

**DESARROLLO DE MODELOS PREDICTIVOS PARA LA
ANTICIPACIÓN DE FENÓMENOS METEOROLÓGICOS EN LA
REGIÓN MEDITERRÁNEA: UN ENFOQUE EN LA TEMPERATURA Y
EL CAPE**

Development of Predictive Models for the Anticipation of Meteorological
Phenomena in the Mediterranean Region: A Focus on Temperature and CAPE

Pablo Martín Parera
pablomartinpar@gmail.com
Grado en Business Analytics
Curso académico 2024-2025

Resumen

Este trabajo aborda el desarrollo de modelos predictivos para la temperatura y la energía potencial convectiva disponible (CAPE) en la región mediterránea, con el objetivo de mejorar la planificación ante altas temperaturas y fenómenos meteorológicos extremos.

Se recopilan datos horarios desde el 2012 hasta el 2018, sobre ellos se aplican técnicas de limpieza, normalización y generación de nuevas variables.

Para la temperatura se comparan dos enfoques distintos, uno clásico (ARIMA/SARIMAX) y uno basado en redes neuronales recurrentes (GRU/LSTM), se busca comparar que enfoque es más efectivo y escalable. La variable CAPE se caracteriza por alta asimetría y abundantes ceros, para su estudio se usó XGBOOST con la función perdida Tweedie, capturando eficazmente las tendencias en los picos convectivos.

Los resultados demuestran la utilidad del uso de los métodos estadísticos y el aprendizaje profundo para la anticipación y mitigación de los riesgos asociados al cambio climático.

Palabras clave: Modelos predictivos, CAPE, Temperatura, LSTM, ARIMA, XGBoost

Abstract

This study focuses on the development of predictive models for temperature and convective available potential energy (CAPE) in the Mediterranean region, aiming to enhance planning for high temperatures and extreme weather events.

Hourly data from 2012 to 2018 were collected and processed using cleaning, normalization, and feature engineering techniques.

For temperature forecasting, two different approaches were evaluated: a classical one (ARIMA/SARIMAX) and another based on recurrent neural networks (GRU/LSTM), with the objective of determining which method is more effective and scalable. Due to its highly skewed distribution and frequent zero values, the CAPE variable was modeled using XGBoost with a Tweedie loss function, effectively capturing trends in convective peaks.

The results demonstrate the usefulness of combining statistical methods and deep learning to anticipate and mitigate risks associated with climate change.

Keywords: Predictive models, CAPE, Temperature, LSTM, ARIMA, XGBoost

ÍNDICE

1.	OBJETIVOS	1
	Objetivo general	1
	Objetivo específico	1
2.	CUERPO DE LA MEMORIA	2
	Introducción	2
	Motivación	2
	Estado del arte	3
	Marco teórico	6
3.	TRABAJO TÉCNICO	11
	Herramientas empleadas	11
	Ingeniería del dato	13
	Análisis del dato	27
	Análisis del Negocio	42
4.	REFERENCIAS BIBLIOGRÁFICAS	48

ÍNDICE DE FIGURAS

Figura 1 Zona acotada en la extracción de datos	13
Figura 2 Valores atípicos Cape.....	17
Figura 3 Valores atípicos sst	18
Figura 4 Valores atípicos T2m.....	18
Figura 5 Relación d2m y cape	19
Figura 6 Relación t2m y sst.....	19
Figura 7 Relación cape y sst.....	20
Figura 8 Relación msl y cape.....	20
Figura 9 Promedio mensual sst, temperatura y cape	21
Figura 10 Matriz correlación	21
Figura 11 Cuatro componentes sst	23
Figura 12 Cuatro componentes t2m	24
Figura 13 Cuatro componentes d2m	24
Figura 14 Cuatro componentes cape	25
Figura 15 Cuatro componentes sp	26
Figura 16 Cuatro componentes wind-magnitude.....	26
Figura 17 Train vs Test GRU	33
Figura 18 Train vs Test LSTM	33
Figura 19 Predicción ARIMA	34
Figura 20 Predicción GRU en test	37
Figura 21 Predicción SARIMAX en test.....	37
Figura 22 Predicción nº1 CAPE	40
Figura 23 Predicción nº2 CAPE	40
Figura 24 Predicción nº3 CAPE	41
Figura 25 Temperatura media anual sin el MAR	44
Figura 26 Temperatura Máxima anual.....	45
Figura 27 Temperatura media anual del mar	45
Figura 28 Evolución del CAPE 2012-2017.....	46

ÍNDICE DE TABLAS

Tabla 1 Formatos variables antes de transformaciones	14
Tabla 2 Ejemplo media y moda antes de normalización.....	16
Tabla 3 Ejemplo media y moda después de normalización	16
Tabla 4 Variables y formatos después de transformaciones	17
Tabla 5 Estacionalidad Variables	22
Tabla 6 Media y desviación estándar Cape y T2m	27
Tabla 7 Comparación métricas GRU y SARIMAX	34
Tabla 8 Métricas XGBoost	39

1. OBJETIVOS

Objetivo general

Desarrollar y validar modelos predictivos de alta precisión para la temperatura y el CAPE en la región mediterránea, con el fin de aportar herramientas robustas que permitan anticipar fenómenos climáticos extremos y respaldar la planificación de actuaciones preventivas por parte de autoridades y ciudadanos. Con ello se pretende minimizar los efectos adversos sobre la salud, los ecosistemas, la agricultura y el turismo, tanto a corto como a medio plazo.

Objetivo específico

- Recopilación y preprocesado de datos: obtención de las series temporales de ERA5, detección y tratamiento de observaciones atípicas, y normalización de variables para asegurar un conjunto limpio y homogéneo.
- Selección y entrenamiento de modelos: identificación de metodologías apropiadas para cada variable (ARIMAX/SARIMAX, XGBoost con devianza Tweedie y LSTM/GRU), división de los datos en conjuntos de entrenamiento y validación temporal, ajuste de parámetros con base en el desempeño en el conjunto de validación y cuantificación del error mediante RMSE, MAE, R y MAPE.
- Afinamiento de hiperparámetros y estructuras: exploración sistemática de configuraciones (órdenes estacionales, parámetros de aprendizaje, número y tamaño de capas) para maximizar la capacidad predictiva, considerando la estacionalidad de la temperatura y la escasez de eventos extremos en el CAPE.
- Propuesta de aplicación operativa: definición del formato de salida , junto con criterios de presentación para facilitar su interpretación en entornos de decisión meteorológica.

2. CUERPO DE LA MEMORIA

Introducción

En los últimos años, se ha hecho cada vez más evidente una transformación en el comportamiento del clima. El aumento continuo de las temperaturas, la aparición de lluvias más intensas y prolongadas, así como la persistencia de sequías más severas, reflejan una alteración en los equilibrios meteorológicos que tradicionalmente caracterizaban distintas regiones, en nuestro caso de estudio, la zona mediterránea. Ante esta realidad, este trabajo tiene como finalidad contribuir al proceso de toma de decisiones en el campo de la meteorología, ofreciendo un enfoque basado en el análisis de datos y en la construcción de modelos que permitan anticipar, situaciones atmosféricas potencialmente críticas.

Motivación

Este proyecto surge ante la creciente necesidad de modelos predictivos más precisos y avanzados, impulsado por la inestabilidad climática que vivimos por el cambio climático. En particular la investigación se centra en el mediterráneo, esta zona está sufriendo temperaturas anormales y fenómenos meteorológicos como olas de calor prolongadas, grandes sequías y fuertes Danas. Estos cambios no solo afectan a la biodiversidad, sino que también afectan gravemente a los pueblos, ciudades, a la agricultura y al turismo. El objetivo principal para las predicciones son el CAPE (Convective available potential energy) y la temperatura, estas son variables importantes ya que el CAPE está muy asociado con la creación de tormentas y fenómenos anormales (danas, tornados o tormentas eléctricas) y la temperatura es una de las variables meteorológicas que más moldean nuestro día a día. Contar con modelos predictivos efectivos facilita la tarea de anticipación para los habitantes de las zonas afectadas. Esta capacidad es crucial ya que con buenas predicciones los gobiernos pueden hacer mejores planes, tomar medidas más concretas y efectivas. Por otro lado, los ciudadanos pueden blindarse ante la predicción de altas temperaturas o grandes tormentas, haciendo su vida más fácil y segura.

Estado del arte

En el presente apartado, principalmente, me voy a centrar en estudios que usen la misma fuente de datos que he usado yo (ERA5), ya que considero que la fuente de datos usada en meteorología es crucial para la precisión y la validez de los resultados obtenidos. El ERA5, se ha consolidado, en los últimos años, como una herramienta fundamental en la investigación meteorológica, debido a la gran cantidad de variables, el largo abanico de tiempo que ofrece y sobre todo la constancia de los datos(Hora a Hora). Además, los estudios que usen el ERA5 ofrecen una base sólida para la construcción de modelos predictivos, por lo que son relevantes para la constitución de mi anteproyecto. También me voy a centrar en los estudios en los que las variables a predecir sean importantes para mi estudio o directamente sean mis variables a predecir (Temperatura y CAPE).

Cualquier trabajo previo que haya utilizado estas variables me proporciona un contexto valioso y una base confiable para mi anteproyecto. Lo más lógico es empezar por el principio, en este caso el principio del estado del arte es este estudio Principles of meteorology and numerical weather prediction (al, 2006). El estudio abarca los principios de la meteorología y de la predicción numérica del clima(NWP), tratando temas como la circulación atmosférica, el caos atmosférico o el impacto de la radiación solar en el clima. La predicción de variables climáticas es un reto que los meteorólogos llevan años intentando perfeccionar, aunque el enfoque principalmente era probabilístico, por esto considero valioso mi enfoque en el campo de la meteorología.

En el estudio, al final, comentan esto y enfocan la predicción meteorológica a ser un problema más de Big Data que un problema puramente probabilístico, así lo podemos ver en el apartado Weather Forecasting as a Big Data Problem:

"Los meteorólogos lidian con las desviaciones de las soluciones exactas mediante el posprocesamiento, adoptando la incertidumbre y abordando el problema como uno probabilístico, lo que subraya la necesidad de planificar cuidadosamente la validación y verificación a través de la definición detallada del problema y el uso de datos observacionales integrados en los modelos." (Haupt et al., Principles of Meteorology and Numerical Weather Prediction) (al, 2006).

El estudio se centra en la predicción numérica del clima, enfocándose en las ecuaciones pronósticas en lugar de algoritmos para predecir las variables, estas ecuaciones

describen como van a evolucionar las variables meteorológicas mediante ecuaciones matemáticas y físicas.

Como he comentado antes, el enfoque de Big Data es diferente a lo que se lleva haciendo los últimos años en la meteorología. A parte de las ecuaciones pronosticas, tenemos las ecuaciones diagnosticas, estas ecuaciones se usan para obtener variables a partir de otras, por ejemplo, la Vorticidad potencial que relaciona la temperatura, la presión y el viento.

"Los modelos NWP están diseñados para predecir directamente variables meteorológicas a través de ecuaciones pronósticas y, en algunos casos, mediante ecuaciones diagnósticas que calculan relaciones derivadas entre las variables observadas." (Haupt et al., Principles of Meteorology and Numerical Weather Prediction, section 2.1) (al, 2006).

Considero que las ecuaciones diagnosticas son fundamentales para mi estudio, ya que la inclusión de un mayor número de variables podría incrementar la precisión de mis modelos. Otro estudio que usa ecuaciones pronosticas es An ERA5-Based Hourly Global Pressure and Temperature (Pedro Mateus, 2020) en este caso añade ecuaciones para que el modelo tenga en cuenta el cambio climático global y también introduce un sistema de periodicidades anuales, semi-anuales y trimestrales. La formulación del modelo, teniendo en cuenta lo comentado, lo hace más robusto.

Es importante tener en cuenta que , de nuevo, es un modelo basado en la física y la probabilidad, no en el Big Data. Los datos que usa el modelo son del ERA5 y del GNSS (The Global Navigation Satellite System), en el caso del ERA5 el estudio resalta la mejora de esta fuente :

"El principal avance se atribuye a un aumento en el espaciamiento de la cuadrícula horizontal (de 79 a 31 km), en el número de niveles del modelo (de 60 a 137) y en la resolución temporal (de 6 a 1 hora), lo que permite una mejor representación atmosférica de los sistemas convectivos, las ondas de gravedad, los ciclones tropicales y otras estructuras atmosféricas de meso a sinóptica escala. Otra mejora del ERA5 es el número de observaciones asimiladas, que pasó de un promedio de 0.75 millones por día en 1979 a aproximadamente 24 millones por día en 2018."(Global Reanalysis: Goodbye ERA-Interim, Hello ERA5. ECMWF Newsletter, (159). (Hersbach, 2019)

El modelo, siendo un modelo que mejora entorno al 5%-30% (dependiendo de la variable) las predicciones comparadas con otros modelos(NWP), es un modelo preciso, pero no mide el error de la misma manera que el próximo modelo que vamos a ver, por

lo que la comparación exacta no la puedo compartir ya que necesitaría más datos del presente estudio.

Como mencioné anteriormente, el problema de las predicciones meteorológicas comúnmente se ha visto como un problema probabilístico con las ecuaciones pronosticas:

"Se han publicado muchos métodos para predecir la temperatura de la superficie del mar. Estos métodos se pueden clasificar generalmente en dos categorías según el punto de vista diferente para crear modelos. Uno se basa en la física, también conocido como modelo numérico. El otro se basa en datos, también llamado modelo impulsado por datos." (Prediction of Sea Surface Temperature using Long Short-Term Memory.) (Zhang, 2017)

Pero este no es el caso de Prediction of Sea Surface Temperature using Long Short-Term Memory, este estudio tiene un enfoque diferente. En este caso los investigadores usan un modelo llamado Long short-term memory(LSTM) para predecir la temperatura del mar(SST), ellos entienden el problema de la predicción como una serie temporal y no como unas variables que adaptar a formulas físicas, atacando el problema desde el Big Data.

Este es el primer estudio que predice una variable meteorológica con LSTM. Estas redes neuronales recurrentes (RNN) están especializadas en series temporales especialmente largas.

Los LSTM están diseñados para solucionar el problema de los gradientes que desaparecen cuando entrenas un RNN en secuencias de datos largas, por lo que usar un LSTM para grandes cantidades de datos meteorológicos es ideal. Este es un gran estudio, encuentra una manera diferente de predecir este tipo de datos, pero su base de datos, aun siendo de gran calidad, es escueta en variables.

Usa datos del NOAA ¹, esta es una fuente de gran calidad, el problema que veo es que solo usan la variable SST, por lo que entienden que para predecir la temperatura del mar solo es relevante la misma temperatura del mar. El clima es un sistema que funciona gracias a la relación de los miles de variables atmosféricas y las numerosas variables oceánicas, el usar solo la variable SST es, como prueba un acierto, pero como modelo meteorológico lo siento débil y susceptible al caos de la atmósfera:

¹ NOAA (Administración Nacional Oceánica y Atmosférica) es una agencia del gobierno de EE. UU. que se encarga de estudiar y monitorear el clima, los océanos y la atmósfera.

"Los científicos del clima a veces se refieren a los efectos del caos como variabilidad intrínseca o no forzada: los cambios impredecibles que surgen de las interacciones dinámicas entre los océanos y la atmósfera." (New Scientist. (2007). Mitos climáticos: Los sistemas caóticos no son predecibles.) (Brooks, 2007)

Aún con esta falta de variables, su modelo se comporta muy bien en las pruebas, mostrando que los LSTM son redes neuronales que tener en cuenta para el estudio meteorológico.

Marco teórico

Primeramente, veamos el modelo que se va a usar para la predicción del cape, este es el XGBoost (Chen, 2016) es un algoritmo de boosting basado en árboles que construye de forma secuencial un ensamblado de árboles de decisión minimizando una función objetivo regularizada. Cada árbol corrige los errores residuales del conjunto previo mediante aproximaciones de primer y segundo orden (gradiente y hessiano), lo que le confiere gran capacidad para capturar relaciones no lineales y manejar ruido en los datos meteorológicos. Para la predicción del CAPE, cuyas observaciones son en su gran mayoría cero o muy bajas y presentan unos pocos valores extremos (días con convectividad² intensa), XGBoost aporta robustez frente a outliers y flexibilidad para modelar interacciones complejas entre variables (por ejemplo, temperatura, humedad y presión).

Dado el característico sesgo hacia ceros y colas pesadas de la distribución de CAPE, seleccionaremos la función de pérdida Tweedie (Dunn, 2005), propia de la familia exponencial de distribuciones compuestas Poisson-Gamma. Esta familia se define por

$$Var(Y) = \phi \mu^p$$

con parámetro de potencia p ($1 < p < 2$) que controla el grado de dispersión y permite una masa en cero simultáneamente con valores continuos positivos. En XGBoost, la optimización de la pérdida de devianza³ Tweedie se realiza derivando su gradiente y

² En meteorología, hace referencia a la capacidad de la atmósfera para desarrollar movimientos verticales que pueden dar lugar a tormentas.

³ «devianza» (del inglés *deviance*) es una medida de ajuste de modelos estadísticos

hessiano específicos, lo que mejora la capacidad del modelo para ajustar la frecuencia de “días tranquilos” frente a los episodios convectivos extremos que genera el CAPE.

Veamos ahora los modelos centrados en la predicción de la temperatura, el primero de ellos es ARIMAX.

Entre los puntos fuertes de emplear un modelo ARIMAX (Box, 1970) en mi trabajo destaca su capacidad para recoger de manera clara cómo la propia serie temporal se relaciona consigo misma (a través de sus rezagos) y, al mismo tiempo, cómo influyen variables externas, por ejemplo, la radiación solar o la humedad. Esto resulta clave cuando sospechamos que factores meteorológicos concretos ejercen un efecto directo sobre la magnitud que queremos predecir.

Además, ARIMAX ofrece una gran claridad a la hora de interpretar resultados. Gracias a los parámetros autorregresivos y de media móvil sabemos cuántos pasos atrás de la serie —o de los errores— necesitamos para explicar el presente, y los coeficientes de los regresores externos nos cuentan hasta qué punto y en qué dirección actúa cada factor. Eso facilita muchísimo la presentación de conclusiones, sobre todo en un entorno académico.

Cuando introducimos también la componente estacional hablamos de SARIMAX (Box, 1976), que añade términos AR y MA estacionales para captar ciclos periódicos (diarios, semanales, anuales...). En meteorología, donde la temperatura o la radiación siguen patrones muy marcados según la época del año, este matiz estacional suele mejorar notablemente la precisión.

Tanto ARIMAX como SARIMAX se ajustan normalmente por máxima verosimilitud o mínimos cuadrados: primero elegimos los órdenes (p , d , q) y, en el caso estacional, (P , D , Q , s) usando criterios como AIC o BIC y el análisis de la autocorrelación, luego se optimizan los parámetros y se contrastan los residuos hasta obtener un modelo fiable.

Entre las pegas, esta que estos modelos asumen relaciones lineales entre variable y regresores, lo que puede quedarse corto si el sistema presenta interacciones no lineales complejas. Además, identificar mal los órdenes o la estacionalidad puede llevar a sobreajustes o a pronósticos poco sólidos.

Por otro lado, a diferencia de métodos “rolling window” que actualizan sus parámetros cada día con sólo datos recientes, ARIMAX/SARIMAX suelen recalibrarse con todo el histórico. Ese respaldo del pasado puede volverse contraproducente si las dinámicas

cambian deprisa. Para evitarlo, a veces conviene reentrenar el modelo sobre una ventana móvil de datos recientes o implementar versiones recursivas que ajusten los coeficientes de forma más ágil.

En resumen, ARIMAX y SARIMAX brindan un enfoque transparente y completo para combinar la memoria propia de la serie, la influencia de variables externas y los ciclos estacionales, siempre que asumamos linealidad y dediquemos el tiempo necesario a elegir bien sus órdenes y a preprocesar la serie.

Otro modelo que está diseñado para las secuencias temporales son las Redes Neuronales Recurrentes (RNN), su mayor diferencia con las redes neuronales tradicionales es que las RNN tienen conexiones recurrentes, esto significa que cada capa en un momento t se retroalimenta como la entrada para el siguiente momento $t+1$, esto capacita que las RNN ‘recuerden’ la información de los pasos anteriores, al tener en cuenta la información de los pasos anteriores la RNN puede capturar patrones y dependencias temporales en los datos. La arquitectura básica de una RNN empieza por la entrada de datos secuenciales, posteriormente la red procesa cada entrada de manera secuencial, por lo que las salidas de los nodos anteriores sirven a los nodos siguientes como entrada. Estos nodos están compuestos por neuronas recurrentes, estas neuronas no solo dependen de t sino que también dependen del estado de la neurona anterior n_{t-1} , para esto las neuronas usan la formula general de una neurona recurrente que es: $h_t = f(W_{ih}x_t + W_{hh}h_{t-1} + b_h)$ siendo:

- h_t el estado oculto ⁴ en el momento t
- x_t la entrada en el momento t
- W_{ih} el peso entre la entrada y el estado oculto
- W_{hh} el peso entre el estado oculto anterior y el estado oculto actual
- b_h representa el sesgo de la neurona

En otras palabras, $W_{ih}x_t$ nos muestra como la entrada influye en el nuevo estado oculto y $W_{hh}h_{t-1}$ muestra como la neurona usa información proporcionada por la neurona anterior para crear el nuevo estado oculto.

Hasta aquí he explicado como las neuronas tienen la capacidad de recordar, el siguiente paso es entender cómo se crea la salida que nosotros esperamos.

⁴ El estado oculto en una RNN representa la memoria interna del modelo en cada instante de tiempo

Para eso usamos esta fórmula: $y_t = f(W_{hy}h_t + b_y)$ donde y_t es la salida y W_{hy} es el peso que se asigna entre el estado oculto y la salida final de la neurona.

Las RNN son significativamente mejores para los datos secuenciales, como las series temporales como es mi caso, pero no son perfectas.

Anteriormente he comentado como las LSTM arreglan el problema de las gradientes de las RNN, pero veamos bien porque las RNN tienen ese problema. Primeramente, veamos qué es un desvanecimiento de gradiente, este fenómeno ocurre cuando los gradientes sufren una reducción exponencial a medida que se propagan a través de las capas de la RNN.

Esto suele ocurrir cuando los gradientes tienen que atravesar muchos pasos, ya que la RNN tiene muchas capas. Cuantas más capas pasan los gradientes⁵ más se van desvaneciendo, este proceso se describe mediante esta fórmula: $\frac{\partial L}{\partial W} = \sum_{t=1}^T \frac{\partial L}{\partial h_t} * \frac{\partial h_t}{\partial W}$ donde:

- ☐ L representa la función de pérdida
- ☐ W son los pesos
- ☐ h_t el estado oculto en determinado t
- ☐ $\frac{\partial L}{\partial h_t}$ es el gradiente de la pérdida con respecto al estado oculto

Si las RNN clásicas tienen este fallo al usar datos con relaciones a largo plazo, como pueden ser mis datos meteorológicos, nos queda una opción que arregla este problema, este modelo se llama LSTM. Las LSTM son un tipo de RNN creada por Hochreiter, S., & Schmidhuber en su estudio Long short-term memory. Neural Computation(1997) (Hochreiter, 1997)

Este modelo consigue arreglar el problema de la desaparición del gradiente de las RNN tradicionales, esto lo consigue con una arquitectura de puertas de entrada, puertas de olvido y puertas de salida. Esta arquitectura permite el flujo constante del gradiente a través del tiempo, esta estructura consigue que la información relevante se mantenga y se propague a través de las capas del modelo.

Veamos cómo funcionan estas puertas, la primera de ellas es la puerta de olvido, esta puerta determina qué información debe ser tomada en cuenta de la neurona anterior, la información relevante se calcula usando la función sigmoide, cuanto más cercano a 0 sea el

⁵ Derivada de la función de pérdida con respecto a los pesos

valor mayor posibilidad de ser olvidada y cuanto mayor sea a 1 mayor es la posibilidad de ser retenida.

$$f_t = \sigma(W_f * [h_{t-1}, x_t] + b_f) \text{ siendo:}$$

- x_t la entrada actual
- h_{t-1} el estado oculto anterior
- σ la función sigmoide ⁶

Cuando la puerta de olvido ha sacado sus resultados, la puerta de entrada determina cual de la información nueva debe almacenarse, de nuevo usa la función sigmoide para escoger los valores y también usa una función tangente hiperbólica para crear un vector de los nuevos valores candidatos.

$$i_t = \sigma(W_i * [h_{t-1}, x_t] + b_i)$$

Antes de actualizar el estado se generan candidatos para la nueva fase.

$$\tilde{C}_t = \tanh(W_C * [h_{t-1}, x_t] + b_C)$$

Cuando ya se han generado los candidatos se combinan con la puerta de entrada para actualizar el estado.

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t$$

Finalmente, la puerta de salida decide qué información será enviada al siguiente estado oculto, usando también la sigmoide y una tangente hiperbólica.

Así calcula el estado oculto:

$$o_t = \sigma(W_o * [h_{t-1}, x_t] + b_o)$$

En el último paso se combina la puerta de salida y el estado de la neurona:

$$h_t = o_t * \tanh(C_t)$$

De esta manera un LSTM arregla el problema del descenso de gradiente de su predecesora la RNN vanilla⁷.

⁶ La función sigmoide es una función matemática que transforma cualquier valor real en un número entre 0 y 1

⁷ El término *vanilla* se utiliza en aprendizaje automático para referirse a una versión básica o sin modificaciones de un modelo

Por otro lado, usaremos GRU (Gated Recurrent Unit) (Cho, 2014) este modelo es una versión simplificada del LSTM que también ataca la desaparición del gradiente, pero con sólo dos puertas en lugar de tres: la puerta de actualización (que combina las funciones de olvido e ingreso) y la puerta de reinicio. Al fusionar la puerta de entrada y la de olvido en una sola, el GRU reduce el número de parámetros y el coste computacional, manteniendo un flujo de gradiente eficiente a través del tiempo.

Mientras que el LSTM gestiona por separado qué información olvidar, qué nueva incorporar y qué enviar al siguiente estado oculto, el GRU decide primero cuánto de la información pasada conservar (update gate) y luego, con la puerta de reinicio, gestiona cuánto de la memoria previa debe olvidarse al generar los candidatos de estado. En la práctica, esto suele traducirse en tiempos de entrenamiento más rápidos y rendimiento similar al LSTM en muchas tareas secuenciales, aunque el LSTM puede ofrecer cierta ventaja cuando hay estructuras de dependencia muy complejas.

En conjunto, ARIMAX y SARIMAX aportan un marco sólido y transparente para explotar tanto la dependencia temporal intrínseca de la serie como el efecto de variables exógenas y patrones estacionales —ideal para capturar la regresión de temperatura y los ciclos convectivos reflejados en el CAPE— siempre que se garantice la estacionariedad y se seleccione cuidadosamente el orden de los procesos. Por su parte, las RNN (en especial LSTM y GRU) superan el problema del desvanecimiento del gradiente de las redes recurrentes clásicas, manteniendo información relevante a largo plazo: las LSTM, con sus tres puertas, ofrecen máxima flexibilidad ante dependencias complejas; las GRU, más ligeras, reducen parámetros y aceleran el entrenamiento con un desempeño comparable. Así, combinando la interpretabilidad estadística de ARIMAX/SARIMAX con la capacidad adaptativa de LSTM/GRU, se obtiene una estrategia integral y robusta para predecir la temperatura.

3. TRABAJO TÉCNICO

Herramientas empleadas

En este Trabajo Fin de Grado se ha empleado Python a través de dos plataformas diferenciadas: una orientada al tratamiento inicial de la información y otra dedicada al desarrollo y optimización de los modelos predictivos.

La fase de preprocesamiento se llevó a cabo en PyCharm, donde se realizaron las tareas de conversión de archivos GRIB a formato CSV, transformación de coordenadas geográficas y depuración de los datos mediante estrategias de normalización y gestión de valores atípicos. Seguidamente, en Google Colab se acometió el entrenamiento de las estructuras analíticas, beneficiándose del acceso a GPU y TPU gratuitas para agilizar la puesta a punto de modelos de series temporales, algoritmos de aprendizaje supervisado y redes neuronales profundas.

De forma transversal, se recurrió a un conjunto de librerías consolidadas en el ámbito de la ciencia de datos:

- pandas: gestión de estructuras tabulares, lectura/escritura de CSV y tratamiento de valores faltantes.
- NumPy: operaciones numéricas eficientes sobre arrays multidimensionales.
- Matplotlib: elaboración de gráficos para la exploración de series temporales y la visualización de resultados.
- statsmodels y pmdarima: análisis de estacionariedad, construcción de modelos ARIMA y búsqueda automática de parámetros.
- scikit-learn: partición de datos en entrenamiento y validación, así como cálculo de métricas (MAE, MSE, R^2).
- XGBoost: implementación de regresión por boosting de gradiente para optimizar la precisión de las predicciones.
- PyTorch: definición y entrenamiento de redes neuronales, gestión de lotes con DataLoader y ajuste dinámico de la tasa de aprendizaje mediante callbacks.
- Módulos estándar de Python como random, time, os, tempfile y warnings para garantizar la reproducibilidad, medir tiempos de ejecución y mantener el entorno limpio.

Ingeniería del dato

Mis datos son del ERA 5, un conjunto de datos meteorológicos creado por el Centro Europeo de Pronósticos a Plazo Medio (ECMWF). Esta base de datos llamada Copernicus Climate Data Storage (CDS) es parte del Programa Copernicus de la unión europea y abarca datos desde el 1950 hasta hoy.

Los datos del ERA5 son de reanálisis, los datos de reanálisis son una reconstrucción del clima global basada en observaciones y modelos meteorológicos. Son ideales para mi propósito ya que proporcionan información detallada sobre variables atmosféricas, terrestres y oceánicas con alta resolución temporal y espacial.

Para acceder a estos datos he usado la API⁸ CDS, lo que ha facilitado la obtención de datos. En esta API he conseguido extraer datos desde 2003 hasta 2018 y finalmente los he acotado de septiembre del 2012 a febrero del 2018. Entre los 6 años de datos y las 12 variables finales y las coordenadas acotadas suman un total de 13.102.272 observaciones.

El output de la API son archivos únicos, por horas y en formato GRIB⁹, por lo que ha sido necesaria la transformación de GRIB a CSV.



Figura 1 Zona acotada en la extracción de datos

⁸ API: interfaz que permite la comunicación entre distintas aplicaciones para intercambiar datos de forma estructurada.

⁹ GRIB (Gridded Binary) es un formato binario utilizado para almacenar y distribuir datos meteorológicos en mallas regulares

Mis variables iniciales eran: time, latitude, longitude, u10, v10, d2m, t2m, msl, sst, sp, MeanSea y Cape.

Veamos más a fondo las variables: U10 representa el componente horizontal del viento en dirección este-oeste (eje x) a 10 metros sobre el suelo, los datos son numéricos y un valor negativo representa viento en dirección oeste y positivos en dirección este. La variable v10 es la componente horizontal de viento en dirección norte-sur (eje y) a 10 metros sobre el suelo, los valores positivos son vientos en dirección norte y los negativos en dirección sur.

La variable d2m, es otra variable numérica, en este caso nos indica la temperatura a punto de rocío a 2 metros de la superficie. Está es la temperatura a la que el aire debe enfriarse (a presión constante) para que el vapor se convierta en líquido, está muy relacionado con la humedad, a su vez la humedad está estrechamente ligada con el CAPE.

T2m, es una de nuestras variables objetivo para las predicciones, es la temperatura medida a dos metros sobre el suelo.

MSL, representa la presión a nivel del mar, son valores numéricos. El MSL es crucial para el estudio de fenómenos meteorológicos como las DANAs, ya que permite identificar la variación de presión que influye en estos fenómenos.

El sst es la temperatura en la capa superior del océano, esta variable numérica también medida en kelvin es importante en el estudio de la temperatura y el CAPE.

Sp es otra variable que mide la presión, en este caso no es exclusivamente a la presión del mar. Es un valor numérico, será muy relevante para las predicciones de CAPE y en menor medida para las de temperatura.

Finalmente, el CAPE, es la energía potencial convectiva disponible, es una medida sobre la cantidad de energía disponible en la atmósfera para la creación de tormentas. Cuanto mayores son los valores, más fuerte es el fenómeno.

Tabla 1 Formatos variables antes de transformaciones

	Formato
Time	Object
Latitude	Float64
Longitude	Float64

U10	Float64
V10	Float64
D2m	Float64
T2m	Float64
msl	Float64
sst	Float64
sp	Float64
MeanSea	Float64
Cape	Float64

Fuente: elaboración propia.

El principal paso en la transformación de las variables era traducir las coordenadas en municipios, para ello escogí un día de datos y lo traduje con una API a *openstreetmap*¹⁰, al ser una API no tenía tokens ilimitados por lo que usé ese día de coordenadas traducidas para filtrar los demás archivos, una vez hecho esto las columnas Latitude y Longitude pueden ser borradas, como también ha sido borrada la columna de MeanSea ya que por algún error en la extracción de los datos salía vacía por lo que era inútil mantenerla.

En cuanto a los valores nulos, en la variable Sst, en las zonas que no son mar el valor es nulo, para arreglar este problema he rellenado los valores nulos con '0' para que posteriormente el modelo no vea como significativo para la temperatura en tierra la temperatura de la superficie del mar.

El siguiente es pasar las variables que están en kelvin a grados Celsius con esta relación $^{\circ}\text{C} = K - 273.15$.

Para que el modelo pueda obtener más información de los datos he creído conveniente hacer dos variables más de las que ya había extraído, estas son : WindMagnitude y Air-Sea Temp Difference.

Para calcular la velocidad del viento he usado la fórmula de la magnitud de un vector en el plano bidimensional $V = \sqrt{v_{10}^2 + v_{10}^2}$ y para la variable Air-Sea Temp Difference es la diferencia de la temperatura del agua y la del aire. Ambas variables son importantes ya que la velocidad del viento influye en la convectividad de la atmosfera y la diferencia de

¹⁰ *OpenStreetMap* es un proyecto colaborativo que ofrece datos geográficos libres, contruidos y mantenidos por una comunidad global de voluntarios.

temperatura entre el aire y el mar ayuda a comprender la dinámica de los procesos de intercambio de energía en la superficie oceánica.

Con las nuevas variables creadas, otras variables suprimidas y otras transformadas lo más conveniente es normalizar los datos. En mi dataset sin normalizar hay valores que oscilan desde -1.19 a 101693 , por lo que la mejor opción tanto para la optimización del proceso de entrenamiento como para que el modelo entienda bien los datos, es la normalización. En la siguiente tabla podemos ver un ejemplo de algunas de mis variables, en la que se ve como los valores son dispares.

Tabla 2 Ejemplo media y moda antes de normalización

	Media	Moda
U10	0.504	-0.912
V10	-0.707	-1.19
D2m	285.92	285.59
T2m	291.05	285.5
msl	101699.47	101693.625

Fuente: elaboración propia.

Una vez normalizados los datos con el MinMaxScaler, los datos se ven así.

Tabla 3 Ejemplo media y moda después de normalización

	Media	Moda
U10	0.486	0.451
V10	0.550	0.538
D2m	0.751	0.745
T2m	0.528	0.460
msl	0.565	0.569

Fuente: elaboración propia.

Una vez hechas todas las transformaciones, veamos con que variables nos quedamos y su tipo.

Tabla 4 Variables y formatos después de transformaciones

	Formato
Time	Object
Air-Sea Temp Difference	Float64
Wind Magnitude	Float64
U10	Float64
V10	Float64
D2m	Float64
T2m	Float64
msl	Float64
sst	Float64
sp	Float64
Nombre_loc	Object
Cape	Float64

Fuente: elaboración propia.

Los valores atípicos, en mi caso, son importantes ya que para la predicción del CAPE son esos valores atípicos de otras variables los que entran en juego. El cape, por lo general, tiene valores cercanos a 0, pero cuando estos valores aumentan es cuando la probabilidad de fuertes tormentas y otros fenómenos se dispara.

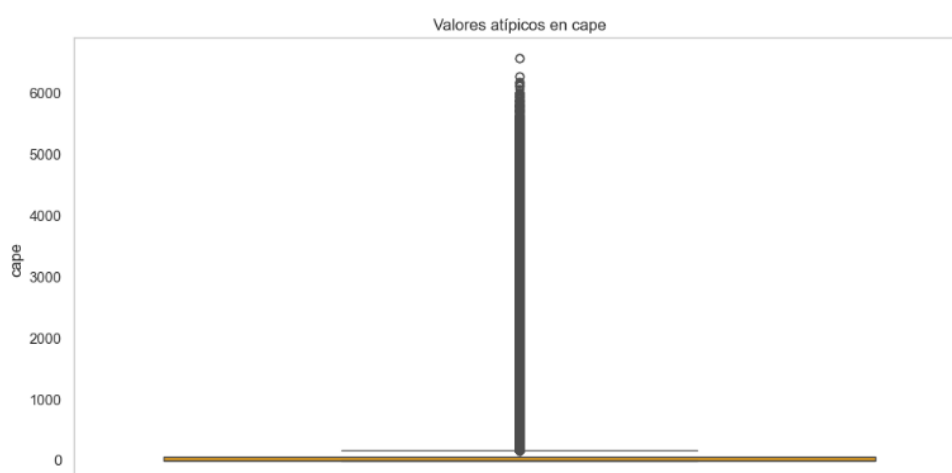


Figura 2 Valores atípicos Cape

Como vemos en la gráfica, el CAPE está formado prácticamente por valores atípicos y ruido, lo que hace que el diagrama parezca “desordenado” o poco informativo a simple vista. Sin embargo, esto es completamente normal: la distribución de CAPE es muy sesgada

y presenta picos extremos únicamente en días con tormentas severas o DANA, mientras que la mayoría de las jornadas tiene valores muy bajos o nulos. En consecuencia, la apariencia “mala” de la gráfica no refleja un fallo en el modelo, sino la realidad física de que los eventos convectivos intensos son poco frecuentes y generan esa acumulación de outliers.

Se ha decidido no cambiar en nada los valores atípicos de mi dataset, ya que para la predicción del cape sin los valores atípicos de otras variables sería prácticamente imposible.

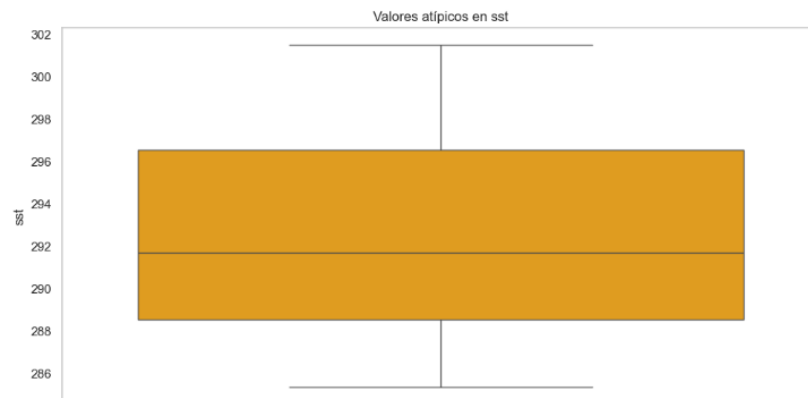


Figura 3 Valores atípicos sst

En concreto la temperatura del mar es muy robusta, se mantiene estable con cierta estacionalidad, aunque eso lo veremos más adelante en el análisis de los cuatro componentes principales de las series temporales.

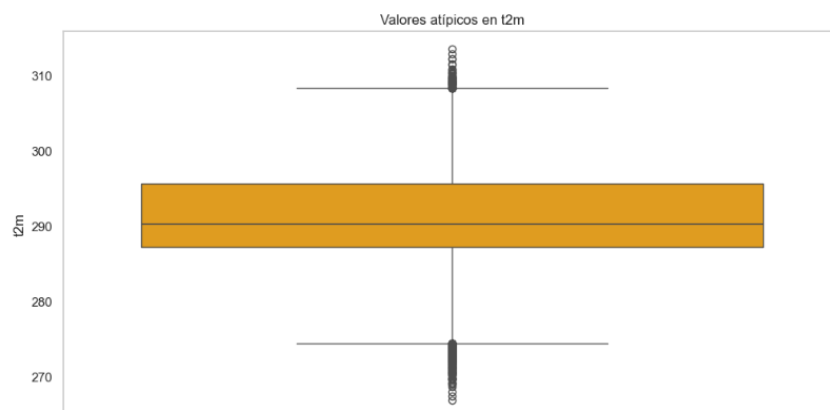


Figura 4 Valores atípicos T2m

La temperatura a dos metros del suelo tiene más variabilidad que la temperatura de la superficie del agua, vemos que la mediana esta exactamente en los 290 kelvin unos 16.85 grados Celsius.

Por lo general los demás gráficos son muy parecidos a los tres que muestro, por lo que veo correcto pasar a los diagramas de dispersión.

En cuanto a los diagramas de dispersión, es importante recalcar que en la meteorología las relaciones suelen ser menos obvias ya que no suele haber una variable que solo dependa de otra, sino que cada variable depende un poco de las otras. Con este contexto, es valioso encontrar relaciones por muy pequeñas que sean.

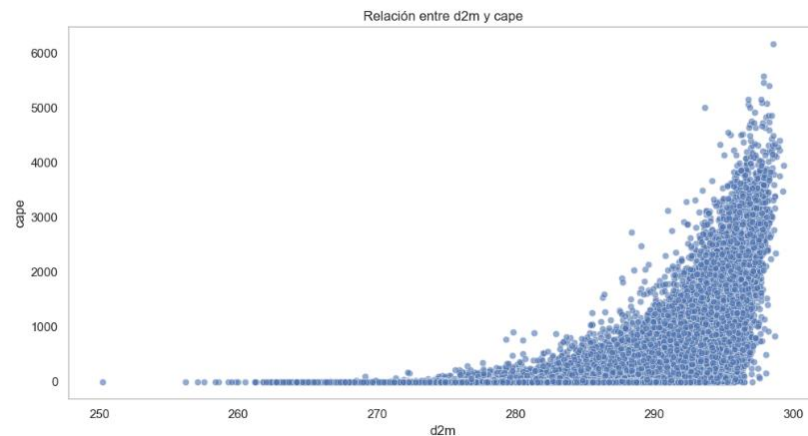


Figura 5 Relación d2m y cape

En este caso vemos que la temperatura a punto de rocío ($d2m$) tiene cierta relación con el cape, y esto tiene sentido ya que cuanto mayor es la temperatura a punto de rocío mayor es la humedad, y cuanto mayor sea la humedad más capacidad convectiva tiene la atmosfera.

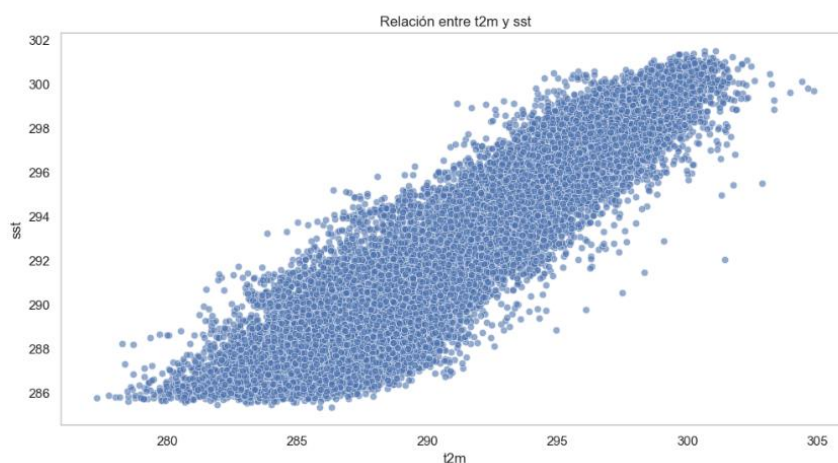


Figura 6 Relación t2m y sst

La temperatura de la superficie del mar y la de la tierra tiene cierta relación, y es algo obvio, cuanto más calor hace mayor será la temperatura del agua.

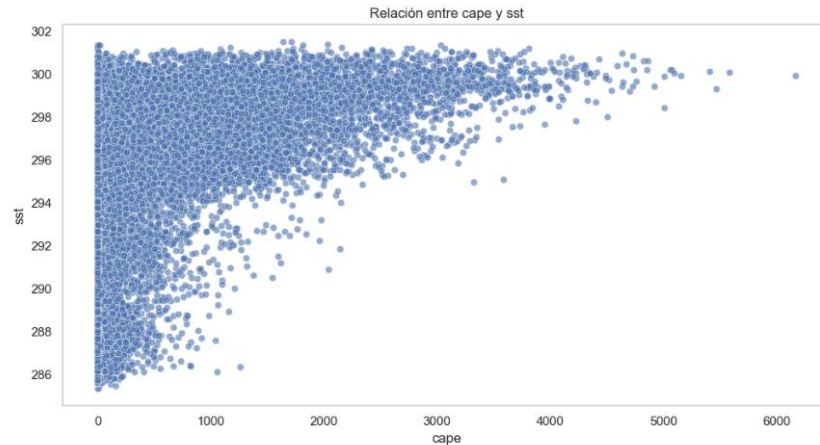


Figura 7 Relación cape y sst

El cape está relacionado con el sst de la misma manera que está relacionado con el d2m, por la humedad. Cuanto mayor sea la temperatura de la superficie del agua más humedad se generará y con ello la posibilidad de valores del Cape más alto.

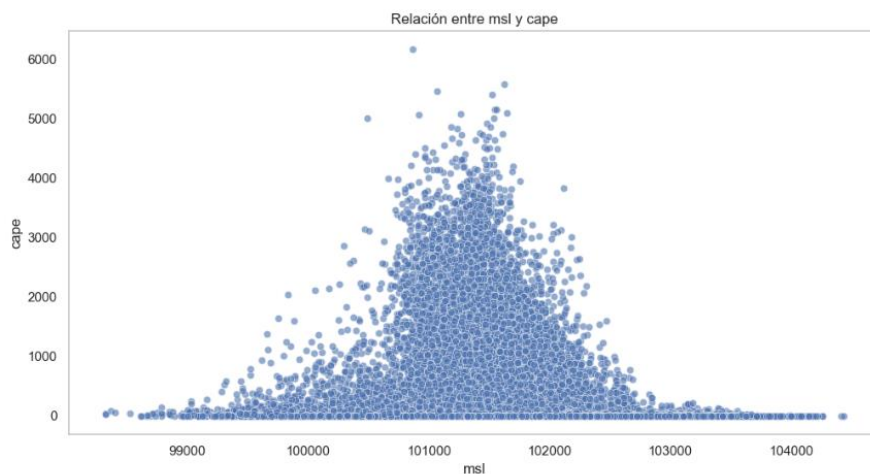


Figura 8 Relación msl y cape

El cape suele registrar mayores valores en el mar, en parte por la cantidad de humedad y en parte por el msl, en torno a los valores de 101500 el msl se considera bajas presiones, a estas presiones el aire sube más fácilmente creando nubes y tormentas. En ingles las danas se llaman *Cut-off low* a lo que se refiere es a una zona en la que el msl ha tenido niveles bajos y esto genera una corriente fuerte de aire que queda aislada en los niveles altos de la atmosfera.

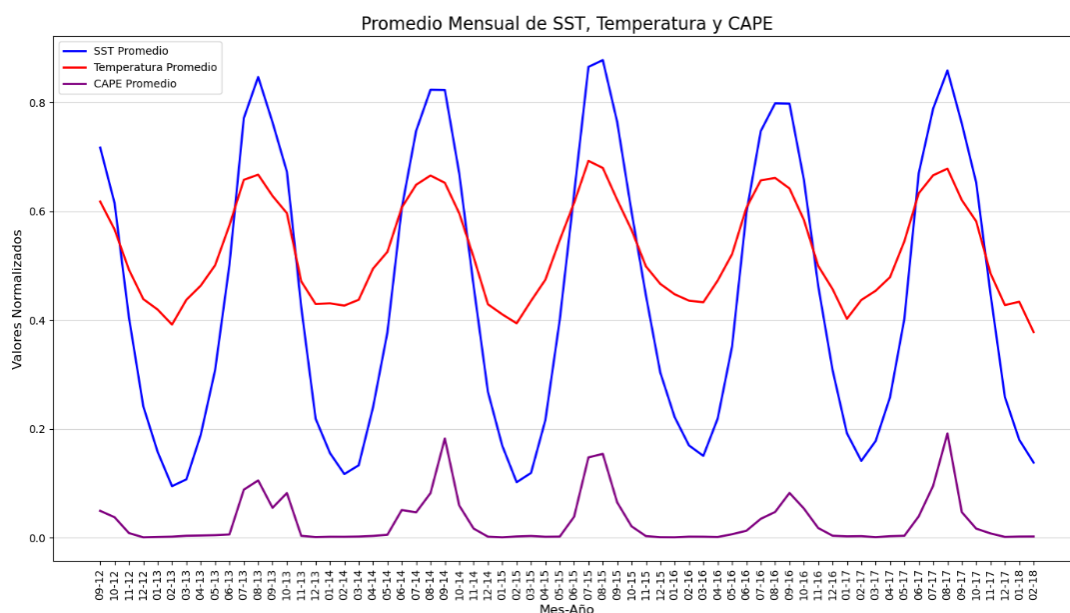


Figura 9 Promedio mensual sst, temperatura y cape

En este caso se ve, como las tres variables mencionadas anteriormente tienen el mismo patrón.

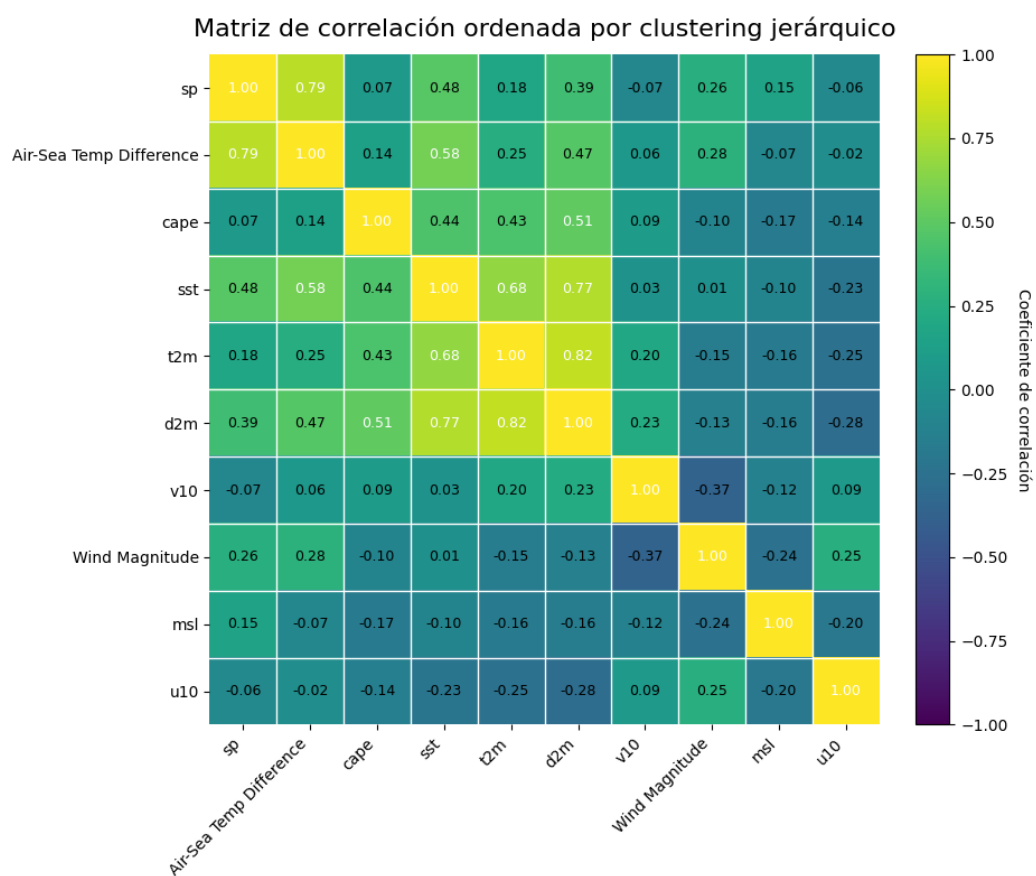


Figura 10 Matriz correlación

En esta matriz de correlación podemos destacar las correlaciones entre el t2m el d2m y el sst, por otro lado, el sp y la diferencia de temperatura del agua y el aire. Estas relaciones son valiosas y ayudaran mucho, pero como he comentado antes, en la meteorología la dificultad se encuentra en las pequeñas relaciones como la del cape y el sst, el d2m y el t2m. Estas relaciones con el cape nos muestran la importancia de la humedad en los valores del Cape, por otro lado, el msl tiene una relación pequeña con el cape(-0.17) pero esto no indica que no influya la una con la otra, ya que en el caso del msl lo que influye en el cape es el cambio de presión no la presión, es cuando esta presión baja drásticamente cuando el cape debería aumentar.

Teniendo en cuenta que mis datos son series temporales, el análisis de las cuatro componentes es crucial. Estas cuatro componentes son: Serie original, tendencia, estacionalidad y residuo.

En la tendencia pretendemos captar la evolución a largo plazo de las variables estudiadas, en la estacionalidad se busca entender si las variables tienen un patrón recurrente en el tiempo y en el residuo podremos encontrar patrones que no son visibles en las demás componentes.

Veamos primeramente la estacionalidad, vamos a usar el método de Dickey-Fuller aumentada (Dickey, 1979), en esta prueba tenemos dos hipótesis h_0 y h_1 . El h_0 es la hipótesis nula, esta establece si la serie tiene una raíz unitaria, esto indicaría la no estacionalidad de la variable, por lo que la hipótesis alternativa h_1 indica la estacionalidad de una variable.

En esta prueba obtenemos el *Estadístico ADF* y el *p-valor* (Fisher, 1925) , el estadístico ADF es el resultado de la prueba, cuanto menor sea el valor mayor es la evidencia de la estacionalidad. En el caso del p-valor, si este es menor a 0.05 también se rechaza la hipótesis nula.

Para la fácil y rápida comprensión sobre la estacionalidad de las variables, se mostrarán en formato de tabla:

Tabla 5 Estacionalidad Variables

	p-valor	ADF	Estacionalidad
Air-Sea Temp Difference	1.2e-22	-12.19	SI
Wind Magnitude	0.0	-20.30	SI
U10	0.0	-20.52	SI
V10	2.6e-28	-15.46	SI

D2m	3.17e-27	-14.62	SI
T2m	4.09e-30	-17.56	SI
msl	7.84e-05	-4.71	SI
sst	0.0	-38.07	SI
sp	0.0	-29.37	SI
Cape	0.0	-24.55	SI

Fuente: elaboración propia.

En mi caso, todas las variables son estacionales, tiene sentido al ser todas meteorológicas.

Veamos las cuatro componentes de algunas de nuestras variables, empecemos con el sst. La tendencia deja claro que a lo largo de los seis años de estudio la temperatura del mar ha subido 0.8 grados kelvin, este hecho es preocupante, ya que como hemos visto anteriormente los valores altos de cape están estrechamente relacionados con la humedad y cuan mayor sea la temperatura del mar mayor será la humedad.

Como hemos visto antes, y el gráfico nos corrobora, la variable sst es estacional.

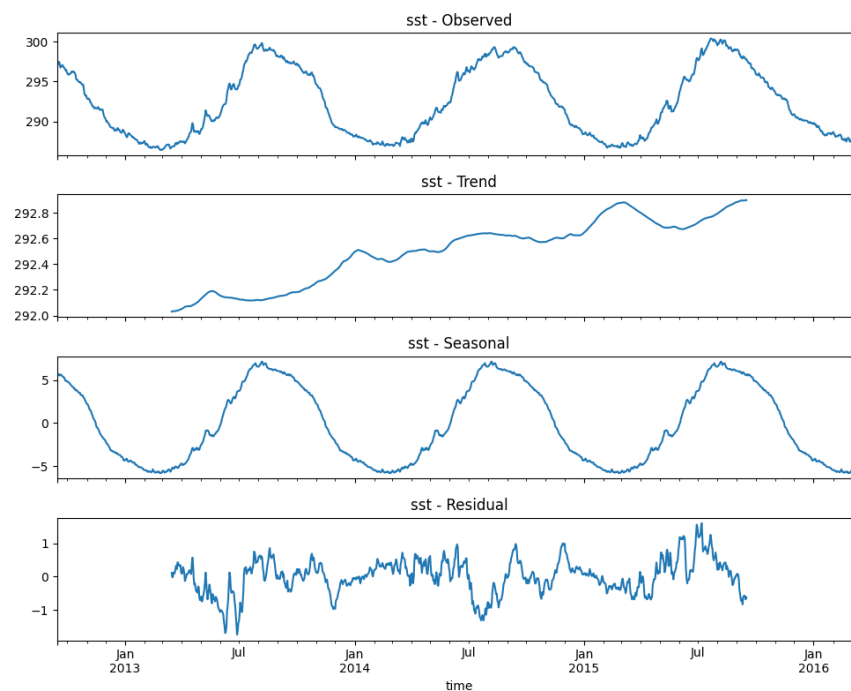


Figura 11 Cuatro componentes sst

Vamos a ver junto al sst las gráficas del d2m y el t2m, que deberían ser muy parecidas.

En cuanto a tendencia el d2m y el t2m están muy parejos, también muestran una tendencia al alza llegando a subir en el caso del t2m 0.75 grados kelvin y en el d2m 1 grado

kelvin. Como ya he comentado esto preocupa especialmente por la posibilidad de aumento de valores del cape que se pueden traducir en catástrofes naturales con mayor frecuencia.

La estacionalidad está clara en estas variables, lo hemos visto en el análisis ADF y se puede volver a ver a simple vista en los gráficos.

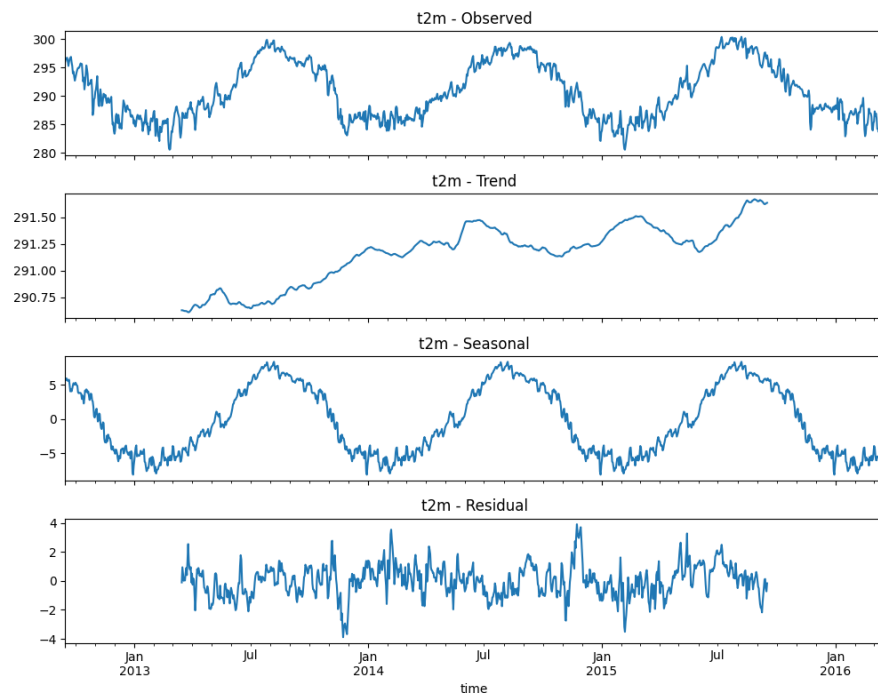


Figura 12 Cuatro componentes t2m

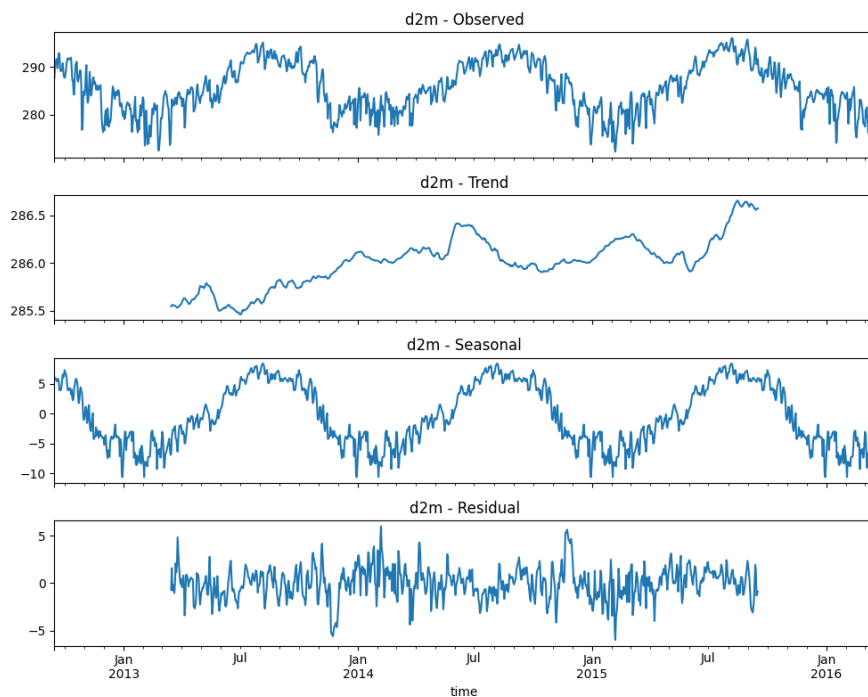


Figura 13 Cuatro componentes d2m

Y si las temperaturas han subido y con ello la humedad, lo lógico sería ver un aumento en la tendencia del cape, veamos:

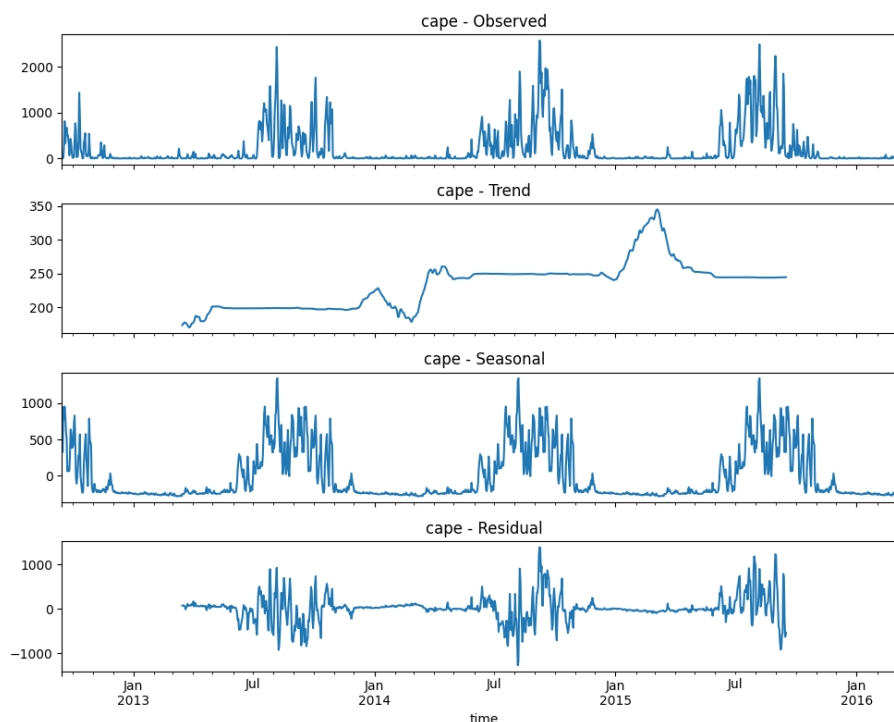


Figura 14 Cuatro componentes cape

Como era de esperar la tendencia está creciendo, es más irregular que la que hemos visto en las tres variables anteriores, pero se aprecia claramente el aumento. Este aumento de temperaturas y de cape aumenta mi afán por intentar conseguir una manera de predecirlo con la mayor exactitud posible, ya que si seguimos con la misma tendencia es posible que la recurrencia de estos fenómenos sea mayor y de mayor agresividad.

Veamos la presión atmosférica, la tendencia vuelve a ser al alza. En este caso el análisis es más complejo, un aumento en la presión sugiere lo que se llama un *desplazamiento de altas presiones* este fenómeno suele venir de la mano de bajadas del viento, como podemos ver en la gráfica de la magnitud del viento en este caso también es así, el desplazamiento de altas presiones afecta directamente a las precipitaciones creando sequías más largas y prolongada, según Mestre y Moreno (Mestre, 2009), el desplazamiento de las altas presiones subtropicales hacia latitudes más elevadas puede conducir a la persistencia de sistemas de alta presión sobre el suroeste de Europa, incluyendo la Península Ibérica, lo que resulta en una disminución significativa de las precipitaciones y favorece la ocurrencia de sequías prolongada. Este panorama empeora, las temperaturas suben y la

probabilidad de fenómenos agresivos también, pero a su vez la presión también sube creando sequias más prolongadas.

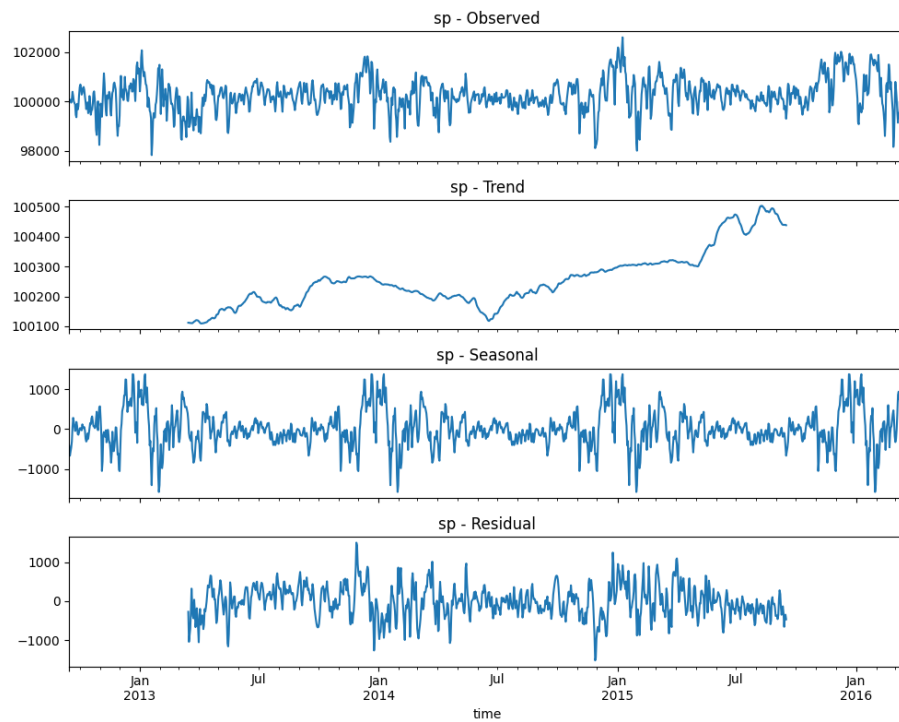


Figura 15 Cuatro componentes sp

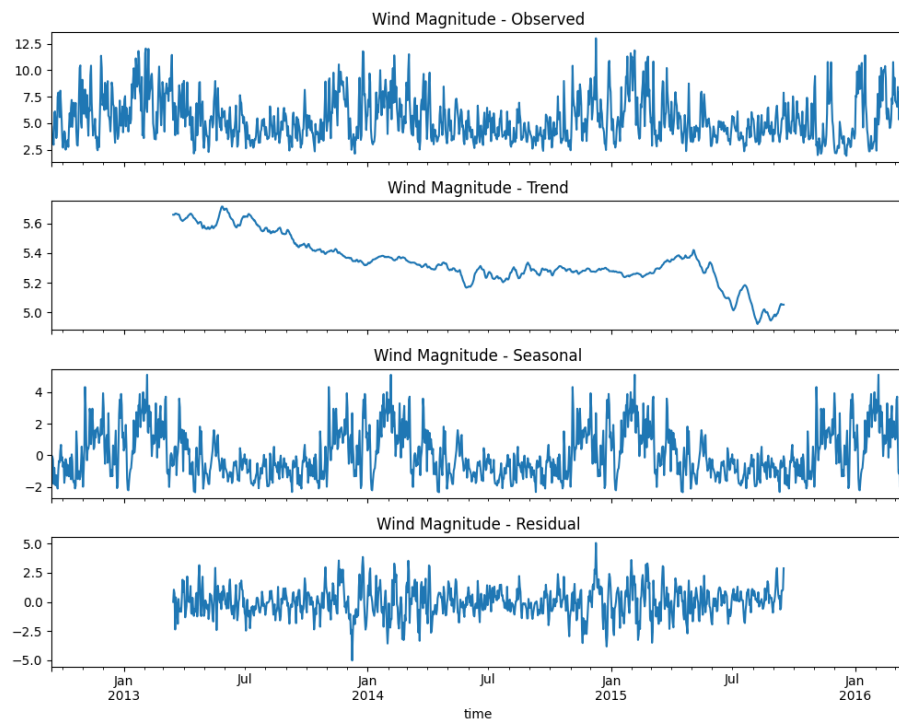


Figura 16 Cuatro componentes wind-magnitude

El trabajo de ingeniería del dato ha sido clave para garantizar que los modelos reciban datos de calidad. Aunque las transformaciones aplicadas han mejorado significativamente la estructura del dataset, siempre existe la posibilidad de que ciertos patrones se escapen o que haya factores no considerados que afecten las predicciones. Aun con estas limitaciones, los datos han sido procesados de manera rigurosa para maximizar la capacidad predictiva de los modelos.

Análisis del dato

Para poder usar mis datos el primer paso ha sido hacer ventanas temporales de los mismos, los modelos tendrán de entrada 7 días para dar una predicción de 5 días.

Las variables que se van a analizar son el t2m(temperatura a dos metros del suelo) y el CAPE, son dos variables muy distintas entre sí por lo que se hará un modelo para cada una de ellas y las comparaciones serán con otros modelos hechos para la misma variable, ya que comparar el modelo de CAPE y T2M entre sí sería injusto por lo diferentes que son los datos en cuanto a su distribución.

El cape es una variable complicada, con muchos valores cercanos a 0 y otros sobre los 1500-3500, estos son los valores que realmente importan en este análisis, ya que si el cape está entre 0-1000 no es relevante pero cuando supera los 1500 empieza a serlo, por lo que no se busca un modelo con una gran precisión sino un modelo que pueda captar las tendencias.

Tabla 6 Media y desviación estándar Cape y T2m

	Media	D. Estándar
T2m	0.52	0.109
Cape	0.031	0.081

Fuente: elaboración propia.

Visto por encima el problema que se plantea y como son los datos veamos que modelos se han usado.

Para la temperatura se han desarrollado cuatro modelos divididos en dos enfoques, el enfoque clásico (ARIMA,SARIMAX) y el enfoque más moderno (GRU,LSTM) para estos modelos se han utilizado cuatro métricas clave para comparar su desempeño: R², MAE,MAPE y RMSE.

Para evaluar la calidad de los modelos, se emplearán diversas métricas de error y ajuste. El coeficiente de determinación R^2 medirá la proporción de la variabilidad explicada por el modelo, siendo su valor un indicador de la capacidad predictiva. Un R^2 cercano a 1 sugiere que el modelo representa fielmente la variabilidad de la variable analizada, lo cual es especialmente útil en estudios meteorológicos donde la precisión en la predicción de tendencias es clave.

El error cuadrático medio (RMSE) evaluará la diferencia promedio entre los valores predichos y los reales, penalizando en mayor medida los errores grandes. En el caso del T2M, esta métrica permitirá detectar desviaciones significativas en la predicción de la temperatura, mientras que para el CAPE resultará crucial para identificar posibles sobrestimaciones o subestimaciones en la predicción de la energía convectiva, un factor determinante en la formación de tormentas.

Por su parte, el error absoluto medio (MAE) calculará el error promedio absoluto entre las predicciones y los valores reales sin penalizar los errores de gran magnitud, proporcionando una visión general del error típico que se puede esperar. En el contexto del T2M, el MAE permitirá cuantificar la desviación media de la temperatura prevista con respecto a la observada, mientras que en el caso del CAPE, indicará la precisión con la que el modelo estima la energía disponible para la convección.

Finalmente, el error porcentual absoluto medio (MAPE) expresará en términos relativos el error porcentual promedio entre las predicciones y los valores reales. Esta métrica será especialmente útil para evaluar la precisión del modelo en función de la magnitud de la variable analizada, facilitando la interpretación del error relativo tanto en la predicción de la temperatura como en la estimación del CAPE, donde la variabilidad de los valores puede ser considerable.

Para abordar el problema de predicción de temperatura, se han implementado dos enfoques contrastados: un ARIMA y SARIMAX (como métodos clásicos de series temporales) y un GRU y LSTM (como representantes de modelos de deep learning). En esta sección, se detallará la aplicación concreta del SARIMAX a nuestro caso de uso.

Inicialmente, la elección de SARIMAX se fundamentó en su capacidad para integrar patrones temporales auto correlacionados, lo que permite capturar la dependencia existente entre observaciones consecutivas, como en el caso de la temperatura, donde los valores registrados en un día determinado están influenciados por los del día anterior. Asimismo, el

modelo es capaz de representar adecuadamente la estacionalidad semanal, lo que resulta fundamental en el análisis de fenómenos que presentan variaciones periódicas en ciclos de siete días.

Además, se consideró su habilidad para incorporar variables exógenas cíclicas, permitiendo codificar información temporal en términos de semanas y meses. Este enfoque contribuye a evitar discontinuidades en la representación del tiempo, lo que mejora la coherencia de las predicciones y permite capturar mejor la evolución de las variables a lo largo del periodo de estudio.

De manera que en el código esto se implementa de esta forma:

1. Selección automática de hiperparámetros: *pm.auto_arima* con *seasonal=true* y *m=7*, con esto busco el mejor SARIMAX que minimice el AIC, por otra parte, se incluyen las variables exógenas con *X=train_exog*
2. Variables exógenas: se añade *dayofweek_sin/cos* y *month_sin/cos* para modelar la estacionalidad diaria y mensual de forma continua.

Veamos ahora el ARIMA, a diferencia de SARIMAX este no incorpora la estacionalidad ni las variables exógenas para capturar los patrones cíclicos de los datos. El modelo ARIMA se ha implementado con el propósito de evaluar si la inclusión de estas componentes tiene un impacto significativo en la calidad de las predicciones.

El modelo ARIMA es una técnica clásica de series temporales que se basa en tres componentes principales:

- AR(AutoRegresivo): Usa valores pasados de la serie para predecir los futuros.
- I (Integrado): Aplica diferenciación a la serie para hacerla estacionaria.
- MA (Media Móvil) : Modela la dependencia de un punto de la serie con los errores pasados

Dado que ARIMA no incorpora estacionalidad ni variables exógenas, su capacidad de modelado depende exclusivamente de la estructura interna de la serie temporal. Por lo que, su implementación permitirá contrastar si la introducción de estacionalidad y variables exógenas en SARIMAX representa una mejora sustancial en la precisión del modelo.

El modelo ARIMA se ha implementado con la siguiente estructura en código:

1. Selección automática de hiperparámetros:

- Se utiliza `pm.auto_arima()` para encontrar los valores óptimos de (p, d, q), donde:
 - p: Número de términos autorregresivos.
 - d: Número de diferenciaciones necesarias para estacionalizar la serie.
 - q: Número de términos de media móvil.
- A diferencia del SARIMAX, aquí `seasonal=False`, ya que no se modela estacionalidad.

2. Entrenamiento del modelo:

- Se ajusta un ARIMA sobre la serie temporal de temperatura (`t2m`), excluyendo las variables exógenas.
- La estimación de los coeficientes se realiza con `ARIMA(train, order=order).fit()`, donde `order` es el conjunto de parámetros determinado por `auto_arima()`.

Comencemos con el segundo enfoque, empezando por el modelo GRU (Gated Recurrent Unit) este es la alternativa avanzada al SARIMAX y el ARIMA, está especialmente diseñada para capturar patrones no lineales y dependencias temporales complejas en los datos de la temperatura.

GRU es una variante de las RNN que usa mecanismos de Gating para controlar el flujo de información a lo largo del tiempo, el gating permite al modelo retener o descartar información relevante. El control adaptativo que ofrece el gating facilita el aprendizaje de patrones temporales en series de datos, lo cual es particularmente útil en la predicción de la temperatura, donde la evolución de los valores depende de las condiciones pasadas y presenta variaciones tanto en el corto como el largo plazo.

GRU se compone de dos puertas, la de actualización y la de reinicio, la puerta de actualización determina la cantidad de información previa que se conserva para la predicción actual y la puerta de reinicio decide cuando no es apropiado retener esa información. Esto permite a GRU manejar de manera efectiva cambios abruptos como los de la temperatura de noche y de día, o los cambios bruscos de temperatura de un día para otro.

En comparación con SARIMAX y ARIMA, GRU tiene un mejor manejo de outliers usando sus mecanismos de gating puede ignorar los datos ruidosos que no aportan información y puede modelar relaciones no lineales más complejas.

Al igual que GRU, LSTM (Long Short-Term Memory) es otra arquitectura basada en RNN diseñada para resolver el problema del desvanecimiento del gradiente y aprender dependencias a largo plazo. LSTM es ligeramente más compleja que GRU, ya que introduce una celda de memoria explícita que actúa como un "canal de información a largo plazo", y tres puertas:

- Puerta de olvido, que decide qué información previa descartar.
- Puerta de entrada, que regula qué información nueva se almacena en la celda.
- Puerta de salida, que controla qué información se usa para generar la salida en cada instante.

Esta arquitectura le da a LSTM una mayor capacidad de memoria, lo que puede ser útil en tareas donde las dependencias entre eventos ocurren con mayor separación temporal. No obstante, también es más pesada computacionalmente. Por eso es interesante compararla con GRU, que tiene una arquitectura más simple, pero, en muchos casos, un rendimiento equivalente.

Para que la comparación sea justa y se evalúen en igualdad de condiciones, el modelo LSTM se configura exactamente con la misma estructura que el GRU, utilizando el mismo esquema Encoder-Decoder y los mismos hiperparámetros clave.

Antes de nada, en el código, se agrupan los datos por localización y por hora, ya que, en alicante, por ejemplo, hay muchas coordenadas y estudiarlas todas no es mi idea ni tengo la capacidad de hardware para hacer entrenamientos de millones de observaciones. Por lo que, después de agruparlas se procede a hacer una división del dataset en ventanas de 7 días de entrada y 5 de salida, posteriormente se divide el dataset en test y train (80/20) pero en esta división no hay shuffle¹¹ ya que perderíamos información e interpretabilidad.

Tanto el modelo GRU como el LSTM siguen una arquitectura tipo Encoder-Decoder, que permite manejar secuencias de entrada y salida de diferente longitud. En este caso, el encoder transforma los 7 días de entrada en un estado interno, que luego el decoder usa para

¹¹ *Shuffle* hace referencia a la acción de mezclar aleatoriamente las observaciones de un conjunto de datos.

generar los 5 días de salida. Veamos la arquitectura, en este caso es igual para el LSTM que para el GRU:

1. Encoder
 - a. Una capa de 2 capas con 128 unidades cada una y un dropout del 0.19
 - b. BatchNormalization para estabilizar el entrenamiento
2. Decoder
 - a. Otra capa con 2 capas de 128 unidades cada una y un dropout de 0.15
 - b. BatchNormalization para normalizar las salidas del decoder
3. Dense Layer
 - a. Capa lineal que mapea las 128 unidades de salida del decoder a las n_ubicaciones

Ambos modelos se entrenan durante un máximo de 200 epochs, pero se implementa early stopping con una paciencia de 35 epochs. Esto significa que, si la pérdida de validación no mejora en 35 iteraciones consecutivas, el entrenamiento se detiene automáticamente para evitar el sobreajuste y reducir el tiempo de cómputo innecesario.

Se utiliza AdamW (Loshchilov, 2017) como optimizador, con una tasa de aprendizaje inicial de 0.0056 y un weight_decay de $1e-5$, lo que ayuda a controlar la regularización L2 y evitar que el modelo memorice los datos de entrenamiento en lugar de generalizar. Además, el ReduceLROnPlateau ajusta la tasa de aprendizaje reduciéndola a la mitad si la pérdida de validación no mejora durante 15 epochs, permitiendo una convergencia más estable.

El entrenamiento se realiza en lotes de tamaño 1028 para acelerar el cálculo y estabilizar las actualizaciones de los pesos. En cada iteración, se calcula la pérdida de entrenamiento utilizando la función de error cuadrático medio (MSE), que mide la diferencia entre las predicciones del modelo y los valores reales de temperatura. Luego, la pérdida se propaga hacia atrás para actualizar los pesos de la red mediante backpropagation¹².

¹² *Backpropagation* es un algoritmo fundamental en el entrenamiento de redes neuronales. Consiste en calcular el gradiente de la función de pérdida con respecto a los pesos del modelo, propagando el error desde la salida hacia las capas anteriores mediante la regla de la cadena. Este gradiente se utiliza para ajustar los pesos a través de un optimizador, minimizando progresivamente el error en sucesivas iteraciones.

Después de cada epoch, el modelo se evalúa en los datos de validación para calcular la pérdida de validación, que indica qué tan bien está generalizando a datos no vistos y nos asegura que el modelo no tiene Overfitting.

Para saber si en nuestros modelos de redes neuronales hay overfitting se compara el val_loss en el conjunto de prueba y en el de entrenamiento en cada epoch, si hubiese overfitting veríamos que en el conjunto de entrenamiento seguiría aprendiendo y en el de prueba dejaría de aprender o ‘desaprendería’ veamos;

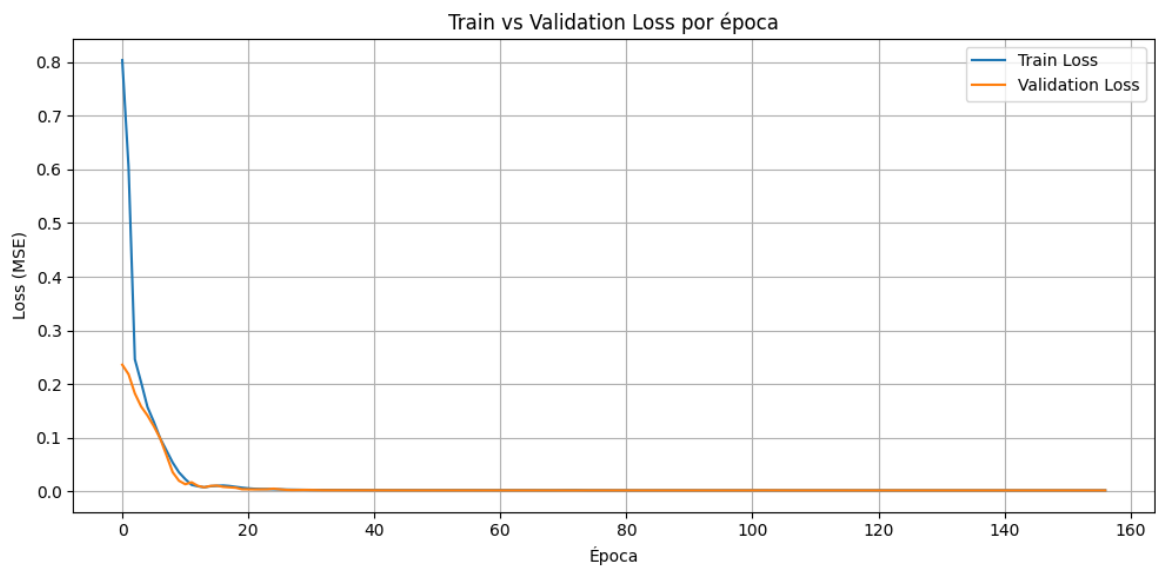


Figura 17 Train vs Test GRU

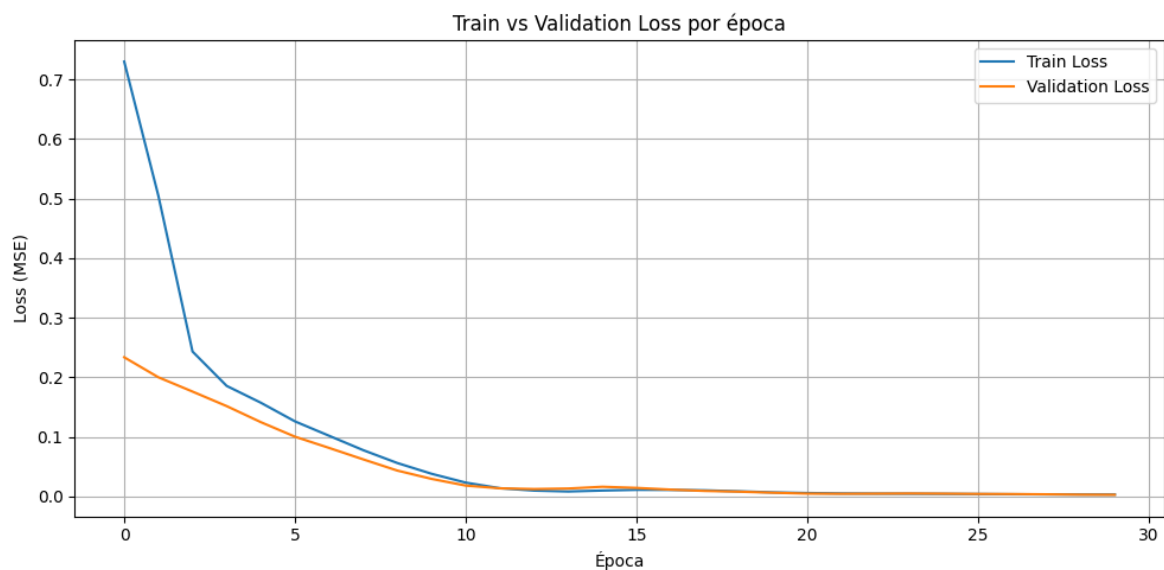


Figura 18 Train vs Test LSTM

En el caso del GRU y del LSTM no están sobreentrenados ya que vemos como las líneas de Train y de Test de las figuras 17 y 18 son parecidas, esto indica que el modelo se está ajustando correctamente tanto en train como en test, por lo que no está sobreentrenado, veamos ahora el caso del ARIMA y el SARIMAX.

Durante la modelización con ARIMA, se encontraron dificultades persistentes con el sobreajuste (overfitting). Se exploraron distintas combinaciones de hiperparámetros, incluyendo el uso de búsqueda en rejilla (grid search), pero no se consiguió un modelo adecuado. El paso lógico siguiente fue transitar a un modelo SARIMAX. Como se puede observar en la figura 19 el modelo ARIMA se aprende perfectamente los datos de entrenamiento y en el test no termina de funcionar.

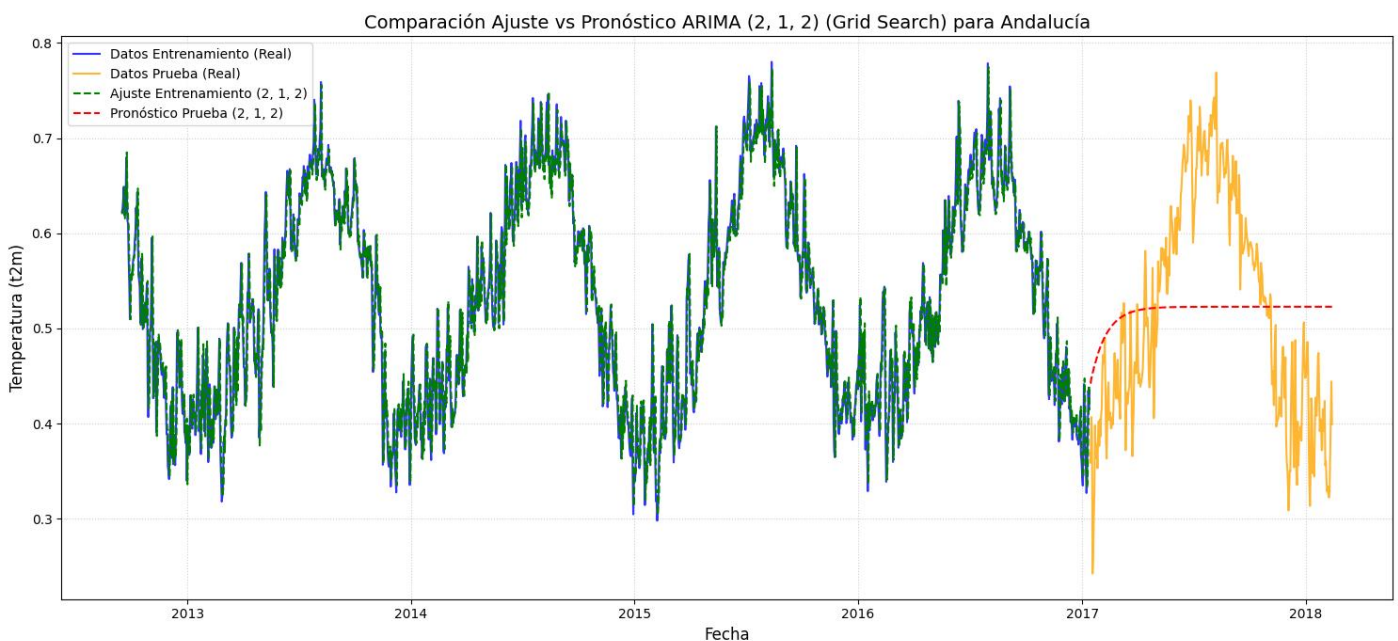


Figura 19 Predicción ARIMA

Tras analizar los modelos y su ajuste (o sobreajuste), veamos su rendimiento mediante las métricas.:

Tabla 7 Comparación métricas GRU y SARIMAX

	SARIMAX	ARIMA	GRU	LSTM
RMSE	0.049	0.118	0.0448	0.045
MAE	0.041	0.986	0.0344	0.347
MAPE	8.67%	19%	8.34%	8.35%
R ²	0.818	-0.034	0.891	0.891

Fuente: elaboración propia.

El primer paso en la comparación de modelos fue evaluar si la inclusión de la estacionalidad y de variables exógenas mejora el rendimiento predictivo. Para ello comparamos un SARIMAX con un ARIMA, donde el primero incorpora estacionalidad semanal y factores cíclicos en la predicción.

Los resultados muestran que SARIMAX supera al ARIMA en todas las métricas que hemos usado, con un RMSE menor (0.049 vs 0.118), un MAE significativamente inferior (0.041 vs 0.956) y un R^2 positivo (0.818 frente al -0.034 en ARIMA). La diferencia en las métricas nos confirma que la incorporación de la estacionalidad y las variables exógenas aportan información relevante a nuestro problema y evitan el sobreajuste, lo que justifica la elección de SARIMAX sobre ARIMA como el modelo clásico de referencia.

Desde el punto de vista de la meteorológica el resultado obtenido tiene sentido, ya que la temperatura no solo sigue una tendencia temporal, sino que también es influenciada por patrones cíclicos y factores externos. Modelos como ARIMA, que solo consideran la relación entre valores pasados sin incorporar información adicional, tienen a tener dificultades para capturar dinámicas complejas. En cambio, SARIMAX permite incluir componentes estacionarios, esto permite al modelo predecir de manera más precisa.

Al comparar los modelos de redes neuronales recurrentes GRU (Gated Recurrent Unit) y LSTM (Long Short-Term Memory), se puede ver que ambos presentan unas métricas muy similares con una ligera ventaja por parte del modelo GRU. Tal como se muestra en la Tabla 9, las métricas de error (RMSE, MAE y MAPE) y el coeficiente de determinación (R^2) son prácticamente iguales para ambos modelos: GRU obtiene un RMSE de 0.0448 y un R^2 de 0.891, mientras que LSTM alcanza un RMSE de 0.045 y el mismo valor de R^2 .

A pesar de esta paridad en términos de rendimiento predictivo, existe una diferencia significativa en cuanto a la eficiencia computacional. El tiempo de entrenamiento del modelo GRU es de 60.49 segundos, mientras que el LSTM requiere 10 minutos y 37 segundos para completar el mismo proceso. Esta diferencia se debe a la mayor complejidad del LSTM, que incorpora una puerta adicional en su estructura.

Dado que la mejora en precisión del LSTM no es sustancial respecto al GRU, pero sí implica un coste computacional considerablemente mayor, se opta por utilizar GRU como modelo de redes neuronales. Esta elección se justifica por su mejor equilibrio entre rendimiento y eficiencia.

A partir de esta selección, la comparación de rendimiento se realiza entre SARIMAX y GRU, ya que el objetivo principal es evaluar si un modelo basado en redes neuronales supera a un enfoque tradicional de series temporales mejorado.

Para interpretar estos resultados es necesario analizar cómo se comportan las diferentes métricas usadas para evaluar el desempeño de los dos modelos. En primer lugar, el error cuadrático medio (RMSE) este nos muestra una diferencia notable en los dos modelos, ya que el valor obtenido para SARIMAX(0.049) es ligeramente superior al del GRU (0.044). Esto indica que, en términos absolutos, las predicciones del modelo GRU se aproximan de manera mejor a los valores reales de nuestra serie temporal, esto es fundamental en la predicción de la temperatura donde la exactitud en la predicción es importante.

Por otra parte, el error absoluto medio (MAE) ,el modelo basado en GRU presenta un MAE relativamente parecido al registrado por el SARIMAX, pero en esta métrica sigue siendo levemente superior el GRU. Este resultado nos sugiere que, en promedio, el modelo GRU comete errores de menor magnitud al predecir los valores de la temperatura, lo que se traduce en una mayor fiabilidad en las predicciones. Para ver el error en términos relativos usamos el MAPE, este nos aporta una perspectiva adicional.

El SARIMAX tiene un error del 8.67%, el GRU logra reducir el porcentaje al 8.3%. Estos datos nos confirman que el modelo GRU es más eficaz en la captura de la variabilidad intrínseca de la serie, este proporciona predicciones que se ajustan de manera más precisa a las fluctuaciones reales.

Por último, el coeficiente de determinación R^2 nos ofrece una visión global del ajuste del modelo, mostrando que el GRU es un 9% mejor que el SARIMAX. Este indicador no solo muestra la capacidad del GRU para explicar la variabilidad de la serie, sino también su efectividad para representar la estructura subyacente de los datos.

En conjunto, estos resultados permiten concluir que el modelo GRU logra reducir el error absoluto y relativo en las predicciones, también se ajusta de forma más robusta a la dinámica de la serie temporal. Antes de dar una conclusión a estos datos veámoslo de forma gráfica.

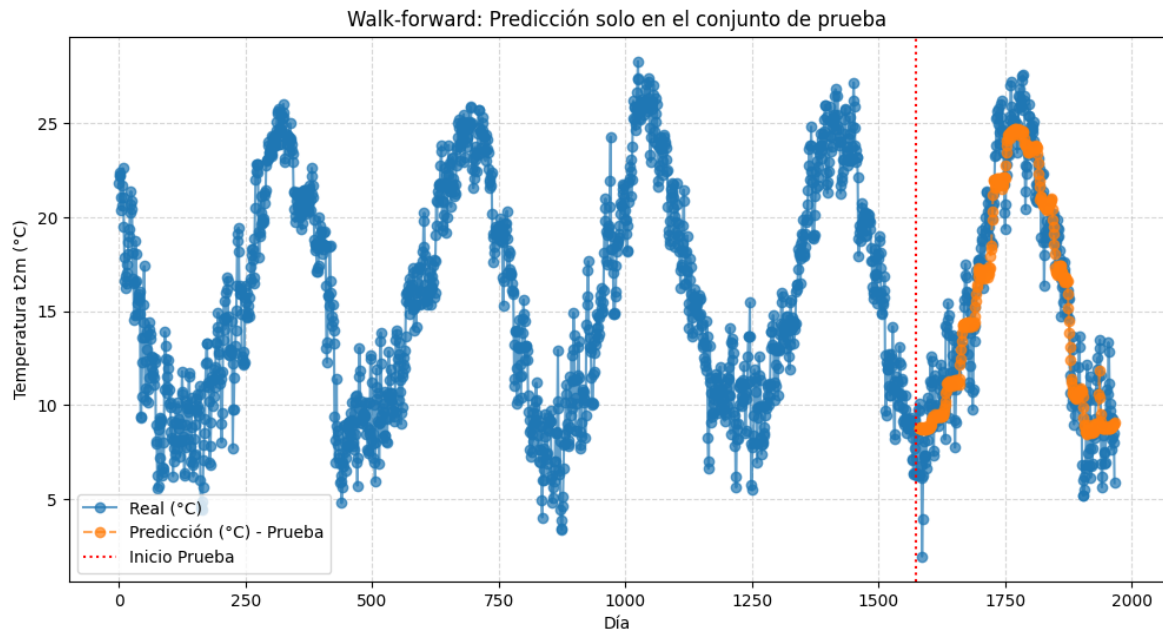


Figura 20 Predicción GRU en test

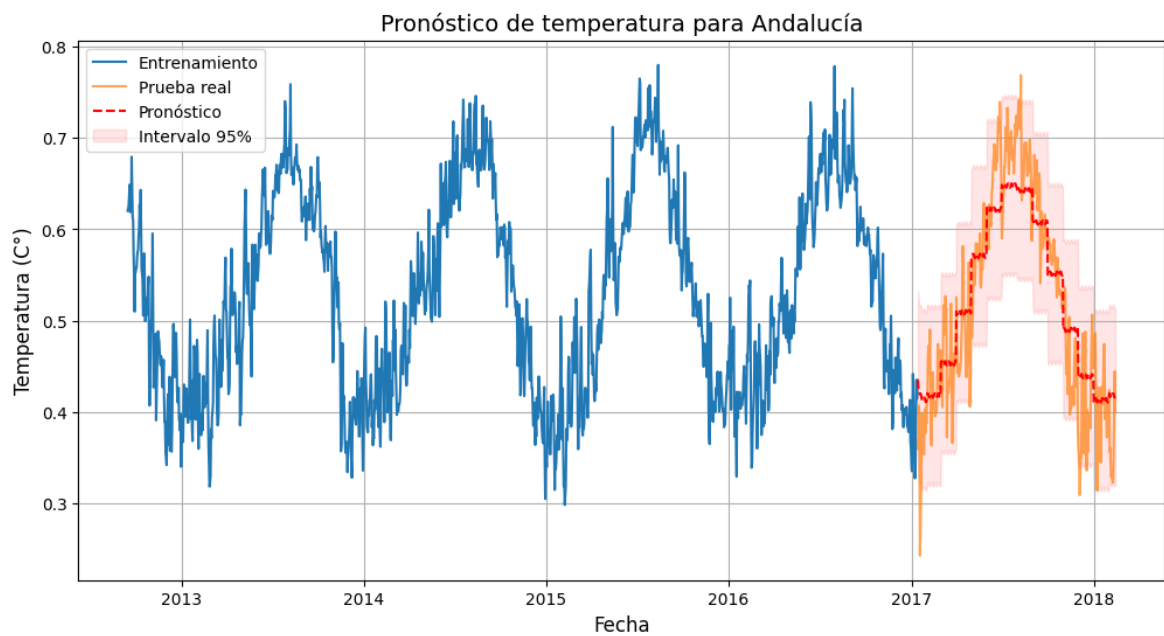


Figura 21 Predicción SARIMAX en test

La elección entre el GRU y el SARIMAX finalmente se decanta hacia el GRU, no tanto por las diferencias en las métricas ya que estas en algunos casos son realmente pequeñas, sino que también el GRU tiene una escalabilidad que el SARIMAX no posee, por lo que para la predicción de temperatura se usa el GRU.

El segundo caso de estudio es el del CAPE(Convective Available Potential Energy), en este caso se usa solo un modelo, un XGBoost con la función de pérdida Tweedie. La elección de esta metodología se fundamenta en la naturaleza de los datos del CAPE , como ya se ha comentado anteriormente, estos datos están caracterizados por una distribución sesgada y una alta proporción de valores cercanos a cero o siendo cero y una variabilidad significativa.

XGBoost es un algoritmo basado en árboles de decisión que optimiza el enfoque de boosting mediante la minimización del error en cada iteración. Lo que lo diferencia de otros métodos de boosting es la introducción de mejoras en eficiencia computacional, regularización y manejo de datos escasos, esto hace que sea especialmente adecuado para datos con alta variabilidad y distribuciones sesgadas, como la del CAPE.

El algoritmo se fundamenta en la construcción secuencial de árboles de decisión, donde cada nuevo árbol corrige el error residual en función de la magnitud del error, dando mayor importancia a aquellas muestras con mayor desviación en la predicción.

El uso de Tweedie es fundamental, ya que la función pérdida de Tweedie está especialmente diseñada para distribuciones sesgadas como la del cape, alta presencia de ceros y varianzas desproporcionadas.

La función perdida de Tweedie en XGBoost se define como : $L(y, \hat{y}) = \frac{y\hat{y}^{-p}}{1-p} - \frac{\hat{y}^{2-p}}{2-p}$ donde p es el parámetro de varianza Tweedie, que en mi caso es de 1.5, este se usa para equilibrar entre una distribución Gamma y una de Poisson.

La idea para este modelo es un modelo que consiga captar las tendencias y que sea eficiente energéticamente, se llegó a probar un GRU, pero por la naturaleza de los datos no conseguí hacer que no se desvaneciese el gradiente, llegue a probar otros modelos más complejos como Transformers pero ninguno conseguía captar la tendencia, todos dibujaban líneas rectas cercanas a 0, y es lógico ya que lo raro es que los valores no estén cercanos a 0.

Fue entonces cuando encontré la función pérdida de Tweedie y decidí probarla, y su desempeño es bastante aceptable para la complejidad de los datos, pero antes de hablar del desempeño del modelo veamos su arquitectura.

Como ya he comentado he usado XGBoost, un algoritmo de boosting basado en árboles de decisión, en mi caso la estrategia usada ha sido ‘lossguide’ para que el árbol crezca buscando la minimización de la pérdida, también he usado histogramas para calcular la mejor división de los nodos posibles con ‘gpu_hist’ y para evitar sobre ajuste he usado regularización L1 y L2.

En cuanto a los hiperparámetros el learning rate es de 0.6, una tasa de aprendizaje alta, la longitud máxima de los árboles es de 15 para poder capturar interacciones no lineales complejas, el máximo número de hojas es de 1028 habiendo probado otros valores este ha sido el mejor sin duda. En el número de boost lo que he buscado es una alta convergencia, ya que el CAPE al tener variaciones tan grandes el modelo necesita un número alto de boost y finalmente añadí un early stopping de 500 sobre el error de validación, previniendo así la posibilidad de sobreajuste.

Como en la temperatura los datos se dividen en 7 de entrada y 5 de salida, con la misma agrupación, se mantiene la división train/test del 80/20 y las iteraciones se evalúan con el RMSE en cada boosting.

Para la evaluación del modelo he elegido de nuevo RMSE, MAE Y R^2 , cabe resaltar que estas métricas van a ser mucho peores que las vistas del modelo GRU, por ello que desde un principio la idea no es comparar los modelos de T2M con el modelo de CAPE.

Tabla 8 Métricas XGBoost

	XGBoost
RMSE	0.0322
MAE	0.0122
R^2	0.6859

Fuente: elaboración propia.

Vamos a interpretar las métricas y después veremos algunas gráficas de las predicciones del modelo. En general los errores son muy bajos, esto no significa que el modelo sea muy bueno sino que aunque el modelo dibujase una línea recta en el 0.1 tendría el mismo RMSE, ya que la gran mayoría de los datos se encuentran muy cerca del 0, siento que las métricas en este caso no son de gran ayuda porque el error que nos arrojan no es

significativo a nuestro problema, y lo mismo pasa con R^2 , el modelo explica un 69% de los datos pero ese 69% es muy posible que sean los datos muy cercanos a 0 o menores a 1000, estas predicciones no nos importan porque no aportan valor a nuestro problema, lo que realmente importa es que capture las tendencias, y no hay mejor forma de ver si consigue capturar tendencias que con unas gráficas.

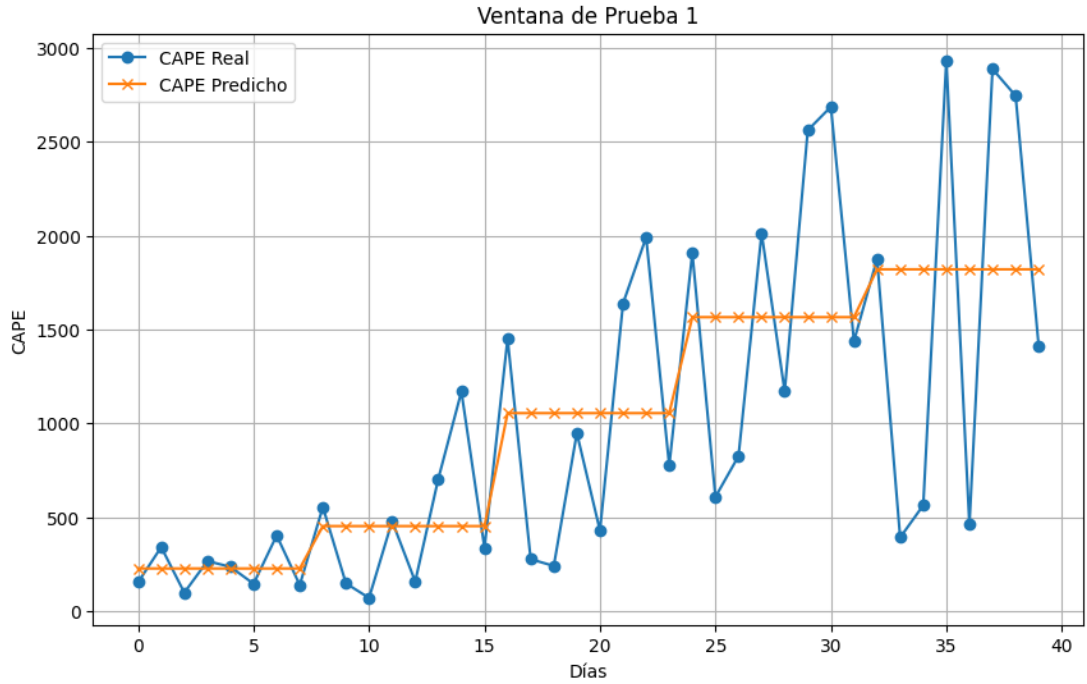


Figura 22 Predicción n°1 CAPE

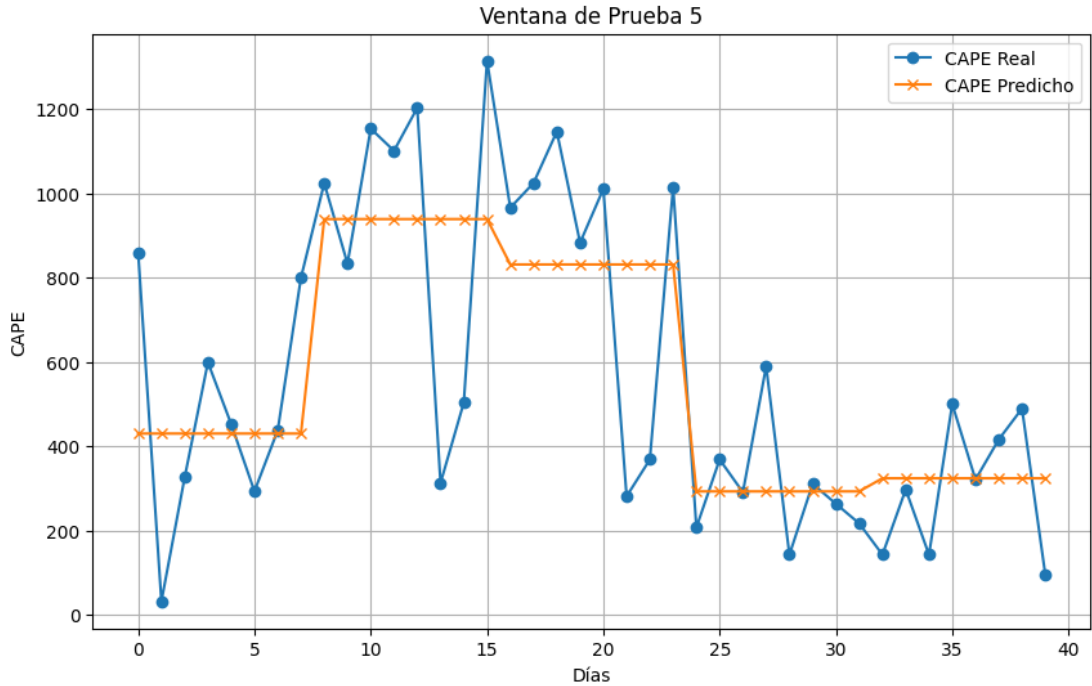


Figura 23 Predicción n°2 CAPE

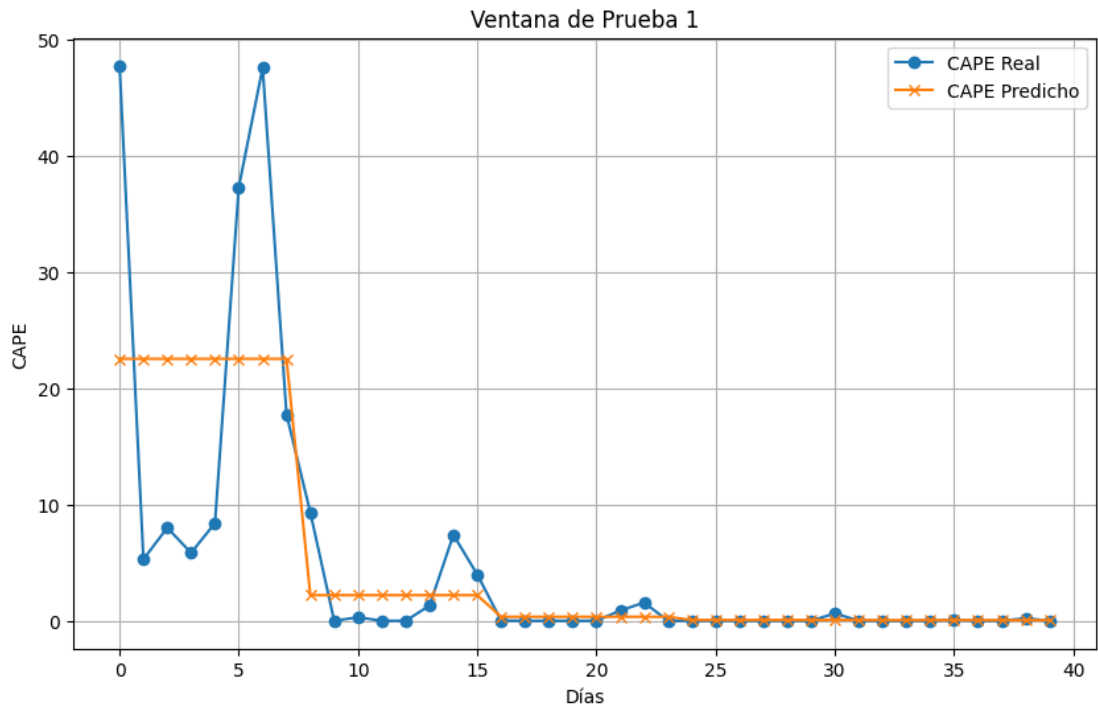


Figura 24 Predicción n°3 CAPE

En estos cuatro gráficos podemos ver que capta la tendencia, con un gran error, pero la consigue captar. En valores del cape en el caso de que el valor real sea 2500 y el predicho 2200, ese error de 300 no se considera relevante porque no implica nada.

El último gráfico es un claro ejemplo de la predicción que no nos interesa, el valor máximo es 50 y el mínimo 0, por eso el RMSE no es nefasto, por estas predicciones en las que dibujando una línea recta en el 0 predice perfectamente, pero fuera de eso el modelo se comporta como debería, entiende las tendencias en los valores importantes del CAPE.

Los resultados obtenidos han sido satisfactorios, mostrando que el modelo GRU ha sido una buena opción para la predicción de temperatura, aunque un LSTM con mayor cantidad de datos podría haber sido superior.

Por otro lado, la predicción del CAPE ha presentado mayores dificultades debido a la complejidad de los datos, aunque el modelo XGBoost con función Tweedie ha logrado captar tendencias relevantes. A pesar de las limitaciones, las predicciones obtenidas pueden ser útiles, y existe margen de mejora mediante el uso de modelos más avanzados y un mayor volumen de datos.

Análisis del Negocio

Inicialmente, se plantearon los siguientes objetivos y preguntas de investigación:

Primeramente, lo que se buscaba era desarrollar la capacidad de predecir temperaturas extremas y niveles elevados de CAPE(Convective Available Potential Energy). El propósito era proporcionar información valiosa para la toma de decisiones tanto de organismos gubernamentales como de las personas de a pie. Estas predicciones son importantes, especialmente considerando la vulnerabilidad de personas de edad avanzada ante el calor extremo y los riesgos que conllevan valores altos de CAPE, los cuales pueden indicar fuertes rachas de viento y precipitaciones repentinas.

Otra de las preguntas estaba enfocada en identificar cuanta mejora hay entre una resolución del problema con modelos más clásicos (ARIMA,SARIMAX) o con modelos más modernos (GRU,LSTM).

Por otro lado, una cuestión esencial que motivó esta investigación fue comprender cómo han variado la temperatura y el CAPE a lo largo del tiempo. Lo que se pretendía esclarecer es si existía una dirección clara, ya fuera un incremento o un descenso, en los valores de estas dos variables meteorológicas durante el lapso temporal considerado.

A continuación, se procederá a analizar los resultados obtenidos para determinar si se han alcanzado estos objetivos iniciales y si se han podido responder satisfactoriamente las preguntas planteadas.

Respondamos a la primera pregunta, ¿He sido capaz de hacer modelos útiles en la predicción de temperatura y CAPE? en cuanto a la temperatura se ha conseguido un modelo que predice de manera correcta, no solo el día de mañana sino los cinco siguientes días. En mi caso, creo que es importante que no solo haga predicciones del día siguiente sino de los próximos 5 días, ya que la temperatura entre otras cosas es un factor determinante en nuestro día a día. Por lo que creo que la respuesta sobre los modelos de temperatura es que si, si se ha conseguido tener un modelo bueno teniendo en cuenta las limitaciones de hardware tanto como para la recolección de datos como para el procesamiento y el entrenamiento de modelos.

La predicción del CAPE presenta desafíos específicos debido a su distribución altamente sesgada. En la mayoría de los días, esta variable adopta valores cercanos a cero, y solo en contadas ocasiones se registran picos elevados. Esta característica dificulta notablemente el ajuste de modelos precisos. Por ello, desde el inicio del proyecto, el objetivo principal no ha sido alcanzar una predicción exacta en términos numéricos, sino identificar correctamente las tendencias generales.

En contextos meteorológicos, un valor de CAPE superior a 2000 ya representa una señal de potencial convectividad², y por tanto lo relevante es que el modelo detecte este cambio de régimen. Más allá del margen de error que pueda haber en torno a ese umbral, lo que realmente importa es si el modelo anticipa o no la posibilidad de un evento significativo. Por ejemplo, una predicción que sitúe el CAPE en torno a 3000 constituye una advertencia clara de condiciones propicias para fenómenos severos, y en ese sentido, se ha logrado construir un modelo que responde adecuadamente a esas situaciones.

En cuanto a la comparación entre modelos, se ha observado una diferencia notable entre los enfoques clásicos utilizados. ARIMA, al no incorporar componentes estacionales, muestra grandes limitaciones a la hora de capturar patrones meteorológicos, que suelen estar marcados por una fuerte estacionalidad.

Este comportamiento era previsible, ya que el modelo tiende a sobreajustar secuencias o, por el contrario, a no aprender nada útil cuando la estacionalidad es dominante. En cambio, SARIMAX, al incorporar tanto la componente estacional como variables exógenas, ha ofrecido resultados considerablemente más sólidos, comparables incluso a los obtenidos con modelos de redes neuronales como GRU y LSTM.

Esto ha sido especialmente llamativo, dado que estos últimos cuentan con arquitecturas más sofisticadas y, en principio, una mayor capacidad de modelar relaciones complejas. No obstante, para que SARIMAX alcance ese nivel, ha sido necesario dedicar un esfuerzo considerable al ajuste de sus hiperparámetros y a la selección cuidadosa de su estructura.

En definitiva, no se ha encontrado una diferencia decisiva entre los modelos clásicos y los más modernos, aunque es importante señalar que las redes neuronales recurrentes poseen una mayor escalabilidad: permiten manejar volúmenes mayores de datos y adaptar su arquitectura mediante capas más profundas o anchas, tamaños de *batch* elevados, y otros

ajustes. Estas mejoras se intentaron durante el desarrollo del proyecto, pero las limitaciones de hardware impidieron explorar configuraciones más complejas.

Esto deja la puerta abierta en un futuro a un escalado de los modelos tanto en arquitectura como en la base de datos, y en ese caso seguramente haya mayor diferencia.

Como he comentado al principio también quería ver cuál es la tendencia de las dos variables objetivos, para saber si en los años estudiados ha habido un incremento o no, y cuál es su tendencia.

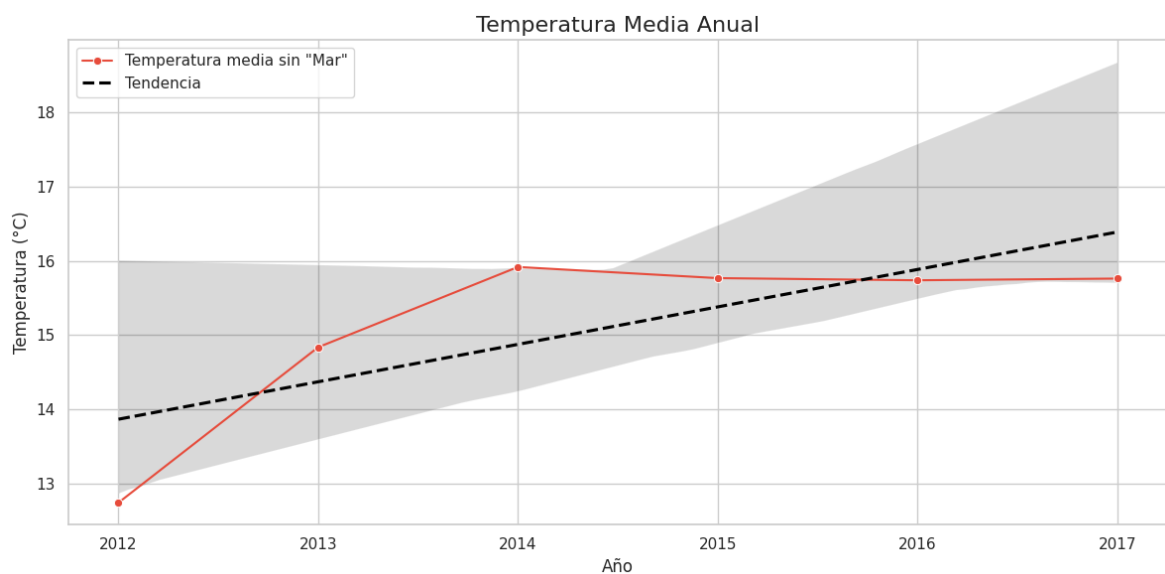


Figura 25 Temperatura media anual sin el MAR

Para la temperatura podemos apreciar que en cuanto la media de temperatura su tendencia es claramente alcista, tan solo en cinco años en la cuenca mediterránea la media de temperatura sube desde los 13 grados a prácticamente los 16, esto se puede deber a que justo esos años han sido excepcionalmente calurosos pero la realidad es que la tendencia parece que claramente sube, veamos las máximas de estos años:

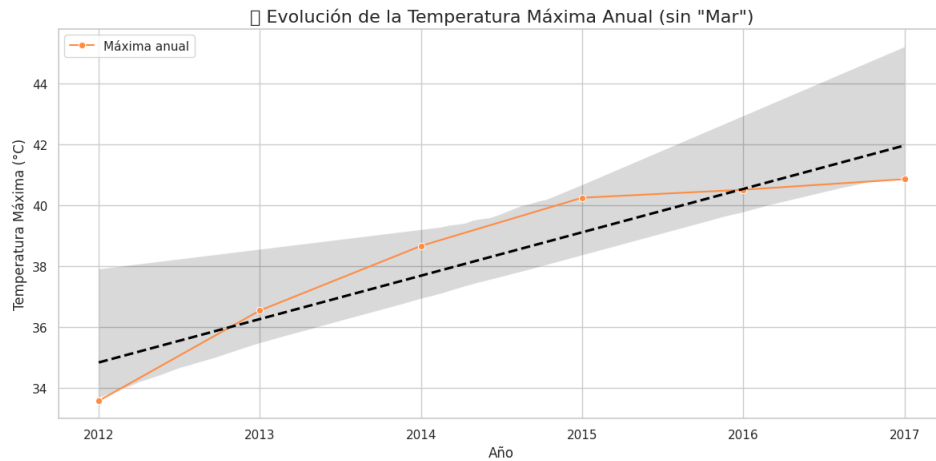


Figura 26 Temperatura Máxima anual

En las temperaturas máximas la lectura es la misma, tendencia más alcista aun y el mismo aumento en las máximas que en el promedio, 3°C.

Esto es importante, porque como ya he comentado en otros momentos en este trabajo, en la meteorología todas las variables están correlacionadas, hay algunas de ellas que son muy sutiles pero aun así todas ellas generan el mundo en el que vivimos, en este caso y llevado a mi estudio el aumento de la temperatura es muy importante, no solo porque implica un riesgo como tal para algunas personas sino porque afecta también al CAPE, veamos primero como ha evolucionado en estos años la temperatura del mar:

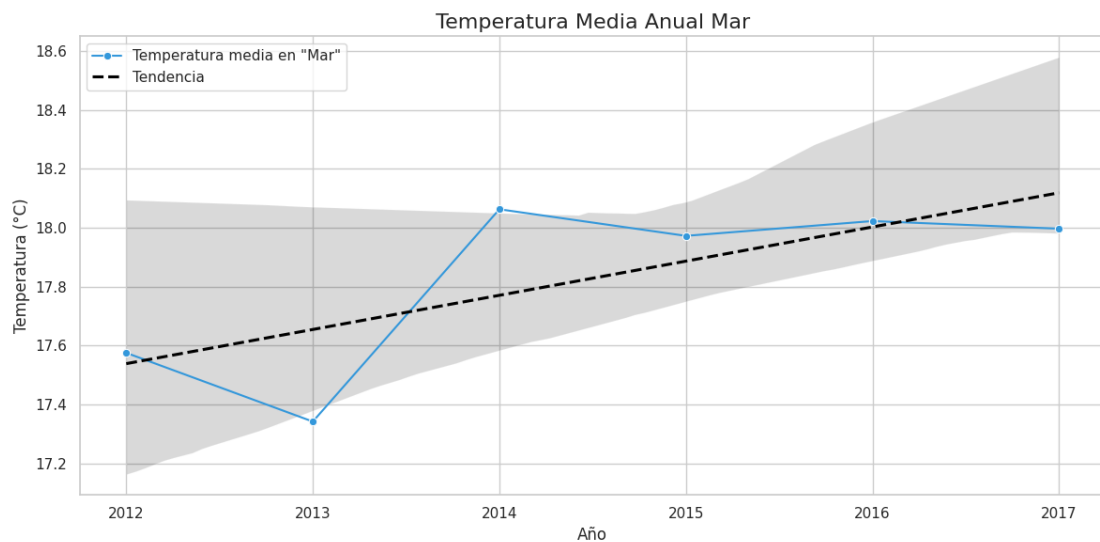


Figura 27 Temperatura media anual del mar

Como era de esperar la tendencia es la misma, es verdad que el aumento de temperatura es menor pero, en el mar, cualquier aumento de temperatura puede generar múltiples problemas, volviendo al CAPE, el aumento en la temperatura del mar está directamente relacionado con la humedad y esta es una variable crítica en el CAPE, cuanto mayor sea la humedad mayores probabilidades de cambios de presión, esto aumenta el valor del cape y genera una ventana de oportunidad a fenómenos como huracanes, danas o fuertes tormentas.

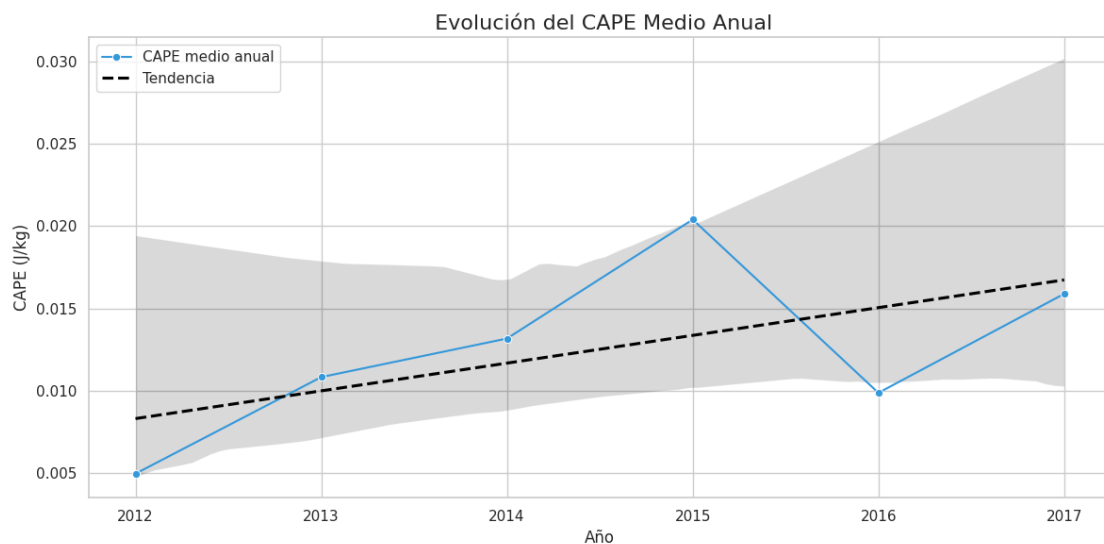


Figura 28 Evolución del CAPE 2012-2017

Este aumento del valor del cape y de la temperatura dan más valor a mis modelos, ya que cada vez tendremos máximas más altas y mayor cadencia de tormentas y fenómenos destructivos, tenemos la necesidad de organizarnos y encontrar la manera de paliar los efectos de estos fenómenos y el primer paso para ello es la anticipación.

Viendo la problemática del aumento de la temperatura y del CAPE, el siguiente paso es conseguir que las predicciones sean accesibles.

Una línea clara de mejora futura consiste en ampliar la base de datos utilizada, tanto en términos temporales como en la inclusión de nuevas variables que puedan aportar mayor riqueza informativa al conjunto. Con una base más extensa y representativa, sería posible abordar un rediseño de los modelos de redes neuronales (GRU o LSTM), incorporando mayor profundidad o complejidad estructural con el fin de obtener predicciones más precisas, especialmente en el caso de la temperatura. En lo que respecta al CAPE, XGBoost ha demostrado ser una opción eficaz, aunque su rendimiento podría beneficiarse notablemente de un mayor volumen de datos. Dado que los valores elevados de CAPE son relativamente poco frecuentes a lo largo del año, ampliar el número de observaciones —

mediante series temporales más largas— permitiría al modelo disponer de más ejemplos representativos, mejorando así su capacidad para anticipar situaciones atmosféricas potencialmente peligrosas.

Para que estos modelos sean útiles habría que crear una API abierta, así cualquier persona podría usarlos en cualquier contexto tanto como para sus proyectos o incluso para mejorarlos, y de esta API se alimentaría una página web en la que se avisa de los momentos con altas temperaturas y posibles fenómenos extremos.

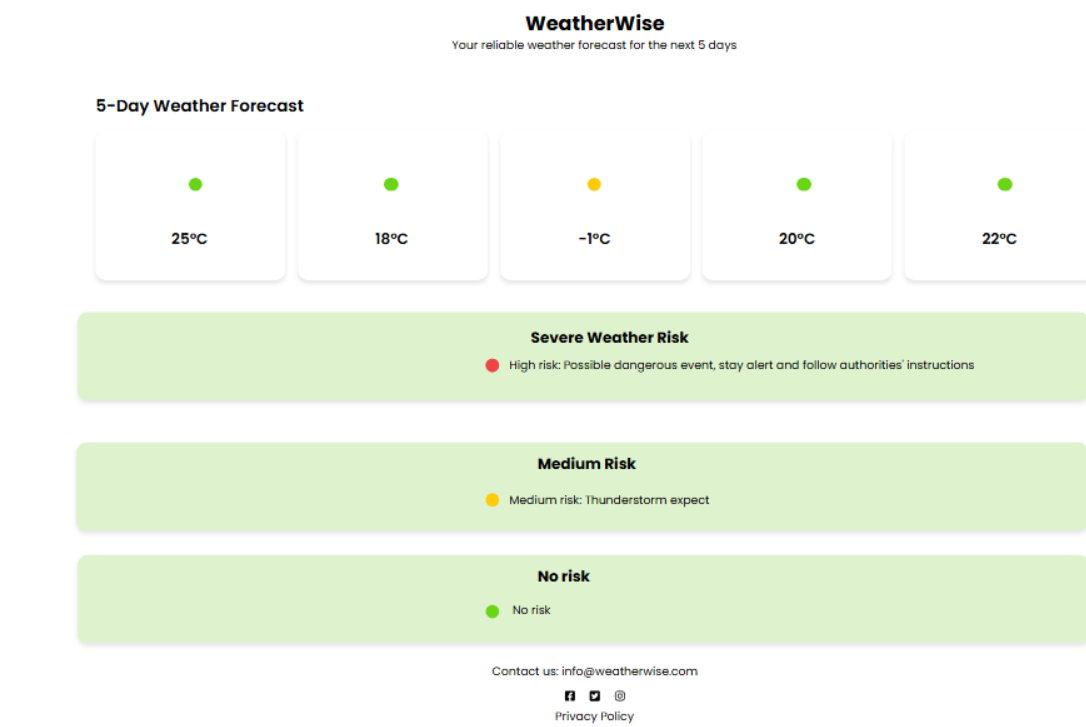


Figura 29 Ejemplo Web

Esta web serviría como plataforma visual e informativa, clasificando los valores predichos del CAPE en tres niveles de riesgo: Verde (0–1250), Amarillo (1250–2500) y Rojo (2500–6000), ofreciendo así una representación clara e intuitiva del riesgo meteorológico a corto plazo.

En definitiva, este trabajo demuestra que es posible desarrollar modelos predictivos útiles que, con los datos adecuados y una implementación accesible, pueden ayudar a mejorar nuestra capacidad de respuesta frente a fenómenos extremos. No resuelve el problema, pero es un paso en la dirección correcta.

4. REFERENCIAS BIBLIOGRÁFICAS

- al, H. e. (2006). Principles of meteorology and numerical weather prediction.
- Box, G. &. (1970). Time series analysis: Forecasting and control. Holden–Day.
- Box, G. &. (1976). Time series analysis: Forecasting and control (Revised ed.). Holden–Day.
- Brooks, M. (2007). Climate myths: Chaotic systems are not predictable.
- Chen, T. &. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, (págs. 785–794).
- Cho, K. v. (2014). Learning phrase representations using RNN encoder–decoder for statistical machine translation. 1724–1734.
- Dickey, D. A. (1979). Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association*, 427–431.
- Dunn, P. &. (2005). Series evaluation of Tweedie exponential dispersion model densities. *Statistics and Computing*. 267–280.
- Fisher, R. A. (1925). Statistical Methods for Research Workers. Edinburgh: Oliver & Boyd.
- Hersbach, H. B.-N. (2019). Global Reanalysis: Goodbye ERA-Interim, Hello ERA5. ECMWF Newsletter. 159.
- Hochreiter, S. &. (1997). Long short-term memory. *Neural Computation*. 1735-1780.
- Loshchilov, I. &. (2017). Decoupled weight decay regularization.
- Mestre, A. &. (2009). El seguimiento de la sequía en estaciones y grandes cuencas usando el índice SPI. *Tiempo y Clima*,.
- Pedro Mateus, J. C. (2020). An ERA5-Based Hourly Global Pressure and Temperature (HGPT) Model. *Remote Sensing*.
- Zhang, Q. W. (2017). Prediction of sea surface temperature using long short-term memory. *Environmental Science and Pollution Research*. 24(14), 12206-12216.