A MACHINE LEARNING MODEL FOR WEATHER FORECASTING

A PROJECT REPORT

Submitted by

PAVAN KALYAN	**(2203A51356)**
JAFAR AHMED	**(2203A51373)**
SAI CHARAN	**(2203A51400)**
SOMASHEKAR	**(2203A51362)**

in partial fulfillment for the award of the degree

of

BACHELORS OF TECHNOLOGY

in

COMPUTER SCIENCE & TECHNOLOGY

SR UNIVERSITY OF ENGINEERING AND TECHNOLOGY (WGL)

ACKNOWLEDGEMENT

I would like to express my special thanks of gratitude to my project guide Dr. Soumik Podder Sir who gave me the golden opportunity to do this wonderful project on the topic "A Machine Learning Model for Weather Predicting", which also helped me in doing a lot of research and I came to know about so many new things I am really thankful to them.

Secondly I would also like to thank my friends who helped me a lot in finalizing this project within the limited time frame.

Date : 03-05-2024

Pavankalyan (2203A51356)
Jafar Ahmed (2203A51373)
Sai Charan (2203A51400)
Somashekar (2203A51363)

TABLE OF CONTENTS

	TITLE	
	ABSTRACT	**4**
	BACKGROUND	**4**
	OBJECTIVE	**4**
1.0	**INTRODUCTION**	**5**
	1.1 Introduction	5
	1.2 Machine Learning	5
	1.3 Use of Algorithms	6
2.0	**METHODOLOGY**	**7**
4.0	**EXPERIMENTATION**	**9**
5.0	**RESULT AND DISCUSSION**	**10**
	5.1 Multiple Linear Regression	10
	5.2 Decision Tree Regression	11
	5.3 Random Forest Regression	12
6.0	**CONCLUSION**	**13**

ABSTRACT

Traditionally, climate assessment has been performed reliably by treating the environment as a liquid. The current wind condition is being observed. The future state of the environment is recorded by understanding thermodynamics and the numerical position of the liquid elements. Nevertheless, this traditional arrangement of differential conditions as observed by physical models is at times unstable under oscillating effects and uncertainties when estimating the underlying states of air. This indicates an insufficient understanding of environmental variations, so it limits climate forecasts to 10-day periods because climate projections are essentially unreliable. But machine learning is moderately hearty for most barometric destabilizing effects compared to traditional techniques. Another favorable position of machine learning is that it does not depend on the physical laws of environmental processes.

Background

For the current situation, India observatory conducts traditional weather forecasting. There are four common methods to predict the weather. The first method is the climatology method that is reviewing weather statistics gathered over multiple years and calculating the averages. The second method is an analog method that is to find a day in the past with weather similar to the current forecast. The third method is the persistence and trends method that has no skill to predict the weather because it relies on past trends. The fourth method is numerical weather prediction the is making weather predictions based on multiple conditions in the atmosphere such as temperatures, wind speed, high-and low-pressure systems, rainfall, snowfall, and other conditions. So, there are many limitations of these traditional methods. Not only it forecasts the temperature in the current month at most, but also it predicts without using machine learning algorithms. Therefore, my project is to increase the accuracy and predict the weather in the future for at least one month by applying machine learning techniques

Objective (Brief)

Purpose of this project is to predict the temperature using different algorithms like linear regression, random forest regression, and K-nearest neighbour. The output value should be numerically based on multiple extra factors like maximum temperature, minimum temperature, cloud cover, humidity, and sun hours in a day, precipitation, pressure and wind speed.

1. 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 K-nearest neighbour. 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, K-nearest neighbour, 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.

2. 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 K-nearest neighbour, 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

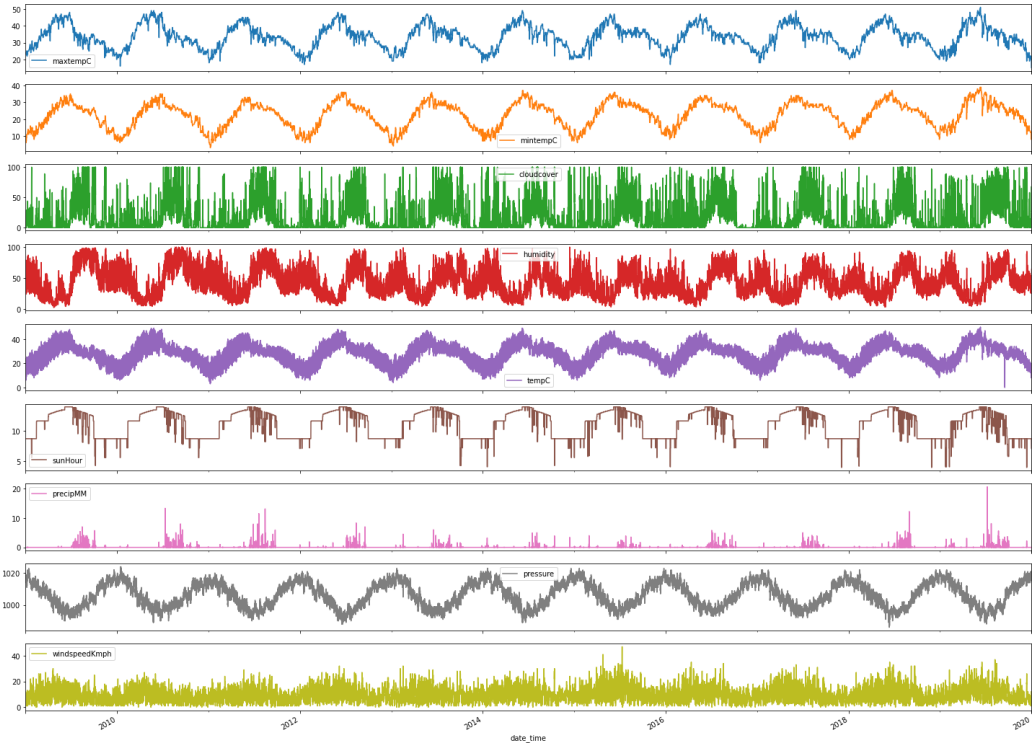


Figure 2.1: Plot for each factor for 10 years

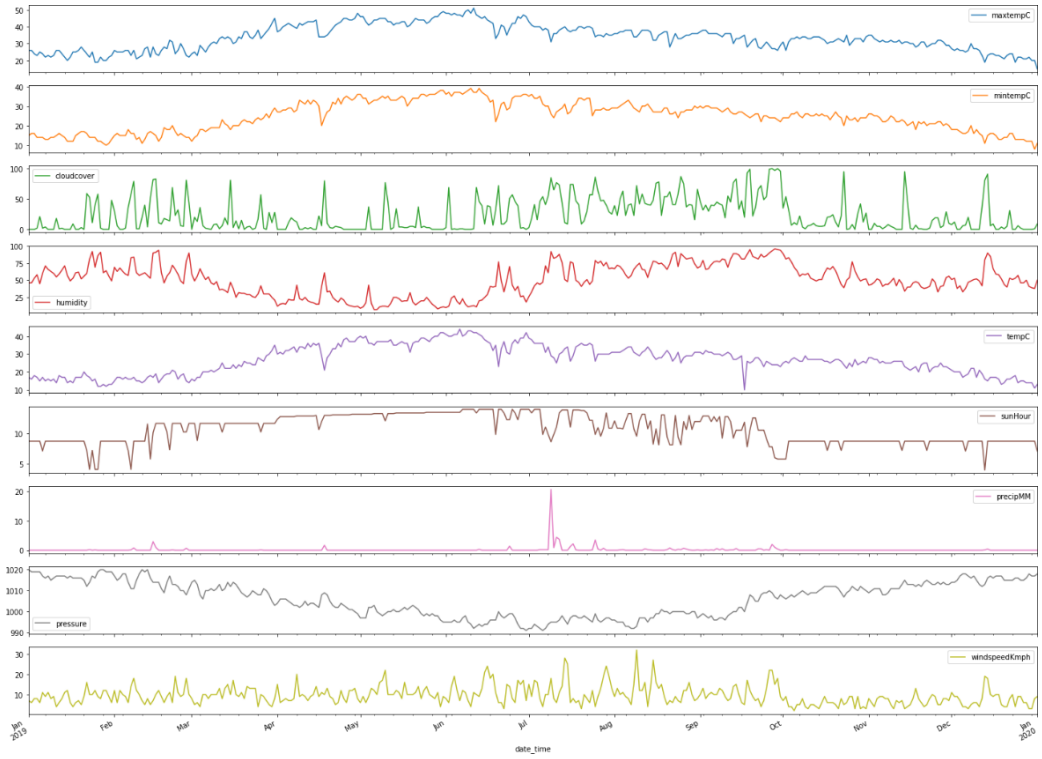
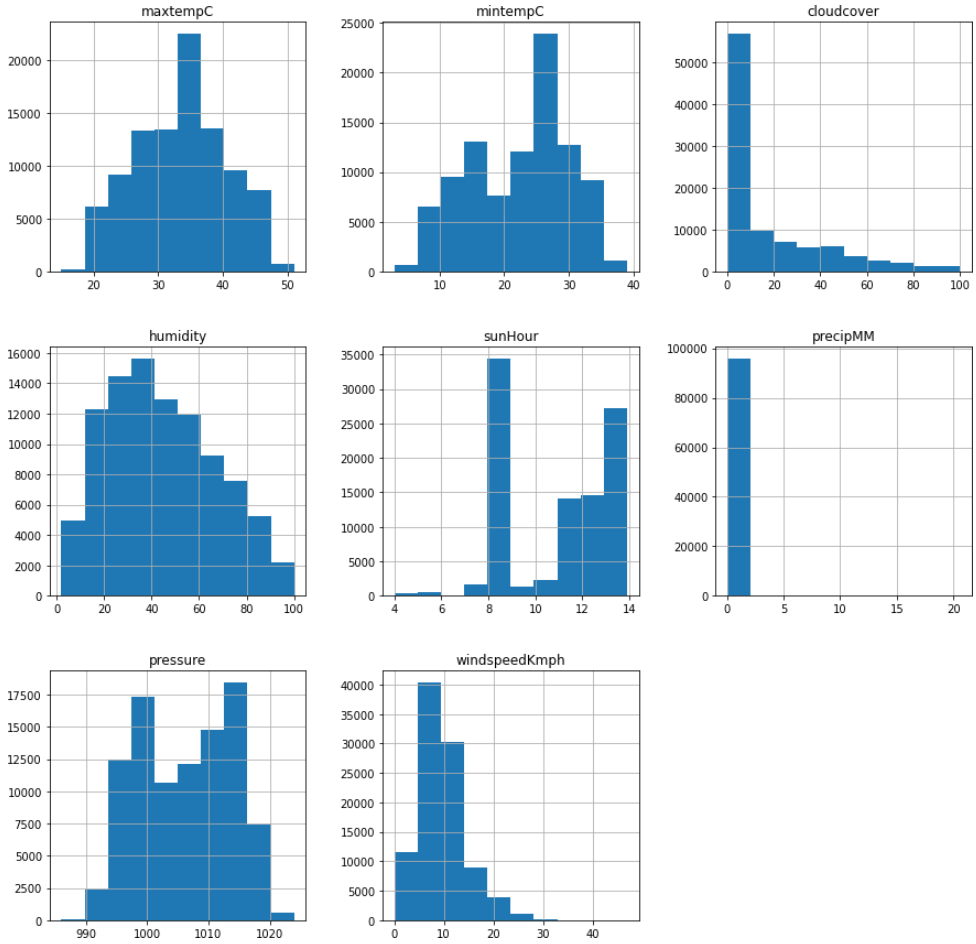


Figure 2.2: Plot for each factor for 1 year

3. 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 Regression, K-nearest neighbour, and 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.



4. RESULT AND DISCUSSION

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

Multiple Linear Regression:

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

K-nearest Neighbour 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.

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

Random Forest Regression:

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	33.94	0.06
2015-11-04 20:00:00	25	24.43	0.57
2015-09-21 09:00:00	34	34.36	-0.36
2017-02-16 11:00:00	28	26.35	1.65
2012-07-21 01:00:00	28	28.17	-0.17
...	...	...	...
2019-03-30 09:00:00	37	32.99	4.01
2015-11-12 12:00:00	32	31.74	0.26
2019-12-31 05:00:00	8	10.62	-2.62
2019-08-02 17:00:00	35	35.72	-0.72
2019-10-22 08:00:00	26	26.85	-0.85

19287 rows × 3 columns

5. CONCLUSION

All the machine learning models: linear regression, various linear regression, K-Nearest neighbour, 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.