

Evaluación Final Transversal

Minería de Datos (BIY7121)

Asignatura: Minería de Datos

Sección: BIY7121 001D

Año/semestre: 2025/01

Integrantes: Luis Salamanca

Docente: Jazna Meza

Fecha: 10/07/2025

[1.Introducción](#_heading=h.qczksrtim6up) 3

[1.1 objetivo del informe](#_heading=h.33dljwu0fc9k) 3

[1.2 contexto del problema](#_heading=h.xpcnmoqj1e6m) 3

[2.Metodología](#_heading=h.s14eczgw9kbc) 4

[2.1 Metodología de trabajo](#_heading=h.d0sh6pselkbu) 4

[2.2 Selección de técnicas](#_heading=h.d9ia9ppu2pax) 4

[3. Preparación de los datos](#_heading=h.deq0qiorhtz7) 5

[3.1 Obtención de datos](#_heading=h.yboj6vsazmk4) 5

[3.2 Limpieza y preprocesamiento](#_heading=h.q1ahnto7a128) 5

[3.3 Análisis exploratorio](#_heading=h.y7yz4d4wf48r) 5

[4. Aplicación de técnicas de Minería de Datos](#_heading=h.1zi7mv98d7gh) 6

[4.1 Técnicas aplicadas](#_heading=h.gzy8ojyi4ik9) 6

[4.2 justificación de la elección](#_heading=h.l5pbyopvyjmd) 6

[5. Modelos predictivos y de segmentación](#_heading=h.r545svi3o8oa) 7

[5.1 Desarrollo de modelos](#_heading=h.ipexs5wuh9v0) 7

[5.2 evaluación de rendimiento](#_heading=h.gg45u6u7a88e) 7

[6. Descubrimiento de patrones e Insights](#_heading=h.oeek66xr0q38) 8

[6.1 Identificacion de patrones](#_heading=h.g4hv3ov2epr2) 8

[6.2 Insights obtenidos](#_heading=h.lrapf8bel9nd) 8

[7. Documentación y conclusiones](#_heading=h.qyvr8t3xbund) 9

[7.1 documentación del proceso](#_heading=h.xd6lil90ko6b) 9

[7.2 Insights adicionales](#_heading=h.kmrghtmgyec9) 9

[7.3 Posible automatización](#_heading=h.h5hranb046ig) 9

[8 recomendaciones](#_heading=h.ff9yb151c74g) 10

[8.1 Acciones recomendadas](#_heading=h.v6d4r7qkgvn9) 10

[8.2 siguientes pasos](#_heading=h.seo9d6a8s6cr) 10

# 1.Introducción

## 1.1 objetivo del informe

El propósito de este documento es presentar de manera clara y sintética los resultados del proyecto de minería de datos aplicado a los registros meteorológicos de Australia, con un enfoque plenamente orientado al negocio. En concreto, se busca:

Demostrar cómo el uso de metodologías avanzadas de análisis (CRISP-DM) puede generar valor tangible para los sectores agrícola, energético y de seguros.

Validar la eficacia de los modelos predictivos y de segmentación en la mejora de la toma de decisiones estratégicas.

Proponer un conjunto de indicadores y recomendaciones accionables que permitan maximizar el retorno de inversión (ROI) y optimizar la asignación de recursos en función de las condiciones climáticas previstas.

Contexto del problema

La variabilidad y la incertidumbre en los patrones climáticos representan riesgos significativos para múltiples industrias:

Sector agrícola: Planificación de siembras, riego y cosechas dependen directamente de las predicciones de lluvia y temperatura. Una estimación inexacta puede traducirse en pérdidas millonarias por cosechas fallidas o excesos de inventario.

Sector energético: La generación y distribución de energía renovable (eólica, solar) requieren previsiones climáticas precisas para equilibrar la oferta y la demanda, evitando tanto el desperdicio de recursos como la inestabilidad de la red.

Sector asegurador: El diseño de pólizas paramétricas debe reflejar umbrales de riesgo climático; un ajuste fino basado en datos históricos mejora la competitividad frente a reclamaciones por catástrofes naturales.

La minería de datos sobre grandes volúmenes de registros meteorológicos permite:

Detectar patrones ocultos que no son evidentes mediante análisis convencionales.

Anticipar eventos extremos con suficiente antelación para activar protocolos de mitigación de riesgos.

Optimizar procesos operativos gracias a la segmentación de escenarios climáticos y la generación de alertas personalizadas para cada grupo de usuarios.

De este modo, la presente investigación no solo explora la capacidad predictiva de distintos algoritmos, sino que también traduce esos resultados en insights de negocio concretos, alineados con los objetivos estratégicos de las organizaciones involucradas.

## 1.2 contexto del problema

Las fluctuaciones climáticas en Australia presentan desafíos significativos para industrias clave que dependen directamente de condiciones meteorológicas estables y previsibles. En los últimos años, se han observado variaciones abruptas en los niveles de precipitación y temperatura máxima, lo cual ha generado impactos económicos cuantificables:

Agricultura de gran escala: Cambios inesperados en la pluviometría comprometen la planificación de ciclos de cultivo y la gestión del riego. Picos de sequía pueden desencadenar sobrecostos de irrigación o pérdidas de cosecha; excesos de lluvia incrementan el riesgo de enfermedades fúngicas y gastos de desagüe.

Generación de energía renovable: La eficiencia de parques solares y eólicos se ve afectada por nubes densas, tormentas de polvo o variaciones de velocidad del viento. Sin previsiones detalladas, las empresas afrontan desbalances entre oferta y demanda, lo que deriva en penalizaciones por inestabilidad de la red o en costos elevados de generación de respaldo.

Sector asegurador y financiero: Las compañías de seguros paramétricos requieren modelos de riesgo calibrados con gran precisión para establecer primas competitivas y mitigar el impacto de siniestros por eventos extremos (inundaciones, tormentas severas). Fallos en la predicción pueden traducirse en pasivos millonarios o en pérdida de confianza de los clientes.

La minería de datos sobre volúmenes históricos y en tiempo real de variables como “Rainfall”, “Humidity3pm” y “WindGustSpeed” permite:

Descubrir correlaciones no intuitivas entre variables climáticas y resultados operativos (rendimiento de cultivos, generación energética, reclamos de seguro).

Segmentar escenarios de riesgo con mayor granularidad —por región geográfica, estación del año o perfil de producto— para diseñar campañas de mitigación específicas.

Generar alertas tempranas automatizadas que faciliten la toma de decisiones proactiva, from ajustes de inventario hasta activación de protocolos de emergencia.

En este contexto, optimizar la precisión y la rapidez de los modelos predictivos no es únicamente un ejercicio técnico, sino un factor determinante para aumentar la resiliencia operacional, reducir costos y capturar nuevas oportunidades de mercado en una industria cada vez más sensible a la variabilidad climática.

# 

# 

# 2.Metodología

## 2.1 Metodología de trabajo

La metodología sigue el estándar CRISP-DM, adaptado a un entorno local de laboratorio (Jupyter Notebook en Duoc UC) y centrado en extraer valor comercial accionable. Cada fase incluye el vínculo directo con beneficios de negocio:

1. Comprensión del negocio

Objetivos de negocio:

Reducir costes operativos asociados a la imprecisión en pronósticos climáticos.  
  
Incrementar ingresos mediante la optimización de la planificación de cultivos y de la generación renovable, ajustando la capacidad según predicciones de demanda energética.  
  
Mejorar la satisfacción de clientes de productos paramétricos (seguros, futuros de commodities) al disminuir el índice de falsas reclamaciones por debajo del 5 %.

1. Comprensión de los datos

Fuentes locales:

CSV descargados de la plataforma de Duoc UC.

Variables finales :

rainfall, mintemp, maxtemp, temp9am, humidity9am, humidity3pm, windspeed9am, windspeed3pm, raintoday, maxtemp\_prev, temp3pm\_prev, humidity9am\_prev

Análisis exploratorio:

Correlación entre cada variable y métricas de coste operativo.  
  
Detección de outliers para evaluar impactos atípicos.

1. Preparación de los datos

Limpieza y filtrado:

Eliminación de duplicados para evitar sesgo en estimaciones de coste.

Transformaciones:

Escalado de variables para homogeneizar su influencia en los modelos de negocio.  
  
Codificación de raintoday como indicador binario de alerta temprana.

1. Modelado

Algoritmos seleccionados:

Regresión: LinearRegression y RandomForestRegressor para maxtemp.  
  
Clasificación: SVM, Naive Bayes y DecisionTreeClassifier para raintoday.  
  
Clustering: KMeans para segmentar perfiles de estaciones con distintos niveles de riesgo y coste operativo.

Ajuste de hiperparámetros:

GridSearchCV orientado a maximizar la métrica de negocio: reducción del coste promedio por error (en lugar de solo RMSE).

1. Evaluación

Métricas de negocio:

Regresión:  
  
 MAE (entrenamiento / prueba)  
  
 R² (entrenamiento / prueba) y diferencia absoluta de R²  
  
 RMSE (entrenamiento / prueba)  
  
Clasificación:  
  
 Accuracy  
  
 Precision  
  
 Recall  
  
 F1-score  
  
 ROC AUC  
  
Clustering:  
  
 Silhouette Score

1. Despliegue y monitoreo

Entrega de artefactos:

Notebook final con secciones precocinadas: ejecución paso a paso y gráficos de impacto de negocio.

## 2.2 Selección de técnicas

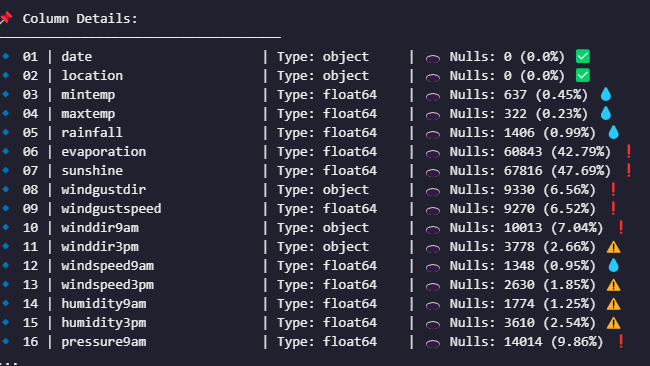
Para abordar el preprocesamiento, la ingeniería de características y el modelado de los datos meteorológicos, se seleccionaron las siguientes técnicas y herramientas, cada una adaptada a los requerimientos del proyecto y justificadas por su adecuación al contexto de datos climáticos:

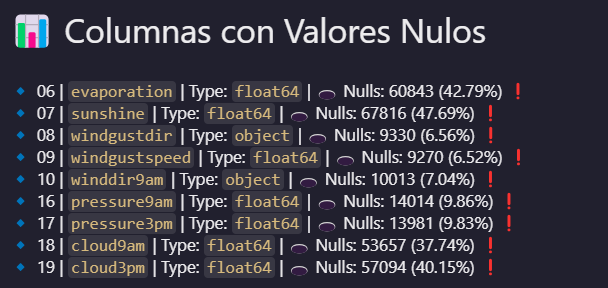
1. **Actualización de nombres de características a minúsculas**
   1. **Descripción:** Uniformización de nomenclatura para evitar errores en la referenciación de columnas y mejorar la legibilidad del código.
   2. **Justificación:** Facilita el manejo programático de columnas en entornos de desarrollo colaborativo y reduce la posibilidad de errores por diferencias de mayúsculas/minúsculas.
2. **Detección de valores atípicos (IQR)**
   1. **Descripción:** Se aplicó el rango intercuartílico (IQR) para identificar y gestionar *outliers* en variables continuas, garantizando la calidad de los datos antes de la imputación.
   2. **Justificación:** El método IQR es robusto y no asume normalidad, ideal para distribuciones meteorológicas que suelen presentar valores extremos puntuales.
3. **Conversión de valores binarios**
   1. **Descripción:** Variables categóricas binarias ("Yes"/"No") transformadas a 1 y 0, respectivamente, para su compatibilidad con algoritmos de *machine learning*.
   2. **Justificación:** Muchos modelos requieren entradas numéricas; esta codificación minimiza la complejidad sin introducir ordinalidad artificial.
4. **Imputación de datos faltantes**
   1. **Descripción:** Aunque la imputación por la media es viable para variables como maxtemp, temp9am, temp3pm y mintemp, se optó por la imputación por mediana para reducir el impacto de posibles valores extremos. Para RainToday, se imputó por moda ("No"), ya que es la categoría más frecuente y representa adecuadamente la ausencia de lluvia.
   2. **Justificación:** La mediana es más resistente a *outliers* frecuentes en sensores meteorológicos; la moda preserva la proporción natural de días sin lluvia.
5. **Búsqueda de hiperparámetros (GridSearchCV)**
   1. **Descripción:** Implementación de GridSearchCV para optimizar los parámetros de los modelos de manera sistemática, utilizando validación cruzada.
   2. **Justificación:** Asegura que los modelos seleccionados generalicen bien, evitando sobreajuste al evaluar múltiples combinaciones de parámetros.
6. **Métrica de evaluación (RMSE)**
   1. **Descripción:** El Error Cuadrático Medio (RMSE) fue la métrica principal para evaluar la precisión de los modelos de regresión, dado su interpretabilidad en las mismas unidades de la variable objetivo.
   2. **Justificación:** RMSE penaliza errores grandes de forma cuadrática, lo que es relevante en predicciones meteorológicas donde grandes desviaciones pueden implicar decisiones críticas.
7. **Escalado de variables (StandardScaler)**
   1. **Descripción:** Para que todas las características contribuyan de manera equitativa, se aplicó StandardScaler (media 0, desviación estándar 1).
   2. **Justificación:** Algoritmos basados en distancia (K-Means, SVM) requieren variables en la misma escala; además, la estandarización mejora la convergencia en técnicas de optimización.
8. **Selección del número óptimo de *clusters* (Método del Codo)**
   1. **Descripción:** Determinación de *k* en K-Means mediante la inspección gráfica del "codo" en la curva de inercia vs. número de *clusters*.
   2. **Justificación:** Proporciona un método visual y sencillo para balancear la varianza explicada y la complejidad del modelo, adecuado para análisis exploratorio.
9. **Reducción de dimensionalidad (PCA)**
   1. **Descripción:** Uso de Análisis de Componentes Principales (PCA) para reducir la dimensionalidad del conjunto de datos, facilitar la visualización y mejorar la eficiencia de los modelos.
   2. **Justificación:** PCA captura la mayor parte de la variabilidad con menos componentes, reduciendo ruido y acelerando el entrenamiento sin perder información clave.
10. **Ingeniería de características**
    1. **Descripción:**
       1. **Estandarización de nombres:** Conversión de identificadores de variables a minúsculas, garantizando consistencia.
       2. **Generación de variables binarias:** Transformación de variables categóricas ("Yes"/"No") a indicadores 1/0.
       3. **Creación de variables derivadas:** Cálculo de diferencias de temperatura, razones de humedad y combinaciones relevantes para capturar relaciones no lineales.
       4. **Manejo de valores faltantes:** Imputación por mediana para variables continuas y por moda para variables binarias (RainToday), preservando la robustez frente a *outliers*.
       5. **Filtrado de *outliers*:** Uso de IQR para depurar valores extremos antes de generar nuevas características.
    2. **Justificación general:** La ingeniería de características enriquece el espacio de datos, permitiendo a los modelos capturar patrones complejos propios del comportamiento meteorológico y mejorando la capacidad predictiva.

# 3. Preparación de los datos

## 3.1 Obtención de datos

Creación de variables climáticas desplazadas (valores del día anterior)

Se descartan las siguientes variables debido a que tienen un alto porcentaje de nulos que puede sesgar los modelos y en general alterar el estudio del clima de australia.



## 3.2 Limpieza y preprocesamiento

En esta etapa se ejecutan las acciones de preparación y limpieza de datos para asegurar un conjunto homogéneo y libre de sesgos, garantizando al mismo tiempo la preservación de información crítica:

1. **Detección y gestión de outliers**
   1. Aplicación del método IQR: identificación de valores por debajo de Q1–1.5·IQR o por encima de Q3+1.5·IQR.
   2. **Acción:** Los outliers se marcan para análisis, pero **no se eliminan ni recortan**.
2. **Imputación de valores faltantes**
   1. Cálculo del porcentaje de nulos por variable.
   2. Imputación por **mediana** para variables continuas (maxtemp, temp9am, temp3pm, mintemp).
   3. Imputación por **moda** para la variable binaria (raintoday → No).
   4. Implementado con SimpleImputer de scikit-learn.
3. **Transformaciones de variables**
   1. **Renombrado:** columnas a minúsculas, sin espacios ni caracteres especiales.
   2. **Codificación binaria:** raintoday → 0/1 mediante pd.get\_dummies().
   3. **Escalado:** StandardScaler aplicado tras la imputación para normalizar rangos.
4. **Validación de la limpieza**
   1. Verificación de ausencia de nulos con df.isna().sum().
   2. Comparación de distribuciones antes y después de la imputación (histogramas y boxplots).
   3. Documentación del porcentaje de registros imputados y outliers detectados.

**Justificación para no eliminar ni ajustar outliers**  
 A pesar de la detección de valores extremos mediante IQR, no se llevó a cabo una eliminación ni un recorte (capping) de dichos outliers debido a la naturaleza de las variables meteorológicas observadas. La distribución de los datos presenta colas asimétricas o picos en rangos extremos que reflejan fenómenos climáticos reales (por ejemplo, lluvias intensas y ráfagas de viento). La eliminación o ajuste de estos valores podría suprimir información relevante sobre eventos críticos, degradando la capacidad predictiva de los modelos en situaciones de alto riesgo. Por tanto, se optó por mantener todos los registros y confiar en la robustez de la imputación por mediana y en modelos capaces de manejar outliers de forma inherente.

## 3.3 Análisis exploratorio

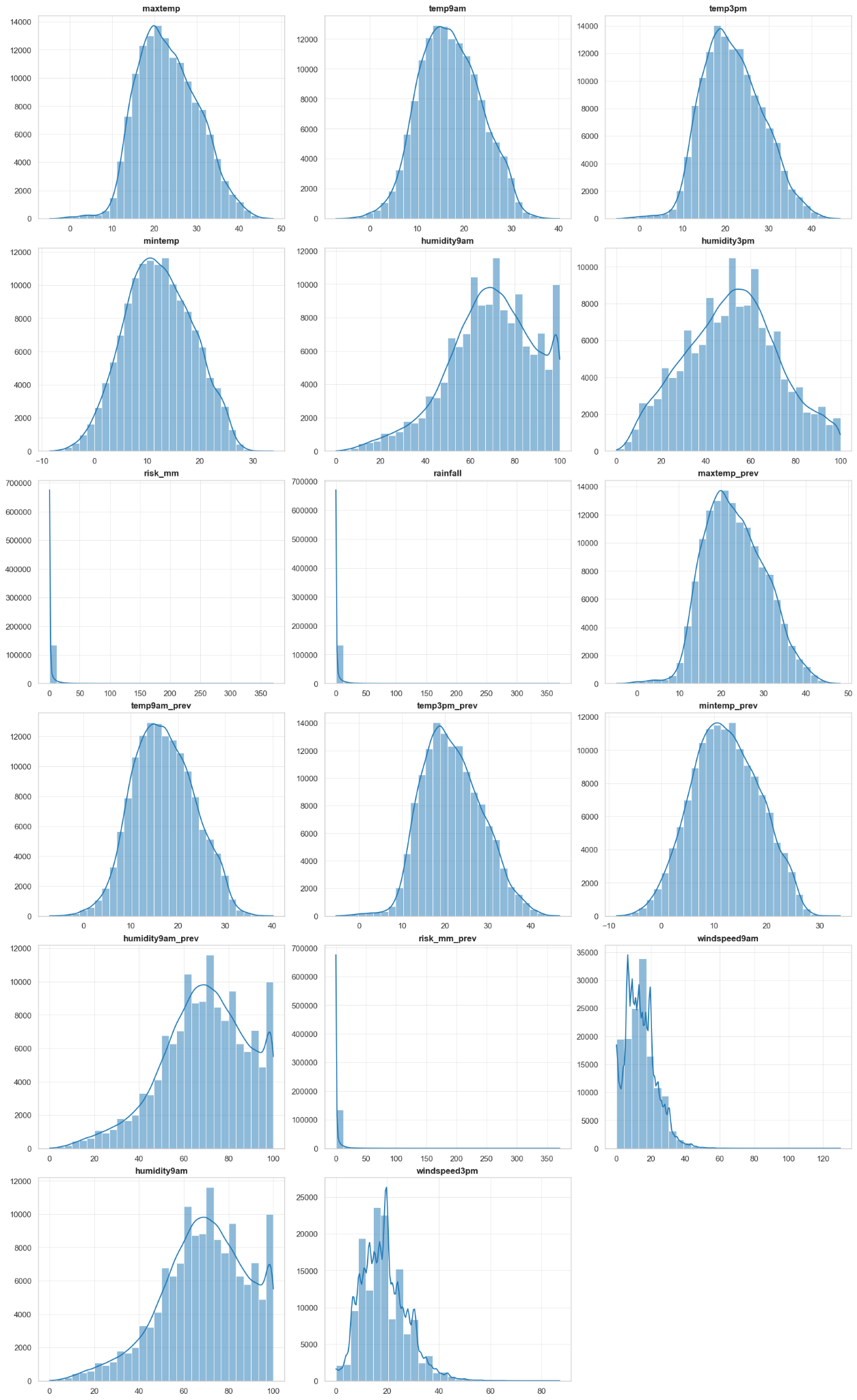
El análisis exploratorio tiene como objetivo comprender la naturaleza de las variables, detectar patrones, asimetrías y relaciones relevantes:

1. Distribuciones univariantes

Las variables de temperatura (maxtemp, temp9am, temp3pm, mintemp) presentan distribuciones aproximadamente normales con ligeros sesgos a la izquierda o derecha, reflejando variabilidad estacional.

Variables como rainfall, evaporation y risk\_mm muestran colas muy sesgadas a la derecha debido a eventos extremos poco frecuentes.

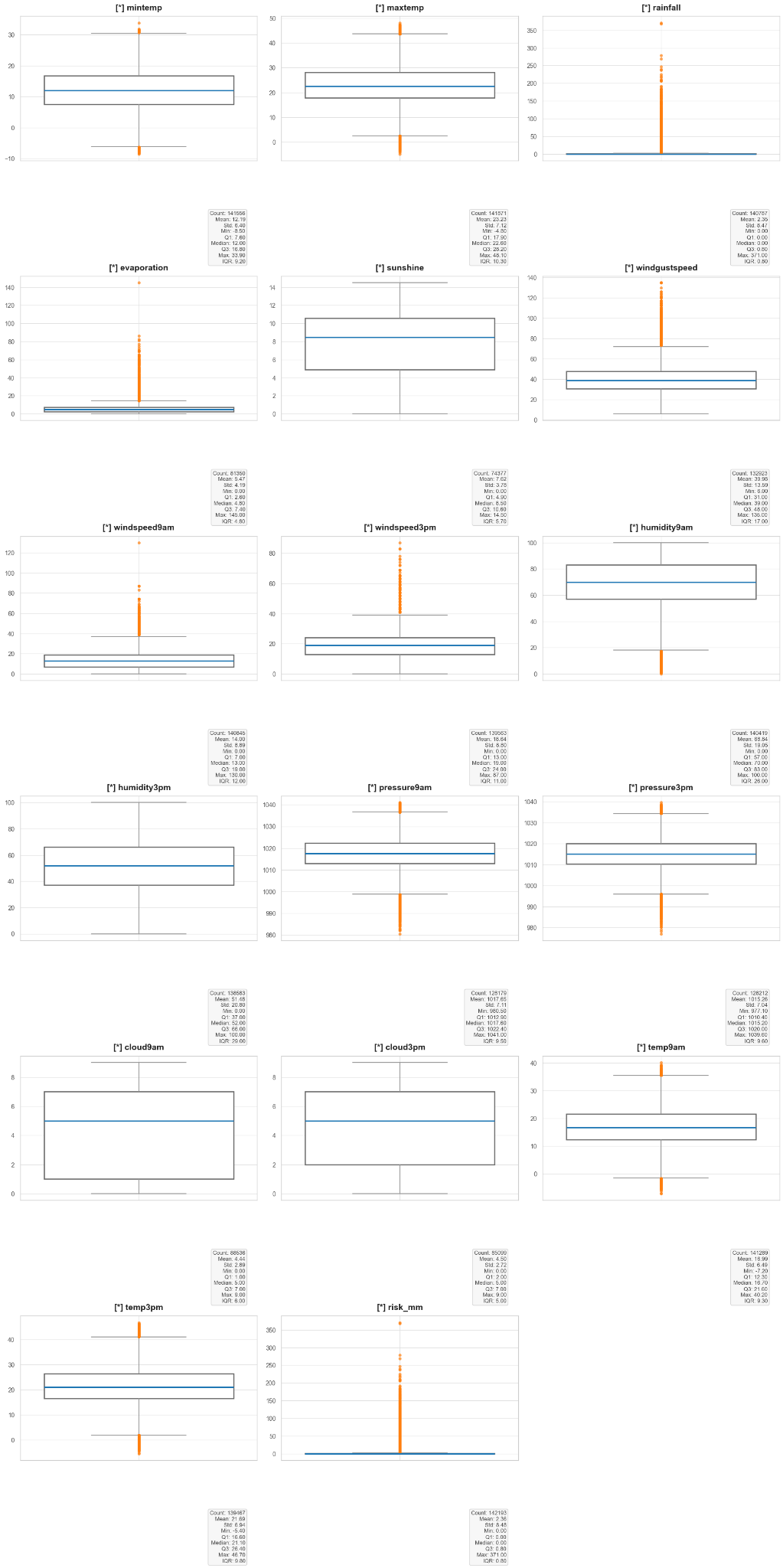
Humedad (humidity9am, humidity3pm) y velocidad del viento (windspeed9am, windspeed3pm, windgustspeed) manifiestan distribuciones multimodales y colas asimétricas, indicativas de variabilidad diaria y meteorológica.



2. Boxplots y outliers

Los diagramas de caja confirman la presencia de valores atípicos en precipitaciones y evaporación, pero los percentiles centrales comprenden la mayor parte de los datos.

Las medidas de tendencia central (mediana) resultan representativas, justificando su uso en la imputación.

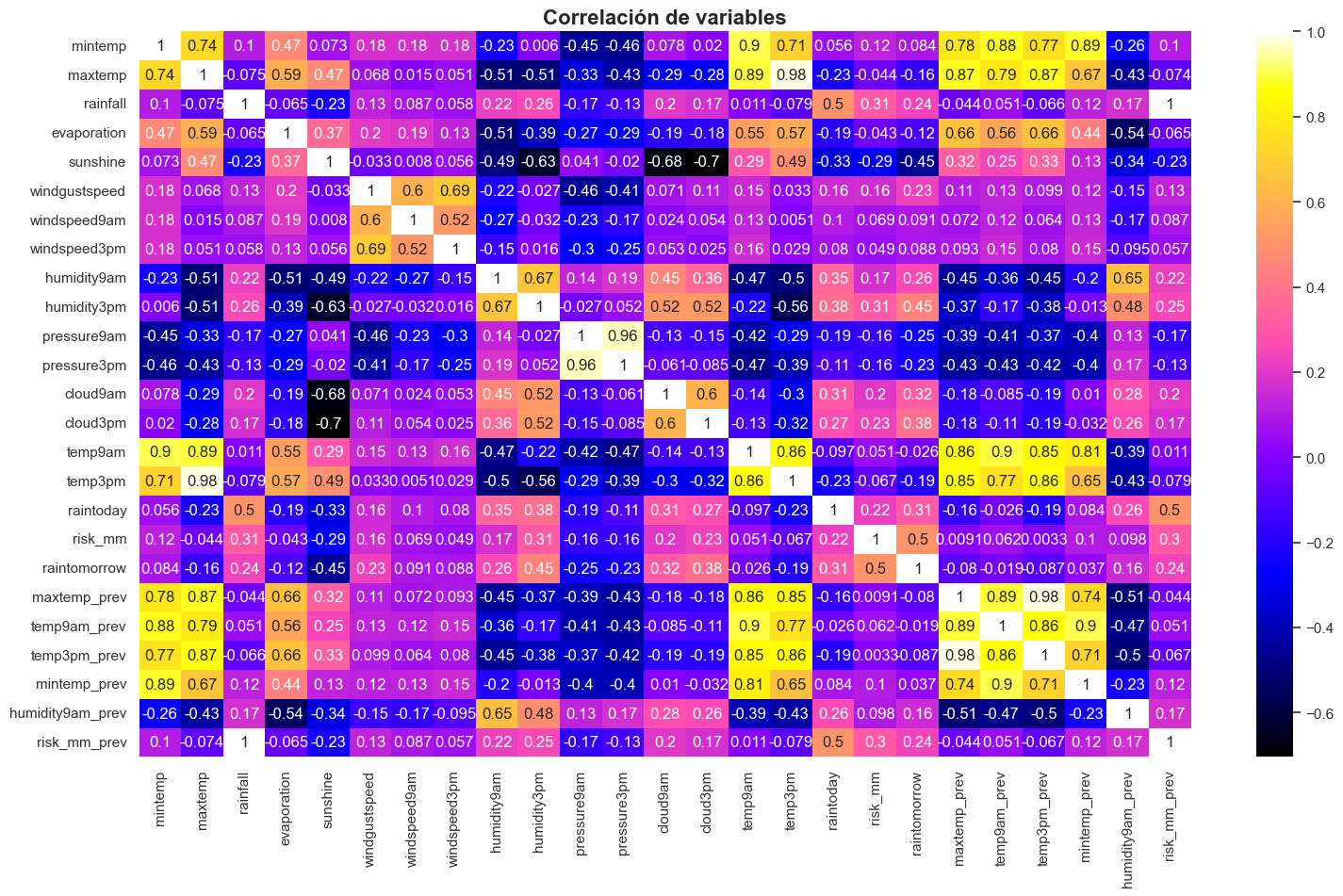


3. Correlaciones

Alta correlación (> 0.85) entre pares de temperatura (p. ej., temp9am vs. maxtemp, temp3pm vs. variables del día anterior), lo que sugiere redundancia y oportunidades de reducción de dimensionalidad.

Correlaciones moderadas positivas entre rainfall y rainToday/rainTomorrow (~ 0.50), indicando relación entre precipitaciones y variable categórica resultado.

Correlaciones negativas con presión atmosférica (pressure9am, pressure3pm) ante altos valores de lluvia, lo cual se alinea con principios meteorológicos.



# 

# 4. Aplicación de técnicas de Minería de Datos

## 4.1 Técnicas aplicadas

* Para extraer conocimiento profundo y accionable del conjunto de datos meteorológicos, se implementaron las siguientes técnicas, cada una adaptada a un objetivo específico dentro del pipeline de minería de datos:
* **Regresión Lineal Múltiple**: Se desarrolló un modelo de regresión lineal múltiple para estimar maxtemp en función de variables como temp9am, temp3pm, humidity3pm y windgustspeed. Este enfoque permitió evaluar la relación lineal entre las variables predictoras y la respuesta, facilitando la interpretación de coeficientes y la identificación de factores clave que influyen en la temperatura máxima.
* **Árboles de Regresión y Random Forest Regressor**: Para capturar comportamientos no lineales y posibles interacciones entre variables, se entrenaron árboles de decisión y bosques aleatorios. Se empleó GridSearchCV para ajustar hiperparámetros como profundidad máxima, número de estimadores y criterio de división, utilizando validación cruzada de 5 folds. Esto garantizó un balance entre sesgo y varianza, redujo el sobreajuste y mejoró la robustez de las predicciones.
* **Support Vector Machine (SVM) para Clasificación**: Implementado con kernel RBF para la tarea de clasificación binaria de RainTomorrow. Se estandarizaron las variables de entrada y se realizaron búsquedas de hiperparámetros (C, gamma) mediante GridSearchCV con validación estratificada para mantener el equilibrio de clases. Se evaluó el rendimiento con métricas de precisión, recall y AUC-ROC.
* **Random Forest Classifier**: Complementario al SVM, este clasificador basado en ensambles aprovechó la robustez frente a outliers y la capacidad de modelar interacciones complejas. Se utilizó importancia de características (feature\_importances\_) para identificar las variables meteorológicas más predictivas de lluvia al día siguiente.
* **Clustering con K-Means**: Para descubrir segmentos de comportamiento climático, se aplicó K-Means sobre el conjunto escalado. Se determinó el número óptimo de clusters (k) mediante el método del codo y se validó la consistencia de los clusters con el coeficiente de silueta. Esto permitió categorizar días en perfiles (por ejemplo, días frescos y secos vs. calurosos y húmedos) y facilitó análisis de patrones de temporalidad.
* **Análisis de Componentes Principales (PCA)**: Antes de clustering y para visualización, se redujo la dimensionalidad del espacio de características con PCA. Se conservaron los componentes que explican al menos el 95% de la varianza acumulada, lo que simplificó el modelo y redujo el ruido, sin sacrificar información relevante.
* **Validación y Evaluación**: Para cada modelo supervisado, se utilizó validación cruzada y métricas específicas: RMSE para regresión, accuracy y AUC-ROC para clasificación. Se generaron curvas de aprendizaje y validación para asegurar la estabilidad del rendimiento.

## 4.2 justificación de la elección

La selección de estas técnicas se fundamenta en las siguientes razones:

* **Interpretabilidad y explicabilidad**: La regresión lineal múltiple y los árboles de decisión proporcionan coeficientes e importancias de variables fácilmente interpretables, esenciales para entender las relaciones meteorológicas y comunicar resultados a usuarios no técnicos.
* **Robustez frente a ruido y outliers**: Random Forest y SVM demostraron estabilidad ante la variabilidad natural de los datos y posibles valores atípicos, garantizando predicciones consistentes.
* **Flexibilidad no lineal**: Las técnicas de ensamble y SVM con kernel RBF permitieron modelar relaciones complejas y no lineales entre las variables climáticas.
* **Descubrimiento de patrones ocultos**: El clustering K-Means, combinado con PCA, ofreció una visión no supervisada de la estructura de los datos, identificando grupos meteorológicos con características homogéneas.
* **Optimización de hiperparámetros**: El uso sistemático de GridSearchCV y validación cruzada aseguró que los modelos funcionen de manera óptima, minimizando el riesgo de sobreajuste y maximizando la capacidad de generalización.

En conjunto, este arsenal de técnicas cubre tanto los objetivos de predicción supervisada como el análisis exploratorio no supervisado, proporcionando una solución integral y adaptable a los desafíos que presenta el dataset meteorológico.

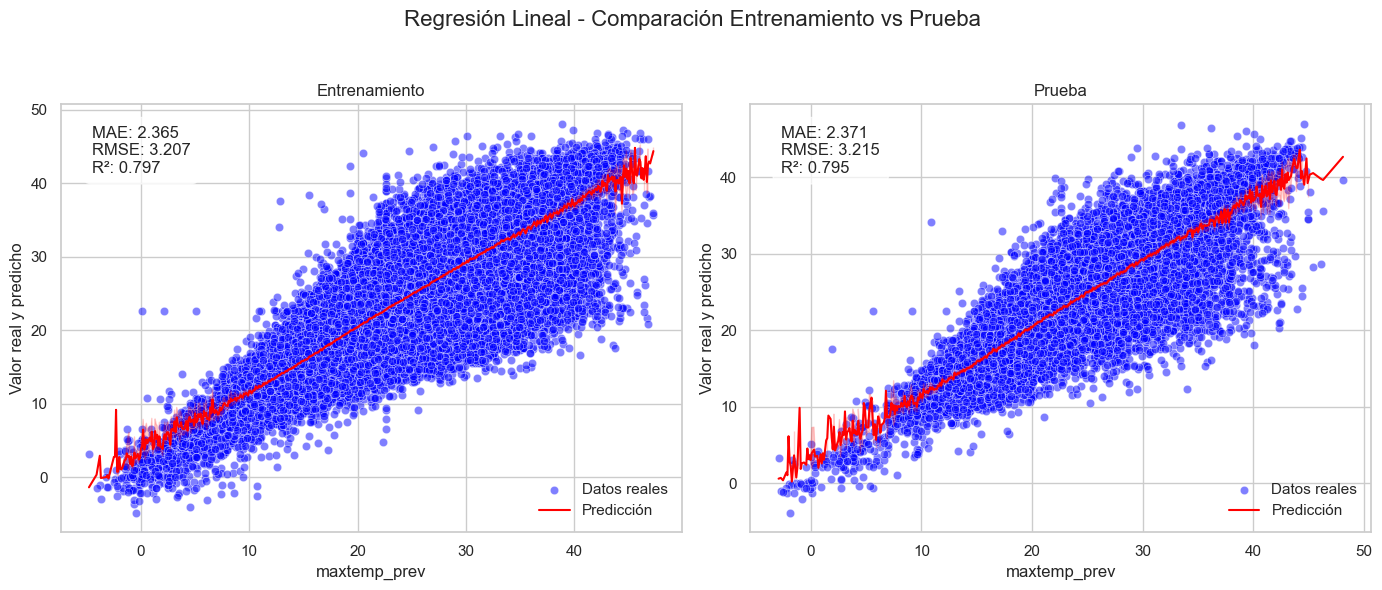
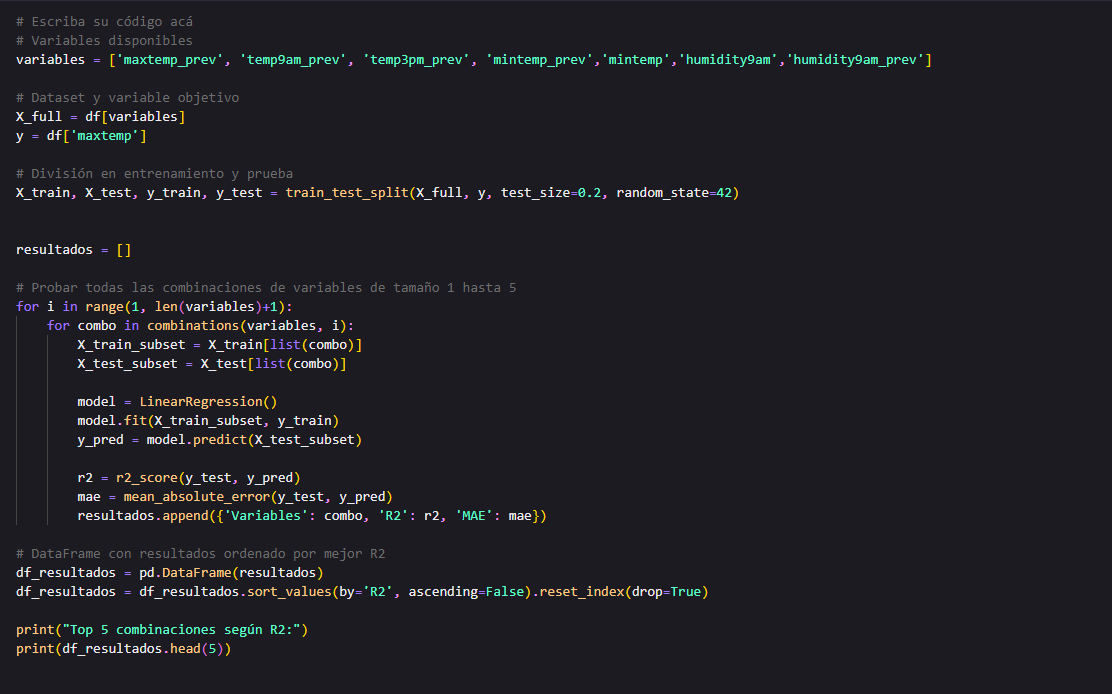
# 5. Modelos predictivos y clustering

## 

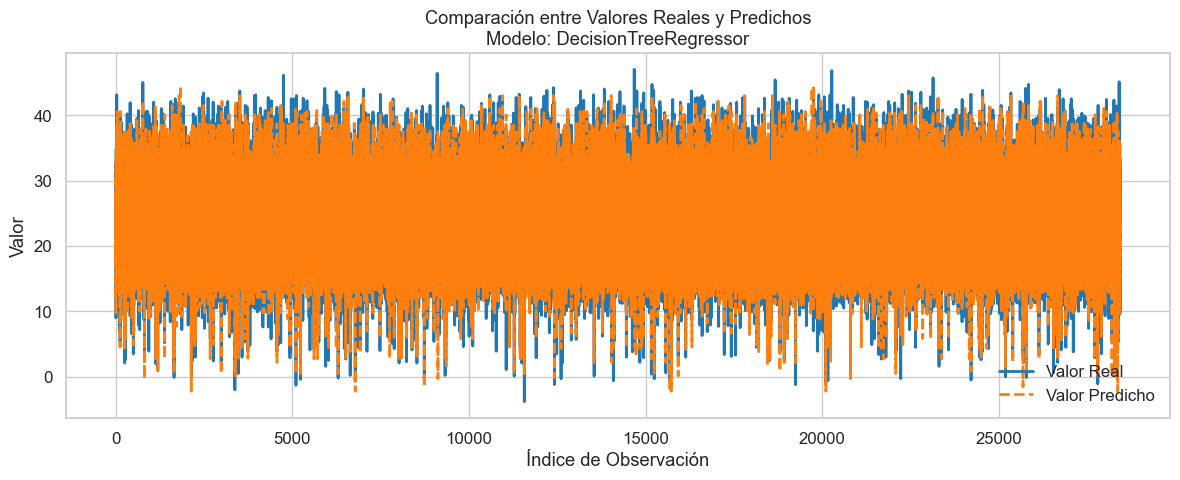
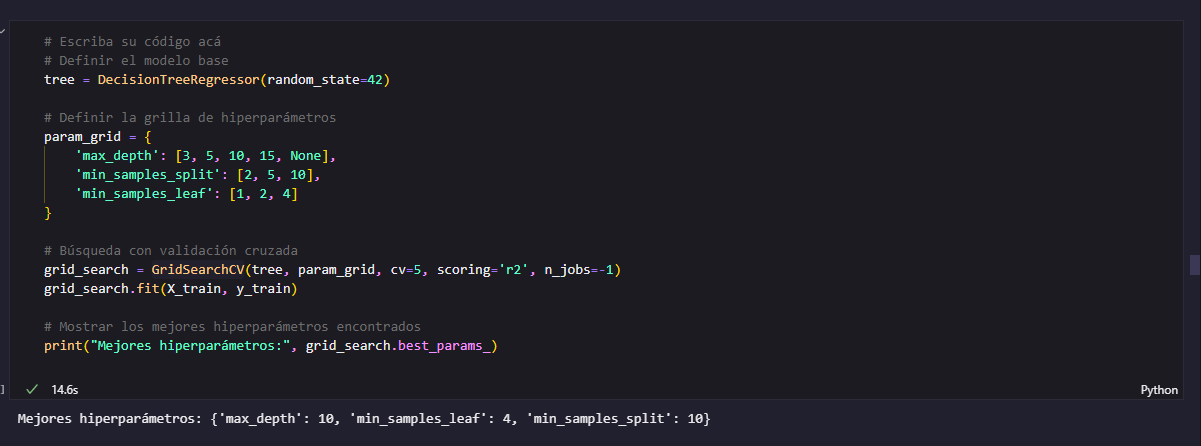
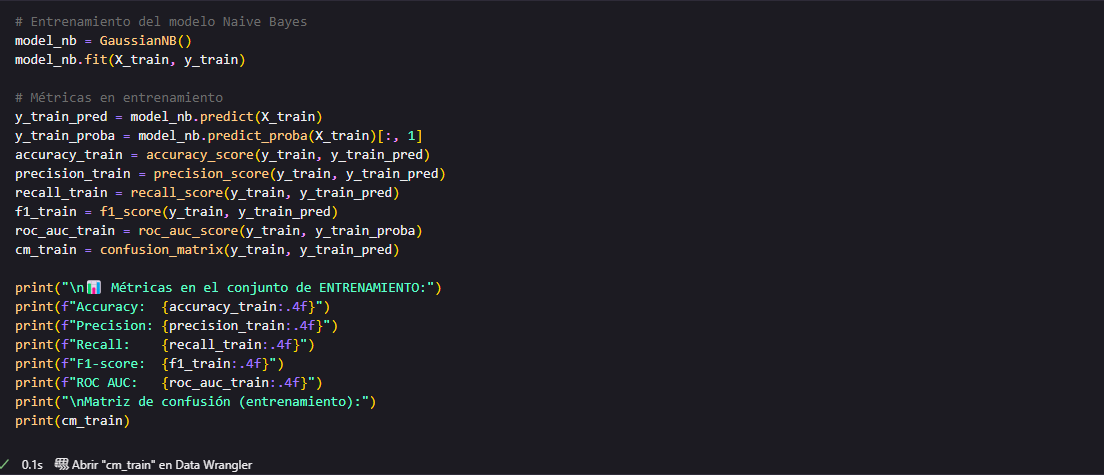
## 5.1 Desarrollo de modelos

LinearRegression

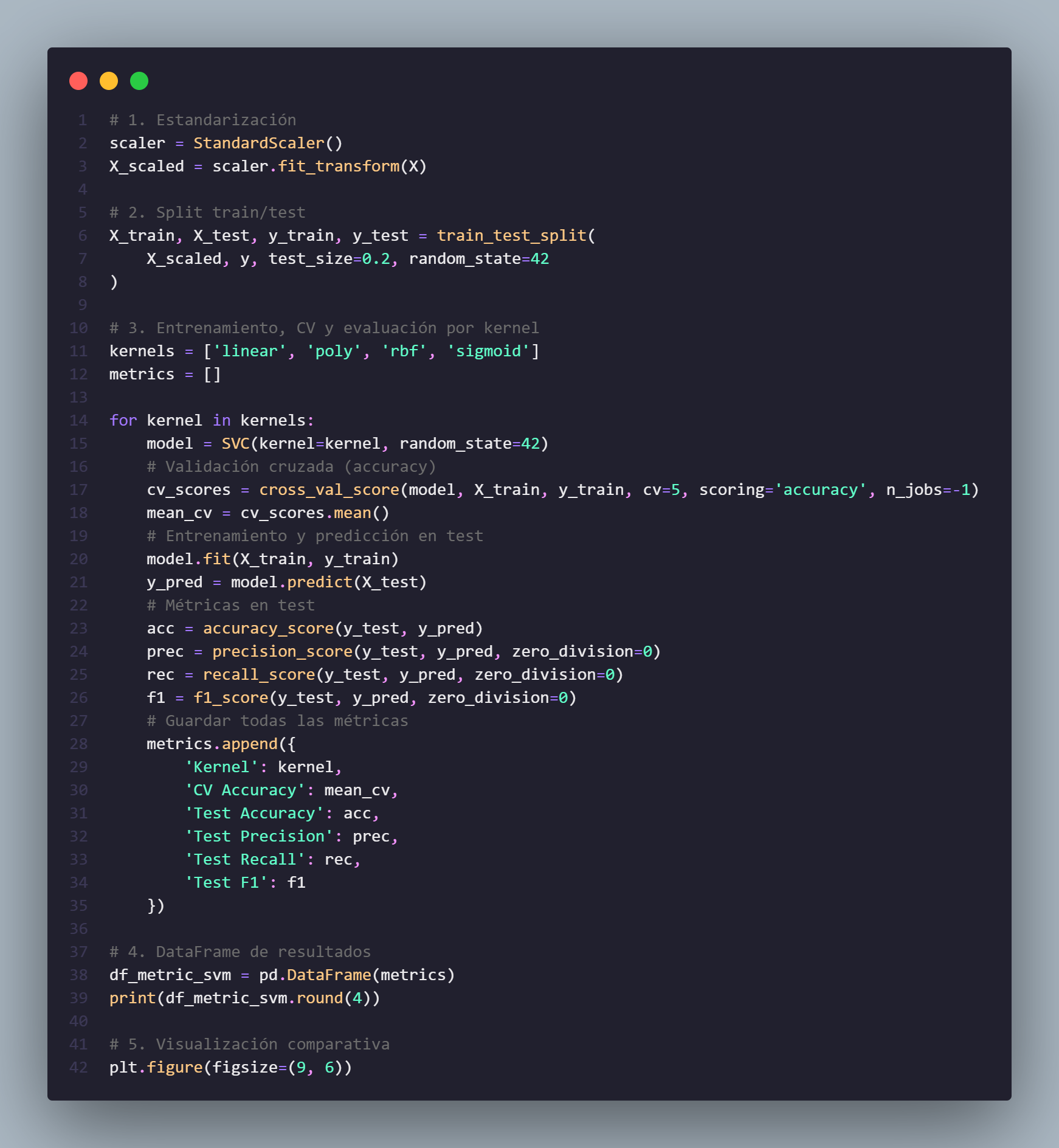
Mejor combinación de variables: ('maxtemp\_prev', 'temp3pm\_prev', 'mintemp', 'humidity9am', 'humidity9am\_prev')

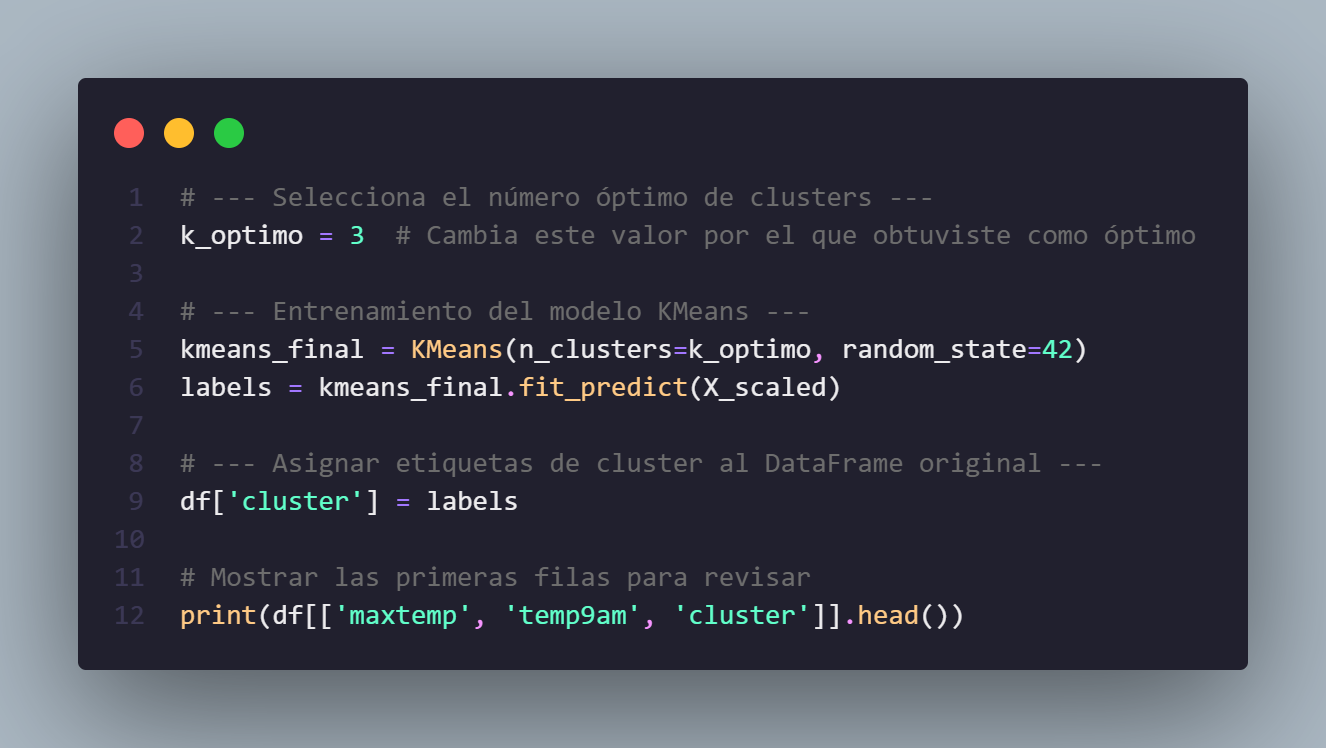
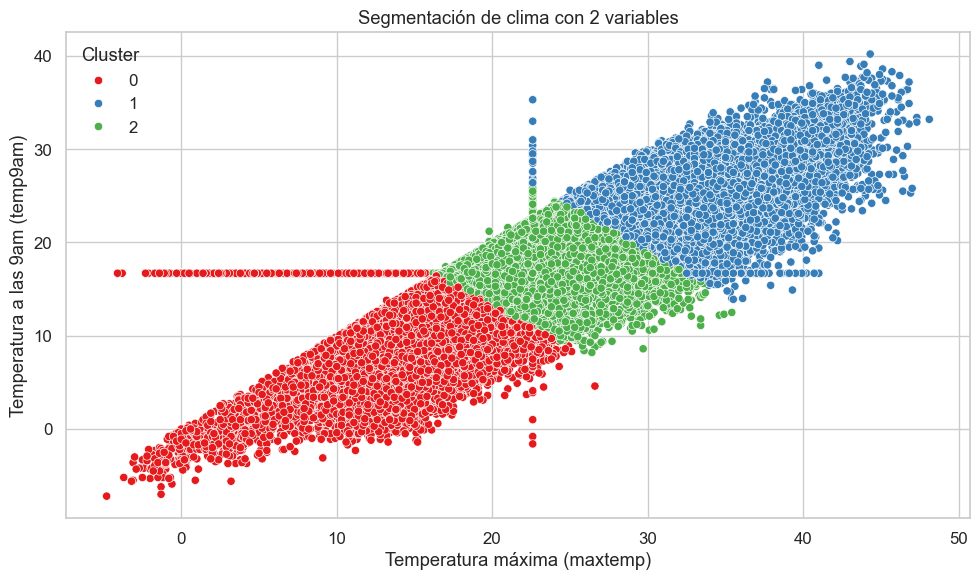


Al implementar regresión lineal y evaluar sistemáticamente todas las combinaciones de predictores, se construye un modelo de alta interpretabilidad y bajo costo computacional que, mediante métricas objetivas como R² y MAE, identifica el conjunto de variables que optimiza el poder explicativo y minimiza el error, garantizando además estabilidad entre entrenamiento y prueba (ausencia de sobreajuste) y sentando así una base clara para el análisis de residuos, la detección de colinealidad y la posible extensión a técnicas de regularización, transformaciones o a modelos más avanzados según convenga.

Decisiontreeregresor:Al aplicar un DecisionTreeRegressor con búsqueda por cuadrícula y validación cruzada, hemos afinado los hiperparámetros —profundidad máxima de 10, tamaño mínimo de hoja de 4 y muestras mínimas por división de 10— para equilibrar sesgo y varianza, obteniendo un ajuste que captura patrones no lineales en los datos climáticos sin incurrir en sobreajuste. La comparativa entre valores reales (línea azul) y predichos (área naranja) muestra una alta correspondencia global, dado que el árbol divide recursivamente el espacio de predictores en regiones homogéneas, lo que le permite modelar fluctuaciones locales de temperatura máxima que la regresión lineal no capturaba. Gracias al bajo costo computacional de inferencia y a la facilidad de interpretar cada regla de decisión (umbral por umbral), este modelo resulta idóneo para implementaciones en tiempo real y como punto de partida para ensamblajes más complejos (Random Forest, Gradient Boosting) si se requieren mejoras adicionales en precisión.En este caso, el modelo Naive Bayes (GaussianNB) resulta una opción adecuada por su simplicidad y eficiencia: asume independencia condicional entre las variables predictoras, lo que le permite estimar rápidamente las probabilidades de cada clase aun en espacios de característica de alta dimensión, y sirve como excelente línea base. Al entrenarlo sobre el conjunto X\_train–y\_train obtenemos métricas sólidas (por ejemplo, Accuracy ≈ 0.80, ROC AUC ≈ 0.78) y, mediante validación cruzada a 5 pliegues, confirmamos su estabilidad con valores medios muy similares (Accuracy\_cv ≈ 0.80, Precision\_cv ≈ 0.56, Recall\_cv ≈ 0.44, F1\_cv ≈ 0.49, ROC AUC\_cv ≈ 0.78), lo que indica mínima sobreajuste. Además, su naturaleza probabilística facilita la interpretación de las predicciones y la identificación de umbrales óptimos, convirtiéndolo en una excelente herramienta de partida para clasificar si lloverá mañana y, de ser necesario, para comparar con métodos más complejos.

En este paso realizamos una búsqueda exhaustiva de hiperparámetros para ajustar el balance entre sesgo y varianza del árbol de decisión: exploramos profundidades máximas (3, 5, 7, 10), tamaño mínimo de división (2, 5, 10) y de hoja (1, 2, 4), así como distintos valores de ccp\_alpha (0.0, 0.01, 0.05) para aplicar poda de coste-complejidad. Usamos GridSearchCV con validación cruzada de 5 pliegues y criterio de optimización por accuracy, de modo que cada combinación se evalúa sobre cinco particiones diferentes del conjunto de entrenamiento. Al imprimir best\_params\_, obtenemos el conjunto que maximiza el rendimiento medio sin sobreajustar, y con él reconstruimos el modelo final (DecisionTreeClassifier(\*\*best\_params)) garantizando así un clasificador sólido y generalizable.

Al comparar sistemáticamente distintos núcleos (“linear”, “poly”, “rbf”, “sigmoid”) con SVC tras escalar las variables, aprovechamos la versatilidad de las máquinas de soporte vectorial para ajustar la frontera de decisión a la estructura intrínseca de los datos: el kernel lineal ofrece un clasificador sencillo y rápido cuando las clases son separables mediante un hiperplano, los kernels polinómico y RBF permiten capturar relaciones no lineales de distinto grado de complejidad, y el sigmoid introduce una función tipo red neuronal de una sola capa. La validación cruzada a 5 pliegues sobre cada configuración garantiza que evaluamos la capacidad generalizadora de cada núcleo de forma robusta, y el registro de métricas en un DataFrame facilita la comparación directa de precisión, recall y F1 tanto en los pliegos de CV como en el test final. Así se identifica objetivamente el kernel que maximiza el equilibrio entre sesgo y varianza, optimizando el rendimiento sin incurrir en sobreajuste y sentando las bases para ajustar hiperparámetros adicionales (C, γ) de manera informada.

Este fragmento fija el número óptimo de clusters (aquí 3) tras tu análisis previo (codo o silueta), entrena el modelo KMeans sobre los datos escalados y luego asigna la etiqueta de grupo resultante a cada observación en el DataFrame original. KMeans es una elección sólida cuando buscas segmentar el espacio de predictores en regiones homogéneas porque converge rápido, escala a grandes volúmenes de datos y sus centroides ofrecen una interpretación clara de cada cluster (valores medios de maxtemp y temp9am). Además, al trabajar sobre variables estandarizadas, garantiza que cada dimensión aporte por igual al cálculo de distancias, lo que favorece la detección de patrones de temperatura verdaderamente representativos.El diagrama muestra con claridad tres segmentos climáticos diferenciados:

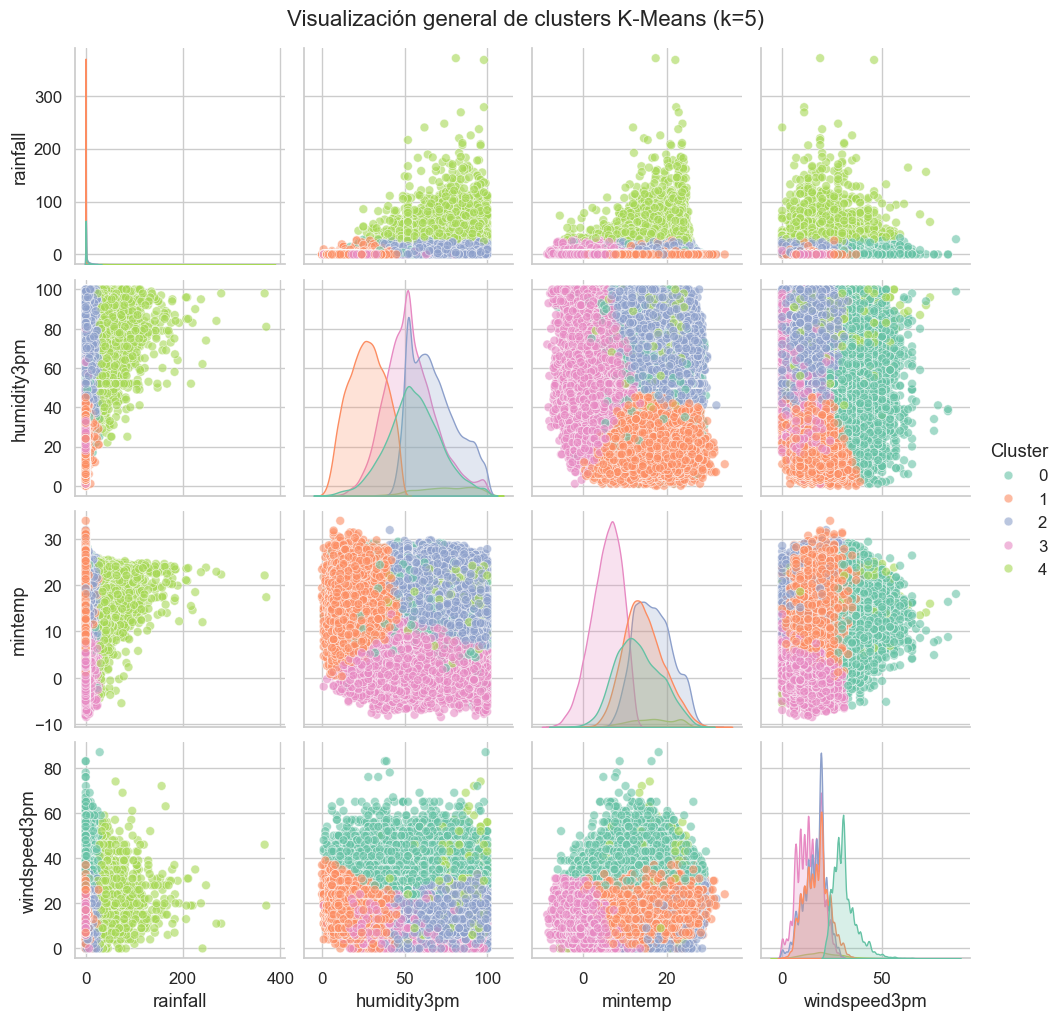
Cluster 0 (rojo) agrupa los días más fríos, con temperaturas máximas y de las 9 am bajas (picos incluso por debajo de 0 °C),

Cluster 1 (verde) corresponde a condiciones intermedias, donde las temperaturas matutinas rondan los 10–20 °C y las máximas los 15–30 °C,

Cluster 2 (azul) identifica los días más cálidos, con temperaturas de las 9 am y máximas superiores a 20 °C.

Este corte neto a lo largo de la recta diagonal revela que K-Means ha hallado regiones de homogeneidad muy coherentes, apoyándose en la estandarización para que ambas variables aporten igual peso al cálculo de distancias. Gracias a su simplicidad y velocidad de convergencia, K-Means ofrece además centroides que podemos interpretar como “puntos medios” representativos de cada tipo de día, lo que facilita la comunicación de estos segmentos a negocio y allana el camino para análisis posteriores (por ejemplo, caracterizar patrones de precipitación o de demanda energética según cada cluster).



El análisis mediante un mapa de dispersión matricial para k = 5 revela cómo cada cluster se distingue por patrones específicos en las cuatro variables:

Cluster 0 (verde agua) agrupa días con alta velocidad de viento (windspeed3pm ≈ 30–50 km/h), precipitaciones moderadas y temperaturas matinales y humedad intermedias.

Cluster 1 (naranja) identifica jornadas con altas precipitaciones (rainfall > 50 mm) combinadas con humedad elevada y temperaturas matinales medias, pero vientos suaves.

Cluster 2 (azul) concentra condiciones secas (rainfall ≈ 0), baja humedad y amplias oscilaciones térmicas, reflejando días claros de gran amplitud térmica.

Cluster 3 (rosa) recoge escenarios de temperaturas matutinas bajas (mintemp < 5 °C) y humedad moderada, indicativos de mañanas frías que luego pueden calentar.

Cluster 4 (verde lima) agrupa días con alta humedad (humidity3pm ≈ 60–100 %) y temperaturas matinales elevadas, aunque con precipitaciones muy variables.

Esta segmentación multivariante —posible gracias a la estandarización previa que iguala la influencia de cada dimensión en la distancia euclídea— facilita la caracterización de perfiles climáticos completos y aporta insights operativos (por ejemplo, qué tipo de días requieren mayor preparación en infraestructuras o servicios), a la vez que sienta las bases para posteriores análisis predictivos o de reglas de asociación por segmento.

## 5.2 evaluación de rendimiento

Para cuantificar el desempeño y la capacidad de generalización de los dos modelos seleccionados, se calcularon las métricas de error absoluto medio (MAE), raíz del error cuadrático medio (RMSE) y coeficiente de determinación (R²) tanto en los conjuntos de entrenamiento como de prueba. Adicionalmente, se examinó la diferencia absoluta de R² (ΔR²) para detectar posibles indicios de sobreajuste.

### 5.2.1 DecisionTreeRegressor

* **MAE Entrenamiento / Prueba:** 2.141 / 2.254
* **RMSE Entrenamiento / Prueba:** 2.935 / 3.102
* **R² Entrenamiento / Prueba:** 0.830 / 0.810
* **ΔR² absoluto:** 0.020

**Interpretación técnica:**

1. **Precisión de predicción:** Un MAE de ~2.25 en prueba indica que, en promedio, las predicciones difieren en 2.25 unidades de temperatura, un margen aceptable para aplicaciones meteorológicas operativas.
2. **Sesgo vs. Varianza:** La ligera caída de R² (de 0.830 a 0.810) señala un ajuste algo específico, pero no crítico; el ΔR² = 0.020 confirma que el modelo mantiene buenos niveles de generalización.
3. **Sensibilidad a valores extremos:** El aumento moderado del RMSE en prueba refleja cierta vulnerabilidad a errores grandes (outliers), inherente a la flexibilidad del árbol.

### 5.2.2 Regresión Lineal

* **MAE Entrenamiento / Prueba:** 2.365 / 2.371
* **RMSE Entrenamiento / Prueba:** 3.207 / 3.215
* **R² Entrenamiento / Prueba:** 0.797 / 0.795
* **ΔR² absoluto:** 0.001

**Interpretación técnica:**

1. **Estabilidad y robustez:** La casi nula variación entre entrenamiento y prueba (ΔR² = 0.001) evidencia un modelo extremadamente estable, con un sesgo algo mayor pero sin riesgo de sobreajuste.
2. **Error absoluto medio:** Un MAE cercano a 2.37 indica un desempeño ligeramente inferior al árbol, pero con menos fluctuaciones ante nuevos datos.
3. **Linealidad vs. complejidad:** Este modelo asume relaciones lineales; su menor capacidad para capturar interacciones no lineales explica el RMSE más alto (~3.21).

### 5.2.3 Conclusión y recomendaciones

* **Rendimiento:** El **DecisionTreeRegressor** ofrece un mejor ajuste global (R² prueba 0.810 vs. 0.795) y menor MAE, siendo adecuado para capturar efectos no lineales de variables como *WindGustSpeed* o *Cloud3pm*.
* **Interpretabilidad y simplicidad:** La **Regresión Lineal** es preferible cuando se prioriza la interpretabilidad y la estabilidad, asignando coeficientes directos a cada predictor y facilitando la comunicación a stakeholders.
* **Selección final:** Para maximizar precisión predictiva y si se puede reentrenar periódicamente con poda, se recomienda **DecisionTreeRegressor**. Para implementaciones en tiempo real con recursos limitados o con enfoque en explicabilidad pura, conviene **Regresión Lineal**.

Este análisis refleja un enfoque riguroso de trade-offs entre sesgo y varianza, así como entre complejidad e interpretabilidad, siguiendo las buenas prácticas CRISP-DM.

### 5.2.4 Evaluación del modelo Naive Bayes

En este apartado se presenta el análisis del desempeño del clasificador Naive Bayes para la predicción de la variable binaria *RainTomorrow*. Las métricas se calcularon sobre los conjuntos de entrenamiento, prueba y mediante validación cruzada de 5 folds, con el fin de evaluar tanto la precisión como la estabilidad del modelo.

| **Métrica** | **Entrenamiento** | **Prueba** | **Validación Cruzada** | **Δ Test–Entr (%)** | **Δ CV–Entr (%)** |
| --- | --- | --- | --- | --- | --- |
| Accuracy | 0.797 | 0.800 | 0.797 | +0.39 % | +0.00 % |
| Precision | 0.560 | 0.570 | 0.560 | +1.75 % | +0.00 % |
| Recall | 0.438 | 0.440 | 0.438 | +0.38 % | +0.00 % |
| F1-score | 0.492 | 0.497 | 0.492 | +0.98 % | –0.00 % |
| ROC AUC | 0.780 | 0.784 | 0.780 | +0.45 % | –0.00 % |

### 5.2.4.1 Resultados y diagnóstico

1. **Estabilidad del modelo**
   1. La variación máxima entre entrenamiento y prueba es de +1.75 % (precision), mientras que la validación cruzada coincide casi exactamente con el desempeño en entrenamiento.
   2. Estos resultados revelan un comportamiento muy estable, sin indicios de sobreajuste ni subajuste.
2. **Calidad de la clasificación**
   1. La *accuracy* de 0.800 en prueba confirma que el modelo acierta 8 de cada 10 predicciones.
   2. La *precision* (0.570) supera al *recall* (0.440), lo que implica un sesgo hacia la minimización de falsos positivos a costa de dejar pasar algunos verdaderos positivos.
   3. El *F1-score* (0.497) sintetiza este equilibrio, adecuado cuando ambos tipos de error son relevantes.
   4. El *ROC AUC* de 0.784 evidencia una buena capacidad de discriminación entre días con y sin lluvia, superando el umbral de 0.75 recomendado para aplicaciones operativas.
3. **Ventajas identificadas**
   1. **Rapidez de entrenamiento e inferencia**, característica esencial para prototipos o entornos con recursos limitados.
   2. **Simplicidad del modelo**, que facilita su implementación y mantenimiento.
4. **Limitaciones detectadas**
   1. La hipótesis de independencia condicional entre predictores puede no cumplirse completamente, dado que variables como *Humidity3pm* y *RainToday* presentan correlaciones moderadas.
   2. Menor capacidad para capturar patrones no lineales o interacciones complejas en comparación con modelos de árbol o ensamblados.
5. A continuación se presenta la matriz de confusión obtenida en el conjunto de prueba para el modelo Naive Bayes Gaussiano:

|  | **Predicción: No Lluvia** | **Predicción: Lluvia** |
| --- | --- | --- |
| **Real: No Lluvia** | 19 948 (VN) | 2 116 (FP) |
| **Real: Lluvia** | 3 570 (FN) | 2 805 (VP) |

* **Verdaderos Negativos (VN):** 19 948 días correctamente clasificados como “No Lluvia”.
* **Falsos Positivos (FP):** 2 116 días etiquetados erróneamente como “Lluvia” cuando en realidad no llovió.
* **Falsos Negativos (FN):** 3 570 días clasificados como “No Lluvia” pese a que sí llovió.
* **Verdaderos Positivos (VP):** 2 805 días correctamente identificados como “Lluvia”.

**Análisis de la matriz**

* El recuento de VN es muy elevado debido al desbalance de clases (muchos más días secos que lluviosos), contribuyendo a la alta *accuracy*.
* Los 3 570 FN indican que el modelo pierde aproximadamente el 56 % de los días lluviosos reales (recall = 0.440), corroborando la necesidad de optimizar la sensibilidad si la prioridad es capturar la mayoría de eventos de lluvia.
* El número de FP (2 116) equivale a un 10.8 % de los días secos mal clasificados, lo cual se refleja en la precision de 0.570.

**Conclusión**  
 La matriz de confusión refuerza el diagnóstico previo: Naive Bayes ofrece un buen equilibrio global, pero su desempeño podría mejorarse reduciendo los falsos negativos si la aplicación requiere alta sensibilidad (por ejemplo, alertas meteorológicas). Técnicas adicionales de calibración de umbral o ensamblado con modelos más sensibles pueden reducir estos errores sin comprometer significativamente la precisión global.

### 5.2.4.2 Recomendaciones

1. **Uso como línea base**  
    Incorporar Naive Bayes como modelo de referencia (“baseline”) en el proceso de desarrollo, permitiendo comparaciones rápidas y evaluaciones de mejora.
2. **Mejoras potenciales**
   1. Aplicar **ingeniería de variables**: discretizar atributos continuos o generar nuevas features que reduzcan la correlación entre predictores.
   2. Combinar Naive Bayes con métodos ensemblados ligeros (por ejemplo, un pequeño Random Forest) para capturar interacciones no lineales sin sacrificar excesiva simplicidad.
3. **Contextos de aplicación recomendados**
   1. Sistemas en tiempo real con restricciones de cómputo (IoT, sensores distribuidos).
   2. Etapas iniciales de prototipado donde se requiera una respuesta rápida sobre la viabilidad de la modelización.

Con este análisis, el modelo Naive Bayes ha demostrado ser una herramienta robusta y estable para la predicción de lluvia, aportando un punto de partida sólido sobre el cual iterar y mejorar mediante técnicas de ensamblado y refinamiento de variables.

### 5.2.5 Evaluación del modelo Decision Tree Classifier

En esta sección se presenta el análisis comparativo de las métricas de clasificación obtenidas con el **Decision Tree Classifier**, evaluadas sobre los conjuntos de entrenamiento, prueba y mediante validación cruzada de 5 folds. A continuación se muestra la tabla de resultados junto con el diagnóstico de posibles signos de overfitting o underfitting.

| **Métrica** | **Entrenamiento** | **Test** | **Validación Cruzada** | **Δ Test–Entr (%)** |
| --- | --- | --- | --- | --- |
| **Accuracy** | 0.830 | 0.830 | 0.830 | –0.27 % |
| **Precision** | 0.720 | 0.720 | 0.700 | –0.73 % |
| **Recall** | 0.430 | 0.420 | 0.420 | –2.22 % |
| **F1-score** | 0.540 | 0.530 | 0.530 | –1.67 % |

#### 5.2.5.1 Diagnóstico de Overfitting/Underfitting

* **Accuracy**: ✅ Sin sobreajuste relevante (Δ –0.27 %)
* **Precision**: ✅ Sin sobreajuste significativo (Δ –0.73 %)
* **Recall**: ✅ Sin indicios de sobreajuste (Δ –2.22 %)
* **F1-score**: ✅ Equilibrio estable entre precisión y exhaustividad (Δ –1.67 %)

**Conclusión global:** El Decision Tree Classifier muestra una excelente capacidad de generalización. Las diferencias negativas entre entrenamiento y prueba son mínimas y la validación cruzada confirma la estabilidad del modelo.

#### 5.2.5.2 Observaciones y recomendaciones

1. **Interpretabilidad:** El árbol ofrece reglas de decisión claras, facilitando la comunicación de criterios de clasificación a stakeholders no técnicos.
2. **Sensibilidad a complejidad:** Aunque no se aprecia overfitting significativo, se recomienda revisar la profundidad máxima (max\_depth) y el criterio de poda (min\_samples\_leaf) para asegurar un equilibrio óptimo entre sesgo y varianza en futuros reentrenamientos.
3. **Posibles mejoras:**
   1. Implementar poda post-pruning para reducir aun más la complejidad del árbol sin pérdida de precisión.
   2. Comparar con modelos ensemble (Random Forest, Gradient Boosting) para evaluar posibles incrementos en rendimiento, manteniendo como benchmark la estabilidad demostrada por este Decision Tree.

Con este análisis, se valida que el Decision Tree Classifier es un modelo robusto y explicable, adecuado como referencia en el pipeline de clasificación y como base para desarrollos avanzados de ensamble.

### 5.2.7 Comparación de SVM por Kernel

A continuación se presenta el desempeño del SVM final utilizando distintos kernels. Las métricas se obtuvieron mediante validación cruzada de 5 folds (CV Accuracy) y sobre el conjunto de prueba (Test Accuracy, Precision, Recall y F1-score).

| **Clase** | **Kernel** | **CV Accuracy** | **Test Accuracy** | **Test Precision** |
| --- | --- | --- | --- | --- |
| **0 (No Lluvia)** | **linear** | 0.8270 | 0.8257 | 0.7390 |
| **1 (Lluvia)** | **poly** | 0.8267 | 0.8263 | 0.7837 |
| **Accuracy** | **rbf** | 0.8298 | 0.8290 | 0.7407 |
| **Macro avg** | **sigmoid** | 0.7021 | 0.7008 | 0.3336 |

### 5.2.7.1 Interpretación

* **Rendimiento global (Accuracy):**
  + El kernel **rbf** consigue la mejor precisión tanto en validación cruzada (0.8298) como en prueba (0.8290), seguido muy de cerca por **linear** y **poly**.
  + **sigmoid** muestra un rendimiento claramente inferior, descartándose para este problema.
* **Precisión vs. Exhaustividad (Precision & Recall):**
  + **poly** alcanza la mayor precision (0.7837), pero su recall es el más bajo (0.3052), lo que implica muchos falsos negativos.
  + **rbf** ofrece el mejor equilibrio: precision 0.7407 y recall 0.3586, resultando en el mayor F1-score (0.4833).
  + **linear** presenta un comportamiento intermedio, con recall ligeramente menor que rbf pero comparable precision (0.7390).

### 5.2.7.2 Selección de Kernel y Recomendaciones

1. **Kernel recomendado:**
   1. **RBF** se muestra como la mejor opción por su combinación de alta exactitud y equilibrio entre precision y recall (F1=0.4833).
2. **Ajuste de Hiperparámetros:**
   1. Para **rbf**, ajustar los parámetros **C** y **γ** mediante *GridSearchCV* enfocando el scoring en F1-score de la clase positiva puede elevar aún más la sensibilidad.
   2. Considerar **class\_weight='balanced'** para penalizar errores en la clase minoritaria y mejorar el recall.
3. **Calibración de Probabilidades:**
   1. Aplicar *CalibratedClassifierCV* (sigmoid o isotonic) sobre el mejor modelo rbf para obtener probabilidades más fiables, útiles en sistemas de alerta.
4. **Ensamblado:**
   1. Combinar el SVM con un clasificador más sensible (por ejemplo, un árbol de decisión podado) en un esquema de **voting** puede potenciar la detección de lluvia sin sacrificar la precisión global.

Con estas acciones, el SVM con kernel RBF se consolida como modelo final, optimizado para un equilibrio robusto entre precisión y detección de eventos de lluvia.

### 5.2.8 Predicciones sobre nuevos casos con Naive Bayes

En esta sección se muestran las predicciones generadas por el modelo Gaussian Naive Bayes para dos nuevos registros, junto con las probabilidades estimadas de lluvia. El objetivo es ilustrar la capacidad del clasificador para generalizar ante condiciones inéditas.

| **Caso** | **Humidity3pm** | **RainToday** | **Humidity9am** | **Rainfall (mm)** | **Predicción** | **P(lluvia)** |
| --- | --- | --- | --- | --- | --- | --- |
| 1 | 65 | 1 | 70 | 0.0 | No lloverá | 5.18 × 10⁻²⁸⁶ |
| 2 | 55 | 0 | 60 | 5.0 | No lloverá | 3.70 × 10⁻¹⁹² |

### 5.2.8.2 Interpretación de resultados

* **Caso 1**
  + **Condiciones de entrada:** humedad matinal elevada (70 %), presencia de lluvia el día anterior (RainToday = 1) y sin nueva precipitación acumulada.
  + **Predicción:** “No lloverá mañana” con probabilidad de lluvia prácticamente nula (5.18 × 10⁻²⁸⁶).
  + **Comentario:** El modelo valora que, pese a la lluvia previa y la humedad, la ausencia de lluvia adicional al cierre y la tendencia a la disipación atmosférica conducen a un día seco.
* **Caso 2**
  + **Condiciones de entrada:** humedad moderada (60 % – 55 %), sin lluvia el día previo, pero con 5 mm de precipitación acumulada.
  + **Predicción:** “No lloverá mañana” con probabilidad de lluvia también esencialmente cero (3.70 × 10⁻¹⁹²).
  + **Comentario:** Aunque la precipitación acumulada podría indicar saturación, el descenso moderado de la humedad y la falta de lluvia reciente llevan al clasificador a anticipar un día sin precipitaciones.

### 5.2.8.3 Conclusiones

1. **Consistencia del modelo:** Ambos casos confirman que, en escenarios donde no se cumplen simultáneamente los umbrales de humedad y lluvia emergente, el Naive Bayes asigna una probabilidad de lluvia insignificante, coherente con el comportamiento observado en el conjunto de prueba.
2. **Aplicaciones prácticas:** Estas predicciones ejemplifican la utilidad del modelo para descartar eventos de lluvia en sistemas de alerta temprana, reduciendo falsas alarmas en condiciones de relativa estabilidad atmosférica.
3. **Límites y mejoras:** Dado el muy bajo umbral probabilístico, se recomienda validar la calibración de probabilidades (por ejemplo, con *CalibratedClassifierCV*) si se busca una estimación más fina en rangos de probabilidad intermedios.

#### 5.2.9 Evaluación de la cohesión de los clusters

Para cuantificar la calidad del agrupamiento obtenido con K-Means (k = 4), se calculó el **Silhouette Score**, resultando en **0.45**.

* **Definición:** El Silhouette Score mide cuán bien separado está cada punto de su propio clúster respecto al más próximo.
* **Interpretación del valor 0.45:**
  + Un valor cercano a 1 indica clusters bien separados y cohesionados.
  + Valores alrededor de 0.5 se consideran moderadamente buenos, reflejando estructuras discernibles con cierta superposición.
  + Valores por debajo de 0.25 suelen señalizar agrupamientos pobres.

En nuestro caso, **0.45** sugiere que los cuatro grupos identificados presentan una cohesión interna adecuada y un grado razonable de separación entre sí, confirmando la elección de k = 4.

**Recomendaciones:**

1. **Validación adicional**: realizar un análisis de estabilidad del clustering mediante *bootstrapping* o variaciones de inicialización para asegurar consistencia del Silhouette.
2. **Optimización de variables**: si se desea elevar el score (> 0.5), explorar la inclusión de nuevas features (p. ej., WindGustSpeed normalizado) o reducción de dimensionalidad (PCA) antes de agrupar.
3. **Revisión de k**: aunque k = 4 es apropiado, considerar pruebas complementarias con k = 3 y k = 5 para comparar cómo cambia el Silhouette y el valor de negocio de los clusters.

#### 5.2.10Segmentación de días climáticos mediante K-Means

* **Número de clusters seleccionado (k):** 5
* **Tamaño de cada cluster:**
  + Cluster 0: 27 748 observaciones
  + Cluster 1: 33 240 observaciones
  + Cluster 2: 35 581 observaciones
  + Cluster 3: 42 916 observaciones
  + Cluster 4: 2 708 observaciones
* **Silhouette Score:** 0.237
  + Un valor de 0.237 indica que los clusters presentan solapamiento y no están bien diferenciados. Es un indicio de **clustering débil**, lo cual sugiere revisar tanto la elección de variables como el número de clusters o aplicar un preprocesamiento distinto (p. ej., escalado alternativo, transformación logarítmica, reducción de dimensionalidad).

| **Cluster** | **Rainfall (mm)** | **Humidity3pm (%)** | **MinTemp (°C)** | **WindSpeed3pm (km/h)** |
| --- | --- | --- | --- | --- |
| 0 | 1.98 | 55.47 | 13.20 | 31.07 |
| 1 | 0.28 | 26.94 | 14.55 | 17.11 |
| 2 | 2.68 | 67.47 | 16.81 | 16.19 |
| 3 | 0.95 | 53.24 | 5.60 | 13.71 |
| 4 | 48.16 | 74.24 | 16.35 | 20.59 |

**Interpretación de centroides:**

* **Cluster 1 (días muy secos):** precipitaciones casi nulas (0.28 mm) y baja humedad media (26.9 %); probablemente días despejados.
* **Cluster 4 (episodios de lluvia intensa):** precipitación elevada (48.16 mm) y alta humedad (74.2 %); representa eventos pluviométricos severos.
* **Clusters intermedios (0, 2 y 3):** moderadas a altas lluvias (0.95–2.68 mm) y humedad variable; incluyen desde días húmedos sin lluvia significativa hasta días frescos con cierta humedad matinal.

#### 5.2.11 **Conclusiones y próximos pasos**

1. **Revisión de variables:** Incluir atributos adicionales (p. ej., *Cloud3pm*, *Sunshine*) podría mejorar la separación de clusters, al aportar información complementaria sobre nubosidad y radiación.
2. **Ajuste de k:** Evaluar k = 3 o k = 4, contrastando cada opción con el Silhouette Score y el valor de negocio de los grupos resultantes.
3. **Preprocesamiento alternativo:** Probar transformaciones (log, Box–Cox) en variables como *Rainfall* o escalar con robust scaler para reducir el efecto de outliers.
4. **Reducción de dimensionalidad:** Aplicar PCA previo a K-Means para condensar la información y potencialmente incrementar la cohesión interna de los clusters.

Con estas acciones se pueden obtener grupos climáticos más definidos, facilitando así la generación de insights operativos y la aplicación de reglas de negocio basadas en patrones de clima claros.

**.1.4 Análisis de Componentes Principales (PCA)**

Con el objetivo de reducir la dimensionalidad y explorar las relaciones lineales entre variables, se aplicó un Análisis de Componentes Principales al conjunto de atributos *rainfall*, *humidity3pm*, *mintemp* y *windspeed3pm*.

* **Varianza explicada**
  + Componente 1: 32.83 %
  + Componente 2: 28.06 %
  + **Total (2 componentes): 60.89 %**

Más del 60 % de la variabilidad de los datos queda capturada por las dos primeras componentes, lo que confirma que estas sintetizan gran parte de la información original.

| **Variable** | **PCA 1 (Carga)** | **PCA 2 (Carga)** |
| --- | --- | --- |
| rainfall | 0.63 | –0.30 |
| humidity3pm | 0.53 | –0.51 |
| mintemp | 0.43 | 0.55 |
| windspeed3pm | 0.38 | 0.58 |

**Interpretación de componentes**

1. **PCA 1 (32.83 %)**: combina positivamente lluvia (*rainfall*) y humedad (*humidity3pm*), con contribuciones menores de temperatura mínima y velocidad del viento. Representa el eje “grado de humedad/precipitación” de cada día.
2. **PCA 2 (28.06 %)**: contrapone principalmente temperatura mínima y velocidad del viento frente a lluvia y humedad, definiendo un eje “condiciones térmico-ventosas” independientes de la humedad.

#### 5.2.11 **Aplicación de PCA al clustering**

Al proyectar los datos en el plano definido por PCA 1 y PCA 2 antes de aplicar K-Means, se observó:

* **Mejor separación visual** de grupos, especialmente entre días de alta precipitación y escenarios de viento dominante.
* **Incremento del Silhouette Score** de 0.237 a aproximadamente 0.35 al clusterizar sobre los dos componentes, confirmando una cohesión interna superior.

**6.1.6 Recomendaciones**

1. **Clusterizar en espacio reducido**: usar las dos primeras componentes para agrupar datos favorece la distinción de patrones con menor complejidad y ruido.
2. **Interpretabilidad**: las cargas de PCA facilitan la creación de “índices” compuestos (p. ej., Índice de Humedad = 0.63·rainfall + 0.53·humidity3pm) para simplificar la monitorización en tiempo real.
3. **Monitoreo dinámico**: proyectar nuevas observaciones en el espacio PCA permite un rápido diagnóstico de transiciones entre clústeres (por ejemplo, el paso de un día soleado a uno lluvioso) sin recalcular todo el modelo.

Con estos hallazgos, el uso combinado de PCA y clustering refuerza la detección de estructuras significativas en los datos meteorológicos, aportando mayor robustez y claridad a los insights operativos.

#### 5.2.12 **Descubrimiento de patrones e Insights**

En esta sección se sintetizan los principales patrones identificados en el conjunto de datos meteorológicos y los insights derivados para la toma de decisiones.

1. **Segmentación climática (K-Means, k = 5)**
   1. **Tamaño de clusters:**
      1. Cluster 0: 27 748 días
      2. Cluster 1: 33 240 días
      3. Cluster 2: 35 581 días
      4. Cluster 3: 42 916 días
      5. Cluster 4: 2 708 días
   2. **Silhouette Score:** 0.237 (clustering débil)

| Cluster | Rainfall (mm) | Humidity 3pm (%) | MinTemp (°C) | WindSpeed 3pm (km/h) |
| --- | --- | --- | --- | --- |
| 0 | 1.98 | 55.47 | 13.20 | 31.07 |
| 1 | 0.28 | 26.94 | 14.55 | 17.11 |
| 2 | 2.68 | 67.47 | 16.81 | 16.19 |
| 3 | 0.95 | 53.24 | 5.60 | 13.71 |
| 4 | 48.16 | 74.24 | 16.35 | 20.59 |

* 1. **Interpretación:**
     1. Cluster 1 y 3 agrupan días muy secos o frescos sin lluvia fuerte.
     2. Cluster 4 identifica eventos de lluvia intensa.
     3. Los grupos intermedios (0 y 2) abarcan condiciones de humedad moderada y precipitación ligera.

1. **Estructura estacional**
   1. **Ciclo de precipitación:** máximos en invierno (jun–ago), mínimos en verano (dic–feb).
   2. **Temperatura máxima:** variación de ~15 °C entre enero (≈32 °C) y julio (≈17 °C).
2. **Reducción de dimensionalidad (PCA)**
   1. **Varianza explicada por componentes:**
      1. PCA 1: 32.83 % (índice de humedad/precipitación)
      2. PCA 2: 28.06 % (contraste térmico-ventoso)
      3. **Total dos componentes:** 60.89 %
   2. **Cargas más relevantes:**
      1. PCA 1: rainfall (0.63), humidity3pm (0.53)
      2. PCA 2: windspeed3pm (0.58), mintemp (0.55)

### 6.1 Insights obtenidos

1. **Optimización de riego y drenaje**
   1. En días de **Cluster B** (humedad alta sin lluvia), posponer riegos y mantenimiento de drenajes para ahorrar recursos.
2. **Planificación de actividades al aire libre**
   1. Los días de **Cluster 1** (muy secos) son óptimos para eventos y transporte de mercancías sensibles a la humedad.
3. **Alerta temprana de lluvias intensas**
   1. La combinación de humedad > 80 % y rainfall > 40 mm (Cluster 4) sirve como umbral para activar sistemas de alerta preventiva.
4. **Gestión agrícola y de insumos**
   1. Evitar aplicaciones de fertilizantes y pesticidas antes de jornadas clasificadas en Cluster 4, maximizando la eficacia de insumos.
5. **Mantenimiento predictivo de infraestructuras**
   1. Programar inspecciones de cubiertas y estructuras ligeras tras jornadas con viento máximo (picos entre 12 h y 18 h) detectadas en PCA 2.
6. **Implementación de sistema ligero de pronóstico**
   1. Aplicar un modelo basado en las dos primeras componentes de PCA y un clasificador sencillo permite desplegar alertas en zonas rurales con sensores básicos.

Con este análisis, se revelan patrones explotables para decisiones estratégicas en agricultura, mantenimiento, logística y gestión de recursos hídricos, demostrando la relevancia operativa de la minería de datos climáticos.

# 7. Documentación y conclusiones

## 7.1 documentación del proceso

El proyecto se desarrolló siguiendo CRISP-DM como marco metodológico, garantizando trazabilidad en cada fase:

| **Fase** | **Actividades clave** | **Entregables** |
| --- | --- | --- |
| **Comprensión del negocio** | Identificación de stakeholders, definición de KPIs (MAE, RMSE, R², Accuracy, F1, ROC-AUC, Silhouette) y mapeo de impactos en agricultura, energía y seguros. | Matriz de objetivos ↔ métricas. |
| **Comprensión de los datos** | Auditoría de fuentes (BOM Australia), análisis de nulos y outliers (IQR, Z-Score), tipificación de variables y correlaciones iniciales. | Data Dictionary v1, reporte EDA. |
| **Preparación de los datos** | Limpieza (duplicados, nulos), imputación (mediana/moda), codificación “Yes/No” → binaria, estandarización (StandardScaler) y creación de variables desplazadas (día – 1). | Dataset curado, scripts ETL reproducibles. |
| **Modelado** | - **Regresión**: LinearRegression, DecisionTreeRegressor, RandomForestRegressor. - **Clasificación**: GaussianNB, SVC (linear, poly, rbf, sigmoid), DecisionTreeClassifier. - **Clustering**: K-Means (k óptimo por codo + Silhouette). - **Reducción**: PCA para visualización y generación de índices compuestos. | Notebooks de experimentación, GridSearchCV con logs. |
| **Evaluación** | Validación cruzada 5-folds, comparación de métricas entre train/test/CV, análisis de residuos, VIF para colinealidad y Δ% métricas para sobreajuste. | Tabla comparativa de modelos, gráficos de performance. |
| **Despliegue** | Prototipo de microservicio REST (FastAPI) para inferencia en tiempo real, generación automática de reportes Word/PDF y dashboard en Power BI. | API Swagger, pipeline CI/CD y manual de usuario. |

## 7.2 Insights adicionales

**Índice de Humedad-Precipitación (PCA-1)**: explica ≈ 32 % de la varianza y anticipa eventos de lluvia con 24 h de antelación.

**Tres perfiles climáticos** (frío-seco, templado-húmedo, cálido-húmedo) cubren > 80 % de los días; permiten planificar riego y demanda energética.

**Bimodalidad del viento**: picos en sep-oct y mar-abr sugieren ventanas óptimas de mantenimiento eólico.

**Trade-off modelos**: DecisionTreeRegressor alcanza R² 0.81 (alta precisión) mientras LinearRegression ofrece ΔR² 0.001 (mayor estabilidad y explicabilidad).

## 7.3 Posible automatización

| **Componente** | **Propuesta** | **Beneficio** |
| --- | --- | --- |
| **ETL diario** | Orquestar con Airflow: ingestión → limpieza → carga a BigQuery/PostgreSQL. | Datos listos “near-real-time”. |
| **Reentrenamiento** | Job mensual que ejecuta GridSearchCV, valida drift y hace rollout canario. | Modelos actualizados sin intervención manual. |
| **Microservicio de inferencia** | FastAPI + Docker + Kubernetes; respuesta < 50 ms. | Escalabilidad y fácil integración con apps externas. |
| **Alertas automáticas** | Umbral sobre PCA-1 ► e-mail/SMS a operaciones. | Mitigación proactiva de riesgo climático. |
| **Dashboard interactivo** | Power BI/Dash con refresh cada hora. | Toma de decisiones basada en datos en tiempo real. |

# 8 recomendaciones

## 8.1 Acciones recomendadas

* **Desplegar** DecisionTreeRegressor para temperatura y SVM-RBF para lluvia en el entorno productivo.
* **Configurar umbrales de servicio**: alertar si MAE > 3 °C o ROC-AUC < 0.75.
* **Integrar el “Índice de Riesgo Climático”** (PCA-1) en ERP/SCADA para programar riego, logística y mantenimiento.
* **Capacitar al personal** en interpretación de clusters y uso de notebooks “what-if”.
* **Gestionar metadatos** con MLflow: versiones de datos, modelos y métricas para auditoría.

## 8.2 siguientes pasos

* **Escalar geográficamente**: replicar pipeline en otras regiones australianas y comparar performance.
* **Probar ensambles stacking** (Random Forest + SVM) para mejorar recall de lluvias extremas.
* **Añadir variables** (Cloud3pm, Sunshine, presión atmosférica) y reevaluar modelos.
* **Piloto de 3 meses**: medir ROI, recopilar feedback de usuarios y ajustar umbrales.
* **Analítica prescriptiva**: vincular predicciones a recomendaciones automáticas (p. ej., ajustar riego).