M.Sc. Data Analytics and Technologies

DAT7303- Data Mining and Machine Learning

Assessment 1- Portfolio 4

Submitted By:

Student ID: 2310413

Submitted To:

Pradeep Hewage

Module Instructor

Date: 10th May 2024

Time of workshop session: 13:00 pm

Table of contents

List of Figures **3**

1.0 Introduction **4**

2.0 Exploratory Data Analysis **5**

3.0 Time series data analytics **8**

3.1 Stationarity 11

3.2 Augmented Dickey-Fuller Test 11

4.0 ARIMA **13**

5.0 Machine learning models for TSA **16**

5.1 Linear Regression Model 16

5.2 Support Vector Regression 17

5.3 Decision Tree Regression 17

5.4 Random Forest 17

5.5 Model Evaluation and Comparison 18

6.0 Reporting Results **19**

7.0 Conclusion **21**

References **22**

List of Figures

Figure 1: Weather Research & Forecasting (WRFdata_May2023) dataset

Figure 2 : The data description of dataset

Figure 3 : Missing values in WRFdata_May2023) dataset

Figure 4 : Histogram of Q2 parameter

Figure 5 : Time series analysis plot for Q2 parameter

Figure 6 : ARIMA model Test.

Figure 7 : HIstogram of ARIMA model Residuals.

Figure 8 : Scatterplots for the actual vs. predicted values

Figure 9: Barplot for RMSE values

Figure 10: Forecasting Q2- 2 metre specific humidity

1.0 Introduction

The ever-growing field of data science offers a plethora of techniques to uncover insights from information. This report explores two key areas of data science: forecasting using historical data in ARIMA models and machine learning approaches. Part 1 focuses on time series analysis involving ARIMA model, a popular forecasting technique for univariate time series data. It breaks down the model, the steps in its creation, and instances illustrating the model's application. Part 2 focuses on machine learning models and subjects to its core components, algorithms, and data mining topics for strong critique. It then discusses the challenges and difficulties that come with applying these techniques in real life through a case study and examples. This report seeks to arm readers with information about the two types of models – ARIMA models and machine learning models – and how and when data science can be applied to address different issues. Time series analysis is critical in describing,revealing and predicting patterns, trends and future states of concern. It is the purpose of the current essay to begin to examine and analyse the potential of the "WRFdata_May2023" dataset. The point is to have a robust assessment that involves the understanding of the problem, EDA, preprocessing of data, choice of a classification model, evaluation of the model, visualisation, and interpretation of the findings.

2.0 Exploratory Data Analysis

Exploratory data analysis or EDA provides a description of the data using statistical and graphical techniques. This involves looking at the variables in the dataset in various ways, presenting an accurate description and summary of them.

1. Data Collection : Data collection is one of the approaches utilised in EDA. It covers the stages of identifying the necessary data and deploying it into our system. Weather forecasting involves using a 'WRFdata_May2023' data file which contains data that is used to predict weather patterns and make decisions accordingly.

	X	X.1	X01.05.2018.00.00	X.2	X.3	X.4	X.5	X.6	X.7	X.8	X.9
1	XLAT	XLONG	TSK	PSFC	"U10"	"V10"	"Q2"	RAINC	RAINNC	SNOW	TSLB
2	48.871	-11.221	NA	101418	6.9	5.5	0.00602	0.0	0.0	0.0	273.2
3	49.010	-11.240	285.2	101388	7.0	5.8	0.00603	0.0	0.0	0.0	273.2
4	49.149	-11.259	285.2	101357	7.0	6.0	0.00604	0.0	0.0	0.0	273.2
5	49.288	-11.278	285.2	101327	7.0	6.3	0.00605	0.0	0.0	0.0	273.2
6	49.427	-11.298	285.2	101296	7.0	6.6	0.00605	0.0	0.0	0.0	273.2
7	49.566	-11.317	285.2	101265	7.0	6.9	0.00606	0.0	0.0	0.0	273.2
8	49.705	NA	285.1	NA	7.0	7.1	0.00607	0.0	0.0	0.0	273.2
9	49.843	NA	285.1	101203	6.9	7.4	0.00608	0.0	0.0	0.0	273.2
10	49.982	-11.377	285.0	101172	6.8	7.7	0.00609	0.0	0.0	0.0	273.2
11	50.121	-11.397	285.0	101141	6.8	8.0	0.00610	0.0	0.0	0.0	273.2
12	50.259	-11.417	284.9	101110	6.6	8.3	0.00610	0.0	0.0	0.0	273.2
13	50.398	-11.437	284.9	101078	6.5	8.5	0.00611	0.0	0.0	0.0	273.2
14	50.537	-11.457	284.9	101047	6.4	8.8	0.00613	NA	0.0	0.0	273.2
15	50.675	-11.478	284.9	101016	6.3	9.0	0.00615	0.0	0.0	0.0	273.2
16	50.814	-11.498	284.9	100985	6.2	9.2	0.00617	0.0	0.0	0.0	273.2
17	50.952	-11.519	284.8	100954	6.1	9.4	NA	0.0	0.0	0.0	273.2
18	51.091	-11.540	284.8	100922	6.1	9.7	0.00621	0.0	0.0	0.0	273.2
19	51.229	-11.561	284.7	100891	6.0	9.9	0.00622	0.0	0.0	0.0	273.2
20	51.368	-11.582	284.7	100859	NA	10.1	0.00623	0.0	0.0	0.0	273.2
21	51.506	-11.603	284.6	100828	6.0	10.3	0.00624	0.0	0.0	0.0	273.2
22	51.644	NA	284.5	100796	5.9	10.5	0.00624	NA	0.0	0.0	273.2

Showing 1 to 22 of 5,452 entries, 2482 total columns

Figure 1: Weather Research & Forecasting (WRFdata_May2023) dataset

Parameter	Description	Measuring Unit
XLAT	Latitude	
XLONG	Longitude	
TSK	Skin temperature or surface temperature	oK (Kelvin)
PSFC	Surface pressure	Pa (Pascal)
U10	X component of wind at 10m	m/s
V10	Y component of wind at 10m	m/s
Q2	2- meter specific humidity	Kg/Kg
Rainc	Convective rain (Accumulated precipitation)	mm
Rainnc	Non-convective rain	Mm
Snow	Snow water equivalent	Kg/m2
TSLB	Soil temperature	oK
SMOIS	Soil Moisture	m3/m3

Figure 2 : The data description of dataset

Selected location for weather forecasting :

Latitude : 51.409

Longitude : -7.344

Predicting : Q2 (2- metre specific humidity)

Geographical Location: North Atlantic Ocean.

2. Data Cleaning : The practice of eliminating unnecessary variables and values from your dataset and eliminating any outliers is referred to as data cleansing. Removing missing values, outliers, and unnecessary rows/ columns.

Pre processing steps:

Generates a fresh data frame which only encompasses 300 rows picked from the initial data. Changes the values of columns (latitude and longitude) to numeric (Numbers). Converts the remaining columns to numeric using as.numeric() function.

Check the missing values using sum(is.na()). Replace the missing values with the mean of the column if exists in the dataset using apply() function and replace() function.

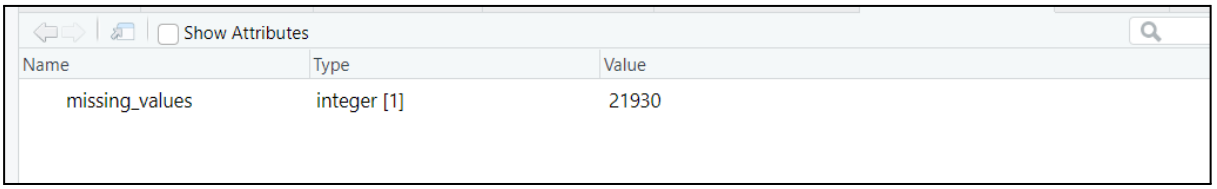

Figure 3 : Missing values in WRFdata_May2023) dataset

3. Univariate Analysis: Univariate analysis in EDA analysis looks at individual variables in order to understand their distributions and summary statistics.

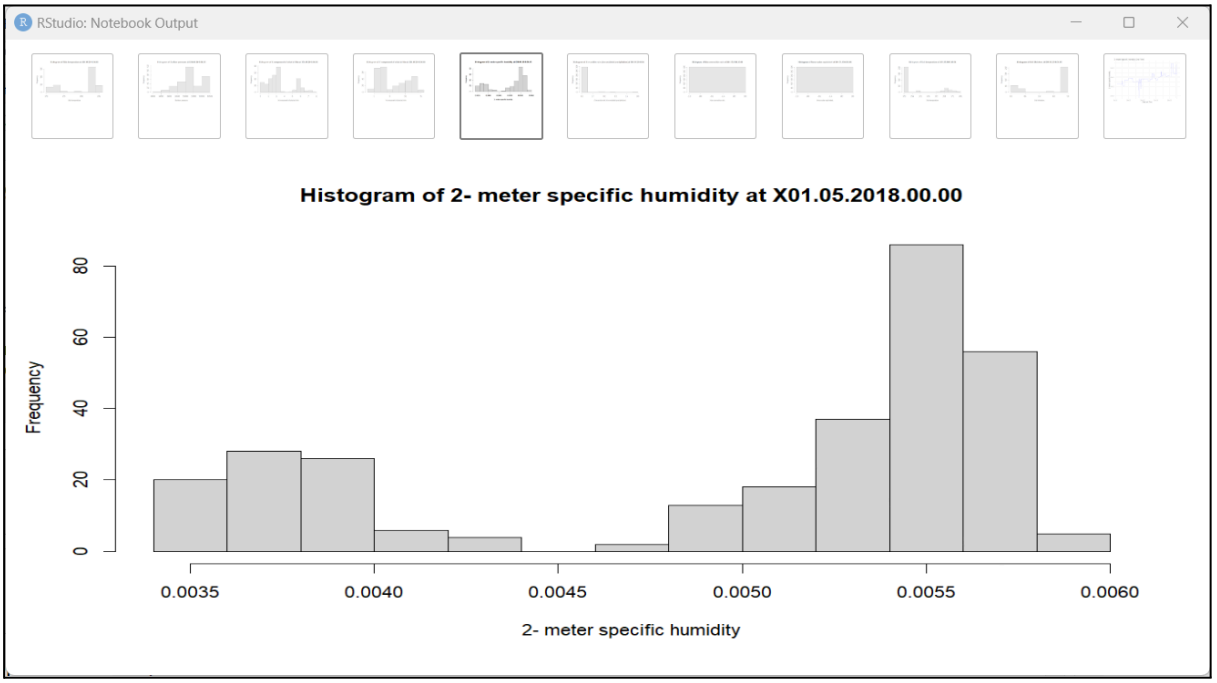

Figure 4 : Histogram of Q2 parameter

3.0 Time series data analytics

According to Pandian, “Time series analysis is a specific way of analysing a sequence of data points collected over time. In time series analysis, analysts record data points at consistent intervals over a set period rather than just recording the data points intermittently or randomly” (Pandian, 2021).

Time Series Analysis Goals :

1. **Uncover patterns and trends:** Time series analysis helps us understand how a variable changes over time and what factors influence those changes. This knowledge can be gleaned from analysing past data.
2. **Extract meaning from change:** By looking at how features within a dataset evolve over time, time series analysis provides insights into the why and how behind those changes.
3. **Forecast the future:** A core objective of time series analysis is to use past trends and patterns to predict future values of the variable being studied.
4. **Stationarity assumption:** In order to make reliable predictions, time series analysis typically assumes that the underlying process generating the data is stationary. This means the statistical properties of the data (like mean and variance) are stable over time.

Here is a time series plot showing the variation of 2-metre specific humidity over time, with DATETIME on the x-axis and Q2 on the y-axis, using the ggplot2 package in R.

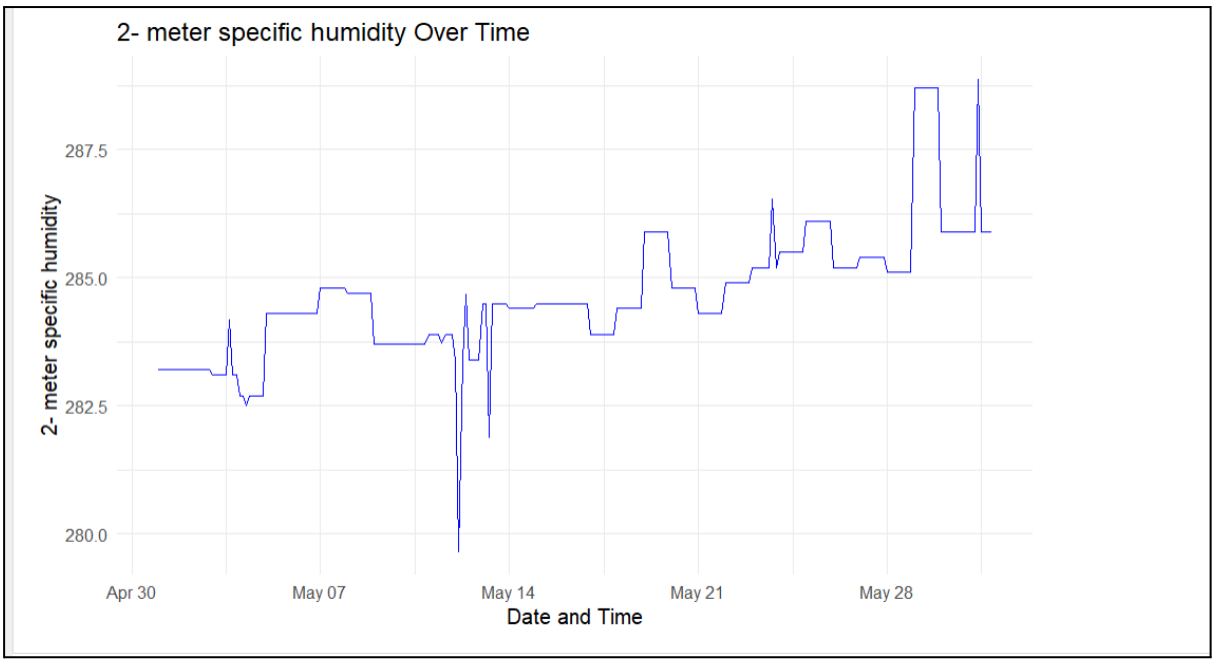

Figure 5 : Time series analysis plot for Q2 parameter

3.1 Stationarity

A series is considered stationary if its statistical characteristics, including mean, variance, covariance, and standard deviation, do not change over time or if they are not time-dependent. Stated differently, time series without trend or seasonal components are also considered stationary. The dataset's stationarity must be determined throughout the time series model preparation phase.

This is done using Statistical Tests. There are two tests available to test if the dataset is stationary:

1. Augmented Dickey-Fuller (ADF) Test
2. Kwiatkowski-Phillips-Schmidt-Shin (KPSS) Test

3.2 Augmented Dickey-Fuller Test

The ADF test, also known as the Augmented Dickey-Fuller test, is a statistical significance test that provides findings for hypothesis testing including null and alternative hypotheses. Consequently, a p-value will be obtained, from which conclusions on the stationary nature of the time series must be drawn (Verma,2021).
ADF test is conducted with the following assumptions:

1. The series has a unit root or is non-stationary, according to the null hypothesis (HO).
2. Alternative Hypothesis (HA): Either the series has no unit root or it is stationary.

Requirements for rejecting the Null Hypothesis (HO):

When the test statistic is less than the critical value and the p-value is less than 0.05, the null hypothesis (HO) is rejected because the time series is stationary and lacks a unit root. Its structure is not dependent on time.

After performing ADF Test p-value = 0.003009 is below a significant level, indicating that the time series is stationary.

4.0 ARIMA

A well-liked statistical technique for time series forecasting is ARIMA. Auto-Regressive Integrated Moving Averages are referred to as ARIMA. The models created by ARIMA operate under the suppositions listed below:

1. Since the data series is stationary, the variance and mean shouldn't change over time. One can use either log transformation or series differencing to make a series stationary.
2. Since ARIMA uses past values to predict future values, the input data must be a univariate series (Singh, 2018).

ARIMA has three sections namely AR (autoregressive term), I (differencing term) and MA (moving average term).

p is the order of autoregression.

d is the order of differencing.

q is the number of periods used in the moving average.

In practice, the parameters p, d and q are selected based on the analysis of the autocorrelation and partial autocorrelation functions of the times series data. In summary, the ARIMA model is a flexible tool for modelling and forecasting time series data with useful application in numerous fields.

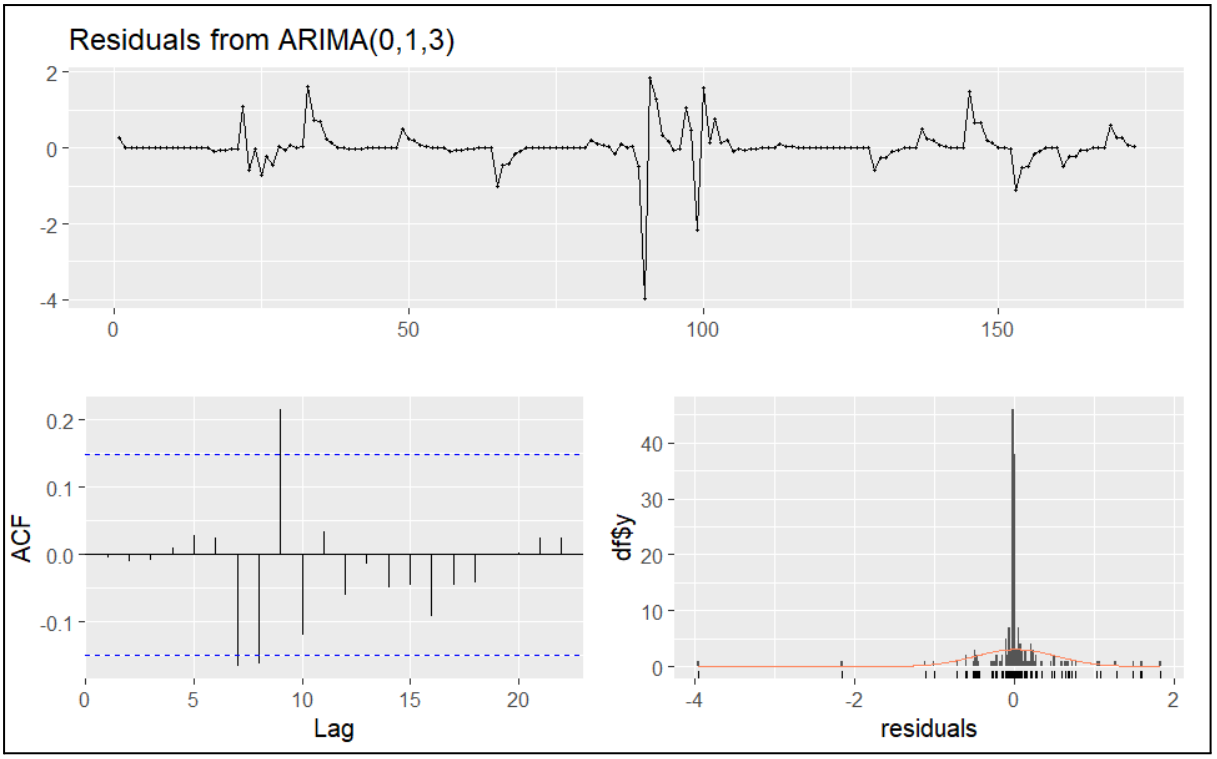

Figure 6: ARIMA model

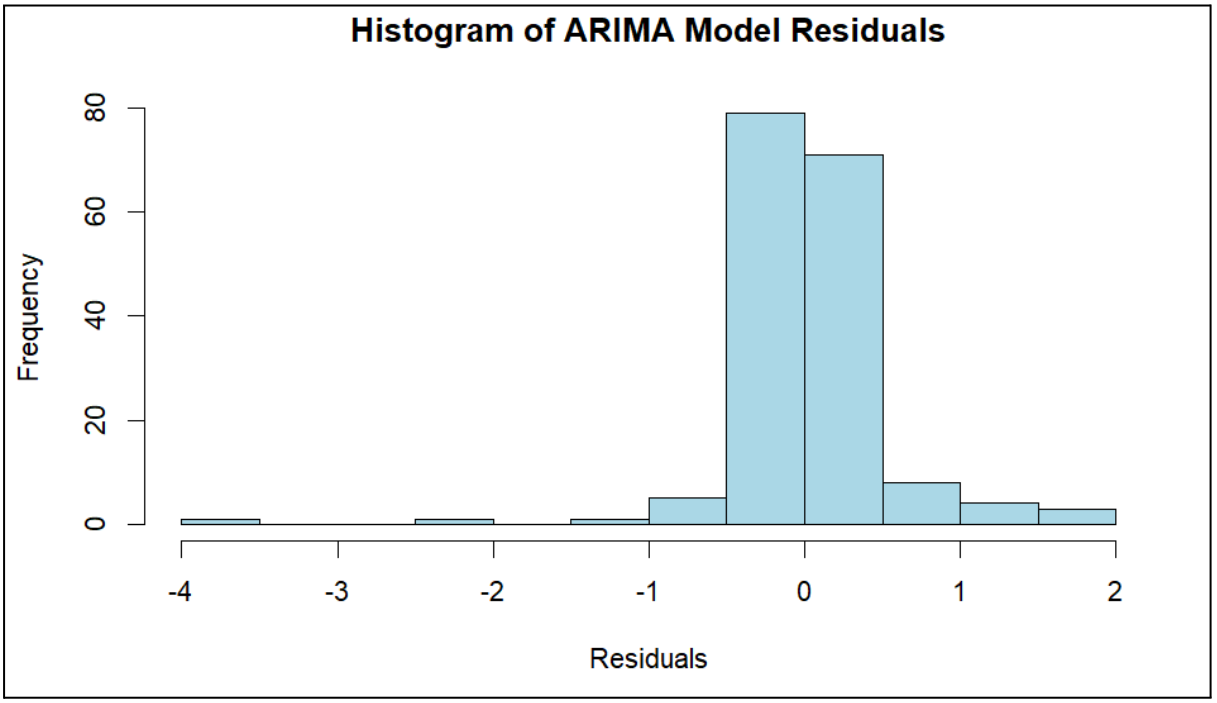

Figure 7: Histogram of ARIMA Residuals model

5.0 Machine learning models for TSA

Machine learning models can extend capabilities and add flexibility to time series analysis if applied. Below are the most used machine learning models in time series analysis

5.1 Linear Regression Model

Regression analysis is a statistical technique that aims to establish and model the association of one or more predictor variables with one response variable. While originally designed for regression problems, linear regression can also be used on time series data by considering the time variable as one of the input features. It can capture linear trends and relationships between time and the target variable. However, it may not be ideal for detecting non-linear trends or seasonality in time series data. It includes applying the linear regression on the training data, predicting the test data, and then calculating the RMSE of the model. This process helps one to check the capability of the model in predicting new unseen data, which is very crucial in determining if the model is fit for real-life use.

5.2 Support Vector Regression

Support Vector Regression is a type of supervised learning technique that is used for handling regression problems. This is an extension of the Support Vector Machine (SVM) technique, which is useful in most classification problems. The main aim of SVR is to minimise the prediction risk to make the system able to identify the most appropriate hyperplane /function for predicting supplied data accurately. To do this the entire algorithm creates a margin around the regression line and looks for error

points that lie outside this margin. In SVR kernels like Linear, Polynomial and Radial Basis functions are mostly preferred.

5.3 Decision Tree Regression

Time series forecasting can also be made using decision trees whereby the historical data is split based on the defined features to form a tree like structure and make the prediction. Decision Tree is one of the most popular and commonly used supervised machine learning algorithms which can handle both regression and classification problems.

5.4 Random Forest

Random forest is one of the most used algorithms for regression problems due to its simplicity and good accuracy. It mostly does very well on several problems, particularly on features where the relationship is non-linear. The above algorithm functions by building many decision trees in the training process. There is bootstrapping of each tree on a random subset of the training data with replacement. Furthermore, at each of the splits, a random selection of features is used for the split. This randomness assists in breaking any strong link between individual trees thus creating a more diverse and strong model.

5.5 Model Evaluation and Comparison

The eight scatter plots comparing the actual vs. the predicted values are shown in the figure below. The predicted Q2 2 metre specific humidity for the Linear Regression model, the SVR model (Poly, linear and radial kernels), and the Random Forest model (for n100,n200,n500). The grid. All these eight plots are displayed side by side using the arrange() function.

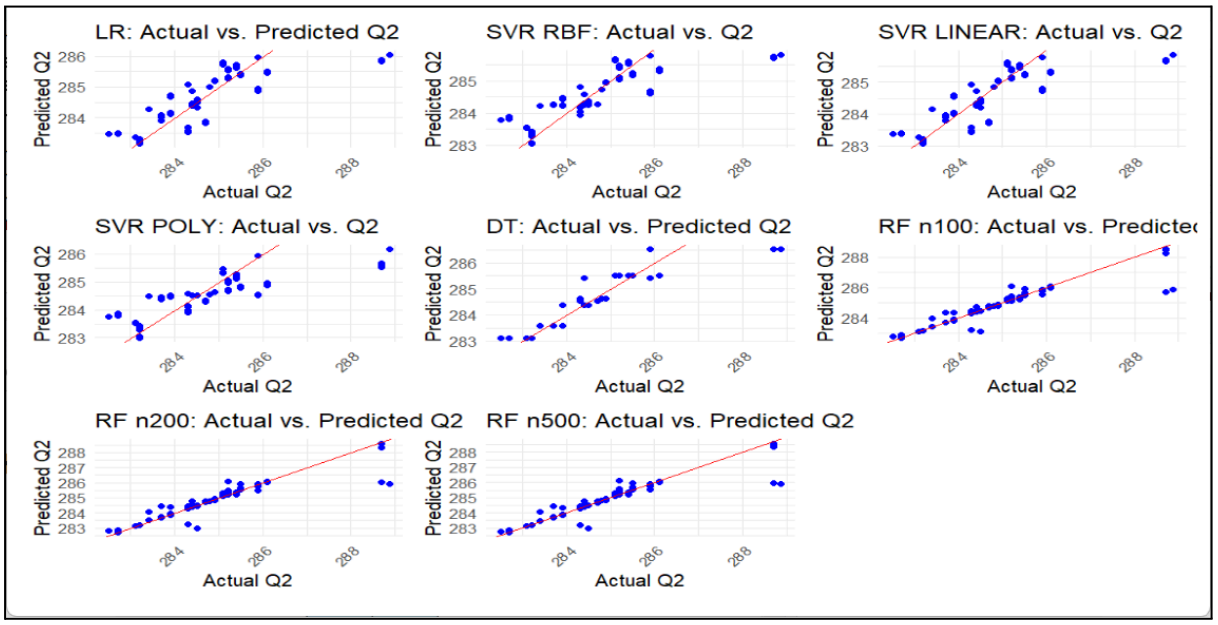

Figure 8 : Scatterplots for the actual vs. predicted values

6.0 Reporting Results

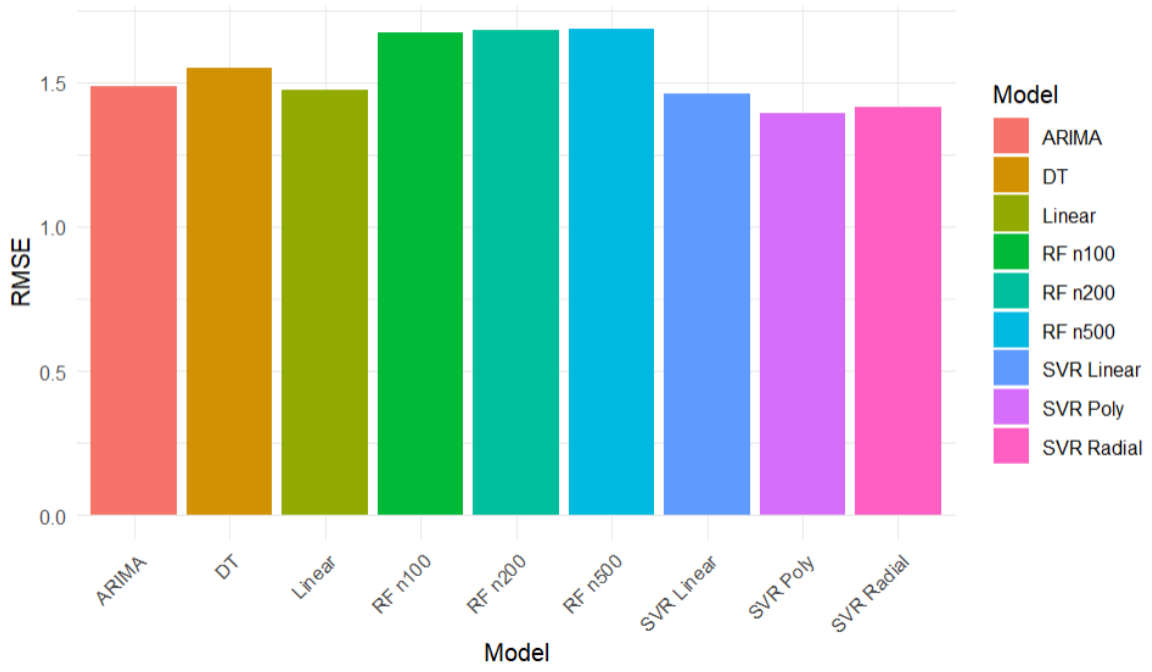

Figure 9: Barplot for RMSE values

Comparing the performance metrics of models to choose the best one. The model with the lowest The Root Mean Squared Error (RMSE) values is considered the best model, and it is the Support vector regression polynomial kernel.

	DATETIME	Q2	TIME
1	2018-05-01 00:00:00	283.2000	0
2	2018-05-01 03:00:00	283.2000	3
3	2018-05-01 06:00:00	283.2000	6
4	2018-05-01 09:00:00	283.2000	9
5	2018-05-01 12:00:00	283.2000	12
6	2018-05-01 15:00:00	283.2000	15
7	2018-05-01 18:00:00	283.2000	18
8	2018-05-01 21:00:00	283.2000	21
9	2018-05-02 00:00:00	283.2000	24
10	2018-05-02 03:00:00	283.2000	27
11	2018-05-02 06:00:00	283.2000	30
12	2018-05-02 09:00:00	283.2000	33
13	2018-05-02 12:00:00	283.2000	36
14	2018-05-02 15:00:00	283.2000	39
15	2018-05-02 18:00:00	283.2000	42
16	2018-05-02 21:00:00	283.2000	45
17	2018-05-03 00:00:00	283.1000	48
18	2018-05-03 03:00:00	283.1000	51
19	2018-05-03 06:00:00	283.1000	54
20	2018-05-03 09:00:00	283.1000	57
21	2018-05-03 12:00:00	283.1000	60
22	2018-05-03 15:00:00	284.1762	63
23	2018-05-03 18:00:00	283.1000	66

Showing 1 to 23 of 248 entries, 3 total columns

Figure 10 : Forecasting Q2- 2 metre specific humidity

7.0 Conclusion

In conclusion, the integration of time series analysis techniques, specifically ARIMA models, into the realm of weather forecasting represents a formidable tool in our arsenal for addressing complex weather-related challenges. By critically reviewing the principles, theories, algorithms, and techniques of data mining and machine learning and creatively applying them to real-world datasets, we can systematically tackle the intricacies of weather forecasting and empower decision-makers with actionable insights.

References

1. Pandian, S. (2021). Time Series Analysis and Forecasting | Data-Driven Insights (Updated 2024). *Analytics Vidhya*. [online]. Available from: https://www.analyticsvidhya.com/blog/2021/10/a-comprehensive-guide-to-time-series-analysis/ [Accessed May 10, 2024].
2. Singh, A. (2018). Build High Performance Time Series Models using Auto ARIMA in Python and R. *Analytics Vidhya*. [online]. Available from: https://www.analyticsvidhya.com/blog/2018/08/auto-arima-time-series-modeling-python-r/ [Accessed May 12, 2024].
3. Verma, Y. (2021). Augmented Dickey-Fuller (ADF) Test In Time-Series Analysis. *Analytics India Magazine*. [online]. Available from: https://analyticsindiamag.com/complete-guide-to-dickey-fuller-test-in-time-series-analysis/ [Accessed May 12, 2024].

Words count : 1799