



# **Universidad de Valladolid**

## **Máster en Informática**

TÉCNICAS ESCALABLES DE ANÁLISIS DE  
DATOS EN ENTORNOS BIG DATA: CLASIFICADORES

### **MEMORIA PRIMERA ETAPA**

#### **PROYECTO DE CLASIFICACIÓN**

##### **RAIN IN AUSTRALIA**

**Elisa Batista Blanco**

**John Jairo Ballestas Payares**

**Jorge Martin Villafruela**

20 de octubre 2025

# Índice

<b>Resumen ejecutivo — Rain in Australia.....</b>	<b>3</b>
Origen de los datos (con enlace).....	3
Recopilación.....	3
Tamaño del conjunto de datos.....	3
Número total de atributos y su naturaleza.....	3
Descripción de la clase (etiquetas y distribución).....	4
Características Relevantes.....	4
<b>Exploración de los datos.....</b>	<b>5</b>
Descripción del problema.....	5
Naturaleza y dificultad del problema.....	5
Descripción del conjunto de datos.....	6
Presentación de los datos.....	6
Valores ausentes.....	6
Atributos numéricos - rangos y estadísticos básicos.....	8
Distribución de frecuencias en variables numéricas.....	10
Atributos categóricos.....	14
Variables Booleanas.....	15
Caso especial: Date.....	16
Mapa de correlaciones.....	16
<b>Preparación de datos.....</b>	<b>19</b>
Creación de conjuntos de entrenamiento y prueba.....	19
Limpieza de los datos.....	20
Tratamiento de valores ausentes.....	20
Eliminación de outliers.....	21
Nubosidad.....	21
Precipitación.....	21
Evaporation.....	22
Viento.....	22
Resumen cuantitativo del proceso de limpieza.....	22
Selección y Transformación de atributos.....	23
<b>Carga de Trabajo:.....</b>	<b>26</b>
<b>Referencias.....</b>	<b>28</b>

# Resumen ejecutivo — *Rain in Australia*

## Origen de los datos (con enlace)

El conjunto se denomina *Rain in Australia* y procede de observaciones de la *Australian Bureau of Meteorology*, compiladas y publicadas en Kaggle en el fichero weatherAUS.csv. Enlace de descarga: [Kaggle – weather-dataset-rattle-package \(autor: jsphyg\)](#).

## Recopilación

Los datos se recopilan con fines de **observación meteorológica operativa** (temperatura, precipitación, humedad, presión, viento, nubosidad).

## Tamaño del conjunto de datos

**Nº instancias:** 145,460 instancias y

**Nº variables:** 23

## Número total de atributos y su naturaleza

Total: 23.

### Atributos Numéricos (16):

- **MinTemp, MaxTemp, Temp9am, Temp3pm:** Temperaturas máxima, mínima, a las 9:00 y a las 15:00, respectivamente (°C).
- **Rainfall:** Precipitación del día (mm).
- **Evaporation:** Cantidad de agua evaporada (mm).
- **Sunshine:** Horas de sol.
- **WindGustSpeed, WindSpeed9am, WindSpeed3pm:** Velocidad del viento máxima, a las 9:00 y a las 15:00; respectivamente (km/h)
- **Humidity9am, Humidity3pm :** Humedad a las 9:00 y a las 15:00 (%)
- **Pressure9am, Pressure3pm:** Presión atmosférica (hPa)
- **Cloud9am, Cloud3pm:** Cobertura de nubes en octavos de cielo cubierto

### Atributos Categóricos (4):

- **Location:** Localización de la medición
- **WindGustDir, WindDir9am, WindDir3pm:** Dirección de la mayor ráfaga de viento, la dirección a las 9:00 y a las 15:00 (km/h)

### Atributos Booleanos (2):

- **RainToday** (Yes/No)
- **RainTomorrow** (Yes/No) - **CLASE A CLASIFICAR**

### Atributo Temporal (1):

- **Date:** Fecha de la observación

Se desglosó la variable de fecha en elementos más específicos, como el día, el mes y el año, para estudiar con mayor detalle los patrones de lluvia a lo largo del tiempo.

- Día: Int
- Mes: Int
- Año : Int

## Descripción de la clase (etiquetas y distribución)

La variable objetivo **RainTomorrow** es una variable binaria que indica si lloverá al día siguiente.

Etiqueta	Significado	Cantidad	Porcentaje
<b>NA</b>	Valores faltantes	3267	2.25%
<b>No</b>	No lloverá mañana	110316	75.84%
<b>Yes</b>	Sí lloverá mañana	31877	21.91%

## Características Relevantes

1. **Total de Instancias Válidas:**
  - 142,193 instancias con valores válidos para la clasificación
  - 3,267 instancias con valores faltantes (NA)
2. **Contexto Meteorológico:**
  - El patrón de ~22% de días con lluvia es consistente con el clima de muchas regiones de Australia

Variables objetivo: *RainTomorrow* (lloverá mañana  $\geq 1$  mm). reporta **desbalanceo marcado** con **predominio de la clase “No”** (~80 % No / 20 % Sí,  $\approx 4:1$ ); el porcentaje exacto puede variar por periodo/ubicación.

# Exploración de los datos

## Descripción del problema

El problema consiste en **predecir si lloverá al día siguiente** en un punto de Australia usando observaciones meteorológicas del día actual. Formalmente, se trata de un problema de **clasificación binaria** sobre la variable objetivo **RainTomorrow** (Yes/No), donde “Yes” indica precipitación  $\geq 1$  mm al día siguiente del registro.

El conjunto de datos incluye observaciones diarias recopiladas durante aproximadamente 10 años en múltiples estaciones o ciudades, e incorpora variables como temperatura, humedad, presión atmosférica, viento, nubosidad y lluvia del día, entre otras.

El análisis de este conjunto de datos no solo permite mejorar la precisión de las predicciones meteorológicas, sino que también proporciona información valiosa para la toma de decisiones en sectores dependientes de las condiciones climáticas, como la agricultura, la gestión del agua, el transporte o la energía. En este sentido, su correcta interpretación contribuye a la planificación y reducción de riesgos asociados a eventos climáticos adversos, favoreciendo una gestión más eficiente y sostenible de los recursos.

## Naturaleza y dificultad del problema.

- **Desbalanceo de clases.** Predominan los días “No” frente a “Sí” (~4:1), lo que penaliza clasificadores ingenuos y exige técnicas de ponderación o muestreo, así como métricas acordes (F1/ROC-AUC).
- **Valores faltantes** en numerosas variables (viento, presión, nubosidad, etc.), que obligan a definir una estrategia de imputación antes del modelado (sin modificar los datos en la fase de “Exploración”).
- **Heterogeneidad espacial y estacionalidad.** Las relaciones pueden variar por **ubicación** y por **época del año**; los materiales muestran que modelos locales por ciudad pueden mejorar el rendimiento, lo que sugiere efectos específicos de zona/estación.
- **Riesgo de data leakage.** Algunas variables están directa o derivadamente ligadas a la lluvia (p. ej., **Rainfall** en  $t$ ); deben tratarse con cautela para no “anticipar” la etiqueta de  $t+1$  durante el entrenamiento.

**Hipótesis y señales meteorológicas relevantes.** A nivel meteorológico, la lluvia de mañana suele asociarse a **mayor humedad** y **mayor nubosidad**, y a **menor presión** y **menor insolación** el día previo; la **dirección/velocidad del viento** también aporta información (p. ej., ciertos rumbos a 9:00/15:00).

En entregas posteriores se compararán distintos clasificadores y se justificarán las decisiones de preparación de los datos (particionado estricto *train/test*, limpieza, selección y transformación de atributos), evitando utilizar el conjunto de prueba en el ajuste de modelos para prevenir cualquier fuga de información.

## Descripción del conjunto de datos

Contiene observaciones meteorológicas diarias recopiladas por la Oficina de Meteorología de Australia (*Australian Bureau of Meteorology*) a lo largo de aproximadamente diez años, en 49 localizaciones distribuidas por todo el país. Este conjunto de datos ofrece una visión representativa de las condiciones climáticas australianas en diferentes regiones y estaciones del año, abarcando tanto zonas costeras como áreas interiores.

A continuación, se presenta el esquema del *DataFrame*, que incluye variables relacionadas con la temperatura, humedad, presión atmosférica, velocidad y dirección del viento, nubosidad y precipitaciones. Todas estas variables constituyen factores determinantes para la predicción de la variable objetivo *RainTomorrow*, la cual indica si lloverá o no al día siguiente.

- **Temporal (1):** *Date* (fecha de medición; **3.436** fechas distintas).
- **Nominales (4):** *Location* (49 niveles), *WindGustDir*, *WindDir9am*, *WindDir3pm* (16 rumbos).
- **Booleanas (2):** *RainToday*, *RainTomorrow* (Yes/No).
- **Numéricas continuas (16):**
  - Temperaturas (°C): *MinTemp*, *MaxTemp*, *Temp9am*, *Temp3pm*;
  - Precipitación/Evaporación (mm): *Rainfall*, *Evaporation*;
  - Insolación (h): *Sunshine*;
  - Viento (km/h): *WindGustSpeed*, *WindSpeed9am*, *WindSpeed3pm*;
  - Humedad (%): *Humidity9am*, *Humidity3pm*;
  - Presión (hPa): *Pressure9am*, *Pressure3pm*;
  - Nubosidad (*oktas* **0–8**, con **9** para “no visible por niebla”).

## Presentación de los datos

### Valores ausentes

El conjunto de datos cuenta con 145.460 registros. Durante el proceso de exploración se detectó que varias variables presentan valores faltantes, lo cual puede influir en la calidad del análisis posterior.

- **filas completas: 56.420 (38,79%);**

- filas con  $\geq 1$  vacío/nulo: 89.040 (61,21%).
- registros duplicados = 0

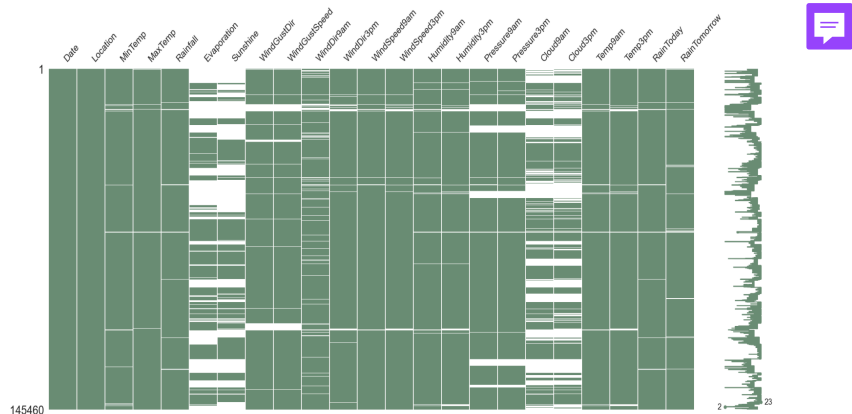
En la siguiente tabla aparecen los diferentes atributos, la cantidad de valores ausentes que tienen, y el porcentaje de los registros totales.

Columna	Valores Nulos/Vacíos	Porcentaje	Columna	Valores Nulos/Vacíos	Porcentaje
<b>Sunshine</b>	<b>69.835</b>	<b>48,01%</b>	<i>RainTomorrow</i>	3.267	2,25%
<b>Evaporation</b>	<b>62.790</b>	<b>43,17%</b>	<i>Rainfall</i>	3.261	2,24%
<b>Cloud3pm</b>	<b>59.358</b>	<b>40,81%</b>	<i>RainToday</i>	3.261	2,24%
<b>Cloud9am</b>	<b>55.888</b>	<b>38,42%</b>	<i>WindSpeed3pm</i>	3.062	2,11%
<b>Pressure9am</b>	<b>15.065</b>	<b>10,36%</b>	<i>Humidity9am</i>	2.654	1,82%
<b>Pressure3pm</b>	<b>15.028</b>	<b>10,33%</b>	<i>WindSpeed9am</i>	1.767	1,21%
<i>WindDir9am</i>	10.566	7,26%	<i>Temp9am</i>	1.767	1,21%
<i>WindGustDir</i>	10.326	7,10%	<i>MinTemp</i>	1.485	1,02%
<i>WindGustSpeed</i>	10.263	7,06%	<i>MaxTemp</i>	1.261	0,87%
<i>Humidity3pm</i>	4.507	3,10%	<i>Date</i>	0	0,00%
<i>WindDir3pm</i>	4.228	2,91%	<i>Location</i>	0	0,00%
<i>Temp3pm</i>	3.609	2,48%			

**Nota:** La tabla está ordenada de mayor a menor porcentaje de valores faltantes. Total de registros en el dataset: 145.460

Antes del análisis, se verifica la presencia de valores ausentes para evaluar la calidad del conjunto de datos. La siguiente matriz muestra su distribución y posibles patrones de ausencia.

**Figura 1: Matriz de valores ausentes**



### Atributos numéricos - rangos y estadísticos básicos.

Se realizó un análisis de los atributos numéricos del conjunto de datos, evaluando su distribución y detectando posibles valores atípicos que podrían afectar los resultados. A continuación, se presentan los principales atributos incluidos en el conjunto de datos, junto con sus unidades de medida y una breve descripción:

Atributo	Unidad / Medición	Descripción breve
<b>MinTemp</b>	°C	Temperatura mínima del día
<b>MaxTemp</b>	°C	Temperatura máxima del día
<b>Rainfall</b>	mm	Precipitación diaria
<b>Evaporation</b>	mm	Nivel de evaporación
<b>Sunshine</b>	horas	Duración del sol
<b>WindGustSpeed</b>	km/h	Velocidad máxima del viento
<b>WindSpeed9am</b>	km/h	Velocidad del viento a las 9 AM



<b>WindSpeed3pm</b>	km/h	Velocidad del viento a las 3 PM
<b>Humidity9am</b>	%	Humedad relativa a las 9 AM
<b>Humidity3pm</b>	%	Humedad relativa a las 3 PM
<b>Pressure9am</b>	hPa	Presión atmosférica a las 9 AM
<b>Pressure3pm</b>	hPa	Presión atmosférica a las 3 PM
<b>Temp9am</b>	C	Temperatura a las 9 AM
<b>Temp3pm</b>	C	Temperatura a las 3 PM

Se procede a examinar las medidas de tendencia central, como la media, mediana y moda, con el objetivo de entender la distribución de las variables meteorológicas. Este análisis ayuda a detectar patrones en temperatura, lluvia, humedad, presión y viento, ofreciendo insumos importantes para la modelación de la probabilidad de precipitación.

Estadísticas descriptivas de las variables numéricas presentes en el conjunto de datos :

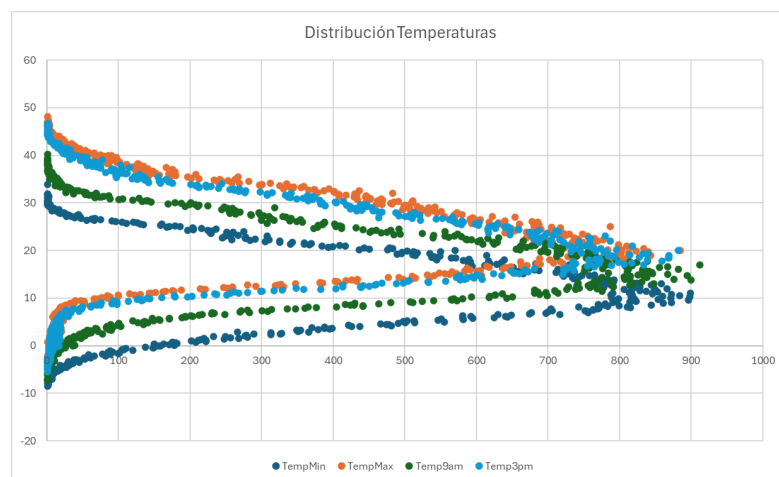
Tabla 2 Estadísticas de los campos numéricos

Campo	Min	Max	Media	Mediana	Std	25%	75%	Count
MinTemp	-8.5	33.9	12.19	12	6.4	7.6	16.9	143975
MaxTemp	-4.8	48.1	23.22	22.6	7.12	17.9	28.2	144199
Rainfall	0	371	2.36	0	8.48	0	0.8	142199
Evaporation	0	145	5.47	4.8	4.19	2.6	7.4	82670
Sunshine	0	14.5	7.61	8.4	3.79	4.8	10.6	75625
Pressure9am	980.5	1041	1017.65	1017.6	7.11	1012.9	1022.4	130395
Pressure3pm	977.1	1039.6	1015.26	1015.2	7.04	1010.4	1020	130432
Temp9am	-7.2	40.2	16.99	16.7	6.49	12.3	21.6	143693
Temp3pm	-5.4	46.7	21.68	21.1	6.94	16.6	26.4	141851
WindGustSpeed	6	135	40.04	39	13.61	31	48	135197
WindSpeed9am	0	130	14.04	13	8.92	7	19	143693
WindSpeed3pm	0	87	18.66	19	8.81	13	24	142398
Humidity9am	0	100	68.88	70	19.03	57	83	142806
Humidity3pm	0	100	51.54	52	20.8	37	66	140953

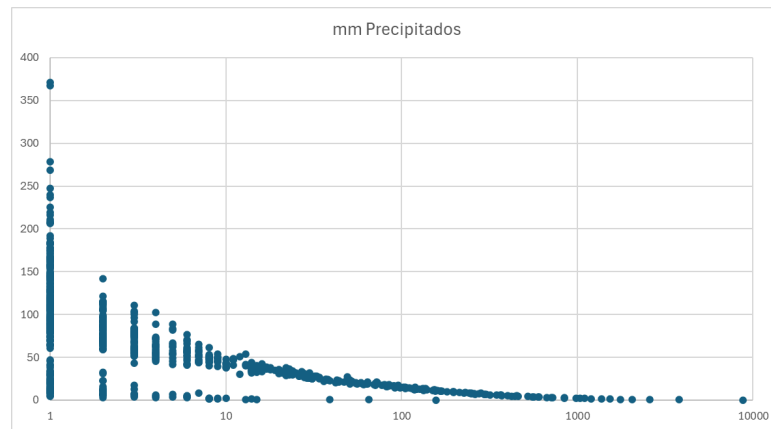
## Distribución de frecuencias en variables numéricas

El análisis de la distribución de frecuencias permitió identificar el comportamiento general de las variables numéricas del conjunto de datos. En términos generales, las variables de temperatura (mínima, máxima, a las 9 am y a las 3 pm) presentan distribuciones aproximadamente simétricas, concentradas en rangos moderados, sin valores extremos significativos.

### Medidas de temperatura (TempMax, TempMin, Temp9am, Temp3pm):

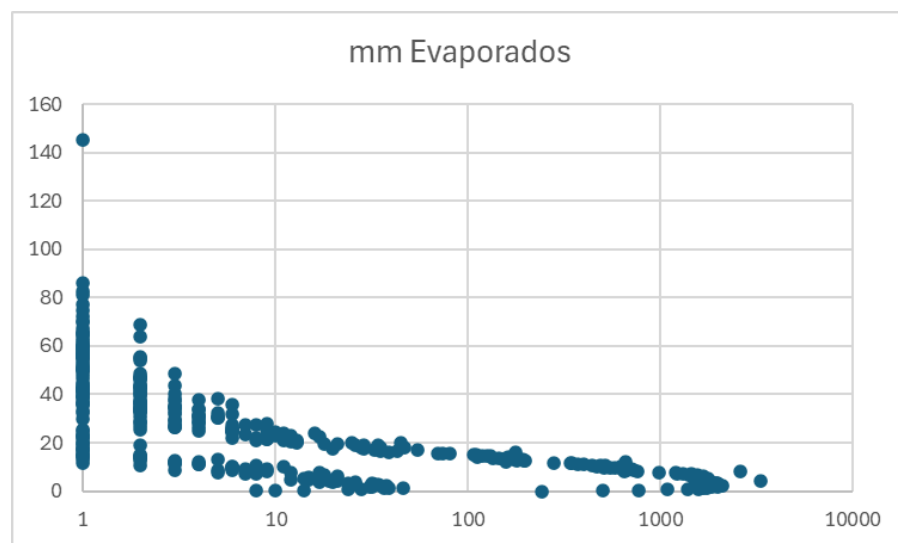


## Rainfall



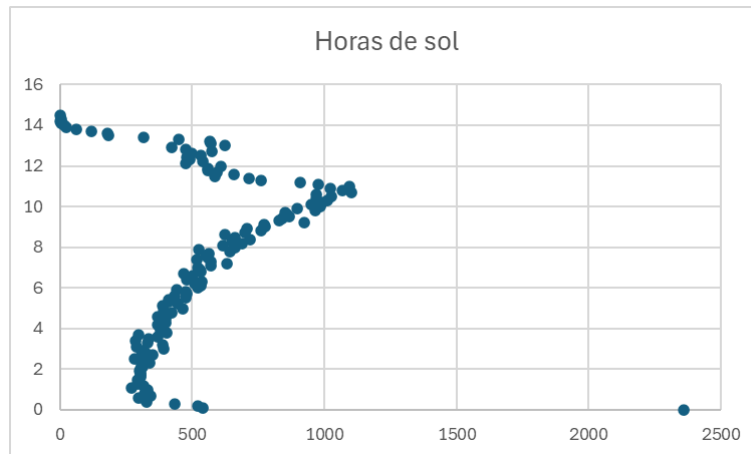
Se ve que tiene una clara presencia de outliers para los valores altos de precipitación. Habrá que estudiar si esos outliers habrá que quitarlos, o son relevantes para estudiar estas situaciones anómalas.

## Evaporation



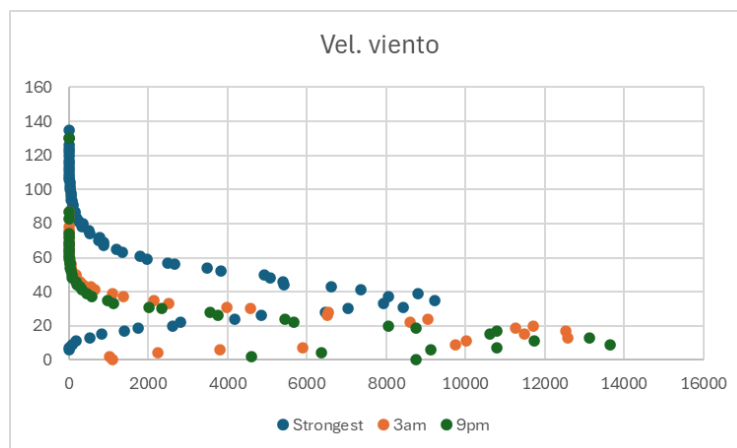
Al igual que *Rainfall*, tiene una clara presencia de outliers, sobre todo porque hay un único valor de más de 140 mm, cuando es físicamente improbable que este valor supere las 40mm.

## Sunshine



La gran cantidad de registros con 0 horas de sol tiene sentido: ahí se encontrarán los días nublados.

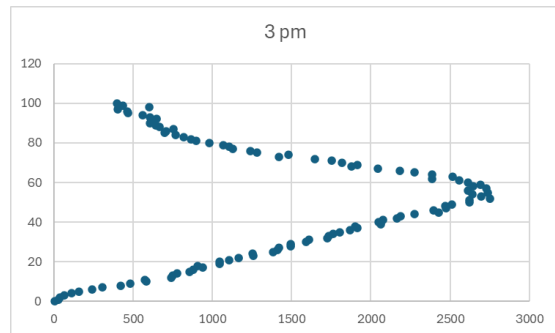
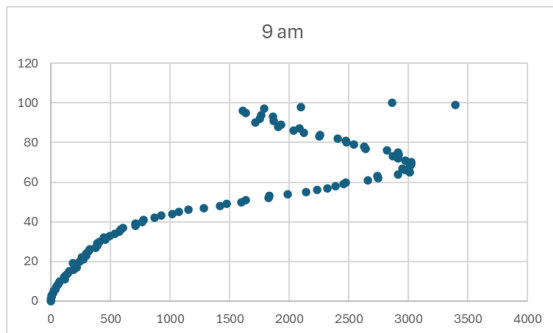
### Velocidad del viento (WindGustSpeed, WindSpeed9am, WindSpeed3am)



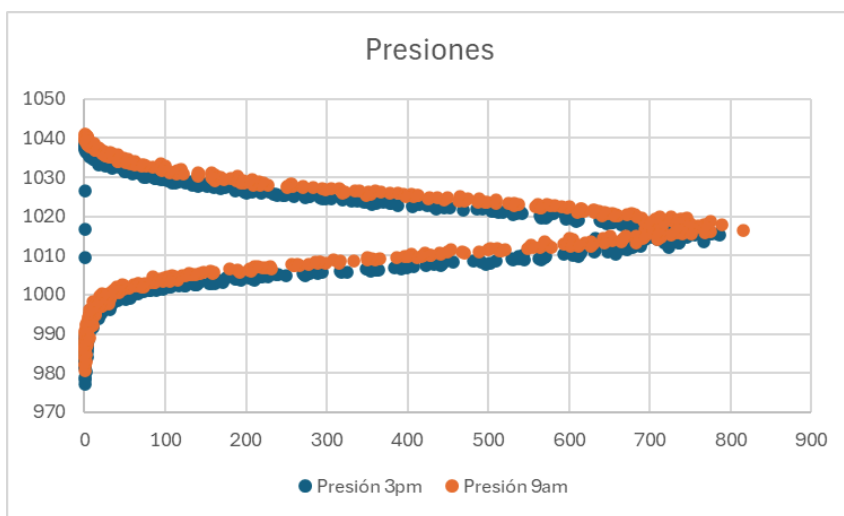
De manera similar a la distribución de precipitaciones, los valores más bajo tienen una mayor frecuencia, mientras que de los mayores valores hay muy pocos registros, por lo que tendremos que analizar si estos outliers hay que quitarlos o preservarlos para estudiar casos específicos.

### Humedad (Humidity9am, Humidity3pm)

- **Valores ausentes:** 62790
- **Rango de valores:** 0 - 100
- **Máximo y mínimo:** [0,100], en ambos casos



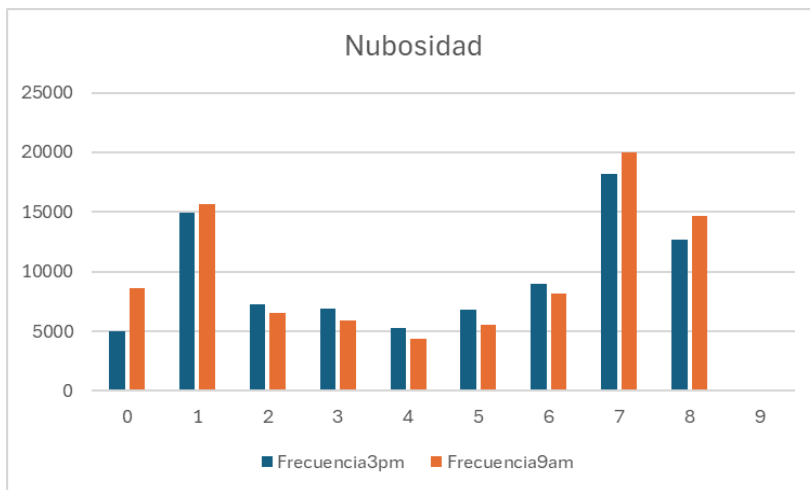
### Presión (Pressure9am, Pressure3pm)



### Nubosidad (Cloud9am, Cloud3pm)

Como se ha definido la nubosidad, los valores posibles son entre 0 (despejado) y 8 (completamente cerrado), de manera cuantitativa. Sin embargo, puede tener además el valor 9 cuando no se puede ver el cielo (niebla). Por ahora lo consideremos así, pero quizás más adelante se tratará como outlier o ruido.

- **Valores ausentes:** 59358 a las 3pm y 55888 a las 9am
- **Rango de valores:** 0-9\* (quizás entre 0-8, depende cómo se considere)
- **Moda:** 7

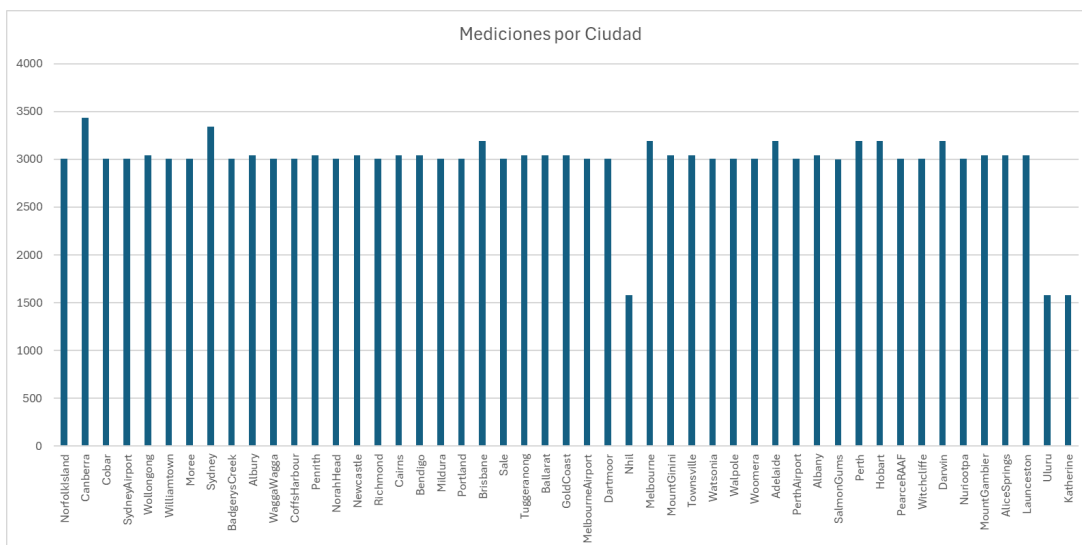


Como se puede observar, los registros con nubosidad 9 son nimios en ambos atributos, lo que probablemente nos permitirá descartar éstos registros, permitiéndonos usarlos como valores cuantitativos.

## Atributos categóricos

### Location

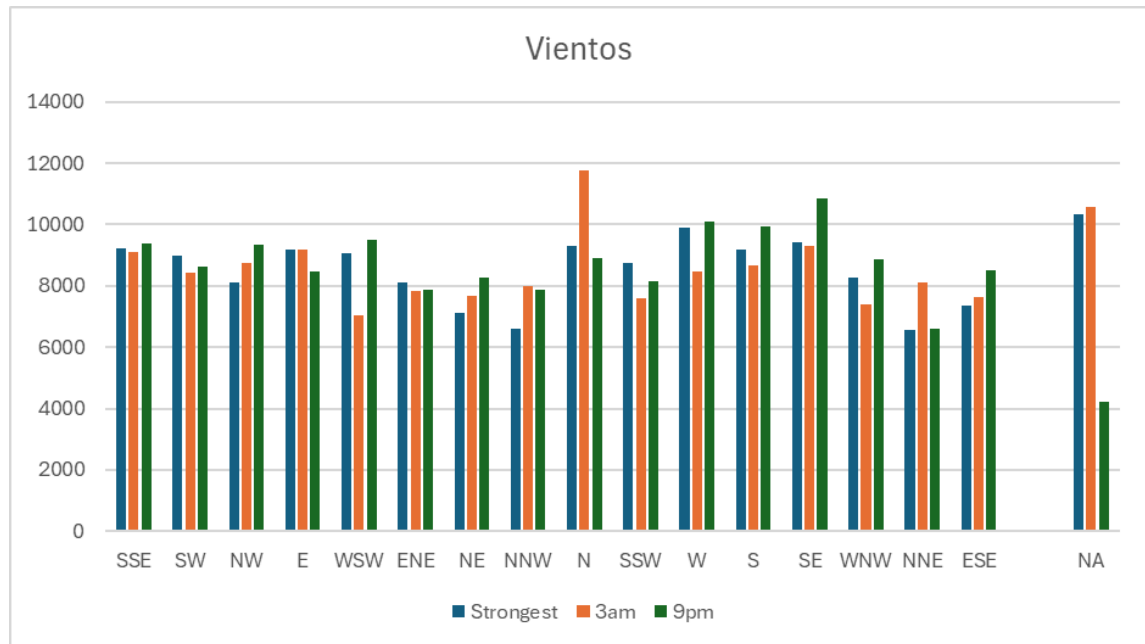
Los datos han sido recogidos en un total de 49 localidades, y están más o menos equilibradas, siendo la ciudad con menos datos Nihl, con 1578 mediciones, y la que más Canberra, con 3436.



Este dato será relevante al correlacionarse con el resto de atributos, puesto que las mediciones, ya que cada estación meteorológica tiene sus propias características (el clima del lugar, qué aparatos de medición tiene, etcétera).

### WindGustDir / WindDir9am / WindDir3pm - Dirección del viento

Los posibles valores que pueden tener son las 16 direcciones cardinales (norte(N), sur(S), este(E) y oeste(W) y sus derivados). Para este atributo, cabe destacar que sí que contienen valores ausentes, y siendo la que menos datos ausentes tiene entre las tres es el de los vientos por la tarde, con 4228 (~3%) de valores ausentes mientras que en los otros casos hay 10326 (~7'1%) y 10566 (~7'26%) valores ausentes respectivamente, a la par con el resto de posibles valores del atributo.



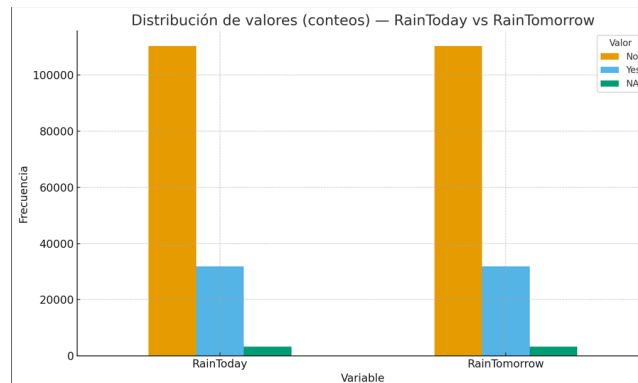
## Variables Booleanas.

### RainToday

Valor	Frecuencia	%
No	110.319	75,84%
Yes	31.880	21,92%
NA	3.261	2,24%

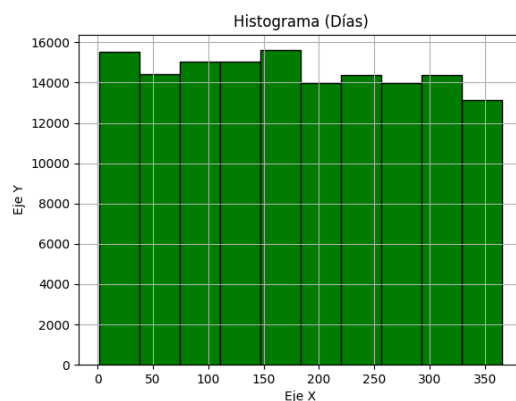
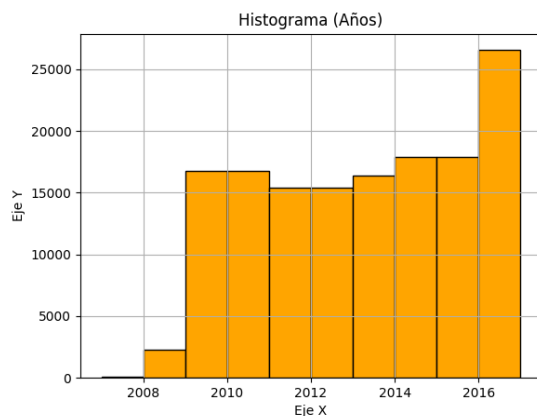
- **Moda:** valor **No**.
- **Ausentes (NA):** 3.261 (2,24%).

Comparando los registros de *RainToday* con *RainTomorrow*, observamos que tienen una distribución bastante similar, lo que da coherencia al conjunto de datos



### Caso especial: Date

La fecha del registro se puede dividir en dos atributos: **Día - Mes y Año**. Mediante esta división, obtenemos dos atributos categóricos diferentes, y que nos permite un mejor estudio de los registros:



Como cabría esperar, los registros son bastante uniformes a lo largo de los diferentes años días, salvo por la excepción de los años 2007 y 2008, donde los registros son bastantes escasos; probablemente por la mayor ausencia de estaciones meteorológicas de ese entonces. Este atributo, sobre todo el día del año, servirá para ver patrones de comportamiento del resto de atributos a lo largo de todo el año.

Esta división del parámetro *Date* se mantendrá a lo largo del estudio, para su mejor manejo.

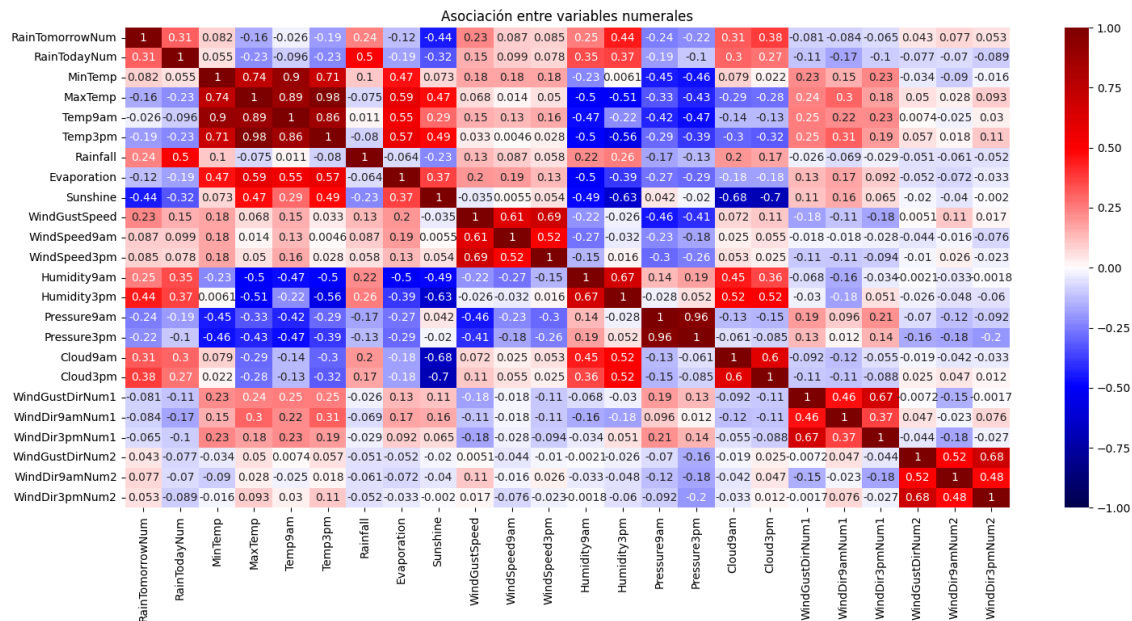
### Mapa de correlaciones



Realizamos un mapa de correlaciones en el que participan, además de los atributos numéricos, los atributos booleanos (False -> 0, True -> 1), y los atributos relacionados con la



dirección del viento, ya que aunque sean categóricos se pueden parametrizar usando dos dimensiones numéricas (Norte-Sur, Este-Oeste) (se explicará en el apartado de la Transformación de los datos):



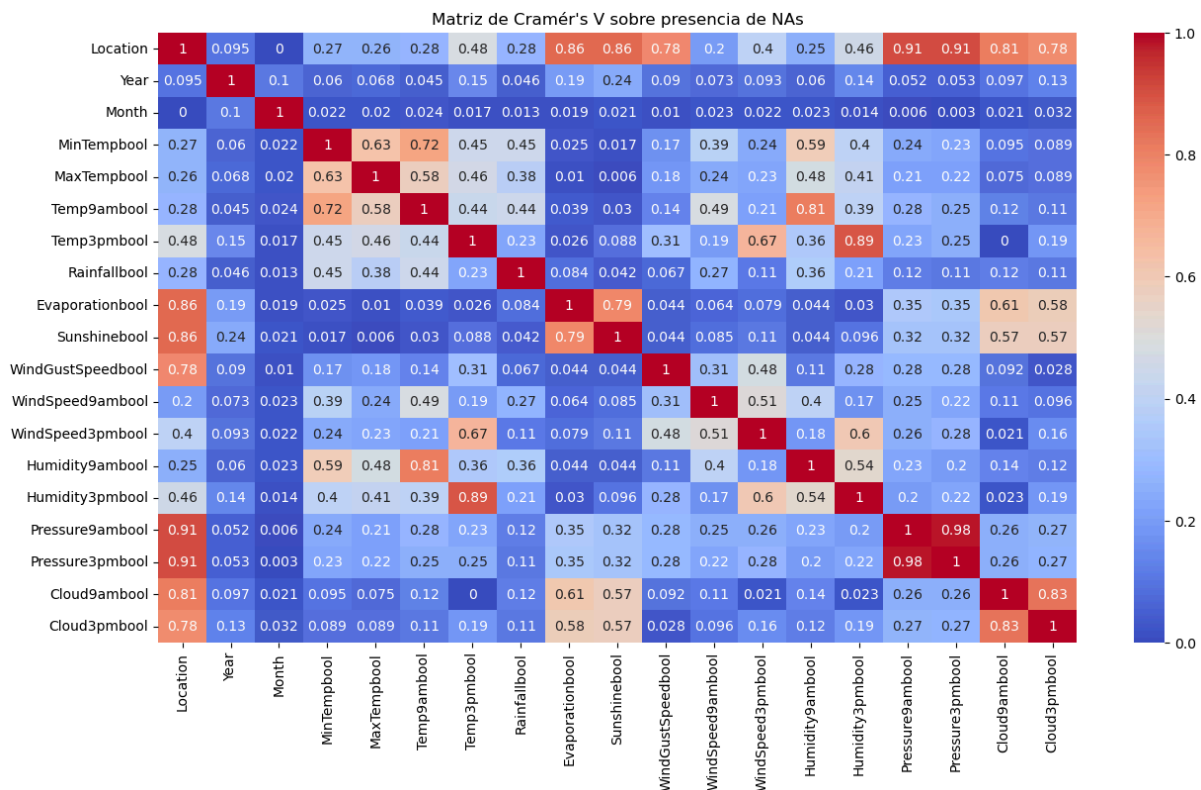
De la matriz de correlación podemos extraer varias observaciones relevantes:

- El atributo más correlacionado con *RainToday* es *Rainfall* - 0.5.
- valores sobre el mismo atributo meteorológico, pero a distinta hora, están altamente relacionados.
- Las horas de sol están altamente relacionadas con las horas de nube (lógicamente) - 0.7.

Además, como comentamos anteriormente, posiblemente la ocurrencia o no de valores ausentes es los atributos está relacionada con la localización de la medida.

Esto nos sugiere realizar una matriz de los valores de Cramer, considerando en vez de los atributos numéricos un categórico que sea “el valor es NaN” -> 0 , “el valor no es NaN” -> 1):

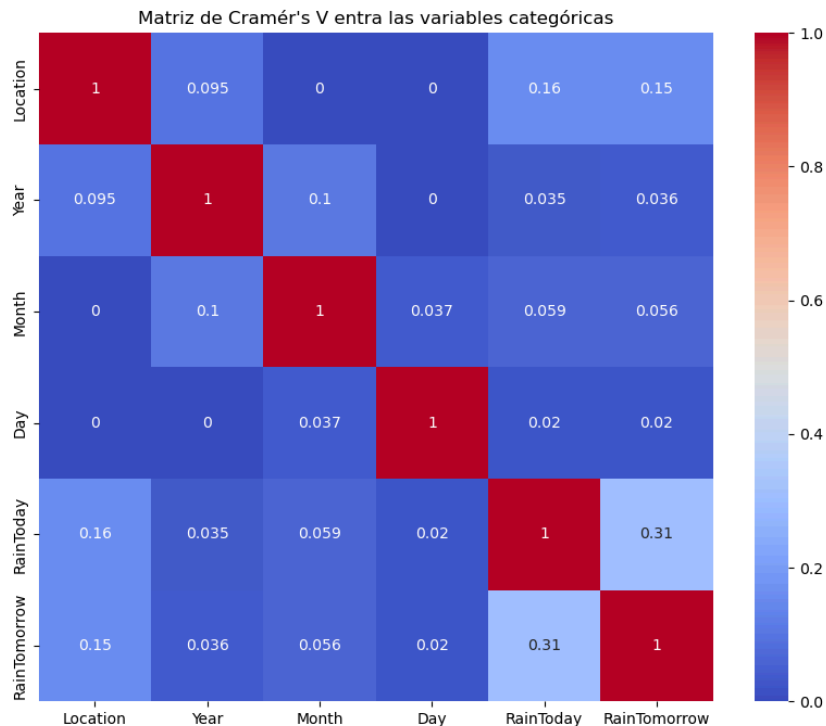




Efectivamente, este mapa de Cramer revela la correlación que **existe entre *Location* y la presencia de valores ausentes en ciertos atributos**, especialmente en los relacionados con la presión, la nubosidad, las horas de sol y la evaporación, que coincide con los atributos con mayor número de valores ausentes. Esto nos da pistas de cómo tratar con los valores ausentes de esos atributos.

Esto contrasta con lo obtenido en las columnas asociadas a las fechas, en las que se ve que apenas hay correlación.

Además, si hacemos un mapa de Cramer sobre los atributos categóricos, se observa que ninguna de ellas está especialmente relacionada con la clase, siendo curioso el caso del mes de la medición



## Preparación de datos.

### Creación de conjuntos de entrenamiento y prueba.

Con el propósito de garantizar una representación equilibrada de las clases en el conjunto de datos, se ha adoptado un esquema de partición estratificada en función de la variable objetivo *RainTomorrow*. Este enfoque busca preservar la proporción de clases ("Yes" y "No") tanto en el conjunto de entrenamiento como en el de prueba, minimizando posibles sesgos en la evaluación del modelo.



El procedimiento seguido consta de los siguientes pasos:

1. **Separación por clase:** Se segmentan los registros en dos subconjuntos según el valor de *RainTomorrow* ("Yes" y "No").
2. **División aleatoria controlada:** Dentro de cada clase, se realiza una partición aleatoria en proporción 80/20 (entrenamiento/prueba), utilizando una semilla fija (*random\_state*) para asegurar la reproducibilidad.
3. **Recomposición estratificada:** Se combinan los subconjuntos *Yes-train* y *No-train* para conformar el conjunto de entrenamiento, y *Yes-test* con *No-test* para el conjunto de prueba. Este procedimiento garantiza que la prevalencia de la clase se mantenga prácticamente constante en ambos conjuntos.

```
seed: Long = 2025
split: Array[Double] = Array(0.8, 0.2)
dfLabel: org.apache.spark.sql.Dataset[org.apache.spark.sql.Row] = [Date: date, Location: string ... 21 more fields]
dfYes: org.apache.spark.sql.Dataset[org.apache.spark.sql.Row] = [Date: date, Location: string ... 21 more fields]
dfNo: org.apache.spark.sql.Dataset[org.apache.spark.sql.Row] = [Date: date, Location: string ... 21 more fields]
yesTrain: org.apache.spark.sql.Dataset[org.apache.spark.sql.Row] = [Date: date, Location: string ... 21 more fields]
yesTest: org.apache.spark.sql.Dataset[org.apache.spark.sql.Row] = [Date: date, Location: string ... 21 more fields]
noTrain: org.apache.spark.sql.Dataset[org.apache.spark.sql.Row] = [Date: date, Location: string ... 21 more fields]
noTest: org.apache.spark.sql.Dataset[org.apache.spark.sql.Row] = [Date: date, Location: string ... 21 more fields]
trainDF: org.apache.spark.sql.Dataset[org.apache.spark.sql.Row] = [Date: date, Location: string ... 21 more fields]
testDF: org.apache.spark.sql.Dataset[org.apache.spark.sql.Row] = [Date: date, Location: string ... 21 more fields]
```

```
Train: 113892 | Test: 28301
Distribución en Global (antes del split) (N=142193):
+-----+-----+
|RainTomorrow|count|pct|
+-----+-----+
|No           |110316|77.58|
|Yes          |31877 |22.42|
+-----+-----+

Distribución en Train (N=113892):
+-----+-----+
|RainTomorrow|count|pct|
+-----+-----+
|No           |88349 |77.57|
|Yes          |25543 |22.43|
+-----+-----+

Distribución en Test (N=28301):
+-----+-----+
|RainTomorrow|count|pct|
+-----+-----+
|No           |21967 |77.62|
|Yes          |6334  |22.38|
+-----+-----+
```

## Limpieza de los datos


### Tratamiento de valores ausentes


Sin tener en cuenta la localización y la fecha, que no tienen datos ausentes, solo quedan los atributos numéricos, y los tres valores categóricos sobre la dirección del viento.

Primero, hay que resolver los valores ausentes debidos a la ausencia de aparatos de medición en ciertas localizaciones. En esos casos, no podemos utilizar los datos de la misma localización. Por tanto, en estos casos, se usará la **media total de las mediciones obtenidas en el mismo mes de la medición** para los atributos numéricos que hemos localizado que tienen una distribución normal (*Temperature, Pressure*). En cambio, para los atributos que no siguen una distribución normal (*Rainfall, Evaporation, Sunshine, Humidity, Cloud, WindSpeed,* ), usamos la **mediana de las mediciones obtenidas en el mismo mes de la medición**.

Como segundo paso, hay que tratar los valores ausentes esporádicos de los atributos numéricos. En estos casos, puede darse dos situaciones:

- Es un valor ausente aislado (es decir, existen registros del día siguiente y anterior sobre el atributo dado en la localización). En este caso, se considerará la media/mediana de las mediciones de los días colindantes (dependiendo de la distribución del atributo).
- Hay varios valores ausentes juntos. En este caso, consideraremos el valor medio/mediana del atributo en la localidad en ese mes (dependiendo de la distribución del atributo).


Este mismo proceso podemos hacer con las variables *WindDir* ya parametrizadas los valores ausentes los asignamos a la media de los valores colindantes (o usar la mediana del mes). Sin embargo, en este caso tenemos que terminar **normalizando el resultado**, para que sea un punto de la circunferencia (sus coordenadas (a,b) cumplan  $a^2 + b^2 = 1$ ). Si en algún caso **el valor resultante es (0,0)**, asumimos que el tratamiento ha fallado y **descartamos el registro**. 

Por último, se tienen que considerar los valores ausentes del atributo booleano *RainToday*. En este caso, debido al bajo número de valores ausentes, consideremos asignarles **el valor de la moda en la localización y mes dado**. De esta manera, no perdemos información sobre el resto de los atributos del registro. 

## Eliminación de outliers

Hay cuatro clases de atributos que, al ver la distribución, muestran que hay outliers: nubosidad, precipitación (*Rainfall*), evaporación y la velocidad del viento. El resto de atributos no muestran datos sospechosos en sus distribuciones, así que se conservarán tal y como están

### Nubosidad

Para medir la nubosidad, se mide los octavos de cielo cubierto, lo que implica que el rango de valores se tiene que encontrar entre 0 y 8. Sin embargo, nos hemos encontrado algunos registros donde la nubosidad ha sido 9. A veces, este valor se suele dar cuando no se puede observar el cielo (por ejemplo, por haber niebla). Estos 3 registros son una parte minúscula del dataset (0,002% del conjunto), por lo que **los eliminamos del conjunto**. 

### Precipitación

Observando los outliers de la precipitación vamos a ver el 0.1% de los valores más elevados (percentil 99.9). Solo miramos como outliers los elementos del extremo superior (no tiene sentido hacerlo sobre el extremo inferior (todos los valores caerán en 0)).

Observando los datos, sobre todo los más elevados, nos fijamos que corresponden con la ocurrencia de tormentas y precipitaciones. Por ejemplo, el segundo mayor registro corresponde a la localidad de Darwin en 16-02-2011. Buscando en internet:




ABC News


abc.net.au > news > 2011-02-16 > cyclone-carlos-batters-darwin-and-top-end > 1945818

#### Cyclone Carlos batters Darwin and Top End - ABC News


16 de febrero de 2011 - In the Darwin suburb of Marrara, a record 435 millimetres of rain fell in 24 hours. There are serious concerns about storm surge in the Rapid Creek, Millner and Nightcliff areas, with a high tide of 6.6 metres expected at 5.40pm CST, Northern Territory...

Prestando atención a su influencia en el valor de la clase, obtenemos que de los 145 registros que forman el 0.1%, en solo 38 llueve el siguiente día, por lo que estos outliers muestran una clara influencia sobre el valor de la clase. Por ese motivo, **decidimos conservar estos outliers**. 

### Evaporation

Tal y como se ha medido la evaporación, implica que, por condiciones físicas, aquellos valores superiores a 40 mm  son muy singulares, incluso podría decirse imposibles. Observando al igual que con las precipitaciones los registros más altos, observamos que están muy por encima de ese nivel físicamente posible (llegando incluso a triplicar su valor en un registro), siendo 35mm el menor de ellos. Por ello, **eliminaremos estos registros**.

### Viento

Sucede algo similar a lo ocurrido con precipitación, donde observamos que en este subconjunto de registros, *RainTomorrow* suele valer "Yes", por lo que son un caso diferenciado. Por tanto, de la misma manera, **decidimos conservar estos outliers**. 

## Resumen cuantitativo del proceso de limpieza

En la ejecución realizada del proceso de limpieza, se han obtenido los siguientes resultados sobre el conjunto de entrenamiento:

Categoría	Descripción	Registros afectados
NTAIE	Por <i>Cloud</i> = 9	0
	De los cuales, vectores de viento nulos (0,0)	0
NTOE	Outliers eliminados	28

Tasa de no clasificados	Proporción del conjunto que no pudo ser clasificado tras limpieza	0,0989%
-------------------------	---	---------

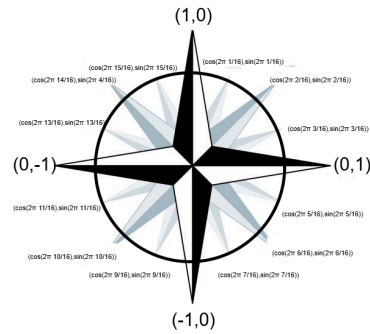
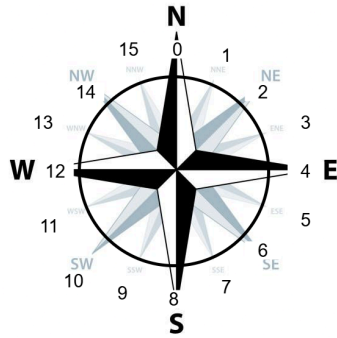
## Selección y Transformación de atributos

### Eliminación de atributos no informativos o redundantes

- **RainToday:** *Rainfall* nos ofrece unos registros más detallados que los de RainToday, puesto que además de indicar si llueve o no también aparece la cantidad de lluvia que ha precipitado, que es importante como hemos observado estudiando los outliers de *Rainfall*.
- **Location y Date:** Estos atributos se utilizaron exclusivamente durante la limpieza de datos (por ejemplo, para imputaciones por localidad y mes). No aportan directamente información útil para la predicción y podrían introducir ruido o sobreajuste si se codifican de forma inapropiada. Por ello, también se eliminan del conjunto final. Otra opción que se había barajado para poder continuar usando las ciudades es agrupar las ciudades entre aquellas que se encuentren cercanas, lo que ayudaría a simplificar su transformación a numérico. Al final se descartó esa idea por falta de tiempo.
- **Sunshine.** En la matriz de correlación no habíamos encontrado además que la nubosidad y las horas de sol se encontraban bastante relacionadas. Puesto que la nubosidad tiene menos valores ausentes, eliminamos la variable *Sunshine*.
- En el caso de los atributos meteorológicos, nos encontramos que las mediciones de lo mismo en distintos momentos del día son **redundantes**, como se puede observar en el mapa de correlaciones, sobre todo en caso de las presiones y las temperaturas. Para todas ellas, crearemos <AtributoMeterologico>, que será la media de estas mediciones (en caso de que sea discreta, como *Cloud*, nos quedamos con su parte entera).

### Transformación de atributos categóricos: dirección del viento

Como ya se comentó anteriormente, pese a ser una variable categórica, podemos parametrizar las mediciones de los vientos (*WindGustDir*, *WindDir9am* y *WindDir3pm*). Asignando cada dirección cardinal a su posición en una circunferencia, podemos asignarles su posición en la circunferencia (lo cual se realiza mediante el seno y el coseno



De esta manera, se realiza la siguiente transformación:

$$DirCardinal \rightarrow n \rightarrow (\cos(2\pi n/16), \sin(2\pi n/16))$$

No nos value únicamente con el primer paso (transformar cardinal a numérico) puesto que se le asigna un orden incorrecto a los valores del atributo: bajo esa asignación, el orden imbuido por ejemplo consideraría que la dirección N (0) es más cercana a S (8) que a NNW (15), lo cual es falso. En cambio, esta parametrización refleja la posición geométrica de las diferentes direcciones. Además, esta transformación es más eficiente a nivel uso de variables, puesto que únicamente necesitamos dos variables para representar los 16 posibles valores del atributo.

Hemos hecho la excepción de realizar esta transformación antes de la limpieza de los datos, ya que para asignar los valores de los valores ausentes, nos resultaba más sencillo realizarlo con la transformación ya hecha.

## Discretización de los atributos continuos

Con el objetivo de mejorar el rendimiento de los modelos que se vayan a utilizar, sobre todo teniendo en cuenta la cantidad de atributos no normales que realizamos, se necesita realizar una discretización de los atributos continuos que tenemos. Esta discretización se hará para que por cada atributo haya un máximo de 15 valores. También se aprovechará para certificar que los valores de los atributos discretos no han sufrido cambios en alguna transformación, y se corrige su valor en dicho caso.

## Normalización de atributos numéricos

Por último, faltaría normalizar los atributos numéricos para que todos los atributos sean igual de importantes.

Para las distribuciones con un rango finito (*Cloud* y *Month*) que siguen una distribución uniforme, bastaría con escalar. Un caso especial son los atributos de la parametrización del viento, que se tendrán que normalizar como vector de características (puede ser que tras la agrupación de las tres medidas se obtenga un vector que no sea de norma 1).



Mientras, para el resto de atributos hace falta estandarizar, que se puede hacer mediante la función `StandardScaler`. Recordemos que esto solo se realiza sobre el conjunto de entrenamiento.



## Carga de Trabajo:

- **Elisa Batista Blanco:** Formateo de la Memoria, y generación de cuadernos de Jupyter.
- **John Jairo Ballestas Payares:** Código Spark.

- **Jorge Martín Villafruela:** Redacción, generación de gráficos (Excel) y funciones *cambiarValorNAModa* y *cambiarValorNAMediana* de Spark, apoyo en la generación de cuadernos.



# Referencias

Hyder, J. (2017). *Weather dataset rattle package*. Kaggle.

<https://www.kaggle.com/datasets/jsphyg/weather-dataset-rattle-package>

Gupta, R., Yadav, A. K., Jha, S. K., & Pathak, P. K. (2022). *Enhancing classification performance of binary class imbalanced data for weather forecasting using machine learning classifiers*. *Research Article*.

<https://doi.org/10.21203/rs.3.rs-1546284/v1>

ABC News. (2011, 16 de febrero). *Cyclone Carlos batters Darwin and Top End*. ABC News.

<https://www.abc.net.au/news/2011-02-16/cyclone-carlos-batters-darwin-and-top-end/1945818>

**Universidad de Valladolid.** (2025). *TESCC-2025-26 (META6940-2025)*. Campus Virtual UVa. <https://campusvirtual.uva.es/course/view.php?id=6940&section=1>

Apache Software Foundation. (n.d.). *ChiSquareTestExample.scala* [Código fuente]. *GitHub*.

<https://github.com/apache/spark/blob/master/examples/src/main/scala/org/apache/spark/examples/ml/ChiSquareTestExample.scala>