



# Proyecto Final Data Science

Entrega Final

## “Predicción de la Lluvia”

Autor

M. Tomás Martínez

20/12/2023

Comisión: 42390

Profesor: Jorge Ruiz.

Tutor: Aldana Ruscitti

# Índice

1. <a href="#"><u>Introducción</u></a> .....	Pag 3
a) <a href="#"><u>Abstract</u></a> .....	Pag 3
b) <a href="#"><u>Contexto Comercial</u></a> .....	Pag 4
c) <a href="#"><u>Problemas comercial</u></a> .....	Pag 4
d) <a href="#"><u>Objetivo</u></a> .....	Pag 4
e) <a href="#"><u>Contexto Analítico</u></a> .....	Pag 4
f) <a href="#"><u>Hipótesis/Preguntas</u></a> .....	Pag 6
2. <a href="#"><u>Exploratory Data Analysis (EDA) &amp; Visualization</u></a> .....	Pag 7
a) <a href="#"><u>Preparación de los datos</u></a> .....	Pag 8
b) <a href="#"><u>Gráficos de los datos</u></a> .....	Pag 10
d) <a href="#"><u>Análisis Estadístico</u></a> .....	Pag 14
e) <a href="#"><u>Obtención de Insights</u></a> .....	Pag 17
3. <a href="#"><u>Data Wrangling</u></a> .....	Pag 18
a) <a href="#"><u>Limpieza de datos</u></a> .....	Pag19
b) <a href="#"><u>Trata de Outliers</u></a> .....	Pag 22
4. <a href="#"><u>Feature Engineering</u></a> .....	Pag 24
a) <a href="#"><u>Encoding</u></a> .....	Pag 24
b) <a href="#"><u>Desbalanceo de datos</u></a> .....	Pag 26
c) <a href="#"><u>Estandarización</u></a> .....	Pag 30
d) <a href="#"><u>PCA</u></a> .....	Pag 31
e) <a href="#"><u>Modelos</u></a> .....	Pag 33
f) <a href="#"><u>Hiperparámetros</u></a> .....	Pag 38
5. <a href="#"><u>Modelo final</u></a> .....	Pag 41
a) <a href="#"><u>Base de Datos</u></a> .....	Pag 41
b) <a href="#"><u>Modelos</u></a> .....	Pag 41
c) <a href="#"><u>Mejores Modelos</u></a> .....	Pag 48
d) <a href="#"><u>Evaluación de nuestro Modelo Final</u></a> .....	Pag 52
e) <a href="#"><u>Evolución de nuestro Modelo Final</u></a> .....	Pag 55
6. <a href="#"><u>Conclusiones</u></a> .....	Pag 57
a) <a href="#"><u>Futuro del proyecto</u></a> .....	Pag 58

# Introducción



## Abstract

La lluvia protagoniza un importante lugar en la vida diaria de las personas, es así que no es solamente un recurso natural por sí mismo, sino que, en muchos sectores de la economía cumple un rol fundamental, ya sea porque muchas de las decisiones que se van a tomar son elegidas mediante la información de que, si se efectuara una lluvia o no, o cuánto va a llover.

Es decir, muchas empresas toman las decisiones o diseñan un plan de negocio teniendo en cuenta esta variable, por lo cual es de suma importancia poder tomar el control de los datos para predecir este fenómeno, que no únicamente beneficia o afecta a grandes empresas, además puede tener influencias en el día a día de medianas y micros empresas, así como también en la elección de cada individuo acerca de qué acciones realizará los días que llueve y los días que no.

## Empresa

Nos contrató una empresa del sector agropecuario de Australia "Kangaroo Crops" en la cual le resolveremos las cuestiones relacionadas con la producción, más específicamente con la lluvia y cómo ésta determina el plan de acción de la empresa.



## **Contexto Comercial**

Nos encontramos al mando del área de Data de una de las mejores empresas en el sector agropecuario de Australia "Kangaroo Crops" en donde sus mayores productos son el trigo, cebada, caña de azúcar y frutas. Por lo tanto, un estudio de la lluvia puede traer consigo numerosos beneficios para la empresa y sus decisiones, y plan de negocio que esta tenga. Nos encargaremos entonces en conjuntos con un equipo multidisciplinario de trabajo para la mejor interpretación y trato de los datos, de realizar el análisis correspondiente de los datos.

## **Problema Comercial**

A nuestro jefe le interesa saber con anterioridad que días y cuánto va a llover, para así tener un plan de acción acorde a lo que el negocio requiere para sacar la máxima ganancia posibles.

Para ello generamos un modelo predictivo para que esto sea posible dadas las características de los datos que tenemos en la base de datos de la empresa referidas al estado del tiempo, de los últimos 10 años.

## **Objetivo**

Obtener los datos más influyentes, para así manejarlos y manipularlos para poder encontrar patrones que nos brinden los datos, con el motivo de predecir las lluvias siguientes o de algún día determinado. Crear un modelo que prediga la probabilidad de que llueva.

## **Contexto Analítico**

La base de datos de la empresa cuenta con aproximadamente 10 años de observaciones meteorológicas diarias de muchos lugares de Australia. En ellas podemos apreciar 145460 Filas y 23 columnas de datos.

La Base de datos requiere una limpieza de los mismo, dado que, en algunos aparatos hay datos faltantes (Por ejemplo, Evaporación, Luz solar, etc), en algunas columnas tenemos también datos NaN la cual puede afectar a nuestro análisis y por consecuencia sería bueno tratar de hacer una limpieza de estos datos, antes de realizar nuestras primeras conclusiones.

Como parte de una base de datos que se especifica en la Lluvia, se encuentra en ellas las siguientes categorías:

**Fecha:** Registro del día/mes/año.

**Lugar:** Ciudades

**Lluvia:** Cantidad de Lluvia en milímetro (mm)

**MinTemp:** Grados (°C)

**MaxTemp:** Grados (°C)

**Evaporacion:** Cantidad en milímetros por día (mm/día)

**Luz Solar:** Cantidad de vatio por metro cuadrado (W/m<sup>2</sup>)

**Direccion Viento:** Dirección del Viento

**Veloc Viento:** Velocidad en Km/h

**Direc Viento 9am:** Dirección del Viento

**Humedad 9am:** Cantidad en g/m<sup>3</sup>

**Humedad 3pm:** Cantidad en g/m<sup>3</sup>

**Presion 9am:** Pascal (Pa)

**Presion 3m:** Pascal (Pa)

**Nubes 9am:** Cantidad de Octas (0.0/8.0)

**Nubes 3am:** Cantidad de Octas (0.0/8.0)

**Temp 9am:** Grados (°C)

**Temp 3pm:** Grados (°C)

**Lluvia Hoy:** No o Yes

**Lluvia Mañana:** No o Yes

## Hipótesis/Preguntas

Los años en los cuales se dieron las mayores cantidades de lluvias fueron los años que mayores días por encima de los 30 Grados( $^{\circ}\text{C}$ ) hubo.

¿Cuáles son los principales elementos para determinar si llueve o no

¿En las ciudades ubicadas más al sur aumenta la probabilidad de que haya lluvias más abundantes?

¿La presión esta correlacionada positivamente con la cantidad de lluvia?

¿Como ha evolucionado la lluvia a lo largo del tiempo? ¿En promedio ha llovido más en los últimos 5 años?

¿En qué estación del año es más propenso de que llueve más?

¿Existe una relación entre la dirección en la que corre el viento y la lluvia?

¿Cuáles son los días que en promedio llueve más, los que se encuentran por debajo del promedio de la temperatura o los que están por encima?

La probabilidad de que llueva dos días seguidos son mayor que el 35%

En el mes de febrero la presión es más baja que el promedio

# Exploratory Data Analysis (EDA) & Visualization



¿Qué implica el Análisis Exploratorio de Datos (EDA)?

El propósito del Análisis Exploratorio de Datos es examinar minuciosamente la información antes de aplicar cualquier método estadístico. Esto permite que un profesional en Ciencia de Datos adquiera una comprensión básica de los datos y las interrelaciones entre las variables analizadas.

¿Cuál es la función del EDA?

El EDA ofrece técnicas simples para ordenar y preparar los datos, identificar posibles deficiencias en el diseño y recolección de la información, así como para manejar y evaluar la presencia de datos faltantes, detectar casos atípicos y más.

Mediante el EDA, es posible responder diversas interrogantes, tales como:

- ✓ ¿Existen tendencias o inclinaciones notables en los datos recopilados?
- ✓ ¿Se han producido errores en la codificación de la información?
- ✓ ¿Cómo se puede resumir y presentar la información contenida en un conjunto de datos?
- ✓ ¿Existen valores atípicos u observaciones inusuales? ¿Cómo se deben abordar?
- ✓ ¿Hay información faltante en los datos? ¿Existe algún patrón sistemático en su ausencia? ¿Cómo se puede manejar esta situación?

El Análisis Exploratorio de Datos (EDA) sigue una secuencia: comienza con la recopilación y limpieza de datos, luego analiza variables individualmente y en relación unas con otras. Emplea visualizaciones para identificar patrones y relaciones clave. Este proceso iterativo permite ajustes y refinamientos, culminando en la presentación clara de los hallazgos y conclusiones importantes.

# Preparación de los datos



## Data Acquisition

Este DataSet contiene aproximadamente 10 años de observaciones meteorológicas diarias de muchos lugares de Australia.

Link de la base de datos: <https://www.kaggle.com/datasets/jsphyg/weather-dataset-rattle-package>

En la cual la misma nos aportó los siguientes datos a priori:

	Date	Location	MinTemp	MaxTemp	Rainfall	Evaporation	Sunshine	WindGustDir	WindGustSpeed	WindDir9am	WindDir3pm	Humidity9am	Humidity3pm	Pressure9am	Pressure3pm	Cloud9am	Cloud3pm
0	2008-12-01	Albury	13.4	22.9	0.6	NaN	NaN	W	44.0	W	...	71.0	22.0	1007.7	1007.1	8.0	
1	2008-12-02	Albury	7.4	25.1	0.0	NaN	NaN	NNW	44.0	NNW	...	44.0	25.0	1010.6	1007.8	NaN	
2	2008-12-03	Albury	12.9	25.7	0.0	NaN	NaN	WSW	46.0	W	...	38.0	30.0	1007.6	1008.7	NaN	
3	2008-12-04	Albury	9.2	28.0	0.0	NaN	NaN	NE	24.0	SE	...	45.0	16.0	1017.6	1012.8	NaN	
4	2008-12-05	Albury	17.5	32.3	1.0	NaN	NaN	W	41.0	ENE	...	82.0	33.0	1010.8	1006.0	7.0	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
145455	2017-06-21	Uluru	2.8	23.4	0.0	NaN	NaN	E	31.0	SE	...	51.0	24.0	1024.6	1020.3	NaN	
145456	2017-06-22	Uluru	3.6	25.3	0.0	NaN	NaN	NNW	22.0	SE	...	56.0	21.0	1023.5	1019.1	NaN	
145457	2017-06-23	Uluru	5.4	26.9	0.0	NaN	NaN	N	37.0	SE	...	53.0	24.0	1021.0	1016.8	NaN	
145458	2017-06-24	Uluru	7.8	27.0	0.0	NaN	NaN	SE	28.0	SSE	...	51.0	24.0	1019.4	1016.5	3.0	
145459	2017-06-25	Uluru	14.9	NaN	0.0	NaN	NaN	NaN	NaN	ESE	...	62.0	36.0	1020.2	1017.9	8.0	

145460 rows × 23 columns

```

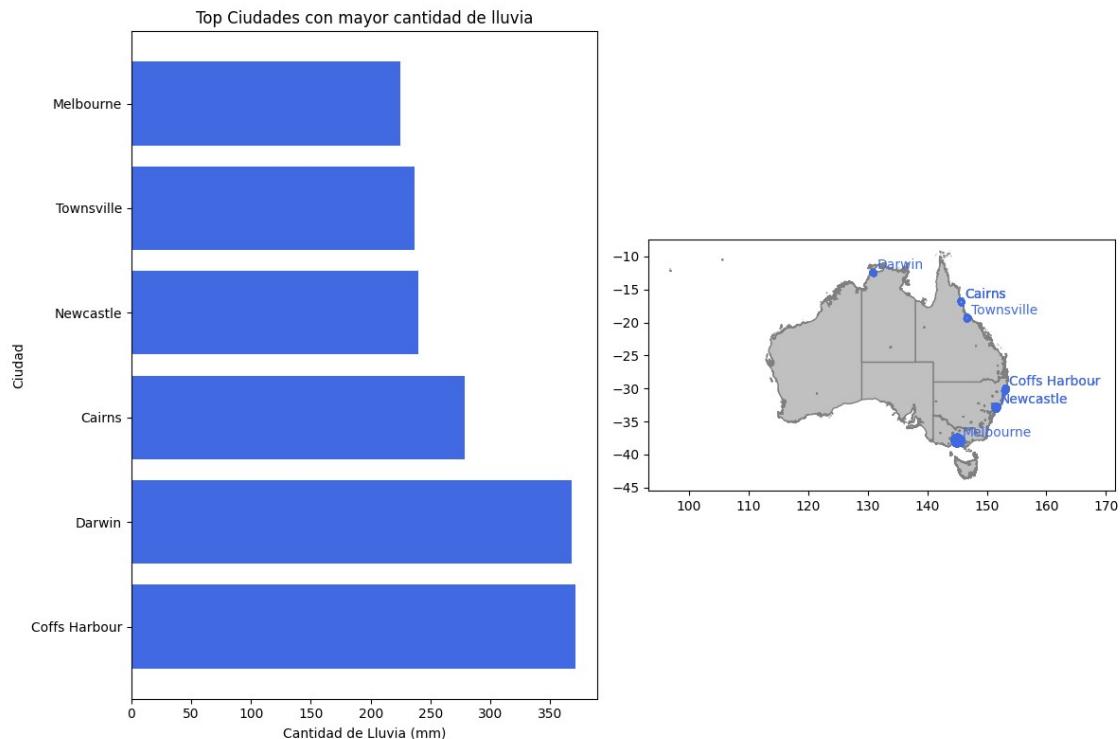
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 145460 entries, 0 to 145459
Data columns (total 23 columns):
 #   Column            Non-Null Count  Dtype  
--- 
 0   Fecha             145460 non-null   object  
 1   Lugar              145460 non-null   object  
 2   MinTemp            143975 non-null   float64 
 3   MaxTemp            144199 non-null   float64 
 4   Lluvia             142199 non-null   float64 
 5   Evaporacion        82670 non-null   float64 
 6   Luz Solar          75625 non-null   float64 
 7   Direccion Viento  135134 non-null   object  
 8   Veloc Viento       135197 non-null   float64 
 9   Direc Viento 9am   134894 non-null   object  
 10  WindDir3pm         141232 non-null   object  
 11  WindSpeed9am      143693 non-null   float64 
 12  WindSpeed3pm      142398 non-null   float64 
 13  Humedad 9am       142806 non-null   float64 
 14  Humedad 3pm       140953 non-null   float64 
 15  Presion 9am       130395 non-null   float64 
 16  Presion 3m        130432 non-null   float64 
 17  Nubes 9am          89572 non-null   float64 
 18  Nubes 3am          86102 non-null   float64 
 19  Temp 9am           143693 non-null   float64 
 20  Temp 3pm           141851 non-null   float64 
 21  Lluvia Hoy          142199 non-null   object  
 22  Lluvia Mañana      142193 non-null   object  
dtypes: float64(16), object(7)
memory usage: 25.5+ MB

```

	MinTemp	MaxTemp	Lluvia	Evaporacion	Luz Solar	Veloc Viento	WindSpeed9am	WindSpeed3pm	Humedad 9am	Humedad 3pm	Presion 9am	Presion 3m
count	143975.000000	144199.000000	142199.000000	82670.000000	75625.000000	135197.000000	143693.000000	142398.000000	142806.000000	140953.000000	130395.000000	130432.000000
mean	12.194034	23.221348	2.360918	5.468232	7.611178	40.035230	14.043426	18.662657	68.880831	51.539116	1017.64994	1015.255889
std	6.398495	7.119049	8.478060	4.193704	3.785483	13.607062	8.915375	8.809800	19.029164	20.795902	7.10653	7.037414
min	-8.500000	-4.800000	0.000000	0.000000	0.000000	6.000000	0.000000	0.000000	0.000000	0.000000	980.50000	977.100000
25%	7.600000	17.900000	0.000000	2.600000	4.800000	31.000000	7.000000	13.000000	57.000000	37.000000	1012.90000	1010.400000
50%	12.000000	22.600000	0.000000	4.800000	8.400000	39.000000	13.000000	19.000000	70.000000	52.000000	1017.60000	1015.200000
75%	16.900000	28.200000	0.800000	7.400000	10.600000	48.000000	19.000000	24.000000	83.000000	66.000000	1022.40000	1020.000000
max	33.900000	48.100000	371.000000	145.000000	14.500000	135.000000	130.000000	87.000000	100.000000	100.000000	1041.00000	1039.600000

## Gráfico de los datos

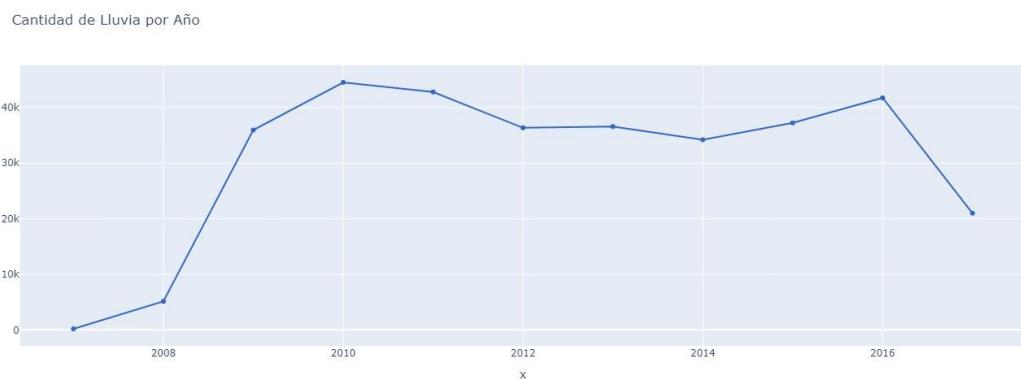
Figura 1: En este grafico se puede visualizar a simple vista el top 10 de Ciudades con más lluvias a lo largo de los 10 años que contiene el DataSet, posicionando a Cairns, Darwin y Coffs Harbour como los primeros 3



¿En las ciudades ubicadas más al sur aumenta la probabilidad de que haya lluvias más abundantes?

En este caso podemos afirmar que no es así, dado que las principales ciudades que se representan en el grafico no son ciudades que se ubiquen en el sur del país.

Figura 2: En el siguiente grafico se puede observar cómo es la progresión a través del tiempo de la cantidad de lluvia por año en Australia, esto da una primera impresión a lo que posteriormente puede ser un análisis más específico de lo que tiene que ver con la lluvia y la variable tiempo.



¿Como ha evolucionado la lluvia a lo largo del tiempo? ¿En promedio ha llovido más en los últimos 5 años?

La distribución de la lluvia en los últimos años ha tenido altos y bajos, por lo que se puede decir que en promedio se ha mantenido una equidad en las cantidades de lluvias, no se destacaron grandes etapas de bajas o altas.

Figura 3: Este grafico representa la relación de la humedad con la lluvia en donde se observa que existe una relación positiva entre el aumento de la humedad con la lluvia. Por lo tanto, entre más humedad también aumenta la posibilidad de que llueva, y de que lo haga de mayor cantidad.

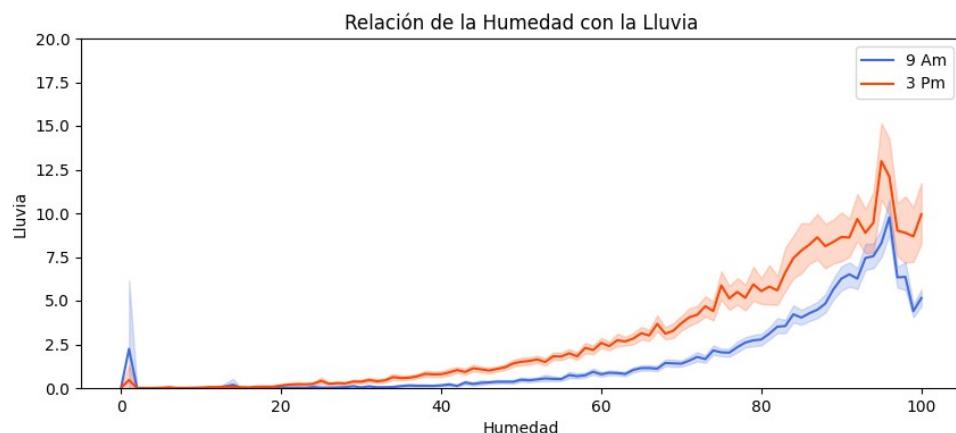
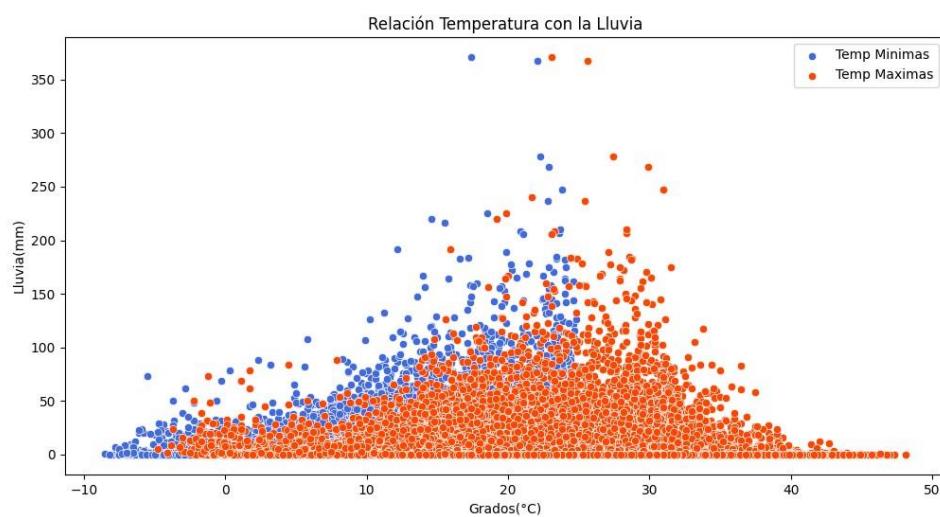


Figura 4: En este grafico se concatenaron los datos de la temperatura mínima con la temperatura máxima, para tener la mayor amplitud de la temperatura para luego realizar un análisis, observando en que rango de temperatura se observa la mayor cantidad de lluvia, y sus puntos más altos para determinar si puede existir cierto vínculo entre grados y lluvia.

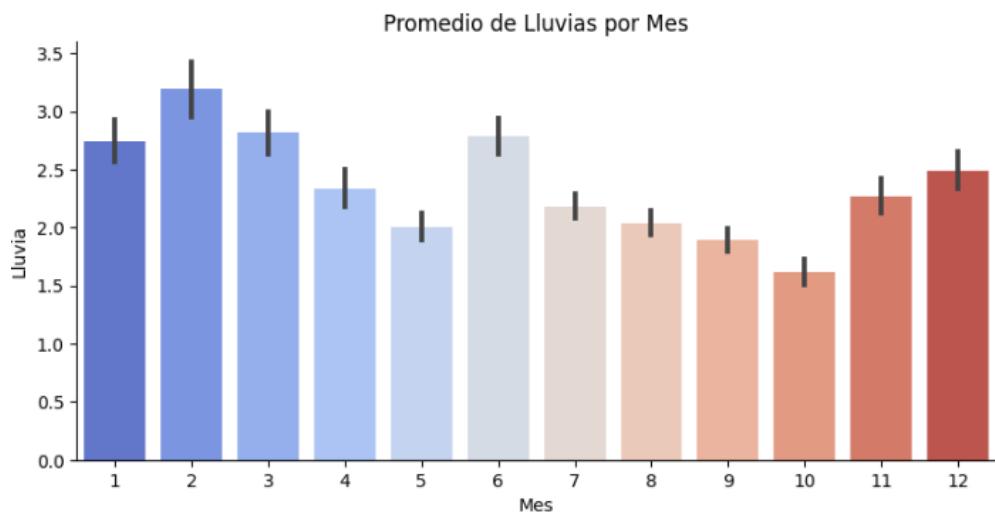


### ¿Cuáles son los principales elementos para determinar si llueve o no?

Incluyendo los dos últimos gráficos podemos apreciar que entre más humedad se alcanza también aumenta la posibilidad de que llueva, y de que lo haga de mayor cantidad.

También se puede observar que la mayor cantidad de lluvia, así como también los picos se dan dentro del rango de 18°C y 28°C. Por lo tanto estos datos pueden no ser los principales elementos para determinar la lluvia, pero podemos determinar que influyen de manera notoria en ella, por ejemplo, podemos determinar que un día que cuenta con 23°C y 80% de humedad es muy propenso a llover.

Figura 5: En este gráfico de barras se ven los meses en orden numérico con el promedio de lluvia en los 10 años de datos, en donde se aprecia que febrero es el mes en donde se han dado las mayores lluvias, y es octubre en donde se observan el promedio más bajo.



### ¿En qué estación del año es más propenso de que llueve más?

La estación del año en la cual se genera más cantidades de lluvias es en el verano, en donde se pueden ver la agrupación de meses con mayores picos de cantidad lluvia.

Figura 6: Grafico en el cual se muestran las 10 principales ciudades con mayor promedio de humedad en Australia en los registros de los 10 años, este grafico tiene el objetivo de comprobar inicialmente unas de las hipótesis que se planteó.

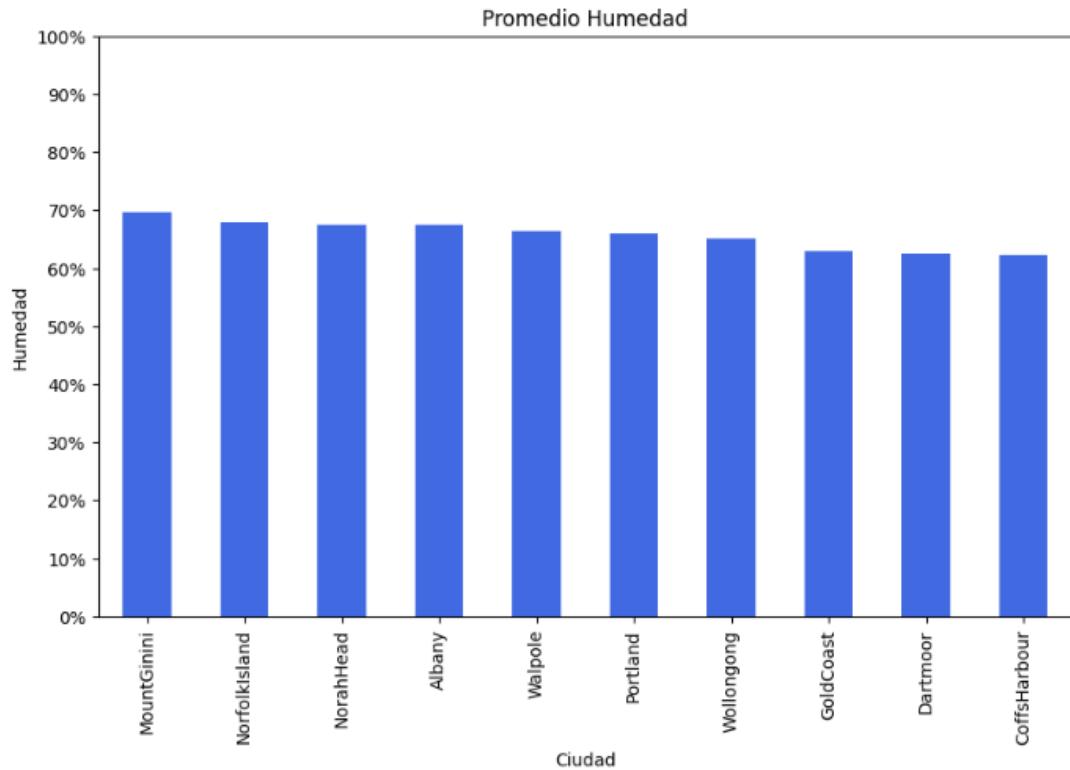
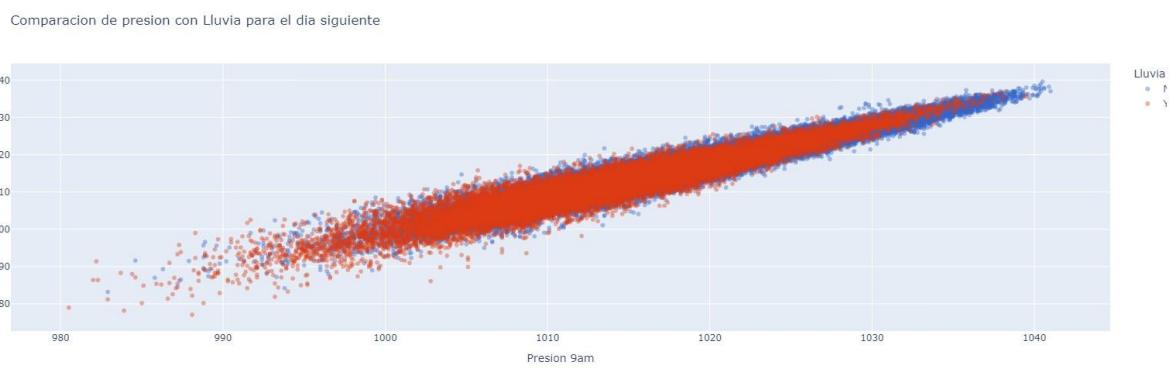


Figura 7: Gráfico en el cual se observan los datos de la Presión en distintos horarios (9am,3pm) con relación a los datos de lluvia para el siguiente día y sus respectivos datos ("si" o "no"). Aquí se muestra como fue la presión los días que llovieron como los que no, en el cual se observa a grandes rasgos una similitud entre ambas variables.



## Análisis Estadístico

- Los años en los cuales se dieron las mayores cantidades de lluvias fueron los años que mayores días por encima de los 30 Grados(°C) hubo.

```
Cantidad de días por encima de 30°C para los años 2010, 2011, 2009 y 2014:  
Año  
2009    3225  
2010    2557  
2011    2347  
2014    3607  
Name: Fecha, dtype: int64
```

Podemos concluir que la hipótesis previamente planteada no se cumple como se determinó, dado que en la realidad se da de la manera contraria.

- ¿Cuáles son los días que en promedio llueve más, los que se encuentran por debajo del promedio de la temperatura o los que están por encima?

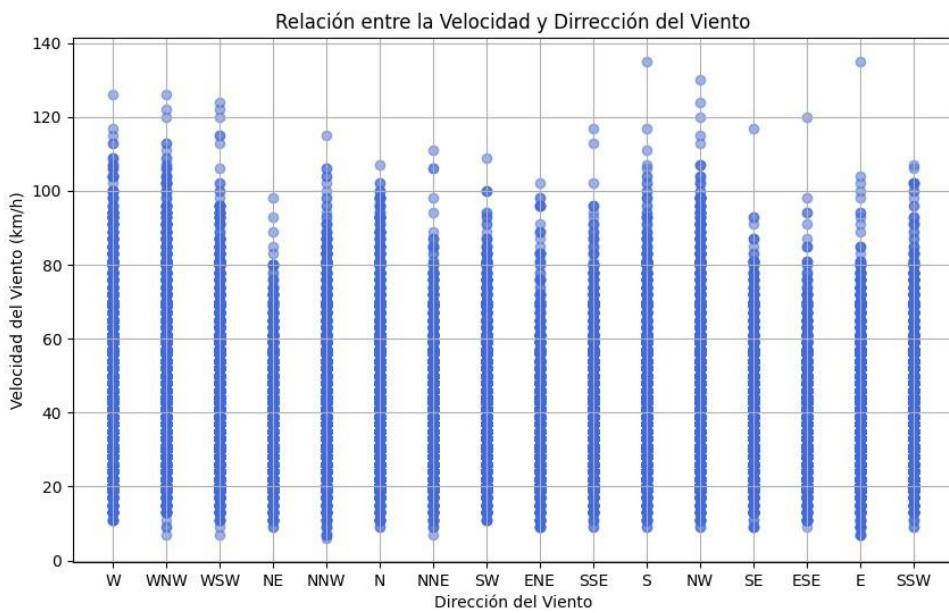
```
Días de lluvias por debajo del promedio de temperatura: 2.782648456531289  
Días de lluvias por encima del promedio de temperatura: 1.8781641293577156
```

Los días que llueven más en promedio son los que están por debajo del promedio de la temperatura

- En el mes de febrero la presión es más baja que el promedio

```
febrero = Lluvias[Lluvias['Mes'] == 2] #Se identifica el mes de febrero para luego hacer el análisis  
  
# Promedio de la presión atmosférica en Febrero  
promedio_presion_febrero = febrero['Presion 9am'].mean()  
  
# Promedio de la presión de todo el DataSet  
promedio_presion_total = Lluvias['Presion 9am'].mean()  
  
if promedio_presion_febrero < promedio_presion_total:  
    print("En el mes de Febrero, la presión es más baja que el promedio.")  
else:  
    print("En el mes de Febrero, la presión no es más baja que el promedio.")
```

En el mes de febrero, la presión es más baja que el promedio.



A primera vista los datos que más pueden captar nuestra atención son los "picos" que existen en la velocidad de viento alcanzada por las distintas direcciones del viento en donde se encuentran las direcciones "S" "NW" "E", pero si analizamos más a fondo lo que se percibe es que hay más vientos por encima de la media en la dirección "W" y "WNW" por lo que se puede decir que aunque no tengan los picos de vientos, los vientos que se producen en esas direcciones estadísticamente son mayormente más rápidos que los que tienen los "picos". A su vez también se destaca que las direcciones en las que se produce menos velocidad en el viento son "NE" y "SE".

- "Lluvia hoy" con "Lluvia Mañana" - **Bivariado - El test de Chi Cuadrado**

Contexto del Problema: El sector de la empresa que se especializa en la producción del trigo, quiere saber si la probabilidad de que llueva dos días seguidos es mayor al 35% para determinar aspectos de producción

HO - No hay relación entre los días y las lluvias HA - Existe relación entre los días y las lluvias

```
# Realizar la prueba de chi-cuadrado
p_valor = chi2_contingency(tabla_contingencia_observados)[1]

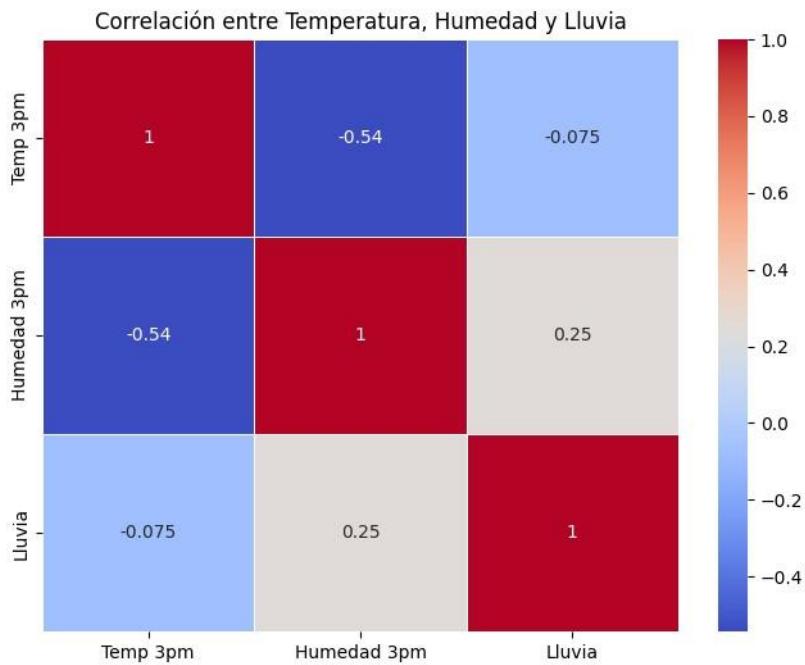
# Redondear el valor de p-valor a 4 decimales
p_valor = round(p_valor, 4)

# Imprimir los resultados
if p_valor < 0.05:
    print(f"el Valor p es de {p_valor:.4f} y se concluye que hay relación entre las variables")
else:
    print(f"el Valor p es de {p_valor:.4f} y se concluye que NO hay relación entre las variables")
```

el Valor p es de 0.0000 y se concluye que hay relación entre las variables

Se determina que si existe un porcentaje mayor al 35% de que llueva al día siguiente de haberse efectuado una lluvia en ese mismo día.

- Temperatura con humedad y Lluvia -**Multivariado**



Se puede apreciar que existe una fuerte correlación entre humedad y lluvia (0.081), no así el caso entre temperatura y lluvia, que es del -0.042 lo que podemos decir que esa relación es débil y por lo tanto entre ellas no hay una relación (el aumento o disminución de una no va a impactar en la otra), lo mismo pasa en el caso de humedad y temperatura en donde su correlación es de -0.56.



## Obtención de Insights



1. Las ciudades con más cantidad de lluvia son Coffs Harbour, Darwin y Cairns
2. A lo largo de estos 10 años los picos de lluvia se dieron en los años 2010, 2011 y 2016
3. Existe una relación positiva entre el aumento de la Humedad con la Lluvia
4. La mayor cantidad de lluvia, así como también los picos se dan dentro del rango de 18°C y 28°C.
5. La estación del año en la cual se genera más cantidades de lluvias es en el verano
  6. En promedio Febrero es el mes en el que más llueve
7. Los vientos que se producen en las direcciones "W" y "WNW" estadísticamente son mayormente más rápidos
8. La ciudad de Mount Ginini es en promedio la ciudad con mayor Humedad
  9. En el mes de Febrero, la presión es más baja que el promedio
10. Los días que llueven más en promedio son los que están por debajo del promedio de la temperatura

# Data Wrangling



El Data Wrangling es un proceso crucial que implica la transformación, limpieza y preparación de conjuntos de datos complejos para facilitar su análisis y uso en modelos. Este procedimiento abarca desde la corrección de errores hasta la unificación de datos dispersos o desordenados, convirtiéndolos en un formato más coherente y apto para su exploración.

Dentro del Data Wrangling, hay diversas tareas que abordan los desafíos inherentes a los datos desordenados. Estas pueden incluir la fusión de múltiples fuentes de datos para crear un único conjunto que sea coherente y comprensible. También implica identificar y tratar los vacíos o valores faltantes en los conjuntos de datos, ya sea llenándolos con información relevante o eliminándolos para preservar la integridad del análisis.

Otra tarea común es la depuración de datos, que consiste en identificar y eliminar datos irrelevantes o inexactos que podrían afectar la calidad del análisis. Además, se enfoca en la detección y manejo de valores atípicos o extremos que podrían distorsionar los resultados, ya sea corrigiéndolos o eliminándolos para mantener la precisión de los análisis posteriores.

El proceso de Data Wrangling puede ser llevado a cabo manual o automáticamente, dependiendo del tamaño de la organización y sus recursos. En empresas con equipos de datos dedicados, los científicos de datos son los responsables de este proceso. En entornos más pequeños, individuos sin especialización en datos suelen asumir esta tarea, lo que puede llevar a desafíos adicionales debido a la falta de experiencia.

La importancia del Data Wrangling radica en la garantía de que los datos estén en condiciones óptimas antes de ser utilizados en análisis posteriores. Si los datos no son limpiados y preparados adecuadamente, los análisis realizados posteriormente podrían verse comprometidos, limitando la fiabilidad y utilidad de los resultados obtenidos. Por lo tanto, este proceso se vuelve fundamental para asegurar que los análisis se basen en datos precisos y confiables, lo que a su vez aumenta la calidad y el valor de las decisiones empresariales o conclusiones científicas que se deriven de ellos.

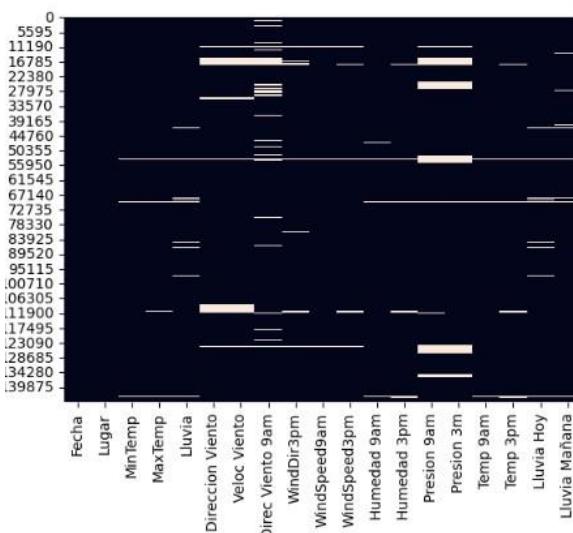
## Inicio del Proceso

```
Fecha          0
Lugar          0
MinTemp       1485
MaxTemp       1261
Lluvia         3261
Evaporacion   62790
Luz Solar     69835
Direccion Viento 10326
Veloc Viento   10263
Direc Viento 9am 10566
WindDir3pm    4228
WindSpeed9am  1767
WindSpeed3pm  3062
Humedad 9am   2654
Humedad 3pm   4507
Presion 9am   15065
Presion 3m    15028
Nubes 9am     55888
Nubes 3am     59358
Temp 9am      1767
Temp 3pm      3609
Lluvia Hoy    3261
Lluvia Mañana 3267
Año            0
Mes            0
dtype: int64
```

### Paso 1

En 1er lugar se realizará la eliminación de cuatro columnas las cuales tienen muchos datos faltantes, y en donde sería una mala práctica llegar una conclusión con estos datos.

Como primera medida de limpieza eliminaremos de forma permanente las columnas de Evaporación, Luz Solar, Nubes 9m y Nubes 3am dado que son columnas que tienen muchos datos NaN. Se eliminará también las columnas de Año y Mes, ya que fueron creadas para utilizarse en un solo gráfico, generadas por la columna Fecha.



Las marcas blancas representan los valores nulos. Mediante este gráfico es más fácil encontrar patrones y vínculos existentes entre los missing values en las diferentes variables.

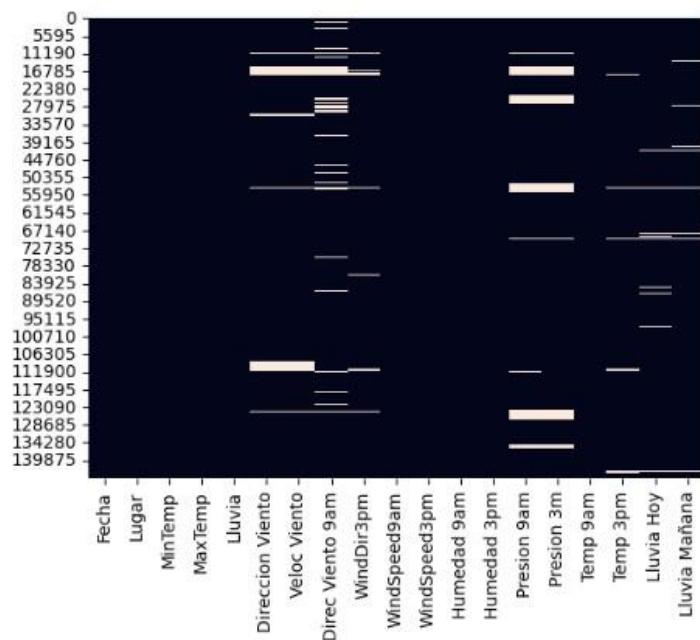
## Paso 2

### Datos Ausentes

En este caso se decidió, ya que hay pocos datos ausentes realizar la acción de Interpolación por el promedio, lo que esto quiere decir es que se toman los dos valores continuos al dato ausente y se le imputa a este dato ausente el promedio de dichos valores. Se escogió esta metodología dado que al faltar menos de 5000 datos en estas columnas nos estaríamos refiriendo a menos del 3.438% de la cantidad de datos que contiene el Data Set.

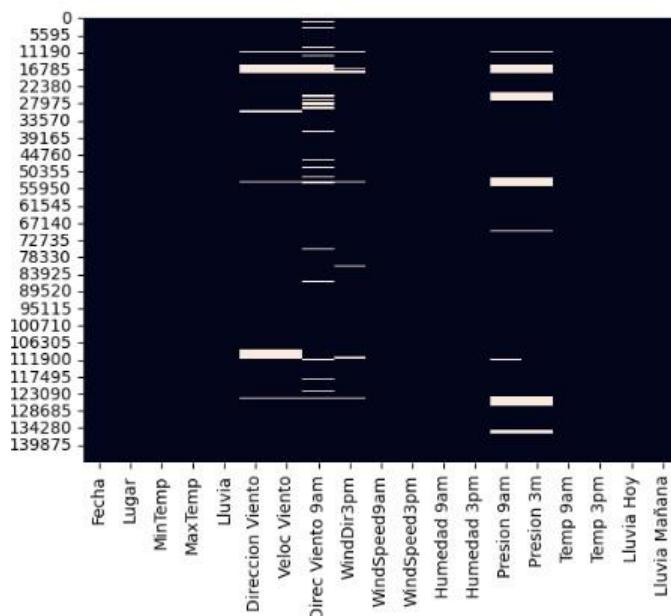
Estas columnas son:

- MinTemp 1485 (Datos Faltantes)
- MaxTemp 1261 (Datos Faltantes)
- Lluvia 3261 (Datos Faltantes)
- WindSpeed9am 1767 (Datos Faltantes)
- WindSpeed3pm 3062 (Datos Faltantes)
- Humedad 9am 2654 (Datos Faltantes)
- Humedad 3pm 4507 (Datos Faltantes)
- Temp 9am 1767 (Datos Faltantes)
- Temp 3pm 3609 (Datos Faltantes)



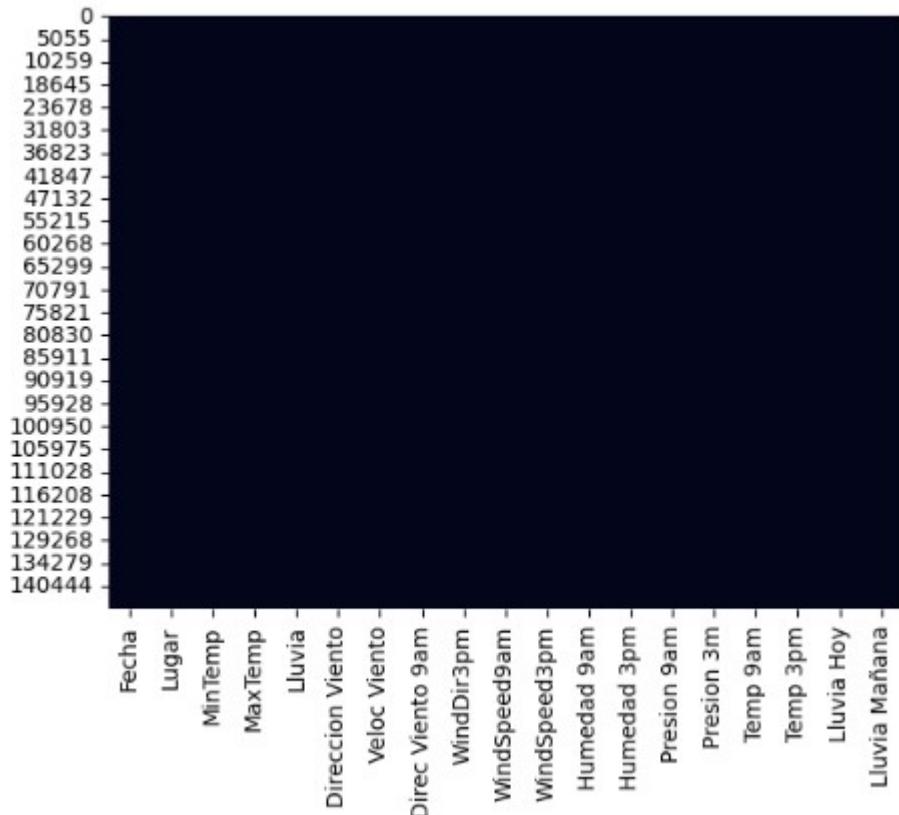
### Paso 3

Dado que no funciona el tratamiento mediante el promedio de los datos ausentes en algunas columnas, se realizará para las columnas de "Lluvia Hoy" y "Lluvia Mañana" la imputación por variable, pues es un método simple pero también efectivo.



### Paso 4

- Remplazamos los datos nulos por la moda en WindDir3pm
- Utilizaremos para la columna de Veloc Viento la mediana para remplazar los datos nulos
- Y para Dirección Viento y Direc Viento 9am usaremos la moda
- Para las columnas de Presión se eliminarán las filas que contienen esos datos nulos, ya que son más del 10% de los datos y remplazarlos no nos garantiza una buena fiabilidad del DataSet



Dataset limpio después de todas las transformaciones

## Outliers

Al examinar una variable numérica, es probable que encontremos una distribución específica de los datos. Es común observar que los datos se agrupen alrededor de un valor medio, aunque también puedan mostrar cierto grado de dispersión. Sin embargo, existen situaciones en las que ciertos valores se encuentran notablemente distantes del conjunto principal de datos. Esto puede indicar una situación excepcional o simplemente reflejar un error en la recolección de datos. Es crucial identificar estos valores extremos, conocidos como outliers, ya que su presencia puede ser significativa. Una vez detectados, se deben tomar medidas apropiadas para manejarlos de manera adecuada según las circunstancias.

Frecuentemente, la gestión de los valores atípicos involucra dos actividades distintas: en primer lugar, identificar y separar los puntos que realmente se ajustan a la definición de valores atípicos, y en segundo lugar, comprender qué significan esos valores atípicos en relación con el resto de los datos. Esta lección se enfoca principalmente en la primera actividad, utilizando herramientas estadísticas y visuales para esta tarea. Sin embargo, también abordaremos brevemente la interpretación de los valores atípicos.

Esta segunda actividad tiende a ser más subjetiva y requiere la intervención de un experto que tenga conocimiento en el campo específico de los datos analizados. En todos los casos, es esencial analizar el valor extremo y determinar si se debe considerar, ajustar manualmente o eliminar. Además, es crucial informar y analizar los valores atípicos con la asistencia de un experto en el campo de los datos, independientemente de la decisión tomada.

#### **Conclusiones de mis datos:**

Para el caso de los Outliers que se encuentran en nuestro DataSet no realizaremos ninguna medida ya sea de imputación o eliminación de los datos atípicos, puesto que en el contexto en el que se ubican los datos, los mismos son reflejo de lo que realmente sucedió en la realidad, por lo tanto, al no ser errores de diferentes motivos, sino que datos que son verídicos, se los tomará como tal y se les otorgará la misma importancia que cualquier otro dato.

# Feature Engineering



Un modelo de machine learning (aprendizaje automático) es un conjunto de algoritmos y parámetros que se entrena utilizando datos para realizar una tarea específica sin una programación explícita. Estos modelos se utilizan para hacer predicciones, tomar decisiones, clasificar datos o realizar tareas similares basadas en patrones y relaciones extraídos de los datos de entrenamiento.

Los modelos de machine learning son un componente esencial en muchas aplicaciones modernas de la inteligencia artificial y la ciencia de datos.

## Encoding

12  
34

Como primera medida en el desarrollo del modelo vamos a realizar un proceso de encoding: se refiere al proceso de convertir datos de un formato a otro, generalmente para que sean compatibles con un determinado tipo de análisis o modelo. En particular, se utiliza comúnmente en el contexto de convertir variables categóricas (como nombres, etiquetas o categorías) en una forma numérica que los algoritmos de machine learning puedan entender y procesar eficientemente.

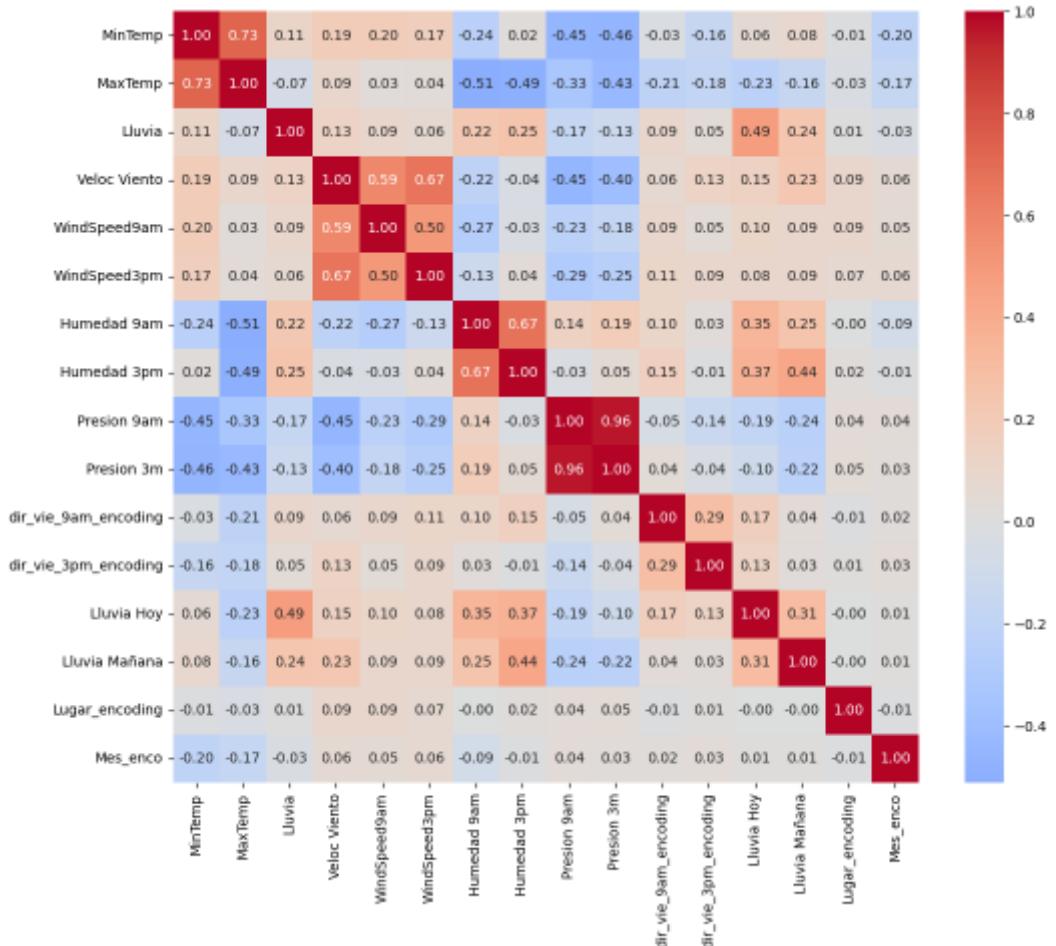
Existen diferentes métodos de encoding, como el "One-Hot Encoding", que convierte cada categoría en una nueva columna binaria, asignando un valor de 1 o 0 según la presencia o ausencia de esa categoría en una observación. Otro método es el "Label Encoding", que asigna un número único a cada categoría, transformando los datos en valores numéricos.

## Mi Dataset

- Dado que las columnas de Dirección de Viento generan gran número de categorías utilizaremos para ellos el método de Encoding Label para este caso. Sin embargo, es importante tener en cuenta que la codificación de etiquetas introduce un orden arbitrario de las categorías, que puede no reflejar necesariamente ninguna relación significativa entre ellas. En algunos casos, esto puede generar problemas en el análisis, especialmente si se interpreta que el ordenamiento tiene algún tipo de relación ordinal.
- Para las columnas de Lluvias hoy y mañana también utilizaremos el método One-Hot Encoding
- Vamos a utilizar el Método Label Encoding para la columna de "Lugar"

- Se procederá ahora a realizar el Encoding de un dato de tipo Date en el cual primero sacaremos los meses, para después si realizar el proceso de encoding.
- Se eliminan las columnas originales dado que, mantener las columnas originales y codificadas podría introducir redundancia y complicar el análisis. Las columnas codificadas ya contienen la información de las columnas originales de manera binaria. Además, podrías introducir multicolinealidad (alta correlación entre características) en los datos, lo que podría afectar negativamente ciertos algoritmos de aprendizaje automático.

### Primera Matriz de Correlación



Se puede apreciar que las relaciones más fuerte que nuestra matriz de correlación grafica son las que tienen que ver con la lluvia, humedad y temperaturas, algunas veces de manera positiva es decir por encima de 0 (lo que quiere decir que cuando una aumenta la otra también) y otras de manera negativa como lo hace con presión y y temperatura o humedad y temperatura(Una de esas dos variables baja mientras la otra sube)

## Desbalanceo de los Datos

En los problemas de clasificación en donde tenemos que etiquetar solemos encontrar que en nuestro conjunto de datos de entrenamiento contamos con que alguna de las clases de muestra es una clase “minoritaria” es decir, de la cual tenemos muy poquitas muestras.

Esto provoca un desbalanceo en los datos que utilizaremos para el entrenamiento de nuestra máquina y por lo general afecta a los algoritmos en su proceso de generalización de la información y perjudicando a las clases minoritarias.

En este caso tenemos más de una técnica la cual podemos usar para balancear nuestros datos por lo que, considerar y aplicar mas de una herramienta y ver sus resultados es lo que va a generar después que nuestro modelo performe mejor, pues puede existir diferencias en los resultados de lo que obtengamos con cada unas de estas técnicas. Por lo tanto, para nuestro Dataset utilizaremos las siguientes técnicas:

### SMOTE

SMOTE (Synthetic Minority Over-sampling Technique) es una técnica de remuestreo. Está diseñada específicamente para abordar el desequilibrio de clases al incrementar el número de muestras en la clase minoritaria generando muestras sintéticas.

Esta técnica puede ayudar a mejorar el rendimiento de los modelos al proporcionarles más ejemplos de la clase minoritaria para aprender y reducir así el sesgo hacia la clase mayoritaria.

Resultados:

	precision	recall	f1-score	support
0	0.98	0.98	0.98	20295
1	0.94	0.95	0.94	5740
accuracy			0.98	26035
macro avg	0.96	0.96	0.96	26035
weighted avg	0.98	0.98	0.98	26035

#### A tener en cuenta:

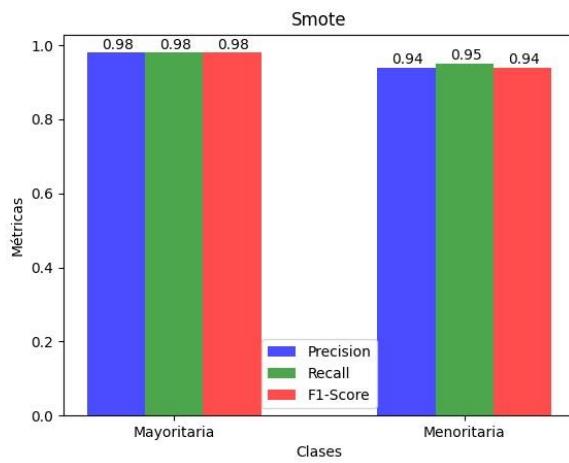
**El accuracy** se refiere a lo cerca que está el resultado de una medición del valor verdadero. En forma práctica la Exactitud es el % total de elementos clasificados correctamente.

**La precision** se refiere a la dispersión del conjunto de valores obtenidos a partir de mediciones repetidas de una magnitud. Cuanto menor es la dispersión mayor la precisión.

**Recall o Sensibilidad** es la proporción de casos positivos que fueron correctamente identificadas por el algoritmo.

**El F1-Score** es otra métrica muy empleada porque nos resume la Precisión (Precisión) y Sensibilidad (Recall) en una sola métrica.

El **support** trata de los casos negativos que el algoritmo ha clasificado correctamente. Expresa cuán bien puede el modelo detectar esa clase.

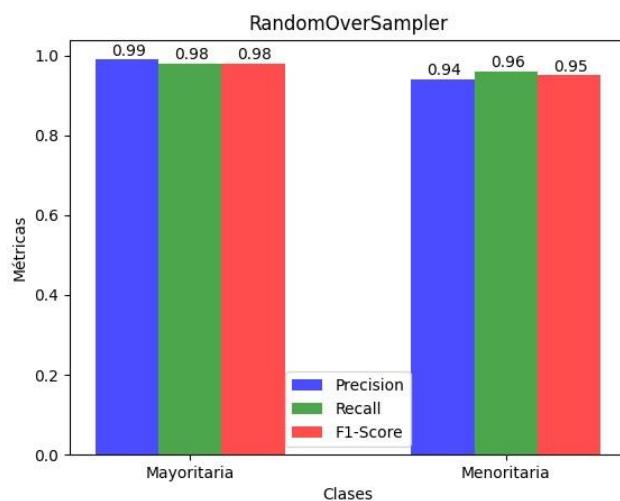


## RandomOverSampler

Le enseño a mi modelo datos repetidos hasta llegar a balancear la clase minoritaria o con la mayoritaria.

Como ventaja de esta técnica es que los mismos datos que ingresan no son datos sintetizados como en SMOTE La desventaja es que todos los datos que me generan son repetidos y para cualquier modelo esto no sería lo óptimo.

	precision	recall	f1-score	support
0	0.99	0.98	0.98	20295
1	0.94	0.96	0.95	5740
accuracy			0.98	26035
macro avg	0.96	0.97	0.97	26035
weighted avg	0.98	0.98	0.98	26035

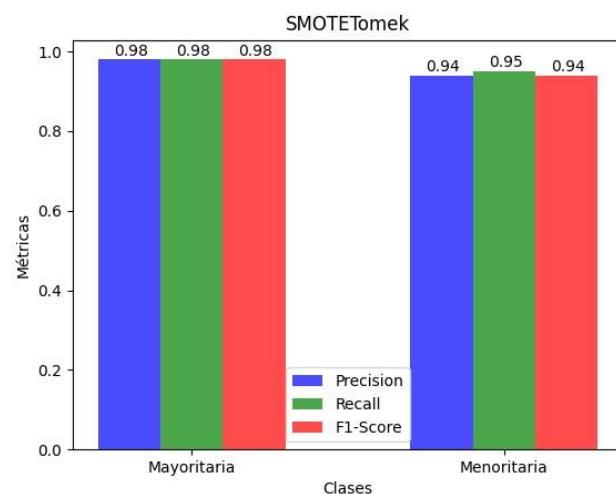


## SMOTETomek

SMOTETomek es una técnica de remuestreo que combina dos métodos: SMOTE (Synthetic Minority Over-sampling Technique) y Tomek Links.

La combinación de estos dos métodos, SMOTE y Tomek Links, como SMOTETomek, busca abordar el desequilibrio de clases al sobre muestrear la clase minoritaria mediante SMOTE y luego limpiar o eliminar muestras cercanas y potencialmente problemáticas utilizando los enlaces de Tomek.

	precision	recall	f1-score	support
0	0.98	0.98	0.98	20295
1	0.94	0.95	0.94	5740
accuracy			0.98	26035
macro avg	0.96	0.96	0.96	26035
weighted avg	0.98	0.98	0.98	26035

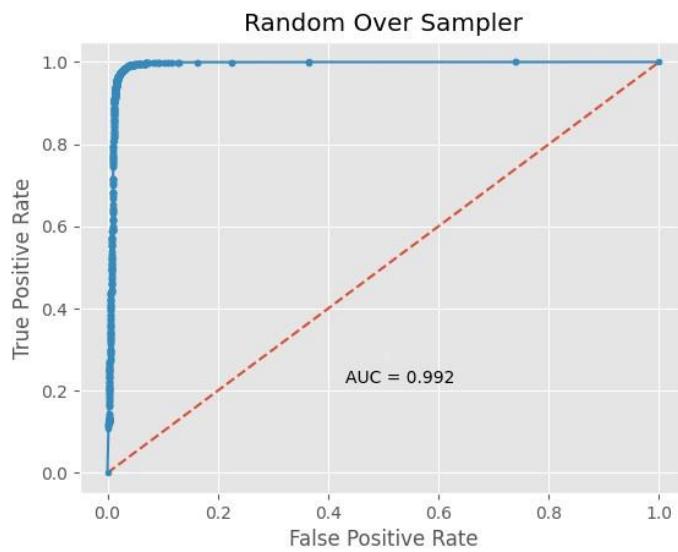
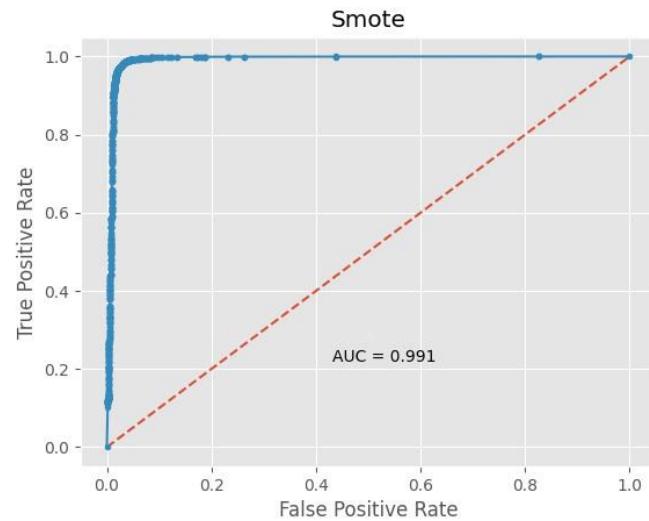


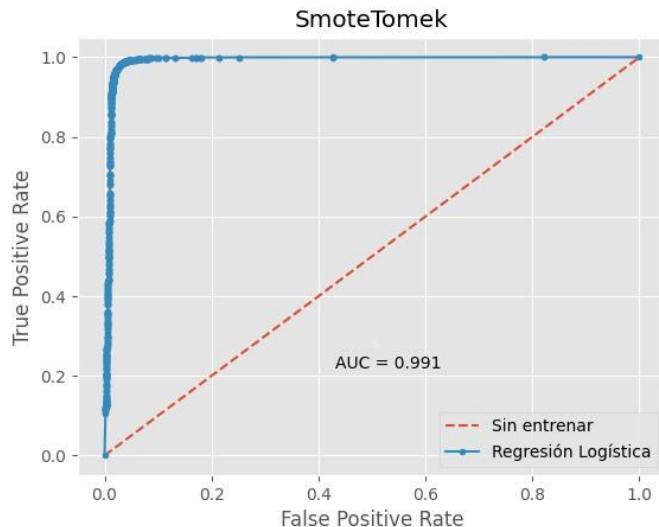
## Curva ROC y AUC

La Curva ROC y el AUC son útiles para evaluar y comparar el rendimiento de modelos de clasificación, especialmente cuando se trabaja con conjuntos de datos desbalanceados o cuando no se quiere depender de un único punto de corte para evaluar el modelo. Estas métricas brindan información sobre el rendimiento del modelo en diferentes umbrales de clasificación.

Cuanto más cerca esté la curva ROC del vértice superior izquierdo, mejor será el rendimiento del modelo, ya que indica una alta TPR y una baja FPR.

Varía entre 0 y 1, donde un valor de 1 indica un modelo perfecto que clasifica perfectamente los puntos positivos y negativos. Cuanto mayor sea el AUC, mejor será la capacidad del modelo para distinguir entre clases.





### Conclusión:

Las evaluaciones muestran métricas bastante similares en términos de precision, recall y f1-score para las clases 0 y 1, en los tres primeros métodos de balanceo (SMOTE, RandomOverSampler y SMOTETomek). Estos métodos parecen haber generado modelos con un rendimiento equilibrado y consistente en la clasificación de ambas clases.

Sin embargo, la evaluación mediante la curva ROC y el área bajo la curva (AUC) muestra una ligera variación en el rendimiento. El modelo que contiene los datos después de realizar el Random Over Sampler, obtuvo una mejor puntuación, pero no con grandes diferencias solo 1 punto por encima de los demás modelos.

Si no hay diferencias significativas en cuanto al rendimiento entre los tres primeros métodos y ninguno sobresale en otros aspectos, se puede optar por el método que sea más fácil de implementar o que tenga mejor eficiencia computacional.

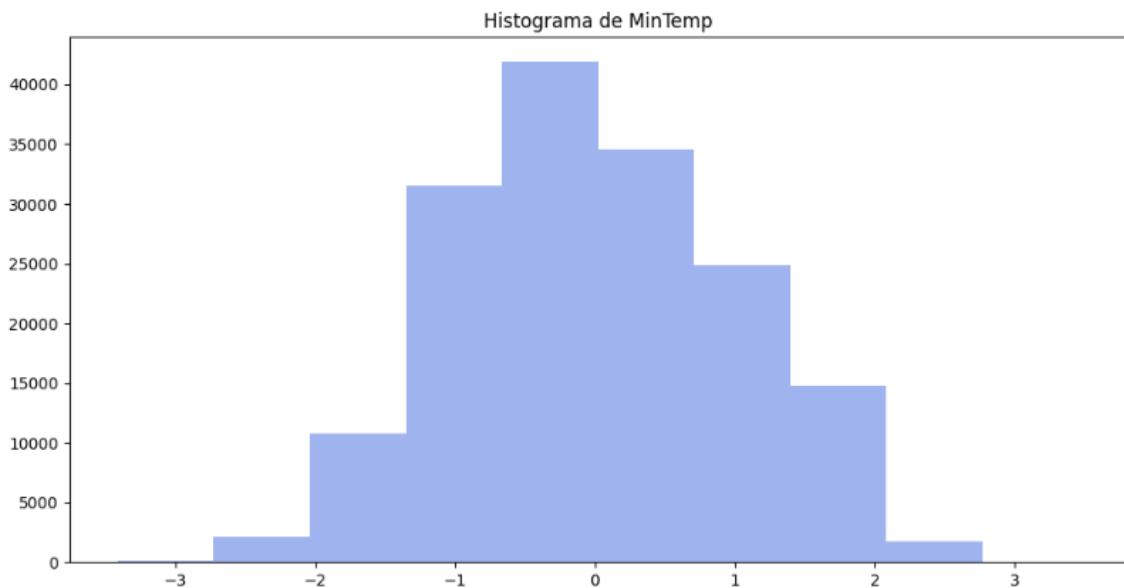
## Estandarización

Este proceso es comúnmente utilizado en técnicas de modelado estadístico y de machine learning, ya que algunos algoritmos pueden ser sensibles a la escala de las características. Al estandarizar, se asegura que todas las características contribuyan de manera equitativa al modelo y que sus valores estén en una escala comparable, evitando así que características con rangos o escalas muy diferentes dominen el proceso de entrenamiento del modelo.

Es proceso de transformar las características de un conjunto de datos para que tengan una media de cero y una desviación estándar de uno. Esto se logra restando la media de cada característica y dividiendo por su desviación estándar.

En este caso utilizaremos el **StandardScaler** para nuestros Dataset. En términos prácticos, StandardScaler aplica la transformación a cada característica, restando su media y dividiendo por su desviación estándar. Esto se hace para cada variable por separado, sin considerar la relación entre las características.

Ejemplo: Unos de los gráficos que se obtuvo posterior al StandarScaler

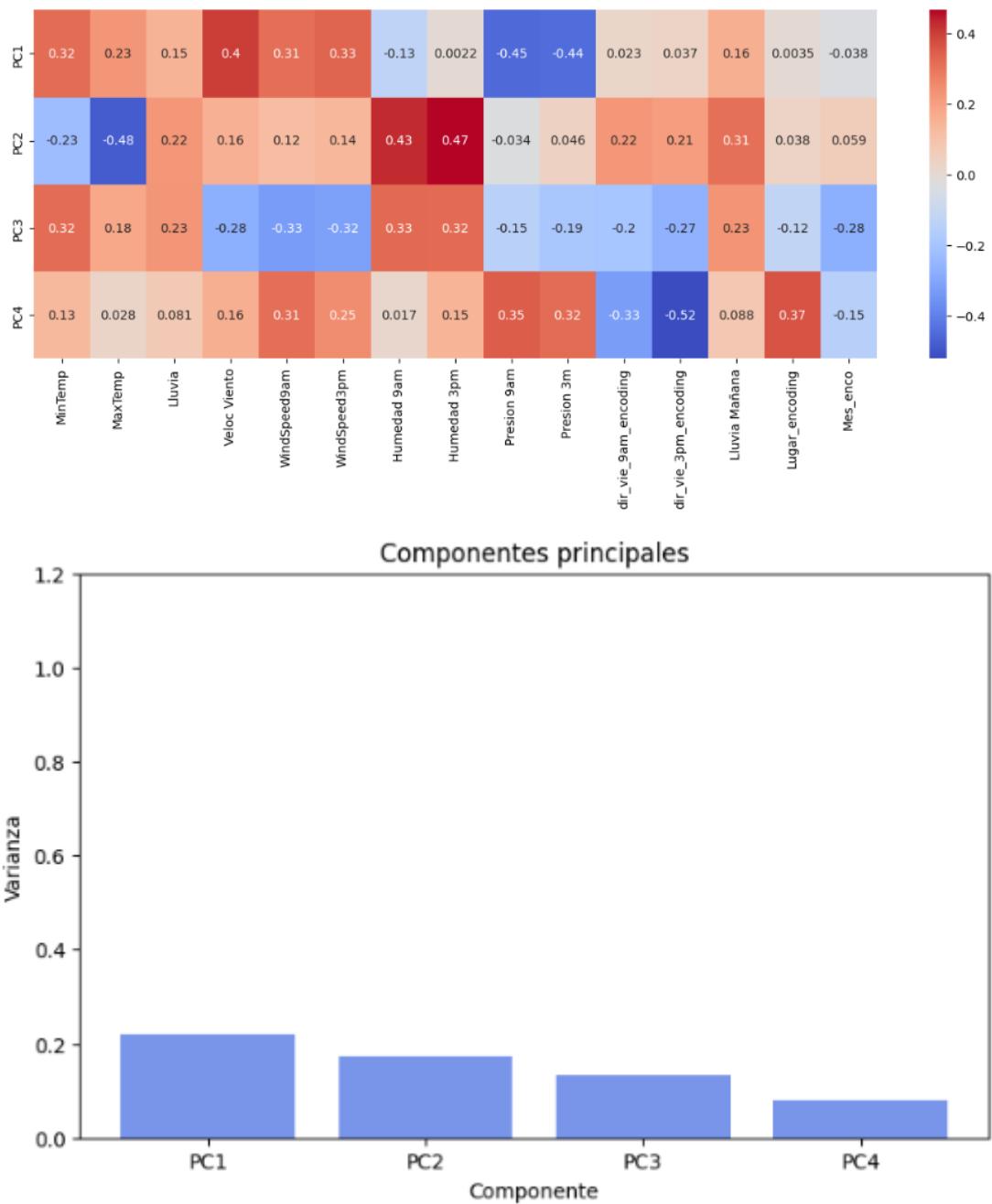


## PCA

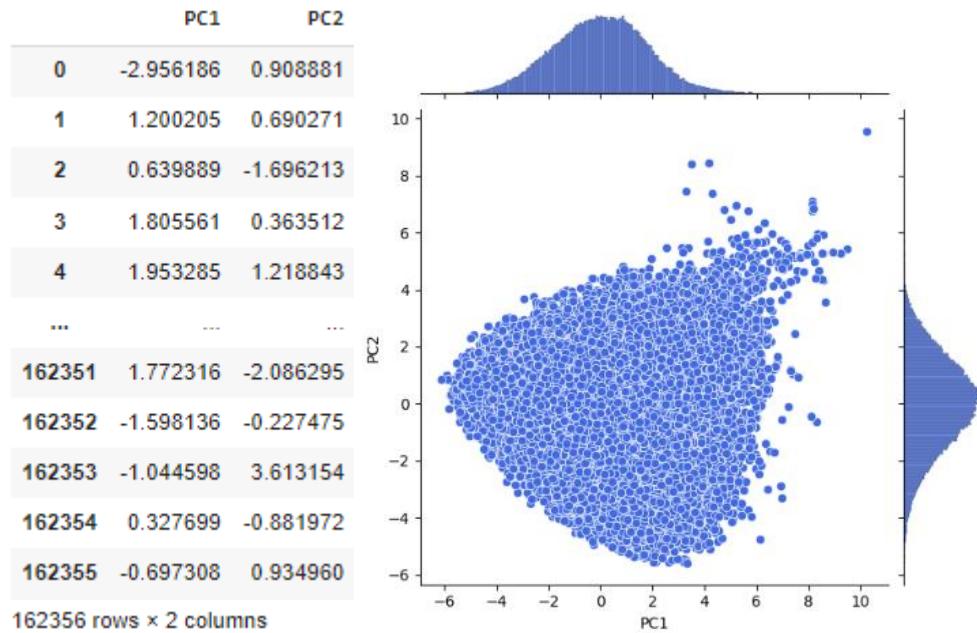
PCA Análisis de Componentes Principales (Principal Component Analysis)

Es una técnica para reducir la dimensionalidad y simplificar la representación de datos complejos mientras se conserva la información esencial. Se utiliza comúnmente en análisis exploratorio de datos, visualización y preparación de datos para modelos de aprendizaje automático.

Para nuestro caso comenzamos el proceso de PCA con 4 componentes, es decir, redujimos inicialmente nuestra base de datos a 4 columnas las cuales explican de igual manera los datos que se encuentran en nuestro dataset original



Aquí se puede apreciar que los PC1 y 2 explican la mayor parte de los datos. Por lo tanto, vamos a quedarnos solamente con estos dos para reducir aún más la dimensionalidad.



## Modelos

Para este proyecto se va a utilizar un Modelo de Clasificación (se trata de problemas que necesitan predecir la clase más probable de un elemento, en función de un conjunto de entradas) dado que lo que se quiere saber es que si va a llover o no.

Y es por eso mismo que el Aprendizaje es Supervisado dado que tiene como objetivo predecir la respuesta que habrá en el futuro, gracias al entrenamiento del algoritmo con datos conocidos del pasado.

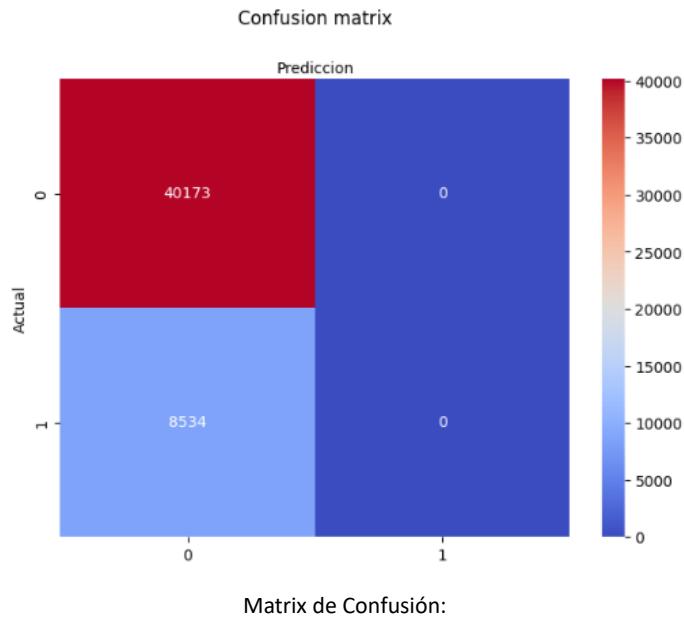
Para este proyecto se utilizaron 4 modelos distintos los cuales son: Regresión Logística, KNN (k-Nearest Neighbors), Random Forests y Árbol de Decisión.

Comenzaremos con el 1ero:

### Regresión Logística

La regresión logística es una técnica que ayuda a predecir cosas que encajan en dos categorías diferentes, como "sí" o "no", "positivo" o "negativo". Usa datos para encontrar la probabilidad de que algo pertenezca a una categoría en particular. Aunque se llama "regresión", en realidad se usa para clasificar cosas en grupos basándose en características que ya conocemos.

**Importante:** En los modelos se trabajó con los datos sin balanceo tal como se produjo después de estandarizarlo y aplicarle PCA. Dado que el orden de los desafíos condujo a eso.



Herramienta que permite visualizar el desempeño de un algoritmo de aprendizaje supervisado. Cada columna de la matriz representa el número de predicciones de cada clase, mientras que cada fila representa a las instancias en la clase real.

En términos prácticos entonces, nos permite ver qué tipos de aciertos y errores está teniendo nuestro modelo.

Verdadero Positivo (posición 0:0): Predice que era positivo y lo era. Verdadero Negativo (posición 1:1): Predice que era falso y lo era. Falso Positivo (posición 0:1): Predice que era positivo, pero resultó ser negativo. Falso Negativo (posición 1:0): Predice que era negativo, pero resultó siendo positivo.

Los Verdaderos Positivos como Negativos son aciertos. Los Falsos Negativos como Positivos son errores.

	precision	recall	f1-score	support
No llueve	0.82	1.00	0.90	40173
Si llueve	0.00	0.00	0.00	8534
accuracy			0.82	48707
macro avg	0.41	0.50	0.45	48707
weighted avg	0.68	0.82	0.75	48707



### K-Nearest-Neighbor (Vecinos cercanos)

KNN (k-Nearest Neighbors) es un algoritmo de aprendizaje automático que se utiliza para clasificar o predecir valores en función de la similitud con ejemplos previos en un conjunto de datos.

Básicamente, busca los ejemplos más cercanos a un nuevo dato y toma decisiones basadas en esos ejemplos cercanos.

En resumen, KNN confía en que puntos similares en el espacio de características tienden a tener resultados similares.

Se lo va a tomar como opción en el proyecto ya que es fácil de entender e implementar, se utiliza principalmente para problemas de clasificación, el valor de K (número de vecinos) se puede ajustar según las características del problema, funciona bien en conjuntos de datos pequeños o cuando hay regiones claras y bien definidas para cada clase y es menos sensible a datos atípicos.

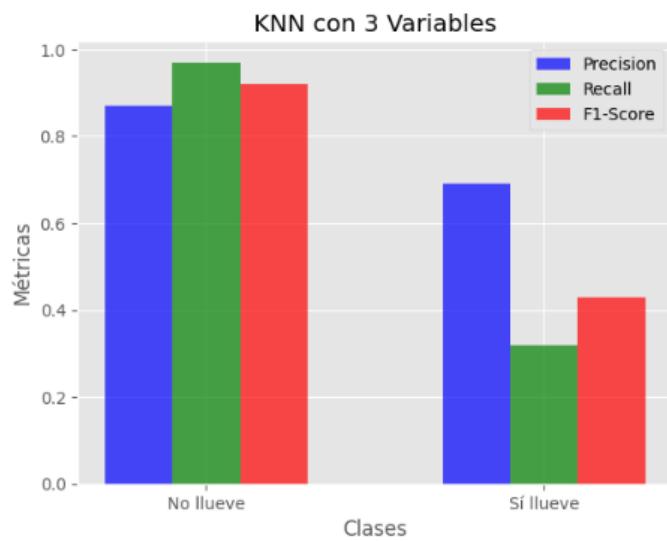
Probamos KNN con 3 variables

```

[[25945    808]
 [ 3950  1769]]
      precision    recall   f1-score   support
          0.0       0.87      0.97      0.92     26753
          1.0       0.69      0.31      0.43      5719

      accuracy         0.78      0.64      0.67     32472
      macro avg       0.78      0.64      0.67     32472
  weighted avg       0.84      0.85      0.83     32472

```



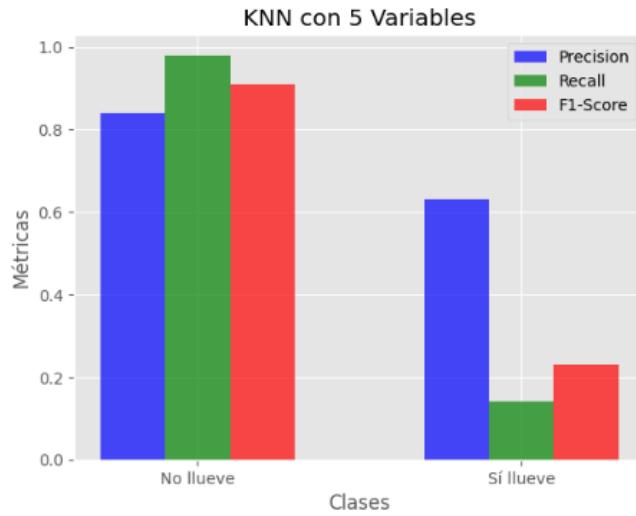
Luego analizamos con 5 variables

```

[[26300    453]
 [ 4913   806]]
      precision    recall   f1-score   support
          0.0       0.84      0.98      0.91     26753
          1.0       0.64      0.14      0.23      5719

      accuracy         0.74      0.56      0.57     32472
      macro avg       0.74      0.56      0.57     32472
  weighted avg       0.81      0.83      0.79     32472

```



## Random Forests

Es un algoritmo de aprendizaje automático que pertenece a la familia de los modelos de conjunto. Su objetivo principal es mejorar la precisión y la robustez de la clasificación y la regresión en comparación con los modelos individuales, como los árboles de decisión.

En este caso lo elijo dado que ofrece una alta precisión en la predicción y es robusto frente al sobreajuste (overfitting) en comparación con otros modelos, permite identificar qué variables son más relevantes para la predicción, tiene la capacidad de manejar datos faltantes sin necesidad de preprocesamiento adicional, no es tan sensible a la elección de hiperparámetros como otros algoritmos, lo que facilita su uso y ajuste.

		precision	recall	f1-score	support	
0.0	0.82	1.00	0.90	26753		
1.0	0.00	0.00	0.00	5719		
		accuracy		0.82	32472	
		macro avg	0.41	0.50	0.45	32472
		weighted avg	0.68	0.82	0.74	32472

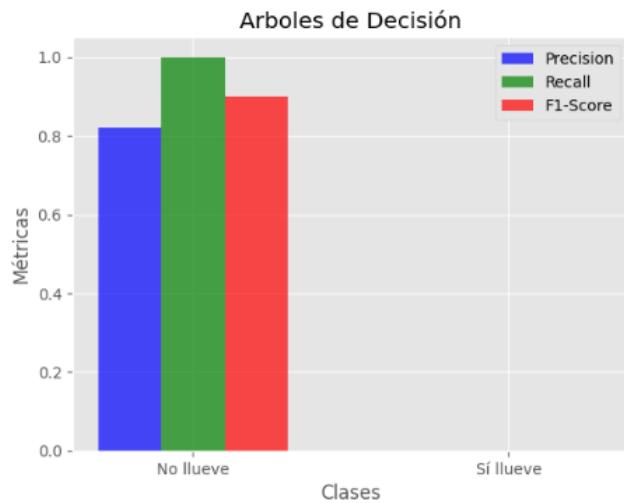


## Árboles de Decisión

Modelo de aprendizaje automático que se utiliza principalmente para problemas de clasificación y regresión. Representa un flujo de toma de decisiones en forma de un árbol, donde cada nodo interno representa una pregunta o una característica, cada rama representa una decisión o una respuesta a la pregunta, y cada hoja representa un resultado, una predicción o una clasificación.

Para este modelo la opción de elegirlo fue: Que es un modelo fácilmente interpretables y comprensibles, eficientes en tiempo de entrenamiento y predicción, permite identificar qué características son más importantes para la predicción y eso ayuda a la selección de variables y comprensión del problema.

	[[26753 0] [ 5719 0]]	precision	recall	f1-score	support
	0.0	0.82	1.00	0.90	26753
	1.0	0.00	0.00	0.00	5719
accuracy				0.82	32472
macro avg		0.41	0.50	0.45	32472
weighted avg		0.68	0.82	0.74	32472



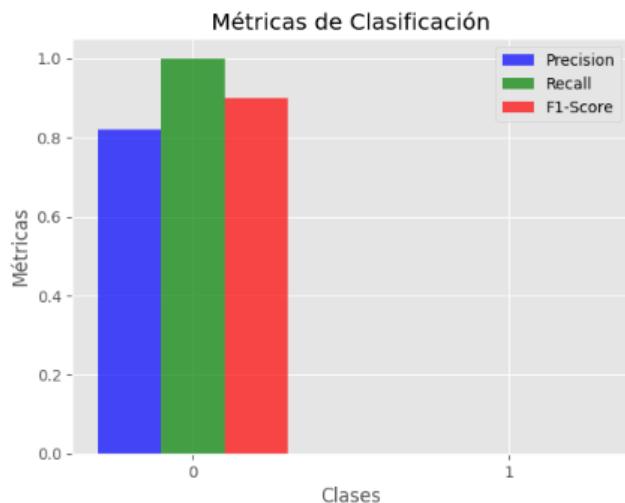
## Hiperparámetros

Los hiperparámetros en ciencia de datos son configuraciones externas al modelo que afectan su comportamiento pero que no se aprenden directamente del conjunto de datos durante el entrenamiento. Estos parámetros son configuraciones que se establecen antes de entrenar un modelo y pueden afectar su rendimiento, complejidad y capacidad de generalización.

Mejores hiperparámetros para los modelos:

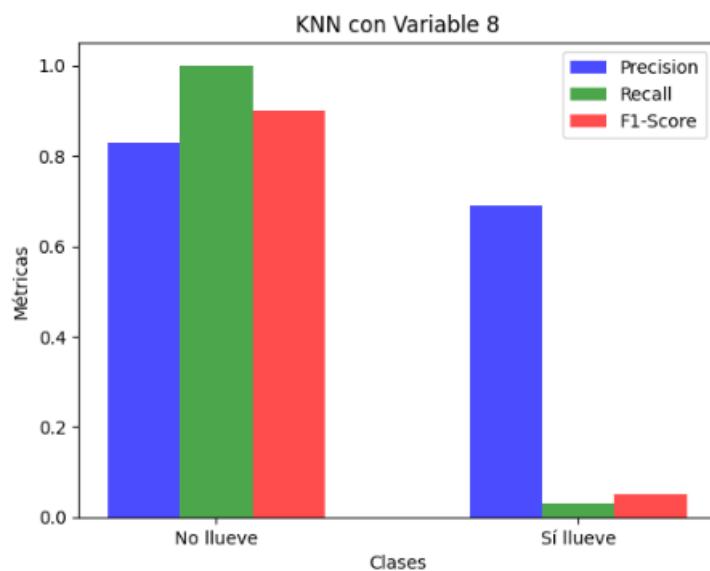
### Regresión logística

Reporte de Clasificación:				
	precision	recall	f1-score	support
0.0	0.82	1.00	0.90	40173
1.0	0.00	0.00	0.00	8534
accuracy			0.82	48707
macro avg	0.41	0.50	0.45	48707
weighted avg	0.68	0.82	0.75	48707



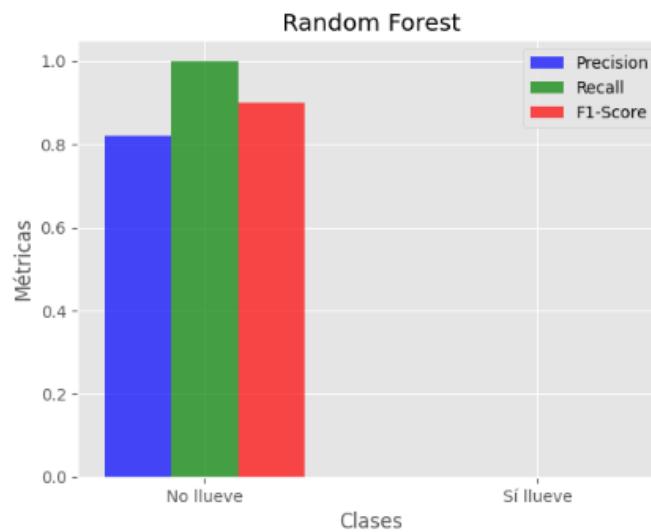
### KNN (k=8)

```
[[26686    67]
 [ 5571   148]]
      precision    recall  f1-score   support
          0.0       0.83     1.00     0.90     26753
          1.0       0.69     0.03     0.05      5719
  accuracy                           0.83     32472
   macro avg       0.76     0.51     0.48     32472
weighted avg       0.80     0.83     0.75     32472
```



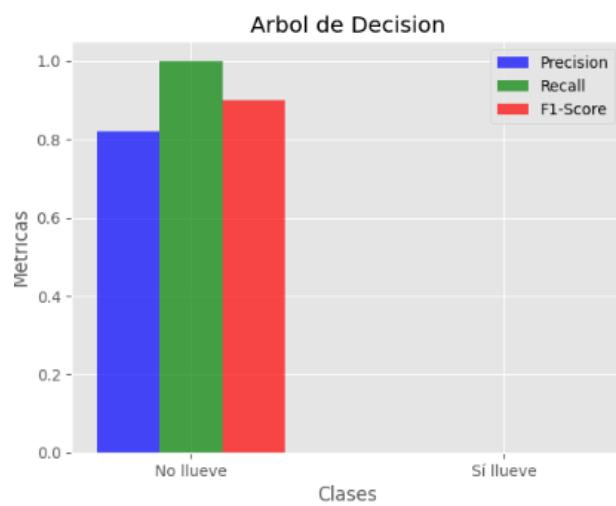
## Random Forest

Reporte de clasificación para datos de prueba:				
	precision	recall	f1-score	support
0.0	0.82	1.00	0.90	26753
1.0	0.00	0.00	0.00	5719
accuracy			0.82	32472
macro avg	0.41	0.50	0.45	32472
weighted avg	0.68	0.82	0.74	32472



## Árbol de Decisión

Reporte de clasificación para datos de prueba:				
	precision	recall	f1-score	support
0.0	0.82	1.00	0.90	26753
1.0	0.00	0.00	0.00	5719
accuracy			0.82	32472
macro avg	0.41	0.50	0.45	32472
weighted avg	0.68	0.82	0.74	32472



# Modelo final



Para este entonces contamos con información muy valiosa de nuestros datos y del performance que tienen los mismos, que métodos y técnicas le hacen mejor, pasamos por el balance de nuestros datos, estandarización de los mismos, reducimos su dimensionalidad, etc. Como así también descubrimos varios modelos y como estos se comportan con esos datos, no solo eso, vimos sus variables como influyen estas en sus rendimientos, hasta logramos obtener sus mejores hiperparámetros.

## Base de datos

En el proyecto la base de datos que mayor rendimiento tuvo fue la base a la cual se le aplicó el desbalanceo con RandomOverSampler (Como se lo pudo ver anteriormente), dado esto partiremos de aquí para buscar nuestro mejor modelo

Por supuesto que seguimos con la Estandarización de nuestros datos, puesto que no es de extrañar que este método se utilice, y más en este dataset en el cual muchas de las columnas de datos que contiene se basan en distintas unidades de medidas, por lo que su uso en nuestro proyecto, doto a los modelos de grandes cambios.

Luego continuamos reduciendo la dimensionalidad con PCA, en este caso la reducción de la dimensionalidad también ha tenido un gran impacto en nuestros modelos es por eso que en combinación con la estandarización son técnicas que no pueden faltar para buscar nuestro mejor modelo.

Estos fueron los pasos a seguir en lo que tiene que ver con nuestro DataSet y que proceso se siguió para poder ayudar de mejor manera al modelo.

## Modelos

### Regresión Logística

	precision	recall	f1-score	support
0	0.95	0.99	0.97	20295
1	0.95	0.80	0.87	5740
accuracy			0.95	26035
macro avg	0.95	0.89	0.92	26035
weighted avg	0.95	0.95	0.94	26035

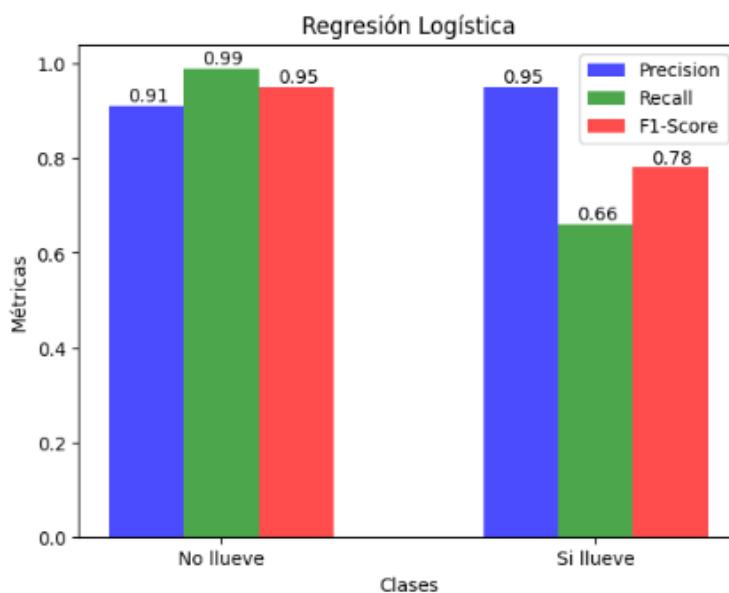


## Hiperparámetros

Anteriormente ya tuvimos sus resultados en este caso son:

Mejores hiperparámetros encontrados: {'C': 0.0001, 'penalty': 'l1', 'solver': 'liblinear'}

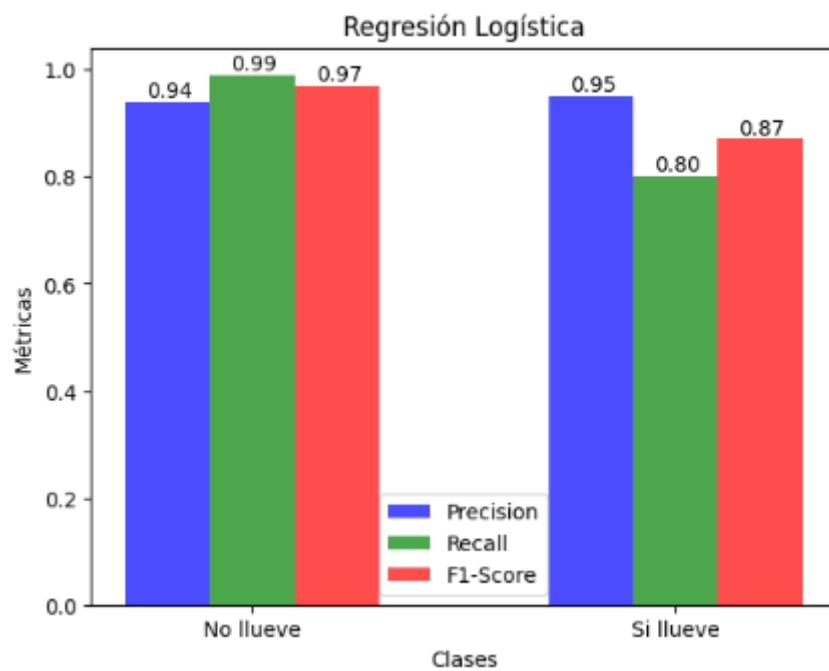
	precision	recall	f1-score	support
0	0.91	0.99	0.95	20295
1	0.95	0.66	0.78	5740
accuracy			0.92	26035
macro avg	0.93	0.83	0.87	26035
weighted avg	0.92	0.92	0.91	26035



Para este modelo usaremos a su vez otro método de hiperparámetros, dados que sus resultados son muy diferentes con los otros modelos le damos una segunda oportunidad, pero con otra técnica de hiperparámetros en este caso BayesSearchCV para ver si la mejora pueda equiparar a los demás modelos.

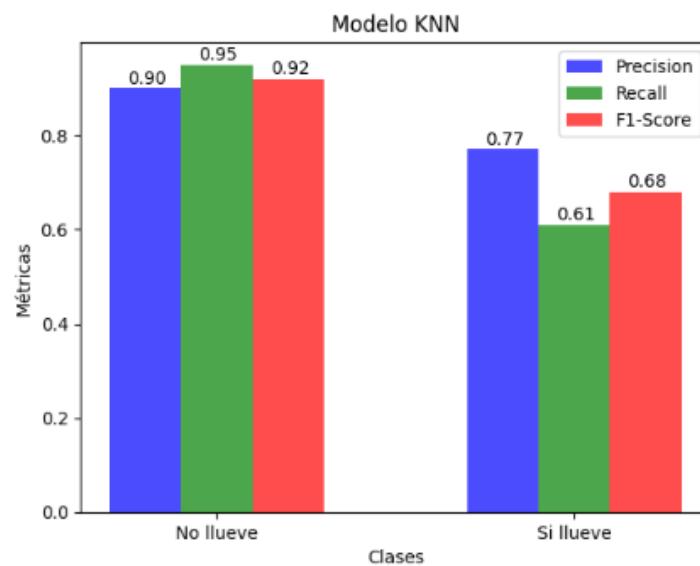
Es una técnica de optimización de hiperparámetros que utiliza el proceso de optimización Bayesiana para encontrar la combinación óptima de hiperparámetros, utiliza información acumulada de las iteraciones anteriores para decidir qué conjunto de hiperparámetros probar a continuación.

	precision	recall	f1-score	support
0	0.94	0.99	0.97	20295
1	0.95	0.80	0.87	5740
accuracy			0.95	26035
macro avg	0.95	0.89	0.92	26035
weighted avg	0.95	0.95	0.94	26035



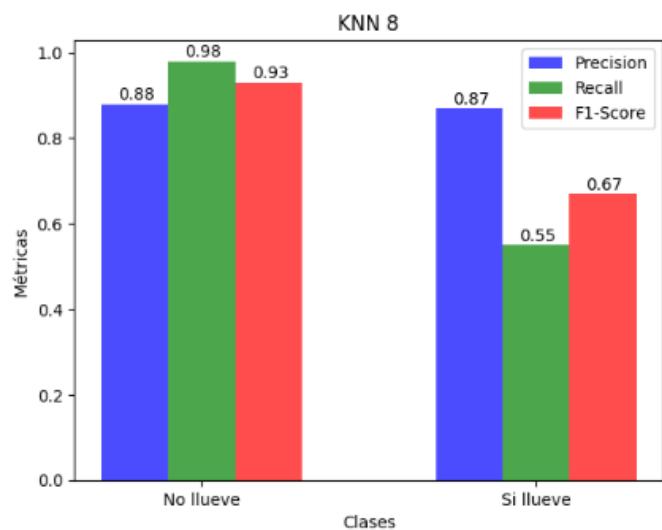
## KNN

	precision	recall	f1-score	support
0	0.90	0.95	0.92	20295
1	0.77	0.61	0.68	5740
accuracy			0.87	26035
macro avg	0.83	0.78	0.80	26035
weighted avg	0.87	0.87	0.87	26035



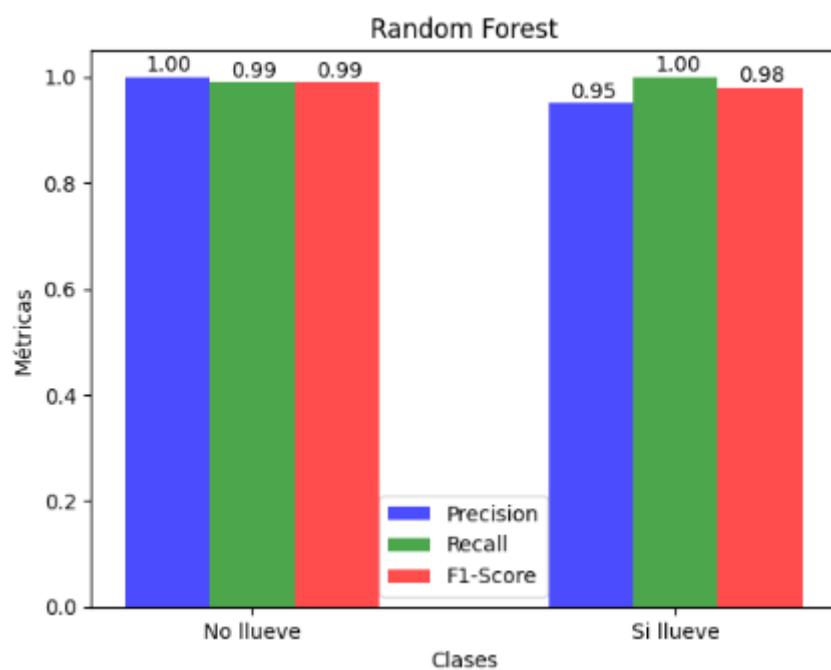
## Hiperparametros:

	precision	recall	f1-score	support
0	0.88	0.98	0.93	20295
1	0.87	0.55	0.67	5740
accuracy			0.88	26035
macro avg	0.88	0.76	0.80	26035
weighted avg	0.88	0.88	0.87	26035



## Random Forest

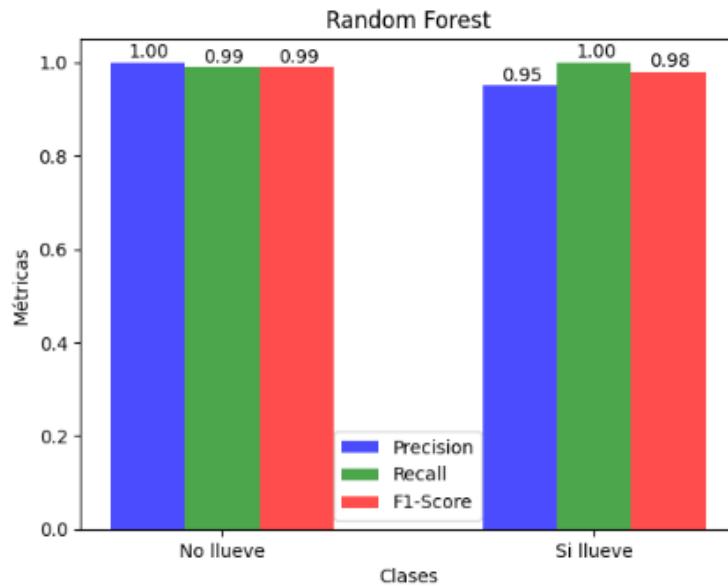
	precision	recall	f1-score	support
0	1.00	0.99	0.99	20295
1	0.95	1.00	0.98	5740
accuracy			0.99	26035
macro avg	0.98	0.99	0.98	26035
weighted avg	0.99	0.99	0.99	26035



## Hiperparametros

```
Mejores hiperparámetros encontrados: {'max_depth': None, 'min_samples_split': 2, 'n_estimators': 20}
      precision    recall   f1-score   support
0         1.00     0.99     0.99     20295
1         0.95     1.00     0.98     5740

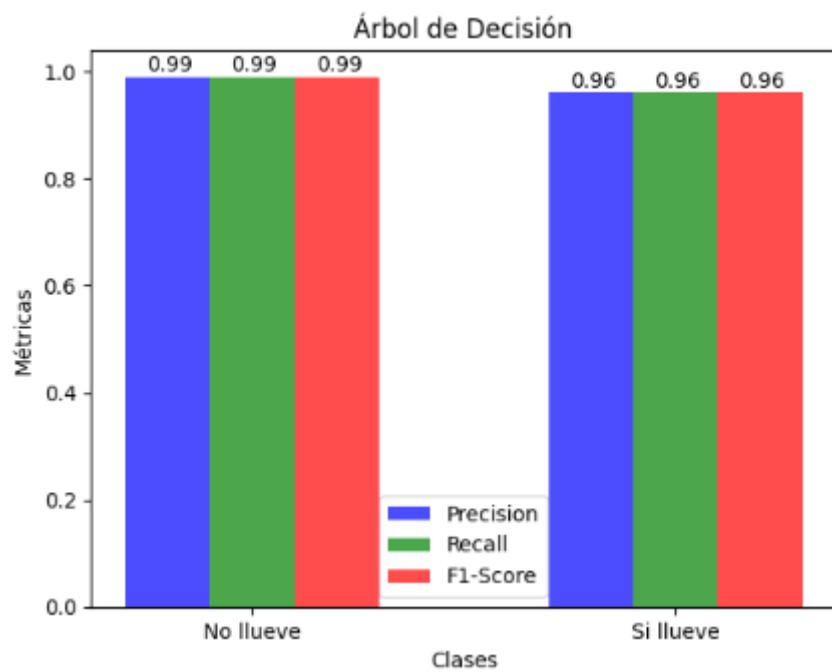
   accuracy       0.99
macro avg       0.98
weighted avg    0.99     0.99     0.99     26035
```



## Árbol de Decisión

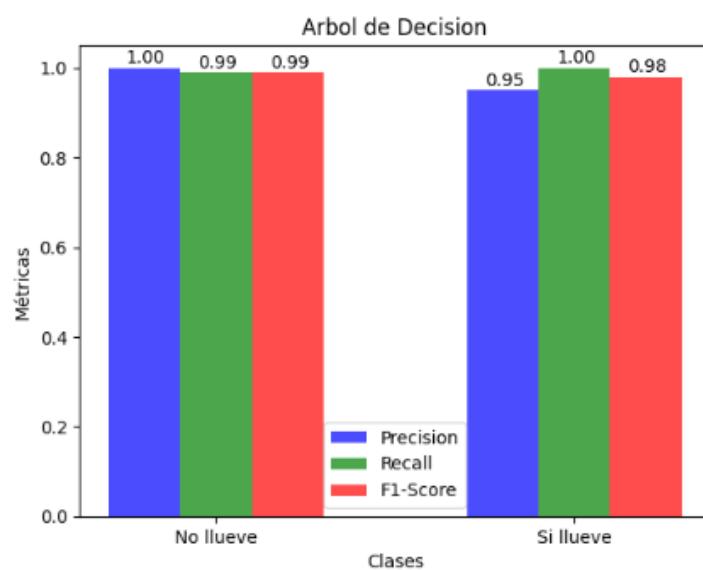
```
      precision    recall   f1-score   support
0         0.99     0.99     0.99     20295
1         0.96     0.96     0.96     5740

   accuracy       0.98
macro avg       0.98
weighted avg    0.98     0.98     0.98     26035
```



## Hiperparametros

```
Mejores hiperparámetros encontrados: {'max_depth': 5, 'min_samples_split': 2}
      precision    recall  f1-score   support
0         1.00     0.99     0.99    20295
1         0.95     1.00     0.98    5740
accuracy                           0.99    26035
macro avg       0.98     0.99     0.98    26035
weighted avg    0.99     0.99     0.99    26035
```



## Mejores Modelos

Podemos observar que los modelos que mejor performaron fueron los modelos de Random Forest y de Árbol de Decisión.

Se le hará a los mismos una validación cruzada y Métodos de Ensamble

La validación cruzada es una técnica fundamental en machine learning para evaluar el rendimiento de un modelo y su capacidad de generalización. Se utiliza para estimar cómo se comportará un modelo en un conjunto de datos independiente, es decir, para entender cómo se comportará ante datos nuevos que no ha visto durante el entrenamiento.

Los métodos de ensamble son técnicas en el campo del Machine Learning donde se combinan múltiples modelos de predicción para mejorar la precisión y el rendimiento general. En lugar de depender de un solo modelo, los métodos de ensamble construyen un conjunto de modelos base y combinan sus predicciones para obtener una predicción final más robusta y precisa.

## Random Forest

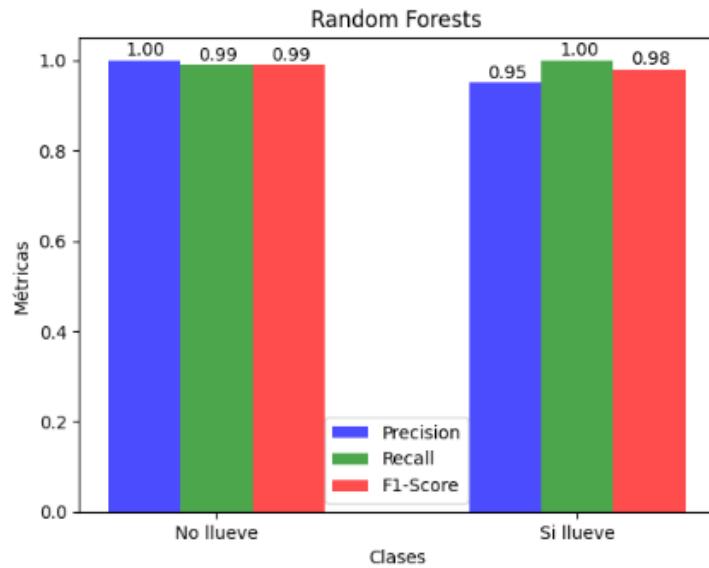
### Cross validation

EL cross-validation es una técnica utilizada para evaluar el rendimiento de un modelo.

Consiste en dividir el conjunto de datos en subconjuntos más pequeños, llamados "folds", y luego realizar múltiples iteraciones donde se entrena el modelo en varios subconjuntos y se evalúa en el restante.

```
Puntajes de la validación cruzada para el mejor modelo de Random Forest (Accuracy): [0.98938928 0.99130936 0.99030105 0.99010899 0.98986892]
Accuracy promedio para el mejor modelo: 0.9901955213013525
Reporte de clasificación con validación cruzada para el mejor modelo de Random Forest en conjunto de prueba:
      precision    recall   f1-score   support
          0         1.00     0.99     0.99    20295
          1         0.95     1.00     0.98    5740

      accuracy                           0.99
      macro avg       0.98     0.99     0.98    26035
  weighted avg       0.99     0.99     0.99    26035
```



## Métodos de ensamble

Random Forest ya utiliza una técnica de ensamblado conocida como Bagging (Bootstrap Aggregating) como parte de su proceso interno. Esta técnica construye múltiples árboles de decisión en paralelo utilizando diferentes subconjuntos de datos (muestreo con reemplazo) y combinando sus predicciones para reducir la varianza y mejorar la generalización del modelo.

Debido a esta naturaleza de Random Forest, no se suele aplicar Boosting a este tipo de modelo, ya que ya maneja eficazmente la variabilidad y el overfitting a través del ensamblado de árboles.

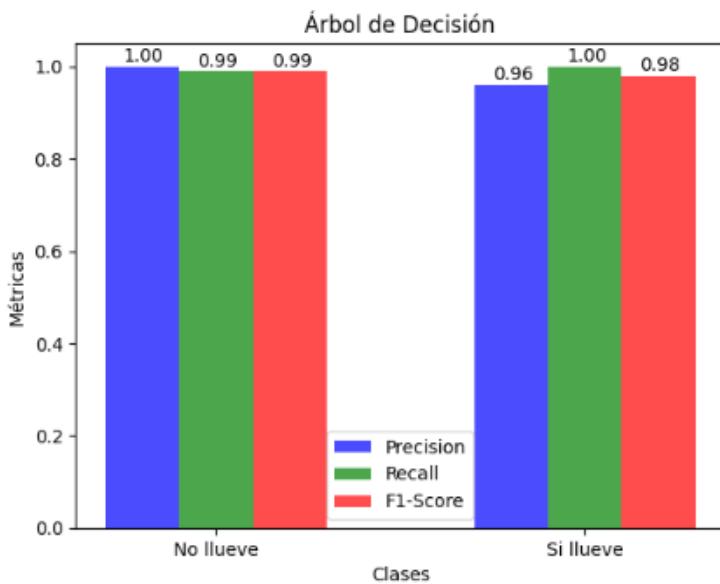
## Árbol de Decisión

### Cross Validation

Puntajes de la validación cruzada para el mejor modelo de Árbol de Decisión (Accuracy): [0.98929326 0.99106929 0.99034907 0.98938877 0.98982091]  
Accuracy promedio para el mejor modelo de Árbol de Decisión: 0.9899842579991447

Reporte de clasificación con validación cruzada para el mejor modelo de Árbol de Decisión:

	precision	recall	f1-score	support
0	1.00	0.99	0.99	81178
1	0.96	1.00	0.98	22958
accuracy			0.99	104136
macro avg	0.98	0.99	0.99	104136
weighted avg	0.99	0.99	0.99	104136



## Métodos de Ensamble

### Baggin

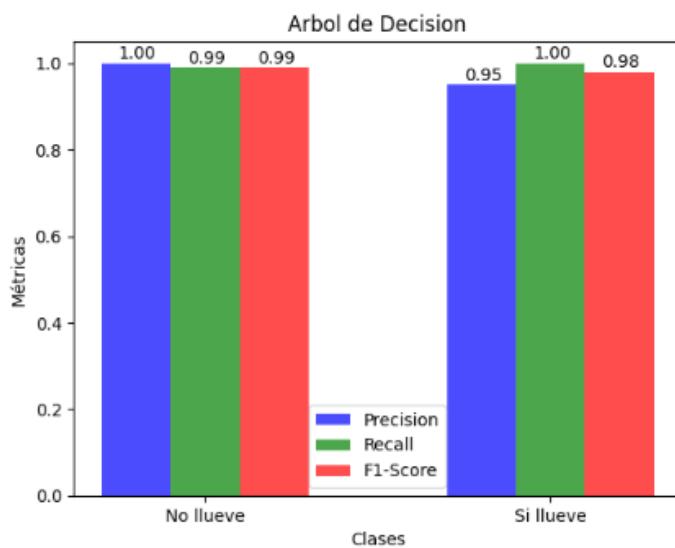
La idea central del bagging es construir múltiples modelos independientes y combinar sus predicciones para obtener una predicción final más precisa y robusta.

```

Puntajes de la validación cruzada para el mejor modelo de Árbol de Decisión con Bagging (Accuracy): [0.98924525 0.99121333 0.99025304 0.98986892 0.98986892]
Accuracy promedio para el mejor modelo de Árbol de Decisión con Bagging: 0.9900898905723649
Reporte de clasificación para el mejor modelo de Árbol de Decisión con Bagging:
precision    recall   f1-score  support
          0       1.00     0.99     0.99    20295
          1       0.95     1.00     0.98    5740

accuracy                           0.99
macro avg       0.98     0.99     0.98    26035
weighted avg    0.99     0.99     0.99    26035

```

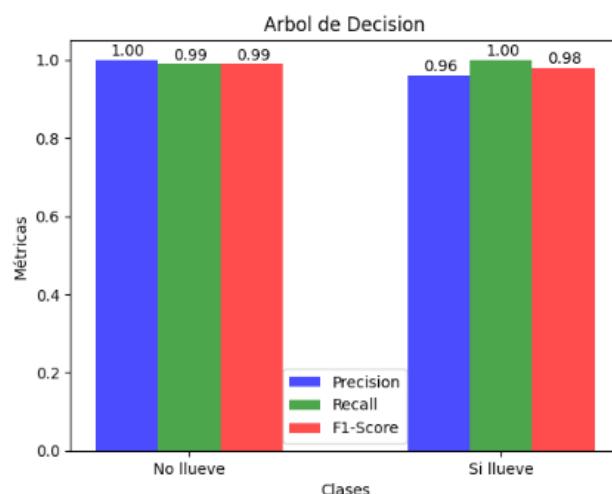


## Boosting

El boosting se centra en la construcción secuencial de modelos, donde cada modelo se entrena para corregir los errores de los modelos anteriores.

Elegí LightGBM dado que tiene buena eficiencia en el tiempo, ya que usa menos memoria y logra una mayor precisión en el modelo en comparación con otros métodos de boosting y te brinda una amplia gama de parámetros para ajustar y optimizar el rendimiento del modelo.

	precision	recall	f1-score	support
0	1.00	0.99	0.99	20295
1	0.96	1.00	0.98	5740
accuracy			0.99	26035
macro avg	0.98	0.99	0.99	26035
weighted avg	0.99	0.99	0.99	26035

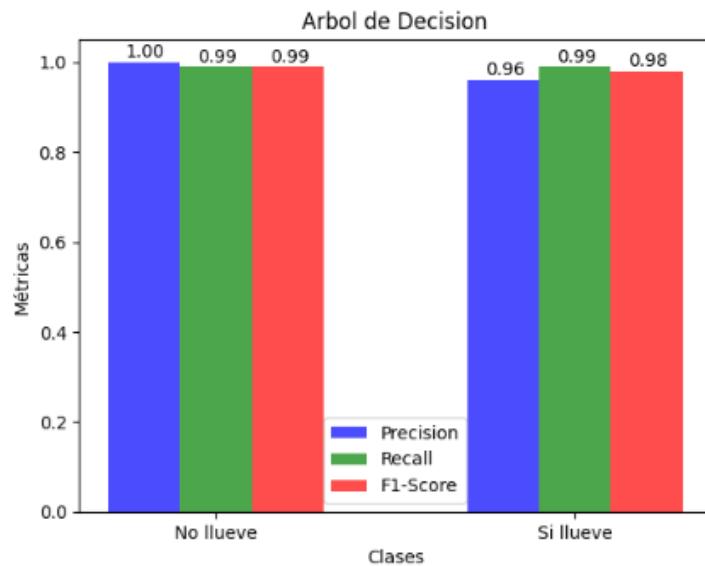


Para el XGBoost se sabe que su rendimiento es bueno por lo tanto va a ayudar en la rapidez del proceso, puede trabajar con grandes cantidades de datos aunque este no es el caso, ayuda a prevenir el sobreajuste por sus técnicas avanzadas, lo que logra como consecuencia que el modelo generalice bien datos que todavía no observo, tiene gran capacidad de escalar, y ayuda a construir un modelo robusto gracias a lo antes mencionado.

Puntajes de la validación cruzada para XGBoost (Accuracy): [0.98938928 0.99202958 0.99073318 0.99068517 0.99034907]  
Accuracy promedio para XGBoost: 0.9906372555885758

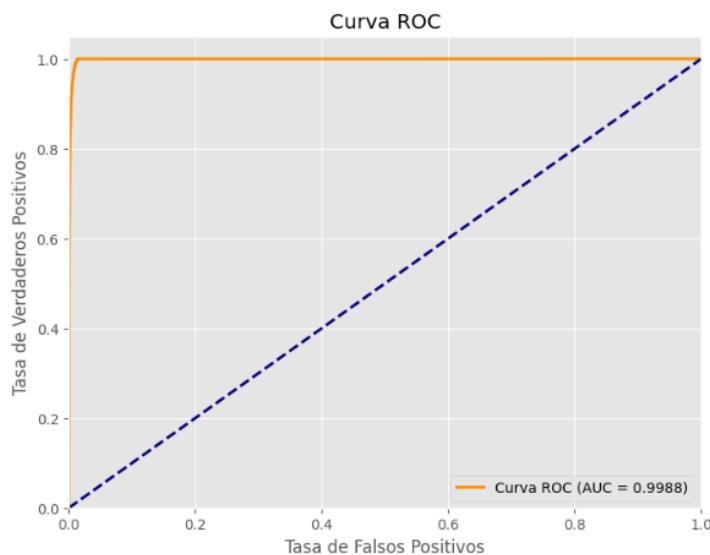
Reporte de clasificación para XGBoost:

	precision	recall	f1-score	support
0	1.00	0.99	0.99	20295
1	0.96	0.99	0.98	5740
accuracy			0.99	26035
macro avg	0.98	0.99	0.99	26035
weighted avg	0.99	0.99	0.99	26035



## Evaluación de nuestro modelo final

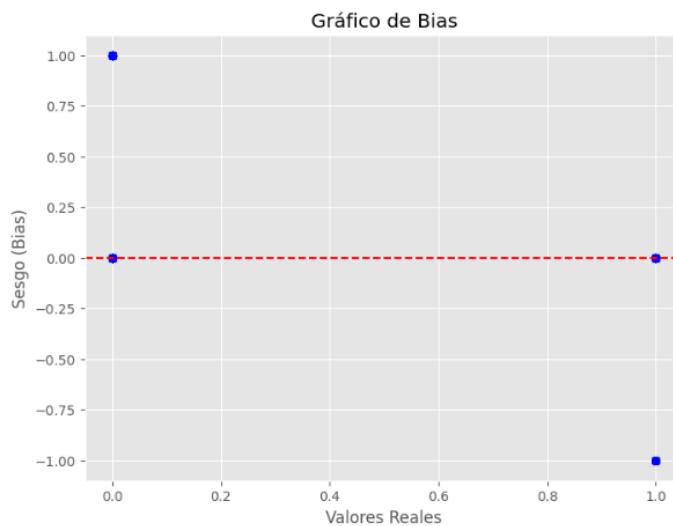
En esta sección pasaremos nuestro mejor modelo por distintas evaluaciones para ver sus resultados, como estos se reflejan y que se puede mejorar posteriormente, para ello usaremos más de una métrica de evaluación.



Un AUC cercano a 1 (o 100%) indica que el modelo tiene una excelente capacidad para distinguir entre las clases. Es decir, la probabilidad de que el modelo clasifique correctamente una instancia de una clase como más probable que la otra es muy alta.

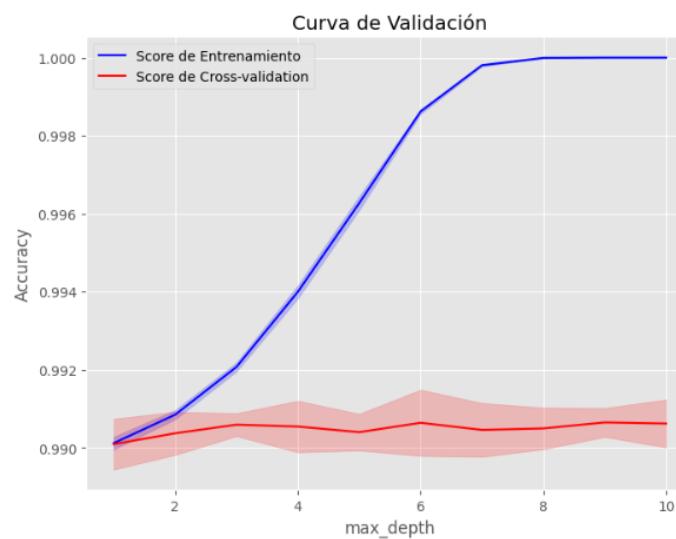
También un AUC tan alto sugiere que las predicciones del modelo están muy en línea con las clases reales (modelo altamente preciso).

Un rendimiento tan alto también puede indicar que el modelo está bien ajustado a los datos y tiene una baja probabilidad de sobreajuste o subajuste. Esto sugiere un buen equilibrio entre sesgo y varianza.



Bias nos muestran cómo se comporta nuestro modelo en función de los valores reales que está prediciendo. Cuando el valor real es 0, el modelo hace predicciones muy cercanas a lo que realmente sucede, lo que es correcto. Pero, cuando el valor real es 1, el modelo parece tener dificultades. Sus predicciones están bastante alejadas de lo que deberían ser, lo que indica un problema cuando se trata de predecir estos casos específicos.

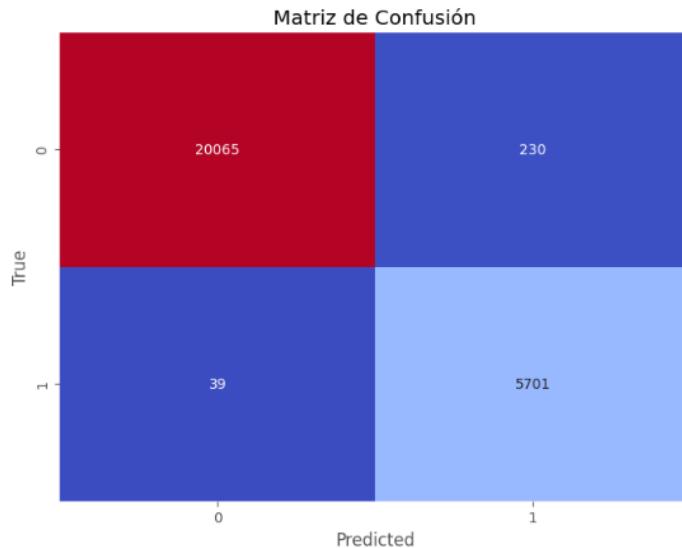
Para resumir la información del grafico nuestro modelo funciona bien para algunos valores, y hay un área específica donde no está haciendo predicciones tan precisas como debería.



Se vio una mejora significativa del modelo en su rendimiento al ajustar las configuraciones. Se partió de un puntaje ya alto, y el mismo, progreso rápido también aunque al acercarse al 1,00 las mejores fueron menos notables por lo tanto capaz que ajustar tanto el modelo puede en este caso no beneficiar al rendimiento, se deben hacer las pruebas para ello del modelo en datos que todavía no vio.

Esta curva muestra cómo cambia el rendimiento del modelo en los conjuntos de entrenamiento y validación a medida que se ajusta un parámetro específico del modelo. En el caso del ejemplo, el parámetro es la profundidad máxima del árbol.

La curva de validación permite visualizar cómo el rendimiento del modelo cambia en función de la variación de un parámetro específico, lo que ayuda a tomar decisiones informadas sobre la configuración óptima del modelo para evitar el sobreajuste y mejorar su capacidad predictiva.



Verdadero Positivo (posicion 0:0): Predice que era positivo y lo era. Verdadero Negativo (posicion 1:1): Predice que era falso y lo era. Falso Positivo (posicion 0:1): Predice que era positivo, pero resultó ser negativo. Falso Negativo (posicion 1:0): Predice que era negativo, pero resultó siendo positivo.

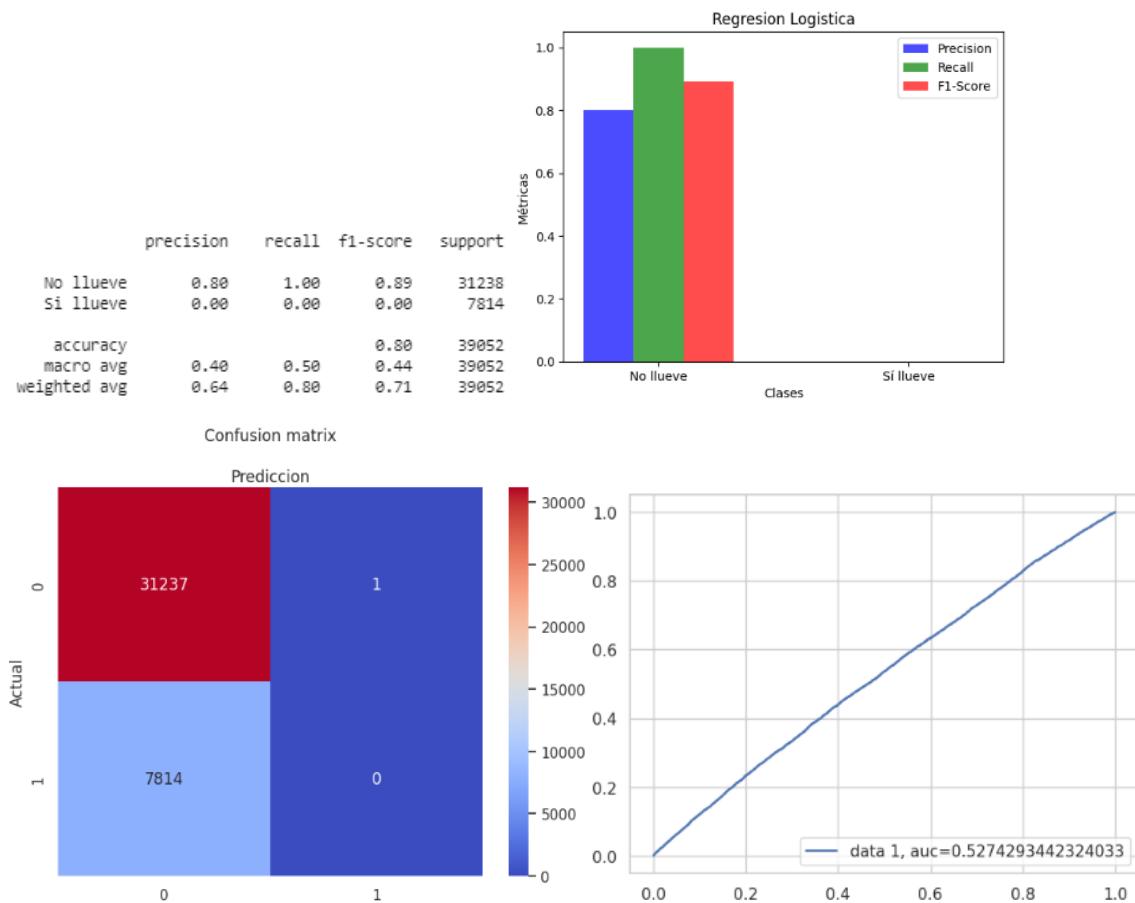
En este caso podemos apreciar que el modelo cuenta con pocos errores y que la mayoría de las veces predice de buena manera en ambos casos Verdaderos Positivos y Verdaderos Negativos, por lo que se puede decir que nuestro modelo cuenta con un margen de error aceptable dado que sus errores son una pequeña parte de sus datos.

# Evolución de nuestro Modelo

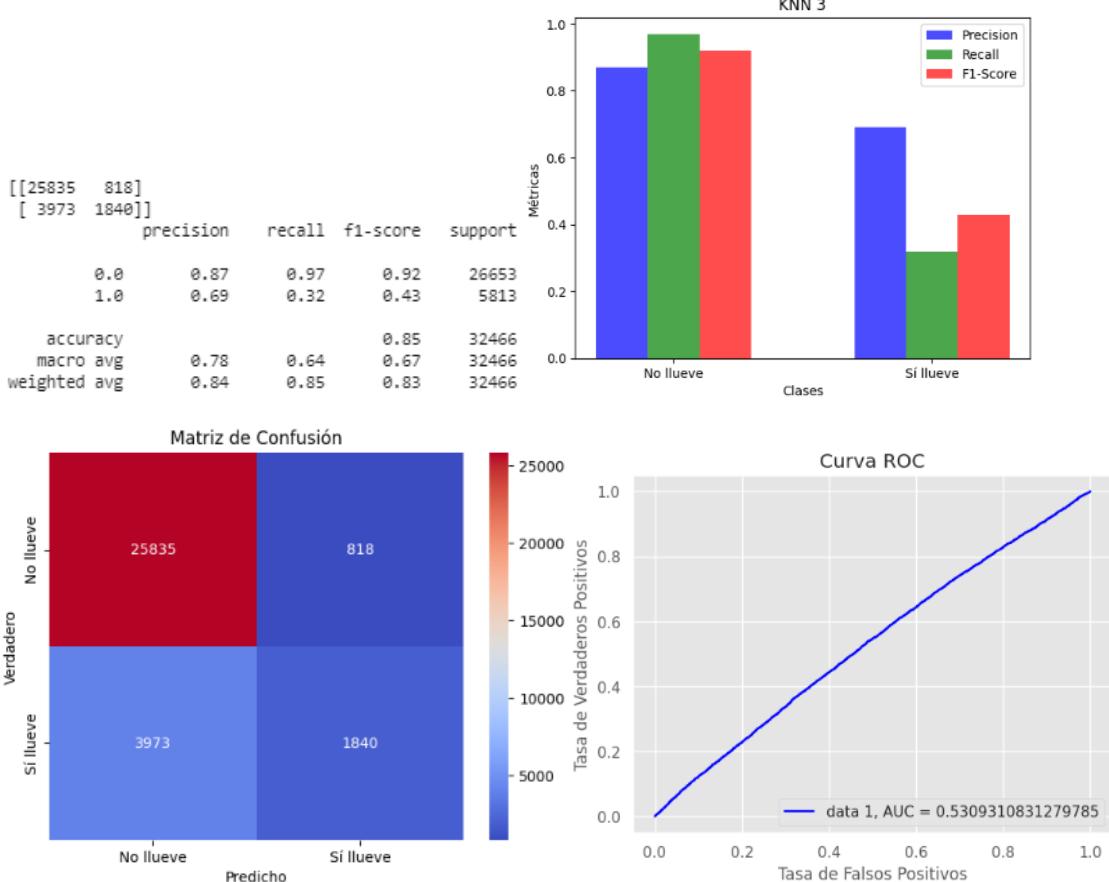
## Final 🏆

- Nuestro primer modelo en este proyecto tenía este aspecto en este caso se utilizó un Modelo de Regresión Logística (Desafío Num 5).

En este caso al modelo no se le realizó ninguna técnica o método para mejorarlo.



- Para la primera entrega nuestro mejor modelo era un modelo de Vecinos Cercanos con  $k=3$ . En este caso el modelo ya contaba con los datos balanceados con el método Smote, con la estandarización de sus datos y con el PCA (Reducción de su dimensionalidad).



- Luego podemos observar los datos de los modelos que se encuentra en la sección de hiperparámetros a partir de la pagina [38](#), en el cual dadas las métricas nuestro modelo de KNN con 3k sigue siendo nuestra mejor opción.
- Por ultimo se sabe que nuestro mejor modelo fue un Árbol de Decisión que paso por todos los procesos anteriormente mencionados, y además se le aplico Random Over Sampler en vez de Smote, y que se le sumo también las técnicas de cross validation y los métodos de ensamble (Baggin y boosting)

# Conclusión Final

## **Entendimiento del problema:**

En lo que respecta al objetivo del proyecto se llegó a un modelo robusto y que su performance es aceptable para aquello en lo que queremos predecir en este caso la lluvia, por lo tanto, las expectativas de nuestro jefe fueron alcanzadas, dados los requerimientos que él nos propuso se cumplieron.

## **Exploración y preparación de datos:**

En cuanto al EDA, se realizó varios gráficos que gracias a ello se descubrieron patrones, relaciones que tiene la lluvia con distintas determinaciones del clima como lo es la humedad, en que época de año llueve más, como es la evolución de la lluvia, no solo eso sino que también se trató varias preguntas y hipótesis con métodos estadísticos ya sea multivariados o no, y demás [Insights](#) que se puede ver en cierta parte del proyecto.

Sobre la limpieza y la detección de Outliers se eliminaron varias columnas de nuestro dataset, algunas se trabajaron con la moda y la mediana, en tanto los outliers no se trataron tanto ya que los mismo se determinaron que era un fiel reflejo de la realidad y en lo que en el día a día puede pasar por eso se los tomo como un dato más sin la consideración de que se lo tome como atípico.

## **Selección y construcción de características:**

En este proceso se mejoró la capacidad predictiva del modelo al utilizar por ejemplo la estandarización de los datos, como así también tratar el desbalanceo de los datos lo cual fue el gran problema que se tuvo que atravesar para este proyecto ya que, naturalmente son más la cantidad de días que no llueve que los que si, y como consecuencia de esto repercute en nuestra dataset y proyecto, aunque se pudo mejorar y genera un conjunto de características óptimo y relevante, lo que conduce a modelos más precisos, interpretables y eficientes (Esto gracias a las técnicas como Smote, RandomOverSampler, Smotetoken). Además de pasar nuestros datos por una reducción de dimensionalidad como lo es PCA.

## **Modelado y evaluación:**

En este caso se tuvo una evolución muy buena de los modelos a lo largo del proyecto, significativamente gracias a nuestro Feature Engineering y al trato que se hizo en el mismo por lo anteriormente mencionado, pero no solo eso, se produjeron buenos

resultados en los distintos modelos con la validación simple, cruzada, hiperparámetros para cada modelo y métodos de ensamble.

Por lo que pasamos de un modelo el cual no era ni apenas aceptable a uno en el cual se puede trabajar y desarrollar, aunque se puede mejorar todavía.

### **Interpretación de resultados:**

Mediante el desarrollo del proyecto se fue evaluando que le resultaba mejor a nuestro modelo, esto puede ser eliminar o no una variable que no influenciaba mucho en el modelo o lo empeoraba, balancear el dataset, reducir su dimensionalidad, que modelo se veía que andaba mejor, ese modelo con que parámetros y que hiperparámetros funcionaba más eficiente,etc.

Por lo que se tuvo la estrategia que al final del proyecto agrupar todos los métodos que se hicieron y fueron los más óptimos para combinar todo eso y generar nuestro modelo final con su mayor performance, y de alguna manera más sencilla y fácil de ver llevar todo el trabajo hacia una parte final del notebook el cual dé un cierre al proyecto.

### **Pasos a futuro para el proyecto**



Para un futuro a nuestro modelo final le podemos realizar varias modificaciones, no solo en lo que respecta al modelo en sí, sino que también al contexto por ejemplo:

- Darle nuevos datos a probar y ver como funciona y se desempeña con esos datos.
- Estudiar como el modelo evoluciona a través del tiempo, por ejemplo en un año puede ser que la realidad haya cambiado o se generen nuevos paradigmas en lo que tiene que ver con clima (Quizá un año de muchas sequias continuas, o lluvias abundantes) y por lo tanto como consecuencia haga que el modelo no se adapte bien a las circunstancias por lo que se tendría que rever.
- Que se generen nuevas preguntas a partir de los datos o del modelo, ya sea en el área de data o por una necesidad de otra área de la empresa o nuestro jefe, dado que un trabajo de Data Science no tiene como tal un ciclo de inicio a fin, sino que es muy de ida y vueltas en las distintas etapas, puede darse que se generen nuevas necesidades, requisitos, situaciones en la realidad a medir y tener en cuenta, etc. Por ejemplo, predecir cuanta cantidad puede llover en cierto mes.
- Analizar y ver nuevamente como se llevó a cabo el progreso de Data Wrangling, ver que otros caminos se puede tomar, realizar una crítica constructiva del proceso en área de data, ver si tal medida era la más correcta o algo que no se tuvo en cuenta pueda ayudar aún más y de mejor manera el trabajo.

- Tener en cuenta además los métodos utilizados en el cross validation, hiperparámetros, métodos de ensamble, dado que muchos de ellos son prueba y error para saber bien cual funciona mejor, tal vez comunicarse con un Machine Learning Engineer para poder optimizar esto mucho mejor.
- Una vez llevado a producción ver como performa en la realidad si al final es aceptable o no su rendimiento, darle un tiempo para ver si funciona o no, tratar de tener un periodo de prueba en el cual se observe detenidamente como es su desarrollo
- Conseguir más fuentes de datos ya sea dentro o fuera de la empresa, que tengan más años de análisis o más variables dentro de tu dataset para enriquecer el análisis.