WEATHER FORECASTING PREDICTION
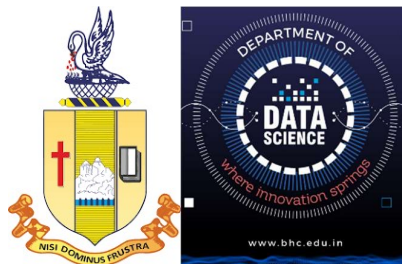
Micro Project for Practical Machine Learning

by
P.Asha Belcilda
225229104

Submitted To
Dr. K. RAJKUMAR
Course Instructor

DEPARTMENT OF DATA SCIENCE
BISHOP HEBER COLLEGE (AUTONOMOUS)
TIRUCHIRAPPALLI 620017

MAY 2022

CERTIFICATE

I hereby acknowledge that this project is the original work done by me for the requirements for Micro Project in Practical Machine Learning Course. This micro project is not copied from internet or whatsoever.

Tiruchirappalli

19 May 2022   Your Name and signature

ACTIVITY: 1

TITLE OF THE PROJECT: WEATHER FORECASTING PREDICTION

ACTIVITY:2

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.

MOTIVATION:

1. Improved Accuracy
2. Early Warning
3. Business Applications
4. Climate Change Mitigation

Overall, the motivation for using machine learning for weather forecasting is to improve the accuracy, reliability, and usefulness of weather forecasts for a range of applications, from public safety to business operations, to environmental protection.

PROBLEM STATEMENT:

The problem statement for weather forecasting using machine learning is to accurately predict weather conditions based on current and historical meteorological data. The goal is to improve the accuracy and reliability of weather forecasts, which can have significant implications for public safety, business operations, and environmental protection.

Overall, the problem statement for weather forecasting using machine learning is to develop accurate, reliable, and adaptable algorithms that can provide early warning of extreme weather events, help businesses make informed decisions, and contribute to efforts to mitigate the effects of climate change.

ACTIVITY 3:

DATA SET DESCRIPTION:

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.

ACTIVITY:4

Existing solution:

- IBM's Deep Thunder: Deep Thunder is a high-resolution weather forecasting system developed by IBM that uses machine learning algorithms to provide real-time weather forecasts. The system combines traditional numerical weather prediction models with machine learning techniques to improve forecast accuracy.
- Google's Global Flood Forecasting Model: Google has developed a machine learning-based global flood forecasting model that predicts the likelihood and severity of flooding in real-time. The model uses satellite imagery and meteorological data to provide accurate predictions.
- The Weather Company: The Weather Company, a subsidiary of IBM, uses machine learning algorithms to provide weather forecasts for businesses and organizations. The company uses a range of data sources, including weather radar and satellite imagery, to provide highly accurate and detailed forecasts.

- AccuWeather: AccuWeather is a weather forecasting company that uses machine learning algorithms to provide highly accurate weather forecasts. The company uses a range of data sources, including weather radar, satellite imagery, and weather station data, to provide real-time and accurate forecasts.

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.

Source code:
https://github.com/ashabelcilda/Weather-Forecasting.git

ACTIVITY:5

Proposed Methodology:

date_time	maxtempC	mintempC	cloudcover	humidity	tempC	sunHour	HeatIndexC	precipMM	pressure	windspeedKmph
2009-01-01 00:00:00	24	10	17	50	11	8.7	12	0.0	1015	10
2009-01-01 01:00:00	24	10	11	52	11	8.7	13	0.0	1015	11
2009-01-01 02:00:00	24	10	6	55	11	8.7	13	0.0	1015	11
2009-01-01 03:00:00	24	10	0	57	10	8.7	13	0.0	1015	12
2009-01-01 04:00:00	24	10	0	54	11	8.7	14	0.0	1016	11

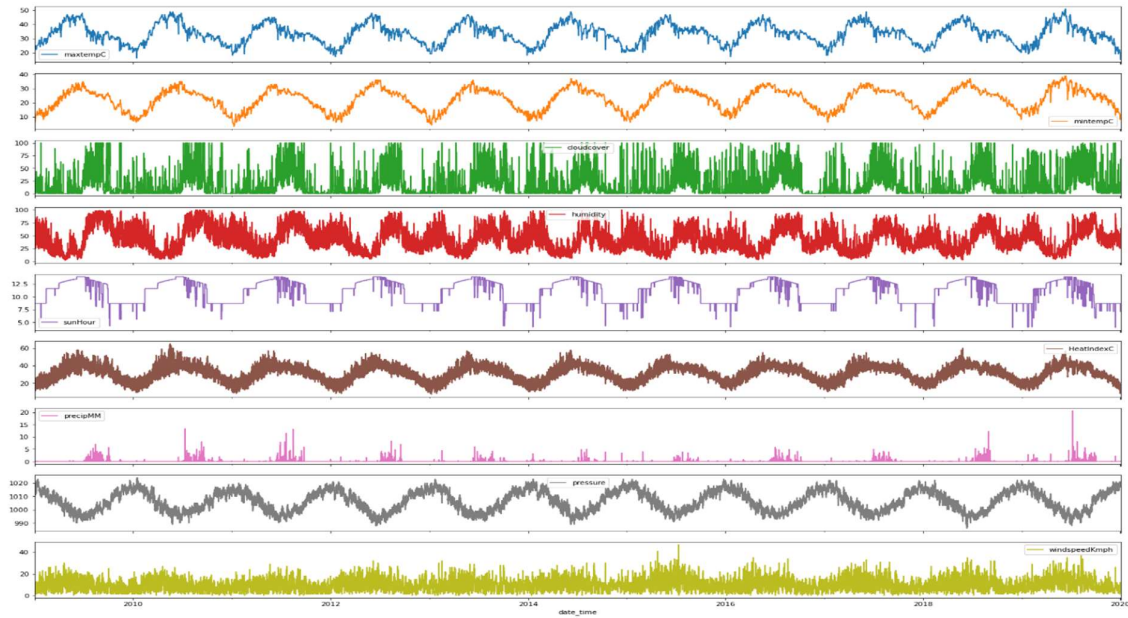Table 1.1: Historical Weather Dataset of Kanpur City

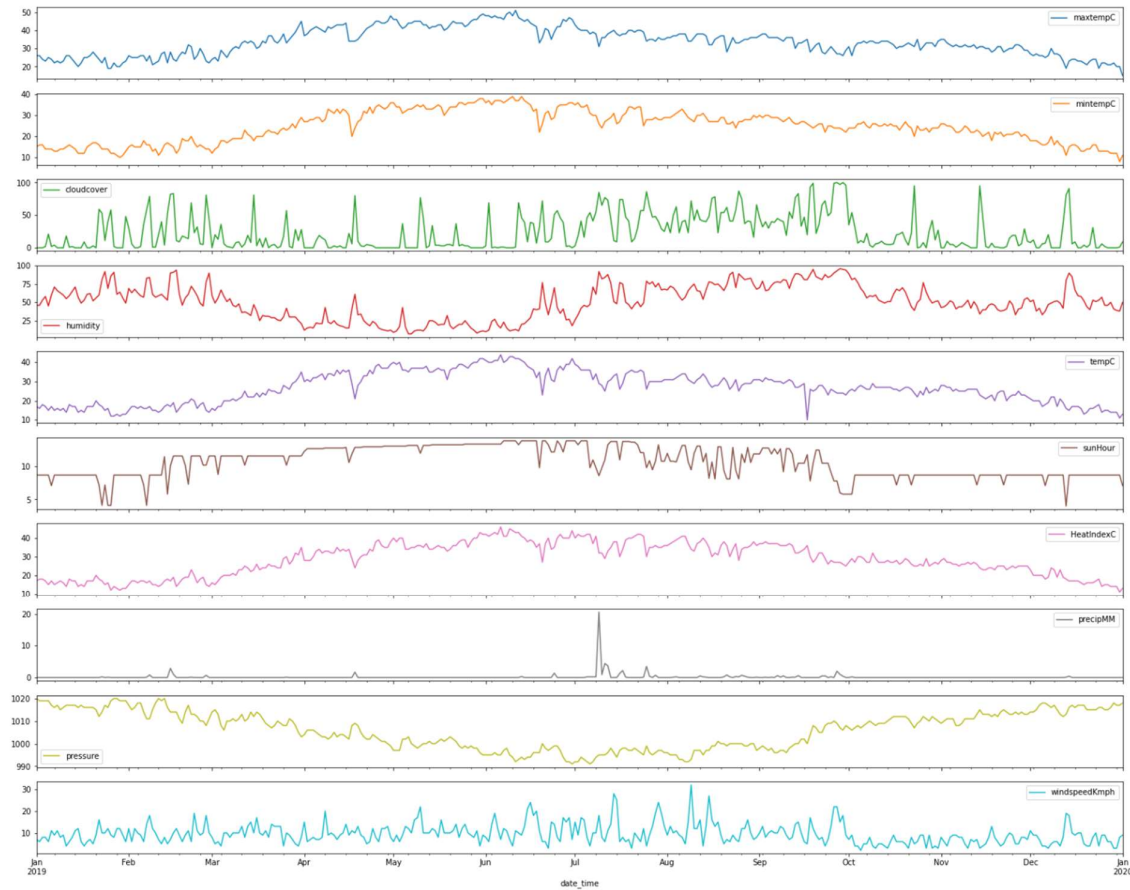Figure 1.1: Plot for each factor for 10 years

Figure 1.2: Plot for each factor for 1 year

ACTIVITY:6

Implementation details:

      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	34.89	-0.89
2015-11-04 20:00:00	25	24.57	0.43
2015-09-21 09:00:00	34	35.08	-1.08
2017-02-16 11:00:00	28	25.22	2.78
2012-07-21 01:00:00	28	28.04	-0.04
...	...	...	...
2019-03-30 09:00:00	37	33.55	3.45
2015-11-12 12:00:00	32	30.36	1.64
2019-12-31 05:00:00	8	9.13	-1.13
2019-08-02 17:00:00	35	35.92	-0.92
2019-10-22 08:00:00	26	25.77	0.23

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.

date_time	Actual	Prediction	diff
2013-07-10 08:00:00	34	34.0	0.0
2015-11-04 20:00:00	25	24.0	1.0
2015-09-21 09:00:00	34	34.0	0.0
2017-02-16 11:00:00	28	27.0	1.0
2012-07-21 01:00:00	28	28.0	0.0
...	...	...	...
2019-03-30 09:00:00	37	32.0	5.0
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	35.0	0.0
2019-10-22 08:00:00	26	26.0	0.0

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	33.92	0.08
2015-11-04 20:00:00	25	24.84	0.16
2015-09-21 09:00:00	34	34.25	-0.25
2017-02-16 11:00:00	28	27.00	1.00
2012-07-21 01:00:00	28	27.99	0.01
...	...	...	...
2019-03-30 09:00:00	37	32.79	4.21
2015-11-12 12:00:00	32	31.91	0.09
2019-12-31 05:00:00	8	8.81	-0.81
2019-08-02 17:00:00	35	34.98	0.02
2019-10-22 08:00:00	26	26.32	-0.32

19287 rows × 3 columns

ACTIVITY:7

Experimental results and evaluation:

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, Decision Tree Regression, 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.

ACTIVITY:8

Details of your code repository in GitHub and demo in Youtube:

Uploaded dataset file,README.md file, Weather Forecasting python file and the whole micro project file on my GitHub account.

GitHub link: gh repo clone ashabelcilda/Weather-Forecasting

ACTIVITY:9

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.