DEVELOPING A SMART WEATHER STATION BY LEVERAGING MACHINE LEARNING ALGORITHMS AND STATISTICAL MODELING FOR AN EFFICACIOUS METEOROLOGICAL PREDICTION

Suchit Lamba, Prof. Jayraj Singh, Prof. Jitendra Joshi

NIIT University, Neemrana, Rajasthan, India

ABSTRACT

Traditionally, climate assessment methods have approached the environment as a liquid, closely observing wind conditions to predict future states. However, estimating the underlying air states has proven challenging due to oscillating effects and uncertainties. As a result, short-term climate forecasts have become increasingly unreliable. In this study, we propose a promising alternative that harnesses the power of machine learning, specifically utilizing the Random Forest model, for weather prediction.

Machine learning techniques, like Random Forest, offer greater robustness in handling the destabilizing effects of barometric conditions compared to traditional methods. What's fascinating is that machine learning doesn't rely on the governing physical laws of environmental processes. Instead, it uses data-driven techniques to reveal patterns and relationships in the available data, enabling more accurate forecasting and forecasting.

By embracing machine learning, we have the potential to overcome the limitations of traditional approaches to climate analysis. The remarkable ability of machine learning to handle complex environmental variables, especially through random forest modelling, makes it an invaluable tool for extends forecasts and increases the reliability of weather forecasting.

This review highlights the implications of machine learning for climate forecasting, specifically focusing on random forest modelling. By conducting our proposed study with random forests we demonstrate the tremendous potential of machine learning to transform climate forecasting by providing more accurate and reliable forecasts.

*In addition, this work highlights the future implications of machine learning in addressing climate challenges, increasing our understanding of the environment, and enabling informed decision-making in the air change and the face of change By embracing machine learning we take an important step towards more reliable weather forecasting***.**

Keywords: *Smart weather station; machine learning; weather forecasting; temperature prediction; linear regression; random forest regression; decision tree regression*

BACKGROUND

Conventional weather forecasting methods used by the Indian Observatory have limitations in accuracy and predictability. These methods include climatology, analogy, sustainability and trends, and statistical climate prediction. Climate science relies on historical weather statistics to calculate averages, while the methods look at past weather days like current forecasts. Numerical weather forecasting requires complex calculations based on weather conditions.

However, these traditional methods have limitations. Short-term forecasts are often provided and machine learning algorithms are not included. To overcome these limitations, my work aims to increase the accuracy of weather forecasting using machine learning. By harnessing the power of machine learning algorithms, we can extend the forecast beyond a month and improve the forecast accuracy. Machine learning algorithms can analyze historical weather data, identify patterns, and make predictions based on various factors such as temperature, wind speed, precipitation, etc. This approach overcomes the limitations of traditional methods and provides the ability to provide accurate and reliable long-term weather forecasts

OBJECTIVE (BRIEF)

The main objective of this project is to predict temperature using various algorithms. To make an accurate forecast, we will consider many other factors such as maximum temperature, minimum temperature, cloud cover, humidity, number of sunshine hours in a day, rainfall, pressure and wind speed etc. These features play an important role in determining temperature and provide valuable information for our predictive models.

By using different regression algorithms and including more input variables, we aim to build a model that can efficiently explore the relationship between these factors and temperature thereby forecasting more accurate future temperatures. The project aims to use machine learning techniques and available data to improve our understanding of how these processes affect temperature and to enhance our ability to predict future temperature changes.

INTRODUCTION

A weather forecast is a forecast of future weather conditions at a specific location. Traditionally, climate prediction was based on physical equations that model climate as a water system. By analyzing the state of the environment, these equations were mathematically solved to predict future climate. But obtaining accurate forecasts beyond about 10 days has proven difficult, advances in science and technology can help solve this issue. Machine learning algorithms offer possible solutions through the historical sky processing situational data and observations, enabling climate models to better account for forecast errors and increase accuracy forecasts in particular, linear regression. Various machine learning algorithms such as polynomial regression, random forest regression, artificial neural networks, recursive neural networks etc can be used. These algorithms are trained with historical site-specific data, with input variables for minimum temperature, maximum temperature, mean air pressure, mean humidity, and two weather distributions. Dates can affect the usefulness of this information Once adopted, forecasts of minimum and maximum temperatures can be made over a seven-day period. This review aims to address existing research gaps and introduce new approaches to weather forecasting using machine learning. By carefully testing different algorithms, integrating physics-based models, and exploring clustering techniques, the goal is to increase prediction accuracy and provide more reliable forecasts for a wide range of applications

MACHINE LEARNING

Machine learning has emerged as a powerful tool in climate forecasting because of its ability to demonstrate resilience to perturbations and not rely solely on physical variables for prediction. In the past, weather forecasters encountered challenges stemming from inaccuracies in satellite data and weather models. However, over the last four decades, the integration of the internet and advancements in data science and artificial intelligence have significantly improved the accuracy and predictability of weather forecasting. These advancements have allowed scientists to achieve more precise and reliable weather forecasts.

USE OF ALGORITHMS

Algorithms play a vital role in temperature forecasting, encompassing regression methods such as linear regression, functional regression, and various other techniques. These algorithms are trained and tested using datasets, typically allocating 80% of the data for training and 20% for testing. For instance, when predicting the temperature in Kanpur, India, using a machine learning algorithm, eight years of historical data may be utilized for training, while two years are reserved for testing. Apart from simulation-based methods based on physics and differential equations, artificial intelligence methods are also employed for temperature prediction. These methods leverage machine learning algorithms to process historical data and make accurate temperature forecasts**.**

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. The choice to use machine learning algorithms like linear regression, random forest regression, and decision tree regression is driven by their ability to handle complex relationships in the data and make accurate predictions. These algorithms have been widely used in weather forecasting and have shown promising results in capturing patterns and trends in historical weather data. Additionally, these algorithms can handle a variety of input factors such as maximum temperature, minimum temperature, cloud cover, humidity, sun hours, precipitation, pressure, and wind speed, allowing for a comprehensive analysis of multiple variables.

RELATED WORKS

Several research papers have been published on weather prediction and forecasting techniques. "The Weather Forecast Using Data Mining Research Based on Cloud Computing" [1] proposes a service-oriented architecture for weather information systems that utilize data mining techniques, including Artificial Neural Networks (ANN) and Decision Tree algorithms. The study shows promising results in generating classification rules for weather variables. Another paper titled "Analysis on The Weather Forecasting and Techniques" [2] explores the use of artificial neural networks and fuzzy logic for weather prediction. The authors consider various attributes such as temperature, humidity, pressure, and wind in their analysis. These techniques demonstrate better accuracy compared to traditional methods.[3]. Research on weather prediction also highlights challenges and issues. "Issues with Weather Prediction" [4] discusses the limitations of weather prediction, including the inherent uncertainties and variations in different regions. While numerical models are used to simulate future weather conditions, they still have inherent errors due to imprecise equations used in the models. Environmental changes have long been emphasized due to rapid changes [5] and therefore climate change is important. Meteorology is responsible for predicting the weather at a future time and in a particular place [6]. Weather plays an important role in many areas. Air pressure plays an important role in the weather [7]. Air pressure is always related to the structure of the body. The current state of the atmosphere is sampled and the future state is calculated by mathematically solving the fluid dynamics and thermodynamic equations.[8]. However, the system of equilibrium equations governing this physical model is unstable under the influence of the atmosphere and uncertain in the initial measurement of the atmosphere [9].

PROPOSED METHODOLOGY

The dataset used in this study was obtained from Kagle, specifically the "Historical Weather Data for Indian Cities" dataset. From this data set, we select data for Kanpur. The purpose of this data set was to provide historical climate information to the community. The dataset includes hourly weather data from 1 January 2009- 1 January 2020 for the top 8 Indian cities based on population. Data was retrieved using the worldweatheronline.com API and wwo.hist package but it should be noted that although the information was obtained from a reliable source the accuracy of the information cannot be guaranteed The main purpose of this list is to forecast the weather for the next day or week by detailed so the information provided as well as rainfall using data We can see the effect of global warming on various weather models such as humidity and water a cold In this work we focused on temperature in Kanpur using various machine learning algorithms and regression methods Applied multilinear regression decision tree regression and random forest regression to historical climate data set in Kanpur.

	maxtempC	mintempC	cloudcover	humidity	tempC	sunHour	HeatIndexC	precipMM	pressure	windspeedKmph
date_time										
2009-01-01 0:00:00	24	10	17	50	11	8.7	12	0	1015	10
2009-01-01 1:00:00	24	10	11	52	11	8.7	13	0	1015	11
2009-01-01 2:00:00	24	10	6	55	11	8.7	13	0	1015	11
2009-01-01 3:00:00	24	10	0	57	10	8.7	13	0	1015	12
2009-01-01 4:00:00	24	10	0	54	11	8.7	14	0	1016	11

Table 3.1: Historical Weather Dataset of Kanpur City

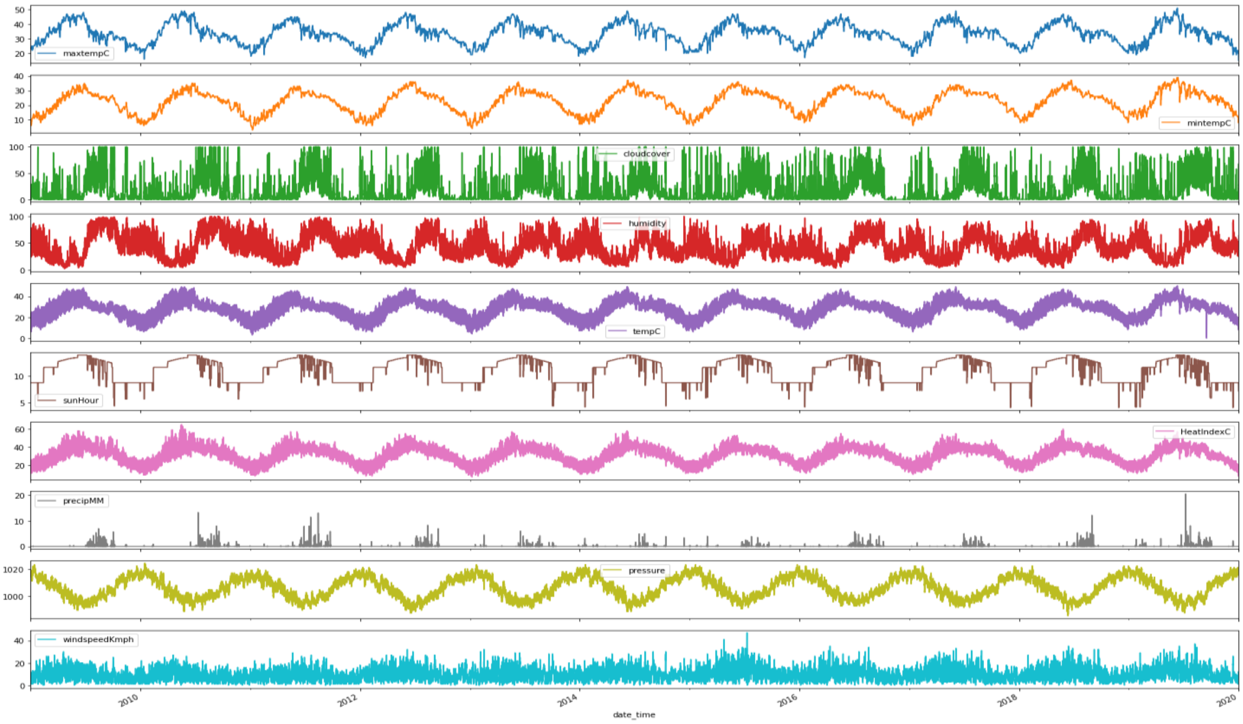


Figure 3.2: Plot for each factor for 10 years

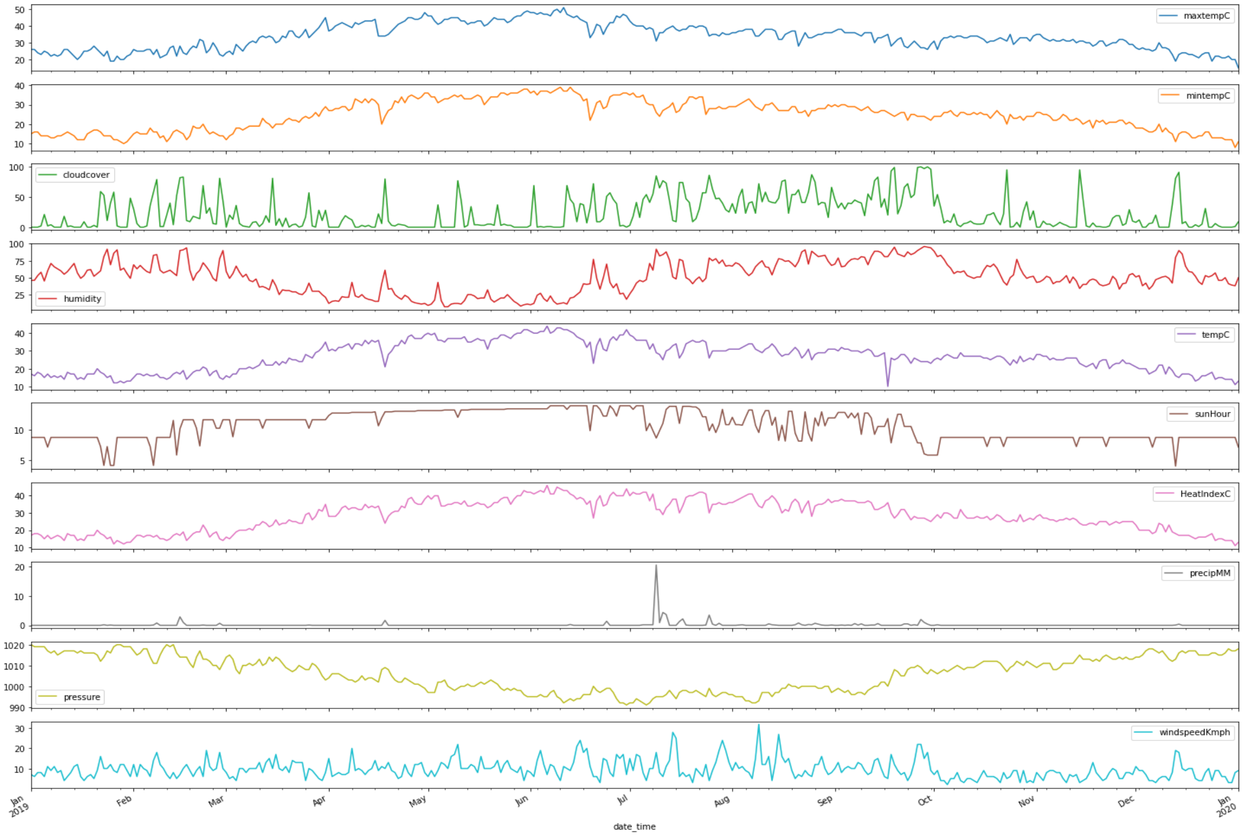


Figure 3.3: Plot for each factor for 1 year

Workflow:

The data set is divided into train sets and test sets and each data point is labelled. We start by analyzing the train system and extracting features from histograms and plots. The extracted features are then stored in a histogram. This process is performed for each data point in the training set. Next, we build our classification models, specifically Random Forest Regression, Decision Tree Regression and Linear Regression, We train our models with histogram data. Tuning of model parameters is important to obtain accurate results. When we train, we use a test set to test the model. For each variable in the test set, we use feature extraction techniques to compare the values in the train set histogram and then predict each day in the test program such as confusion matrix and R2 scores that compare predicted and labelled about values.

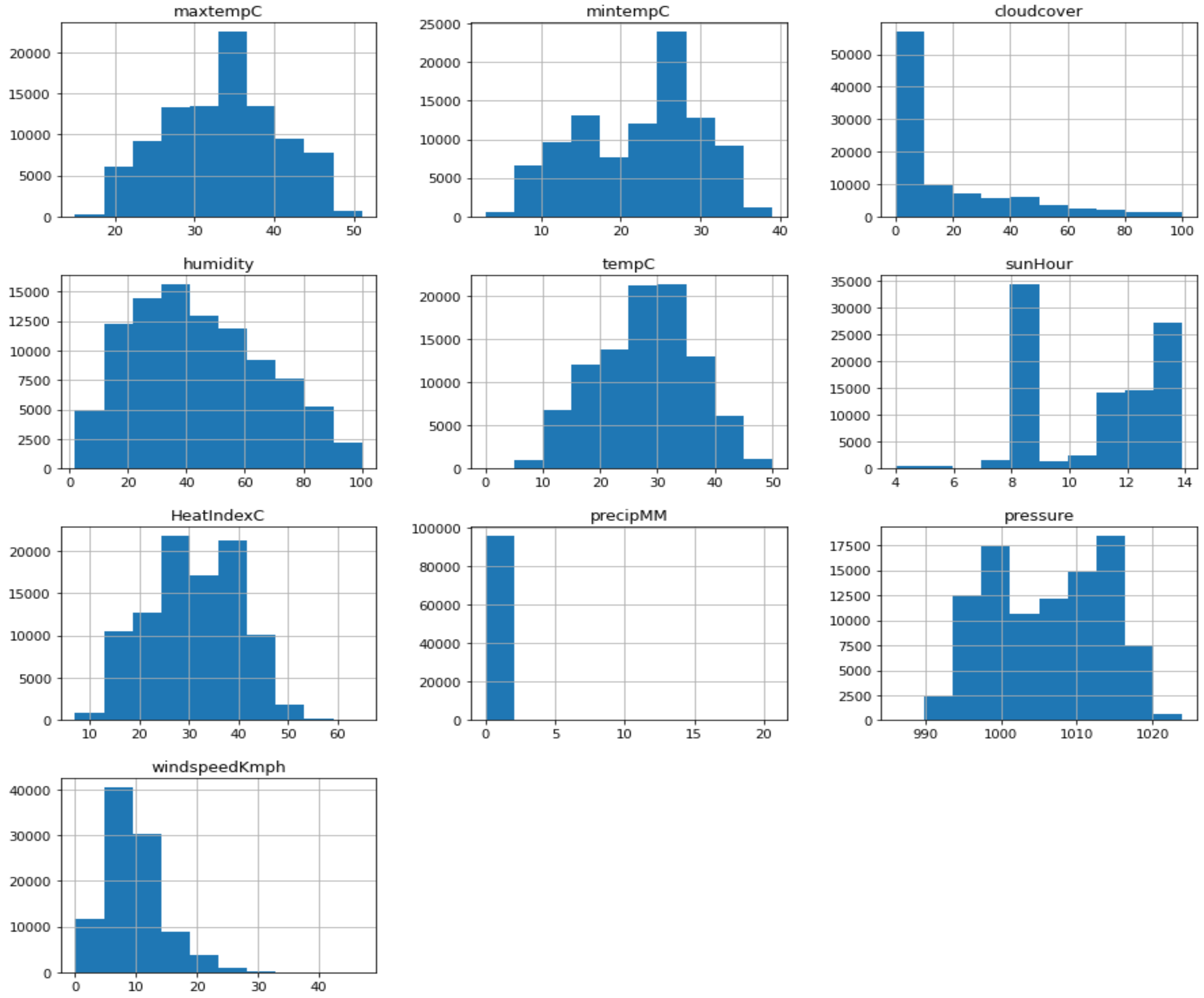


Figure 3.1.1 Histogram of Variables

Technology

This project uses the following technology. Several libraries and packages will be used in Python for our analysis. Pandas, NumPy, among other required packages will allow us to carry out calculations along with operations on arrays.

We will use Matplotlib and Seaborn to visualize an array of data structures as well as its dimensions. These libraries provide a wide range of graphs and images in order to create informative and insightful visual weather data.

As far as building forecasting models go, we make use of several algorithms. Specifically, for temperatures, we employ multiple linear regression, decision tree regression, and random forest regressor Algorithms. These algorithms allow us to capture relationships and underlying patterns revealed in the dataset and predict it’s perfect.

A normal personal computer with enough processing power and memory may do this task in terms of hardware requirements. The size of the data collection and the complexity of the models we utilize will determine the precise hardware requirements.

To speed up our development process, we'll employ open-source software (OSS) tools like Jupyter Notebook or Google Colab. These platforms make it simpler to collaborate and share our work while also offering a user-friendly environment for coding, testing, and documenting.

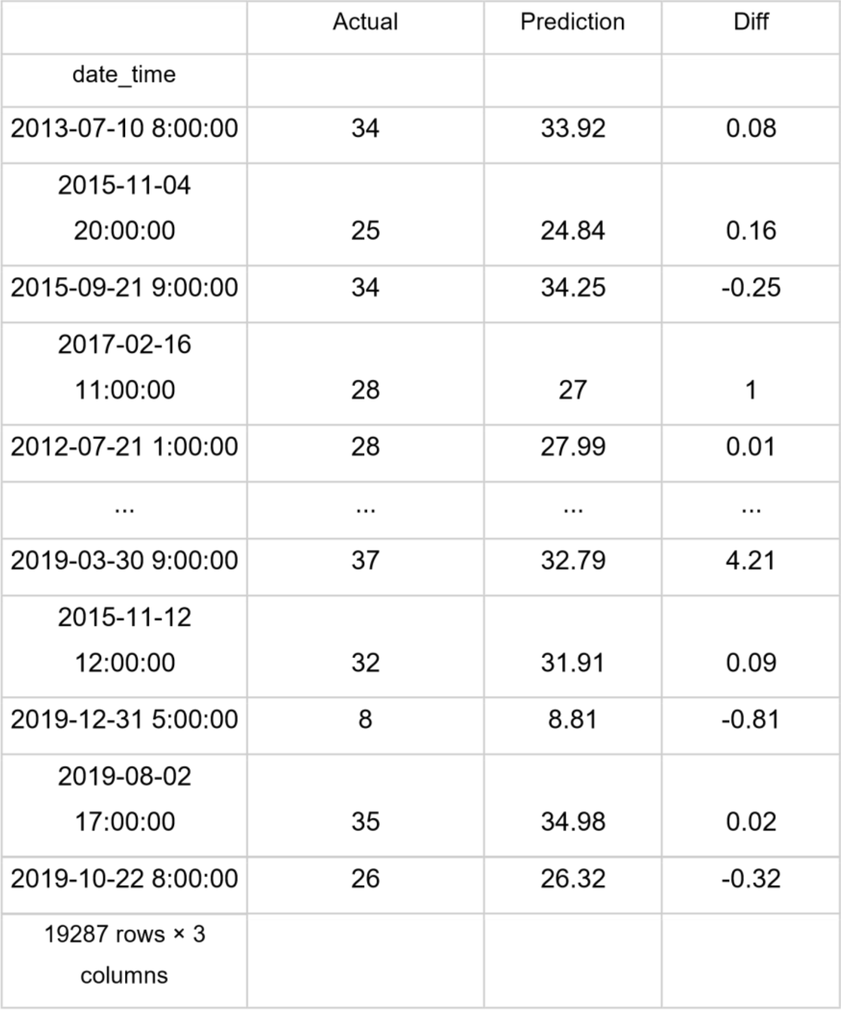
We will be prepared to preprocess and extract some data sets by utilizing the capabilities of Python, along with associated libraries, machine learning frameworks, and open-source software tools.

RESULTS AND ANALYSIS

The project implementation yielded the following results:

Multiple Linear Regression:

The multiple linear regression model showed a high error rate, indicating a very imprecise model. Below is a plot of the actual results obtained from running multiple linear regressions of the task.

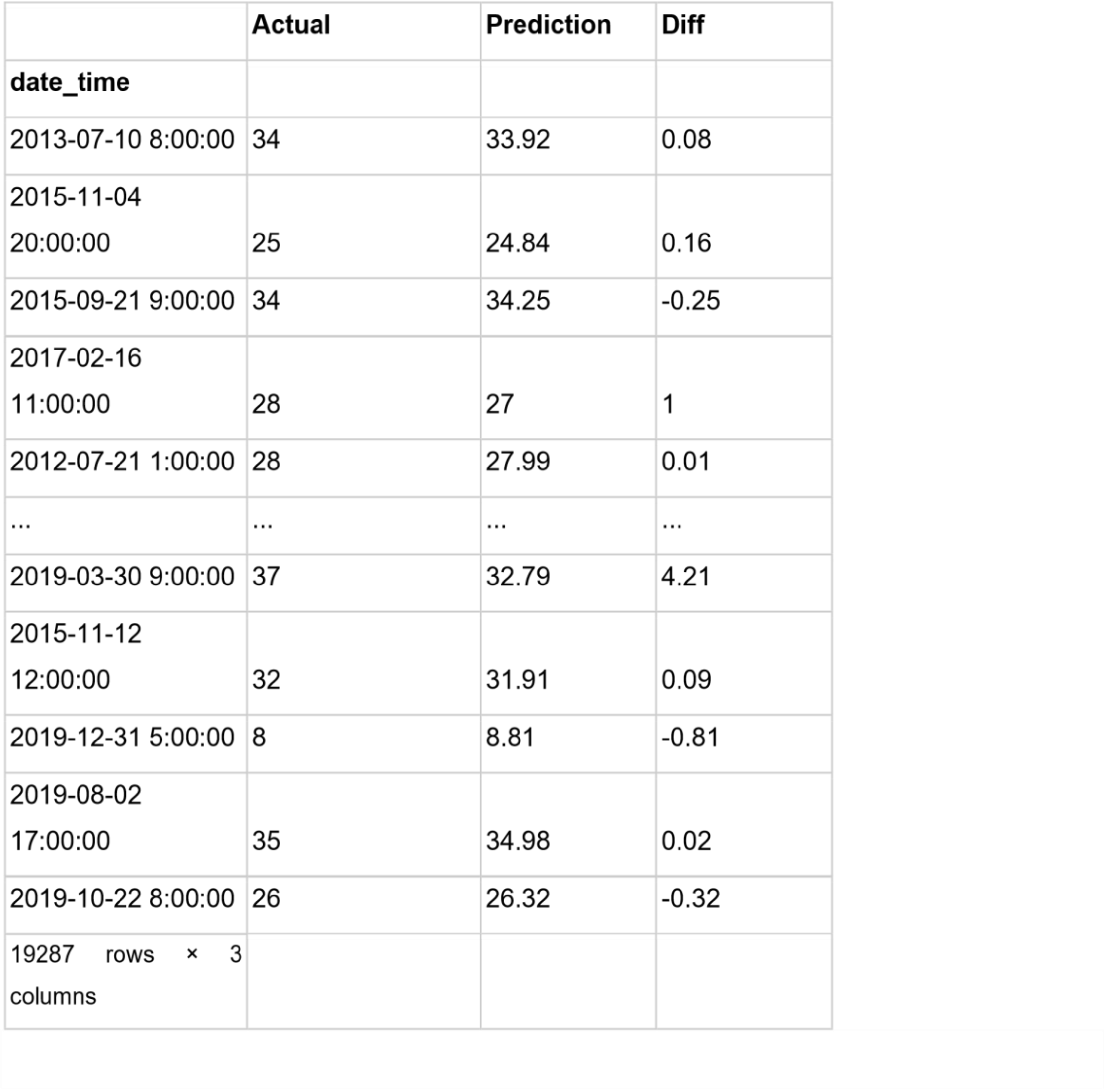


	Actual	Prediction	Diff
date_time			
2013-07-10 8:00:00	34	33.92	0.08
2015-11-04 20:00:00	25	24.84	0.16
2015-09-21 9:00:00	34	34.25	-0.25
2017-02-16 11:00:00	28	27	1
2012-07-21 1:00:00	28	27.99	0.01
...	...	...	...
2019-03-30 9:00:00	37	32.79	4.21
2015-11-12 12:00:00	32	31.91	0.09
2019-12-31 5:00:00	8	8.81	-0.81
2019-08-02 17:00:00	35	34.98	0.02
2019-10-22 8:00:00	26	26.32	-0.32
19287 rows × 3 columns			

Table 4.1 Results of Multiple Linear Regression

Decision Tree Regression:

The mean absolute error was revealed, indicating a moderate accuracy in the decision tree regression model. Below is a screenshot of the actual results using the project decision tree.

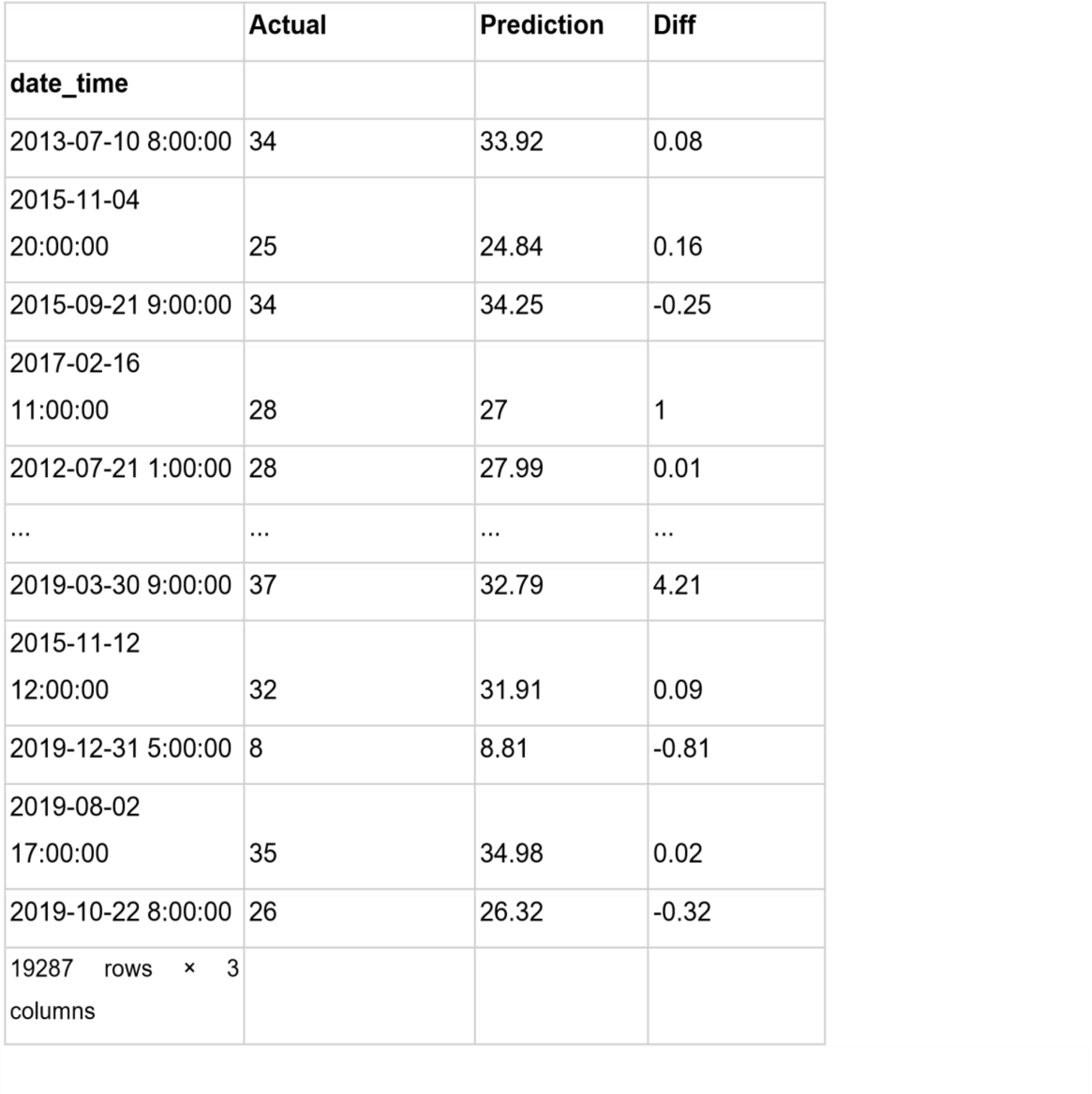


	Actual	Prediction	Diff
date_time			
2013-07-10 8:00:00	34	33.92	0.08
2015-11-04 20:00:00	25	24.84	0.16
2015-09-21 9:00:00	34	34.25	-0.25
2017-02-16 11:00:00	28	27	1
2012-07-21 1:00:00	28	27.99	0.01
...	...	...	...
2019-03-30 9:00:00	37	32.79	4.21
2015-11-12 12:00:00	32	31.91	0.09
2019-12-31 5:00:00	8	8.81	-0.81
2019-08-02 17:00:00	35	34.98	0.02
2019-10-22 8:00:00	26	26.32	-0.32
19287 rows × 3 columns			

Table 4.2 Results of Decision Tree Regression

Random Forest Regression:

The random forest regression model exhibited lower absolute errors, indicating better accuracy than the other models. Below is a plot of the actual results obtained from the project's random regression forest implementation.



	Actual	Prediction	Diff
date_time			
2013-07-10 8:00:00	34	33.92	0.08
2015-11-04 20:00:00	25	24.84	0.16
2015-09-21 9:00:00	34	34.25	-0.25
2017-02-16 11:00:00	28	27	1
2012-07-21 1:00:00	28	27.99	0.01
...	...	...	...
2019-03-30 9:00:00	37	32.79	4.21
2015-11-12 12:00:00	32	31.91	0.09
2019-12-31 5:00:00	8	8.81	-0.81
2019-08-02 17:00:00	35	34.98	0.02
2019-10-22 8:00:00	26	26.32	-0.32
19287 rows × 3 columns			

Table 4.3 Results of Random Forest Regression

CONCLUSIONS AND FUTURE SCOPE

Applied regression models including, decision tree regression, linear regression, and random forest regression showed promising results in predicting temperature based on climatic conditions The models achieved reasonable accuracy in predicting temperature within, with low absolute error and root mean squared error values. The R-squared score indicated a good fit of the model to the data. These findings highlight the effectiveness of machine learning in weather forecasting and show how temperature forecasts can be improved over conventional methods

LIMITATIONS AND SHORTCOMINGS

Despite the positive results, there were some limitations and shortcomings in this work. One of the limitations was the availability and quality of the dataset used. The dataset may not have captured all factors that could have an impact on temperature forecasting, which could have affected the accuracy of the models. In addition, the dataset contained missing or erroneous data requiring preprocessing. The limited availability of real-time data was another obstacle, as climate can change rapidly and affect temperature forecasting. Another limitation is the choice of regression algorithms. Although linear regression, decision tree regression, and random forest regression performed well in this study, there may be other algorithms that can provide even better results and consider more advanced algorithms such as gradient boosting if deep learning techniques are to be explored in future research.

Scope for Improvement

Many more areas could be explored to improve the accuracy and reliability of temperature forecasts. First, adding additional features to the models can improve their performance. Factors such as air pollution levels, geological features, and climate can be considered influential variables. By including a wider range of variables, models can capture more complex relationships and improve temperature prediction.

Another way to improve is to explore team strategies. Combining predictions from multiple models, each trained with different algorithms or subsets of features, can often improve overall performance Ensemble techniques, such as stacking or boosting, can power models have been successfully used to improve the robustness of temperature estimates.

Moreover, the accuracy of temperature forecasts can be greatly increased by the integration of real-time data. Access to up-to-date weather information such as satellite imagery, aerial measurements, and sensor data can enable accurate and timely input into images Real-time data integration can enable dynamic and adaptive forecasting, to enable adaptation based on changing weather conditions.

Future Directions

Collecting detailed and diverse data is important for future weather forecasting. Including data from different geographical, climatic, and seasonal regions can help develop general models that can accurately predict temperatures under different conditions Furthermore, historical climate data and the inclusion of multi-year data can facilitate long-term climate analysis and forecasting.

Exploring other machine learning algorithms specifically optimized for time series analysis capabilities is another direction for future research. Algorithms such as recurrent neural networks (RNNs) or short-term memory networks (LSTM) can capture time dependence and patterns in climate data, providing improved forecasting capabilities.

In addition, the range of comprehensive climate prediction models that can be applied is vast. Industries such as agriculture, transportation, renewable energy, and disaster management can benefit from more accurate temperature forecasts. Combining decision support systems with the development of user-friendly interfaces can enable stakeholders to make informed decisions based on reliable weather forecasts.

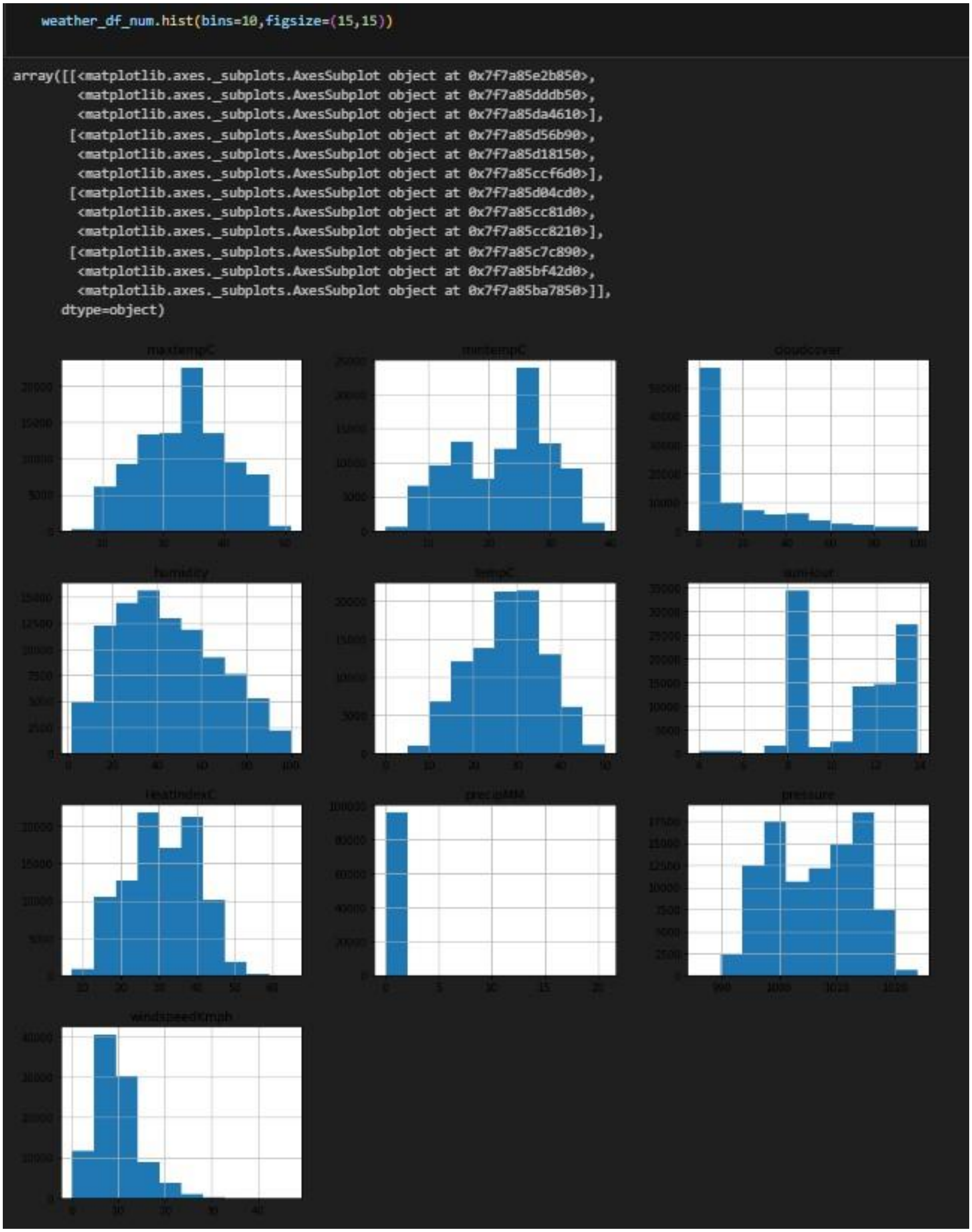
In conclusion, this work demonstrates the potential of machine learning in weather forecasting, especially in temperature forecasting, with high accuracy and still areas for improvement and future research. By addressing constraints, adding additional features, exploring clustering methods, and using real-time data, more accurate and reliable temperature forecasts can be provided to so has further improved, leading to more efficient decision-making and improved outcomes across industries

REFERENCES

1. A B M Mazharul Mujib Dalian University of Technology. The Weather Forecast Using Data Mining Research Based on Cloud Computing.
2. Jabani and Priyanka Sebastian. (2014). Analysis of The Weather Forecasting and Techniques.
3. Samenow and Frirz. (2015). Issues with weather prediction.
4. Shubham Madan, Praveen Kumar, Seema Rawat, Tanupriya Choudhury, “Analysis of Weather Prediction using Machine Learning & Big Data,” International Conference on Advances in Computing and Communication Engineering (ICACCE-2018) Paris, France 22-23 June 2018. [5]
5. Munmun Biswas, Tanni Dhoom, Sayantanu Barua “Weather Forecast Prediction: An Integrated Approach for Analyzing and Measuring Weather Data” International Journal of Computer Applications (0975– 8887) Volume 182 – No. 34, December 2018.
6. Aris Pujud Kurniawan, Agung Nugroho Jati, Fairuz Azmi “Weather Prediction Based on Fuzzy Logic Algorithm for Supporting General Farming Automation System,” International Conference on Instrumentation, Control, and Automation (ICA) Yogyakarta, Indonesia, August 9-11, 2017.
7. Nasimul Hasan, Md. Taufeeq Uddin, Nihad Karim Chowdhury “Automated Weather Event Analysis with Machine Learning,”.

8. Mark Holmstrom, Dylan Liu, Christopher Vo, “Machine Learning Applied to Weather Forecasting” Stanford University (Dated: December 15, 2016).

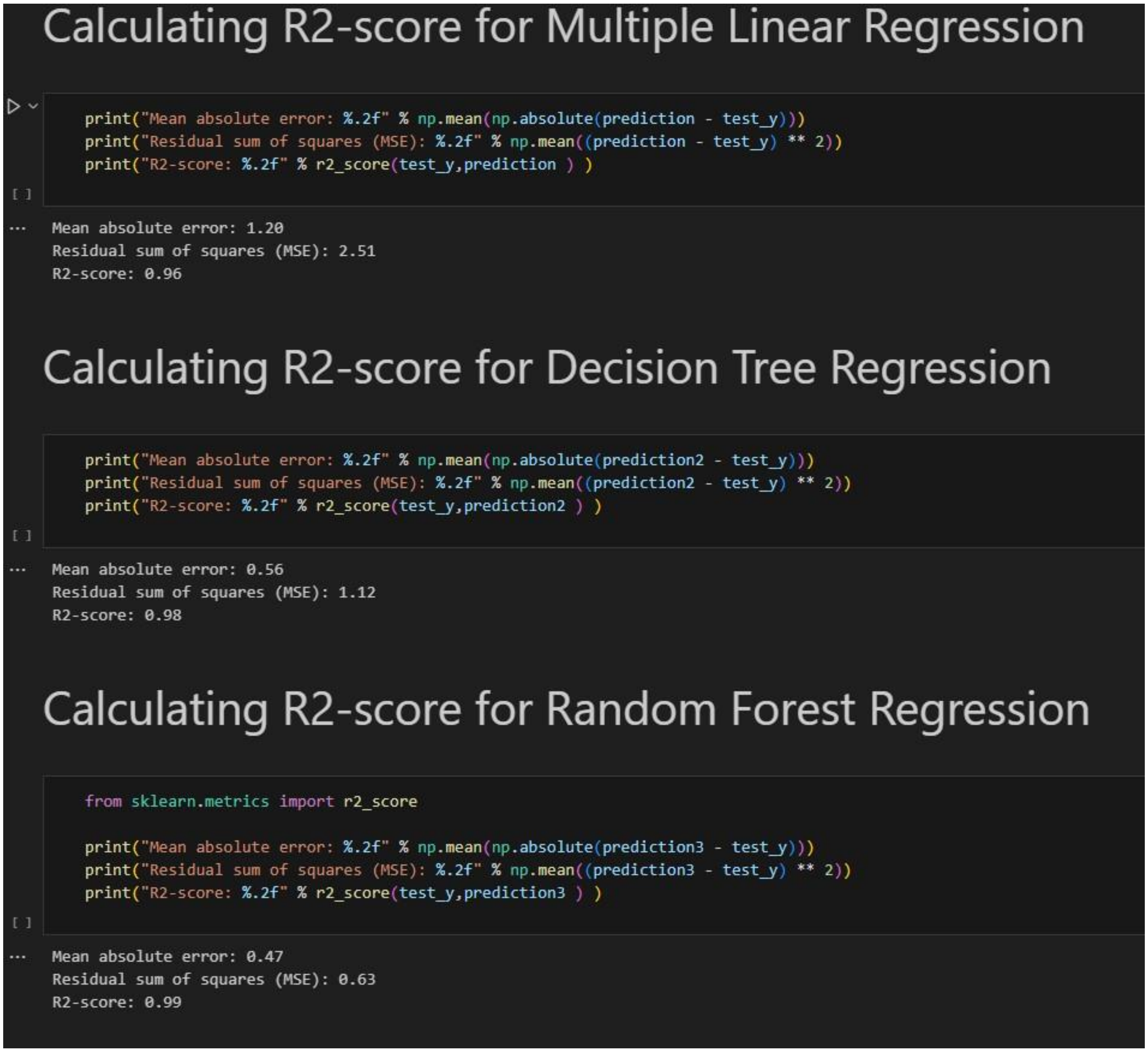
9. J. Wu, L. Huang, and X. Pan, "A novel Bayesian additive regression trees ensemble model based on linear regression and nonlinear regression for torrential rain forecasting,” in Computational Science and Optimization (CSO), 2010 Third International Joint Conference on, vol. 2. IEEE, 2010, pp. 466–470.



```
weather_df_num.hist(bins=10,figsize=(15,15))
```

```
array([[<matplotlib.axes._subplots.AxesSubplot object at 0x7f7a85e2b850>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x7f7a85dddb50>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x7f7a85da4610>],
       [<matplotlib.axes._subplots.AxesSubplot object at 0x7f7a85d56b90>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x7f7a85d18150>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x7f7a85ccf6d0>],
       [<matplotlib.axes._subplots.AxesSubplot object at 0x7f7a85d04cd0>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x7f7a85cc81d0>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x7f7a85cc8210>],
       [<matplotlib.axes._subplots.AxesSubplot object at 0x7f7a85c7c890>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x7f7a85bf42d0>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x7f7a85ba7850>]],
      dtype=object)
```

Fig6.1 Finding the Histograms of variable



Calculating R2-score for Multiple Linear Regression

```
print("Mean absolute error: %.2f" % np.mean(np.absolute(prediction - test_y)))
print("Residual sum of squares (MSE): %.2f" % np.mean((prediction - test_y) ** 2))
print("R2-score: %.2f" % r2_score(test_y,prediction ) )
```

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

Calculating R2-score for Decision Tree Regression

```
print("Mean absolute error: %.2f" % np.mean(np.absolute(prediction2 - test_y)))
print("Residual sum of squares (MSE): %.2f" % np.mean((prediction2 - test_y) ** 2))
print("R2-score: %.2f" % r2_score(test_y,prediction2 ) )
```

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

Calculating R2-score for Random Forest Regression

```
from sklearn.metrics import r2_score

print("Mean absolute error: %.2f" % np.mean(np.absolute(prediction3 - test_y)))
print("Residual sum of squares (MSE): %.2f" % np.mean((prediction3 - test_y) ** 2))
print("R2-score: %.2f" % r2_score(test_y,prediction3 ) )
```

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

Fig6.2 Finding the R2,MAE,MSE