CAN THO UNIVERSITY SOFTWARE CENTER

MEKONG DELTA - APTECH

PROJECT – ANALYZING DATA WITH R
WEATHER FORECAST FOR OUTDOOR ACTIVITIES

Instructor:	**Class: CP2397M12**
Lam Chi Nguyen	Students:
	L23001 – Huynh Anh Thu

Can Tho, 09/2024

TABLE OF CONTENTS

FIGURES

TABLES

PHASE 1: PROBLEM DEFINITION OF PROJECT

1.1 PROBLEM

A seemingly straightforward outdoor pastime, picnicking has significant effects on people's health. It provides a necessary escape from the limits of modern living, which is defined by rising urbanization, technological reliance, and fast-paced lifestyles, beyond the simple enjoyment of food and company. Studies have repeatedly shown that engaging in outdoor activities is positively correlated with better mental health, lower levels of stress, and increased cognitive function. Because of its natural beauty and peacefulness, nature is a healing environment that can mitigate the harmful effects of burnout and chronic stress.

Especially when it comes to picnicking, there's no better way to strengthen ties with the natural world, build community, and establish a connection with others. For thousands of years, communal meals outside have been an integral part of human civilization, signifying joy, harmony, and thankfulness. In a time of increasing individualism and technological seclusion, the picnic stands in as a counterweight by encouraging in-person communication and shared experiences.

That being said, social or psychological considerations are not the only ones that influence the decision to go on a picnic. Weather circumstances have a significant influence on people's decisions.

1.2 PURPOSE

The allure of outdoor activities can be greatly influenced by variables including temperature, humidity, precipitation, wind speed, and cloud cover. Perfect weather makes for a welcoming setting, but unfavorable weather might turn off even the most ardent picnickers.

This project intends to analyze a dataset including weather conditions and corresponding picnic decisions in order to obtain a greater knowledge of the intricate relationship between weather and human behavior in the context of leisure planning. Through the identification of the primary weather-related parameters that impact picnic planning, predictive models can be developed to help people make well-informed decisions regarding outdoor activities.

Moreover, the knowledge gained from this research could benefit other industries. The tourism and hospitality sectors, for example, can use this data to improve consumer experiences and maximize outdoor offerings. Incorporating weather-resistant picnic spots into park designs is a way for urban planners to improve public health and well-being. Additionally, developing measures for climate adaptation and catastrophe preparedness can benefit from an understanding of how human behavior is influenced by weather.

Through investigating the nuances of weather-related picnic decision-making, this study aims to add to the expanding corpus of information on the interaction between humans and the environment. In the end, the results will guide the formulation of policies that support outdoor leisure, improve public health, and cultivate a society that is more resilient and sustainable.

PHASE 2: REQUIREMENTS SPECIFICATION

2.1 PROBLEM DEFINITION

As a picnic and BBQ business, we understand that a picnic is not just a relaxing occasion, but also an opportunity to create memorable memories with family and friends. However, things don't always go smoothly, and one of the factors that determines the success of a picnic is the weather. Good weather can turn a picnic into a great experience, while bad weather can disrupt or ruin the whole plan.

With the goal of bringing the most perfect experiences to customers, we have combined the picnic product business with research and analysis of weather data. By understanding factors such as humidity, pressure, global radiation, rainfall, sunlight, average and maximum temperature, we can help our customers prepare thoroughly for every situation. Whether it is sunny or rainy, we have the right solutions and products to ensure your picnic is always perfect.

This research not only helps us understand the weather better, but also helps us make accurate recommendations to our customers on which days to go camping and what to pack. With this information, we can provide products such as tents, raincoats, moisture-proof mats, coolers and many other tools to meet all needs, regardless of the weather conditions. Along with that, based on weather analysis, we can come up with weather-appropriate business strategies.

2.2 SPECIFIC REQUIREMENTS

1. What is the relationship between temperature and precipitation, and how does it affect the decision to picnic?
2. What is the impact of the relationship between humidity and sunshine on the likelihood of engaging in outdoor leisure activities?
3. What changes in global radiation influence picnic preferences and timing?
4. What are the effects of various weather patterns (e.g., clear skies, overcast, rainy) on the overall success and enjoyment of a picnic?
5. What methods can be used to build and evaluate predictive models that recommend ideal picnic days based on weather conditions?
6. What role does weather data analysis play in making recommendations for the best customer experience and creating appropriate business strategies?

PHASE 3: PROJECT PLAN

3.1 PROJECT DETAILS

3.1.1. Date of the project plan

30th July, 2024

3.1.2. Project vision/objective

1. What is the relationship between temperature and precipitation, and how does it affect the decision to picnic?
2. What is the impact of the relationship between humidity and sunshine on the likelihood of engaging in outdoor leisure activities?
3. What changes in global radiation influence picnic preferences and timing?
4. What are the effects of various weather patterns (e.g., clear skies, overcast, rainy) on the overall success and enjoyment of a picnic?
5. What methods can be used to build and evaluate predictive models that recommend ideal picnic days based on weather conditions?
6. What role does weather data analysis play in making recommendations for the best customer experience and creating appropriate business strategies?

3.1.3. Scope

The "weather prediction dataset" is intuitively accessible weather observations from 18 locations in Europe. While all selected locations provide data for the variables ‘mean temperature’, ‘max temperature’, and ‘min temperature’, we also included data for the variables 'cloud_cover', 'wind_speed', 'wind_gust', 'humidity', 'pressure', 'global_radiation', 'precipitation', 'sunshine' wherever those were available.

Location: Europe

Number of observations: 3654

Time: from 2000 to 2010.

Source: Klein Tank, A.M.G. and Coauthors, 2002. Daily dataset of 20th-century surface air temperature and precipitation series for the European Climate Assessment. Int. J. of Climatol., 22, 1441-1453.

3.2 KEY STEPS

Table 1. Key steps

Task	30/7	6/8	13/8	17/8	20/8	24/8	30/8	4/9
1. Requirement gathering	▨	▨						
1.1. Problem definition	▨	▨						
1.2. Requirement Specification	▨	▨						
1.3. Project details	▨	▨						
2. Database		▨						
3. Build and run the model			▨	▨	▨			
3.1. Data visualization			▨	▨	▨			
3.2. Descriptive analysis			▨	▨	▨			
3.3. Build model based on input datasets			▨	▨	▨			
4. Develop test models			▨	▨	▨			
5. Test (quality plan)					▨	▨		
6. Draft report						▨	▨	
7. Final report							▨	▨

3.2.1. QUALITY PLAN

1. Review Activities (Review meeting participants, frequency, and so on)

2. Testing Activities (Final Test)

Step 1: Run the chosen test models.

Step 2: Select the model with the highest accuracy

Step 3: Re-test to the selected model

Step 4: Start to write the draft report

3. Backup and Recovery Strategies (In case of disk crashes, network failures, and so on)

PHASE 4: TASK SHEET

1. Project Title: Weather Forecasting for Picnic

2. Activity Plan Prepared by: Anh Thu

3. Project start date is: 30th July 2024

4. Project end date is: 31st August 2024

Table 2. Task sheet

Task	30/7	6/8	13/8	17/8	20/8	24/8	30/8	4/9	Status
1. Requirement gathering	■	■							Done
1.1. Problem definition	■	■							Done
1.2. Requirement Specification	■	■							Done
1.3. Project details	■	■							Done
2. Database		■							Done
3. Build and run the model			■	■	■				Done
3.1. Data visualization			■	■	■				Done
3.2. Descriptive analysis			■	■	■				Done
3.3. Build model based on input datasets			■	■	■				Done
4. Develop test models			■	■	■				Done
5. Test (quality plan)					■	■			Done
6. Draft report						■	■		Done
7. Final report							■	■	Done

Prepared By:	Date	Approved by:
Huynh Anh Thu	September 2024	Lam Chi Nguyen

PHASE 5: EXPLORATORY DATA ANALYSIS

5.1 DATA SOURCES

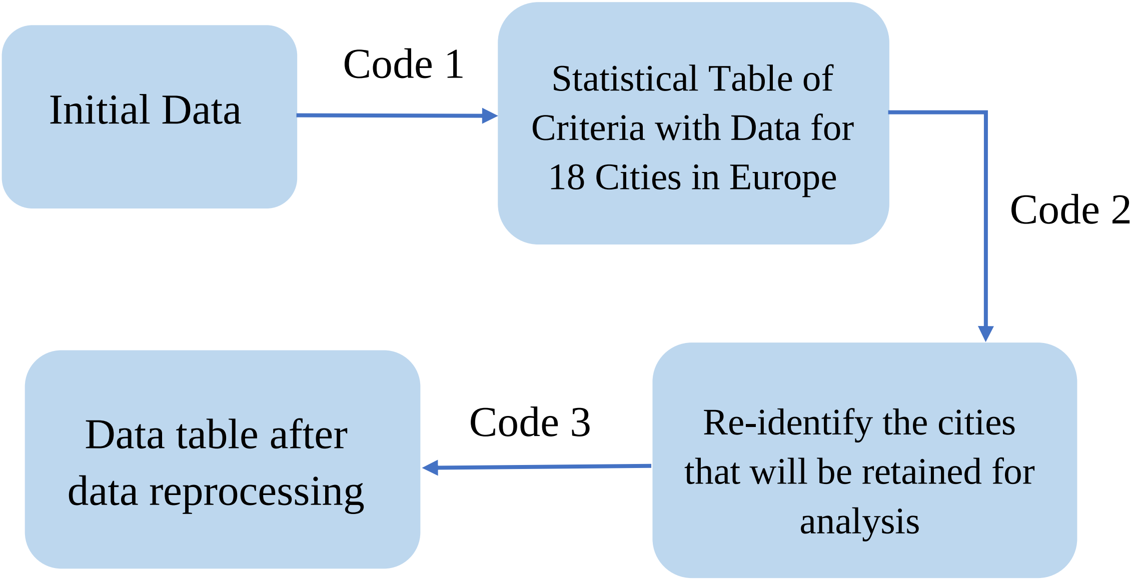

Initial Data
Code 1
Statistical Table of Criteria with Data for 18 Cities in Europe
Code 2
Re-identify the cities that will be retained for analysis
Code 3
Data table after data reprocessing

Figure 1. Data processing steps

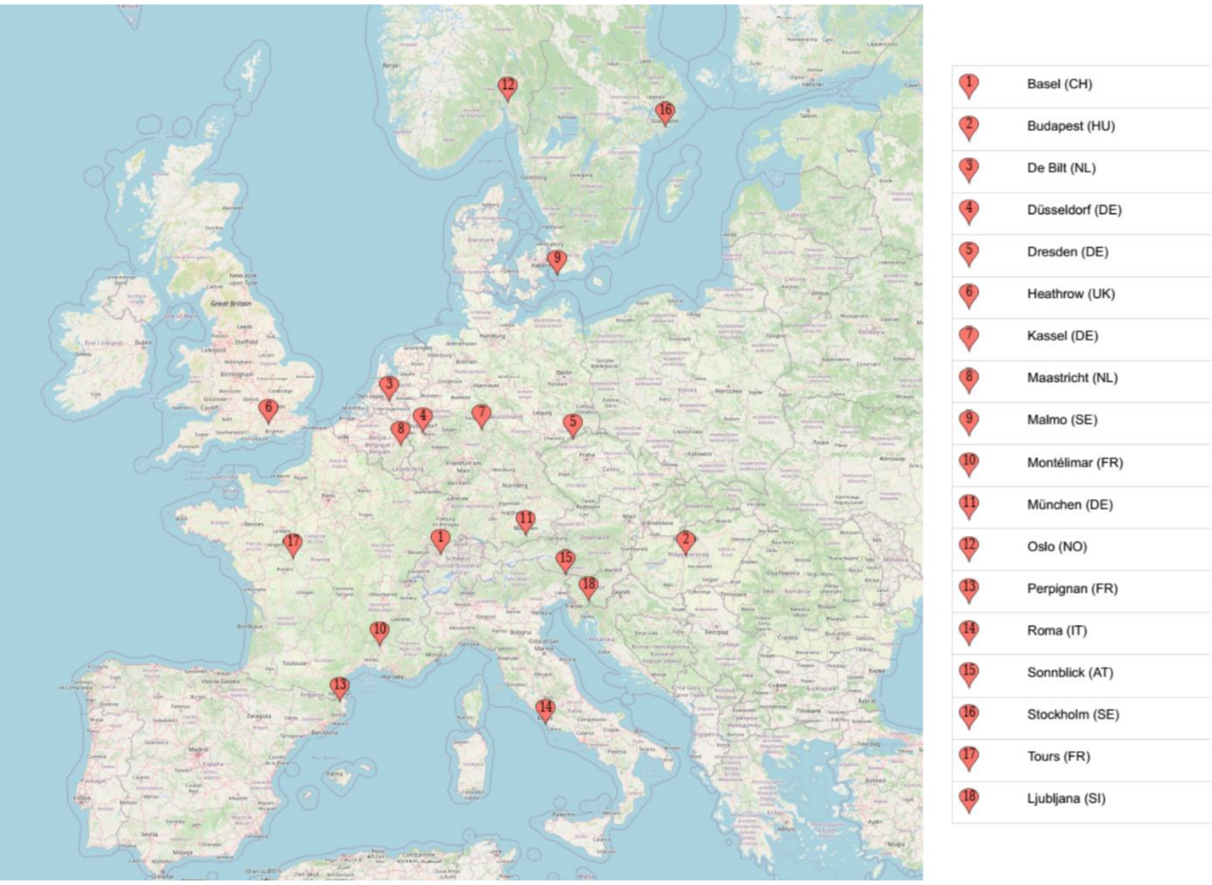
5.1.1. Initial data

The "weather prediction dataset" is intuitively accessible weather observations from 18 locations in Europe. While all selected locations provide data for the variables ‘mean temperature’, ‘max temperature’, and ‘min temperature’, we also included data for the variables 'cloud_cover', 'wind_speed', 'wind_gust', 'humidity', 'pressure', 'global_radiation', 'precipitation', 'sunshine' wherever those were available.

Location: Europe

Number of observations: 3654

Time: from 2000 to 2010.



Source: Klein Tank, A.M.G. and Coauthors, 2002. Daily dataset of 20th-century surface air temperature and precipitation series for the European Climate Assessment. Int. J. of Climatol., 22, 1441-1453.

Dataset source path: Weather Prediction (kaggle.com)

Table 3. A portion of the dataset

DATE	MONTH	BASEL_cloud_cover	BASEL_humidity	BASEL_pressure	BASEL_global_radiation	BASEL_precipitation	BASEL_sunshine	BASEL_temp_mean	BASEL_temp_min	BASEL_temp_max
20000101	1	8	0.89	1.0286	0.2	0.03	0.0	2.9	1.6	3.9
20000102	1	8	0.87	1.0318	0.25	0.0	0.0	3.6	2.7	4.8
20000103	1	5	0.81	1.0314	0.5	0.0	3.7	2.2	0.1	4.8
20000104	1	7	0.79	1.0262	0.63	0.35	6.9	3.9	0.5	7.5
20000105	1	5	0.9	1.0246	0.51	0.07	3.7	6.0	3.8	8.6
20000106	1	3	0.85	1.0244	0.56	0.0	5.7	4.2	1.9	6.9
20000107	1	8	0.84	1.0267	0.2	0.0	0.0	4.7	1.8	6.2

20000108	1	4	0.79	1.0248	0.54	0.0	4.3	5.6	4.1	8.4
20000109	1	8	0.88	1.0243	0.11	0.65	0.0	4.6	3.8	5.7
20000110	1	8	0.91	1.0337	0.06	0.09	0.0	2.4	1.4	3.8

5.1.2. Statistics of variables in each city

To determine which variables are considered in each city, use R commands to split the name of each column to identify the variable names.

```
> data <- read.csv("/Users/Documents/poject/weather_prediction_dataset.csv")

> col_names <- colnames(data)

> city_prefixes <- unique(sub("_.*", "", col_names))

> city_prefixes <- city_prefixes[city_prefixes != "DATE" & city_prefixes != "MONTH"]

> city_variables <- list()

> for (city in city_prefixes) {

+   city_cols <- grep(paste0("^", city, "_"), col_names, value = TRUE)

+   variables <- sub(paste0("^", city, "_"), "", city_cols)

+   city_variables[[city]] <- variables

+ }

> city_variables
```

From the above code, we can identify the variables present in each city. Then, we proceed to create a classification table and mark "x" for the variables for which the city has data. From this, we obtain the following table:

Table 4. Statistical Table of Criteria with Data for 18 Cities in Europe

	cloud_cover	wind_speed	wind_gust	humidity	pressure	global_radiation	precipitation	sunshine	temp_mean	temp_min	temp_max
BASEL	X			X	X	X	X	X	X	X	X
BUDAPEST	X			X	X	X	X	X	X		X
DE_BILT	X	X	X	X	X	X	X	X	X	X	X
DRESDEN	X	X	X	X		X	X	X	X	X	X
DUSSELDORF	X	X	X	X	X	X	X	X	X	X	X
HEATHROW	X			X	X	X	X	X	X	X	X

KASSEL		X	X	X	X	X	X	X	X	X	X
LJUBLJANA	X	X		X	X	X	X	X	X	X	X
MAASTRICHT	X	X	X	X	X	X	X	X	X	X	X
MALMO		X					X		X	X	X
MONTELIMAR		X		X	X	X	X		X	X	X
MUENCHEN	X	X	X	X	X	X	X	X	X	X	X
OSLO	X	X	X	X	X	X	X	X	X	X	X
PERPIGNAN		X		X	X	X	X		X	X	X
ROMA	X			X	X			X	X	X	X
SONNBLICK	X			X		X	X	X	X	X	X
STOCKHOLM	X				X		X	X	X	X	X
TOURS		X		X	X	X	X		X	X	X

The data table above provides a detailed view of the weather indicators recorded in different cities. Each city is represented by a separate set of criteria, such as cloud cover, wind speed, precipitation, air pressure, humidity, average temperature, maximum temperature, and other factors. However, not all cities have complete data for all 12 of these criteria. For example, the city of MALMO has data for only 5 criteria, while cities such as HEATHROW and DE BILT have data for most of the criteria, except for a few factors such as global radiation or sunshine.

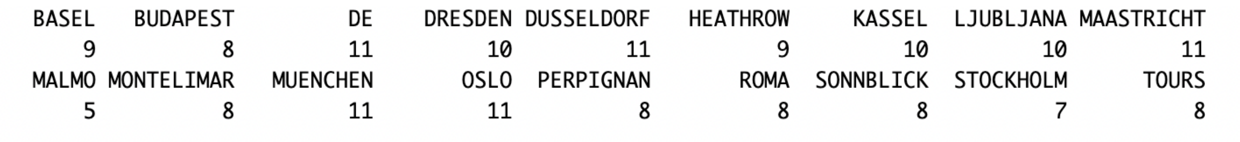
This creates a significant challenge when processing the data for analysis, especially when the goal is to compare common criteria between cities. When using all 18 cities in the dataset to identify common weather criteria, we can only select two criteria, "temp_mean" (average temperature) and "temp_max" (maximum temperature), because these are the only two criteria that are present in all cities. This leads to a serious limitation in terms of information and is not enough to conduct a comprehensive analysis or draw high-precision conclusions.

Therefore, to ensure that there are at least 6 criteria needed for a more in-depth and comprehensive data analysis, it is necessary to consider removing some cities with less data. This not only helps to increase the usability of the data, but also ensures that the analyses and comparisons are performed on a rich and meaningful set of criteria. For example, if we only retain cities with at least 6 common criteria, it will help to focus on important information and improve the reliability of the research results. Thus, excluding some cities from the analysis is not only a technical issue but also an important step to ensure the quality and depth of weather data analysis.

5.1.3. Re-identify the cities that will be retained for analysis

From the data above, proceed to determine the number of variables in each city to examine the number of variables present in each city.

```
> city_var_counts <- sapply(city_variables, length)

> print(city_var_counts)
```

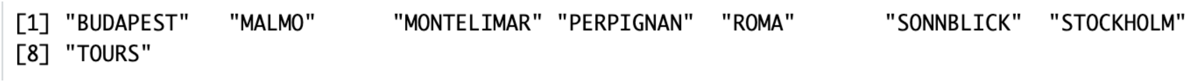


```
     BASEL   BUDAPEST         DE    DRESDEN DUSSELDORF   HEATHROW     KASSEL  LJUBLJANA MAASTRICHT 
         9          8         11         10         11          9         10         10         11 
     MALMO MONTELIMAR   MUENCHEN       OSLO  PERPIGNAN       ROMA  SONNBLICK  STOCKHOLM      TOURS 
         5          8         11         11          8          8          8          7          8 
```

Figure 2. The number of variables in each city in Europe

From Figure 2 most cities have more than 9 variables. Therefore, it might be considered to exclude cities with fewer than 9 variables to increase the diversity of the data. The code below is used to identify the names of cities with fewer than 9 variables.

```
> cities_under_9_vars <- names(city_variables)[city_var_counts < 9]

> cities_under_9_vars
```



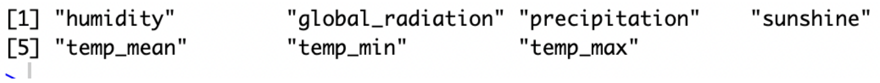
```
[1] "BUDAPEST"   "MALMO"      "MONTELIMAR" "PERPIGNAN"  "ROMA"       "SONNBLICK"  "STOCKHOLM" 
[8] "TOURS"
```

Figure 3. Cities with less than 9 variables

From Figure 3 a total of 8 cities has fewer than 9 weather criteria, specifically: Budapest (8 criteria), Malmo (5 criteria), Montelimar (8 criteria), Perpignan (8 criteria), Rome (6 criteria), Stockholm (7 criteria), Sonnblick (8 criteria), and Tours (8 criteria). This means that these cities do not provide the necessary amount of data for a comprehensive analysis if the goal is to include at least 9 common criteria across the cities.

Remove these 8 cities to see how many common variables are considered by R.

```
> data <- read.csv("/Users/Documents/poject/weather_prediction_dataset.csv", header = T)
> # Delete columns containing city names
> cleaned_data <- data %>%
+   select(
+     -starts_with("Budapest"),
+     -starts_with("Malmo"),
+     -starts_with("Montelimar"),
+     -starts_with("Perpignan"),
+     -starts_with("Roma"),
+     -starts_with("Sonnblick"),
+     -starts_with("Stockholm"),
+     -starts_with("Tours")
+   )
> # Get the column names excluding 'DATE' and 'MONTH'
> column_names <- colnames(cleaned_data)[-c(1, 2)]
> # Extract city names and variable names
> split_names <- strsplit(column_names, "_")
> cities <- sapply(split_names, function(x) x[1])
> variables <- sapply(split_names, function(x) paste(x[-1], collapse = "_"))
> # Find unique cities
> unique_cities <- unique(cities)
> # Initialize a list to store the variables for each city
> variables_by_city <- lapply(unique_cities, function(city) {
+   unique(variables[cities == city])
+ })
> # Find the intersection of variables across all cities
> common_variables <- Reduce(intersect, variables_by_city)
> # Print the common variables
> print(common_variables)
```

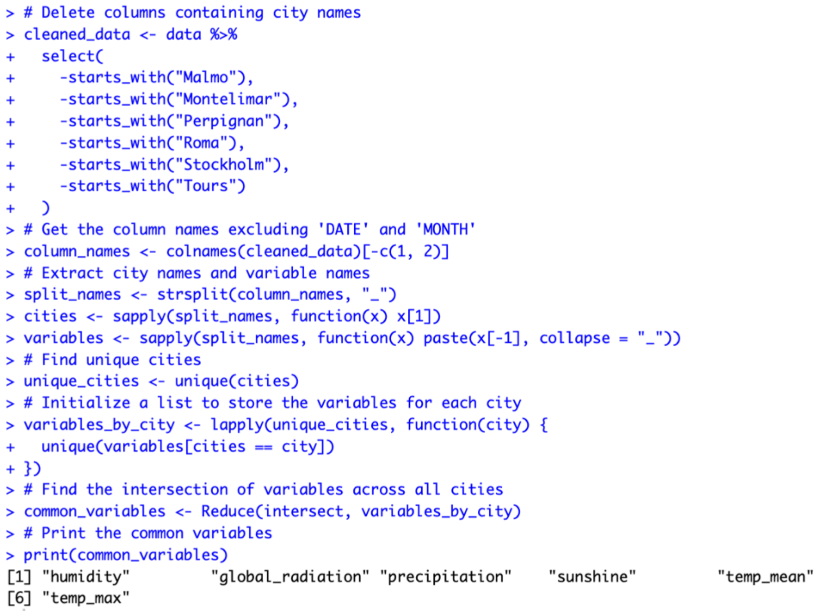


```
[1] "humidity"         "global_radiation" "precipitation"    "sunshine"
[5] "temp_mean"        "temp_min"         "temp_max"
```

Figure 4. Common variables when eliminating cities with less than 9 criteria

From figure 4 if these 8 cities are excluded, the number of common criteria that can be analyzed increases to 8 criteria, including: "humidity," "global_radiation," "precipitation," "sunshine," "temp_mean," "temp_min," "temp_max," and "Go_picnic."

However, retaining all three temperature criteria (mean temperature, minimum temperature, maximum temperature) is unnecessary if the goal is to narrow down the list to more distinct criteria. Therefore, keeping only two of the three temperature criteria, such as "temp_mean" (mean temperature) and "temp_max" (maximum temperature), is sufficient to represent the temperature factor. The exclusion of "temp_min" (minimum temperature) can be justified for several important reasons. Firstly, "temp_mean" provides an overall view of the general temperature conditions of a region over a period, reflecting the average temperature experienced by the region and helping us understand the overall weather situation. On the other hand, "temp_max" provides information about extreme maximum temperatures, which is crucial for assessing severe weather conditions such as extreme heat waves. This data is often used in studies on long-term temperature trends and climate change. Retaining all three temperature criteria may lead to information redundancy due to their high correlation. Thus, "temp_min" is less useful for assessing general temperature changes and extreme conditions compared to "temp_mean" and "temp_max." By focusing on "temp_mean" and "temp_max," we can optimize data analysis, avoid information redundancy, and concentrate on indicators that provide higher informational value. Therefore, combining with the Statistical Table of Criteria with Data for 18 Cities in Europe we can retain the cities of Budapest and Sonnblick.



```
> # Delete columns containing city names
> cleaned_data <- data %>%
+   select(
+     -starts_with("Malmo"),
+     -starts_with("Montelimar"),
+     -starts_with("Perpignan"),
+     -starts_with("Roma"),
+     -starts_with("Stockholm"),
+     -starts_with("Tours")
+   )
> # Get the column names excluding 'DATE' and 'MONTH'
> column_names <- colnames(cleaned_data)[-c(1, 2)]
> # Extract city names and variable names
> split_names <- strsplit(column_names, "_")
> cities <- sapply(split_names, function(x) x[1])
> variables <- sapply(split_names, function(x) paste(x[-1], collapse = "_"))
> # Find unique cities
> unique_cities <- unique(cities)
> # Initialize a list to store the variables for each city
> variables_by_city <- lapply(unique_cities, function(city) {
+   unique(variables[cities == city])
+ })
> # Find the intersection of variables across all cities
> common_variables <- Reduce(intersect, variables_by_city)
> # Print the common variables
> print(common_variables)
[1] "humidity"         "global_radiation" "precipitation"    "sunshine"         "temp_mean"
[6] "temp_max"
```

Figure 5. Common variables after retaining the two cities Budapest and Sonnblick

After removing the 6 cities with the fewest criteria—namely Malmo, Montelimar, Perpignan, Rome, Stockholm, and Tours—we will be left with 12 cities with a sufficient number of criteria to ensure meaningful analysis. The remaining cities are: Basel, Budapest, De Bilt, Dresden, Dusseldorf, Heathrow, Kassel, Ljubljana, Maastricht, Muenchen, Oslo, and Stockholm. With 6 criteria considered as follows: "humidity", "global_radiation" "precipitation", "sunshine", "temp_mean", "temp_max"

5.1.4. Create a new dataset

From the analysis and excluding the cities above, use R to consolidate the data into a new 'Main Data' using the following code:

```
> library(readxl)
> library(xlsx)
> library(dplyr)
> library(tidyr)
>data2<-read.csv("/Users/Documents/poject/weather_prediction_bbq_labels.csv",
header = T)
>data1<read.csv("/Users/Documents/poject/weather_prediction_dataset.csv",
header = T)
> # Loại bỏ các cột có chứa tên thành phố
> cleaned_data <- data1 %>%
+   select(
+     -starts_with("Malmo"),
+     -starts_with("Montelimar"),
+     -starts_with("Perpignan"),
+     -starts_with("Roma"),
+     -starts_with("Stockholm"),
+     -starts_with("Tours"))
> # Get the column names excluding 'DATE' and 'MONTH'
> column_names <- colnames(cleaned_data)[-c(1, 2)]
> # Extract city names and variable names
> split_names <- strsplit(column_names, "_")
> cities <- sapply(split_names, function(x) x[1])
> variables <- sapply(split_names, function(x) paste(x[-1], collapse = "_"))
> # Find unique cities
> unique_cities <- unique(cities)
> # Initialize a list to store the variables for each city
> variables_by_city <- lapply(unique_cities, function(city) {
+   unique(variables[cities == city])})
> # Find the intersection of variables across all cities
> common_variables <- Reduce(intersect, variables_by_city)
> # Lấy tên cột hiện tại
> new_colnames <- colnames(cleaned_data)
> # Xóa tên thành phố khỏi tên cột, chỉ giữ lại tên biến
> new_colnames <- gsub("^[A-Za-z]+_", "", new_colnames)
> # Cập nhật lại tên cột trong cleaned_data
> colnames(cleaned_data) <- new_colnames
> # Khởi tạo danh sách để lưu các cột đã gộp
> stacked_columns_list <- list()
> # Gộp các cột có cùng tên và lưu vào danh sách
> for (col in common_variables) {
+   # Tìm các cột có cùng tên cần gộp
+   matching_columns <- cleaned_data[, grep(col, names(cleaned_data))]
+   # Xếp chồng dữ liệu của các cột
+   stacked_column <- unlist(matching_columns)
+   # Chuyển vector thành dataframe với tên cột đúng
+   stacked_column_df <- data.frame(stacked_column)
+   colnames(stacked_column_df) <- col
+   # Thêm cột đã gộp vào danh sách
+   stacked_columns_list[[col]] <- stacked_column_df
+ }
> # Kết hợp các cột đã gộp vào dataframe mới
> df_new <- do.call(cbind, stacked_columns_list)
> # Nhân dữ liệu của cột date lên gấp 12 lần
> data1_date <- data.frame(DATE = data1$DATE)
> data1_date <- data.frame(DATE = rep(data1_date$DATE, 12))
> # Nhân dữ liệu của cột city2 lên gấp 12 lan
> data1_month <- data.frame(MONTH = data1$MONTH)
> data1_month <- data.frame(MONTH = rep(data1_month$MONTH, 12))
> # Tạo cột mới "sesion" dựa trên giá trị của cột city2
> data_session <- data.frame(SEASION = data1$session)
> data_session <- data.frame( SEASION = ifelse(data1_month$MONTH < 5 &
data1_month$MONTH > 1 , "Spring",
+                                              ifelse(data1_month$MONTH < 8 &
data1_month$MONTH > 4, "Summer",
+                                              ifelse(data1_month$MONTH < 11 &
data1_month$MONTH >7, "Autumn", "Winter"))))
> # Loại bỏ các cột có chứa tên thành phố
> cleaned_data2 <- data2 %>%
+   select(
+     -starts_with("Malmo"),
+     -starts_with("Montelimar"),
+     -starts_with("Perpignan"),
+     -starts_with("Roma"),
+     -starts_with("Stockholm"),
+     -starts_with("Tours"))
```

```
> # Lấy tên cột hiện tại
> new_colnames2 <- colnames(cleaned_data2)
> # Xóa tên thành phố khỏi tên cột, chỉ giữ lại tên biến
> new_colnames2 <- gsub("^[A-Za-z]+_", "", new_colnames2)
> # Cập nhật lại tên cột trong cleaned_data
> colnames(cleaned_data2) <- new_colnames2
> # Tên gốc của các cột cần gộp
> base_name <- "BBQ_weather"
> # Tìm các cột có cùng tên gốc cần gộp
> matching_columns <- cleaned_data2[, grep(base_name, names(cleaned_data2))]
> # Xếp chồng dữ liệu của các cột
> stacked_column <- unlist(matching_columns)
> # Tạo dataframe mới với cột đã gộp
> df_new2 <- data.frame(GO_PICNIC = stacked_column)
> Main_data <- cbind(data1_date,data1_month,data_session,df_new,df_new2)
> # Lưu kết quả vào file CSV mới
> write_xlsx(Main_data, "/Users/Documents/poject/Main Data.xlsx")
```

Table 5. Data table after data reprocessing

DATE	MONTH	SEASON	humidity	global_radiation	precipitation	sunshine	temp_mean	temp_max	GO_PICNIC
20000101	1	Winter	0.89	0.2	0.03	0	2.9	3.9	NO
20000102	1	Winter	0.87	0.25	0	0	3.6	4.8	NO
20000103	1	Winter	0.81	0.5	0	3.7	2.2	4.8	NO
20000104	1	Winter	0.79	0.63	0.35	6.9	3.9	7.5	NO
20000105	1	Winter	0.9	0.51	0.07	3.7	6	8.6	NO
20000106	1	Winter	0.85	0.56	0	5.7	4.2	6.9	NO
20000107	1	Winter	0.84	0.2	0	0	4.7	6.2	NO
20000108	1	Winter	0.79	0.54	0	4.3	5.6	8.4	NO

5.2 DATA DESCRIPTION

5.2.1. Structure data

No.	Field Name	Data Types	Level of measurements	Description
1	DATE	Character	Obligatory	Date, month, and year where the observation took place.

2	MONTH	Character	Obligatory	The month where the observation took place.
3	SEASON	Character	Obligatory	The season where the observation took place.
4	humidity	Numeric	Obligatory	Humidity (unit: %)
5	global_radiation	Numeric	Obligatory	Global radiation (unit: 100 W/m2)
6	precipitation	Numeric	Obligatory	Precipitation amount (unit: 10 mm)
7	sunshine	Numeric	Obligatory	Sunshine in 1 Hours
8	temp_mean	Numeric	Obligatory	Mean temperature (unit: Celsius)
9	temp_max	Numeric	Obligatory	Max temperature (unit: Celsius)
10	GO_PICNIC	Character	Obligatory	Decide if the weather is suitable for a picnic (Yes/No)

Table 6. Structure data

5.2.2. Dataset used for analyzing

Table 7. Dataset used for analyzing

DATE	MONTH	SEASON	humidity	global_radiation	precipitation	sunshine	temp_mean	temp_max	GO_PICNIC
20000101	1	Winter	0.89	0.2	0.03	0	2.9	3.9	NO
20000102	1	Winter	0.87	0.25	0	0	3.6	4.8	NO
20000103	1	Winter	0.81	0.5	0	3.7	2.2	4.8	NO
20000104	1	Winter	0.79	0.63	0.35	6.9	3.9	7.5	NO
20000105	1	Winter	0.9	0.51	0.07	3.7	6	8.6	NO
20000106	1	Winter	0.85	0.56	0	5.7	4.2	6.9	NO
20000107	1	Winter	0.84	0.2	0	0	4.7	6.2	NO
20000108	1	Winter	0.79	0.54	0	4.3	5.6	8.4	NO

5.3 VISUALIZATION AND BASICS EVALUATION (DESCRIPTIVE ANALYTICS)

5.3.1. Humidity

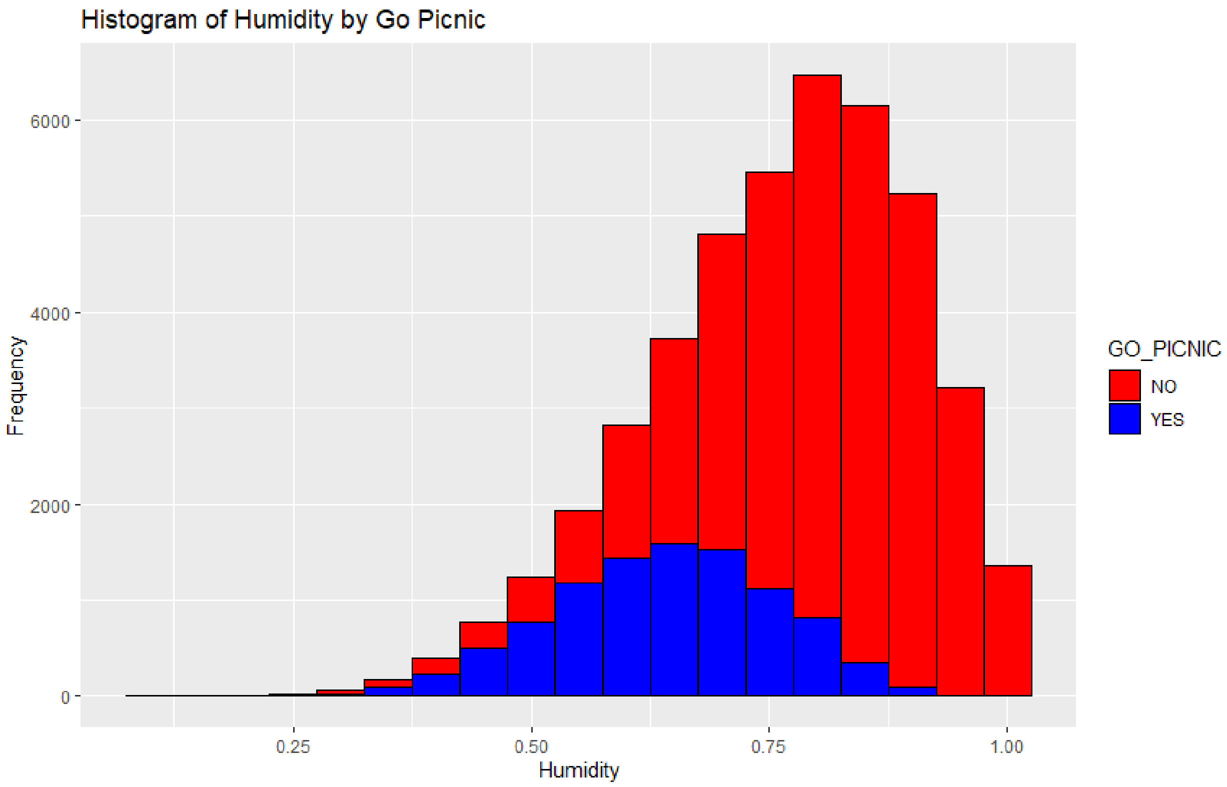


Figure 6. Histogram of Humidity by Go Picnic

The humidity distribution chart shows that during the survey period, humidity was generally concentrated around 0.75, with most days having humidity levels between 0.5

and 0.8. The number of no picnic days (“NO”) peaked at humidity levels between 0.7 and 0.8, indicating that as humidity increased, the likelihood of people deciding not to picnic also increased significantly. Although the number of no-picnic days decreased gradually at humidity levels between 0.8 and 1, it remained relatively high. In contrast, days with humidity levels below 0.5 were the days when people were more likely to decide to picnic (“YES”), reflected in the higher number of “YES” days at lower humidity levels. However, at higher humidity levels, the number of “YES” days decreased significantly. The chart is skewed to the right, with high humidity being more prevalent, and this strongly influenced the decision not to picnic. In summary, humidity is an important factor, with people tending to avoid outdoor activities when humidity is too high, especially 0.7 or higher.

5.3.2. Global Radiation

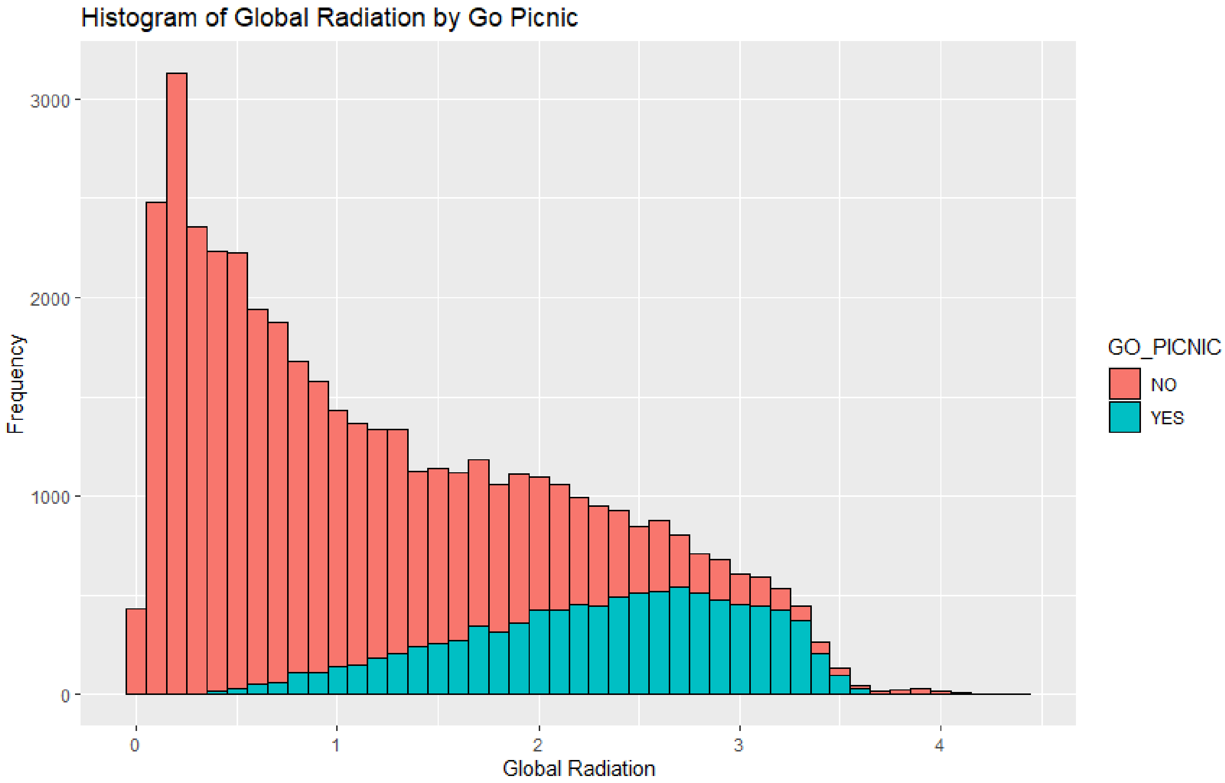


Figure 7. Histogram of Global Radiation by Go Picnic

The global radiation distribution shows that during the survey period, global radiation was generally concentrated around levels below 1, with most days having global radiation between 0 and 1. The number of no picnic days ("NO") was dominant at low radiation levels, especially below 1, peaking at around 0.3 to 0.4. This suggests that when global radiation was low, people tended not to picnic more.

However, at higher radiation levels (1 and above), the number of picnic days ("YES") began to increase. This reflects that when global radiation increased, especially from 1 and above, people tended to picnic more. However, the frequency of picnic days did not completely dominate over no picnic days but increased only in the radiation range between 1 and 3. The number of no-picnic days remained significant even at high radiation levels but began to decrease as radiation increased.

The graph is skewed to the left, with low radiation levels being more common, and this significantly influences the decision not to picnic. In general, when global radiation is low, people tend to avoid outdoor activities such as picnicking. Conversely, when global radiation increases, the frequency of people deciding to picnic also increases, especially at medium to high radiation levels (1 to 3).

5.3.3. Precipitation

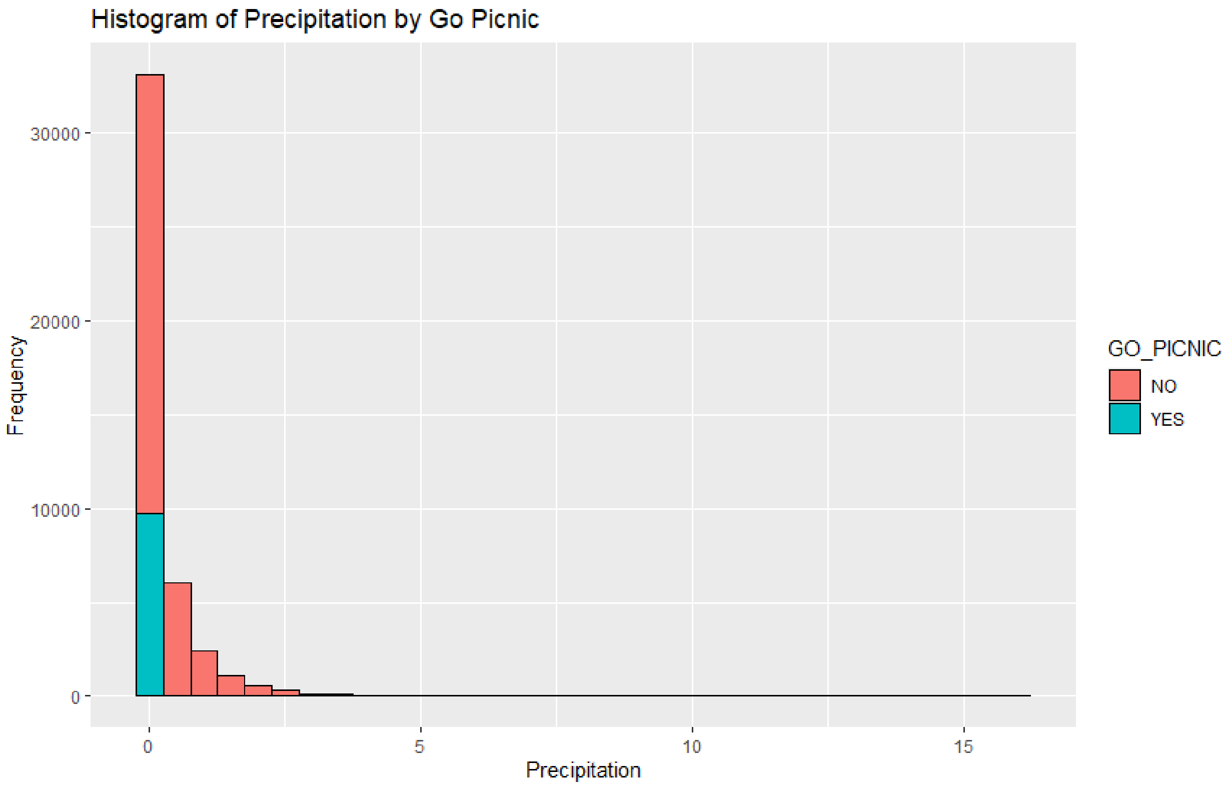


Figure 8. Histogram of Precipitation by Go Picnic

The Precipitation distribution chart shows that during the survey period, rainfall was mainly concentrated around the level below 1 mm. The number of no picnic days ("NO") was dominant at low rainfall levels, especially below 1 mm, peaking at around 0 mm. This suggests that people tended not to picnic more on the majority of the surveyed days.

However, there were still a few days when people decided to picnic when rainfall was between 0 and 2 mm, but this frequency was insignificant compared to the number of no picnic days. And at higher rainfall levels (from 1 mm and above), the number of picnic days ("YES") began to decrease sharply. This reflects that as rainfall increased, especially from 1 mm and above, people tended not to picnic. The number of no picnic days remained significant even at high rainfall levels, but gradually decreased as rainfall increased.

The graph is skewed to the left, with low rainfall being more common, and this significantly influences the decision not to picnic. In general, when rainfall is low, people are still able to go for outdoor activities such as picnics. Conversely, as rainfall increases, the frequency of people deciding to picnic decreases, especially at moderate to high rainfall levels (above 1 mm).

5.3.4. Sunshine

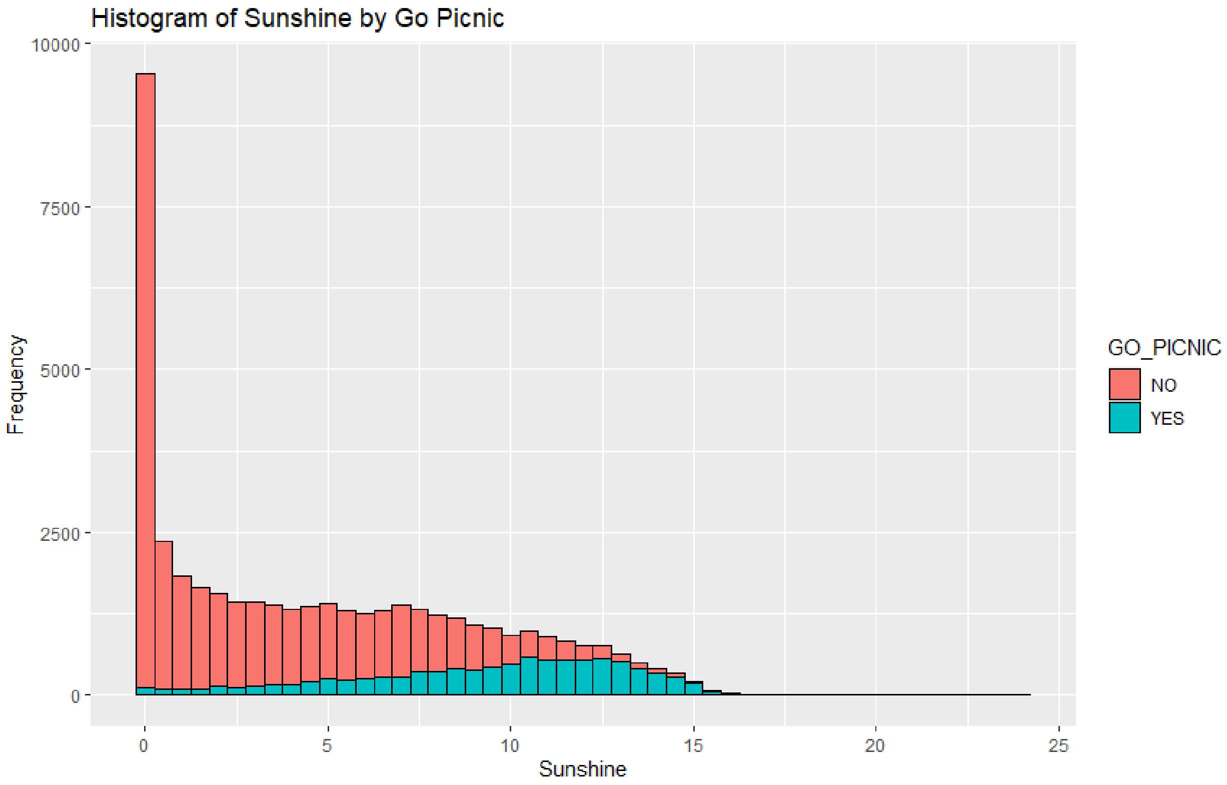


Figure 9. Histogram of Sunshine by Go Picnic

The Sunshine distribution chart shows that during the survey period, the number of no picnic days ("NO") is dominant at low sunshine levels, especially at level 0, which is equivalent to little or no sunshine for 1 hour. This shows that when there is no sunshine or very little sunshine for 1 hour, people tend not to picnic more.

As the sunshine level starts to increase (from 1 to 5), the number of picnic days ("YES") also starts to increase, but the number of no picnic days still dominates. This reflects that even when the sunshine level increases for 1 hour, many people still choose not to picnic.

However, from sunshine levels of about 5 and above, especially from 10 to 15, the number of picnic days increases significantly. This is the period when the sunshine level is strong enough to encourage people to participate in outdoor activities such as picnics. This shows that when the sunshine reaches 10 to 15 degrees per hour, many people feel that the weather conditions are ideal for picnicking. In this range, the number of picnicking days ("YES") is almost equal to or even exceeds the number of non-picnic days ("NO"), indicating the positive impact of high sunshine on the decision to picnic.

The distribution is highly left-skewed, suggesting that sunny days are relatively uncommon during this time period, with most days having little to no sunshine. The graph has a wider distribution at high sunshine levels, showing that when the sunshine is between 10 and 15, not only does the frequency of picnicking increase but it can also exceed the number of non-picnic days. This clearly reflects the trend that when the weather is sunny, especially when there is a significant amount of sunshine between 10 and 15, people tend to picnic more.

5.3.5. Mean Temperature

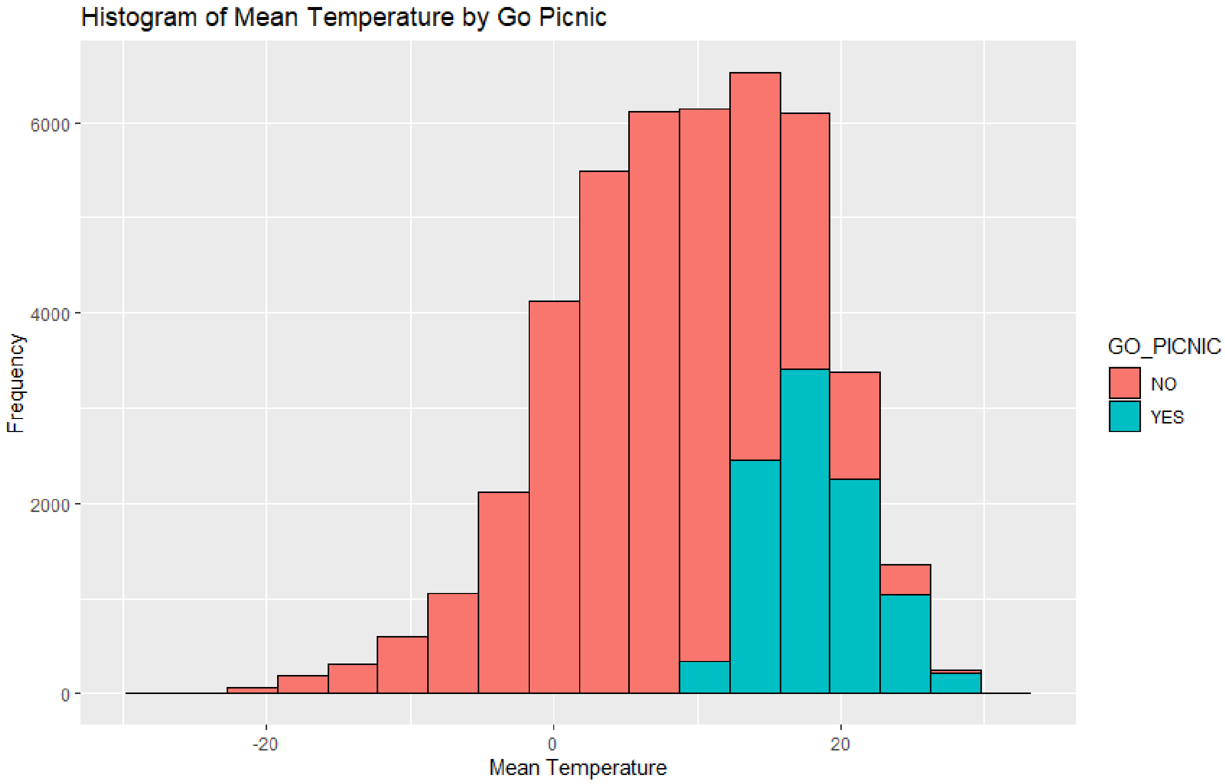


Figure 10. Histogram of Mean Temperature by Go Picnic

The Mean Temperature distribution chart shows the relationship between temperature and people's decision to picnic.

At low temperatures, especially below 0 degrees Celsius, the number of no picnic days ("NO") is almost absolute. This is understandable because when the temperature is too cold, people tend to avoid outdoor activities such as picnics. As the temperature starts to warm up, from about 5 to 15 degrees Celsius, the number of picnic days ("YES") also increases, but the number of no picnic days still accounts for the majority. This shows that when the weather is warmer but still quite cold, many people still choose not to picnic.

However, from about 15 degrees Celsius onwards, the number of picnic days increases sharply, especially at temperatures from 15 to 25 degrees Celsius. In this temperature range, the number of picnic days is almost equal to or even exceeds the number of no picnic days. This reflects those temperatures between 15 and 25 degrees Celsius are ideal for picnics, as the weather is warm and pleasant.

In contrast, at higher temperatures, from around 25 degrees Celsius onwards, the number of picnic days begins to decrease, although they are still present but not as dominant as before. This may be due to the fact that the temperature is too high, making people feel uncomfortable when participating in outdoor activities.

In summary, the chart shows a clear trend: when the average temperature increases to a pleasantly warm level (around 15-25 degrees Celsius), the frequency of picnics increases. Conversely, when the temperature is too low or too high, people tend to limit outdoor activities, and the number of days without picnics becomes dominant.

5.3.6. Max Temperature

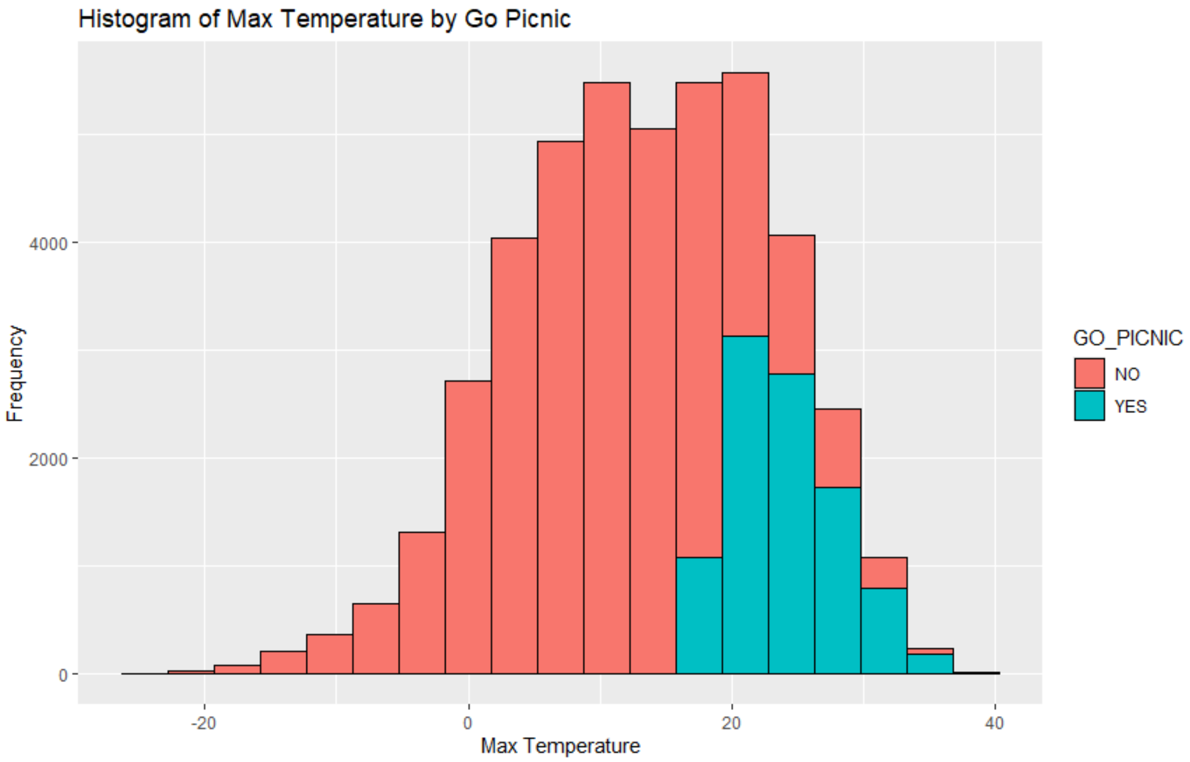


Figure 11. Histogram of Max Temperature by Go Picnic

The Max Temperature frequency chart by people's picnic decision shows the relationship between the highest temperature of the day and the decision to picnic ("YES" or "NO").

At low temperatures, especially below 0 degrees Celsius, the number of days without picnics ("NO") is absolutely dominant. This makes sense because when the temperature is too cold, people tend to avoid outdoor activities such as picnics due to harsh weather conditions. The number of days without picnics is still the majority, but there have started to be some days recorded as picnics. This shows that although some people may enjoy outdoor activities at lower temperatures, the majority still choose not to picnic when the temperature is not warm enough.

From about 10 to 20 degrees Celsius, the number of days without picnics is still the majority, but there have started to be some days recorded as picnics. This shows that although some people may enjoy outdoor activities at lower temperatures, the majority still choose not to picnic when the temperature is not warm enough.

When the temperature rises above 20 degrees Celsius, especially in the range of 20 to 30 degrees Celsius, the number of picnic days continues to remain high, sometimes even close to or more than the number of non-picnic days. This shows that temperatures between 20 and 30 degrees Celsius are considered ideal for outdoor activities such as picnics, people tend to enjoy picnics when the weather is warm but not hot.

In summary, the chart shows a clear trend: when the maximum temperature rises to a comfortable warm level (around 20-30 degrees Celsius), the frequency of picnics also increases. Conversely, when the temperature is too low or too high, people tend to avoid outdoor activities, and the number of non-picnic days is dominant.

5.3.7. Relationship between weather variables

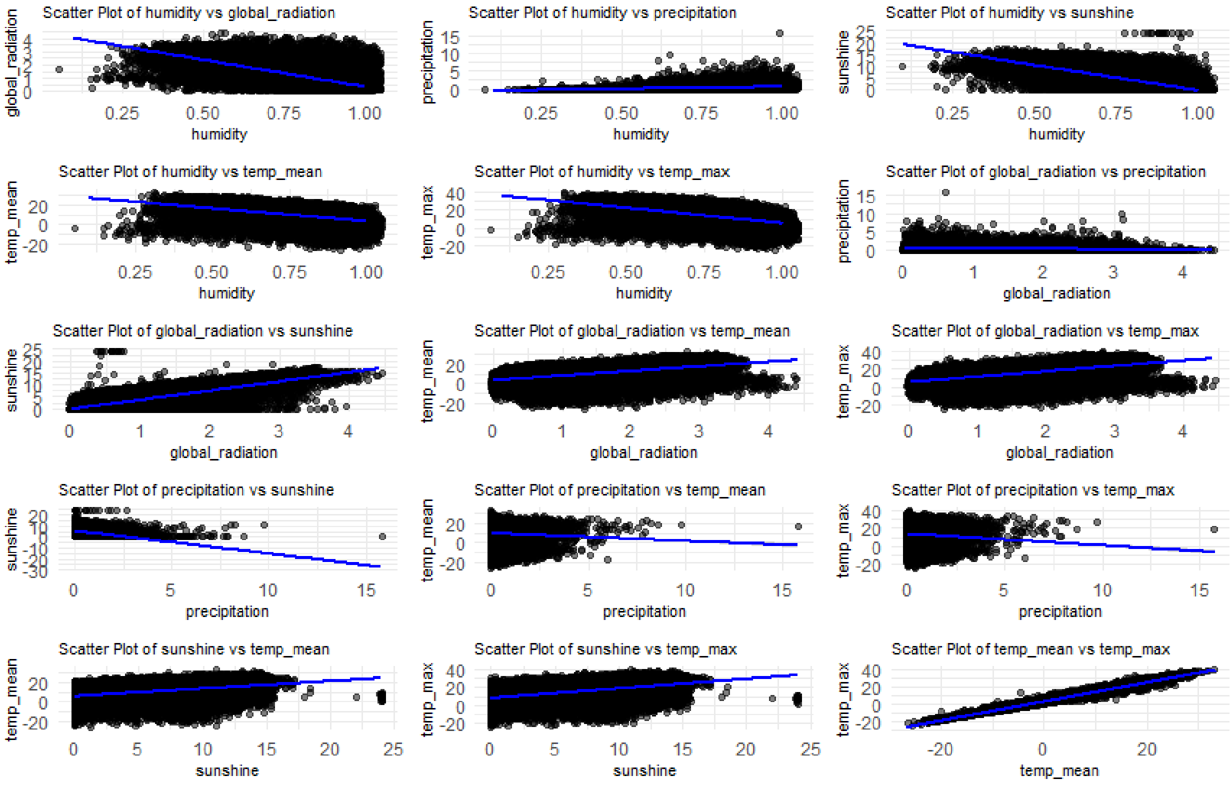


Figure 12. Scatter plots between variables

The image shows a series of scatter plots analyzing the relationships between various weather-related variables: humidity, global radiation, precipitation, sunshine, temp_mean (mean temperature), and temp_max (maximum temperature). Each plot includes a regression line to highlight the trend between the variables.

Humidity generally shows a negative correlation with other variables like global radiation, sunshine, temp_mean, and temp_max. This suggests that higher humidity is associated with lower levels of these factors.

Global radiation has a positive correlation with sunshine and temperature-related variables (temp_mean, temp_max), indicating that as global radiation increases, so do sunshine and temperature.

Precipitation tends to have a negative relationship with sunshine and temperature, suggesting that higher precipitation is associated with lower sunshine and cooler temperatures.

The scatter plot in the last position shows the relationship between temp_mean (mean temperature) and temp_max (maximum temperature). There is a strong positive linear relationship, as indicated by the tight clustering of points along the regression line. Given the strong linear correlation between temp_mean and temp_max, removing either of these variables during model building becomes complicated, as they both provide similar information about temperature. Keeping both variables in the model could result in a situation where the explanatory variables are too similar, reducing efficiency and making it difficult to determine the true influence of each variable. However, removing one variable is not as straightforward, as each variable can be important in its own way—temp_mean provides a snapshot of average temperature, while temp_max reflects maximum temperatures, which are important in assessing weather extremes. Therefore, rather than removing one variable outright, running two separate models, one using temp_mean and one using temp_max, would allow for a more accurate assessment of the contribution of each variable. Comparing the performance of these two models through metrics such as accuracy and coefficient of determination will provide a solid basis for deciding which variable is better suited for a particular model objective, or determining whether both variables can be substituted for each other without losing predictive value.

PHASE 6: MACHINE LEARNING MODEL SPECIFICATION

6.1 MODEL NAME

Weather forecast for outdoor activities

6.2 MODEL SPECIFICATION

Input: The weather prediction dataset has been processed. This dataset includes 12 cities in Europe: BASEL (Switzerland), BUDAPEST (Hungary), DE BILT (the Netherlands), DRESDEN (Germany), DUSSELDORF (Germany), HEATHROW (UK), KASSEL (Germany), LJUBLJANA (Slovenia), MAASTRICHT (the Netherlands), MUENCHEN (Germany), OSLO (Norway), and SONNBLICK (Austria). It contains 43,848 observations from these 12 cities with variables such as DATE, MONTH, SEASON, humidity, global_radiation, precipitation, sunshine, temp_mean, temp_max, and GO_PICNIC. The timeframe is from the year 2000 to 2010.

Output: YES or NO Classification (Go or not go on a picnic).

Approach: Decision Tree Classification.

6.3 DATASET USED FOR MODEL

Data preprocessing is essential before taking the next steps in the model building process.

Table 8. Datasets used for modeling

Field name	**Description**	**Data Type**	**Range of data**
DATE	Date of observation	Numeric	20000101 – 20100101
MONTH	Month of observation	Numeric	1 – 12
SEASON	Season of observation	Character	Spring, Summer, Autumn, Winter
humidity	Humidity	Numeric	0.1 – 1
global_radiation	Global radiation	Numeric	0.01 – 4.42
precipitation	Daily precipitation	Numeric	0 – 15.8
sunshine	Sunshine hours	Numeric	0 – 24
temp_mean	Mean daily temperature	Numeric	-26.6 – 33.1
temp_max	Max daily temperature	Numeric	-24.7 – 40.1

CITY	City of observation	Character	BASEL, BUDAPEST, DE_BILT, DRESDEN, DUSSELDORF, HEATHROW, KASSEL, LJUBLIANA, MAASTRICHT, MUENCHEN, OSLO, SONNBLICK
GO_PICNIC	Decide to go on a picnic	Character	YES, NO

6.4 PROCESS AND CODE ANALYSIS

6.4.1. Decision Tree

A decision tree classifier is similar to flowcharts. It is a map of the possible outcomes of related choices. A decision tree classifier breaks down a dataset into smaller subsets and resembles the shape of a tree.

Decision tree algorithms use the training data to segment the predictor space into non-overlapping regions, the nodes of the tree. Each node is described by a set of rules which are then used to predict new responses. The predicted value for each node is the most common response in the node (classification), or mean response in the node (regression).

The algorithm splits by recursive partitioning, starting with all the observations in a single node. It splits this node at the best predictor variable and best cut point so that the responses within each subtree are as homogenous as possible, and repeats the splitting process for each of the child nodes until a stopping criterion is satisfied.

This much results in a large tree that provides a good fit to the training data, but it likely overfits the data. The solution is to “prune” leaves from the tree. The most common pruning method is cost-complexity pruning. Cost-complexity pruning minimizes the cost complexity:

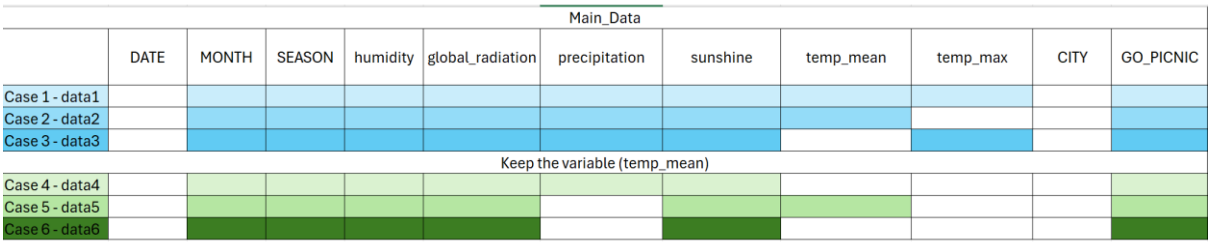
$$CC(T) = R(T) + cp|T|$$

Where $|T|$ is the size of tree (complexity), $R(T)$ is the misclassification rate (decision trees) or RSS (regression trees), and $cp|T|$ is the complexity parameter.

It is expensive to evaluate the error on all possible subtrees, so instead the algorithm defines a sequence of nested trees by successively pruning leaves from the tree, repeating until only the root node remains. The complexity parameter yielding the lowest cost complexity is the optimal tree size

6.4.2. Process and code analysis

The analysis process is divided into 6 cases as follows:



Main_Data

	DATE	MONTH	SEASON	humidity	global_radiation	precipitation	sunshine	temp_mean	temp_max	CITY	GO_PICNIC
Case 1 - data1		🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦		🟦
Case 2 - data2		🟦	🟦	🟦	🟦	🟦	🟦	🟦			🟦
Case 3 - data3		🟦	🟦	🟦	🟦	🟦	🟦		🟦		🟦
Keep the variable (temp_mean)											
Case 4 - data4		🟩	🟩	🟩	🟩	🟩	🟩				🟩
Case 5 - data5		🟩	🟩	🟩	🟩		🟩	🟩			🟩
Case 6 - data6		🟩	🟩	🟩	🟩		🟩				🟩

Step 1: Set working environment and import data

```
>library(readxl)

>Main_Data <- read_excel("D:/Báo cáo dự án R/dataset/Main Data.xlsx")

>View(Main_Data)
```

Step 2: Install decision tree related library packages

```
> install.packages("rpart")

> install.packages("rpart.plot")

> install.packages("caret")

> library(rpart)

> library(rpart.plot)

> library(caret)
```

Step 3: Divide data into training and testing group and run the demo test

Using Rpart library for build Decision tree model.

Case 1: Building a Decision Tree model with columns in Main_Data

Remove columns: DATE, CITY

Build a model with the following columns:

Case 1- data1								
MONTH	SEASON	humidity	global_radiation	precipitation	sunshine	temp_mean	temp_max	GO_PICNIC

```
# Remove DATE, CITY columns

data1<- Main_Data[, !(names(Main_Data) %in% c("DATE", "CITY"))]

#Create Decision Tree model

> indicies<-sample(1:nrow(data1), 0.8*nrow(data1))

> training.data1<-data1[indicies, ]

> test.data1<-data1[-indicies, ]

> decisiontree_fit1 <- rpart(GO_PICNIC ~ ., data = training.data1, method = 'class')

# Plot a Decision Tree
```

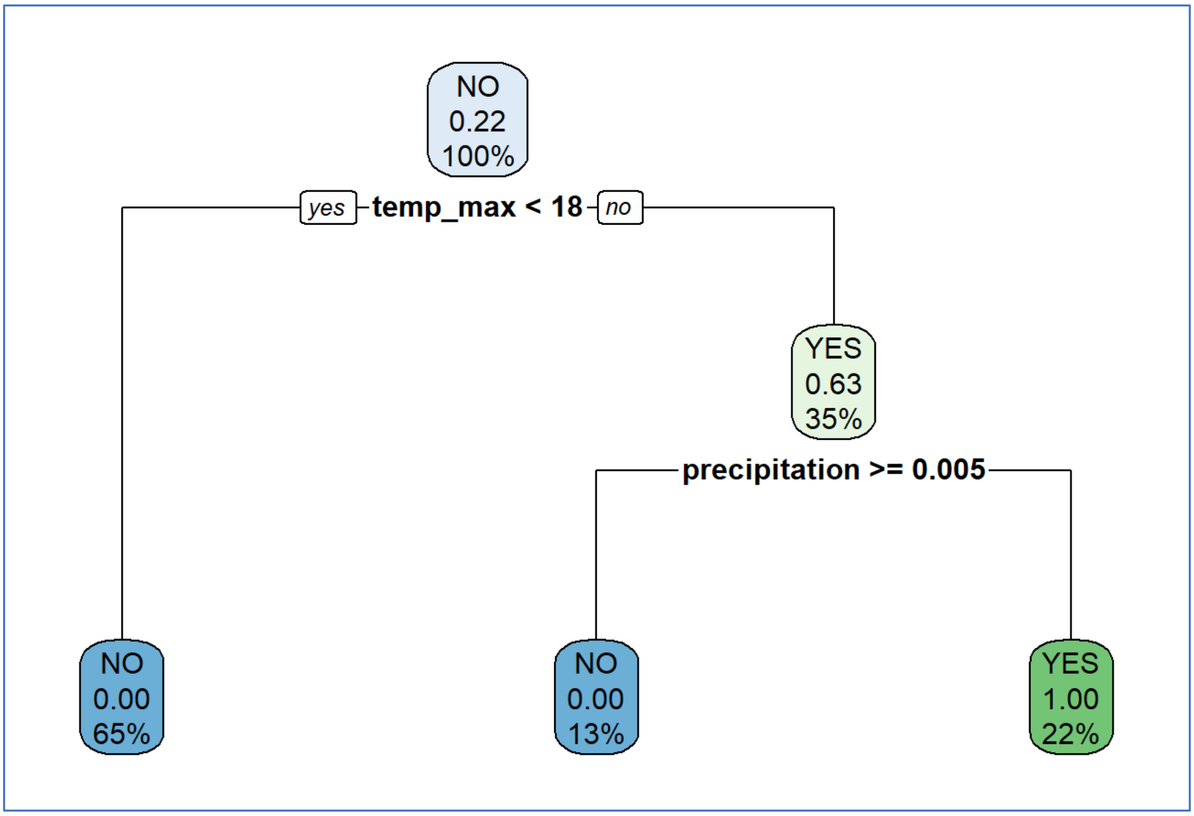


Figure 13. Decision Tree when drawing columns of Case 1 - data1

Based on Figure 6.1, it can be seen that the data is split based on maximum temperature (temp_max). When the maximum temperature is < 18°C, the likelihood of deciding not to go for a picnic is 65%. This could indicate that the data observed in the 12 cities (BASEL, BUDAPEST, DE BILT, DRESDEN, DUSSELDORF, HEATHROW, KASSEL, LJUBLJANA, MAASTRICHT, MUENCHEN, OSLO, SONNBLICK) mainly corresponds to regions with temperate or continental climates, where summers are usually warm. People in these areas may be accustomed to warmer temperatures in summer and feel uncomfortable in cooler weather. When the maximum temperature is < 18°C, it may not be ideal for outdoor activities such as picnics or BBQs.

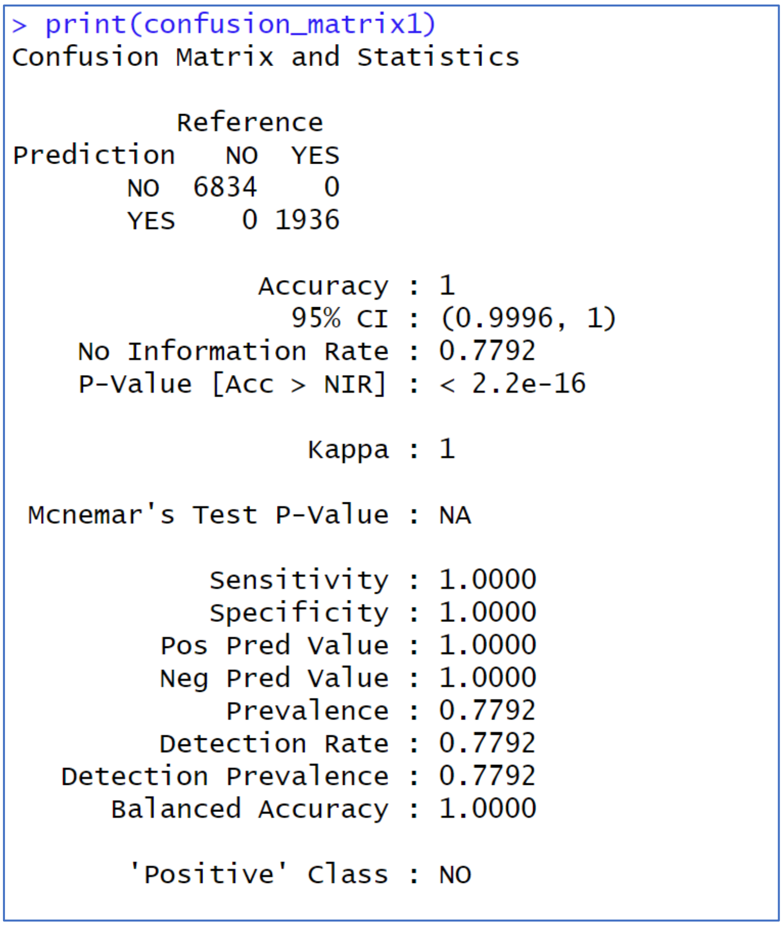
- If the maximum temperature is ≥ 18°C, the next factor considered is precipitation. If precipitation is ≥ 0.005, the likelihood of a "NO" decision is 13%. This indicates that rain may deter people from participating in outdoor activities. Rain can significantly reduce comfort when organizing events like picnics or BBQs.

Even if the temperature is warm enough, having to endure rain may lead people to stay home or choose indoor activities instead.

- If the maximum temperature is ≥ 18°C and precipitation is < 0.005, the likelihood of a "YES" decision is 22%. This may suggest that, although the weather appears ideal with warm temperatures and little rain, only a small proportion (22%) of people are willing to go for a picnic. This could be due to other factors (humidity, global radiation, sunshine) that are not reflected in this decision tree. Failing to consider these factors could lead to an overfitting model that overlooks other weather-related elements.

Step 4: Create Confusion Matrix

```
# Make predictions on the test dataset
> predictions1 <- predict(decisiontree_fit1, test.data1, type = 'class')
# Convert to factor
> predictions1 <- as.factor(predictions1)
> test.data1$GO_PICNIC <- as.factor(test.data1$GO_PICNIC)
# Synchronize levels
> levels(predictions1) <- levels(test.data1$GO_PICNIC)
# Create Confusion Matrix
> confusion_matrix1 <- confusionMatrix(predictions1, test.data1$GO_PICNIC)
# Print Confusion Matrix
> print(confusion_matrix1)
```



```
> print(confusion_matrix1)
Confusion Matrix and Statistics

          Reference
Prediction   NO  YES
       NO  6834    0
       YES    0 1936

               Accuracy : 1
                 95% CI : (0.9996, 1)
    No Information Rate : 0.7792
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 1

 Mcnemar's Test P-Value : NA

            Sensitivity : 1.0000
            Specificity : 1.0000
         Pos Pred Value : 1.0000
         Neg Pred Value : 1.0000
             Prevalence : 0.7792
         Detection Rate : 0.7792
   Detection Prevalence : 0.7792
      Balanced Accuracy : 1.0000

       'Positive' Class : NO
```

Figure 14. Confusion Matrix of Case 1 - data1

- Accuracy: 1.0 (100%): This is an absolute accuracy, which may indicate that the model has overfitted, meaning it has memorized the training data and cannot generalize to new data.
- Kappa: 1.0: A Kappa value of 1 also reflects the perfection of the model and reinforces the likelihood that the model has overfitted.
- Other metrics: All metrics (Sensitivity, Specificity, Positive Predictive Value, Negative Predictive Value) are 1.0, indicating there are no prediction errors.

The 100% accuracy suggests that the model appears very strong in prediction; however, using this result to draw conclusions about the impact of weather conditions on the decision to go for a picnic or not is impossible. Since the model relies only on two factors (temp_max and precipitation), it is insufficient to draw a comprehensive conclusion. To make a more reliable conclusion, further consideration and testing with other scenarios are necessary.

Case 2: Building a Decision Tree model without the temp_max column

Remove columns: DATE, temp_max, CITY

Build a model with the following columns:

Case 2 – data2							
MONTH	SEASON	humidity	global_radiation	precipitation	sunshine	temp_mean	GO_PICNIC

```
# Remove DATE, temp_max, CITY columns
> data2<- Main_Data[, !(names(Main_Data) %in% c("DATE", "temp_max",  "CITY"))]
#Create Decision Tree model
> indicies<-sample(1:nrow(data2), 0.8*nrow(data2))
> training.data2<-data2[indicies, ]
> test.data2<-data2[-indicies, ]
> decisiontree_fit2 <- rpart(GO_PICNIC ~ ., data = training.data2, method = 'class')
# Plot a Decision Tree
>rpart.plot(decisiontree_fit2)
```

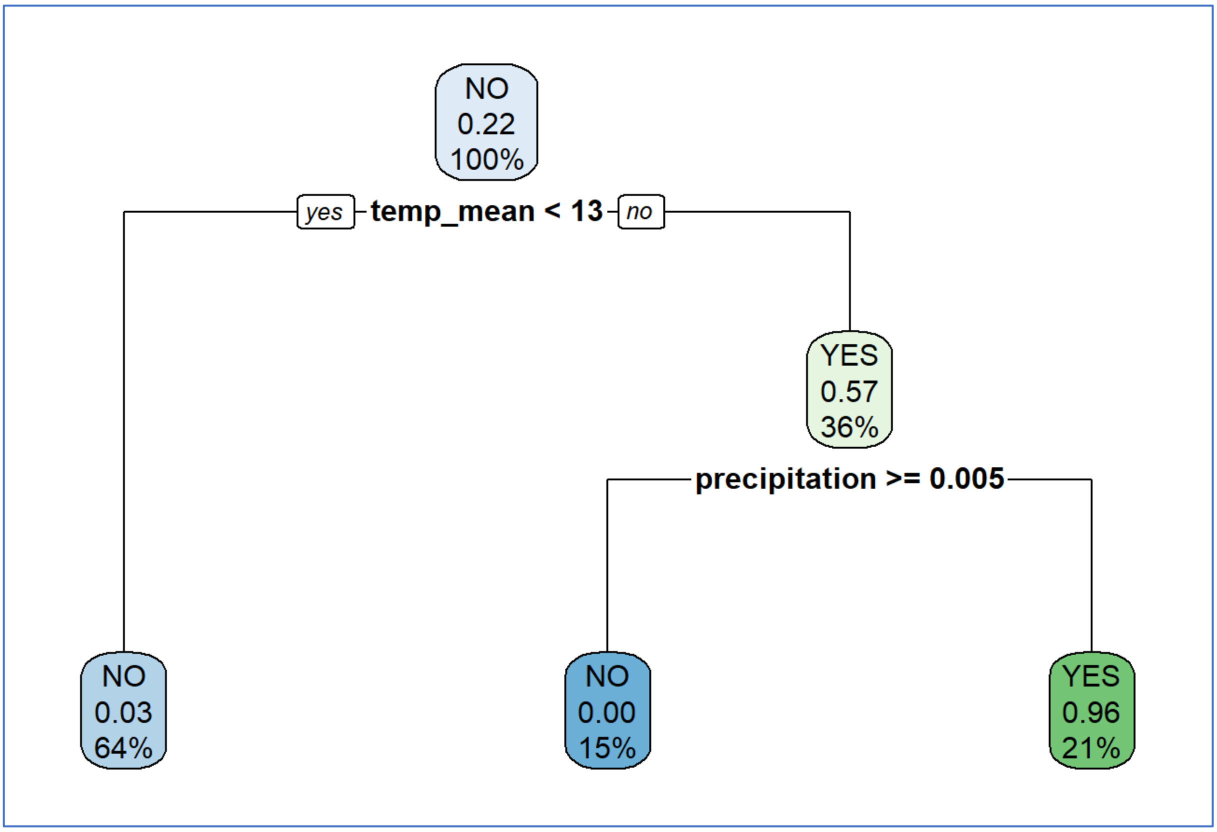


Figure 15. Decision Tree when drawing columns of Case 2 – data2 (without temp_max column)

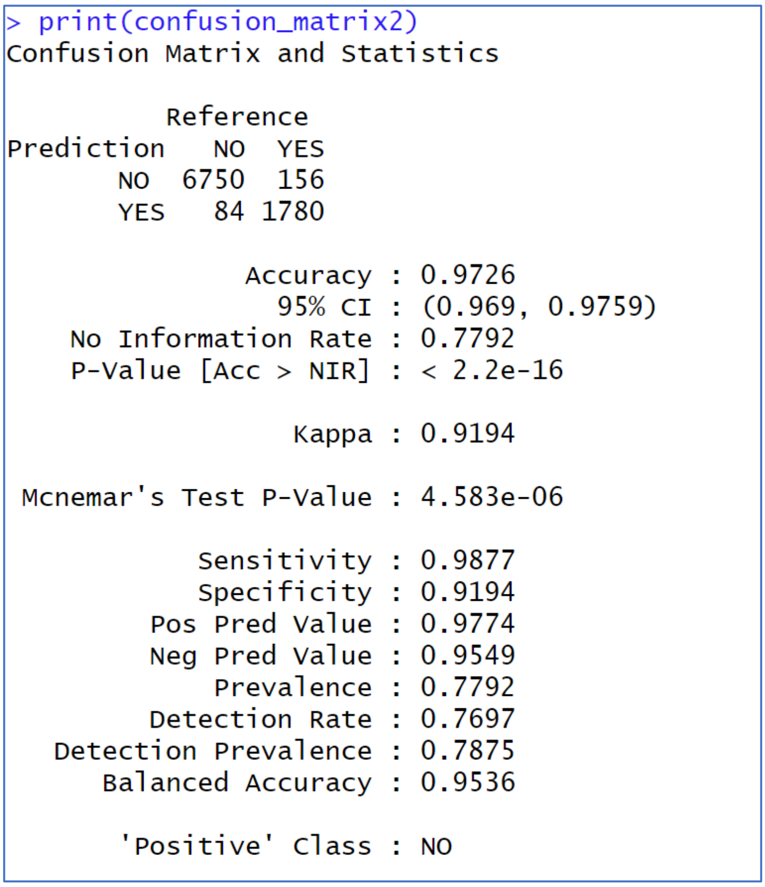
Based on Figure 6.3, when creating a Decision Tree model without the maximum temperature column (temp_max), the data is split based on the mean temperature (temp_mean).

- If the mean temperature is < 13°C, the likelihood of deciding "NO" is 64%. This is similar to the analysis of maximum temperature above case 1, showing that when the weather is cold, people tend to not go for a picnic.
- If the temperature >= 13°C, precipitation is further examined. When precipitation is >= 0.005 the likelihood of deciding "NO" is 15%. This is also similar to the analysis above, where people tend to not go for a picnic when it rains.
- If the temperature >= 13°C and precipitation is < 0.005, the likelihood of deciding "YES" is 21%. This suggests that people tend to go for a picnic when the weather is warm and dry.

Thus, it can be seen that when the maximum temperature column (temp_max) is removed, the analysis is based on the mean temperature (temp_mean), and there is no significant difference when removing the maximum temperature column (temp_max) compared to Case 1.

Step 4: Create Confusion Matrix

```
# Make predictions on the test dataset

> predictions2 <- predict(decisiontree_fit2, test.data2, type = 'class')

# Convert to factor

> predictions2 <- as.factor(predictions2)

> test.data2$GO_PICNIC <- as.factor(test.data2$GO_PICNIC)

# Synchronize levels

> levels(predictions2) <- levels(test.data2$GO_PICNIC)

# Create Confusion Matrix

> confusion_matrix2 <- confusionMatrix(predictions2, test.data2$GO_PICNIC)

# Print Confusion Matrix

> print(confusion_matrix2)
```



```
> print(confusion_matrix2)
Confusion Matrix and Statistics

          Reference
Prediction   NO  YES
       NO  6750  156
       YES   84 1780

               Accuracy : 0.9726
                 95% CI : (0.969, 0.9759)
    No Information Rate : 0.7792
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.9194

 Mcnemar's Test P-Value : 4.583e-06

            Sensitivity : 0.9877
            Specificity : 0.9194
         Pos Pred Value : 0.9774
         Neg Pred Value : 0.9549
             Prevalence : 0.7792
         Detection Rate : 0.7697
   Detection Prevalence : 0.7875
      Balanced Accuracy : 0.9536

       'Positive' Class : NO
```

Figure 16. Confusion Matrix of Case 2 – data2

- Accuracy: 0.9726 (97.26%), which is a very high level of accuracy, though not perfect. It shows that the model has improved compared to the first model.
- Kappa: 0.9194, indicating a very high agreement between the model’s predictions and the actual outcomes, though there is still a slight difference.
- Sensitivity for the NO class: 0.9877, showing that the model is very effective in identifying cases where people do not go for a picnic.
- Specificity for the YES class: 0.9194, indicating that the model is also relatively good at identifying suitable days for a picnic, though some cases are misclassified.
- Precision for NO (Pos Pred Value): 0.9774
- Precision for YES (Neg Pred Value): 0.9549
- Balanced Accuracy: 0.9536

Removing the maximum temperature variable (temp_max) may have led to the loss of some important information, resulting in the model not achieving perfect accuracy. This suggests that (temp_max) could be a crucial factor in the picnic decision. However, the model still performs well, demonstrating that other weather conditions aside from (temp_max) still provide strong enough information to accurately predict this decision.

Case 3: Building a Decision Tree model without the temp_mean column

Remove columns: DATE, temp_mean, CITY

Build a model with the following columns:

Case 3 – data3							
MONTH	SEASON	humidity	global_radiation	precipitation	sunshine	temp_max	GO_PICNIC

```
# Remove DATE, temp_mean, CITY columns
> data3<- Main_Data[, !(names(Main_Data) %in% c("DATE", "temp_mean",  "CITY"))]
#Create Decision Tree model
> indicies<-sample(1:nrow(data3), 0.8*nrow(data3))
> training.data3<-data3[indicies, ]
> test.data3<-data3[-indicies, ]
> decisiontree_fit3 <- rpart(GO_PICNIC ~ ., data = training.data3, method = 'class')
# Plot a Decision Tree
>rpart.plot(decisiontree_fit3)
```

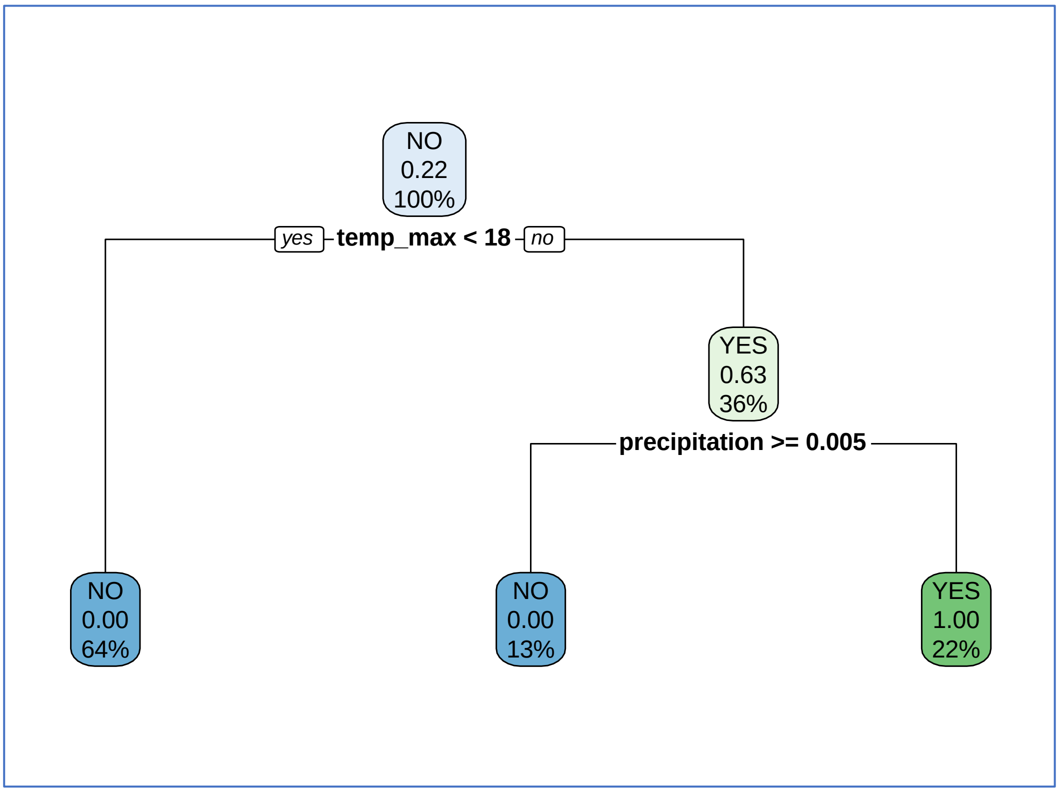
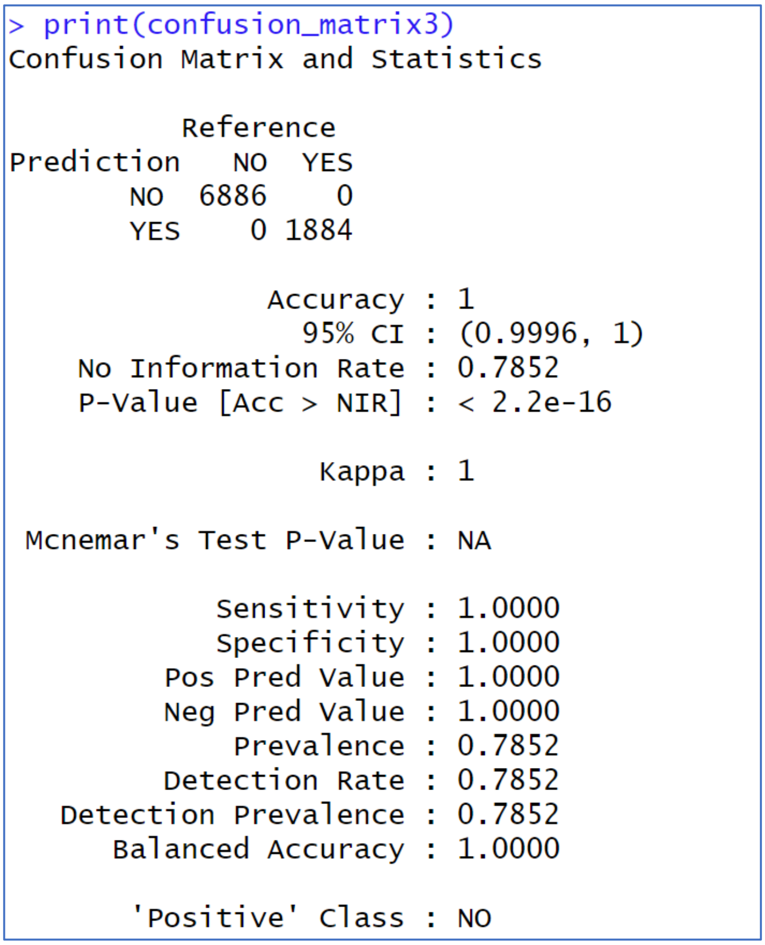


Figure 17. Decision Tree when drawing columns of Case 3 – data3 (without temp_mean column)

Based on Figure 6.6, when creating a Decision Tree model without the mean temperature column (temp_mean), the model created is similar to the case 1 Decision Tree.

Step 4: Create Confusion Matrix

```
# Make predictions on the test dataset
> predictions3 <- predict(decisiontree_fit3, test.data3, type = 'class')
# Convert to factor
> predictions3 <- as.factor(predictions3)
> test.data3$GO_PICNIC <- as.factor(test.data3$GO_PICNIC)
# Synchronize levels
> levels(predictions3) <- levels(test.data3$GO_PICNIC)
# Create Confusion Matrix
> confusion_matrix3 <- confusionMatrix(predictions3, test.data3$GO_PICNIC)
# Print Confusion Matrix
> print(confusion_matrix3)
```



```
> print(confusion_matrix3)
Confusion Matrix and Statistics

          Reference
Prediction   NO  YES
       NO  6886    0
       YES    0 1884

               Accuracy : 1
                 95% CI : (0.9996, 1)
    No Information Rate : 0.7852
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 1

 Mcnemar's Test P-Value : NA

            Sensitivity : 1.0000
            Specificity : 1.0000
         Pos Pred Value : 1.0000
         Neg Pred Value : 1.0000
             Prevalence : 0.7852
         Detection Rate : 0.7852
   Detection Prevalence : 0.7852
      Balanced Accuracy : 1.0000

       'Positive' Class : NO
```

Figure 18. Confusion Matrix of Case 3 – data3

The results from this confusion matrix are similar to those in case 1. The 100% accuracy suggests that the model appears very strong in prediction; however, using this result to draw conclusions about the impact of weather conditions on the decision to go for a picnic or not is impossible. Since the model relies only on two factors (temp_max and precipitation), it is insufficient to draw a comprehensive conclusion. To make a more reliable conclusion, further consideration and testing with other scenarios are necessary.

From cases 1, 2 and 3:

We can keep the variable “temp_mean” for the following reasons:

- Practical nature of average temperature (temp_mean): Average temperature provides general information about weather conditions throughout the day. People often care about average temperature to make decisions about whether to participate in outdoor activities or not. Keeping the variable (temp_mean) in the model helps ensure that the model is using a familiar and highly applicable weather index, easy to verify in daily life and make reasonable decisions.
- Helps the model to be more general: Using (temp_max) can lead to overfitting, making the model unable to draw general conclusions. Keeping the variable (temp_mean) helps the model to be more general, avoiding overfitting.

After keeping the temp_mean variable unchanged, we can continue to look at the following cases to see how the remaining variables influence the decision to participate in outdoor activities.

Case 4: Continue building the Decision Tree model without the temp_mean column.

Remove columns: DATE, temp_mean, CITY

Build a model with the following columns:

Case 4 – data4						
MONTH	SEASON	humidity	global_radiation	precipitation	sunshine	GO_PICNIC

```
# Remove DATE, temp_mean, temp_max, CITY columns
> data4<- Main_Data[, !(names(Main_Data) %in% c("DATE", "temp_mean", "temp_max", "CITY"))]
#Create Decision Tree model
> indicies<-sample(1:nrow(data4), 0.8*nrow(data4))
> training.data4<-data4[indicies, ]
> test.data4<-data4[-indicies, ]
> decisiontree_fit4 <- rpart(GO_PICNIC ~ ., data = training.data4, method = 'class')
# Plot a Decision Tree
>rpart.plot(decisiontree_fit4)
```

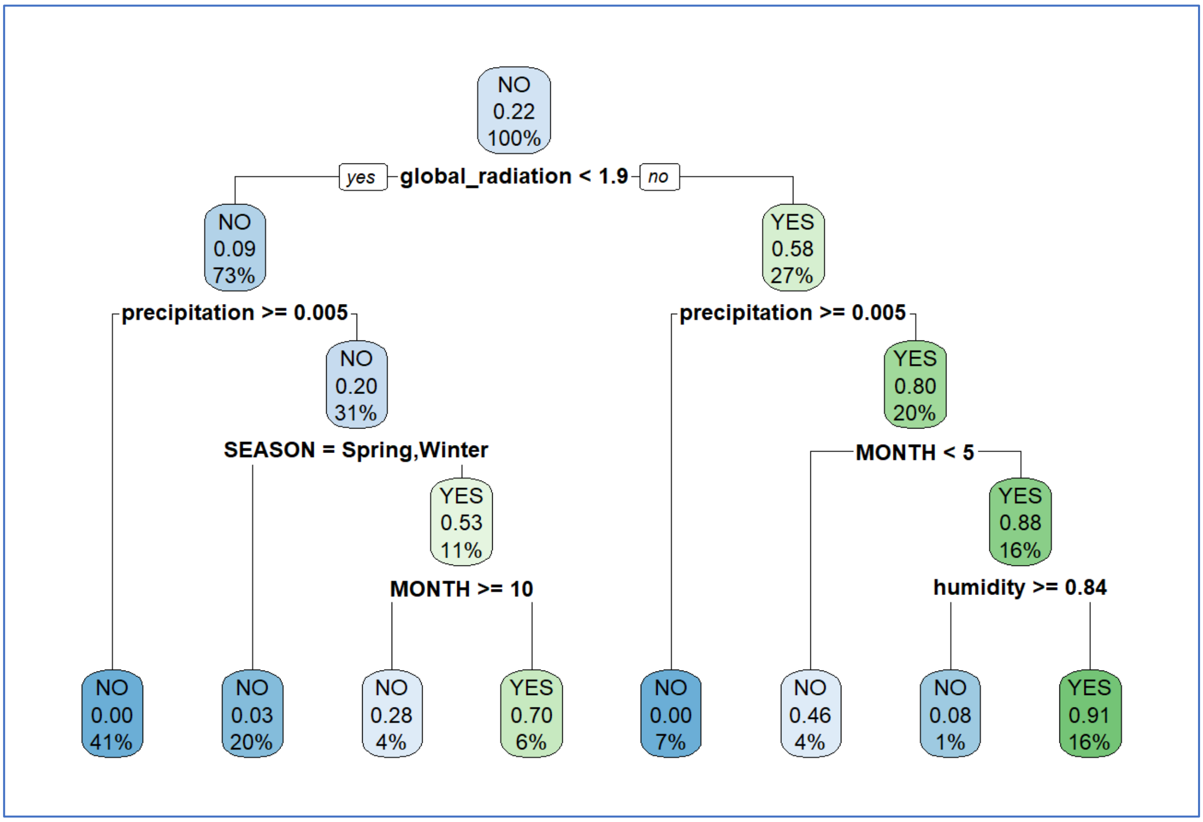


Figure 19. Decision Tree when drawing columns of Case 4 – data4 (without temp_mean column)

Based on Figure 6.8, when the mean temperature column (temp_mean) is also removed, other weather conditions are considered, with global_radiation being the first important factor examined.

On the left branch:

- When global radiation is < 1.9, the likelihood of deciding "NO" is 73%. This suggests that when solar radiation is low, the weather may not be favorable (cloudy, with little sunlight), reducing the likelihood of going for a picnic.
- Precipitation is the second important factor in the decision tree.
  - If precipitation is >= 0.005 the likelihood of deciding "NO" is 41%. This indicates that if the weather is cloudy and rainy, people tend not to go for a picnic.
  - On the other hand, if precipitation is < 0.005, the likelihood of deciding "NO" is 31%. This shows that when it is cloudy or there is little sunlight, but there is no rain or very little rain, the likelihood of going for a picnic decrease.

- The factors of season (SEASON) and month (MONTH) are then considered in deciding whether to go for a picnic.
  - If it is Spring or Winter, the likelihood of deciding "NO" is 20%.
  - If it is not Spring or Winter, the likelihood of deciding "YES" is 11%, and the month (MONTH) is further considered. From October onwards, the likelihood of deciding "NO" is 4%. Conversely, before October, the probability of deciding "YES" is 6%.

In summary, from the left branch, it can be seen that when when global radiaion is low, with little sunlight, and rainy, people tend not to go for a picnic. When it is not raining or only lightly raining, it is not Spring or Winter and falls before October, the likelihood of participating in outdoor activities increases.

On the right branch:

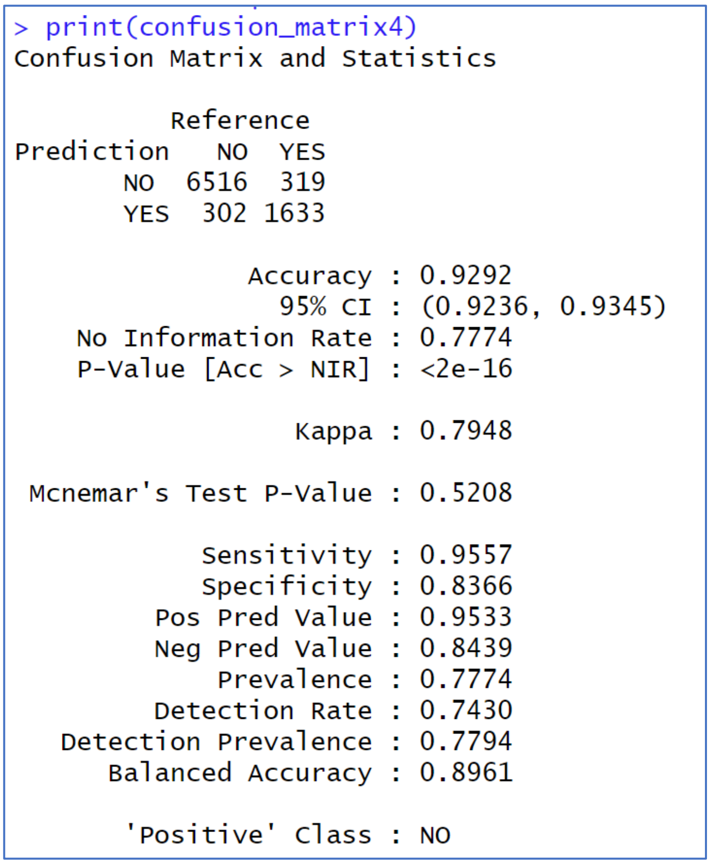
- If global radiation is >= 1.9, the likelihood of deciding “YES” is 27%. This indicates that when solar radiation is high, there is plenty of sunlight, the weather becomes more favorable for going on a picnic.
- Precipitation is the second important factor in the decision tree.
  - If precipitation is >= 0.005, the likelihood of deciding “NO” is 7%. This shows that even though there is sunlight, people tend not to go for a picnic when it is raining.
  - On the other hand, if precipitation is < 0.005, the likelihood of deciding “YES” is 20%. This suggests that when there is sunlight and little or no rain, the weather becomes favorable for outdoor activities.
- The third factor considered is the month (MONTH).
  - If MONTH is < 5, the likelihood of deciding “NO” is 4%, and if MONTH is greater than or equal to 5, the likelihood of deciding “YES” is 16%. Before May is late Winter or Spring, which suggests that despite high solar radiation and little rain, the weather can still be cold, hindering outdoor activities. From May onwards, which is Summer or Autumn, the weather may become warmer and more favorable for outdoor activities like picnics or BBQs.
- The fourth factor considered is humidity.
  - If MONTH is >= 5 and humidity is >= 0.84, the likelihood of deciding “NO” is 1%. This indicates that when radiation is high, there is little or no rain, and the weather from May onwards is warmer, high humidity does not significantly impact the decision to say "NO," and people tend to still participate in outdoor activities even with high humidity.
  - If MONTH is >= 5 and humidity is < 0.84, the likelihood of deciding “YES” is 16%. This shows that lower humidity can increase comfort, and

along with high radiation, little or no rain from May onwards, the weather conditions become more ideal for people to engage in outdoor activities.

In summary, from the right branch, it can be seen that when solar radiation is high, the weather becomes more favorable for outdoor activities, especially if precipitation is low or nonexistent. The weather from May onwards (Summer and Autumn) also provides better conditions for picnics due to warmer temperatures. Although humidity has an impact, when the weather is already warm and has little rain, high humidity does not significantly reduce the likelihood of deciding to participate in outdoor activities.

Step 4: Create Confusion Matrix

```
# Make predictions on the test dataset

> predictions4 <- predict(decisiontree_fit4, test.data4, type = 'class')

# Convert to factor

> predictions4 <- as.factor(predictions4)

> test.data4$GO_PICNIC <- as.factor(test.data4$GO_PICNIC)

# Synchronize levels

> levels(predictions4) <- levels(test.data4$GO_PICNIC)

# Create Confusion Matrix

> confusion_matrix4 <- confusionMatrix(predictions4, test.data4$GO_PICNIC)

# Print Confusion Matrix

> print(confusion_matrix4)
```



```
> print(confusion_matrix4)
Confusion Matrix and Statistics

          Reference
Prediction   NO  YES
       NO  6516  319
       YES  302 1633

               Accuracy : 0.9292
                 95% CI : (0.9236, 0.9345)
    No Information Rate : 0.7774
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.7948

 Mcnemar's Test P-Value : 0.5208

            Sensitivity : 0.9557
            Specificity : 0.8366
         Pos Pred Value : 0.9533
         Neg Pred Value : 0.8439
             Prevalence : 0.7774
         Detection Rate : 0.7430
   Detection Prevalence : 0.7794
      Balanced Accuracy : 0.8961

       'Positive' Class : NO
```

Figure 20. Confusion Matrix of Case 4 – data4

- Accuracy: 0.9292 (92.92%) the model achieves lower accuracy than the model in case 2. This indicates that the model's accuracy has decreased when the mean temperature variable (temp_mean) was removed.
- Kappa: 0.7948, showing that the level of agreement between predictions and reality has decreased compared to case model 2.
- Sensitivity for the NO class: 0.9557 (95.57%), indicating that the model is still very effective in identifying cases where people do not go for a picnic.
- Specificity for the YES class: 0.8366 (83.66%) a significant decrease from 91.94% in case 2. This suggests that the model has lost some of its ability to accurately identify days suitable for a picnic. Specifically, there were many incorrect predictions, with 319 cases predicted as "NO" when in reality, they were "YES," and 302 cases predicted as "YES" when in reality, they were "NO."
- Precision for NO (Pos Pred Value): 0.9533
- Precision for YES (Neg Pred Value): 0.8439
- Balanced Accuracy: 0.8961

The decline in accuracy and Kappa indicates that the variables (temp_max) and (temp_mean) play important roles in the predictive model. Removing both "temp_max" and "temp_mean" has significantly reduced the model's performance, suggesting that temperature is a critical factor in the decision to go or not go for a picnic.

Case 5: Continue building the Decision Tree model without the precipitation column and keep temp_mean column.

Remove columns: DATE, precipitation, CITY

Build a model with the following columns:

Case 5 – data5						
MONTH	SEASON	humidity	global_radiation	sunshine	temp_mean	GO_PICNIC

```
# Remove DATE, precipitation, CITY columns
> data5<- Main_Data[, !(names(Main_Data) %in% c("DATE", "precipitation", "CITY"))]
#Create Decision Tree model
> indicies<-sample(1:nrow(data5), 0.8*nrow(data5))
> training.data5<-data5[indicies, ]
> test.data5<-data5[-indicies, ]
> decisiontree_fit5 <- rpart(GO_PICNIC ~ ., data = training.data5, method = 'class')
# Plot a Decision Tree
>rpart.plot(decisiontree_fit5)
```

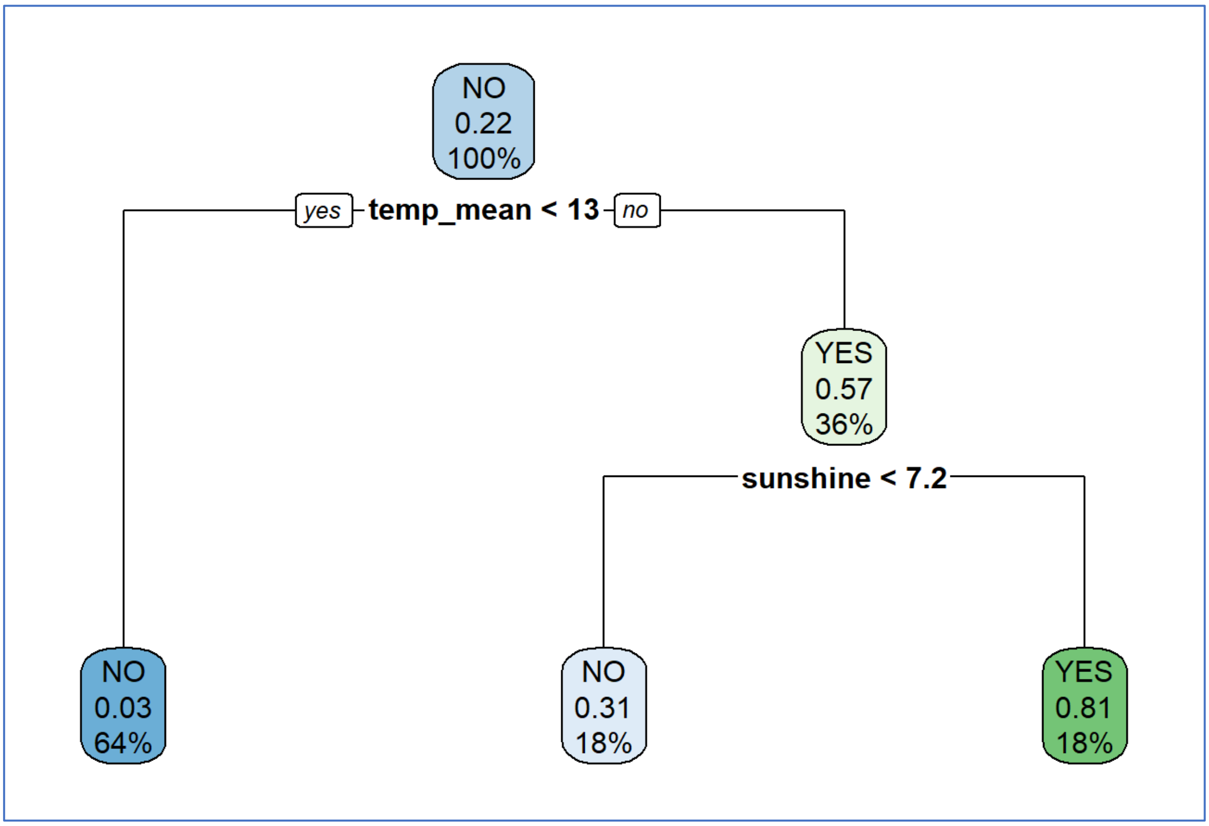


Figure 21. Decision Tree when drawing columns of Case 5 – data5 (without precipitation column)

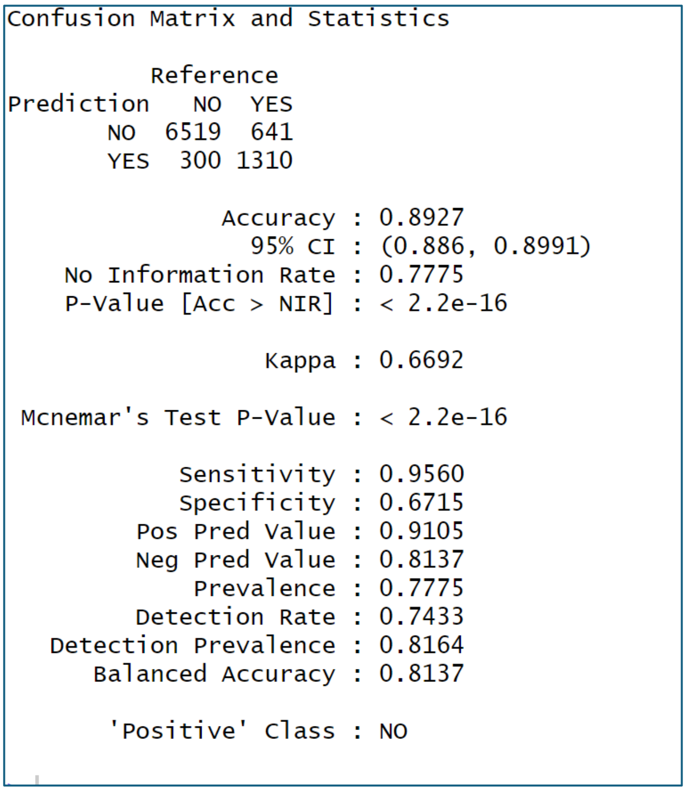
Based on Figure 6.10, when the precipitation column is removed and temp_mean column is retained, the data is first split based on the mean temperature.

- If the mean temperature < 13°C, the likelihood of deciding "NO" is 64%. This indicates, similar to the case 2 above, that when the weather is cold, people tend not to go for a picnic.
- If the mean temperature >= 13°C the likelihood of deciding "YES" is 64%, with further consideration of the sunshine factor. When sunshine is < 7.2, the likelihood of deciding "NO" is 18%. This suggests that although the temperature may be higher, if there is not enough sunlight, the weather is still not ideal. Conversely, when sunshine is >= 7.2 the likelihood of deciding "NO" is 18%, indicating that when there is plenty of sunlight along with warm temperatures, the conditions are more favorable for outdoor activities.

When the precipitation column is removed, the decision to go for a picnic strongly depends on both the mean temperature and sunshine. Warm temperatures and abundant sunlight are two important factors that create ideal conditions for outdoor activities.

Step 4: Create Confusion Matrix

```
# Make predictions on the test dataset

> predictions5 <- predict(decisiontree_fit5, test.data5, type = 'class')

# Convert to factor

> predictions5 <- as.factor(predictions5)

> test.data5$GO_PICNIC <- as.factor(test.data5$GO_PICNIC)

# Synchronize levels

> levels(predictions5) <- levels(test.data5$GO_PICNIC)

# Create Confusion Matrix

> confusion_matrix5 <- confusionMatrix(predictions5, test.data5$GO_PICNIC)

# Print Confusion Matrix

> print(confusion_matrix5)
```



```
Confusion Matrix and Statistics

          Reference
Prediction   NO  YES
       NO  6519  641
       YES  300 1310

               Accuracy : 0.8927
                 95% CI : (0.886, 0.8991)
    No Information Rate : 0.7775
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.6692

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.9560
            Specificity : 0.6715
         Pos Pred Value : 0.9105
         Neg Pred Value : 0.8137
             Prevalence : 0.7775
         Detection Rate : 0.7433
   Detection Prevalence : 0.8164
      Balanced Accuracy : 0.8137

       'Positive' Class : NO
```

Figure 22. Confusion Matrix of Case 5 – data5

- Accuracy: 0.8927 (89.27%) The model achieves lower accuracy than the case model 4. This indicates that the model’s accuracy decreased after removing the precipitation variable.
- Kappa: 0.6692.
- Sensitivity for the NO class: 0.9560 (95.60%), showing that the model is still very good at identifying cases where people do not go on a picnic.
- Specificity for the YES class: 0.6715 (67.15%), indicating that the model has more difficulty accurately classifying cases where people do go on a picnic. Specifically, there were many incorrect predictions, including 641 cases predicted as "NO" but were actually "YES," and 300 cases predicted as "YES" but were actually "NO."
- Precision for NO (Pos Pred Value): 0.9105.
- Precision for YES (Neg Pred Value): 0.8137.
- Balanced Accuracy: 0.8137.

The decrease in accuracy and Kappa suggests that the precipitation variable plays an important role in the prediction model. Removing precipitation significantly reduced the model's performance, lowering overall accuracy. Precipitation is a crucial factor in the decision to go on a picnic or not.

Case 6: Continue building the Decision Tree model without both precipitation and temp_mean columns.

Remove columns: DATE, precipitation, CITY

Build a model with the following columns:

Case 6 – data6					
MONTH	SEASON	humidity	global_radiation	sunshine	GO_PICNIC

```
# Remove DATE, precipitation, temp_mean, CITY columns
> data6<- Main_Data[, !(names(Main_Data) %in% c("DATE", "precipitation", "temp_mean", "CITY"))]
#Create Decision Tree model
> indicies<-sample(1:nrow(data6), 0.8*nrow(data6))
> training.data6<-data6[indicies, ]
> test.data6<-data6[-indicies, ]
> decisiontree_fit6 <- rpart(GO_PICNIC ~ ., data = training.data6, method = 'class')
# Plot a Decision Tree
>rpart.plot(decisiontree_fit6)
```

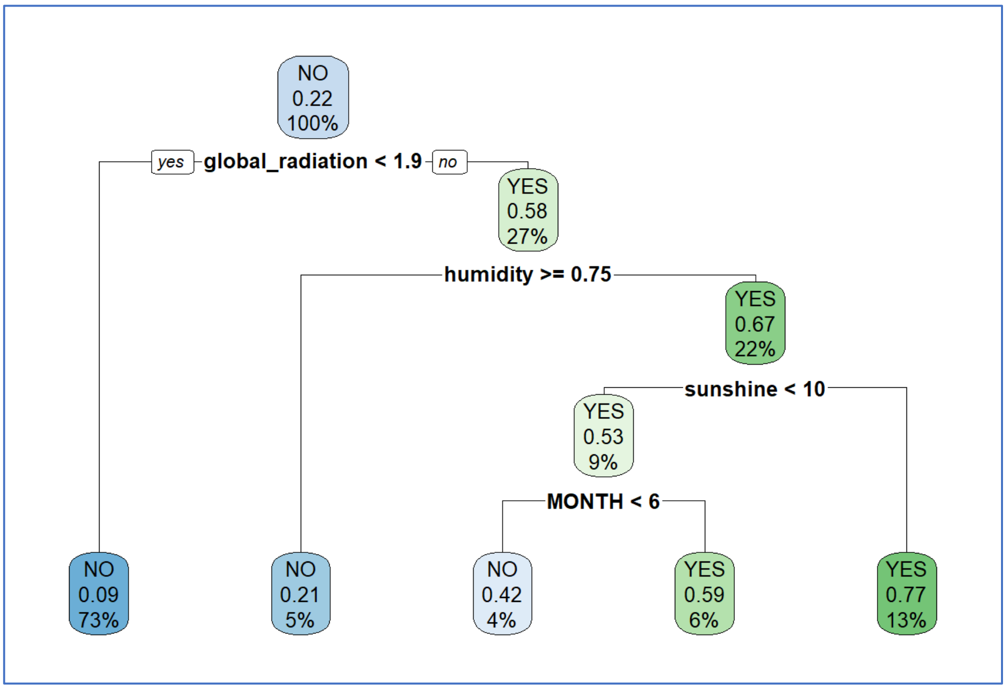


Figure 23. Decision Tree when drawing columns of Case 5 – data5 (without both precipitation and temp_mean columns)

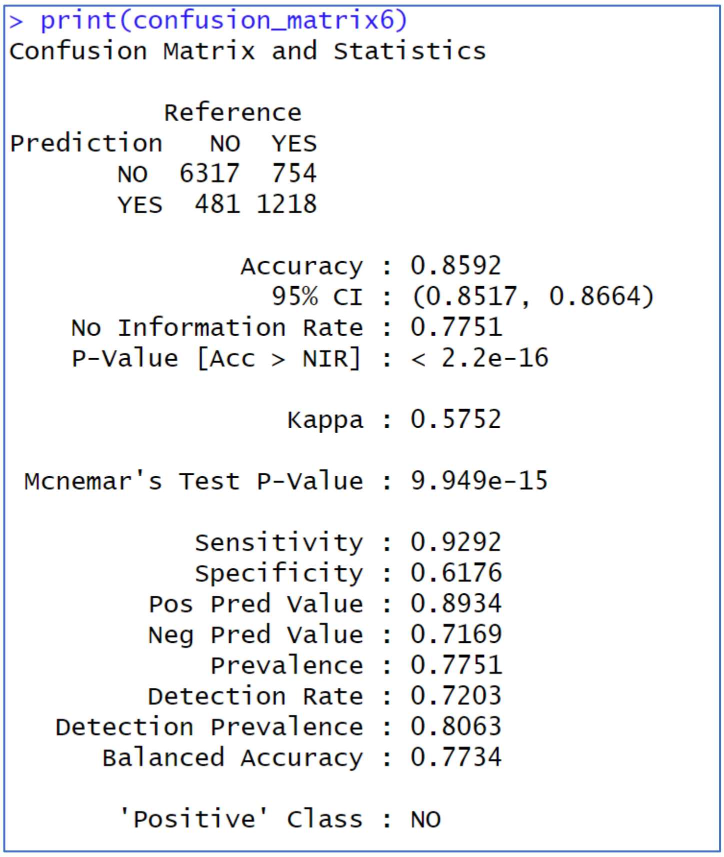
Based on Figure 6.12, after removing the temp_mean and precipitation columns, the data is first split based on global radiation.

- If the global radiation is < 1.9, the likelihood of deciding "NO" is 73%. This suggests that when solar radiation is low, the weather might not be favorable (cloudy skies, low sunlight), reducing the likelihood of going on a picnic.
- If the global radiation is >= 1.9, the likelihood of deciding "YES" is 27%. This indicates that when solar radiation is high, with more sunlight, the weather becomes more favorable for going on a picnic.
- Humidity is the second most important factor in the decision tree:
    - If humidity is >= 0.75, the likelihood of deciding "NO" is 5%. This suggests that when humidity is >= 0.75, people tend to avoid outdoor activities, possibly due to the wet feeling despite high radiation (more sunlight).
    - If humidity is < 0.75, the likelihood of deciding "YES" is 22%. When humidity is < 0.75 and there is plenty of sunlight, the conditions might be more favorable for people to engage in outdoor activities like picnics and BBQs.
- Sunshine is the third most important factor in the decision tree:
    - If sunshine is >= 10, the likelihood of deciding "YES" is 13%. This suggests that people are more likely to participate in outdoor activities when radiation is high, humidity is low, and there is a high amount of sunshine.
    - If sunshine is < 10, the likelihood of deciding "YES" is 9%, with further consideration of another factor MONTH. When sunshine is < 10 and MONTH is < 6, the likelihood of deciding "NO" is 4%. This indicates that in the early months of the year, the weather in the 12 cities might still be cold, with less sunshine, despite high radiation, making people less inclined to go on picnics.
    - If sunshine is < 10 and MONTH is > 6, the likelihood of deciding "YES" is 6%. In the middle months of the year, the weather might become warmer, with high radiation during the summer or autumn, making people more likely to enjoy outdoor activities even with less sunshine.

When both temp_mean and precipitation are removed, the decision to go on a picnic strongly depends on factors like global radiation, humidity, and sunshine. Low humidity and the presence of sunshine increase the likelihood of going on a picnic.

Step 4: Create Confusion Matrix

```
# Make predictions on the test dataset

> predictions6 <- predict(decisiontree_fit6, test.data6, type = 'class')

# Convert to factor

> predictions6 <- as.factor(predictions6)

> test.data6$GO_PICNIC <- as.factor(test.data6$GO_PICNIC)
```



```
> print(confusion_matrix6)
Confusion Matrix and Statistics

          Reference
Prediction   NO  YES
       NO  6317  754
       YES  481 1218
                                          
               Accuracy : 0.8592          
                 95% CI : (0.8517, 0.8664)
    No Information Rate : 0.7751          
    P-Value [Acc > NIR] : < 2.2e-16       
                                          
                  Kappa : 0.5752          
                                          
 Mcnemar's Test P-Value : 9.949e-15       
                                          
            Sensitivity : 0.9292          
            Specificity : 0.6176          
         Pos Pred Value : 0.8934          
         Neg Pred Value : 0.7169          
             Prevalence : 0.7751          
         Detection Rate : 0.7203          
   Detection Prevalence : 0.8063          
      Balanced Accuracy : 0.7734          
                                          
       'Positive' Class : NO              
```

Figure 24. Confusion Matrix of Case 6 – data6

- Accuracy: 0.8592 (85.92%), indicating that the model achieved lower accuracy compared to the case model 5. This suggests that the model's accuracy has decreased when both precipitation and temp_mean variables are removed.
- Kappa: 0.5752.
- Sensitivity for the NO class: 0.9292 (92.92%), showing a decrease in correctly predicting cases where people don't go on a picnic, indicating that the model's

performance in predicting NO cases is affected when temp_mean variable is also removed.

- Specificity for the YES class: 0.6167 (61.67%), also showing a decrease, indicating that the model struggles even more with accurately classifying picnic cases. Specifically, there were many misclassifications, with 754 cases predicted as "NO" but actually "YES," and 481 cases predicted as "YES" but actually "NO."
- Precision for NO (Pos Pred Value): 0.8934.
- Precision for YES (Neg Pred Value): 0.7169.
- Balanced Accuracy: 0.7734.

The decrease in accuracy and Kappa suggests that both temp_mean and precipitation play crucial roles in the predictive model. Removing these two variables has led to a less effective model, reducing its accuracy compared to previous cases. Temp_mean and precipitation are essential factors in determining whether to go on a picnic or not.

PHASE 7: CONCLUSION

The two variables, average temperature and precipitation, are important factors. The accuracy of the model decreases in each case when either of these variables or both are removed, indicating their significance in predicting the likelihood of going on a picnic. Specifically, when the mean temperature (temp_mean) variable is removed, the model's accuracy decreases significantly, which shows that mean temperature has a major impact on this decision. When the precipitation variable is subsequently removed, the model's performance further declines, indicating that precipitation also plays an equally important role in influencing outdoor activity decisions.

The decline in model accuracy when both variables are removed suggests that these factors not only have independent value but also interact with each other in influencing human decisions. Mean temperature affects comfort levels during outdoor activities, while precipitation can disrupt such activities. The combination of these two factors plays a crucial role in determining whether a day is suitable for outdoor activities. When both of these factors are removed, the model no longer has sufficient information to accurately predict, leading to poor performance.

People tend to engage in outdoor activities when the weather meets certain conditions:

- **Ideal Temperature:** An average temperature of ≥ 13°C is ideal for outdoor activities. Warm weather helps people feel comfortable, not too hot and not too cold.
- **No Rain or Light Rain:** Precipitation < 0.005. Dry weather is always ideal for outdoor activities because rain can wet belongings and hinder activities. In some cases, light rain may not significantly affect outdoor activities like picnics or BBQs if it occurs briefly and doesn't reduce the temperature significantly.
- **Ideal Season:** Summer and Fall are the most popular seasons for outdoor activities, even if there is light rain.
- **Low Humidity:** Humidity < 0.84 is ideal for outdoor activities. Low humidity makes the air feel fresher and more pleasant.
- **Global Radiation:** People tend to participate in outdoor activities when global radiation is > 1.9, indicating warmer weather. As the 12 cities are primarily located in temperate or continental climate zones, where summers are usually warm, residents of these areas may be accustomed to warmer temperatures in the summer and may feel uncomfortable in colder weather.
- **Sunshine:** When sunshine is ≥ 10, people tend to engage in outdoor activities as there is more sunlight.

Based on the above analysis, from a business perspective, selecting products for BBQs or picnics requires special attention to weather forecasts before deciding to restock or develop pricing strategies tailored to each season.

Business Strategy:

Business Strategy called “Weather Adaptation Strategy”

Weather Adaptation Strategy:

1. Developing the ability to balance opportunities and risks:

Leveraging weather opportunities to increase revenue: According to the report, people tend to engage in outdoor activities during Summer and Fall, when there is sunshine and warm temperatures.

- Update a diverse range of products to meet weather conditions: Products for picnics and BBQs such as mats, folding tables and chairs, coolers, portable grills, etc.
- Use weather forecasts to enhance advertising campaigns that align with weather conditions, encouraging people to engage in outdoor activities: Provide suggestions for ideal locations for picnic activities and recommend picnic products people should prepare.
- Pricing strategy and promotions: Slightly increase prices during peak seasons, and offer promotional packages with suitable product combos.
- Offer flexible warranty and return policies if the weather changes suddenly.

Managing business risks when the weather is unfavorable: People tend to avoid outdoor activities during Winter and Spring, when it rains, temperatures are low, and there is little sunlight.

- Limit off-season product imports.
- Boost efforts to clear out inventory: Apply the “Just in Time” technique, ensuring the right time, right place, and right quantity.
- Enhance advertising campaigns before unfavorable weather periods: Promote articles about the fun of picnicking and camping in the rainy season.
- Pricing strategy and promotions: Offer discounts on remaining inventory and promotions for products before Winter, Spring, or the rainy months to encourage early shopping. At the same time, take advantage of holidays such as Halloween, Christmas, etc., to create marketing campaigns for products suited for family gatherings, indoor BBQs, and similar events.

2. Integrating sales and distribution channels:

- Enhance online sales channels to make it easier for customers to access products regardless of weather.

- Expand distribution with local partners such as stores, tourist areas, or event-organizing companies.

3. Marketing communication:

- Use weather forecasts as a tool to guide marketing strategies and inventory management.
- Develop digital advertising and marketing platforms to create flexibility and adaptability to changing business environments according to weather conditions.