Table of Contents

[Table of figures and tables 2](#_Toc132877002)

[Abstract. 3](#_Toc132877003)

[Objective 4](#_Toc132877004)

[Introduction. 5](#_Toc132877005)

[Data Cleaning and Analysis. 5](#_Toc132877006)

[Modeling 11](#_Toc132877007)

[ARIMA and SARIMA 11](#_Toc132877008)

[Multi Linear Regression 11](#_Toc132877009)

[Deep Learning (LSTM) 14](#_Toc132877010)

[H-test Predictions Visualization 15](#_Toc132877011)

[Forecast Function 16](#_Toc132877012)

[Summary and Conclusion 18](#_Toc132877013)

# Table of figures and tables

[Figure 1: First 5 rows 5](#_Toc132877021)

[Figure 2: Missing Values 5](#_Toc132877022)

[Figure 3: Descriptive Statistics 6](#_Toc132877023)

[Figure 4: Meantemp between 2013-2017 6](#_Toc132877024)

[Figure 5: Plots of All features 7](#_Toc132877025)

[Figure 6: Meanpressure Outlier fixed 8](#_Toc132877026)

[Figure 7: PACF 8](#_Toc132877027)

[Figure 8: Heatmap 9](#_Toc132877028)

[Figure 9: Pearson Correlation Matrix 9](#_Toc132877029)

[Figure 10: Decomposition 10](#_Toc132877030)

[Figure 11:Feature Importance 10](#_Toc132877031)

[Figure 12: Baseline Models 11](#_Toc132877032)

[Figure 13:ARIMA and SARIMA Evaluations 11](#_Toc132877033)

[Figure 14: Multi-Linear Regression 11](#_Toc132877034)

[Figure 15: 1 Step Prediction 11](#_Toc132877035)

[Figure 16: Multi-Linear Regression 12](#_Toc132877036)

[Figure 17: ACF Residuals 12](#_Toc132877037)

[Figure 18: Q-Variance 13](#_Toc132877038)

[Figure 19: Variance and Mean residual 13](#_Toc132877039)

[Figure 20: LSTM Summary 14](#_Toc132877040)

[Figure 21: LSTM Evaluation 14](#_Toc132877041)

[Figure 22: H-test Prediction on SARIMA 15](#_Toc132877042)

[Figure 23: Arima h-test 16](#_Toc132877043)

[Figure 24: SARIMA Forecast Function 16](#_Toc132877044)

# Abstract.

The main aim for this project is to explore various Machine Learning algorithms with a sole purpose of analyzing and developing models for Indian climate. This project will have the following parts; data loading, data cleaning and visualization, modeling and evaluation. Some of the models we are going to train here are ARIMA, SARIMA, and LSTM.

# Objective

The main objective of this project is to predict future meantemp in Delhi.

# Introduction.

Time series refers to tracing data incrementally over a period of time. Data points in time series are recorded sequentially. Time series applies in every aspect of life since time is everywhere. For instance, an e-commerce firm can record daily sales and use this data in future to predict profits. In our case, we are using time series data to predict Indian climate between 2013 and 2017. This report will contain analysis of time series, cleaning, feature engineering, performing various modeling algorithms such as ARIMA and LSTM, and forecasting.

# Data Cleaning and Analysis.

I am using DailyDelhiClimateTrain.csv which has date, meantemp, humidity, wind\_speed, and meanpressure columns. The date is of type object while the other three features of the type float64. While working with this data, we are going to use set date as index and then work with other 4 features. Below is a section of data we will be working with:

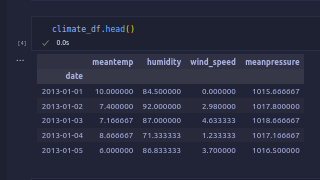


Figure 1: First 5 rows

The csv has no missing values as shown in the figure 2 below.

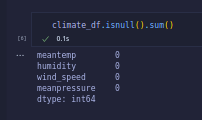


Figure 2: Missing Values

When we perform descriptive statics in our data, we observe that the data is well balanced. Figure 3 below shows the descriptive statics;

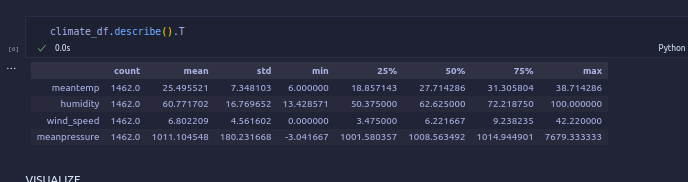


Figure 3: Descriptive Statistics

Let’s now do some plots on dependent variables with date.

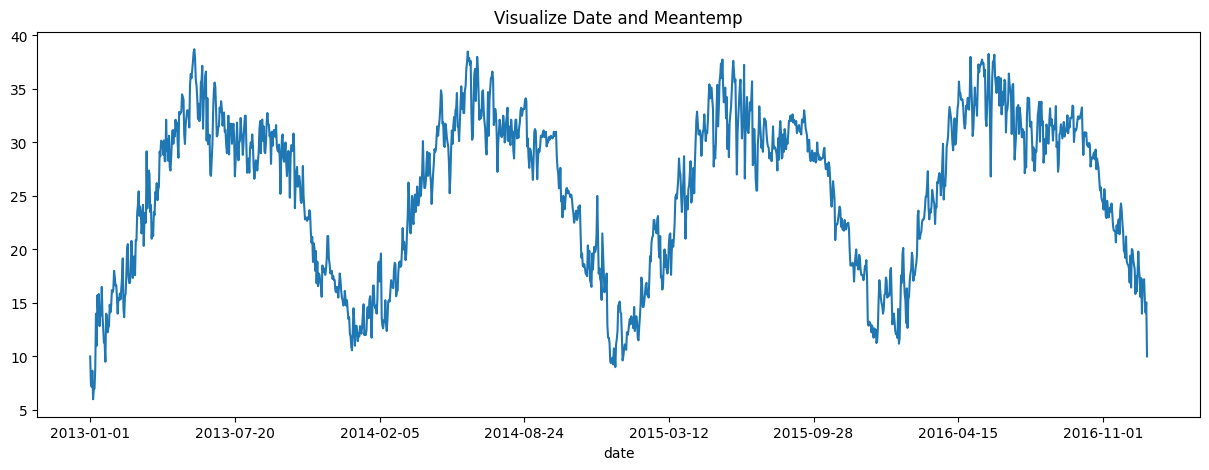


Figure 4: Meantemp between 2013-2017

From Figure 4 above, we observe that temperature rises exponentially in the first quarter of the year, remains high during the second quarter, seems to be constant during third quarter, and starts to reduce during the last quarter. This is the case in all yearly between 2013 and 2027.

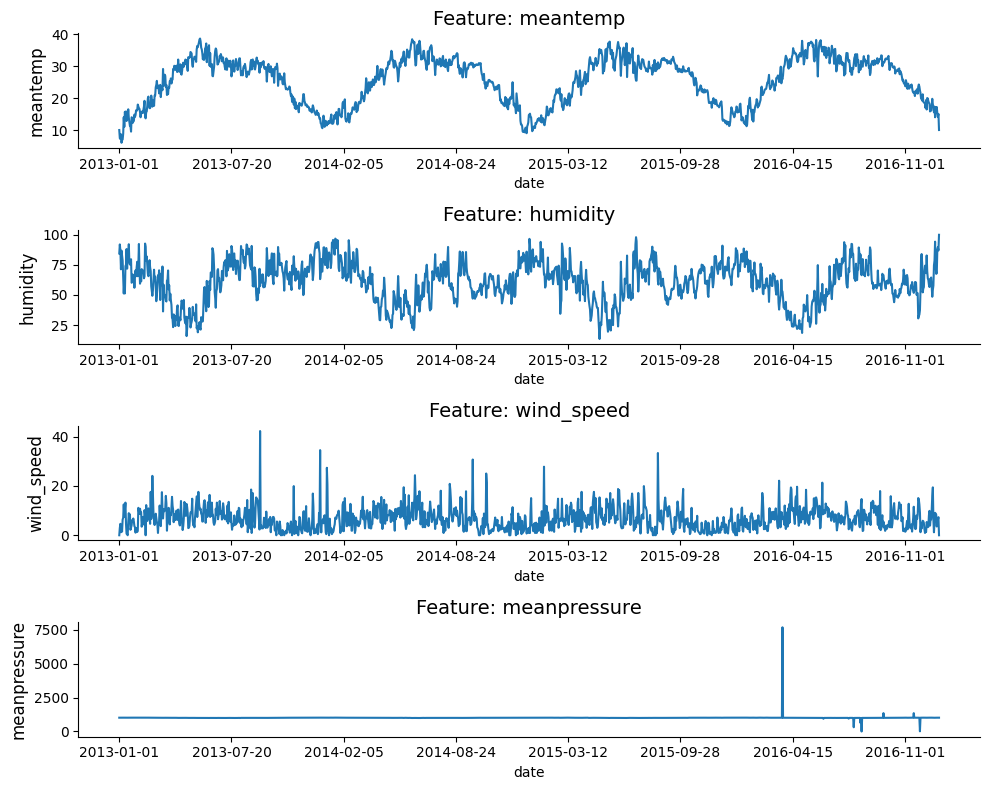


Figure 5: Plots of All features

From figure 5 above, we observe that:

The time series has constant variation and is stationary. Notably, meanpreassure seems to have some abnormal behavior between 2016 and 2017. This abnormality is due to the fact there are outliers in the time series.

The plot below shows the outlier problem fixed from the meanpressure features. This is because the plot seems to have assumed normal behaviour as humidity, meantemp, and wind\_speed.

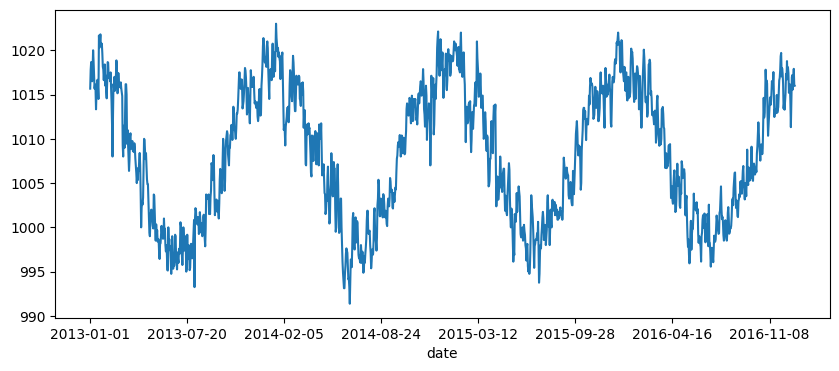


Figure 6: Meanpressure Outlier fixed

PACF of the dependent variable is as shown below.

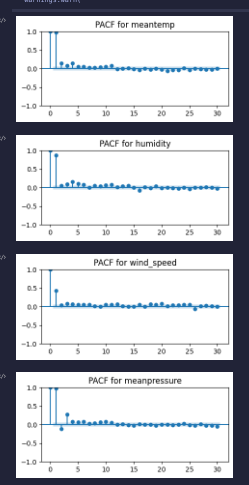


Figure 7: PACF

PACF observation:

From PACF of the dependent variables above, we can see that, correlation decays to zero and this shows that it is constant and stationery.

Let’s now have a look heatmap and Pearson Correlation matrix:

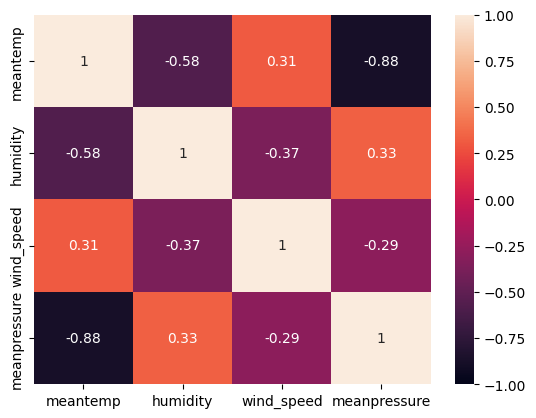


Figure 8: Heatmap

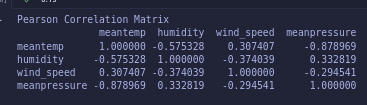


Figure 9: Pearson Correlation Matrix

From above heatmap and pearson correlation matrix, humidity, meanpressure, and to meantemp have high correlation as compared to other features. We can also see that, wind\_speed increases with decrease in humidity and meanpressure. Let's confirm this by performing decomposition below;

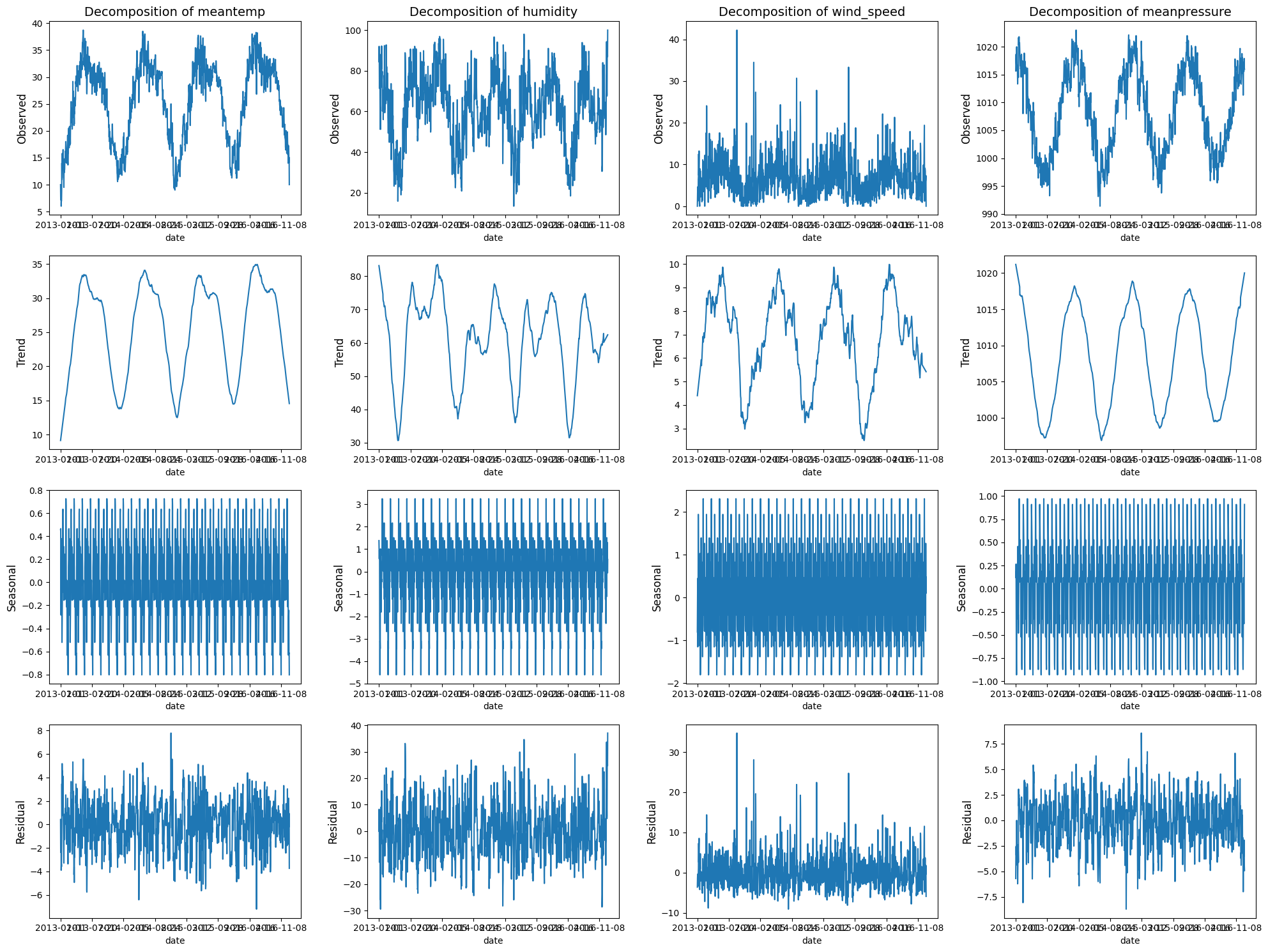


Figure 10: Decomposition

When we check feature importance for, meanpressure and humidity seems to have more effect on the target variable.

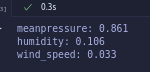


Figure 11:Feature Importance

# Modeling

In this part, I did base models, ARIMA and SARIMA, and LSTM. For base models, I fit the data in the average, naïve, drift, and simple algorithms. I then evaluated the respective models using Mean Squared Error (MSE) and the following were the results.

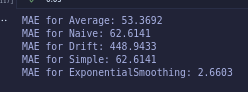


Figure 12: Baseline Models

## ARIMA and SARIMA

The following figure shows evaluation of ARIMA and SARIMA.



Figure 13:ARIMA and SARIMA Evaluations

When we compare evaluated values between Baseline models and SARIMA, we find out that, the only model that performs better than SARIMA is Average and Exponential Smoothing.

## Multi Linear Regression

I did a multi-linear modelling on a time series dataset and got an accuracy of 91.74% as shown in the figure below:

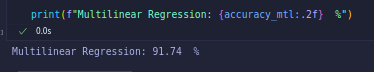


Figure 14: Multi-Linear Regression

I performed 1 step prediction and the following were the results:

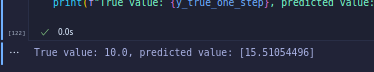


Figure 15: 1 Step Prediction

When we compare the true value and predicted value in Figure 15 above, we observe that there is a difference of 5 which is not very large.

Here is an evaluation of Multi-Linear Regression

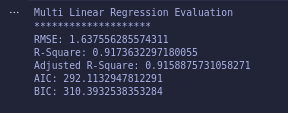


Figure 16: Multi-Linear Regression

ACF of residuals:

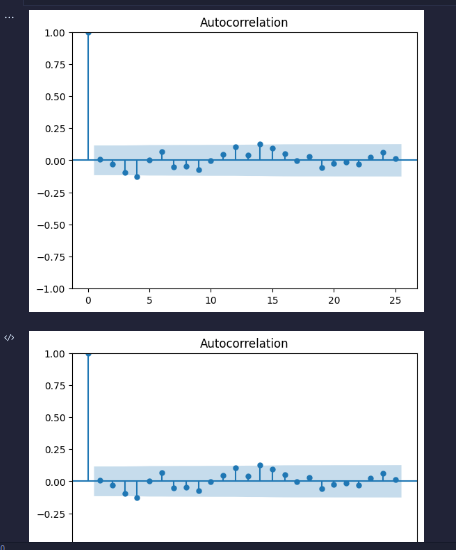


Figure 17: ACF Residuals

Q-Variance results:

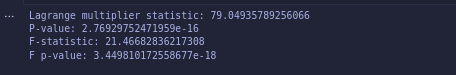


Figure 18: Q-Variance

From the Q-variance, we can see that p-value is greater than 0.05, implying that our time series is stationary.

Variance and Mean Residual Results:

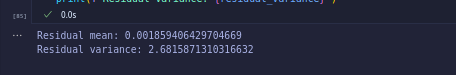


Figure 19: Variance and Mean residual

## Deep Learning (LSTM)

I designed an LSTM network with 2 LSTM Layers of 50 neurons each, a 1 Dense Layer of 25 neurons and 1 Dense Layer of 1 neuron. The figure below shows a summary of the LSTM network:

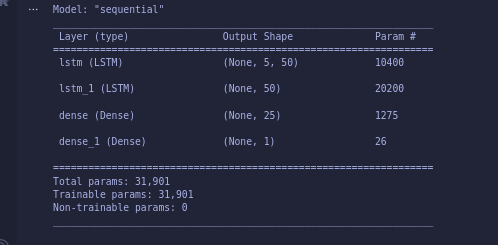


Figure 20: LSTM Summary

Evaluation of LSTM model is as shown in the Figure below:

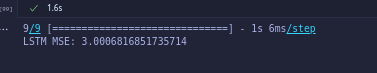


Figure 21: LSTM Evaluation

When we compare all models trained, I would choose LSTM model since it performs better than both SARIMA and ARIMA. This is as evident in the Mean Squared Error where by LSTM has an MSE of 3, ARIMA has MSE of 59, and SARIMA 56. The smaller the error the better the model and thus LSTM is by far the best.

When we compare ARIMA and SARIMA, obviously I will choose SARIMA since it has a lower Mean Squared Error.

## H-test Predictions Visualization

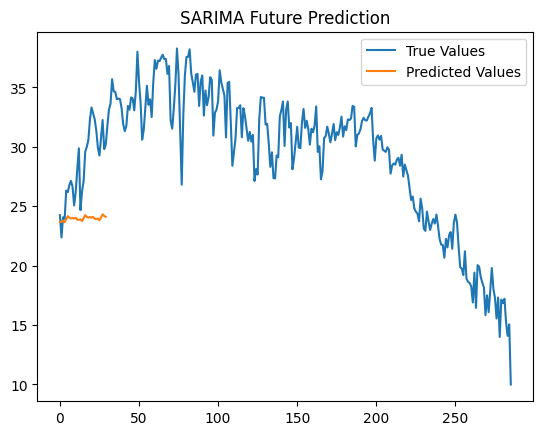


Figure 22: H-test Prediction on SARIMA

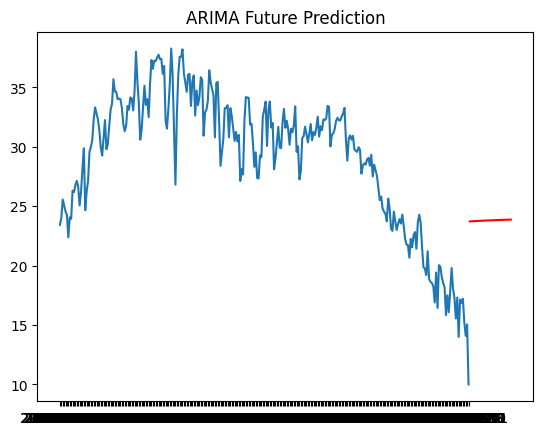


Figure 23: Arima h-test

## Forecast Function

The following is a SARIMA forecast function to predict future meantemp.

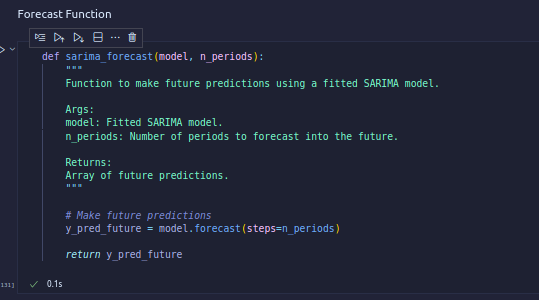


Figure 24: SARIMA Forecast Function

After applying the above function, I got the following results with 24 steps

[23.65598622 23.69852869 23.76537511 23.65500745 23.90114782 24.1694459

24.03721306 23.96640614 24.01917101 23.95806399 24.02882253 23.8666022

23.84941897 23.88193965 23.75184421 23.98664657 24.24842843 24.11245056

24.0394913 24.09101917 24.02920121 24.09955116 23.937096 23.91977781]

## Summary and Conclusion

SARIMA model has various limitations such us being unable to handle complex variances and non-stationary time series data and also being not able to handle data with outliers. The best method to handle such problems in future is to use Deep Learning model such as LSTM.