Tequila meets Technology

A new look at Tequila

Robert Gravelle

2023-12-06

Contents

Figure 1: Background Image

Information to begin

To establish a correlation between temperature and tequila ratings, it is imperative to examine the predicted temperature values. These predictions will serve as a foundational element in uncovering potential connections between temperature fluctuations and the quality assessments of tequila.

Introduction

Introduction

Temperature prediction plays an instrumental role in various fields, including agriculture, energy management, and urban planning. Accurate temperature forecasts empower decision-makers to plan and respond effectively to weather conditions. It is imperitive in Agave growing for Tequila. One powerful tool for temperature prediction is the Random Forest modeling algorithm, a versatile and robust machine learning technique.

What is Random Forest Modeling

Random Forest is an ensemble learning method that builds a multitude of decision trees during training. Each tree in the forest independently predicts the target variable (temperature, in this case), and the final prediction is an average or a vote of all the individual tree predictions. This ensemble approach enhances the model's accuracy and generalizability.

Data Summary

The initial files we are going to be using are the weather1.csv and weather2.csv, a meteorological data source of tequila and agave growth regions. Location, sunshine hours, humidity, temperature – this file paints a vivid meteorological portrait sourced from multiple online platforms.

In our dataset, we've merged two files, distilling pertinent information into a concise set of columns. These include datetime, temp, tempmin, tempmax, dew (indicating the dewpoint), humidity, windspeed (representing wind speed in the Agave growing area known as Tequila), winddir (denoting wind direction in the Tequila agave growing region), and solar energy.

While not all these values contribute directly to the temperature prediction calculations, they play crucial roles in other analyses. The resulting dataframe encompasses a comprehensive array of meteorological parameters.

Initially, I had a variable labeled 'X' in the primary dataframe, intended for a potential new column in the merged dataset. However, I opted for an alternative approach, retaining 'X' for future use. To facilitate the merger into a new cohesive dataset named 'weather_total,' I need to exclude 'X' from the initial table, ensuring the seamless integration of these datasets.

Data Analysis and Modeling

Loading Libraries

```
## Tequila Exploratory Analysis
The first thing we need to do is load the libraries and packages.

# Install the packages if not already installed

if (!require(randomForest)) {
  install.packages("randomForest")
}

if (!requireNamespace("dendextend", quietly = TRUE)) {
  install.packages("dendextend")
}

if (!requireNamespace("ggdendro", quietly = TRUE)) {
  install.packages("ggdendro")
}
# Load the libraries
library(randomForest)
library(dendextend)
library(ggdendro)
```

Here we have loaded the needed packages if required and now we can move on to the loading the data and the dataframes.

Loading the data from the csv files and merging them into one. Here I also remove the column X from the first dataframe as I am using that in future editions of this project.

```
# Load the first data source (weather.csv)
weather1 <- "C:\\RProj\\Tequila\\weather.csv"
weather_data1 <- read.csv(weather1)

# Load the second data source (weather1.csv)
weather2 <- "C:\\RProj\\Tequila\\weather1.csv"
weather_data2 <- read.csv(weather2)

# remove column X from weather_data1
weather_data1 <- subset(weather_data1, select = -X)

# combine month, day, year into datetime
weather_data1$datetime <- as.POSIXct(paste(weather_data1$DateYear,
                                           weather_data1$DateMonth,
                                           weather_data1$DateDay,
                                           sep = "-"),
                                     format="%Y-%m-%d")

# create dataset called weather_total

# Rename columns in weather_data2 to match the column names in weather_data1
colnames(weather_data2) <- colnames(weather_data1)
```

```
# Combine the data frames while preserving datetime format
weather_total <- data.frame(do.call(rbind, lapply(list(weather_data1, weather_data2), function(df) {
  df$datetime <- as.POSIXct(paste(df$DateYear, df$DateMonth, df$DateDay, sep = "-"), format="%Y-%m-%d")
  df[, !(colnames(df) %in% c("DateMonth", "DateDay", "DateYear"))]
})))

# Select specific columns from weather_total
selected_columns <- c("datetime", "temp", "tempmin", "tempmax", "dew", "humidity", "windspeed", "winddi
weather_total <- weather_total[, selected_columns]
```

Random Forest

In the next step, we'll leverage Random Forest modeling to build a decision tree, enabling us to predict forthcoming temperature values. Subsequently, we'll explore the correlation between established reviews and temperatures. By eliminating this correlation, we aim to evaluate the viability of utilizing predicted temperatures as a method for forecasting future ratings.

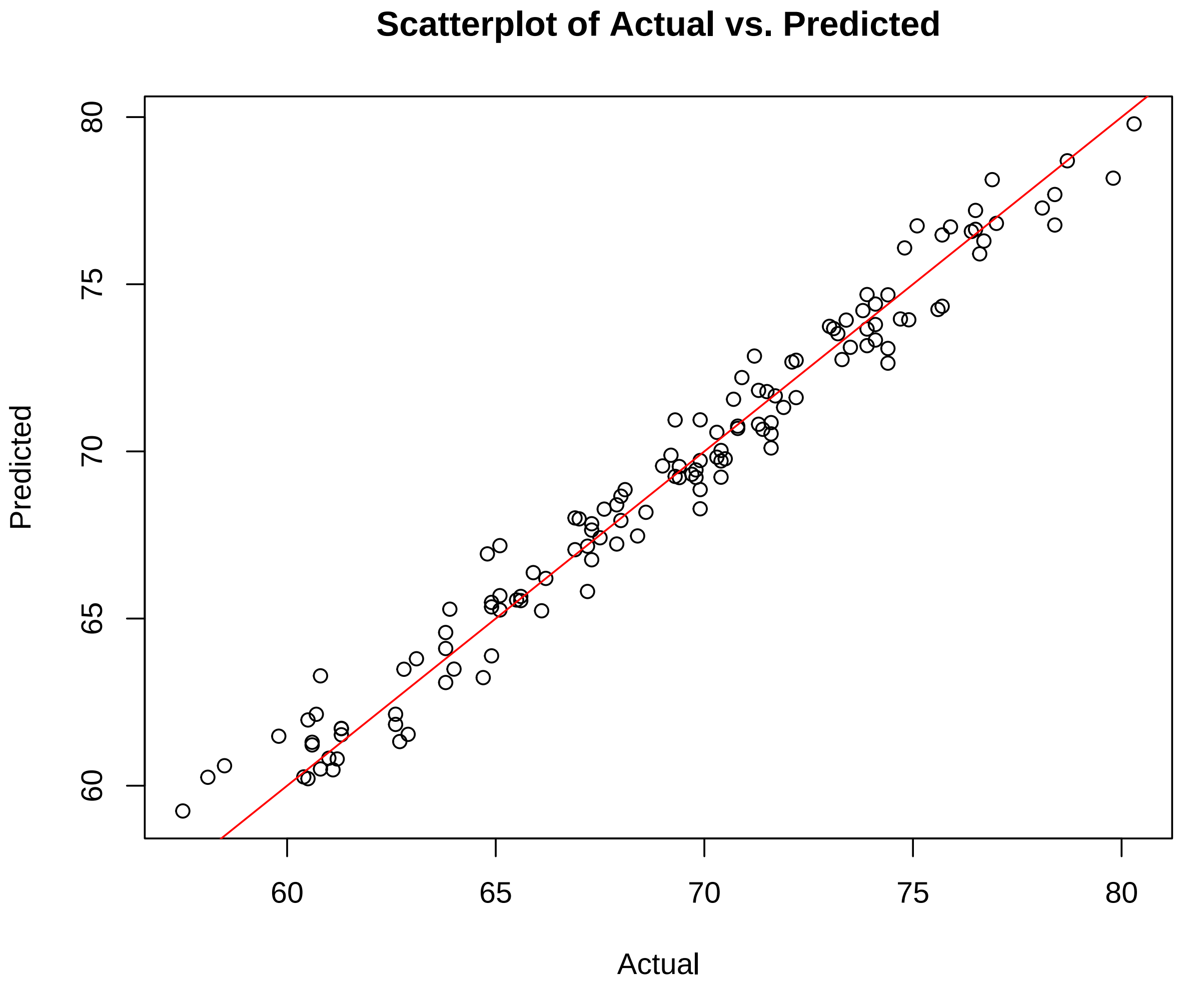
```r
# Use Forest tree to predict the future weather
# Set a random seed for reproducibility
set.seed(123)

# Split the data into training and testing sets
train_indices <- sample(1:nrow(weather_total), 0.8 * nrow(weather_total))
train_data <- weather_total[train_indices, ]
test_data <- weather_total[-train_indices, ]

# Check for missing values in the training dataset
missing_values <- sum(!complete.cases(train_data))

# If there are missing values, remove rows with missing values
if (missing_values > 0) {
  cat("Removing", missing_values, "rows with missing values.\n")
  train_data <- train_data[complete.cases(train_data), ]
}

# Train the random forest model
model <- randomForest(temp ~ ., data = train_data)

# Make predictions on the test set
predictions <- predict(model, test_data)

# Compare predictions to actual values
comparison <- data.frame(Actual = test_data$temp, Predicted = predictions)
head(comparison)

# Remove rows with missing values in the test set
test_data <- test_data[complete.cases(test_data), ]

# Make predictions on the cleaned test set
predictions <- predict(model, test_data)

# Calculate RMSE
rmse <- sqrt(mean((test_data$temp - predictions)^2))
cat("Root Mean Squared Error (RMSE):", rmse, "\n")

# Visualization the results of random forest model

# Variable Importance Plot
varImpPlot(model)

# Scatterplot of Actual vs. Predicted
plot(test_data$temp, predictions, main = "Scatterplot of Actual vs. Predicted", xlab = "Actual", ylab =
abline(0, 1, col = "red")
```

Now, let's explore a subset of the temperature predictions and compare them to the actual values. Afterward, we'll examine the Root Mean Squared Error (RMSE) values and present a scatter plot that visually contrasts the actual and predicted temperature values.

```
## Removing 552 rows with missing values.
```

	Actual	Predicted
3	64.7	63.23356
14	64.0	63.48804
15	62.8	63.48391
21	60.6	61.29906
22	59.8	61.48067
27	63.1	63.79687

```
## Root Mean Squared Error (RMSE): 0.912249
```



## Random Forest modeling

Now, we'll employ Random Forest modeling to construct a decision tree and subsequently forecast future temperature values. These values will then be used to test the correlation between the known reviews and the temperatures which we can remove and then test to see if this is a viable way of predicting future ratings.

An ensemble model is a machine learning model that combines the predictions of multiple individual models to improve overall performance and robustness.

Now, let's move on to clustering our results. Clustering, a sophisticated data analysis technique, entails the grouping of similar data points into distinctive subsets or clusters based on specific characteristics or features. The primary aim of clustering is to unveil inherent patterns or structures residing within the data, seeking out natural groupings among observations that exhibit shared similarities.

In our clustering endeavor, we'll delve into two distinct models: K-Means Clustering and Hierarchical Clustering. Understanding the conceptual underpinnings of each model is needed for grasping their distinct approaches:

K-Means Clustering: This method divides the data into a predetermined number of clusters (k), focusing on minimizing the variance within each cluster.

Hierarchical Clustering: Operating akin to the growth of a tree, this technique organizes data points into a hierarchical structure, unveiling nested clusters at varying levels of granularity.

By comprehending these clustering models, we gain insight into their unique methodologies and how they can be applied to discern meaningful patterns within our temperature prediction results.

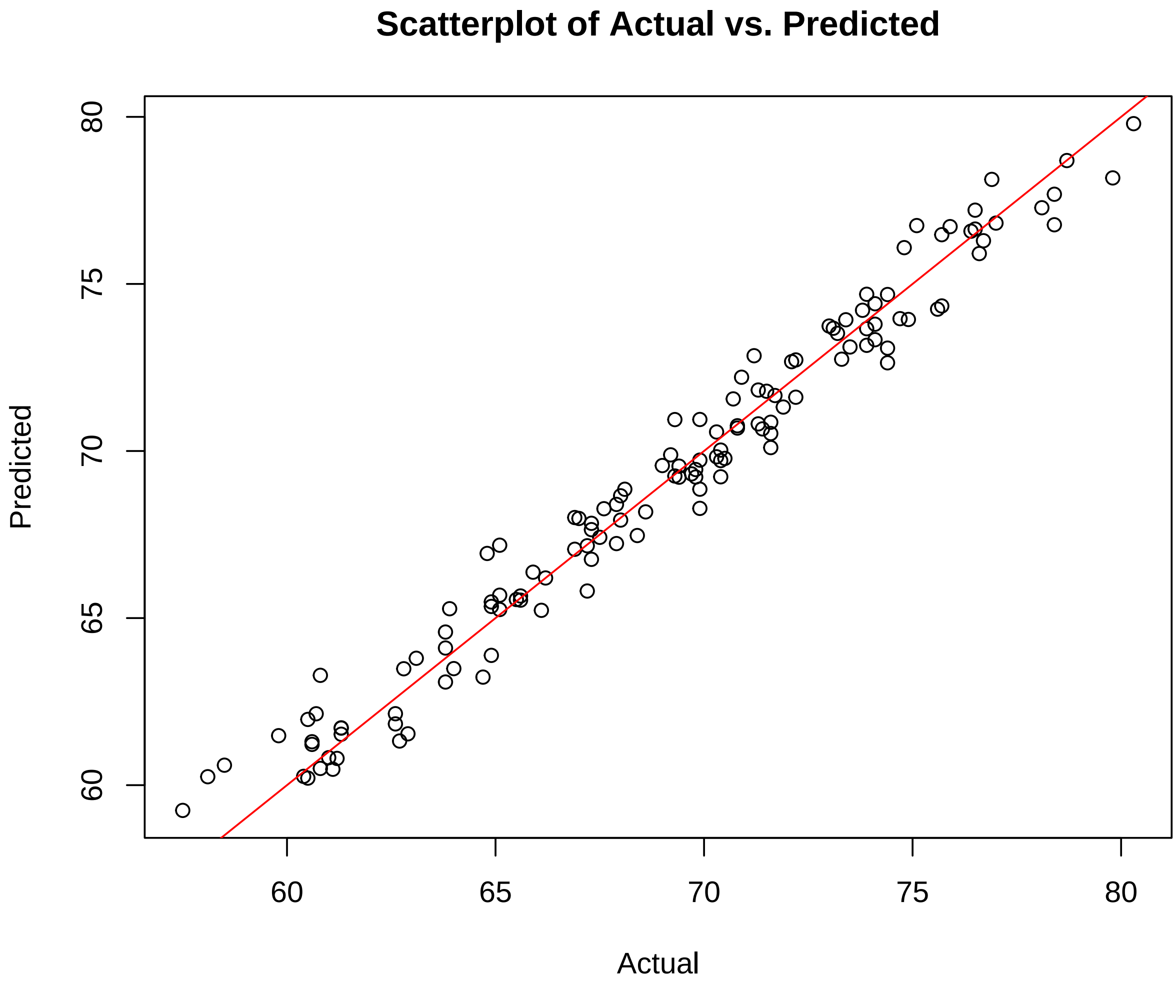
We now have a foundational understanding of K-Means Clustering and Hierarchical Clustering, let's now delve into the visual representation of our clustered results. This exploration will provide a tangible look at how these methods organize and group the predicted temperature values, allowing us to discern patterns and relationships within the clustered data.

Now, let's explore the visual outcomes of our computations. Initially, we'll examine the Clustered Results through a scatter plot, offering a clear representation of the interplay between predicted and actual temperatures. Subsequently, we'll delve into the intricacies of the dendrogram, providing a visual depiction of the hierarchical clustering structure.

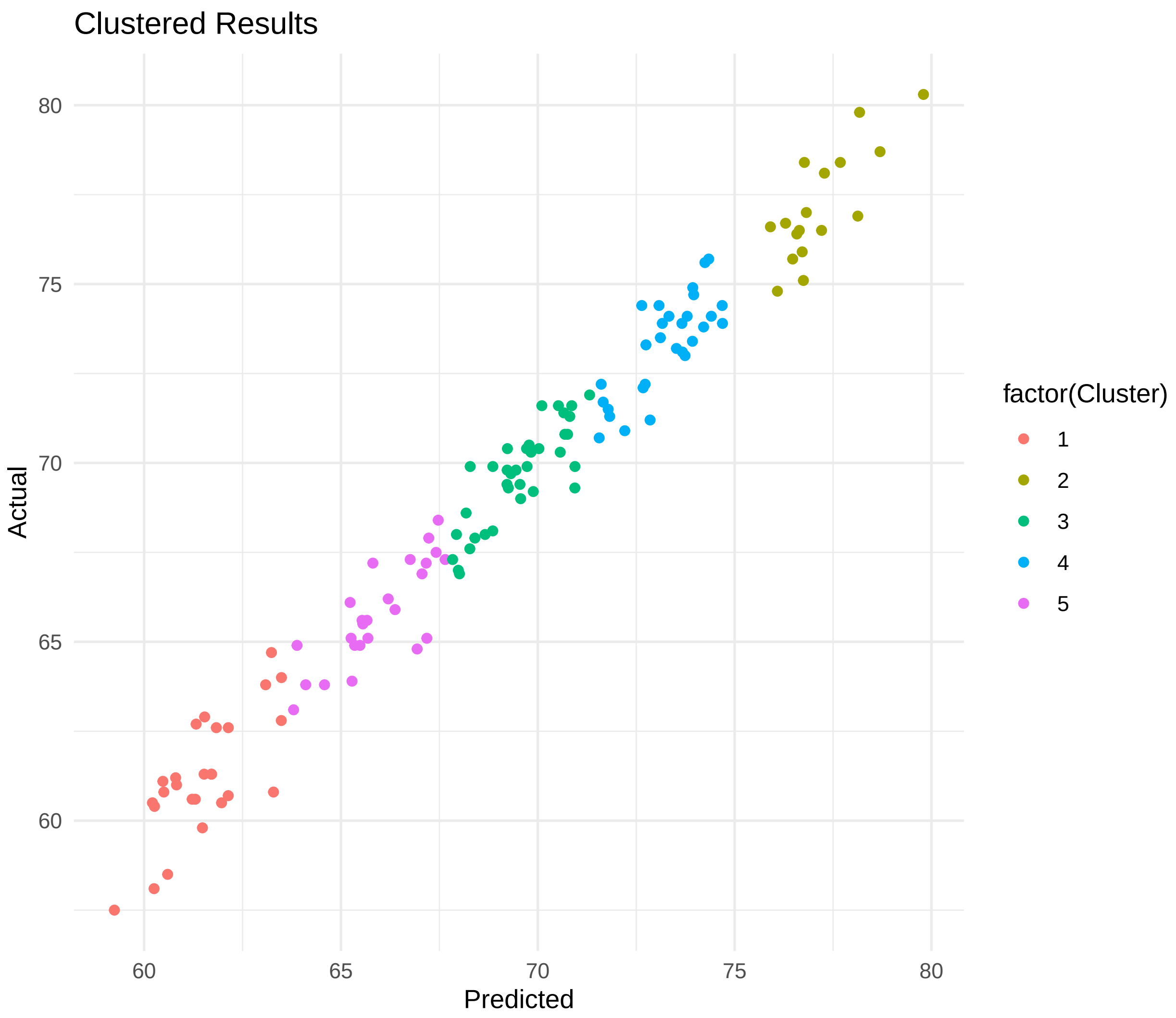
```
## Removing 552 rows with missing values.
```

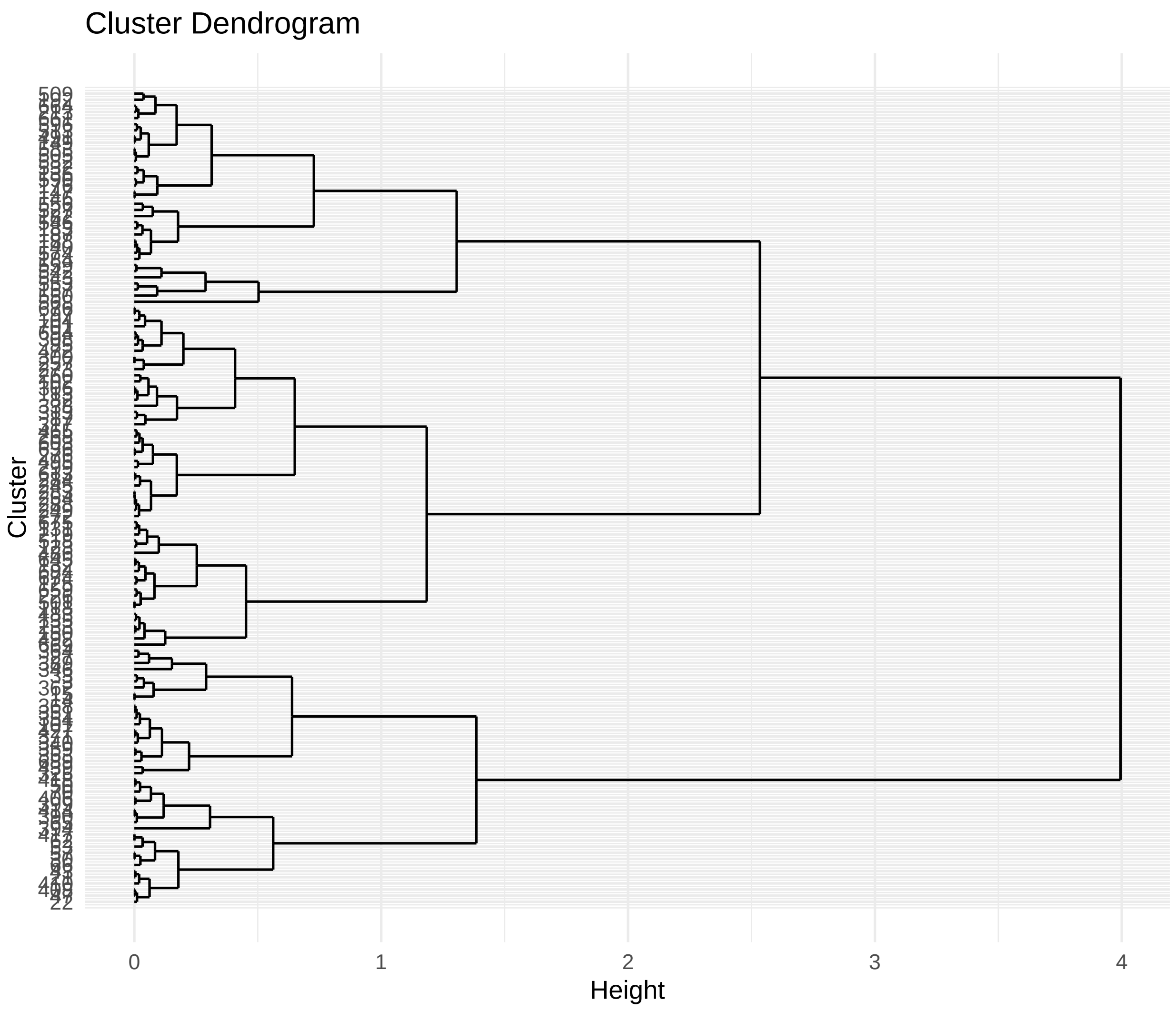
	Actual	Predicted
3	64.7	63.23356
14	64.0	63.48804
15	62.8	63.48391
21	60.6	61.29906
22	59.8	61.48067
27	63.1	63.79687

```
## Root Mean Squared Error (RMSE): 0.912249
```



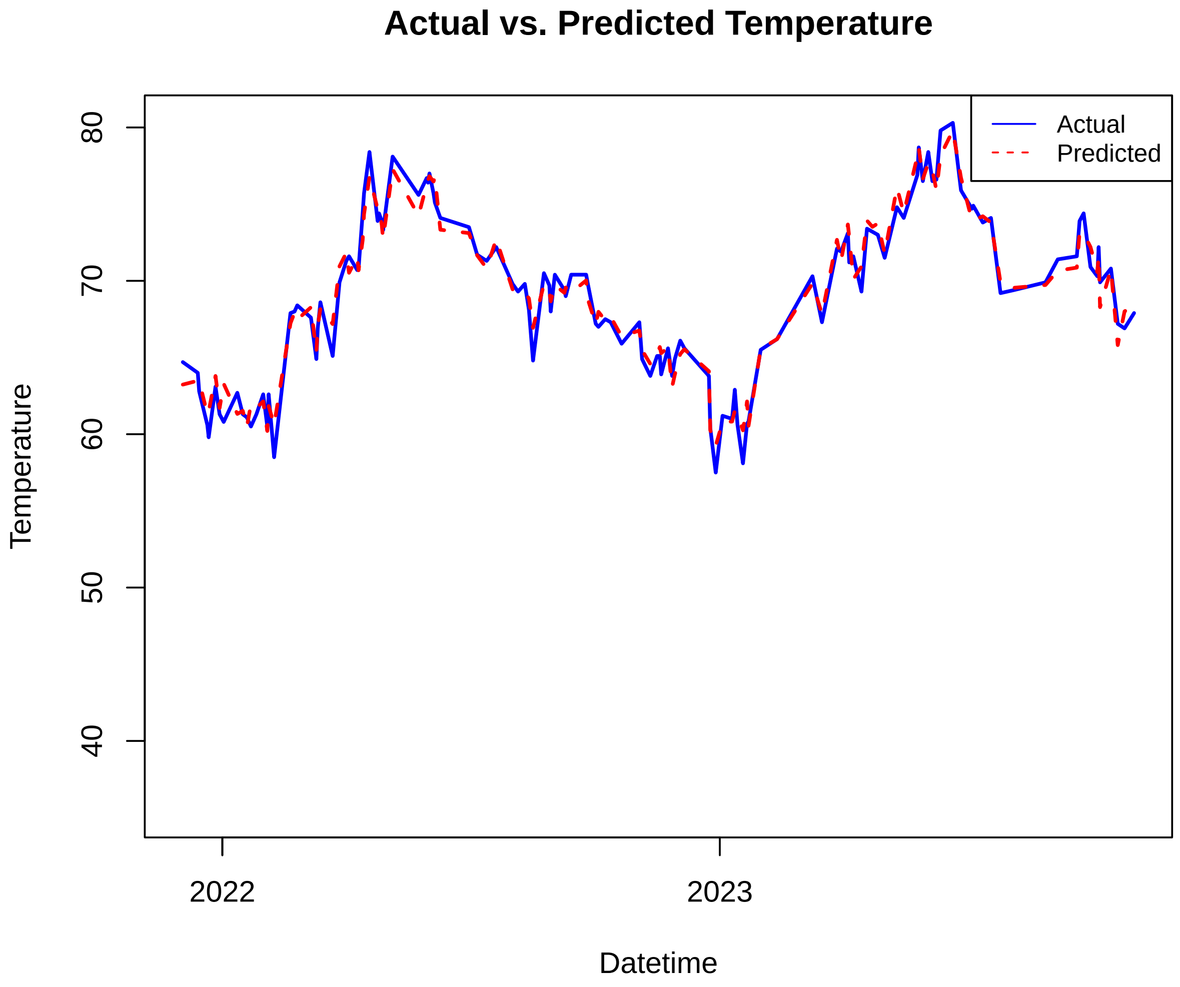
	Actual	Predicted	Cluster
3	64.7	63.23356	1
14	64.0	63.48804	1
15	62.8	63.48391	1
21	60.6	61.29906	1
22	59.8	61.48067	1
27	63.1	63.79687	5


Clustered Results
Actual
Predicted
80
75
70
65
60
60
65
70
75
80
factor(Cluster)
1
2
3
4
5


Cluster Dendrogram
Cluster
509
22
0
1
2
3
4
Height

Results

Now, we will execute a Random Forest model to compare actual versus predicted temperatures, incorporating the updated information. Subsequently, we'll visualize the results through a graph, offering a representation of the similarities between the actual and predicted temperature values.



The graphical representation reveals that the two lines, corresponding to actual and predicted temperatures, closely overlap, indicating a high degree of similarity. This visual alignment underscores the effectiveness of the algorithm, reinforcing its suitability for predicting future temperatures with a notable level of accuracy.

We are now incorporating these predicted temperatures into the main “Tequilafinalsbmt” file, enriching our dataset for the purpose of forecasting ratings and validating the data within that context.

Now, let’s examine a concise sample of the data that will be imported into the primary R source file. This sample will play a crucial role in the upcoming ratings predictions.

```
##       datetime predicted_temp actual_temp
## 36  2022-01-05       62.62506        63.0
## 426 2023-01-30       64.38092        64.6
## 617 2023-08-09       69.96094        70.0
## 82  2022-02-20       67.55558        67.9
## 615 2023-08-07       72.38680        72.4
## 470 2023-03-15       69.41591        68.0
## 591 2023-07-14       73.72744        73.3
## 182 2022-05-31       76.59972        76.7
## 283 2022-09-09       69.22652        69.4
## 226 2022-07-14       71.07551        71.3
## 582 2023-07-05       74.45833        74.7
## 394 2022-12-29       58.33952        57.5
## 546 2023-05-30       76.44367        76.5
## 214 2022-07-02       73.00402        73.0
## 275 2022-09-01       70.02048        70.3
## 495 2023-04-09       71.03947        71.6
## 686 2023-10-17       65.80148        65.9
## 14  2021-12-14       63.64553        64.0
## 412 2023-01-16       58.31863        57.4
## 482 2023-03-27       71.56695        71.6
## 7   2021-12-07       63.43370        63.8
## 49  2022-01-18       62.33420        62.2
## 663 2023-09-24       73.98893        74.8
## 570 2023-06-23       79.31206        80.3
## 193 2022-06-11       73.66966        73.6
```

Resources and References

1. Rafael Irizarry. 2018. Introduction to Data Science.
2. Jared Lander, 2017, R for Everyone Advanced Analytics and Graphics.
3. Norman Matloff, 2011 & 2019, The Art of R Programming

Data Resources

The weather data came from across the Internet but mainly from a website.

1. https://www.visualcrossing.com/ This platform imposes limitations on the number of monthly calls for its free users, but since I currently utilize the free version, I'm mindful not to incorporate any elements exclusive to the paid account in my reports. Despite having a premium subscription, my intention is to ensure that all information included remains accessible to everyone without any cost.