A blue and white sign with white text

Description automatically generated

**M.Sc. Data Analytics & Technologies**

**2023 – 2024**

**AN IN-DEPTH EXPLORATION OF WEATHER PATTERNS IN SCARCLIFFE, UK: A COMPREHENSIVE ANALYSIS AND VISUALISATION**

**Course: DAT7006 – Data Science**

**Student ID: 2315053**

**Lecturer-in-Charge: Dr. Anchal Garg**

**Submission Date: 15 May 2024 - 11:59 PM**

**Workshop Session:** **Monday, 10:00 a.m**

**TABLE OF CONTENTS**

[**ABSTRACT** 6](#_Toc168323943)

[**1.0. INTRODUCTION** 7](#_Toc168323944)

[**1.1.** **ABOUT THE LOCATION** 7](#_Toc168323945)

[**1.2.** **SIGNIFICANCE OF SCARCLIFFE LOCATION** 8](#_Toc168323946)

[**1.3.** **PROBLEM STATEMENT** 8](#_Toc168323947)

[**1.4.** **SIGNIFICANCE OF DATA ANALYSIS** 9](#_Toc168323948)

[**1.4.1.** **Rationale for Data Analysis** 9](#_Toc168323949)

[**1.4.2.** **Who are the Stakeholder?** 10](#_Toc168323950)

[**1.4.3.** **Impact on Specific Organisation Decision Making** 10](#_Toc168323951)

[**2.0. LITERATURE REVIEW** 10](#_Toc168323952)

[**3.0. METHODOLOGY** 11](#_Toc168323953)

[**3.1.** **BUSINESS UNDERSTANDING:** 11](#_Toc168323954)

[**3.1.1.** **Objectives** 11](#_Toc168323955)

[**3.1.2.** **Research Questions** 12](#_Toc168323956)

[**3.2.** **DATA UNDERSTANDING** 12](#_Toc168323957)

[**3.2.1.** **About Dataset** 12](#_Toc168323958)

[**3.3.** **DATA PREPARATION:** 13](#_Toc168323959)

[**Brief:** 13](#_Toc168323960)

[**3.3.1.** **Importing and Restructuring the Dataset** 14](#_Toc168323961)

[**3.3.2.** **Explore the Data** 15](#_Toc168323962)

[**3.3.3.** **Handling Missing Values** 16](#_Toc168323963)

[**3.3.4.** **Data Cleaning** 19](#_Toc168323964)

[**3.3.4.1.** **Detecting and Handling of Outliers** 20](#_Toc168323965)

[**4.0. STATISTICAL ANALYSIS** 24](#_Toc168323966)

[**4.1. UNIVARIATE ANALYSIS** 24](#_Toc168323967)

[**4.2. BIVARIATE ANALYSIS** 26](#_Toc168323968)

[**4.3. MULTIVARIATE ANALYSIS** 27](#_Toc168323969)

[**5.0. MACHINE LEARNING MODEL** 29](#_Toc168323970)

[**5.1. DATA MODELLING OF MLR, SVM AND RFR** 29](#_Toc168323971)

[Figure 12: 2/2 Diagnostic Plots 31](#_Toc168323972)

[Figure 13: Check for Linearity 31](#_Toc168323973)

[Figure 14: Check of Heteroscedasticity 32](#_Toc168323974)

[Figure 15: Shapiro-Wilk Test 32](#_Toc168323975)

[Figure 16: No Multicollinearity 32](#_Toc168323976)

[**5.2. TIME SERIES FORECASTING** 33](#_Toc168323977)

[**5.3. DATA EVALUATION** 35](#_Toc168323978)

[**CONCLUSION** 36](#_Toc168323979)

[**REFERENCES** 37](#_Toc168323980)

**LIST OF FIGURES**

[Figure 1. Snippet of the Raw Weather Research & Forecasting Dataset 13](#_Toc168323981)

[Figure 2. Importing the raw dataset 14](#_Toc168323982)

[Figure 3. Restructuring of the dataset 14](#_Toc168323983)

[Figure 4. Creating new variable “Datetime” using the time as.POSIXlt 15](#_Toc168323984)

[Figure 5. head() function. 15](#_Toc168323985)

[Figure 6. tail() function. 16](#_Toc168323986)

[Figure 7. Display of Missing Value per Column 16](#_Toc168323987)

[Figure 8. Plot for our missing values on each column. 17](#_Toc168323988)

[Figure 9. Bar chart of missing values 17](#_Toc168323989)

[Figure 10. Visualisation of missing values using vis\_miss() function 17](#_Toc168323990)

[Figure 11 Linear Interpolation of all NA values. 18](#_Toc168323991)

[Figure 12. Handled missing values after deploying Linear Interpolation function. 18](#_Toc168323992)

[Figure 13. Visualisation of columns after handling the missing values 18](#_Toc168323993)

[Figure 14. Creating “Windspeed” variable using the X and Y components of windspeed 19](#_Toc168323994)

[Figure 15. Dropping of Columns 19](#_Toc168323995)

[Figure 16. Conversion of Some Parameters 19](#_Toc168323996)

[Figure 17. head() view of my cleaned dataset. 20](#_Toc168323997)

[Figure 18.tail() view of my cleaned dataset. 20](#_Toc168323998)

[Figure 19. Detecting outliers in TSK column using IQR 21](#_Toc168323999)

[Figure 20. Boxplot Visualisation of Surface Temperature (TSK) Column 21](#_Toc168324000)

[Figure 21. Handling of outliers in PSFC column using Tukey’s Fences method 22](#_Toc168324001)

[Figure 22. Boxplot Visualisation of Surface Pressure (PSFC) Column 22](#_Toc168324002)

[Figure 23. Boxplot Visualisation of Humidity (Q2) Column 22](#_Toc168324003)

[Figure 24. Histogram of RAINC values 23](#_Toc168324004)

[Figure 25. Boxplot Visualisation of Soil Temperature (TSLB) Column 23](#_Toc168324005)

[Figure 26. Boxplot Visualisation of Soil Moisture (SMOIS) Column 23](#_Toc168324006)

[Figure 27. Boxplot Visualisation of Windspeed Column 24](#_Toc168324007)

[Figure 28. Summary Statistics, Q-Q plot and Histogram plot of RAINC 24](#_Toc168324008)

[Figure 29. Histogram of RAINC 25](#_Toc168324009)

[Figure 30. Q-Q Plot of RAINC 25](#_Toc168324010)

[Figure 31. Shapiro-Wilk Test of RAINC 26](#_Toc168324011)

[Figure 32. Scatter Plot of Surface Temperature vs. Soil Temperature 26](#_Toc168324012)

[Figure 33. Spearman's rank correlation rho 27](#_Toc168324013)

[Figure 34. Correlation Matrix 28](#_Toc168324014)

[Figure 35. Multiple Linear Regression Model 30](#_Toc168324015)

[Figure 36. MLR Model Accuracy Test 30](#_Toc168324016)

**ABBREVIATIONS**

ST - Statistical Analysis

AI - Artificial Intelligence

MLR - Multiple Linear Regression

ML - Machine Learning

TSF - Time Series Forecasting

RNN - Recurrent Neural Network

DA - Data Analysis

WF - Weather Forecast

SVM - Support Vector Machine

RNN - Recurrent Neural Network.

MLR - Multiple Linear Regression

RFR - Random Forest Regression

TSAF - Time Series Analysis and Forecasting

MAE - Mean Absolute Error

RMSE - Root Mean Squared Error

CRISP-DM - Cross-Industry Standard Process for Data Mining

MAE - Mean Absolute Error

ARIMA - Autoregressive Integrated Moving Average

XLAT - Latitude

XLONG - Longitude

TSK - Surface Temperature

PSFC - Surface Pressure

U10 - X Component of wind at 10mm

V10 - Y Component of wind at 10mm

Q2 - Humidity

RAINC - Convective Rain (Accumulated Precipitation)

RAINNC - Non-convective Rain

SNOW - Snow Water Equivalent

TSLB - Soil Temperature

SMOIS - Soil Moisture

# **ABSTRACT**

This study explores the use of three different machine learning models, Multiple Linear Regression (MLR), Support Vector Regression (SVR), and Random Forest Regression (RFR), for weather forecasting in the Scarcliffe region. This project will document the preprocessing procedures used to clean the unprocessed weather data before the models are trained and tested. Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) metrics are used to assess the performance and accuracy of the model.

The project, which is written in the R programming language, uses several libraries for data manipulation, visualisation, and analysis, including zoo, DescTools, ggplot2, tidyverse, dplyr, and naniar. The goal of the study is to produce insightful information about weather forecasting that will be useful to a variety of stakeholders, such as residents, farmers, tourism operators, flight navigation services, and policymakers.

This research compares the performance of the MLR, SVR, and RFR models to determine which method is best for efficient weather prediction in the Scarcliffe region. It is anticipated that the results would improve weather forecasts' accuracy and usefulness, which will help with resource management and well-informed decision-making in a variety of industries.

# **1.0. INTRODUCTION**

As climatic variability becomes more visible, understanding its implications on agricultural production and tourism is critical to ensuring economic stability and growth. This research aims to examine local weather patterns and their effects on agricultural and tourism sector in Scarcliffe, UK. The aim of this research is to offer a thorough analysis, relevant correlations, trends, and patterns by assessing the direct relationships between certain weather parameters and economic outcomes, agricultural production and tourism flows in Scarcliffe. It will equip stakeholders, particularly local farmers, tourism providers, and residents, with adequate data-driven insights that will help boost their operations and strategic objectives. Accurate weather forecasting plays a significant role in practically every sector and society, as it allows businesses to make informed decisions, emergency personnel to prepare for natural tragedies and people to organise their activities (Weather Forecast Models: The Ultimate Guide (2024), no date). Pliske et al. (2004) discussed that technological advancement have greatly influenced the technologies accessible to weather forecasters in recent years. Today's forecasters are frequently overwhelmed by the enormous variety of technological products, the development of these tools has made it simple for us to conduct research on weather in various locations for the benefit of many stakeholders, including businesses, farmers, and people. However, in this study, I will be primarily focusing on Scarcliffe data to understand how different climate variables relates with one another.

## **ABOUT THE LOCATION**

Scarcliffe offers an interesting case study for investigation as it has a population of 5,288 in the UK Census of 2011 (‘Scarcliffe’, 2024). It is situated 127 miles north-west of London, two miles south-east of Bolsover, and six miles north-west of Mansfield. Two miles to the northwest of Nottinghamshire's boundary is Scarcliffe. Scarcliffe could be found in latitude 53.227°N and longitude -1.224°E. The town is within the jurisdiction of the Bolsover District Council, which is part of the Derbyshire County Council (*Where is Scarcliffe? Scarcliffe on a map*, no date). Gaining a comprehensive understanding of the weather patterns in Scarcliffe is of utmost importance for residents of the town, local farmers, tourist providers, meteorologists, and policymakers. This project aims to utilise a comprehensive dataset containing weather parameters for Scarcliffe to address various research questions and provide insights into local weather patterns.

## **SIGNIFICANCE OF SCARCLIFFE LOCATION**

Despite not being well-known worldwide, Scarcliffe, an English village in Derbyshire, is significant locally and historically for several reasons. The village was referenced in the Domesday Book, a handwritten document finalised in 1086, that offers a glimpse into mediaeval England. This historical documentation highlights Scarcliffe’s existence and importance during the Norman conquest.

Scarcliffe, being a small village, epitomises the rural and community-oriented way of life that is characteristic of several English settlements. The customs, social dynamics, and community events in Scarcliffe contribute to the cultural tapestry of Derbyshire. The village’s past includes connections to agriculture and local industries that have shaped the economic history of the region. Understanding these local industries can provide a deeper appreciation of the area’s development over time. Scarcliffe’s location in Derbyshire places it near other historical and cultural sites, such as Bolsover Castle and the Peak District National Park, which enhance its significance by association. Its geographical position in the East Midlands makes it vulnerable to a wide range of weather patterns, which are impacted by its closeness to the Pennines and the North Sea. Its weather impacts various sectors, including agriculture, tourism, and local infrastructure.

## **PROBLEM STATEMENT**

Scarcliffe is a significant location, but its weather patterns have not been thoroughly studied, which makes it difficult to forecast and prepare for bad weather. To close this gap and give stakeholders useful information, a thorough examination of past weather data is required. A closer analysis of Scarcliffe's weather data is needed to improve prediction accuracy. This would lead to better agricultural outcomes as well as enhanced ecological management and preparedness for emergencies. The potential effects of global climate change make it essential to update and enhance weather prediction models often to maintain their relevance and usefulness.

To address these issues, more precise and locally applicable weather forecasting models will be developed. Sophisticated statistical and machine learning techniques will be employed, together with comprehensive data analysis, to provide a deeper knowledge that will enable Scarcliffe residents, local farmers, and businesses to make informed decisions.

## **SIGNIFICANCE OF DATA ANALYSIS**

### **Rationale for Data Analysis**

For variety of reasons, it is essential to analyse data on a wide range of weather factors, including soil moisture, wind speed, humidity, surface pressure, surface temperature. Every parameter is essential to comprehending and forecasting weather patterns, which has significant implications for many industries like urban planning, agriculture, tourist sector and disaster management. This is the complete rationale behind the data analysis (DA) of these weather forecasting parameters:

Understanding temporal patterns and fluctuations in these weather parameters enables us to forecast daily and seasonal weather trends. This approach is critical for understanding climate change and spotting anomalies that may indicate major climatic shifts. Futhermore, investingating the links among various weather parameters allows us to identify complex pehenomena that cause weather occurences. For example, the interaction of humidity and temperature may have a substantial impact on precipitation patterns, whereas the combined effect of temperature and soil moisture might affect agricultural prodduction.

Assessing the multivariate impacts of different parameters offers a comprehensive understanding of atmospheric dynamics, allowing us to create more accurate forecasting models. These models can estimate future situations with better accuracy. This integrated data analysis (DA) can produce improved weather forecasts that provide useful insights for a range of industries in Scarcliffe. Accurate forecasting in agriculture may boost crop yields by enhancing irrigation planning, planting and harvesting schedules, and irrigation efficiency. Extreme weather conditions can be lessened in disaster management by using early warning systems that are based on reliable weather data.

The goal of this project is to improve weather forecast accuracy by analysing and integrating variety of factors. By doing this, we can foster better decision making, boost the effectiveness of resources, and strengthen industry vulnerability to weather related problems. A detailed analysis of these climatic traits will provide significant new understandings that will improve living standards and inspire innovation in Scarcliffe, UK.

### **Who is the Stakeholder?**

* Farmers and Agricultural Sector: Accurate weather forecasts are essential for farmers and other agricultural workers to plan planting, watering, and harvesting. Predictive weather analysis reduces the likelihood of crop failure while assisting in output growth.
* Local governments and emergency management agencies: They depend on accurate weather forecasts to help them anticipate and react to extreme weather events, ensure public safety, and reduce financial losses.
* Tourism Industry: The weather has a significant effect on tourism. Precise weather data is helpful to Scarcliffe's tourism sector regarding marketing and management of operations.
* Environmental Agencies: The study can support further attempts to track and comprehend how local weather patterns are affected by climate change.

### **Impact on Specific Organisation Decision Making**

* Transportation: Weather trends and forecasts have an impact on transportation time management, route planning, and safety protocols. This helps to minimise weather-related hazards and ensure on-time delivery.
* Building & Infrastructure: Planning time frames for maintenance and allocation of resources for building projects, as well as designing and building resilient infrastructure, may all be facilitated by an understanding of local weather dynamics.
* Farmer’s Community: Fluctuations in temperature, precipitation, and extreme weather phenomena can all influence the duration of growing seasons, the amount of crops produced, and the existing agricultural methods.
* Tourist Operators: The tourist industry depends significantly on reliable weather for activities and attractions, is subject to uncertainties that may have an impact on visitor numbers and earnings.

# **2.0. LITERATURE REVIEW**

ML models for WF has gained a lot of significance in recent years because of their capacity to enhance prediction accuracy and reliability. Several researchers have investigated the use of different methods of ML for WF, drawing on historical weather data and meteorological variables. Previous research has demonstrated the value of ML, time-series modelling, and statistical analysis when examining weather data. many ML models including random forests and neural networks, as well as statistical models, such as SARIMA and ARIMA, have been studied for WF. However, studies on Scarcliffe's weather patterns have been limited, providing an opportunity for further investigation.

In one study, Mohandes et al. (2004) developed a reliable forecasting model by applying ensemble learning techniques including gradient boosting machines and random forests. Their research made clear how crucial it is to integrate many data sources and use feature engineering to improve wind speed prediction. Singh et al. (2019) also utilised three ML models, ANN, RNN, and SVM for weather prediction in one of their research. Lagerquist et al. (2019), on the other hand, tested multiple machine learning algorithms, including logistic regression, decision trees, RFR, neural networks and gradient-boosted forests, for forecasting straight-line wind alone.

# **3.0. METHODOLOGY**

The project will employ the CRISP-DM methodology, it encompasses these subsequent stages:

## **BUSINESS UNDERSTANDING:**

### **Objectives**

* Understand the correlations and patterns between weather parameters in the Scarcliffe dataset.
* Create statistical models that measure correlations among variables and test hypotheses.
* Create time-series forecasting models to anticipate future weather conditions in Scarcliffe.
* Deploy ML regression approaches to identify complicated, non-linear correlations between weather parameters and provide Scarcliffe with accurate weather predictions.
* Use appropriate assessment metrics and cross-validation procedures to check the performance and accuracy of the produced statistical models, SVR, and ML predictions.

### **Research Questions**

* Using Shapiro-Wilk normality test and visualisation like histogram and Q-Q plot, does the convective rainfall (RAINC) follow a normal distribution?
* Is there a substantial correlation between Surface Temperature (TSK) and Soil Temperature (TSLB) at Scarcliffe?
* How can humidity (Q2), soil temperature (TSLB), Soil Moisture (SMOIS), and windspeed influence surface temperature (TSK) in the scarcliffe region?
* What is the best machine learning model for predicting surface temperature (TSK) in Scarcliffe?
* Can machine learning regression models capture the complex relationship between meteorological variables and produce accurate predictions for temperature, pressure, humidity, and wind speed at Scarcliffe?

## **DATA UNDERSTANDING**

### **About Dataset**

The data chosen for this assessment talks about weather research and Forecasting in May 2018 of various regions from which I extracted my Scarcliffe, UK dataset for my analysis. The dataset raw dataset was downloaded from my Data Science course Moodle:

URL: (<https://moodle.bolton.ac.uk/mod/resource/view.php?id=2241359>)

There are 5,451 rows and 2,482 columns in the raw dataset, notable columns are: XLAT, XLONG, TSK, PSFC, U10, V10, Q2, RAINC, RAINNC, SNOW, TSLB, SMOIS. The remaining columns are repetition of the above listed columns. First row contains the spread out of the time and date. I will be showing step by step of how I reconstruct the whole dataset, and how I got my specific location: Scarcliffe. I made use of google map to check the XLAT and XLONG of various location before getting my Scarcliffe location, XLAT = 53,227 and XLONG = -1,224.

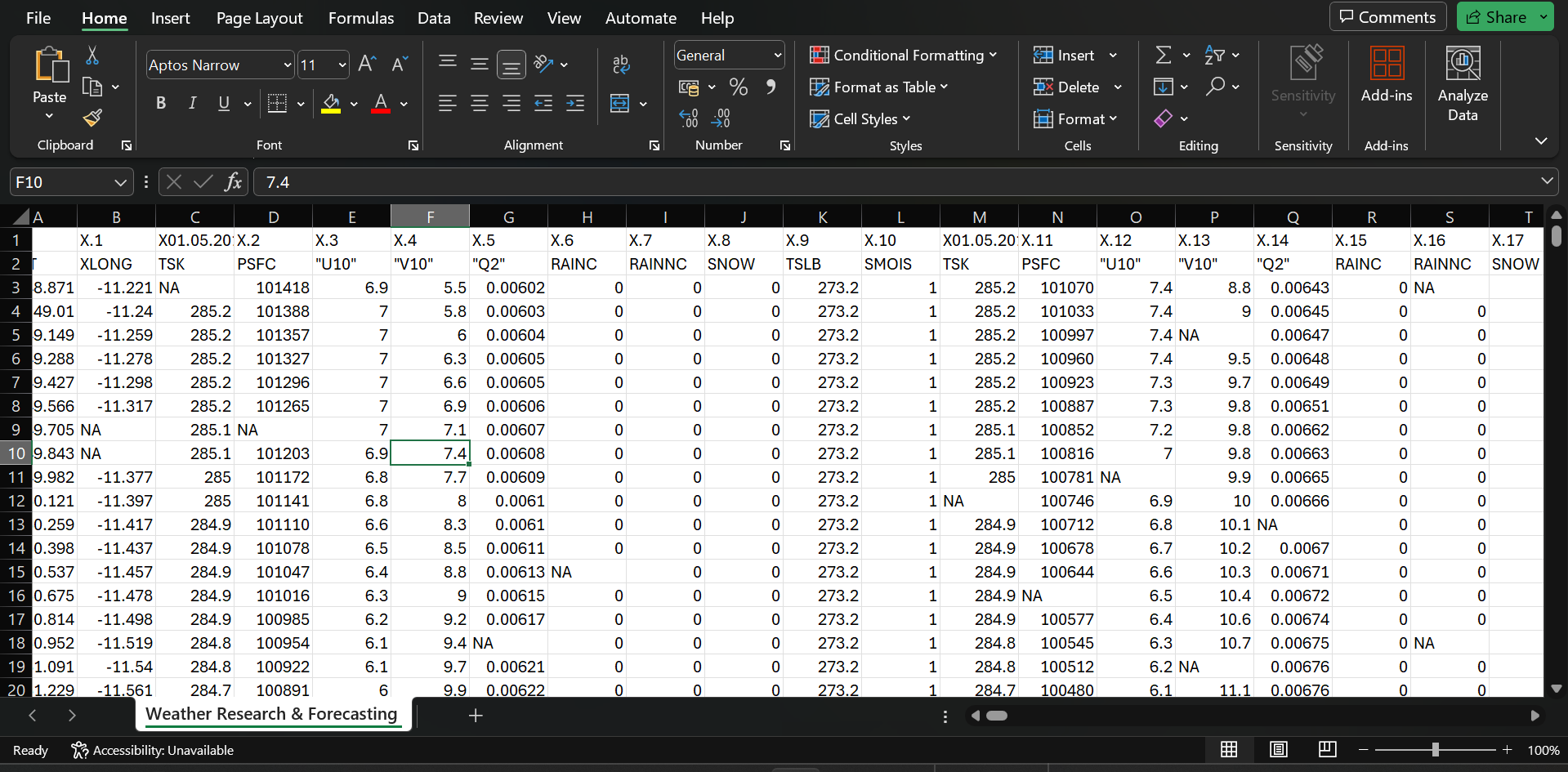


Figure 1. Snippet of the Raw Weather Research & Forecasting Dataset

Source: Course Moodle (<https://moodle.bolton.ac.uk/mod/resource/view.php?id=2241359>)

## **DATA PREPARATION:**

Step by step documentation and visualisation of how we transformed the raw data into meaningful data suitable for our analysis will be done here. During this phase, the data is cleaned, transformed, and pre-processed as needed before being analysed. Handling missing values, encoding categorical variables, scaling features, and separating the data into training and testing sets are some of the possible tasks. I carried out all my data preprocessing and analysis on RStudio, an integrated Development Environment (IDE), It offers an intuitive interface that simplifies the writing, running, and debugging of R code.

## **Brief:**

Looking at the “Weather Research & Forecasting” dataset, I understand the first thing I need to do is to restructure the data as it was not well arranged at the point of download. Firstly, I used google map to get my preffered location using the Longitude and Latitude, my location is Scarcliffe, Derbyshire. The dataset contains total of 2,482 columns as explained earlier, my understanding of the dataset is that, we have just 12 major columns – XLAT, XLONG, TSK, PSFC, u10, v10, Q2, RAINC, RAINNC, SNOW, TSLB, and SMOIS, other columns are repition of this 10 columns excluding the XLAT and XLONG columns, which means we have to restructure it in a way that, we will break the whole 2,482 columns into list of groups with each group having 10 columns each, after which, we will bind them altogether. In the raw dataset, the first row is not our header, those are our time and date column which we have to restructure as a column, step by step on how I achieved this will be shown below.

### **Importing and Restructuring the Dataset**

Importing the downloaded raw data into R is the initial step in every data analysis job. We can use many functions to read the data into R depending on the format of the data. In this instance, the read.csv() function can be utilised as the downloaded file is in CSV format, we can use View() function, this will pop out in another table for us to view the values. To use R seamlessly, there are libraries and packages we must download some of which are tidyverse, dplyr, ggplot2, zoo, naniar, corrplot and so on.

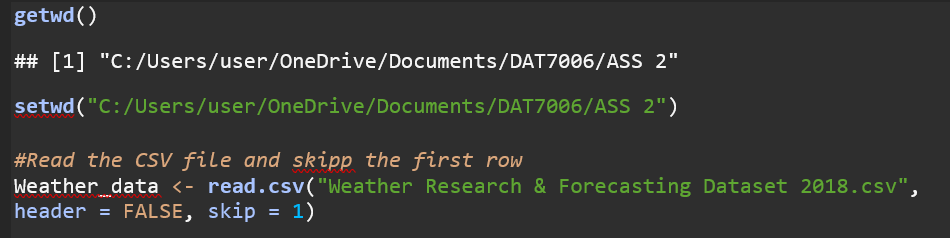


Figure 2. Importing the raw dataset

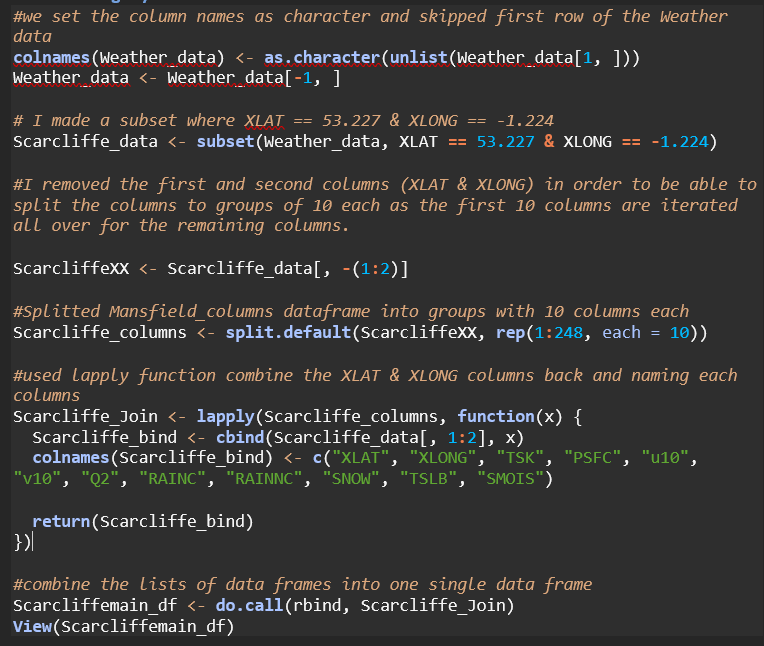


Figure 3. Restructuring of the dataset

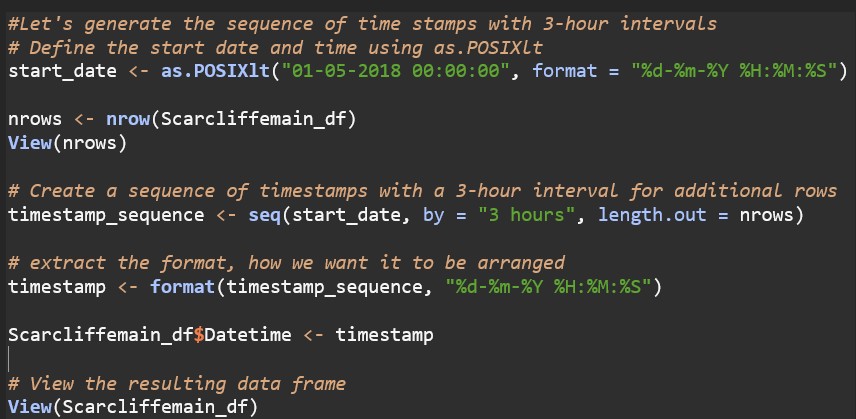


Figure 4. Creating new variable “Datetime” using the time as.POSIXlt

### **Explore the Data**

It is very important to explore and comprehend the data's variables, structure, and possible problems after it has been imported. To obtain a summary of the data, we can utilise functions like head() - this will show the first 5 rows of the dataset, tail() - this will show last 5 rows of the datset, str() - this function will show the data type of all variable, this is very important as it gives us insight of the variable we are working on and to know if to change accordingly.

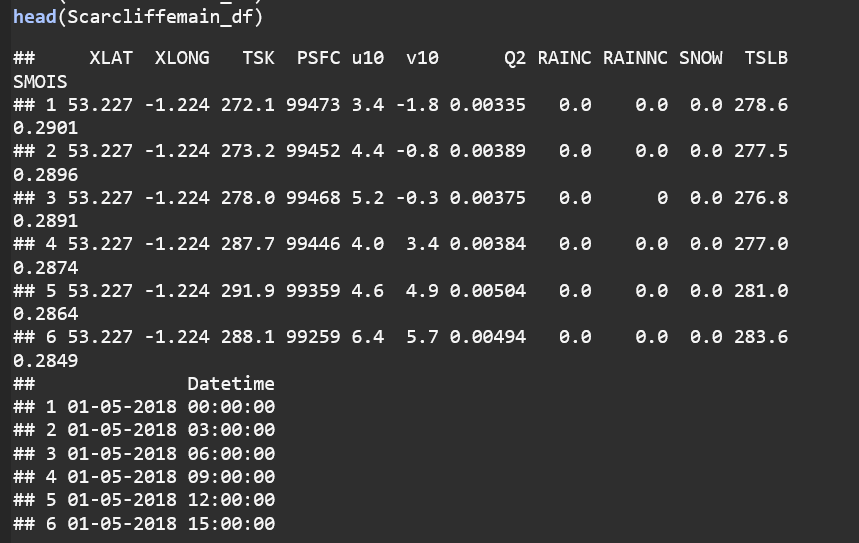


Figure 5. head() function.

A screenshot of a computer

Description automatically generated

Figure 6. tail() function.

### **Handling Missing Values**

As stated by (Horton and Lipsitz, 2001) and corroborated by (Yadav and Roychoudhury, 2018), Missing data typically affects data analysis in scientific research. In recent decades, there has been a lot of study on statistical strategies for dealing with missing data. Missing values are common in datasets, and they must be handled carefully. Missing values can be identified using functions such as is.na() and sum(is.na()). Depending on the analysis needs and the nature of the missing data, we can either eliminate the rows or columns with missing values or impute them using appropriate procedures. In this project, I made use of approximation (na.approx) to handle my missing values, using mean or dropping the NA cells is not ideal for my analysis. After thorough restructuring of my dataset, I realised we I have total 67 missing values.

A screen shot of a computer code

Description automatically generated

Figure 7. Display of Missing Value per Column

A black screen with colorful text

Description automatically generated

Figure 8. Plot for our missing values on each column.

A graph of missing values

Description automatically generated

Figure 9. Bar chart of missing values

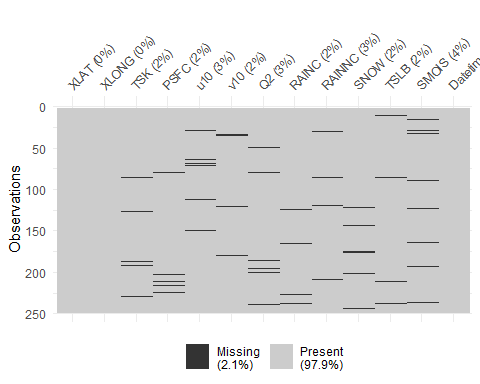


Figure 10. Visualisation of missing values using vis\_miss() function

The above figure is a bar chart visualising the number of missing values for each variable. As explained by Huang (2021), One way to interpolate one-dimensional data is using linear generate the data value. After using Linear interpolation (na.approx), interpolation, which uses two neighbouring points in the one-dimensional data sequence to all missing values were handled appropriately.

A screen shot of a computer

Description automatically generated

Figure 11 Linear Interpolation of all NA values.

The figure below shows the missing value after deploying linear interpolation to handle the NA values.

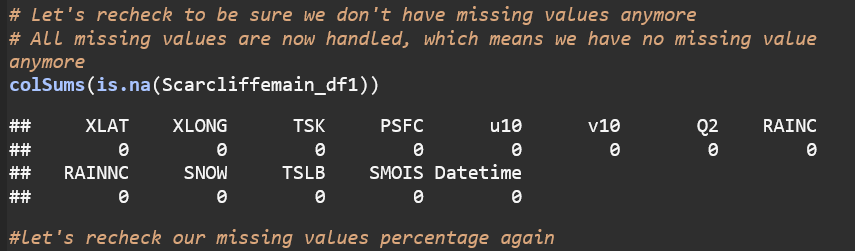


Figure 12. Handled missing values after deploying Linear Interpolation function.

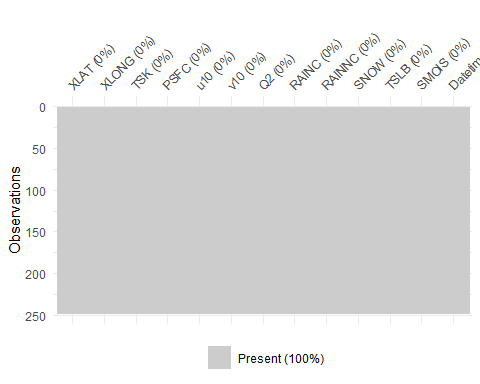


Figure 13. Visualisation of columns after handling the missing values

### **Data Cleaning**

Data cleansing is a crucial process that guarantees the integrity and uniformity of the data. This may involve removing duplicates, handling outliers, correcting data entry errors, and standardizing formats. Detecting and handling unclean data is one of the recurring issues in data analytics, and failing to do so might result in incorrect insights and erroneous choices (Chu et al., 2016).

I will handle my outliers one by one and show visualisations for each column. Before checking for my outliers, I need to calculate my windspeed, I use X and Y components of wind to get my windspeed which will be added to my dataset as a Windspeed variable. Additionally, I did some unit conversion, I changed Surface Temperature (TSK) from Kelvin to degree Celsius (°C), I also changed Surface Pressure (PSFC) from Pascals to hectopascals (hPa). Humidity (Q2) unit was also changed from kg/kg to g/kg, and lastly, Soil Temperature was changed from Kelvin to degree Celcius (°C).

A screenshot of a computer

Description automatically generated

Figure 14. Creating “Windspeed” variable using the X and Y components of windspeed

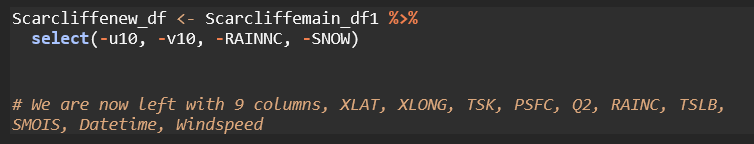


Figure 15. Dropping of Columns

A screenshot of a computer

Description automatically generated

Figure 16. Conversion of Some Parameters

A screenshot of a computer

Description automatically generated

Figure 17. head() view of my cleaned dataset.

A screenshot of a computer

Description automatically generated

Figure 18.tail() view of my cleaned dataset.

### **Detecting and Handling of Outliers**

In Surface Temperature (TSK) column, I made use of Interquartile Range (IQR). The interquartile range is a statistical measure that quantifies the extent of variability or dispersion in a dataset. It is described as the variation between an ordered data set's top and lower quartiles (Clark-Carter, 2005). First thing to do is to use summary() function, this will give us the the descriptive statistics such as “minimum number, maximum number, mean, median, the first quartile, and third quartile. As defined, IQR = Q3 – Q1, where Q3 is the 3rd quartile or 75th percentile while Q1 is the first quartile or 25th percentile. I also made use of Tukey’s Fences method and boxplot to detect and handle some outliers in my project. I will be showing visualisation for each column below.

A screen shot of a computer program

Description automatically generated

Figure 19. Detecting outliers in TSK column using IQR

A diagram of a box plot

Description automatically generated

Figure 20. Boxplot Visualisation of Surface Temperature (TSK) Column

From the IQR detection method and the boxplot visualisation, TSK column has no outlier.

A screen shot of a computer

Description automatically generated

Figure 21. Handling of outliers in PSFC column using Tukey’s Fences method

A diagram of a box plot with numbers and a line

Description automatically generated

Figure 22. Boxplot Visualisation of Surface Pressure (PSFC) Column

A diagram of a box plot

Description automatically generated

Figure 23. Boxplot Visualisation of Humidity (Q2) Column

A graph of a graph

Description automatically generated

Figure 24. Histogram of RAINC values

A diagram of a graph

Description automatically generated

Figure 25. Boxplot Visualisation of Soil Temperature (TSLB) Column

A diagram of a box plot with text

Description automatically generated

Figure 26. Boxplot Visualisation of Soil Moisture (SMOIS) Column

A diagram of a box plot

Description automatically generated

Figure 27. Boxplot Visualisation of Windspeed Column

Certain outliers in my dataset reflect inherent variances within the population, I decided to retain them instead of removing them because they are true outliers with natural variations. They are genuine observations and very vital for my analysis.

# **4.0. STATISTICAL ANALYSIS**

## **4.1. UNIVARIATE ANALYSIS**

**Research Method 1:** Using Shapiro-Wilk normality test and visualisation like histogram and Q-Q plot, does the convective rainfall (RAINC) follow a normal distribution?

I will be doing Univariate analysis on Convective rain (Accumulated precipitation) – RAINC. To determine the RAINC central tendency and distribution, Summary() function in R will give us the basic descriptive statistical analysis such as the minimum value, 1st quartile, Mean, Median, 3rd quartile and the Maximum number. Histogram will also be used to visualise RAINC frequency distribution to understand its distribution shape. Additionally, I will use Shapiro-Wilk to test the normality of RAINC to determine if it follow a normal distribution.

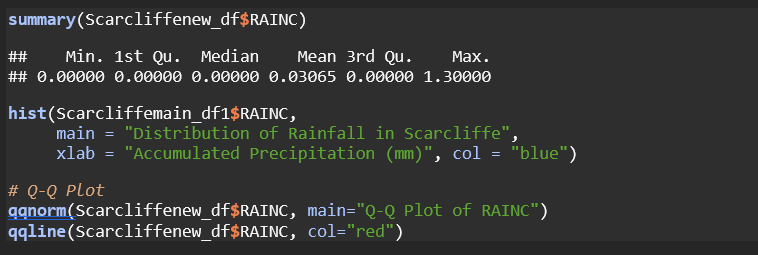


Figure 28. Summary Statistics, Q-Q plot and Histogram plot of RAINC

A graph of rainfall in a number of times

Description automatically generated with medium confidence

Figure 29. Histogram of RAINC

A graph of a number of numbers and a line

Description automatically generated with medium confidence

Figure 30. Q-Q Plot of RAINC

It can be seen from the Q-Q plot that RAINC es not follow normal distribution. The histogram shows a highly skewed distribution with a significant peak at 0.0 mm, indicating that the most frequent observation is no precipitation. There are a few smaller bars to the right, showing that there are fewer occurrences of higher precipitation values. This distribution suggests that over the period or area studied, dry conditions are much more common than wet conditions. The shape of the distribution, heavily skewed with a long tail to the right, is typical for precipitation data, where many dry days and a few very wet days are common.

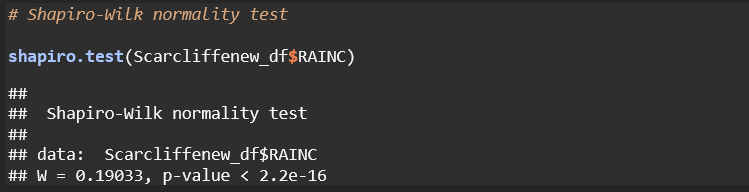


Figure 31. Shapiro-Wilk Test of RAINC

The Shapiro-Wilk test result shows that test statistic (W) is 0.19033. Values close to 1 suggest that the data is approximately normally distributed. Values far from 1 indicate deviations from normality. In this case, 0.19033 is very far from 1, suggesting a significant deviation from normality. P-value < 2.2e-16 is extremely low; it is less than 0.05 (threshold for statistical significance).

**Hypothesis:** Given that the p-value < 0.05, I reject the null hypothesis that RAINC is normally distributed. Both the test statistic (W) being far from 1 and the extremely low p-value indicate that the RAINC data does not follow a normal follow a normal distribution.

## **4.2. BIVARIATE ANALYSIS**

Research Question 2: Is there a substantial correlation between Surface Temperature (TSK) and Soil Temperature (TSLB) at Scarcliffe?

This is analysis of two variable. Let’s examine the correlation between Surface Temperature and Soil Temperature in Scarcliffe by plotting a scatter plot.

A graph showing a diagram of a surface temperature

Description automatically generated

Figure 32. Scatter Plot of Surface Temperature vs. Soil Temperature

The plot above shows a scatter plot of surface temperature on the x-axis plotted against the soil temperature on the y-axis. The scattered data points on the graph clearly demonstrate a pattern that shows a rise in soil temperature in relation to surface temperature.

The plots illustrate how the surface temperature rises from 0°C to around 30°C while the soil temperature rises, typically ranging from 5°C to 20°C. The distribution of points indicates a positive correlation between surface and soil temperatures, implying that warmer surface conditions are often associated with warmer soil temperatures.

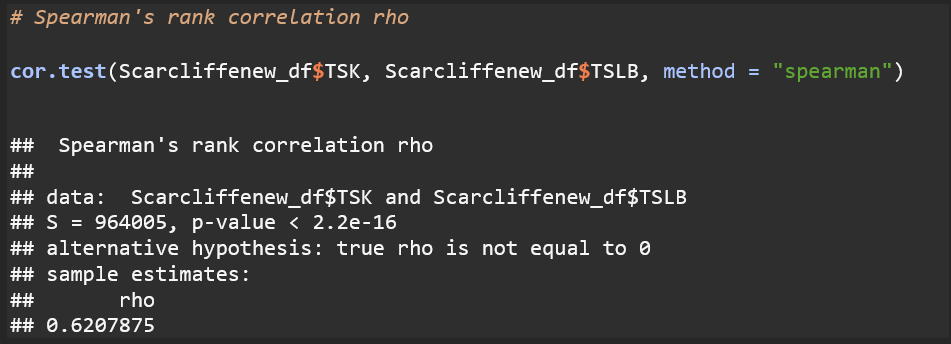


Figure 33. Spearman's rank correlation rho

The rho coefficient is 0.6207875, indicating a moderate to high positive correlation between TSK and TSLB variables. The spearman’s rank correlation coefficient runs from -1 to 1, with 1 indicating a perfect positive connection, -1 indicating a perfect negative correlation, and 0 denoting no association.

The weak p-value (<2.2e-16) indicates that the observed connection is unlikely to have occurred by chance. Therefore, we reject the null hypothesis which claims that there is no correlation between TSK and TSLB.

## **4.3. MULTIVARIATE ANALYSIS**

Research Question 3: How do humidity (Q2), soil temperature (TSLB), Soil Moisture (SMOIS) and windspeed influence surface temperature (TSK) in the scarcliffe region?

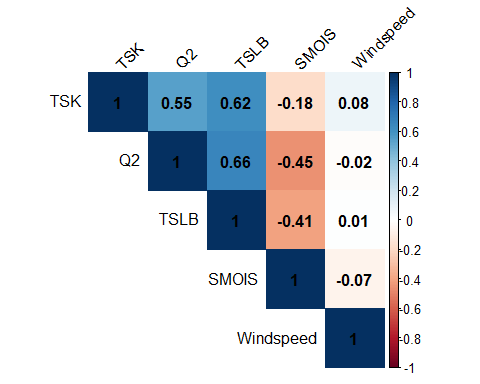


Figure 34. Correlation Matrix

The above correlation matrix shows correlation coefficient between the variables, the values is between -1 to 1.

* 1 indicates a perfect positive correlation which means as one variable increases, the other variable tends to increase.
* -1 indicates a perfect negative correlation, as one variable increases, the other variable tends to decrease.
* 0 indicates no correlation.

**TSK and Q2:** Correlation coefficient = 0.55 which indicate that there is a moderate positive relationship between surface temperature (TSK) and humidity (Q2). TSK rises as Q2 increases.

**TSK and TSLB:** Correlation coefficient = 0.62. TSK and TSLB have a moderate positive association. Higher soil temperatures can contribute to higher surface temperature.

**TSK and SMOIS:** Surface temperature and soil moisture show a weak negative correlation (-0.18). As soil moisture levels rise, surface temperature will fall.

**TSK and Windspeed:** The relationship between these 2 variables is a weak positive correlation (0.07794999).

# **5.0. MACHINE LEARNING MODEL**

## **5.1. DATA MODELLING OF MLR, SVM AND RFR**

MLR, which has its origins in the early 20th century, is a vital technique in statistical data analysis. Regression analysis was made possible by statisticians like Fisher (1922). The objective of multiple linear regression MLR analysis is to investigate the correlation between a dependent variable and two or more independent variables (Uyanık and Güler, 2013).

In this analysis, four predictor variables humidity (Q2), surface pressure (TSLB), soil moisture (SMOIS) and Windspeed are used in this Multiple Linear Regression (MLR) model to predict surface temperature (TSK). Each variable represents a factor that may influence the dependent variable, and the regression analysis allows us to quantify the strength and direction of these relationships. To find out how each independent variable affects the variability of the dependent variable, many independent variables must be used. This is done while making sure that the uniformity of the other variables is maintained.

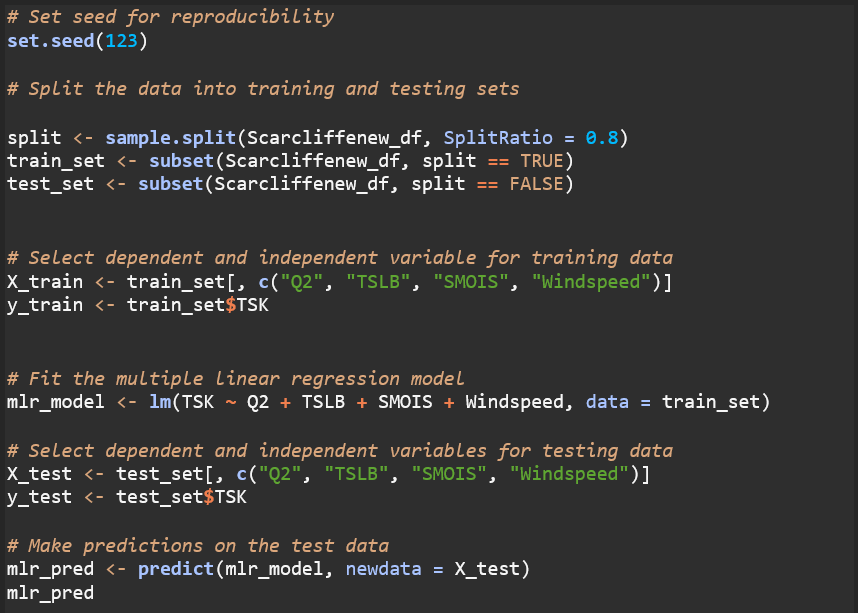


Figure 35. Multiple Linear Regression Model

A screenshot of a computer

Description automatically generated

Figure 36. Model Summary Statistics

By analysing the coefficients associated with each independent variable, we can determine the extent to which they influence the dependent variable, controlling for the effects of other variables in the model. Highly significant predictors (\*\*\* and \*\*) provide strong evidence that their coefficients are different from zero, indicating a reliable effect on surface temperature (TSK). Marginal significance (.) suggests a weaker but still potentially meaningful effect.

The model’s overall fit is summarised using metrics such as R-squared (0.4431) and adjusted R-squared (0.4315), which indicates how well the model explains the variability in TSK.

This analysis is valuable for identifying significant predictors of the dependent variable and understanding the overall pattern and relationships among the variables. Additionally, it provides insights into the relative importance of different factors in explaining the variability in the dependent variable.

A computer screen shot of a black screen

Description automatically generated

Figure 37. MLR Model Performance Metrics (RMSE and MAE)

The MLR model’s accuracy is proven by the above performance metrics. Lower values indicate better model performance.

* RSME: 5.518934
* MAE: 4.632473

A graph with blue dots and red line

Description automatically generated

Figure 38. Actual Vs Predicted Surface Temperature

**Linear Regression Assumptions**

To ensure the validity and reliability of this model, it is crucial to satisfy all linear regression assumptions, like linearity check, homoscedasticity, normality of residuals, and Multicollinearity test. These assumptions help to ensure that the estimates produced by the regression model accurately and impartially represent the true relationships found in the data.

A group of graphs showing different values

Description automatically generated

Figure 39. 2/2 Diagnostic plots

Above plot helps in evaluating the linear regression model's overall fit and underlying assumptions. It analyses the residuals (the discrepancies between actual and expected values) to look for trends or violations of assumptions. This plot is a set of diagnostic charts typically used to analyse the assumptions and performance of a linear regression model.

A white background with black text

Description automatically generatedBy examining these diagnostic plots, you can assess evaluate the validity of the linear regression assumptions and identify any potential issues that may require further investigation or adjustments to the model. As stated by Alguraibawi (2015), identifying high leverage points is essential because they can lead to erroneous predictions and invalid inferential statements due to their significant impact on the calculated values of various estimations. It is essential to classify the high leverage points into good and bad leverage points because only the bad leverage points have an undue effect on the parameter estimates.

### Figure 13: Check for Linearity

The p-value of my model Linearity check is 0.916 which is bigger than the generally accepted significance level of 0.05. This means that we fail to reject the null hypothesis of linearity. In other words, based on the Rainbow Test, there is no evidence to suggest that the linearity A white rectangle with black text

Description automatically generatedassumption is violated in the multiple linear regression model.

### Figure 14: Check of HeteroscedasticityA white background with black text Description automatically generated

### Figure 15: Shapiro-Wilk Test

The Shapiro-Wilk test carried on my model indicates that the residuals of the linear regression model lack characteristics of a normal distribution as shown above. The assumption of normality of residuals is one of the key assumptions of linear regression models. Violation of this assumption can have implications for the validity of statistical inferences and the reliability of the model’s predictions.

When the residuals are not normally distributed, it may be necessary to consider alternative modelling approaches, such as transforming the response variable, using robust regression techniques, or exploring non-parametric methods that do not rely on the assumption of normality.

A close up of a number

Description automatically generated

### Figure 16: No Multicollinearity

The VIF values for all the independent variables (PSFC, Q2, RAINC, TSLB, and Windspeed) are less than 10, indicating that the model lacks substantial multicollinearity. Multicollinearity is the condition in which two or more predictor variables are highly linked, resulting in unstable and incorrect coefficient estimations.

## **5.2. TIME SERIES FORECASTING**

TSAF are increasingly important in a variety of fields, including weather prediction. Accurate WF is critical for a variety of industries, including agriculture, transportation, and disaster management. This dissertation seeks to investigate the use of TSF techniques, namely the ARIMA models, to predict weather variables such as temperature, precipitation, and wind speed.

TSAF are increasingly important in a variety of fields, including weather prediction. Accurate WF is critical for a variety of industries, including agriculture, transportation, and disaster management. This dissertation seeks to investigate the use of TSF techniques, namely the ARIMA models, to predict weather variables such as temperature, precipitation, and wind speed.

For TSF, particularly WF, ARIMA models are extensively utilised. These models may identify trends, autocorrelation, seasonality, and other patterns in time series data. When working with non-stationary time series data, ARIMA models are very helpful since they employ differencing techniques to make the data stationary.

The degree of differencing, d, and q denote the number of moving average terms, p the number of autoregressive terms, and q the number of differencing terms in the ARIMA model. ARIMA models can also incorporate seasonal components, represented by P, D, and Q, in addition to the seasonal period (s), for seasonal data. Based on historical data, ARIMA models may be used to identify patterns and forecast future meteorological variables like wind speed, precipitation, and temperature.

A graph showing a number of times

Description automatically generated with medium confidence

A graph of a blue line

Description automatically generatedA graph of a graph

Description automatically generated with medium confidence

A graph of different types of waves

Description automatically generated with medium confidence

## **5.3. DATA EVALUATION**

We used the Scarcliffe weather dataset in this work to do an in-depth evaluation of three regression models: MLR, SVR, and RFR. The goal was to determine which model, based on several meteorological criteria, forecasted weather variables, including temperature, the most accurately and consistently.

The dataset was divided into subsets for training and testing, with the testing data being used to assess the models' performance and the training data being used to construct and optimise the models. To evaluate the reliability and robustness of the models, we utilised two commonly used assessment metrics: MAE and RMSE. The Scarcliffe dataset was used to train the MLR model, a traditional linear regression approach, which assumed a linear connection between the goal variable, temperature, and the predictor variables, pressure, humidity, rain, precipitation, and wind speed. The MLR model performed rather well on the testing data, with an RMSE of [5.4573] and an MAE of [4.62373534213109], despite its simplicity. Given that the SVR model was trained with suitable kernel functions and hyperparameters, the testing data showed an RMSE of [4.337471] and an MAE of [3.219531].

Furthermore, we evaluated the RFR model, which produced an RMSE of [4.992907] and an MAE of [4.180711] on the testing data. SVR was found to be the best accurate and dependable model for predicting weather variables in the Scarcliffe dataset based on the assessment measures.

# **CONCLUSION**

The present research has investigated the use of ML methods in the field of WF, particularly SVR , RFR, and MLR. By means of an extensive examination of existing literature and practical investigation, this study has enhanced comprehension of the potential and constraints of these models in forecasting diverse meteorological occurrences.

The results of this study have shown how ML models may improve the precision and dependability of WF. Despite its ease of use, the MLR model has shown to be a useful instrument for determining linear correlations between weather parameters and events.

Recognising these models' limits as well as the difficulties in WF is crucial, though. Achieving extremely precise forecasts can be somewhat hampered by the intrinsic complexity and chaotic character of atmospheric processes as well as the availability and calibre of weather data. For these models to be deployed effectively in real-world settings, it is still necessary to consider the interpretability of the model, hyperparameter tweaking, and feature selection.

The results of this dissertation open the door for more investigation and advancement in the field of machine learning based WF, despite these obstacles. Future research may examine how to combine these models with more established numerical weather prediction models and create ensemble methods that include many machine learning models.

In the end, effective use of ML techniques in WF might improve decision-making procedures, lessen the effects of extreme weather, and improve societal well-being in general. We may explore new areas in our comprehension and forecasting of Earth's dynamic atmospheric processes by pursuing this line of inquiry.

**TOTAL WORD COUNT: 4,263**

# **REFERENCES**

Alguraibawi, M., Midi, H. and Imon, A.H.M.R. (2015) ‘A New Robust Diagnostic Plot for Classifying Good and Bad High Leverage Points in a Multiple Linear Regression Model’, *Mathematical Problems in Engineering*, 2015, p. e279472. Available at: https://doi.org/10.1155/2015/279472.

Chu, X. *et al.* (2016) ‘Data Cleaning: Overview and Emerging Challenges’, in *Proceedings of the 2016 International Conference on Management of Data*. New York, NY, USA: Association for Computing Machinery (SIGMOD ’16), pp. 2201–2206. Available at: https://doi.org/10.1145/2882903.2912574.

Clark-Carter, D. (2005) ‘Interquartile Range’, in *Encyclopedia of Statistics in Behavioral Science*. John Wiley & Sons, Ltd. Available at: https://doi.org/10.1002/0470013192.bsa311.

Fisher, R.A. (1922) ‘The Goodness of Fit of Regression Formulae, and the Distribution of Regression Coefficients’, *Journal of the Royal Statistical Society*, 85(4), pp. 597–612. Available at: https://doi.org/10.2307/2341124.

Horton, N.J. and Lipsitz, S.R. (2001) ‘Multiple Imputation in Practice: Comparison of Software Packages for Regression Models With Missing Variables’, *The American Statistician*, 55(3), pp. 244–254. Available at: https://doi.org/10.1198/000313001317098266.

Huang, G. (2021) ‘Missing data filling method based on linear interpolation and lightgbm’, *Journal of Physics: Conference Series*, 1754(1), p. 012187. Available at: https://doi.org/10.1088/1742-6596/1754/1/012187.

Lagerquist, R., McGovern, A. and Ii, D.J.G. (2019) ‘Deep Learning for Spatially Explicit Prediction of Synoptic-Scale Fronts’, *Weather and Forecasting*, 34(4), pp. 1137–1160. Available at: https://doi.org/10.1175/WAF-D-18-0183.1.

Mohandes, M.A. *et al.* (2004) ‘Support vector machines for wind speed prediction’, *Renewable Energy*, 29(6), pp. 939–947. Available at: https://doi.org/10.1016/j.renene.2003.11.009.

Pliske, R., Crandall, B. and Klein, G. (2004) ‘Competence in weather forecasting’, *Psychological Investigations of Competence in Decision Making*, pp. 40–70.

‘Scarcliffe’ (2024) *Wikipedia*. Available at: https://en.wikipedia.org/w/index.php?title=Scarcliffe&oldid=1205119906 (Accessed: 5 May 2024).

Singh, S. *et al.* (2019) ‘Weather Forecasting using Machine Learning Techniques’, *SSRN Electronic Journal* [Preprint]. Available at: https://doi.org/10.2139/ssrn.3350281.

Uyanık, G.K. and Güler, N. (2013) ‘A Study on Multiple Linear Regression Analysis’, *Procedia - Social and Behavioral Sciences*, 106, pp. 234–240. Available at: https://doi.org/10.1016/j.sbspro.2013.12.027.

*Weather Forecast Models: The Ultimate Guide (2024)* (no date) *Climavision*. Available at: https://climavision.com/resources/the-ultimate-guide-to-weather-forecast-models/ (Accessed: 13 May 2024).

*Where is Scarcliffe? Scarcliffe on a map* (no date). Available at: https://www.getthedata.com/scarcliffe/where-is-scarcliffe (Accessed: 4 May 2024).

Yadav, M.L. and Roychoudhury, B. (2018) ‘Handling missing values: A study of popular imputation packages in R’, *Knowledge-Based Systems*, 160, pp. 104–118. Available at: https://doi.org/10.1016/j.knosys.2018.06.012.