Contents

[**1.** **ABSTRACT** 2](#_Toc175392127)

[**2.** **PROBLEM STATMENT** 2](#_Toc175392128)

[**3.** **Justification for Actions Performed** 3](#_Toc175392129)

[**3.1 Data Cleaning and Preprocessing:** 3](#_Toc175392130)

[**3.2 Script in R For Data Cleaning and Pre-Processing** 3](#_Toc175392131)

[**3.3 Exploratory Data Analysis (EDA):** 5](#_Toc175392132)

[**3.3.1 EDA Script** 6](#_Toc175392133)

[**4.** **Visualization of Cleaned Data** 7](#_Toc175392134)

[**5.** **Model Building** 9](#_Toc175392135)

[**5.1 Model Evaluation:** 9](#_Toc175392136)

[**5.2 Script of Models** 10](#_Toc175392137)

[**5.2.1 Linear Regression Model** 10](#_Toc175392138)

[**5.2.2 Random Forest Model** 11](#_Toc175392139)

[**5.2.3 Gradient Boosting Model** 11](#_Toc175392140)

[**6.** **Results and Interpretation: Model Comparison and Analysis** 12](#_Toc175392141)

[**6.1** **Interpretation** 12](#_Toc175392142)

[**Linear Regression:** 12](#_Toc175392143)

[**Random Forest:** 12](#_Toc175392144)

[**Gradient Boosting Machine (GBM):** 13](#_Toc175392145)

[**7.** **Usefulness for Organizations** 15](#_Toc175392146)

[**7.1 Agriculture:** 15](#_Toc175392147)

[**7.2 Environmental Monitoring:** 15](#_Toc175392148)

[**7.3 Resource Management:** 15](#_Toc175392149)

[**8.** **Conclusion** 15](#_Toc175392150)

[**9.** **References** 16](#_Toc175392151)

[**9.1 Journal References:** 16](#_Toc175392152)

[**9.2 Book References:** 16](#_Toc175392153)

# **ABSTRACT**

The report highlights data-analysis ranging from a brief to the depth and predictive modelling of soil moisture (SMOIS) by variables from the provided data set that are skin temperature (TSK) and 2-meter specific humidity (Q2). These variables which have been used to forecast soil moisture through simple predictive models. A weather dataset was used with over five thousand entries about variables like Skin temperature, Surface pressure, Soil Temperature among others. I performed analysis of these data by dividing my work into chunks. At First, the data was cleaned and Pre-Processed. After that I performed EDA on the cleaned dataset and then do different types of visualization such as histograms box plots and timeseries plot on it for my better understandings. After that I define my Problem statement based on given dataset and data description. It was to predict Soil Moisture. The machine learning techniques helps me to predict soil moisture. Choosing three different models on my preferences as per the given requirement. These models are Linear Regression, Random Forest, and Gradient Boosting Machine (GBM). From these three models, Random Forest model came out as the best model that is most accurate in predicting soil moisture with an R-squared value 0f 0.932 approximately 93%, followed by GBM at 91% while Linear Regression shows the least accuracy of 83%. This whole Process from cleaning of data to prediction the soil moisture is meant to help particularly for those companies that are specially working in agriculture environmental monitoring or resource management; thus enhance its ability to predict soil moisture and make prediction about it.

# **PROBLEM STATMENT**

Soil moisture prediction is a very important and critical part of many areas such as farming, environmental monitoring and resource management. Soil moisture greatly affects crops growth, irrigation needs as well as ecosystem balance. As Soil moisture connects in influencing factors like temperature and humidity in complex ways, we must have a same model to predict it.

To solve the problem, I want to create a model able of predicting soil moisture (SMOIS) using cleaned data on skin temperature (TSK) and 2-meter specific humidity (Q2). The dataset that is given includes the environmental stimulus of skin temperature or surface temperature. It is the recorded temperature, surface pressure and rainfall at intervals of time. The main goal is to find effective statistical and machine learning models to predict soil moisture, at its best.

**My major tasks include:**

1. Clean and process data to deal with missing values and outliers.
2. Perform Exploratory Data Analysis (EDA) to understand data distributions and

relationships.

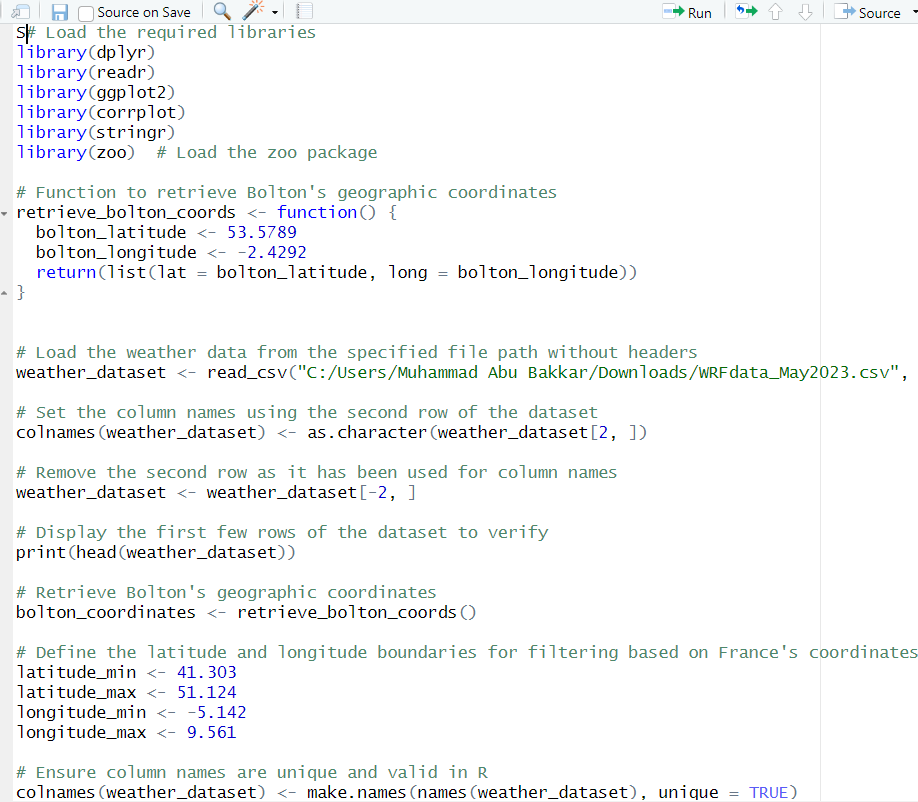
1. Build Model using Linear Regression, Random Forest, and Gradient Boosting Machine (GBM).
2. Run each model and then choose the model with most accuracy percentage.

# **Justification for Actions Performed**

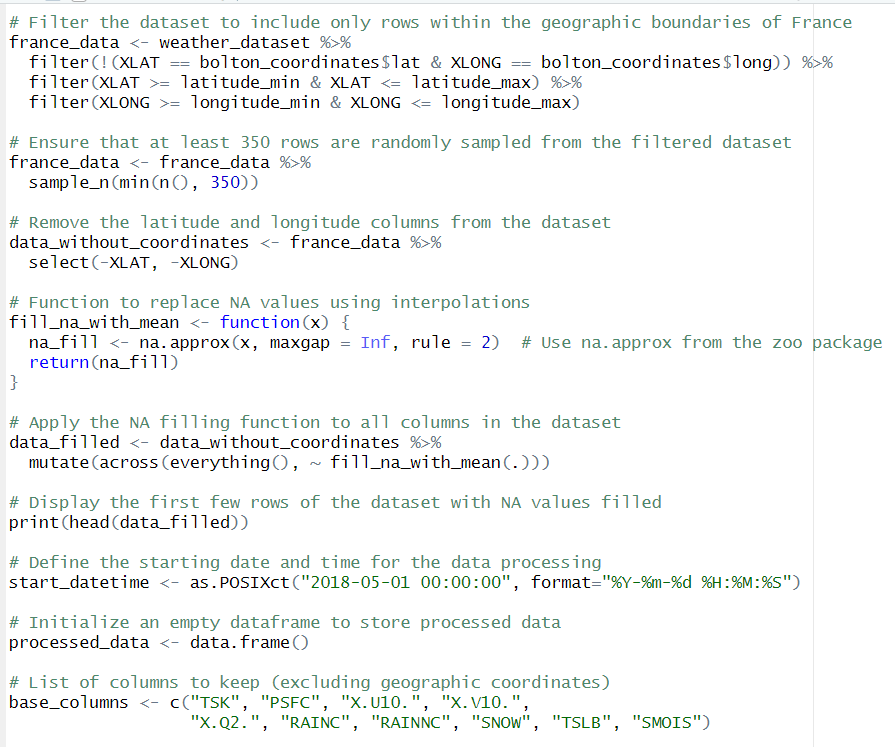
## **3.1 Data Cleaning and Preprocessing:**

* **Handling Missing Values:** In data cleaning process handling missing values are very important A technique named as imputation technique is used to adjust missing values and replace NA values with the average of the previous non-missing two components (taking the mean) in the given dataset. This technique will help to make the dataset complete so it can be utilized for modelling.
* **Removing Columns:** For the soil moisture prediction, I just focus on the independent variables that are (TSK and Q2) for my prediction task, I also removed columns of XLAT and XLONG after getting the France coordinate that are approximately ranges between latitude 41°N and 51°N and longitude 5°W to 8°. Now my problem statement become more dominant for prediction at specific area.
* **Outlier Detection:** Outliers can affect my model, so I use the IQR method to identify and to prevent extreme values. So, I Calculated IQR and removed the outliers.

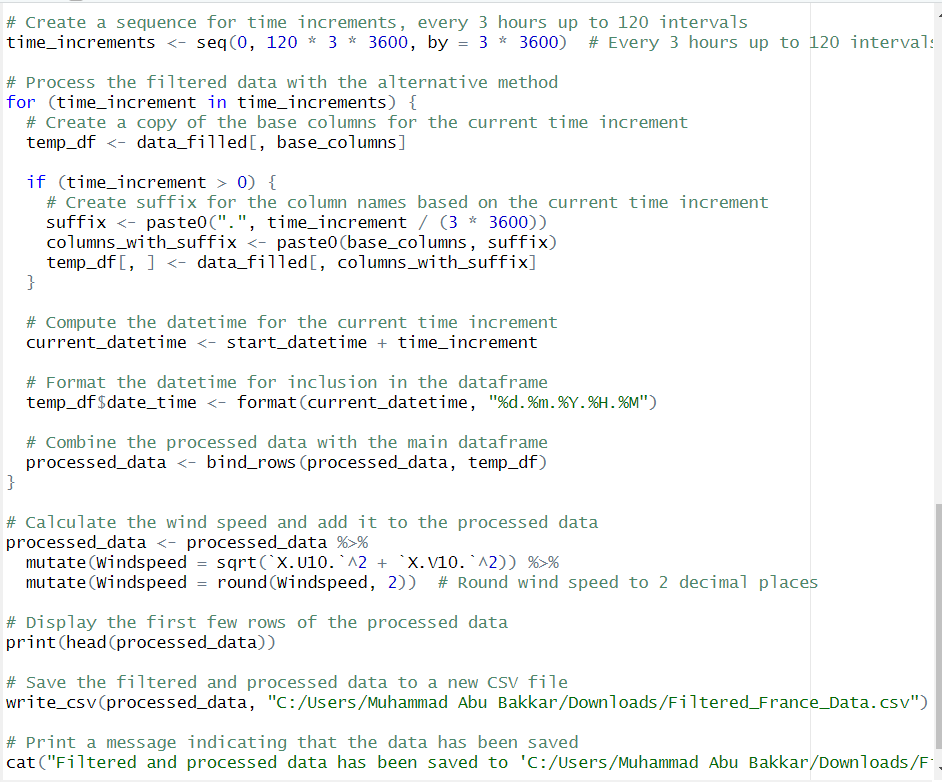
## **3.2 Script in R For Data Cleaning and Pre-Processing**



**Figure 1: First Script of Data Cleaning and Pre-Processing**

****

**Figure 2: Second Script of Data Cleaning**



**Figure 3: Third Script of R for Data Cleaning**

## **3.3 Exploratory Data Analysis (EDA):**

* **Univariate investigation:** For this purpose, each feature was analysed seperately as to understand their dispersion, central location and variability. Histograms and summary statistics were used.
* **Finding Correlations:** Relationships between variables were identified by calculating the correlation matrix. This allows us to examine predictors' relationship with soil moisture and with each other.

### **3.3.1 EDA Script**

**Figure 4: Plotting Histogram**

**Figure 5: Boxplot and Time Series Plot**

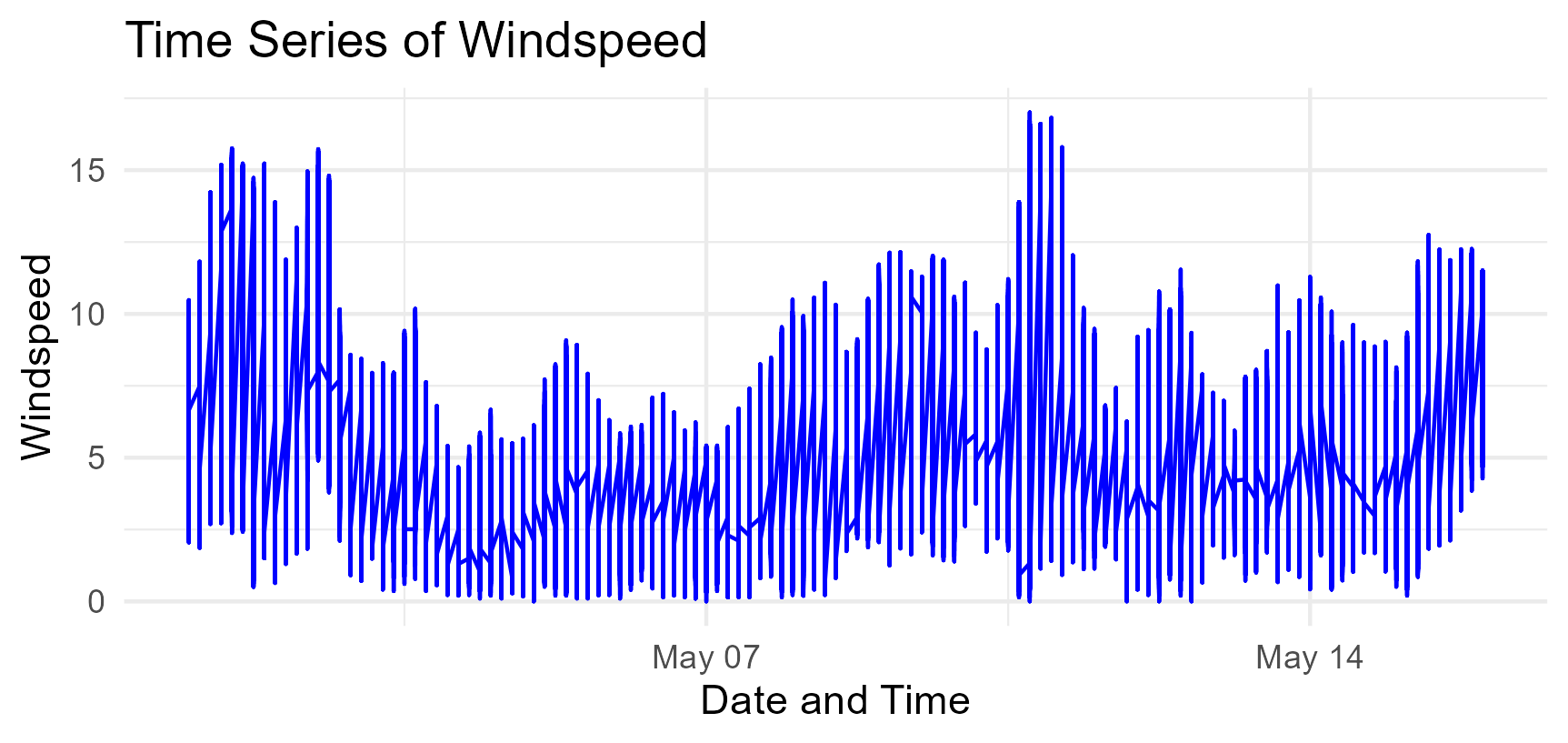
## **Visualization of Cleaned Data**

* 1. **Wind-Speed**

Wind Speed is calculated by using the X and Y component. In dataset the variables X.U10 and X.V.10 represent these components that is 10 meters above ground level in X (east-west) and y (north-south) respectively.

Then I calculated the wind speed by applying Pythagorean theorem:

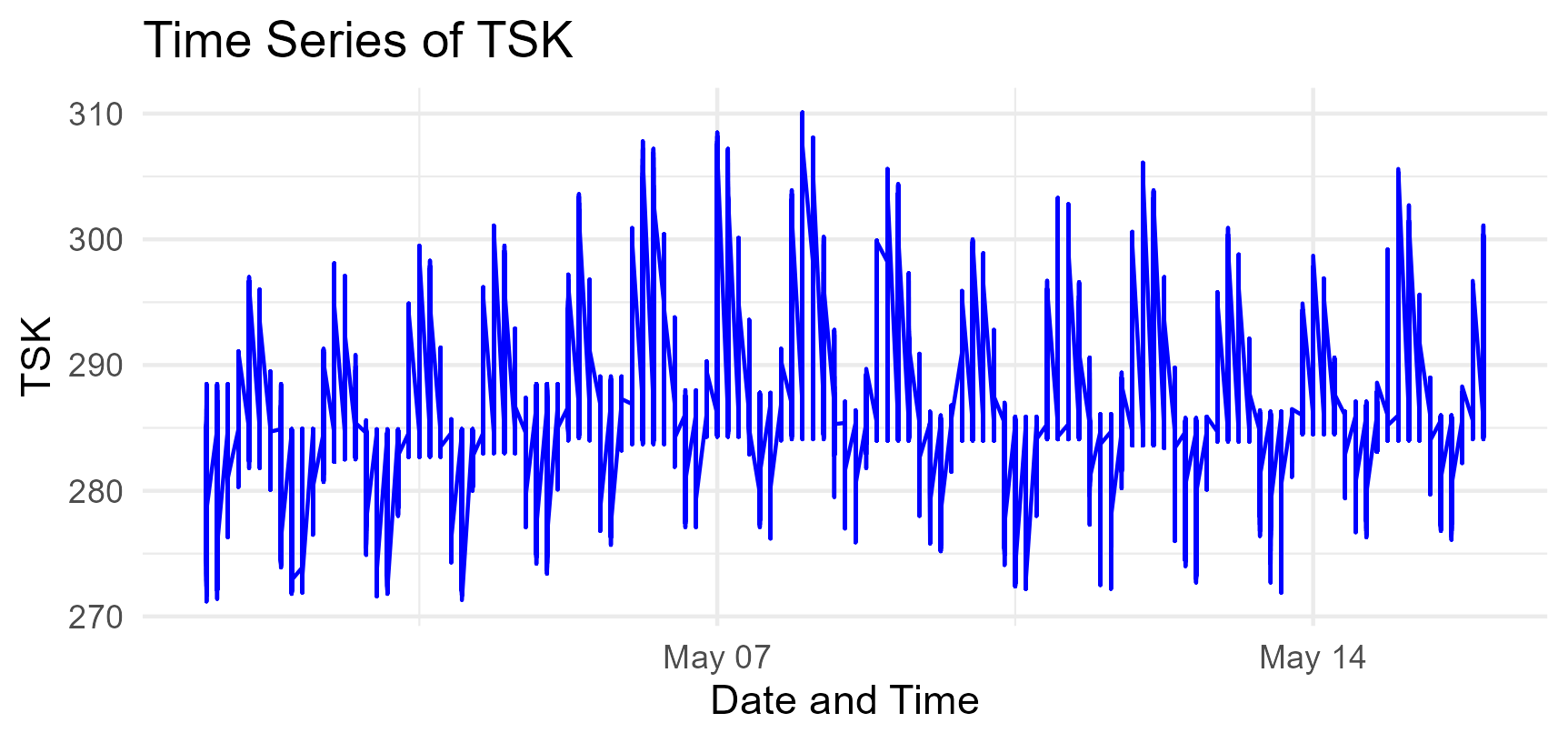
**Formula:** Wind Speed=√U2+V2

****

**Figure 6: Graphical Representation of Wind-Speed**

* 1. **TSK (Skin temperature or surface temperature)**

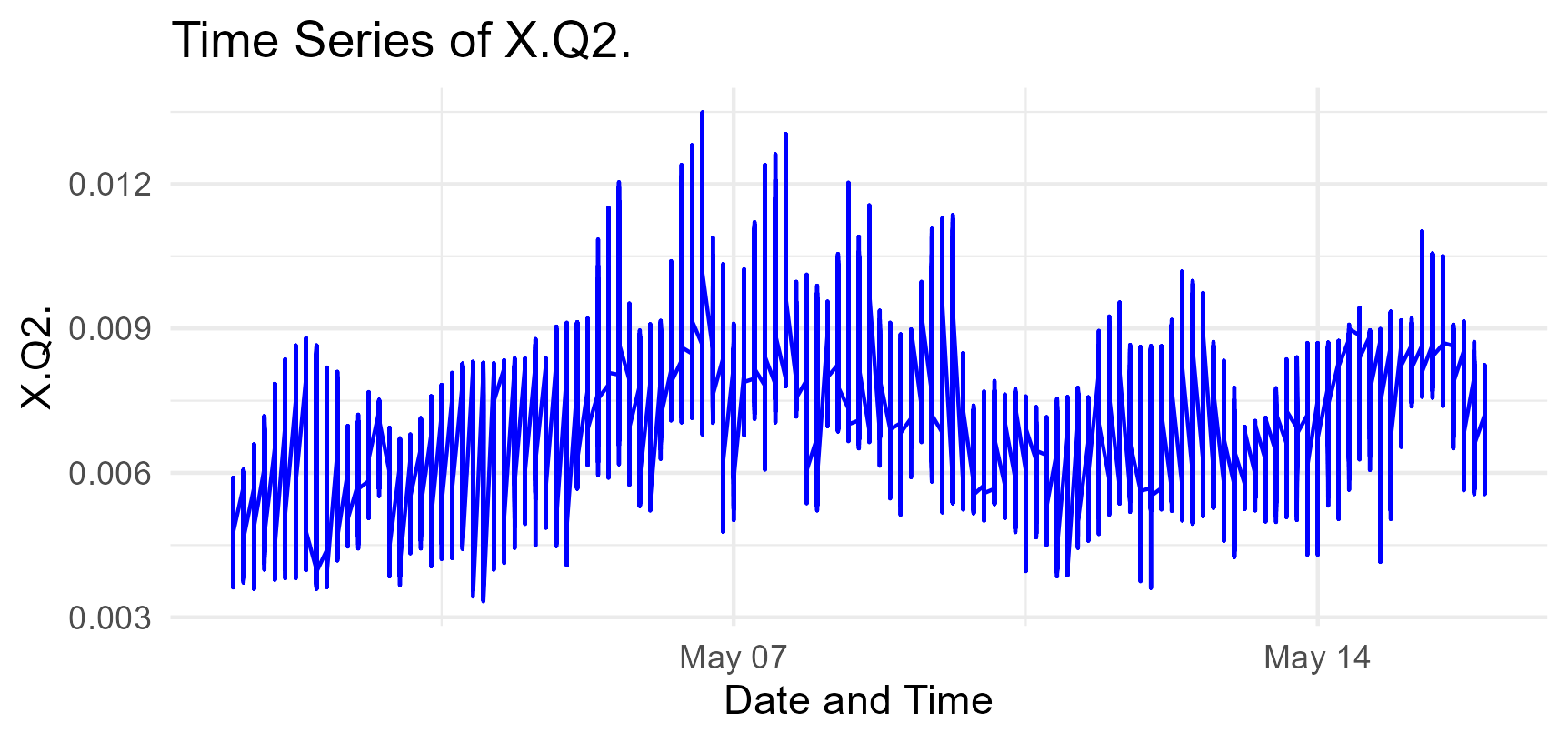
After cleaning the data as per my defined problem, I plotted the TSK for better understanding.

****

**Figure 7: TSK Plotting**

* 1. **Q2 (2- meter specific humidity)**

This helps me to understand and visualize the data in a better way also to understand my problem statement by visualizing and then choosing the model for predication.



**Figure 8: Plotting of Humidity (Q2)**

# **Model Building**

The below three models were choose to predict soil moisture based on TSK and Q2 .Their justification is mentioned below

* **Linear Regression:** To describe linear relationships between predictors and SMOIS, an initial simple model was constructed. Linear regression is fairly easy; it tells how each predictor affects SMOIS.
* **Random Forest:** This composite approach captures intricate interactions involving predictors and soil moistures. It can also handle nonlinearities due to its resistance against overfitting.
* **Gradient Boosting Machine (GBM):** GBM is a sequential tree building algorithm that improves the prediction accuracy by correcting errors made in the previous model. For this reason, it is highly predictive and flexible.

## **5.1 Model Evaluation:**

* **Performance Metrics:** I evaluated models using Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared (R2). These metrics help us understand how accurate and precise the models are.
* **Comparison of Models:** The Random Forest model showed the best performance in accuracy and fit. GBM came in second, and Linear Regression third.

## **5.2 Script of Models**

### **5.2.1 Linear Regression Model**

**Figure 9: R code of Linear Regression Model**

### **5.2.2 Random Forest Model**

**Figure 10: Code of Random Forest Model**

### **5.2.3 Gradient Boosting Model**

**Figure 11: GMB Model**

# **Results and Interpretation: Model Comparison and Analysis**

**Table 1: Comparison of Models**

|  |  |  |  |
| --- | --- | --- | --- |
| **Results** | **Linear Regression** | **Random Forest** | **Gradient Boosting Machine** |
| **MAE (Mean Absolute Error):** | 0.0899 | 0.0292 | 0.0327 |
| **MSE (Mean Squared Error)** | 0.0189 | 0.0079 | 0.0094 |
| **RMSE (Root Mean Squared Error)** | 0.1376 | 0.0891 | 0.0971 |
| **R² (R-Squared)** | 0.8384 | 0.9324 | 0.9196 |

**Best Fit Model: Random Forest, Accuracy 93%**

## **Interpretation**

**MAE:** Lower the MAE value better the accuracy.

**MSE:** Lower the MSE value better the precision.

**RMSE:** This shows the magnitude of prediction errors.

**R-Squared**: Higher R2 value better model performance

## **Linear Regression:**

**MAE: 0.0899:** This means that on average, the model's predictions are off by 0.0899 units.

**MSE: 0.0189:** This shows the average squared difference between predicted and actual values.

**RMSE: 0.1376:** This shows how much the predictions deviate from the actual values pointing to a fair amount of error.

**R2: 0.8384:** The model accounts for 83.84% of the changes in soil moisture. This suggests a good match, but there's still room to get better.

## **Random Forest:**

**MAE: 0.0292:** This reveals a smaller average error than Linear Regression hinting at improved accuracy.

**MSE: 0.0079:** This reflects a lower average squared error suggesting higher precision.

**RMSE: 0.0891:** This gives a smaller typical deviation of prediction errors showing higher accuracy.

**R2: 0.9324:** The model accounts for 93.24% of the variance showing a great fit and strong ability to predict.

## **Gradient Boosting Machine (GBM):**

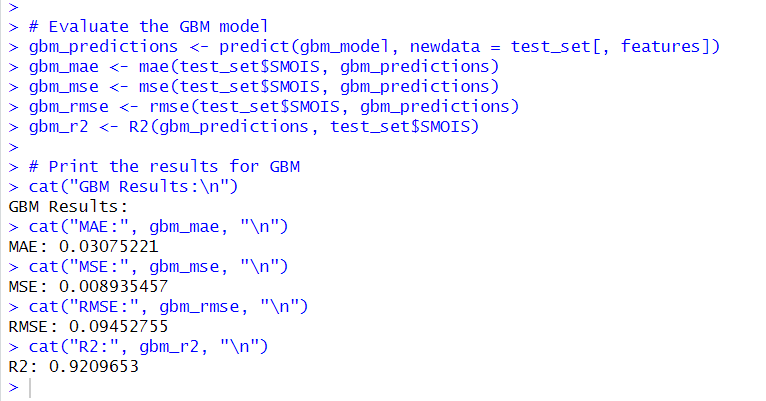
**MAE: 0.0327:** A bit higher than Random Forest pointing to a bigger average error.

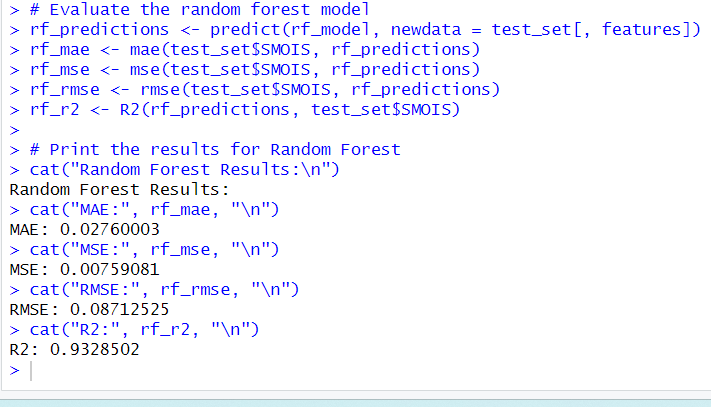
**MSE: 0.0094:** Shows a moderate average squared error.

**RMSE: 0.0971:** Points to a higher spread of errors compared to Random Forest.

**R2: 0.9196:** The model explains 91.96% of the variance demonstrating solid performance but not quite as effective as Random Forest.

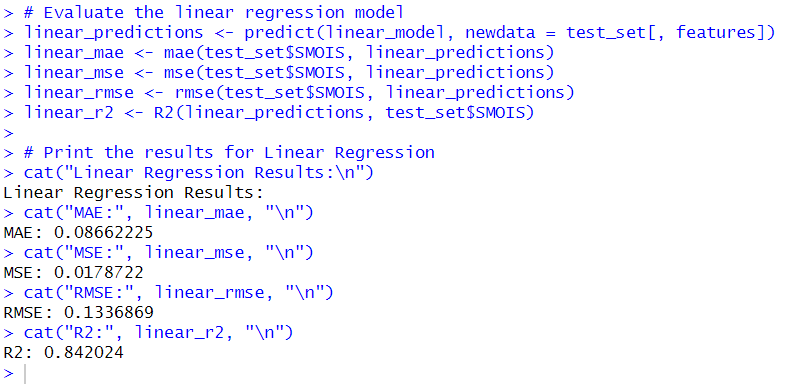
**Output of Model**



****

**Figure 12: GMB Model Result**

**Figure 13: Random Forest Results**



**Figure 14: Linear Regression Model**

# **Usefulness for Organizations**

Regarding agriculture, monitoring of the environment, and resource management as they relate to soil moisture predictions, irrigate effectively, manage water resources well and have an assessment of the health status of crops. Soils moisture levels can be understood by means of models established in this study thus giving organizations an opportunity to make informed decisions. The areas covered here are.

## **7.1 Agriculture:**

Forecasts on soil moisture could improve scheduling of irrigation, minimize water wastage and boost crop production.

## **7.2 Environmental Monitoring:**

Proper estimates about soil moisture enable a better understanding of ecological situations for better natural resource management and ability to respond appropriately to environmental changes.

## **7.3 Resource Management:**

Soil moisture forecasts help organizations reduce risks from water scarcity and land degradation through proper resource allocation.

This is how they can benefit from enhanced prediction capabilities for managing soil moisture through adoption Random Forest model which demonstrated the highest level of accuracy and fit among other regression models used.

# **Conclusion**

This report presents a process of predicting soil moisture using different statistical and machine learning methods. This is a detailed approach that encompasses data cleansing, EDA, as well as model evaluation to guarantee in-depth analysis. Using the Random Forest model resulted in the best prediction of soil moisture which has tremendous advantages for agricultural operations, environmental monitoring agencies, resource management firms. The knowledge obtained from this analysis can help to improve decision making procedures and operational effectiveness in these areas.

# **References**

## **9.1 Journal References:**

1. Schafer, J.L. (1999). Multiple imputation: a primer. Statistical Methods in Medical Research, 8(1), pp.3-15.
2. Hodge, V.J. and Austin, J. (2004). A survey of outlier detection methodologies. Artificial Intelligence Review, 22(2), pp.85-126.
3. Schober, P., Boer, C. and Schwarte, L.A. (2018). Correlation coefficients: appropriate use and interpretation. Anesthesia & Analgesia, 126(5), pp.1763-1768.
4. Montgomery, D.C., Peck, E.A. and Vining, G.G. (2012). Introduction to Linear Regression Analysis. 5th ed. New York: Wiley.
5. Breiman, L. (2001). Random forests. Machine Learning, 45(1), pp.5-32.
6. Friedman, J.H. (2001). Greedy function approximation: a gradient boosting machine. Annals of Statistics, 29(5), pp.1189-1232.
7. Willmott, C.J. and Matsuura, K. (2005). Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. Climate Research, 30(1), pp.79-82.

## **9.2 Book References:**

* Van der Aalst, W. (2016). Process Mining: Data Science in Action. 2nd ed. Berlin: Springer.
* Tukey, J.W. (1977). Exploratory Data Analysis. Reading, MA: Addison-Wesley.
* Bishop, C.M. (2006). Pattern Recognition and Machine Learning. New York: Springer.
* Mucherino, A., Papajorgji, P.J. and Pardalos, P.M. (2009). Data Mining in Agriculture. New York: Springer.
* Wainwright, J. and Mulligan, M. (eds.) (2013). Environmental Modelling: Finding Simplicity in Complexity. 2nd ed. Hoboken: Wiley-Blackwell.
* Mays, L.W. (ed.) (2010). Water Resources Engineering. 2nd ed. Hoboken: John Wiley & Sons.

**Word count = 1425**