on advanced computer science and information systems (ICACSIS)* (pp. 281-285). Ieee.
- Hewage, Pradeep, et al. "Deep learning-based effective fine-grained weather forecasting model." *Pattern Analysis and Applications* 24.1 (2021): 343-366.
- Hewage, P., Trovati, M., Pereira, E., & Behera, A. (2021). Deep learning-based effective fine-grained weather forecasting model. *Pattern Analysis and Applications*, *24*(1), 343-366.
- Paras, Sanjay Mathur, Avinash Kumar, and Mahesh Chandra. "A feature based neural network model for weather forecasting." *International Journal of Computational Intelligence* 4.3 (2009): 209-216.
- Paras, S. M., Kumar, A., & Chandra, M. (2009). A feature based neural network model for weather forecasting. *International Journal of Computational Intelligence*, *4*(3), 209-216.