

Clima Australia

Cristian Arroyo

Nibaldo Quezada

Nicolas Peña



Entendiendo el negocio

Australia presenta un clima mayormente desértico o semiárido, con el 40% de su territorio cubierto por dunas de arena. Sin embargo, su gran extensión geográfica le confiere una diversidad climática significativa. El norte tiene un clima tropical con estaciones seca e húmeda, el sureste y suroeste tienen climas templados. Las lluvias son escasas en el interior, aumentando en las zonas costeras, donde la tierra es más fértil. La temporada de lluvias monzónicas ocurre en el verano en el norte, mientras que en el sur la lluvia es moderada durante todo el año en la franja costera entre Sídney y Adelaida.

Clima



Entendiendo y preparando datos

General

142.193 / 24
Object 7 / Float 17

Data time

Convertimos a tipo de datetime
y agrupamos registros por años

Location

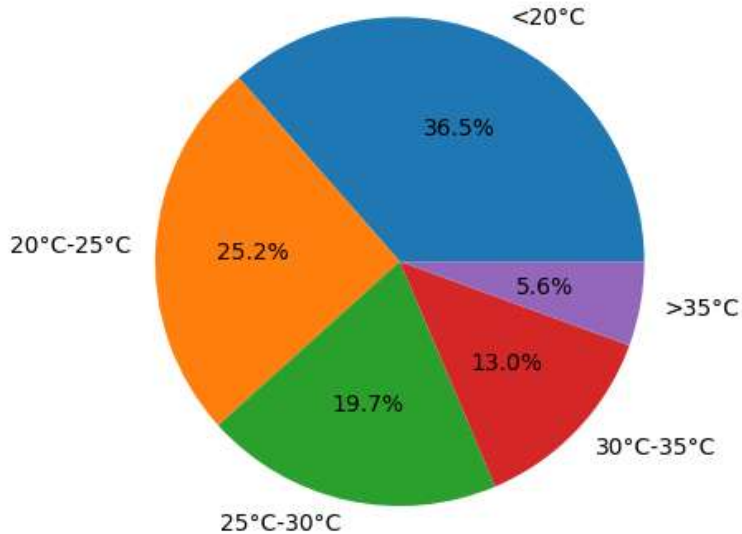
Para comenzar a trabajar la
data realizamos un `.groupby`,
para saber cuántos registros
existen por localidad

Código

Realización de `.describe`:
podemos obtener información
detallada sobre las columnas
numéricas.
`describe(include=[np.object])` :
Para apreciar las variables
categóricas del dataset

Temperaturas

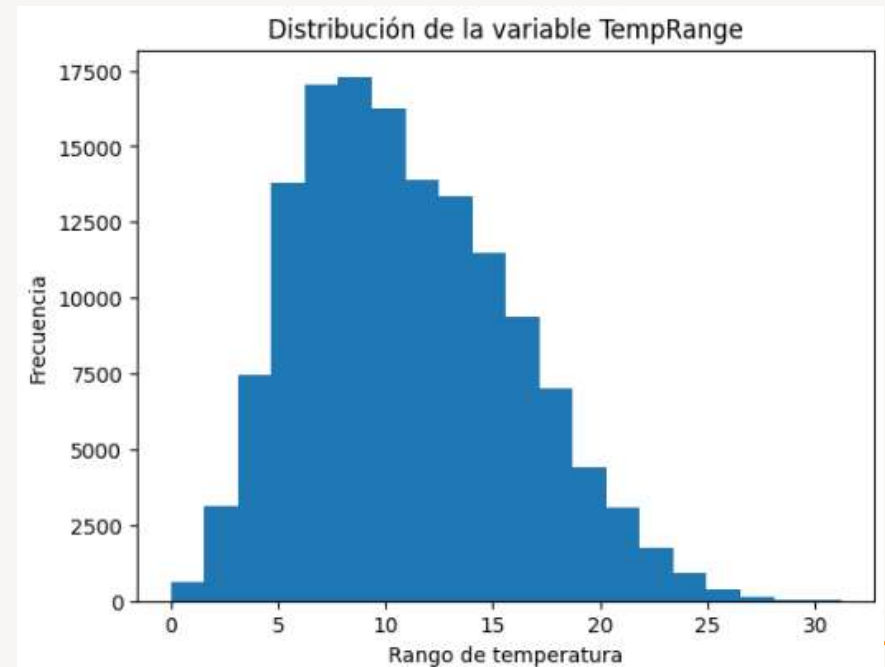
Distribución de la temperatura máxima en Australia



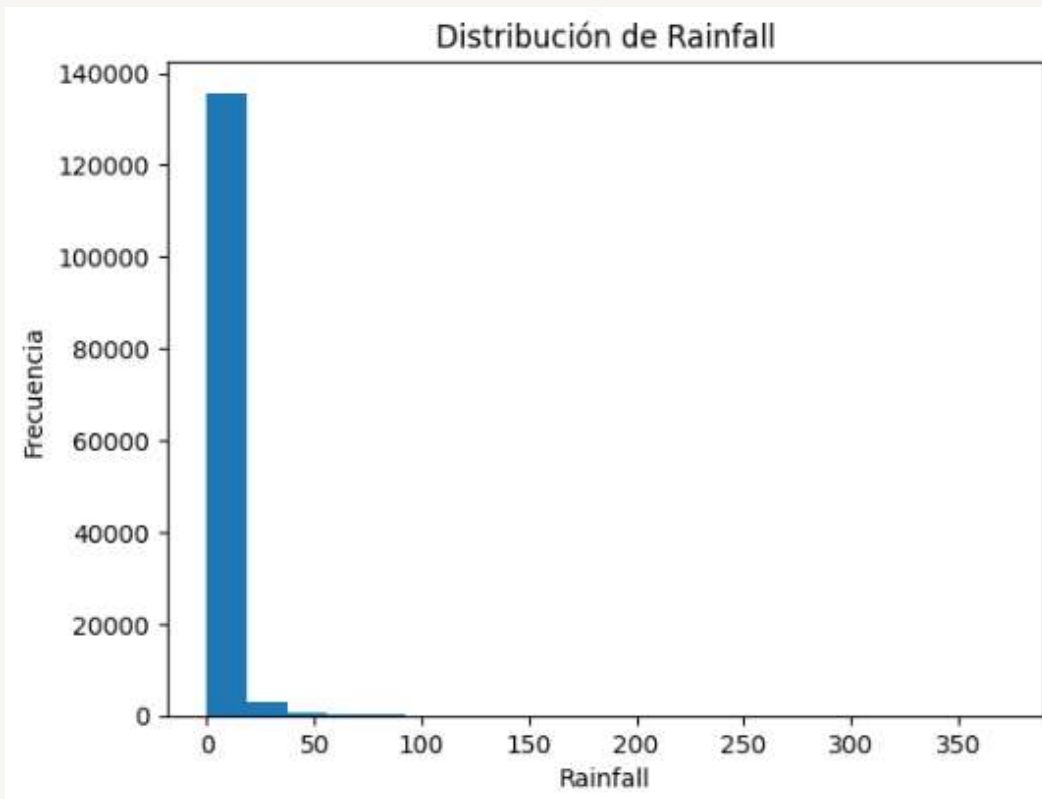
A continuación, se muestra el siguiente gráfico de torta o pie con el que intentamos representar la distribución de temperaturas máximas en Australia, agrupando en intervalos 20°C o menor, 20°C a 25°C, 25°C a 30°C, 30°C a 35°C y 35°C o mayor. A modo de hipótesis, se puede inferir que las zonas que han presentado temperaturas más bajas son las zonas en las que más se puede presentar lluvia y a modo contrario, las zonas que han presentado mayores temperaturas se pueden deducir que son las localidades más áridas. También a modo de hipótesis podríamos decir que más del 64% de la temperatura

Temperaturas

El histograma muestra que la mayoría de los días tienen una diferencia de temperatura máxima y mínima entre 5°C y 15°C, ya que la mayor concentración de los datos se encuentra en ese rango. También se puede apreciar que la distribución es asimétrica con valores atípicos, lo que significa que la diferencia de temperaturas en algunos días es muy alta. Además, en la parte izquierda del gráfico hay una pequeña cantidad de días en los que la temperatura es muy baja. Se podría deducir que en la gran mayoría de las localidades de Australia hay una variación moderada entre la temperatura máxima y mínima. Donde también se observa una pequeña cantidad de valores extremos de la distribución de los datos que indica que en algunas zonas la diferencia de temperatura es muy alta.

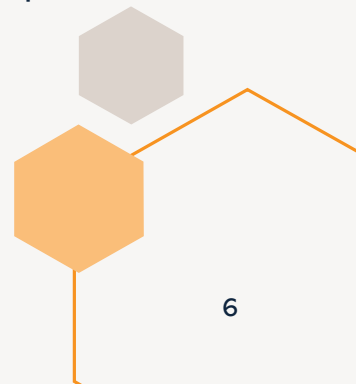


Datos Anómalos



Clima

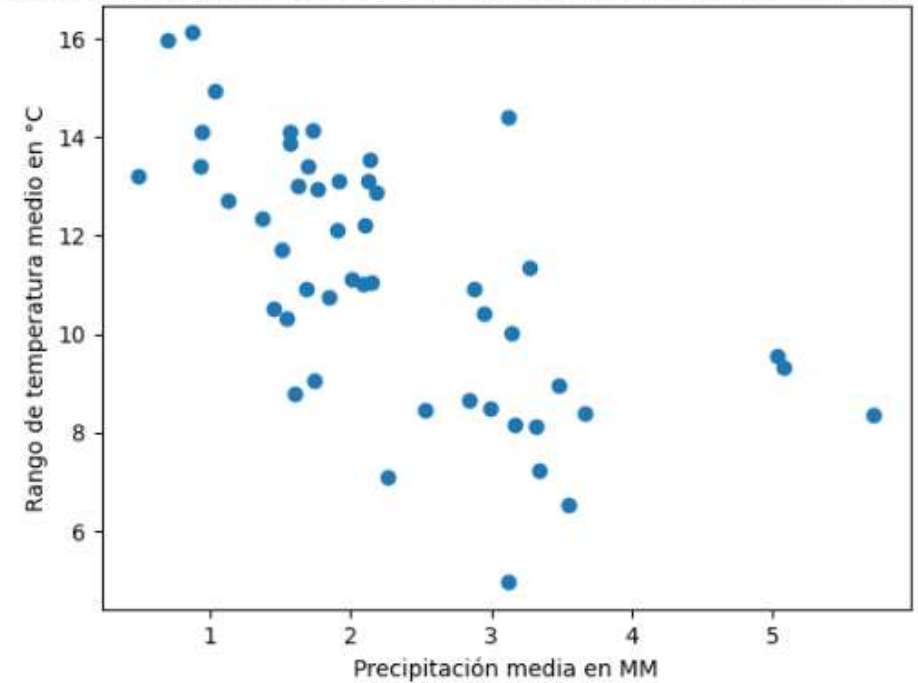
Se puede apreciar que la variable Rainfall tiene una distribución atípica con muchos datos anómalos debido a que en 2 o 3 localidades cuando llueve, lo hace en grandes cantidades a diferencia del resto de las ciudades donde la precipitación se acerca mucho a 0. Las localidades donde más lluvia se presenta encuentran en las zonas costeras del país. lo que explica que exista semejante diferencia pues a diferencia del clima presente en las costas y debido a su geografía, gran parte del resto del país corresponde a zonas desérticas rodeadas por dunas de arena.



Precipitación y rango temperatura

Se puede observar que no hay una correlación clara entre las cantidades de precipitación y el rango de temperatura en las diferentes ciudades. Porque hay algunas ciudades que tienen una cantidad relativamente alta de precipitaciones, pero también tienen un rango de temperatura medio alto, mientras que otras ciudades con una baja cantidad de precipitación tienen un rango de temperatura medio bajo. Esto podría ser que hay otros factores que influyen en la variación de la temperatura en diferentes ciudades.

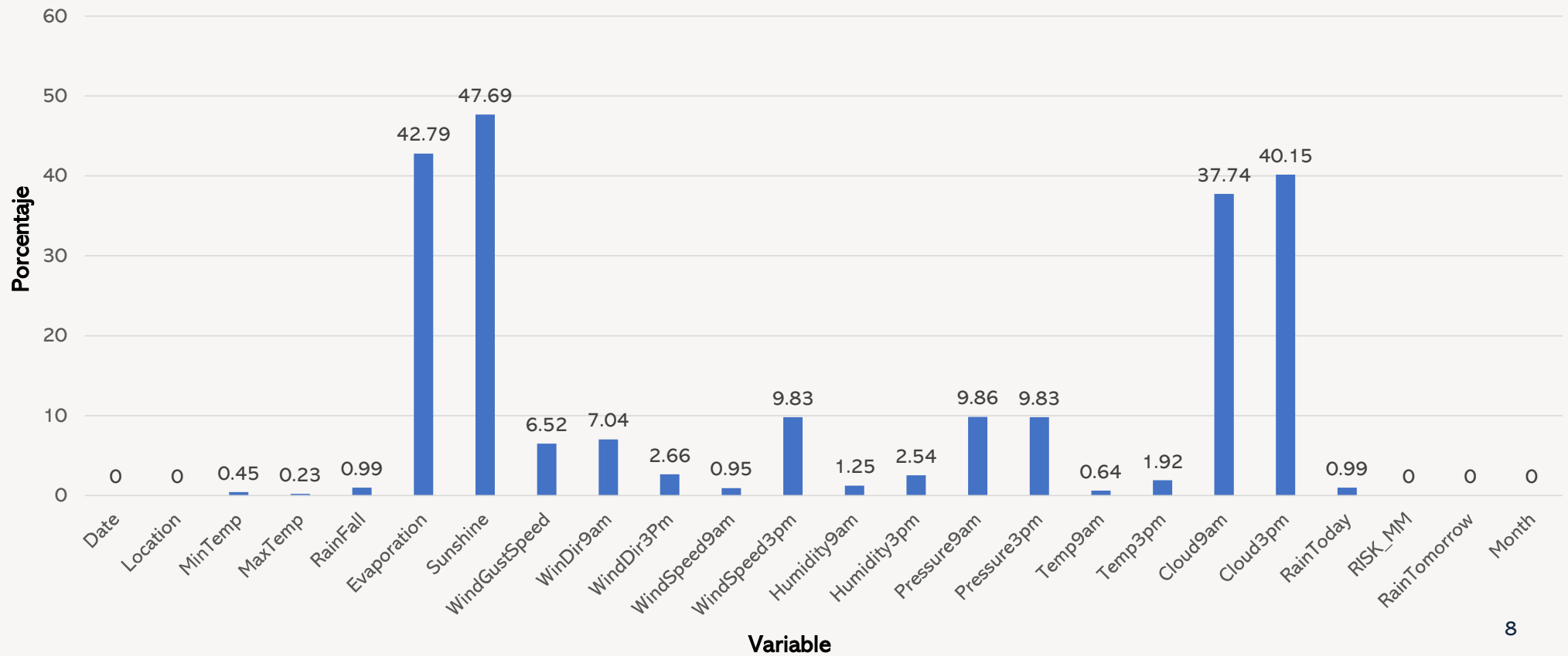
Relación entre la precipitación y el rango de temperatura en diferentes ciudades

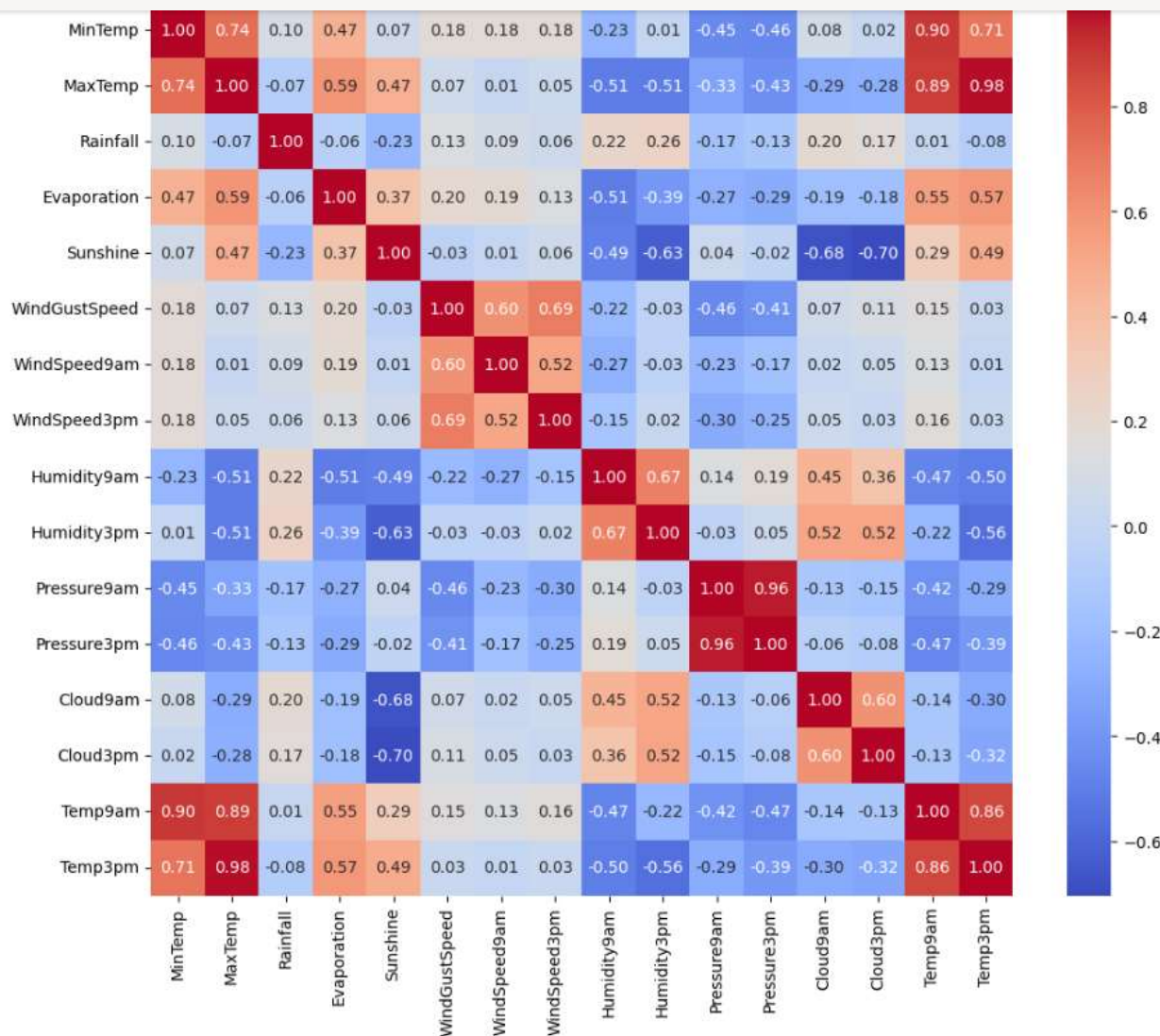


Limpieza de Valores Nulos

Determinar el porcentaje y cantidad de valores nulos para cada columna.

Porcentaje de Valores Nulos de cada Variable





Relación

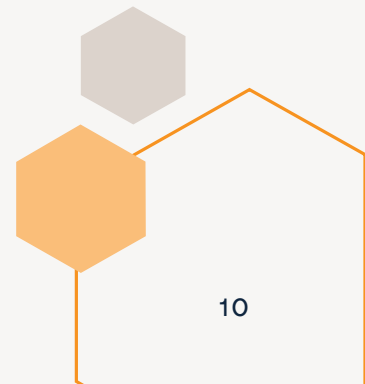
El fin del siguiente gráfico es representar la correlatividad entre la temperatura máxima y las precipitaciones, la variable RainFall(cantidad de lluvia caída ese día) y Sunshine(número de horas de sol brillante en el día) estas variables no tienen mucha relación entre sí, en base al mapa, debería ser una relación del -0.2 , o sea que dependiendo de la cantidad de lluvia que caiga ese día, será el número de horas de sol que exista ese día. Esto cambia cuando hablamos de las variables MaxTemp y Temp3Pm, deja en evidencia que a las 3 de la tarde se ha presenciado la más alta temperatura y todo en base a la relación positiva que existe según el gráfico.

También podemos ocupar el gráfico para interpretar que en Australia no existe mucha nubosidad, tiende a tener resultados más negativos que positivos y de acuerdo a la variable Rainfall se puede afirmar que Australia es un país semiarido

Detección de nulos

Columna	Porcentaje	Cantidad
DATE	0.0	0
Location	0.0	0
MinTemp	0.45	637
Evaporation	42.79	60843
RainToday	0.99	1406
WindSpeed9am	0.95	1348
WindGustSpeed	6.52	9270
Sunshine	47.69	67816

Es útil para determinar si hay variables que contengan una gran cantidad de valores nulos, si hay variables con un alto porcentaje de valores nulos, es recomendable excluir esas variables y no tomarlas en consideración.



Transformación de datos

RainToday	Location Raintoday RainTomorrow	MinTemp MaxTem	RainTomorrow	RainFall WindSpeed WindGustSpeed
Decidimos reemplazar los valores nulos de RainToday con la moda ya que al ser una variable categórica es una forma común de tratar valores nulos	Se utilizo labelEncoder solamente para simplificar la variable categórica y tener una representación numérica de la misma	Aquí se calcula primero la media de las variables MinTemp y MaxTemp, para luego crear una copia del dataframe original y reemplazar los valores nulos por la media de las variables	Reemplazamos los valores nulos con la moda ya que al ser una variable categórica es una forma común de tratar valores nulos , ya que la moda representa el valor más frecuente en la distribución de la variable	Reemplazo de los valores nulos por la mediana en estas variables debido a que el porcentaje de valores nulos es bajo . Esta técnica se utiliza para evitar que los valores faltantes tengan un impacto significativo en los modelos


Transformación de datos

Temp9am Temp3pm	Pressure9am Pressure3pm	Demasiados nulos	WindGustDir WindDir9am, WindDir3pm
<p>algoritmo KNNImputer para reemplazar los valores nulos. Estas variables tienen aproximadamente un 2% de valores nulos, y hemos decidido utilizar esta técnica con el fin de asignar un valor a cada campo nulo, el cual será designado realizando una comparación con sus cinco vecinos más cercanos.</p>	<p>Técnica de interpolación lineal se utiliza la información disponible de los valores observados antes y después de los valores faltantes para estimar los valores intermedios.</p>	<p>Las variables Cloud9am = 37.74 % Esto siendo equivalente a: 53657 datos nulos, Cloud3pm = 40.15 % Esto siendo equivalente a: 57094 datos nulos, Evaporation = 42.79 % Esto siendo equivalente a: 60843 datos nulos y Sunshine = 47.69 % Esto siendo equivalente a: 67816 datos nulos, no las tomaremos en cuenta dado a la gran cantidad de nulos presentes en el dataframe.</p>	<p>Para las variables categóricas WindGustDir, WindDir9am y WindDir3pm, se aplicó el método de simple imputer con la estrategia most frequently, el cual realiza un remplazo de los valores nulos por los más frecuentes dentro de la data.</p>

Transformación de datos

LabelEncoder

Las variables categóricas Localidad, WindGustDir, WindDir9am, WindDir3pm, RainToday, RainTomorrow se aplicó el método LabelEncoder, el cual reemplaza los valores categóricos por numéricos, asignando por ejemplo un numero a cada localidad, o un valor 0 y 1 a las variables llueve hoy o llueve mañana.



¿TIENE RELACIÓN LA CANTIDAD DE LLUVIA
QUE CAYÓ DURANTE EL DÍA Y LA
TEMPERATURA MÁXIMA PRESENTADA, CON LA
PROBABILIDAD DE QUE LLUEVA MAÑANA?

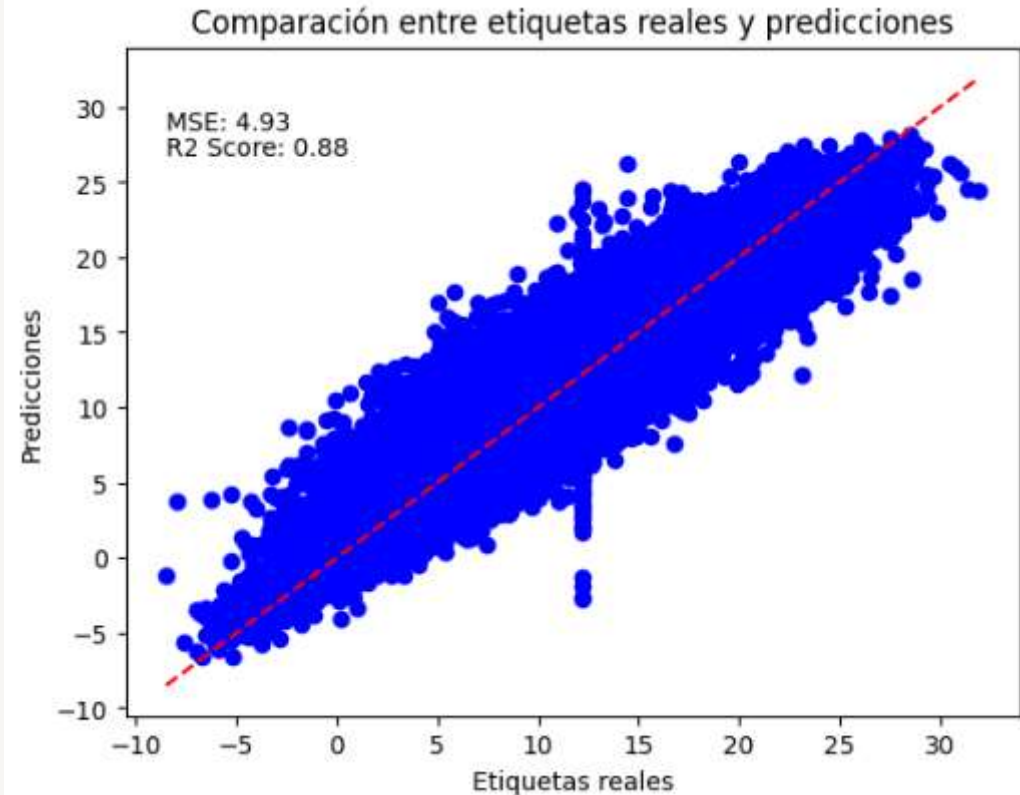
¿CUÁL ES LA UBICACIÓN Y EN QUE PERIODOS
SE PRESENTAN MÁS RIESGOS DEBIDO A LA
CANTIDAD DE LLUVIA CAÍDA?

¿PUEDEN LAS VARIABLES CLIMÁTICAS COMO
LA HUMEDAD, VELOCIDAD DEL VIENTO Y LA
PRESIÓN PRESENTADA, AYUDARNOS A
PREDECIR CUAL SERÁ LA TEMPERATURA
MÍNIMA DEL DÍA?

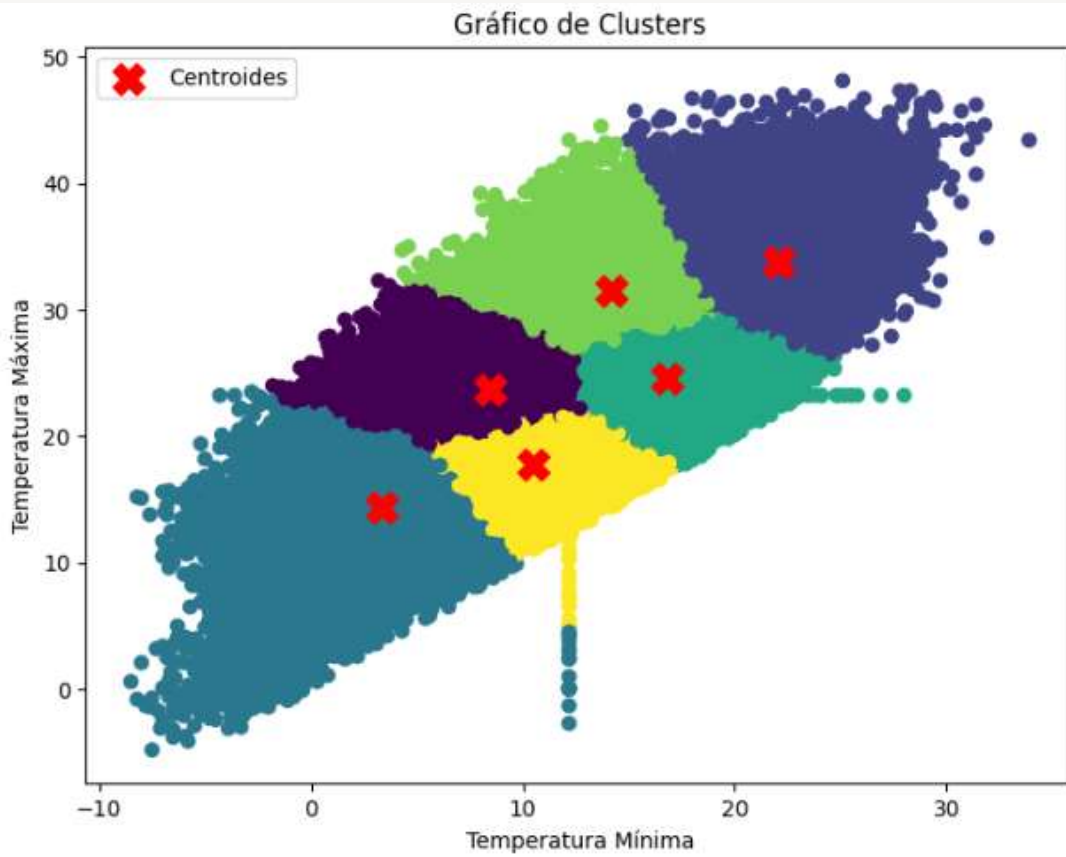
Objetivos Para alcanzar

Modelo regresión

- Este modelo demostró un rendimiento superior en la predicción de la temperatura mínima, con un MSE más bajo y un coeficiente de determinación (R2) más alto. Su capacidad para capturar relaciones complejas y no lineales entre las variables climáticas lo hace ideal para este propósito.
- modelo de regresión Random Forest muestra un buen rendimiento en la predicción de la temperatura mínima. Con un MSE de 4.93, un RMSE de 2.22 y un R2 de 0.88, el modelo demuestra una alta capacidad para ajustarse a los datos reales, capturando una gran parte de la variabilidad de los datos.



Modelo K-Means



El modelo k-Means agrupó efectivamente los datos en 6 clusters distintos, identificando patrones naturales basados en las temperaturas mínimas y máximas. Cada cluster representa un grupo de puntos con características similares, y los centroides proporcionan información sobre el promedio de las temperaturas dentro de cada cluster. Este análisis puede ser útil para identificar comportamientos climáticos y realizar predicciones más precisas basadas en las características agrupadas.



Gracias