

**4th grade**

# Regresión Logística

## Hoja de Trabajo 6

Diego Alberto Leiva  
Gustavo Andrés González  
Maria Marta Ramírez  
José Pablo Orellana  
Gabriel Estuardo García

# Introducción al DataSet



# Orígen de los datos y Propósito principal

- El dataset fue extraído de Kaggle, la cual contiene varios datos sobre el clima en Australia
- El principal objetivo de este trabajo es poder analizar todas las variables del dataset, para poder predecir el clima.




# Datos

- Se cuenta con 145460 observaciones y 23 variables
- Datos de diciembre del 2008 hasta julio del 2017

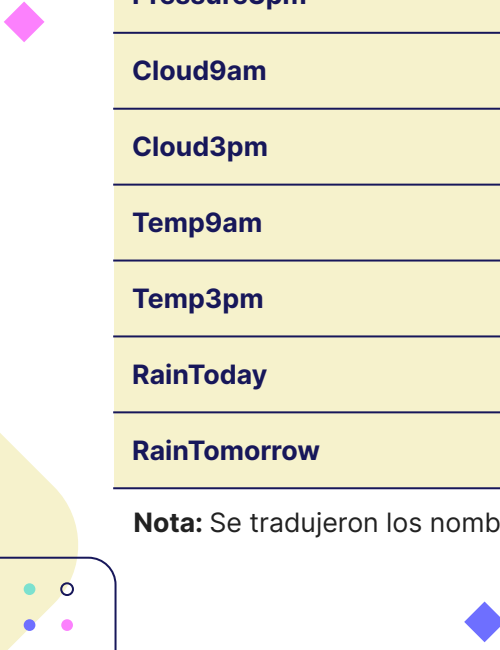


<b>Date</b>	Fecha de la observación
<b>Location</b>	Ubicación geográfica
<b>MinTemp</b>	Temperatura mínima del día en grados Celsius.
<b>MaxTemp</b>	Temperatura máxima del día en grados Celsius
<b>Rainfall</b>	Cantidad de precipitación registrada en el día en milímetros
<b>Evaporation</b>	Tasa de evaporación del día en milímetros
<b>Sunshine</b>	Número de horas de sol durante el día
<b>WindGustDir</b>	Dirección de la ráfaga de viento más fuerte en 16 puntos cardinales.
<b>WindGustSpeed</b>	Velocidad de la ráfaga de viento más fuerte en kilómetros por hora
<b>WindDir9am</b>	Dirección del viento a las 9am en puntos cardinales
<b>WindDir3pm</b>	Dirección del viento a las 3pm en puntos cardinales
<b>WindSpeed9am</b>	Velocidad del viento a las 9am en kilómetros hora



<b>WindSpeed3pm</b>	Velocidad del viento a las 3pm en kilómetros hora
<b>Humidity9am</b>	Humedad a las 9am porcentaje
<b>Humidity3pm</b>	Humedad a las 3pm porcentaje
<b>Pressure9am</b>	Presión atmosférica a las 9am en hectopascales
<b>Pressure3pm</b>	Presión atmosférica a las 9am en hectopascales
<b>Cloud9am</b>	Porcentaje de cobertura nubosa a las 9am en octavos
<b>Cloud3pm</b>	Porcentaje de cobertura nubosa a las 3pm en octavos
<b>Temp9am</b>	Temperatura a las 9am en grados Celsius
<b>Temp3pm</b>	Temperatura a las 3pm en grados Celsius
<b>RainToday</b>	Indicador si llovió o no (Sí/No)
<b>RainTomorrow</b>	Indicador si lloverá al día siguiente (Si/No)

**Nota:** Se tradujeron los nombres de las columnas



# Variables

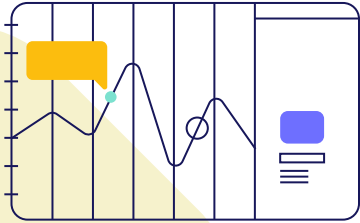
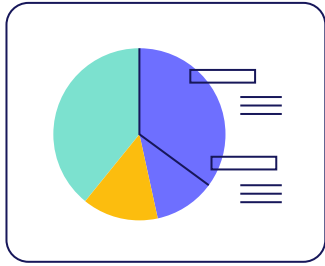
## Cuantitativas

- MinTemp
- MaxTemp
- Rainfall
- Evaporation
- Sunshine
- WindGustSpeed
- WindSpeed9m
- WindSpeed3pm
- Humidity9am
- Humidity3pm
- Pressure9am
- Pressure3pm
- Cloud9am
- Cloud3pm
- Temp9am
- Temp3pm

## Categoricas

- Date
- Location
- WindGustDir
- WindDir9Am
- WindDir3pm
- RainToday
- RainTomorrow

# Objetivo de Análisis





# Proposito Comercial

Este análisis proporciona información relevante para una amplia cantidad de industrias y negocios como:

- Agricultura: Ya que estos puede utilizar el pronóstico del tiempo para planificar la siembra, el riego y la cosecha de los cultivos, teniendo así una mayor eficiencia.
- Turismo: Las empresas turísticas requieren estos datos para poder anticipar patrones climaticos, cambiando así sus estrategias de marketing y operaciones en función de las condiciones.
- Seguros: Estas compañías pueden utilizar los pronósticos para evaluar los riesgos asociados con los eventos climáticos, ofreciendo una cobertura adecuada.


Entre otros tipos de negocios





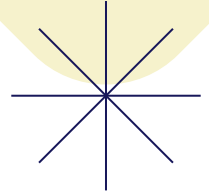
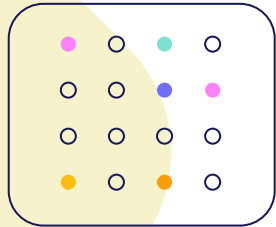
# Objetivo de la Regresión Logística

El objetivo principal de nuestra regresión logística es el saber el pronóstico del clima, con ayuda de las variables climáticas que tenemos disponibles.



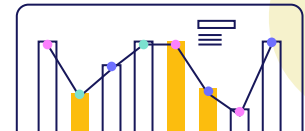
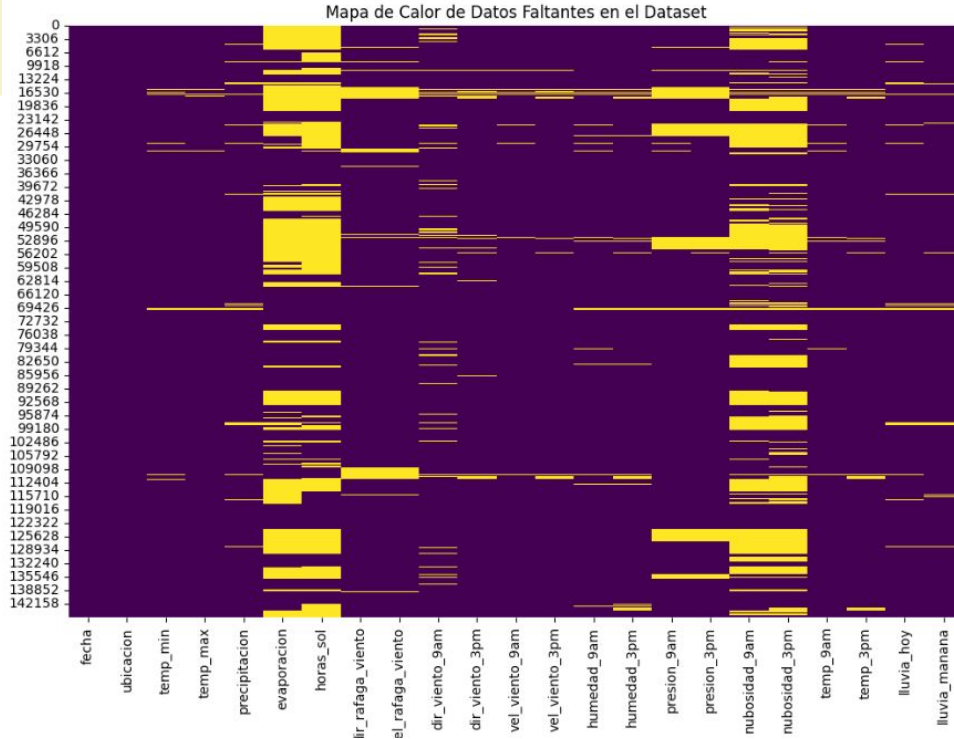
Se buscó crear un modelo que pueda clasificar por medio de dos estado “Lluvia” o “No lluvia”. Este modelo puede utilizarse para tomar decisiones informadas sobre actividades al aire libre, planificación de viajes, etc. La precisión del modelo se evalúa utilizando dos métricas como la exactitud del test y el entrenamiento, así como el área bajo la curva ROC, que indica la capacidad del modelo para distinguir entre clases positivas y negativas.

# Metodología



# Visualizar data faltante

Se buscó realizar un mapa de calor para poder observar los datos faltantes en el dataset, en donde podemos encontrar que las columnas numéricas como horas\_sol, evaporacion, nubosidad\_3pm y nubosidad\_9am tienen data faltante de hasta el 40%



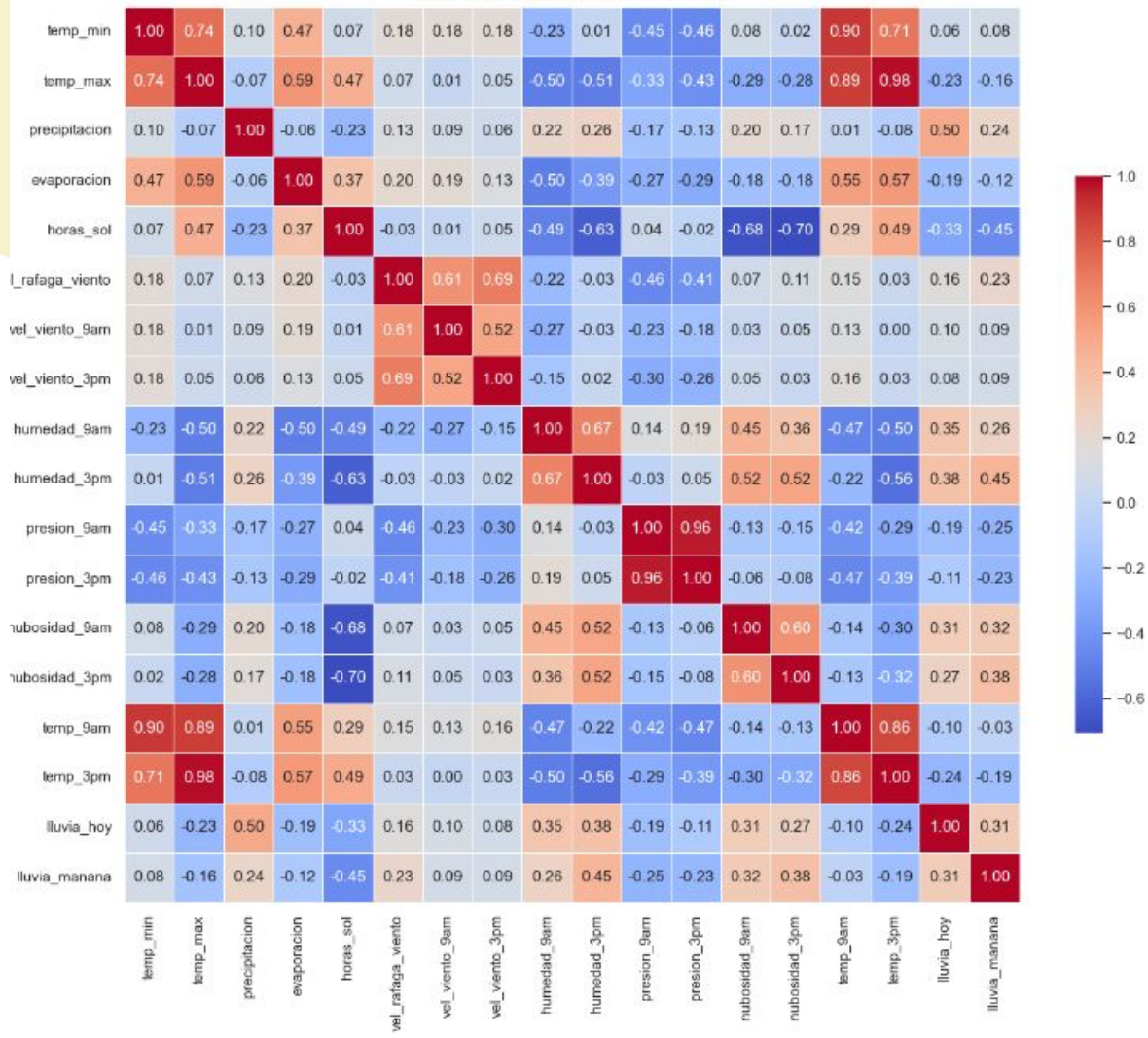
## Eliminar datos nulos

El eliminar todos los datos nulos tendría un gran impacto en el dataset, ya que esto nos dejaría con un 38% de los datos originales, en consecuencia se ha determinado que esta no es una acción viable para el análisis

## Evaluación de correlación entre las variables numéricas

Se asume que la variable objetivo será la variable lluvia\_manana, que servirá para poder predecir si el día de mañana habrá lluvia. Es por ello que es necesario que esta sea una variable numérica para poder evaluar la correlación de esta con las demás variables.

Matriz de Correlación de variables numericas

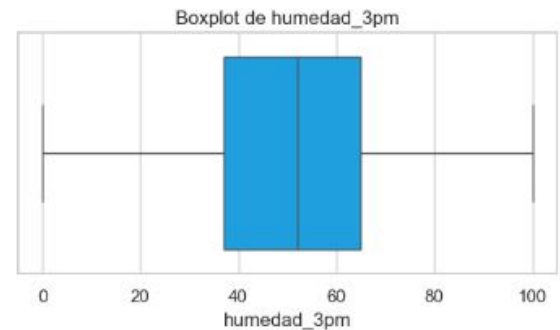
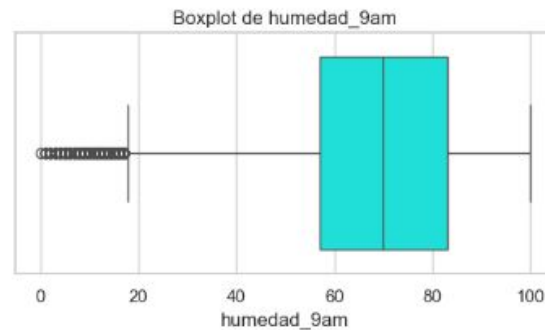
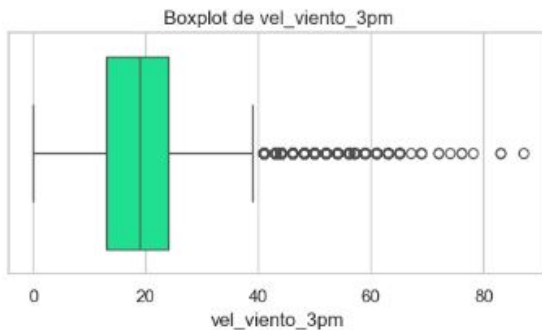
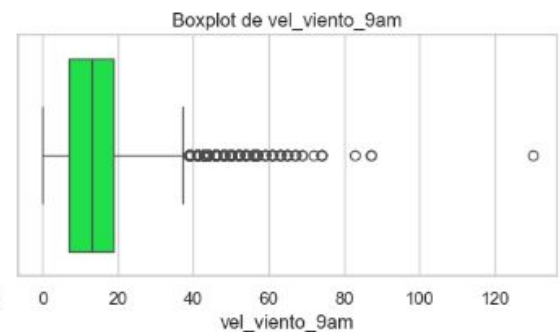
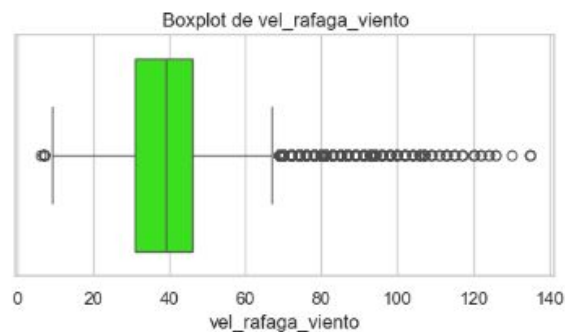
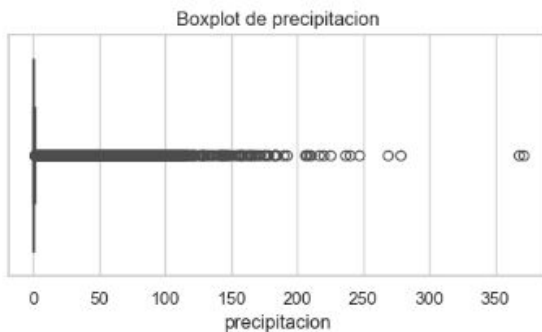
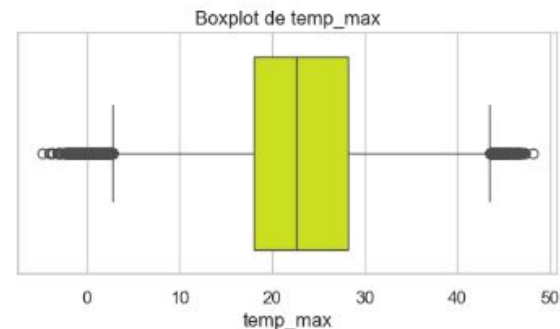
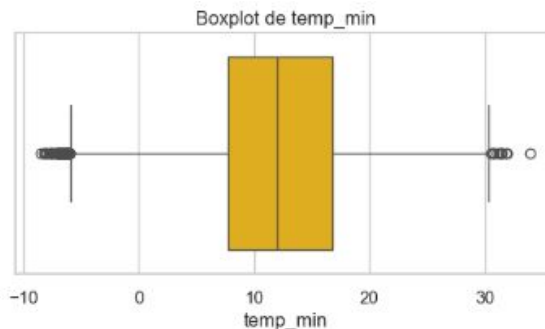
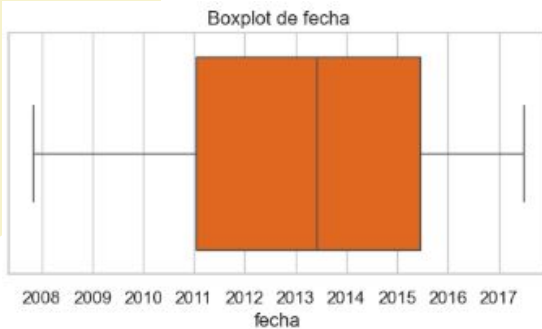


Aquí podemos observar como las variables problemáticas detectadas anteriormente tiene si mucho una correlación ligeramente moderada con las variables objetivo de lluvia.

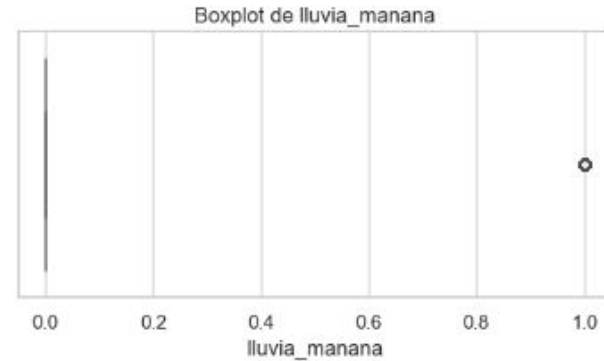
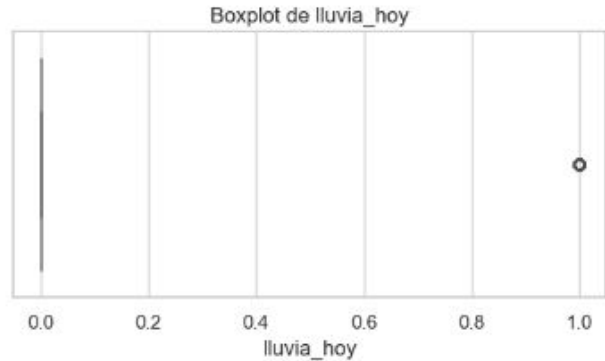
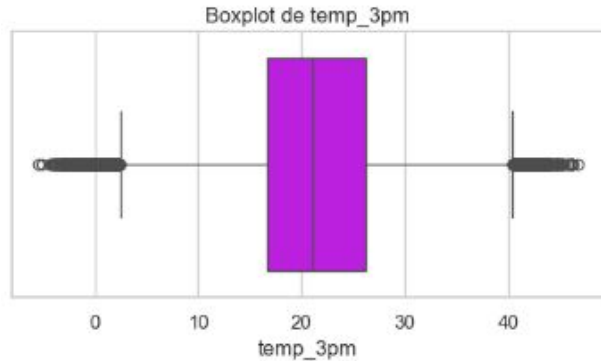
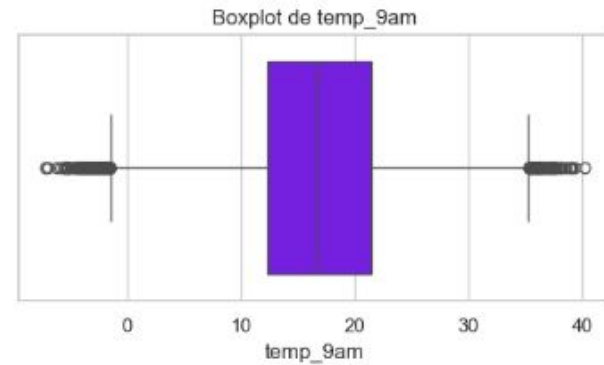
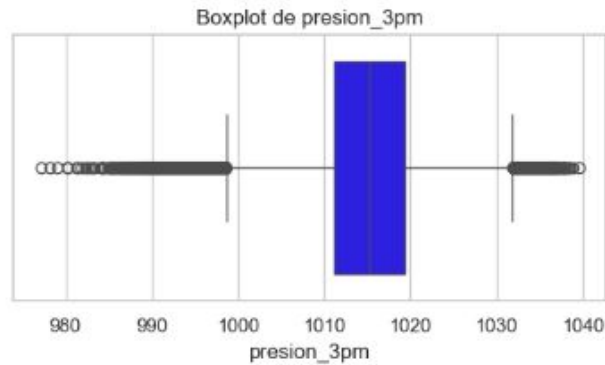
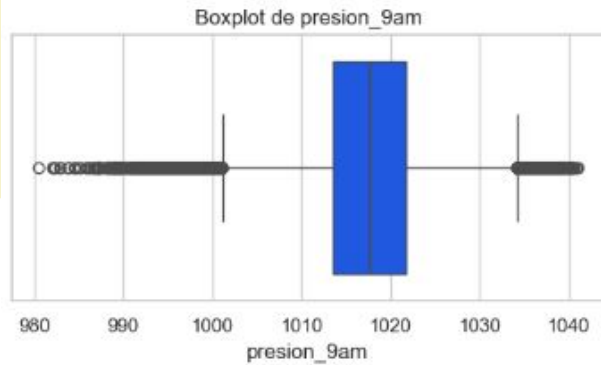
Dado que imputar más del 40% de los valores a estas variables puede llevar a un sesgo significativo, se concluye que la mejor opción es eliminar 4 variables horas\_sol, Evaporacion, nubosidad\_3pm, nubosidad\_9am.

# Datos atípicos

- 'fecha' tiene 0 (0.00%) valores atípicos
- 'temp\_min' tiene 82 (0.06%) valores atípicos
- 'temp\_max' tiene 544 (0.37%) valores atípicos
- 'precipitacion' tiene 28938 (19.89%) valores atípicos
- 'vel\_rafaga\_viento' tiene 5523 (3.80%) valores atípicos
- 'vel\_viento\_9am' tiene 1817 (1.25%) valores atípicos
- 'vel\_viento\_3pm' tiene 2523 (1.73%) valores atípicos
- 'humedad\_9am' tiene 1425 (0.98%) valores atípicos
- 'humedad\_3pm' tiene 0 (0.00%) valores atípicos
- 'presion\_9am' tiene 2758 (1.90%) valores atípicos
- 'presion\_3pm' tiene 2524 (1.74%) valores atípicos
- 'temp\_9am' tiene 307 (0.21%) valores atípicos
- 'temp\_3pm' tiene 988 (0.68%) valores atípicos
- 'lluvia\_hoy' tiene 31880 (21.92%) valores atípicos
- 'lluvia\_manana' tiene 31877 (21.91%) valores atípicos

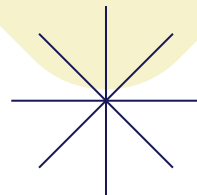
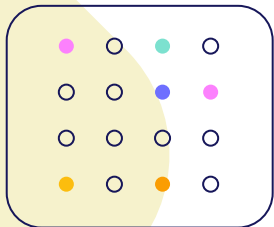




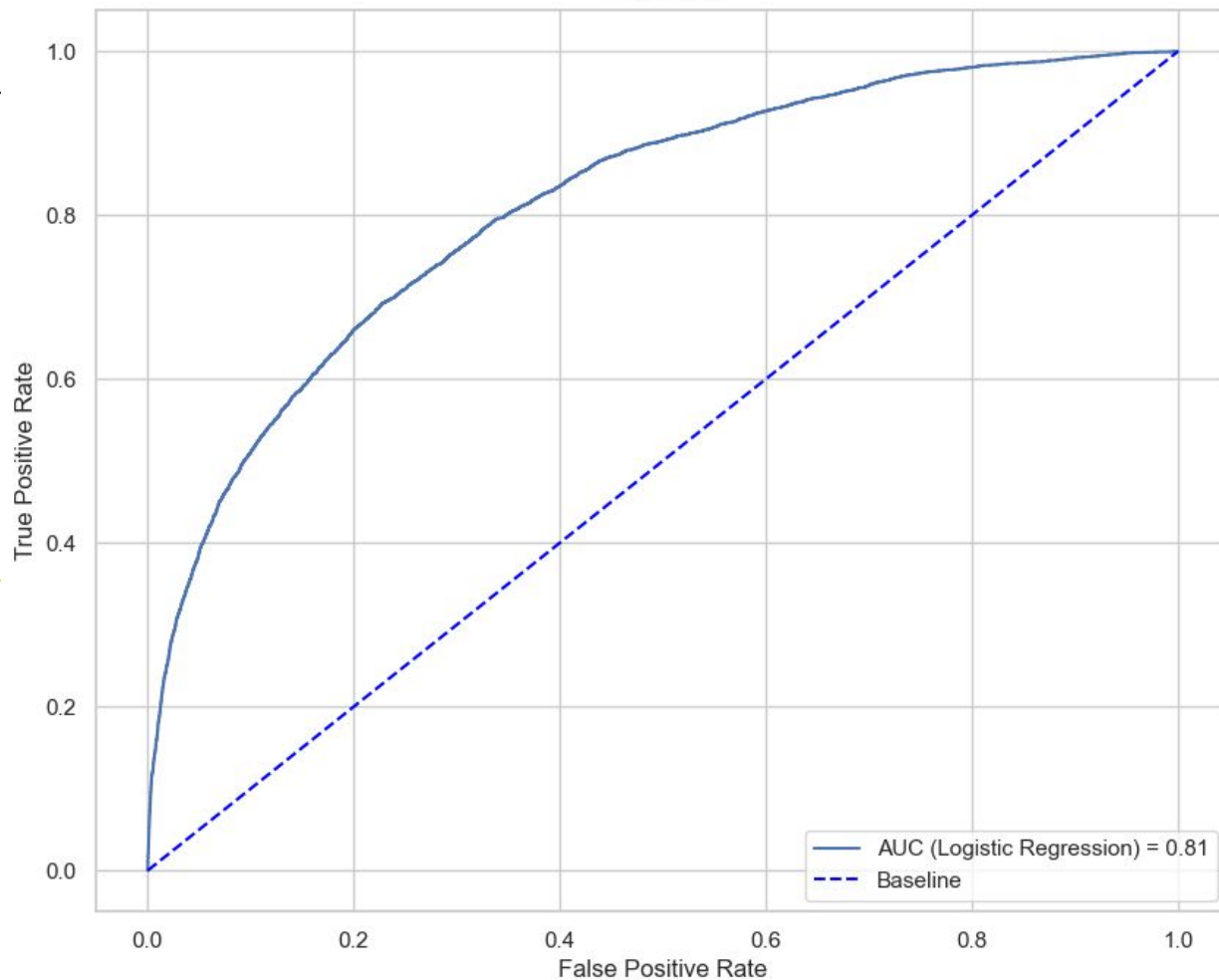


Se utilizará el método de Turkey para identificar los valores atípicos en las variables numéricas. Este método indica que un valor atípico cuando se encuentra más de 1.5 veces el rango intercuartílico por encima del tercer cuartil o por debajo del primer cuartil.

# Resultados



ROC Curve



**AUC:**  
**0.81338**

# Precisión

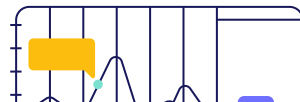
**Precisión de la prueba**

87%

**Precisión del entrenamiento:**

87%

Esta métrica indica la proporción de predicciones positivas correctas (lluvia) en relación con todas las predicciones positivas realizadas por el modelo.



Matriz de Confusión

Valores Verdaderos

No Lluvia

Lluvia

26205

540

3479

1220

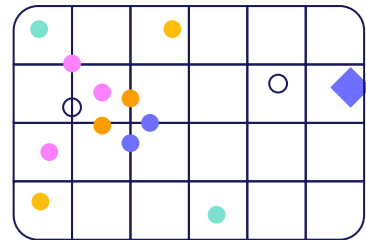
No Lluvia

Lluvia

Predicciones

**Matriz de  
Confusión**

# Interpretación de la Matriz de Confusión



## Verdaderos positivos (TP):

El modelo predijo correctamente que llovería y efectivamente llovió al día siguiente en 1,220 casos.

## Falsos positivos (FP):

El modelo predijo incorrectamente que llovería cuando en realidad no lo hizo en 540 casos.

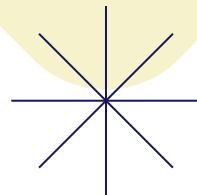
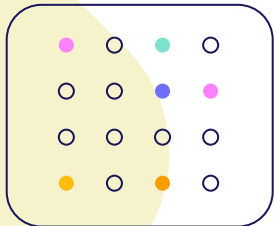
## Falsos negativos (FN):

El modelo predijo incorrectamente que no llovería cuando sí llovió al día siguiente en 3,479 casos.

## Falsos negativos (FN):

El modelo predijo correctamente que no llovería y efectivamente no llovió al día siguiente en 26,205 casos.

# Implicaciones para negocio



# Agricultura



En base a los resultados obtenidos, se puede deducir que los agricultores pueden tomar decisiones proactivas sobre cuándo regar los cultivos o protegerlos contra posibles inundaciones o erosión del suelo. Y existen ciertas regiones que son más óptimas para poder tener ciertos cultivos, como lo son las siguientes regiones.



# Regiones agrícolas

## 1. Albury

Esta se considera una región desfavorable para poder tener cultivos por las siguientes razones:

- **Alta Humedad (100% a las 9 am y 3 pm):** Puede promover la proliferación de enfermedades fúngicas, lo que dificulta el cultivo de especies sensibles a la humedad.
- **Alta Velocidad de Viento (37 km/h):** Riesgo de daño físico a los cultivos y dificultad en la aplicación de tratamientos como pesticidas.
- **Temperaturas y Presiones Variadas:** Condiciones fluctuantes que pueden afectar la previsibilidad del crecimiento de los cultivos.

## 2. Cobar

Esta se considera una región muy favorable para poder tener cultivos por las siguientes razones:

- **Baja Humedad (19% a las 9 am):** Menor riesgo de enfermedades fúngicas, favorable para cultivos que requieren ambientes más secos.
- **Presión Estable y Alta:** Menor riesgo de eventos climáticos severos que podrían afectar los cultivos.
- **Temperaturas Altas:** Adecuado para cultivos resistentes al calor, como el trigo o el sorgo (maicillo).

# Turismo



En el sector turístico, el poder adaptarse a las condiciones climáticas es crucial para poder garantizar tanto la seguridad con la satisfacción del cliente, los datos climáticos como las temperaturas extremas, presión atmosférica y velocidades del viento son de mucha importancia para empresas dedicadas al turismo y en base a los datos se pudo deducir lo siguiente.

# Regiones turísticas

## 1. Cobar

Esta se considera una región desfavorable para poder tener cultivos por las siguientes razones:

- **Temperaturas extremas de hasta 43.4°C**
- En este caso se deben Limitar las actividades al aire libre durante las horas de máximo calor. Promover actividades en interiores o en áreas con suficiente sombra y climatización.

## 2. Sydney

Esta se considera una región muy favorable para poder tener cultivos por las siguientes razones:

- **Velocidad del Viento:** Aunque las velocidades de viento no suelen ser extremas, es importante monitorizarlas especialmente para actividades como el turismo náutico
- **Temperaturas Moderadas:** La relativa moderación en las temperaturas, comparado con otras regiones más interiores, hace de Sydney un lugar ideal para actividades turísticas durante todo el año.
- Presiones atmosféricas bajas que pueden indicar clima más templado

# Seguros



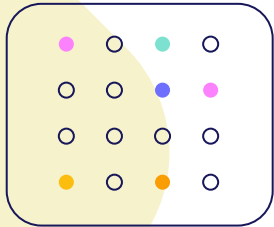
la integración de datos climáticos detallados es fundamental para evaluar y gestionar los riesgos asociados con eventos climáticos extremos. Los resultados obtenidos de un análisis minucioso de variables como temperaturas extremas, presión atmosférica, humedad y velocidad del viento proveen a las aseguradoras una base sólida para ajustar las políticas de cobertura, en base a los resultados algunas de dichas políticas pueden ser aplicadas en estas regiones:

## 1. Albury y Cobar

Son regiones con altas temperaturas y velocidades de viento bajas, por lo que son regiones con mayor riesgo de incendios y daños por tormentas.

- **Presiones Bajas:** Indicativo de posibles tormentas y eventos climáticos severos.
- **Recomendación:** Ajustar las primas de seguros y revisar las coberturas para incluir adecuadamente los riesgos asociados con estas variables.

# Conclusiones y Recomendaciones



## Rendimiento

El modelo de regresión logística tiene una precisión del 87.22% en el conjunto de prueba y del 87.39% en el conjunto de entrenamiento, indicando una buena capacidad de generalización.

El área bajo la curva ROC (AUC-ROC) es del 81.34%, lo que sugiere que el modelo es eficaz para distinguir entre las clases positiva y negativa.

## Matriz de Confusión

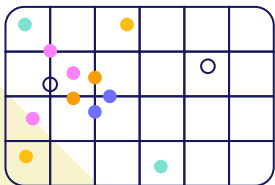
Se observa un número significativo de falsos positivos y falsos negativos, lo que indica cierta dificultad del modelo para clasificar correctamente algunas instancias.

## Recomendaciones

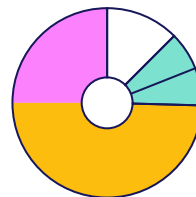
Explorar nuevas variables o técnicas de ingeniería de características para mejorar la captura de la relación entre las variables meteorológicas y la lluvia.

Investigar modelos más avanzados y analizar detalladamente los errores de clasificación para encontrar patrones comunes y áreas de mejora.

Mantener una actualización y recalibración regular del modelo para adaptarse a los cambios en los datos y condiciones ambientales.



# ¡Gracias !



CREDITS: This presentation template was created by [Slidesgo](#), and includes icons by [Flaticon](#), and infographics & images by [Freepik](#)