# Descripción del Set de Datos

# Alumno: Mariano Buet

## Origen de los datos

El conjunto de datos utilizado proviene del Servicio Meteorológico Nacional (SMN) de la República Argentina, bajo el título “Estadísticas Climatológicas Normales (Período 1991–2020)”.
Estos registros están disponibles públicamente a través del portal de datos oficiales del SMN “ https://www.smn.gob.ar/descarga-de-datos” y también se encuentran referenciados en datos.gob.ar, dentro del apartado de climatología nacional.

El archivo original fue publicado como un documento de texto tabulado y contiene las medias mensuales de variables meteorológicas registradas en distintas estaciones del país durante el período de 30 años comprendido entre 1991 y 2020.

## Descripción general del contenido

El dataset contiene información de más de 50 estaciones meteorológicas distribuidas a lo largo del territorio argentino.
Cada estación cuenta con observaciones mensuales de distintas variables climáticas, las cuales son:

| Variable | Descripción | Unidad |
|---|---|---|
| Temperatura | Promedio mensual de temperatura | °C |
| Temperatura máxima | Promedio mensual de las temperaturas máximas diarias | °C |
| Temperatura mínima | Promedio mensual de las temperaturas mínimas diarias | °C |
| Humedad relativa | Promedio mensual de humedad en el aire | % |
| Velocidad del viento | Velocidad media del viento | km/h |
| Nubosidad total | Promedio mensual de cobertura nubosa | octavos |
| Precipitación | Precipitación mensual acumulada | mm |
| Frecuencia de días con precipitación > 0.1 mm | Número de días con lluvias o nieve | días |

## Estructura de los datos

Cada registro corresponde a una combinación única de estación meteorológica, variable y meses registrados.

Por ejemplo:

| Estación | Valor Medio de | Ene | Feb | Mar | Abr | May | Jun | Jul | Ago | Sep | Oct | Nov | Dic |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| LA QUIACA OBSERVATORIO | Temperatura (°C) | 13,2 | 13 | 12,8 | 11,3 | 7,3 | 4,8 | 4,5 | 7 | 10 | 12,4 | 13,4 | 13,9 |

Para el modelado, se prevé reformatear los datos a un formato de tabla unificada, donde cada fila represente una instancia mensual con todas las variables meteorológicas de una estación:

| ESTACION | MES | Temperatura (°C) | Temperatura máxima (°C) | Temperatura mínima (°C) | Humedad relativa (%) | Velocidad del Viento (km/h) (2011-2020) | Nubosidad total (octavos) | Precipitación (mm) | Frecuencia de días con Precipitación superior a 1.0 mm |
|---|---|---|---|---|---|---|---|---|---|
| LA QUIACA OBSERVATORIO | ENERO | 13,2 | 20,6 | 7,7 | 62,6 | 6,5 | 4,9 | 101,9 | 11,5 |

## Pre procesamiento planificado

Eliminación de valores “S/D” (sin dato) y reemplazo por NaN para permitir imputación o eliminación posterior.

Transformación de variables: cálculo de promedios anuales o trimestrales según necesidad.

Tratamiento adecuado de variables categóricas

**Generación de la variable objetivo:**

Riesgo de helada: temperatura mínima < 3 °C

Clima de confort: 5 °C ≤ temperatura media ≤ 30 °C

Riesgo de sobrecalor: temperatura máxima > 30 °C

## Cantidad de registros y características

Luego del pre procesamiento el set de datos quedaría de la sig manera:

*Cantidad de registros*: Aproximadamente 600 a 700 registros, considerando 50 estaciones × 12 meses.

*Características (columnas):* 11 Variables, 2 nominales + 8 numéricas + Variable Objetivo.

*Tipos de variables:*

*Numéricas continuas*: temperaturas, humedad, viento, precipitación.

*Categóricas nominales*: nombre de estación, mes.

*Variable objetivo categórica: estado térmico esperado* (helada, confort o sobrecalor).