MINOR PROJECT REPORT

Submitted in partial fulfillment for the award of the degree of

BACHELOR OF TECHNOLOGY
(Department of Information Technology)

Submitted to

INDIAN INSTITUTE OF INFORMATION TECHNOLOGY BHOPAL (M.P.)

Submitted by

Harshit Chaudhary (20U03071)
Jimish Prajapati (20U03072)

Under the supervision of

Dr. Neeraj Kumar

Designation & Affiliation

INDIAN INSTITUTE OF INFORMATION TECHNOLOGY BHOPAL (M.P.)

Certificate

This is to certify that the minor project report entitled “Whether Prediction Model Using Machine Learning” is submitted by 1. Harshit Chaudhary (20U03071) 2. Jimish Prajapati (20U03072) – Indian institute of Information Technology in fulfilment of the requirements for the degree of Bachelor of Technology in Department of Information Technology. This project is an authentic work done by them under my supervision and guidance.

This project has not been submitted to any other institution for the award of any degree.

Date: 19/04/2023

Dr. Neeraj Kumar

Department of Information Technology

IIIT Bhopal

Minor Project Co-ordinator

Dr. Vishakha Chourasia

Department of Information Technology

IIIT Bhopal

INDIAN INSTITUTE OF INFORMATION TECHNOLOGY BHOPAL (M.P.)

Student Declaration

I hereby declare, that the work presented in the project report entitled **"Whether Prediction Model Using Machine Learning"** in partial fulfilment of the requirement for the award of degree of **"Bachelor of Engineering"** from **INDIAN INSTITUTE OF INFORMATION TECHNOLOGY, BHOPAL** is record of my own work.

I, with this, declare that the facts mentioned above are true to the best of our knowledge. In case of any unlikely discrepancy that may occur, we will be the ones to take responsibility.

Date: 19/04/2023

Place: IIIT Bhopal

Harshit Chaudhary
(20U03071)

Jimish Prajapati
(20U03072)

INDIAN INSTITUTE OF INFORMATION TECHNOLOGY BHOPAL (M.P.)

Acknowledgement

With immense pleasure I, **Mr. Harshit Chaudhary and Mr. Jimish Prajapati** presenting "Whether Prediction Model Using Machine Learning" minor project report as a part of curriculum of "Bachelor of Engineering". I wish to thank all the people who gave me unending support.

I express my profound thanks to my project supervisor "Dr. Neeraj Kumar" and all those who have indirectly guided and helped me in preparation of the report.

Harshit Chaudhary (20U03071)

Jimish Prajapati (20U03072)

Contents

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 favourable position of machine learning is that it does not depend on the physical laws of environmental processes.

[1]Weather forecasting provides numerous societal benefits, from extreme weather warnings to agricultural planning. In recent decades, advances in forecasting have been rapid, arising from improved observations and models, and better integration of these through data assimilation and related techniques. Further improvements are not yet constrained by limits on predictability. Better forecasting, in turn, can contribute to a wide range of environmental forecasting, from forest-fire smoke to bird migrations.

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 Decision tree regression and compare the predicted values to the actual values and error estimate though plots and graphs and R2 scores. 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 forecasting is the task of predicting the atmosphere in a future time and area. This was done using physics equations in the early days when the atmosphere was considered a fluid. The current state of the environment is checked and the future state is predicted by solving these equations numerically, but we are not able to determine very accurate weather for more than 10 days and this can be improved with science and technology.

Machine learning can be used to process instantaneous comparisons of historical weather forecasts and observations. Machine learning weather models can better account for forecast inaccuracies, such as overestimated precipitation, and produce more accurate forecasts. Temperature prediction is essential in a large number of applications, including studies related to climate, energy, agriculture, medicine, etc.

There are many types of machine learning calculations that are linear regression Polynomial regression, random forest regression, artificial neural network and Recurrent Neural Network. These models are prepared depending on the authenticity provided information about any area. The contribution to these models is given e.g. if forecast temperature, lowest temperature, average air mass, highest temperature, average humidity and order for 2 days. In light of this minimum temperature and maximum A temperature of 7 days will be reached.

Machine Learning

Machine learning is a rapidly evolving field that has the potential to revolutionize many aspects of society, from healthcare to finance to transportation. At its core, machine learning involves training algorithms on large datasets to make predictions or decisions based on new data. This approach has already shown promise in a wide range of applications, including image and speech recognition, natural language processing, and recommender systems.

One area where machine learning has particularly significant potential is in predictive modelling. By training models on large datasets, machine learning algorithms can identify patterns and relationships that might be difficult for human analysts to discern. This enables more accurate predictions and insights that can inform decision-making and drive innovation.

However, there are also challenges associated with machine learning, such as the need for large amounts of high-quality data, the potential for bias, and the difficulty of interpreting complex models. Nonetheless, as the field continues to advance and new techniques are developed, machine learning has the potential to transform how we approach many of the world's most pressing problems.

Weather prediction is an important field that has a significant impact on many aspects of society, from agriculture to transportation to disaster response. However, accurately predicting weather patterns is a complex and challenging task that has historically relied on statistical models and expert analysis. With the growth of machine learning, there is now potential for significant improvement in the accuracy of weather predictions.

According to a study by the National Oceanic and Atmospheric Administration (NOAA), machine learning algorithms can improve the accuracy of precipitation forecasts by up to 25% (NOAA, 2020). Machine learning algorithms can be trained on large datasets of historical weather data to identify patterns and relationships that can help to inform future predictions. In addition, machine learning algorithms can incorporate a wide range of data sources, including

satellite imagery, weather station data, and social media data, to provide more accurate and timely predictions.

While machine learning has shown significant promise in weather prediction, there are also challenges associated with the approach, such as the need for large amounts of high-quality data and the potential for overfitting. Nonetheless, with continued development and refinement, machine learning has the potential to significantly improve our ability to predict and prepare for weather events, which could have important implications for public safety and economy.

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 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

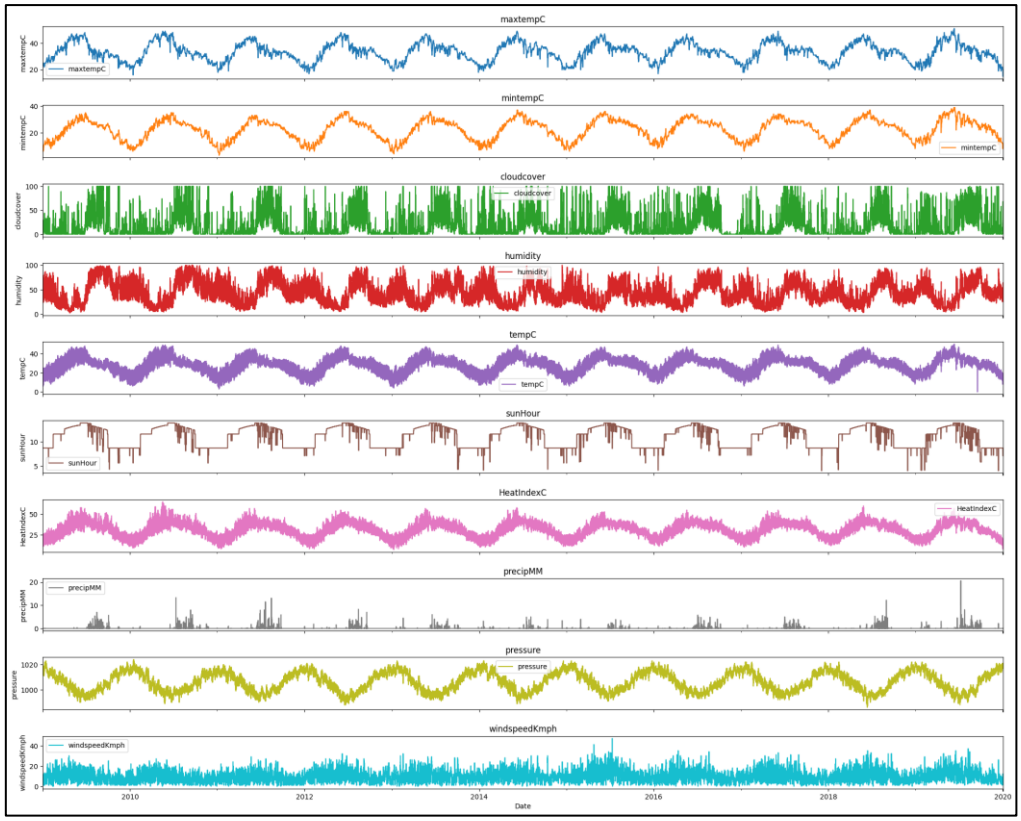


Figure 2.1: Plot for each factor for 10 years

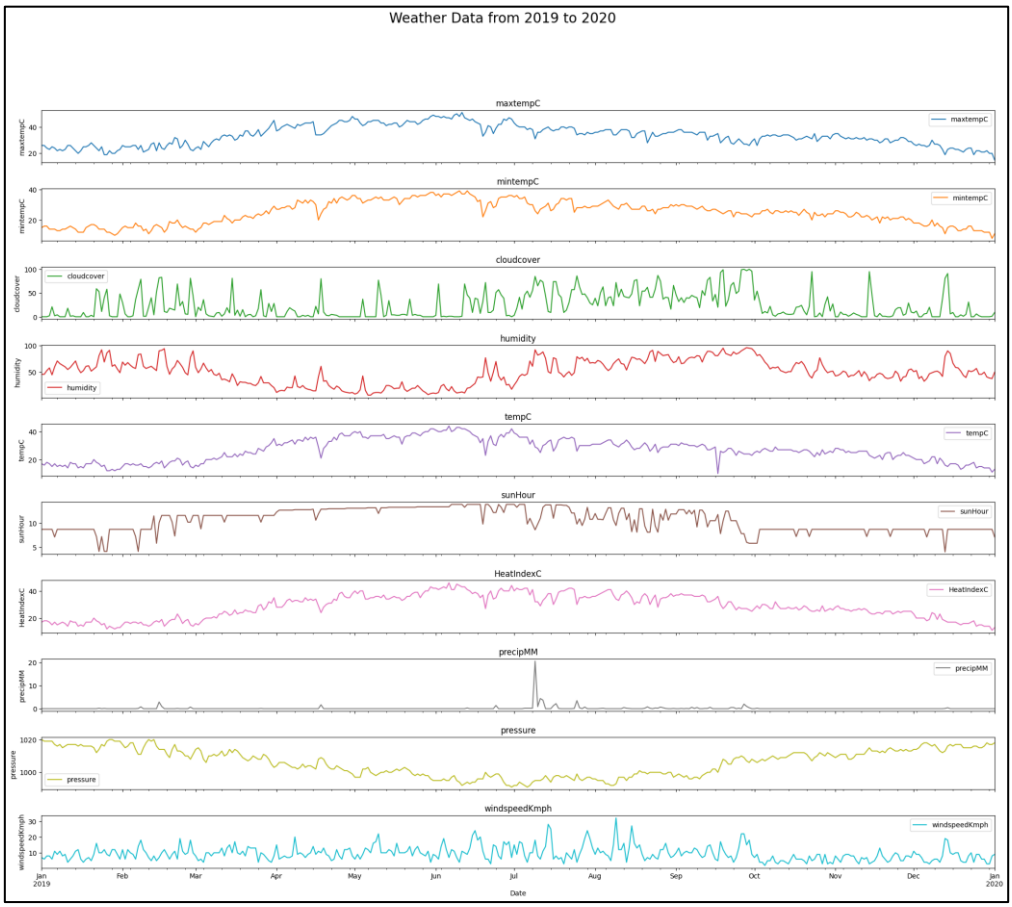


Figure 2.2: Plot for each factor for 1 year

Multiple Linear Regression

[3] Multiple linear regression refers to a statistical technique that is used to predict the outcome of a variable based on the value of two or more variables. It is sometimes known simply as multiple regression, and it is an extension of linear regression. The variable that we want to predict is known as the dependent variable, while the variables we use to predict the value of the dependent variable are known as independent or explanatory variables.

Multiple linear regression Formula

The formula for a multiple linear regression is:

$$y = \beta_0 + \beta_1 X_1 + \ldots + \beta_n X_n + \epsilon$$

- y = the predicted value of the dependent variable
- B_0 = the y-intercept (value of y when all other parameters are set to 0)
- $B_1 X_1$ = the regression coefficient (B_1) of the first independent variable (X_1) (a.k.a. the effect that increasing the value of the independent variable has on the predicted y value)
- … = do the same for however many independent variables you are testing
- $B_n X_n$ = the regression coefficient of the last independent variable
- ϵ = model error (a.k.a. how much variation there is in our estimate of y)

To find the best-fit line for each independent variable, multiple linear regression calculates three things:

- The regression coefficients that lead to the smallest overall model error.
- The *t* statistic of the overall model.
- The associated *p* value (how likely it is that the *t* statistic would have occurred by chance if the null hypothesis of no relationship between the independent and dependent variables was true).

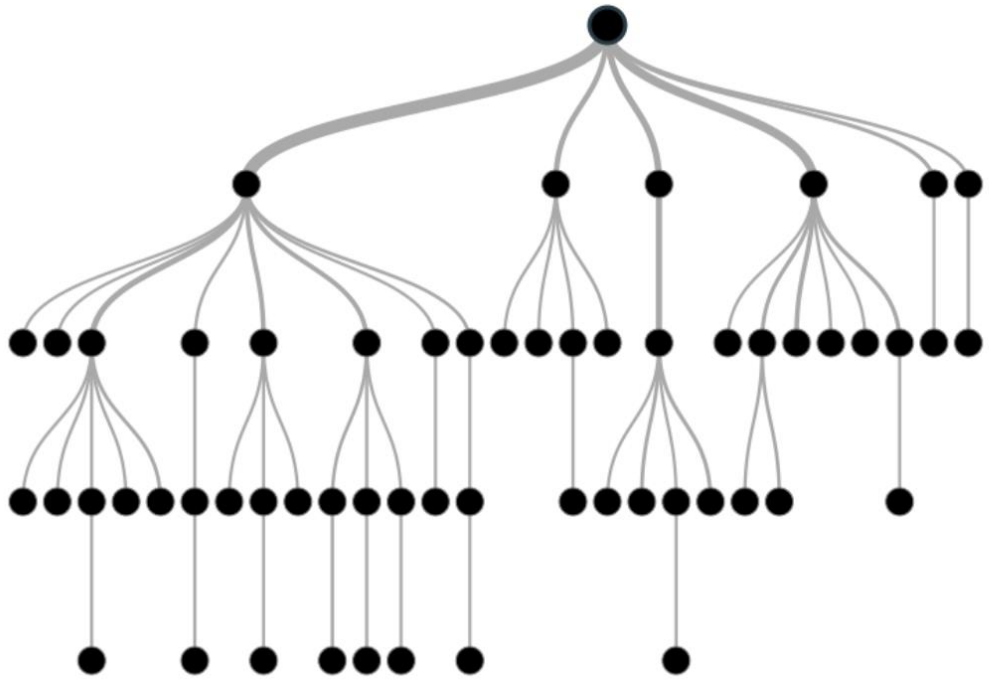
It then calculates the *t* statistic and *p* value for each regression coefficient in the model.

Decision Tree Regression

[2]A decision tree is a decision bolster device that uses a treelike graph or model of decisions and their conceivable results, including chance occasion results, asset expenses, and utility. It is one approach to show a calculation. A Decision Tree is a stream outline like tree structure. Each node indicates a test on a property. Every branch speaks to a result of the test. Leaf nodes speak to class appropriation. The choice tree structure gives an express arrangement of "assuming then" guidelines making the outcomes simple to translate. In the tree structures, leaves speak to groupings and branches speak to conjunctions of components that prompt those arrangements. Formally, data increase is characterized by entropy. In other to enhance the precision and speculation of arrangement and relapse trees, different systems were presented like boosting and pruning. Boosting is a procedure for enhancing the precision of a prescient capacity by applying the capacity over and again in an arrangement and consolidating the yield of every capacity with weighting so that the aggregate blunder of the forecast is minimized or growing various free trees in parallel and join them after all the trees have been created. Pruning

is completed on the tree to enhance the span of trees and along these lines decrease overfitting which is an issue in extensive, single-tree models where the model starts to fit commotion in the information. At the point when such a model is connected to information that was not used to construct the model, the model won't have the capacity to sum up. Numerous decision tree calculations exist and these include: Alternating Decision Tree, Logit help Alternating Decision Tree (LAD), C4.5 and Classification and Regression Tree (CART).

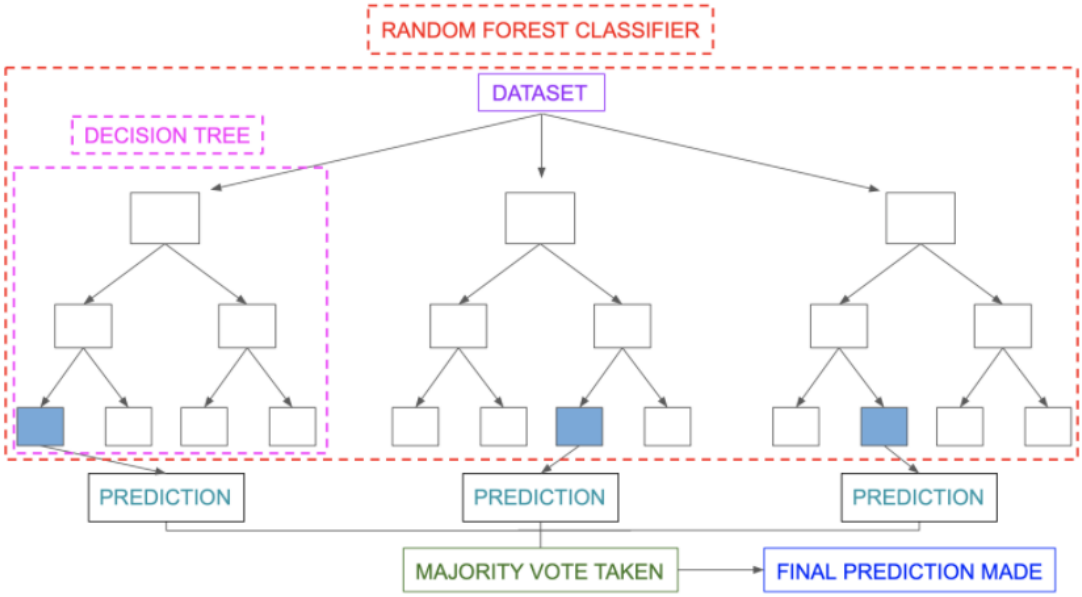
Decision tree constructs arrangement or relapse models as a tree structure. It separates a datasets into littler and littler subsets while in the meantime a related decision tree is incrementally created. The last result is a tree with decision nodes and leaf nodes. A decision node (e.g., Outlook) has two or more branches (e.g., Sunny, Overcast and Rainy). Leaf node (e.g., Play) speaks to an order or choice. The highest decision node in a tree which relates to the best indicator called root node. Decision trees can deal with both clear cut and numerical information.



Random Forest Regression

[4] Random Forest Regression is a supervised learning algorithm that uses ensemble learning method for regression. Ensemble learning method is a technique that combines predictions from multiple machine learning algorithms to make a more accurate prediction than a single model.

A Random Forest Regression model is powerful and accurate. It usually performs great on many problems, including features with non-linear relationships. Disadvantages, however, include the following: there is no interpretability, overfitting may easily occur, we must choose the number of trees to include in the model.



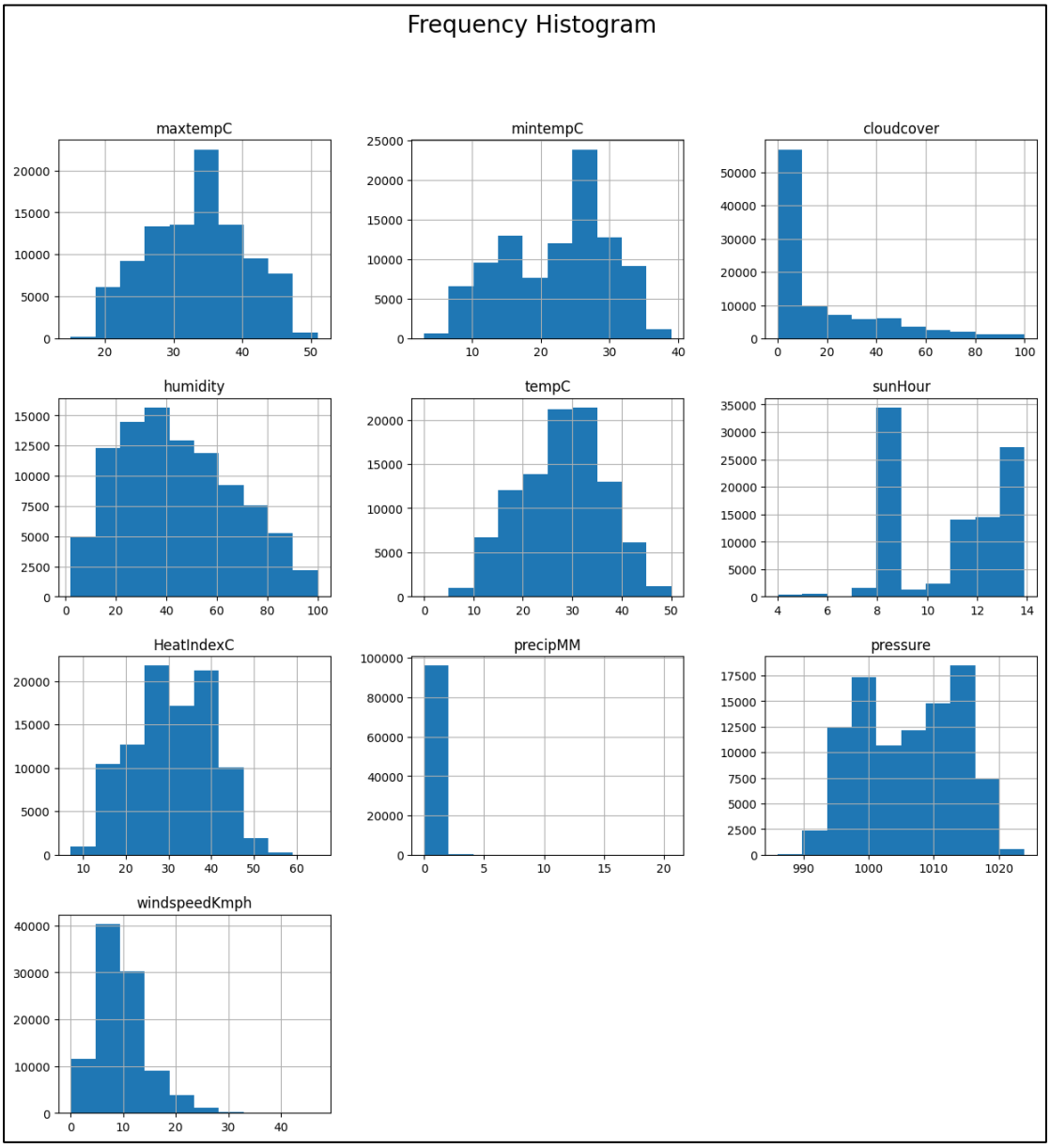
Difference between Decision Tree regressor and Random Forest Regressor

Decision Trees	**Random Forests**
Decision trees normally suffer from the problem of overfitting if it's allowed to grow till its maximum depth.	Random forests use the bagging method. It creates a subset of the original dataset, and the final output is based on majority ranking and hence the problem of overfitting is taken care of.
A single decision tree is faster in computation.	It is comparatively slower.
When a data set with features is taken as input by a decision tree it will formulate some set of rules to do prediction.	Random forest randomly selects observations, builds a decision tree and the average result is taken. It doesn't use any set of formulas.

Hence, we can come to a conclusion that random forests are much more successful than decision trees only if the trees are diverse and acceptable.

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, 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.



4. RESULT AND DISCUSSION

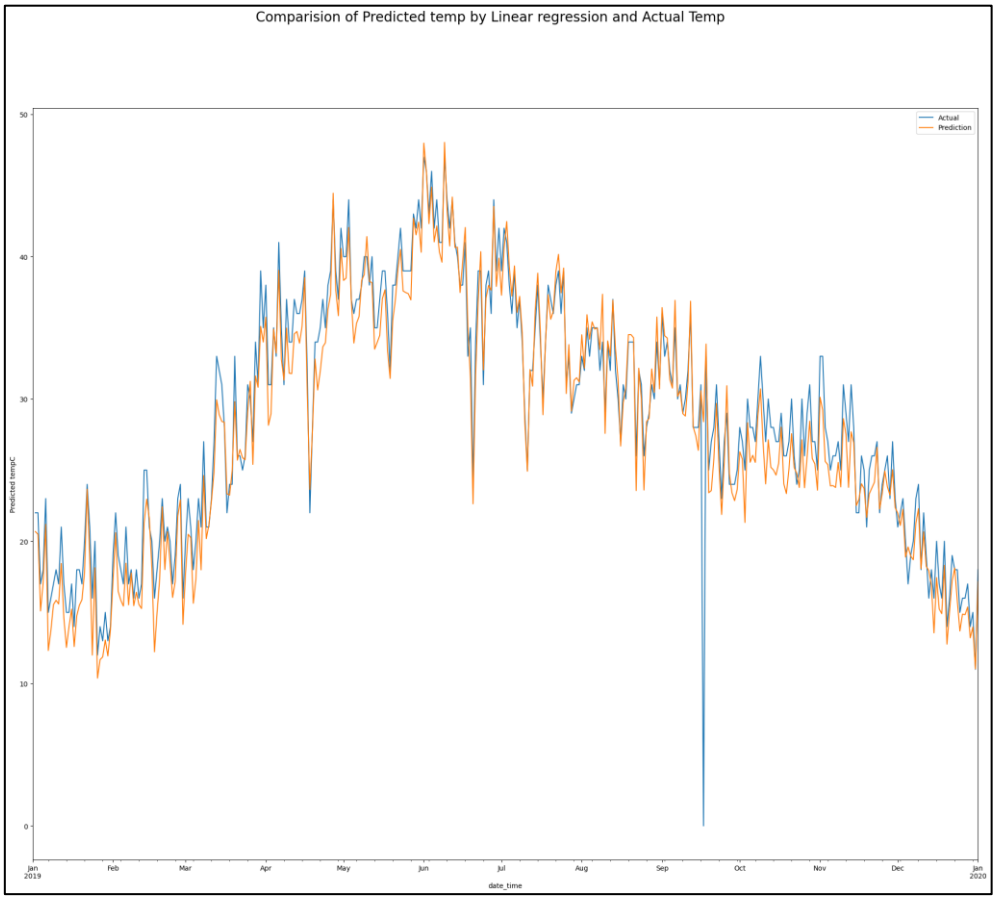
The results of the implementation of the project are demonstrated below.

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

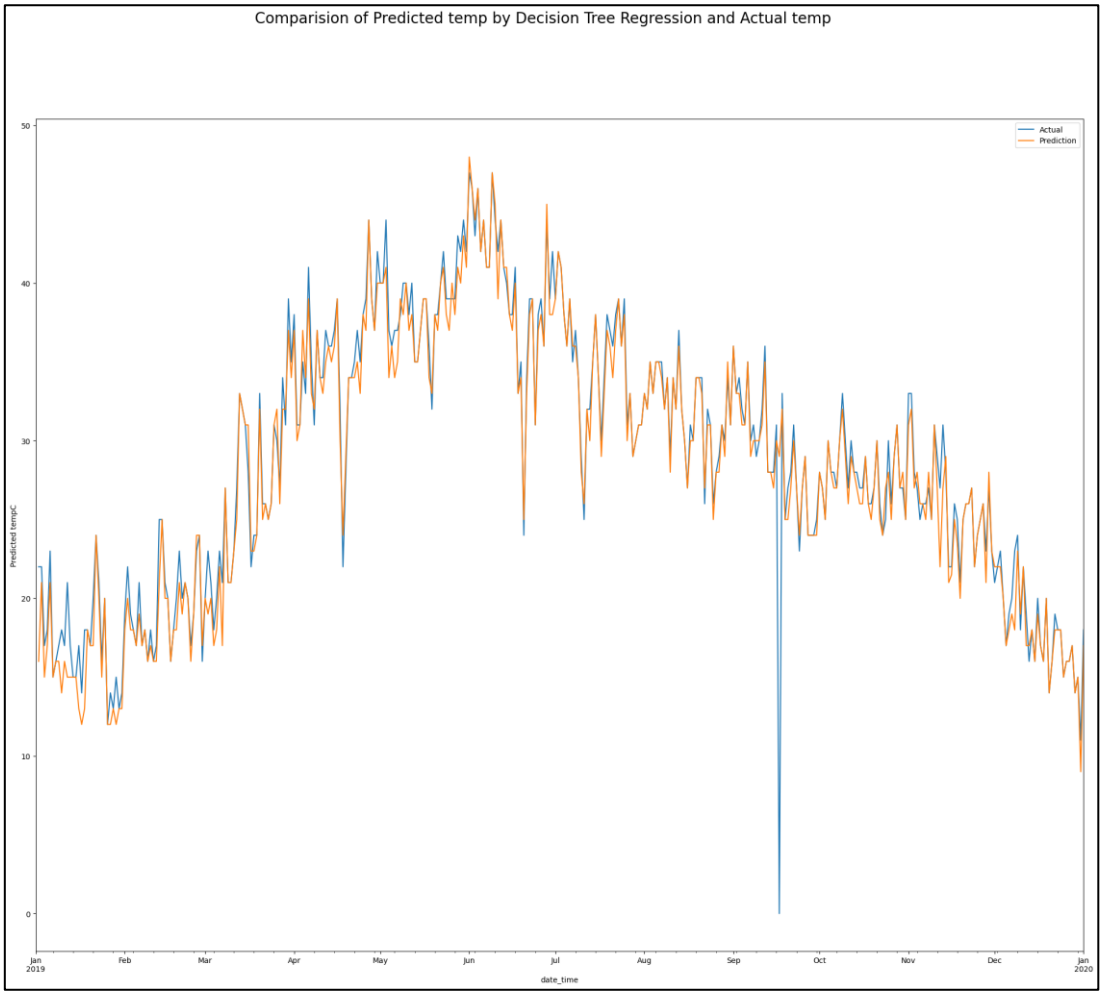


Decision Tree Regression:

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	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

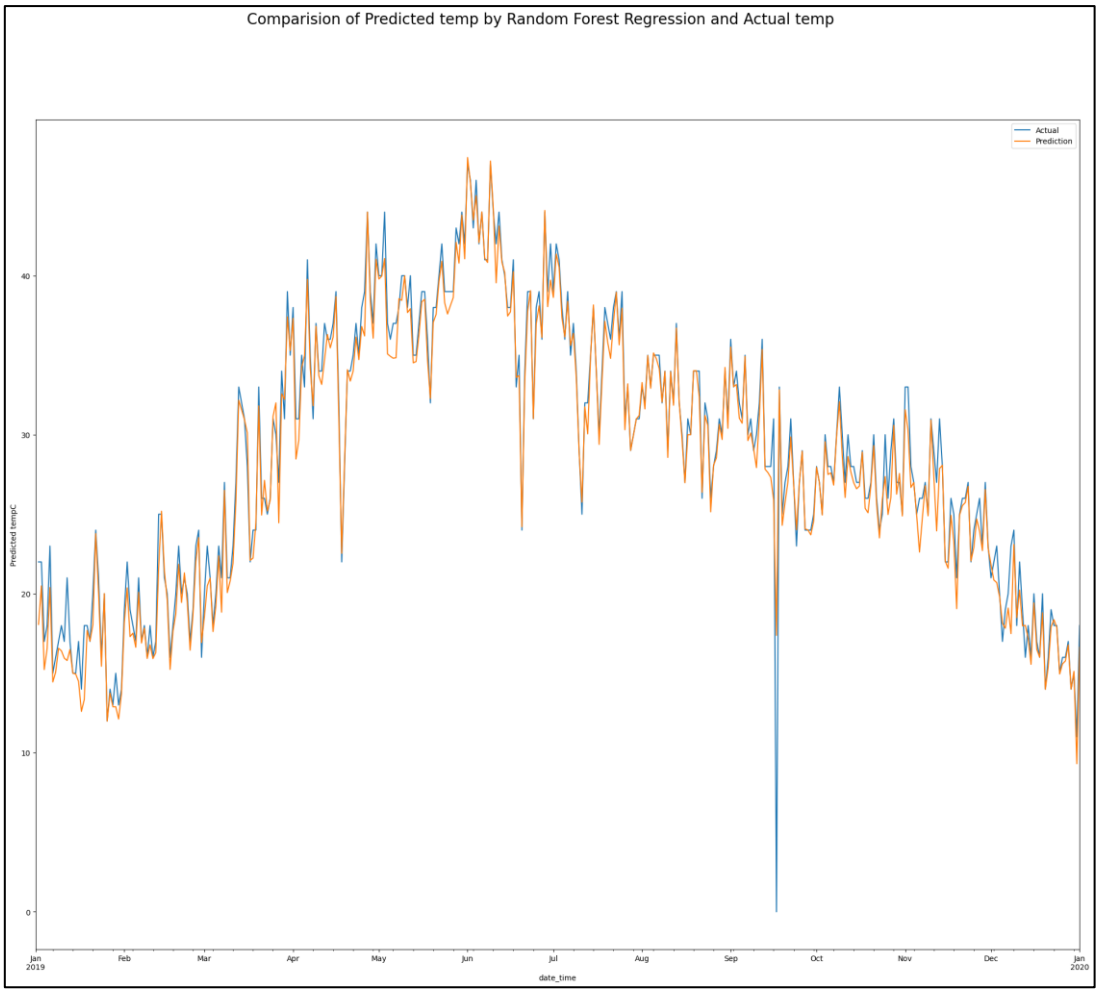


Random Forest Regression:

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

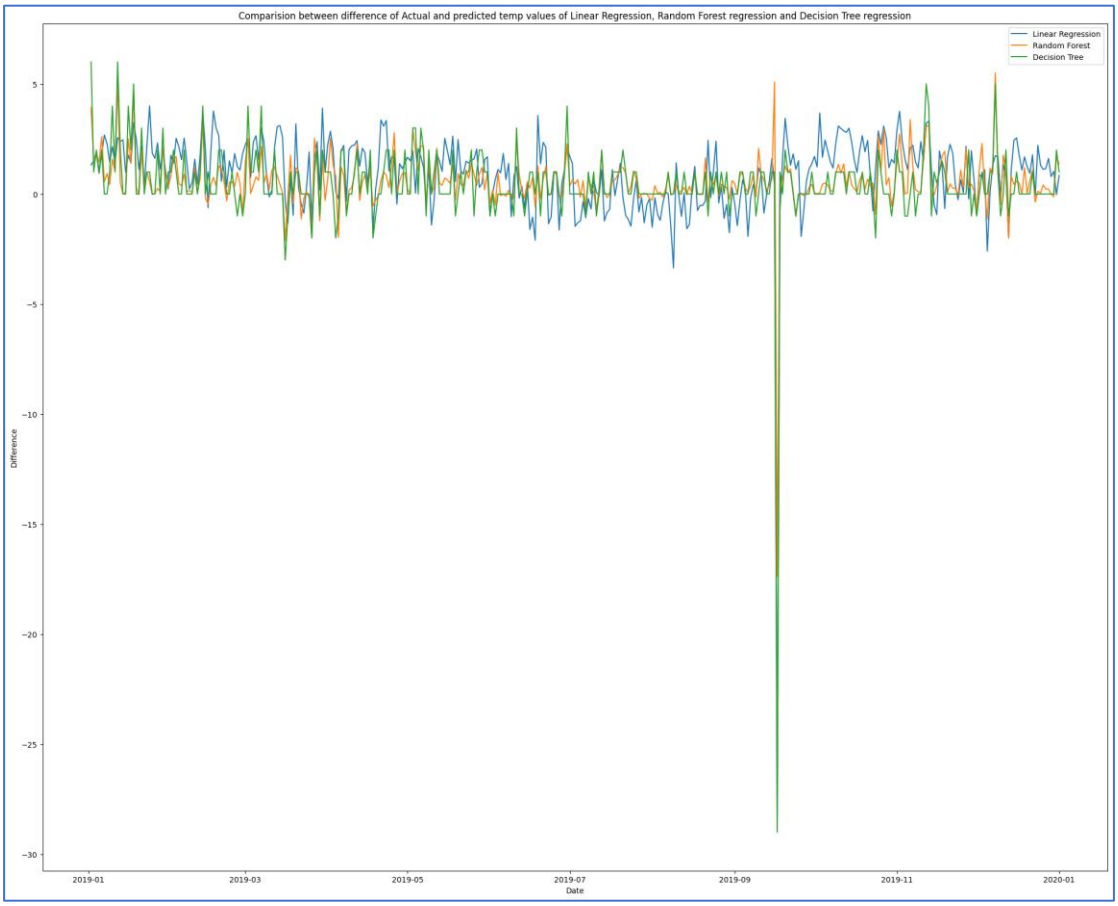


Comparison

The comparison of the difference between the actual and predicted values is a common way to evaluate the performance of a machine learning model. In this analysis, we computed the difference between the actual and predicted values for each model (linear regression, random forest, and decision tree) and plotted the results over time.

The results show that all three models are able to predict the target variable reasonably well, with relatively small differences between the actual and predicted values. However, there are some periods where the differences are larger, suggesting that the models may be less accurate during these times.

Overall, the comparison of the difference between the actual and predicted values provides a useful way to evaluate the performance of a machine learning model and can help identify areas where the model may need improvement. It is important to note that this is just one of many ways to evaluate a machine learning model, and other metrics such as accuracy, precision, recall, and F1-score may also be useful depending on the specific problem and application.



5. 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.

REFERENCES

[1] RICHARD B. ALLEY, KERRY A. EMANUEL, AND FUQING ZHANG Advances in weather prediction pp.342-343, 2019

[2] Siddharth S. Bhatkande, Roopa G. Hubballi, "Weather Prediction Based on Decision Tree Algorithm Using Data Mining Techniques", IJARCCE, pp.485, 2016

[3] Sebastian Taylor " Multiple Linear Regression" , March 6, 2023

[4] chaya "Random Forest Regression" , Jun 9 , 2020.