PROYECTO METEOROLÓGICO

1. INTRODUCCIÓN

El clima influye en casi todo lo que hacemos por lo que predecirlo con precisión se ha vuelto esencial. Por eso, se ha creado un simulador de predicción meteorológica que utiliza técnicas de Big Data para procesar y analizar grandes cantidades de información. Este sistema no solo almacena y gestiona datos de manera eficiente, sino que también genera pronósticos a corto plazo con un alto nivel de precisión, algo clave para tomar decisiones informadas en sectores como la energía, el transporte o la gestión de emergencias.

Para hacerlo posible, se han recopilado datos de distintos municipios de España, cubriendo desde las zonas costeras hasta los municipios del norte y del centro peninsular. Esto permite entender cómo varía el clima en diferentes zonas del país. Para manejar toda esta información, se han usado tecnologías como PySpark, que facilita el procesamiento de grandes volúmenes de datos de manera rápida y eficiente.

2. BUSINESS UNDERSTANDING

2.1 OBJETIVOS

A. OBJETIVO GENERAL

Predecir datos meteorológicos mediante varios modelos analíticos basados en datos obtenidos a través de apis de aemet, open meteo y weatherapi, aprovechando técnicas de análisis de datos y aprendizaje automático.
Habrá dos tipos de datos: diarios y horarios.

- Diarios: información meteorológica de cada día del que se harán predicciones para el día siguiente y para los siguientes 7 días.

- Horarios: información meteorológica de cada hora de cada día del que se harán predicciones para la hora siguiente y las 24 horas posteriores.

B. OBJETIVOS ESPECÍFICOS

- Extraer información relevante de las plataformas seleccionadas, mediante el uso de apis.
- Estandarizar y depurar los datos recopilados para asegurar su calidad y consistencia, preparándolos para el análisis posterior.
- Consolidar la información de las distintas fuentes en una serie de archivos con misma estructura de datos.
- Diseñar y entrenar varios modelos analíticos o de aprendizaje automático capaz de predecir los distintos datos meteorológicos con alta precisión, validar y optimizar los modelos desarrollados mediante técnicas de evaluación, asegurando su fiabilidad en escenarios reales.

3. DATA UNDERSTANDING

Proporciona la base para tomar decisiones informadas sobre cómo procesar, transformar y analizar los datos en las etapas posteriores del proyecto. Una comprensión sólida de los datos asegura un análisis más efectivo y resultados de mayor calidad.

En este proyecto se han recolectado datos de dos páginas web (aemet y open meteo y weatherapi) de distintos municipios de España (en el caso de los datos diarios de Alcalá de Henares, Marbella y Asturias y por la parte de datos horarios de Madrid y Marbella) con el objetivo de obtener información sobre los diferentes datos meteorológicos, tanto diarios como horarios para su posterior predicción.

Para el análisis de datos diarios, se recopilaron registros correspondientes a un período de cinco años, mientras que, para los datos horarios, se decidió utilizar información de un lapso de dos años.

En nuestro caso, hemos recopilado información mediante la utilización de apis que proporcionan dichas webs. En el caso específico de los datos diarios de Alcalá de Henares, se han ido obteniendo los datos de seis en seis meses ya que no se podían obtener más datos de seguido. Cada fichero generado se ha pasado a formato JSON para mejor visualización de la estructura de datos y cuando ya se obtuvieron todos los datos necesarios se optó por juntar todos los ficheros en un solo CSV. En los demás casos simplemente se llamó a una API introduciendo los datos que se querían obtener y el rango de fechas y se pasó a un CSV.

Este proceso de recolección y transformación de datos asegura una base sólida para el análisis posterior. La utilización de APIs permite acceder a información precisa y actualizada, mientras que la conversión de los datos a formatos como JSON y CSV facilita su manejo y análisis. Además, la recopilación de datos de diferentes municipios y periodos de tiempo ofrece una visión más amplia, lo que es esencial para obtener predicciones meteorológicas más precisas y adaptadas a diferentes contextos. En cuanto a Marbella, se ha utilizado la api de weatherAPI, que tan solo dispone datos históricos por horas para un periodo de un año, por lo que se ha recopilado datos a lo largo del tiempo previo al proyecto y concatenándolos con el motivo de obtener el máximo número de datos posibles.

Entre los datos obtenidos, en el caso de los datos diarios se incluyen la fecha, temperatura media, temperatura máxima, temperatura mínima, humedad relativa media, humedad relativa máxima y mínima, velocidad del viento, hora de la temperatura máxima y mínima, racha de viento, hora de la racha y hora de la humedad relativa máxima y mínima, la presión mínima y máxima, la hora media y la sensación térmica.

Por otro lado, en el caso de la información para datos horarios se han obtenido temperatura, temperatura del suelo, temperatura ambiente, humedad relativa, precipitación, probabilidad de precipitación y velocidad del viento.

Aunque toda la información obtenida es útil, será necesario prescindir de algunos detalles, ya que o no hay suficientes datos en dichas columnas o son innecesarias.

4. DATA PREPARATION

En cuanto a la transformación de los datos, se han realizado utilizando el framework de **pyspark** que permite procesar grandes cantidades de datos de manera distribuida, aunque en nuestro caso no tenemos un volumen de datos excesivamente grande.

Para el caso de los datos diarios se han realizado las siguientes operaciones:

Datos Alcalá de Henares: antes de realizar ninguna transformación, se eligieron una serie de variables que podían ser elegidas para ser predichas en las que se encuentran la temperatura media, temperatura máxima, temperatura mínima, humedad relativa media y velocidad del viento además de la fecha. Ya elegidas dichas variables, se optó por eliminar las demás y de transformar las elegidas al formato pertinente como a formato date o a números decimales.

Acto seguido para comprobar que no faltaba ninguna fecha en el conjunto de datos se creó una lista de fechas del rango de 5 años y restarla con el rango de fechas del dataframe original y se observó que faltaban fechas, las cuales se introdujeron en el original. Después de ver si había duplicados, para poder rellenar nulos se optó por coger el valor anterior no nulo y ponérselo al actual.

Datos Marbella: para la limpieza de datos, antes de nada, se ha comprobado si existen datos en todas las fechas, e insertar con valores nulos aquellos datos para las fechas que no se hayan obtenido de la llamada a api. Se crea un Profiling Report sobre los datos para una mayor compresión y poder trabajar para la limpieza de datos. Gracias al PR encontramos valores nulos para limpiar y las relaciones entre variables, de esta manera podemos filtrar por las columnas que pueden aportar para la predicción de las variables. Luego ajustamos los formatos para las columnas de

datos numéricos o de tipo fecha, imputamos datos para los valores nulos basándonos en la media del dato anterior y el dato siguiente que tengan valor distinto de nulo y eliminamos las filas duplicadas.

Para los datos por horas, se sigue el mismo procedimiento para obtener datos de todos los días dentro del rango de fechas sin valores nulos.

Datos Asturias: antes de realizar las transformaciones necesarias, se eligieron una serie de variables para ser predichas en las que se encuentran la temperatura media, temperatura máxima, temperatura mínima, hora media y la fecha. Ya elegidas, eliminamos las demás y se transformaron las elegidas al formato date o a números decimales.

A continuación, para asegurar que no faltara ninguna fecha en el conjunto de datos, se generó una lista de fechas correspondiente al periodo de 5 años. Al detectar fechas ausentes, estas se añadieron al conjunto de datos. Posteriormente, para manejar los valores nulos, se seleccionaron tanto el valor no nulo anterior como el posterior, y se calculó la media de ambos para reemplazar los datos faltantes.

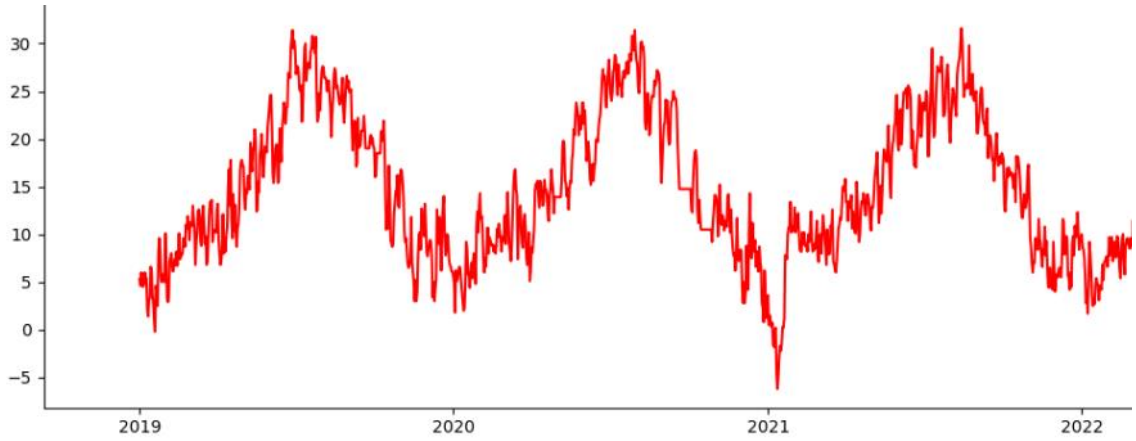
En el caso de los datos horarios se han seguido os distintos pasos:

Datos Madrid: antes de realizar ninguna transformación, se eligieron una serie de variables que podían ser elegidas para ser predichas en las que se encuentran la temperatura, humedad relativa, precipitación y velocidad del viento además de la fecha. Ya elegidas dichas variables, se optó por eliminar las demás y de transformar las elegidas al formato pertinente como a formato date o a número decimales.

Acto seguido para comprobar que no faltaba ninguna fecha en el conjunto de datos se creó una lista de fechas del rango de 2 años y restarla con el rango de fechas del dataframe original y se observó que no faltaban fechas. Después de ver si había duplicados, se observó que no había nulos por lo que no hizo falta hacer ninguna imputación.

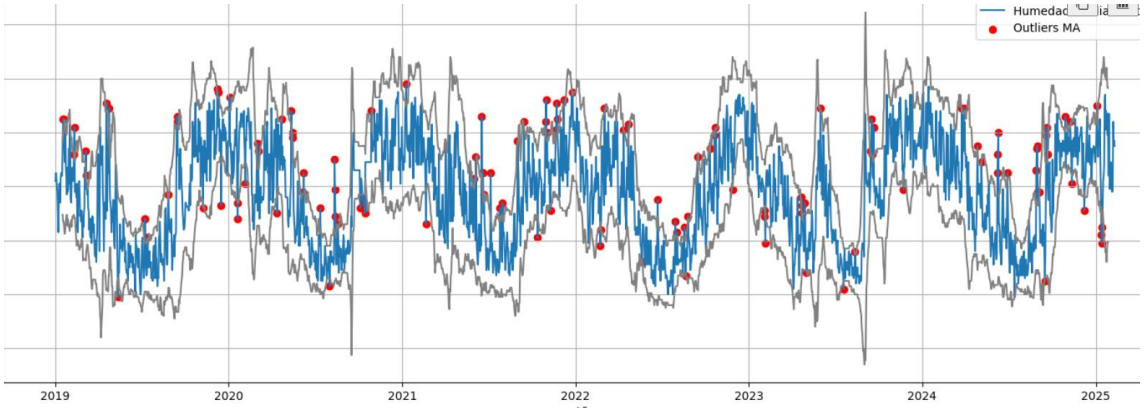
Respecto al estudio del EDA, en cada municipio se han realizado las siguientes observaciones:

Datos diarios Alcalá de Henares: por cada variable se ha hecho una descomposición estacional para ver sus componentes principales como su tendencia, estacionalidad y residuo. Se observa que en enero 2021 hubo un descenso drástico de las temperaturas y fue debido al fenómeno meteorológico de la Filomena como se observa en la siguiente imagen.



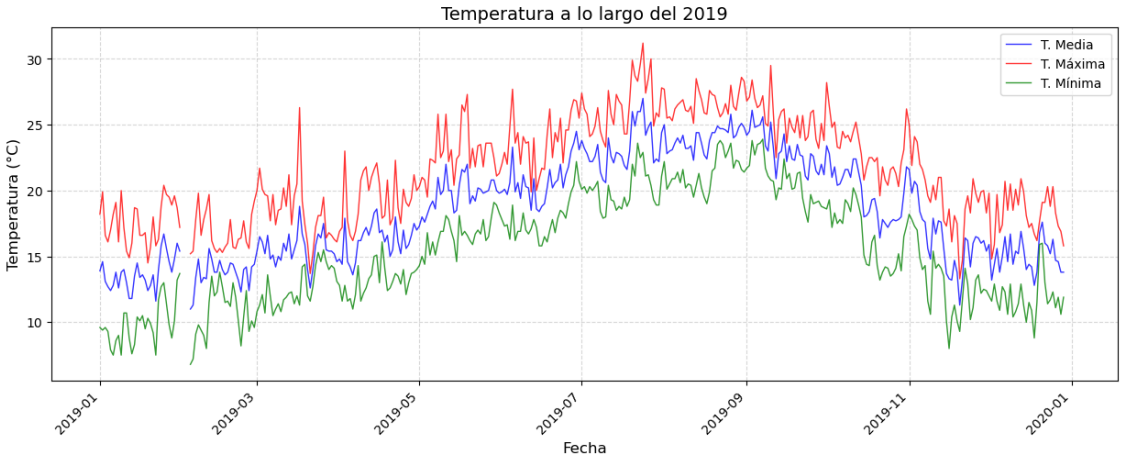
En el caso de los datos de temperatura media, máxima y mínima se observa que hay estacionalidad anual. Para poder observar como varían la media y la varianza, se ha escogido una ventana temporal y se observa en el caso de las variables de las temperaturas no varían, pero en la humedad la varianza sí que varía de forma mínima.

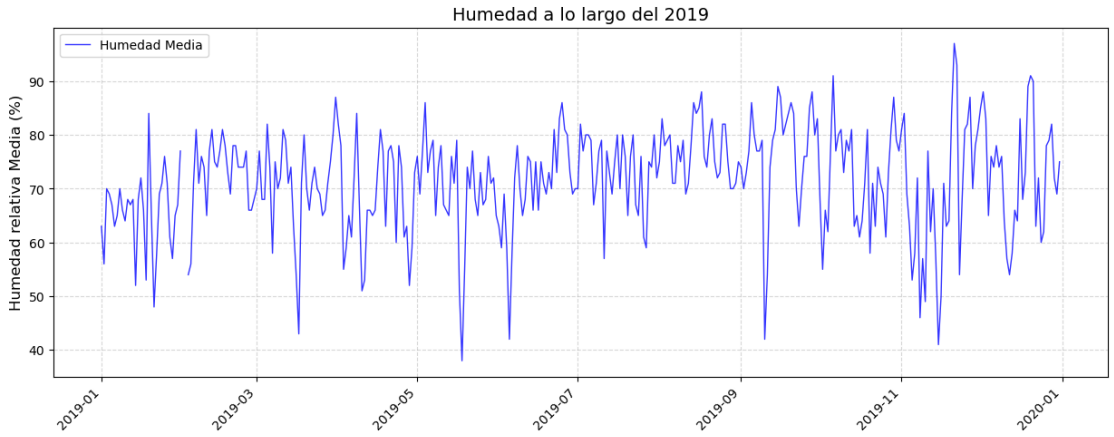
En relación con los outliers, para detectarlos se ha utilizado el método de filtro de hampel, que detecta los outliers dentro de una ventana utilizando la mediana, ya que la mediana no suele ser muy afectada por outliers. Un ejemplo de visualización de outliers de la humedad relativa es el siguiente en el que los puntos rojos son los outliers ya que sobrepasan las líneas grises:



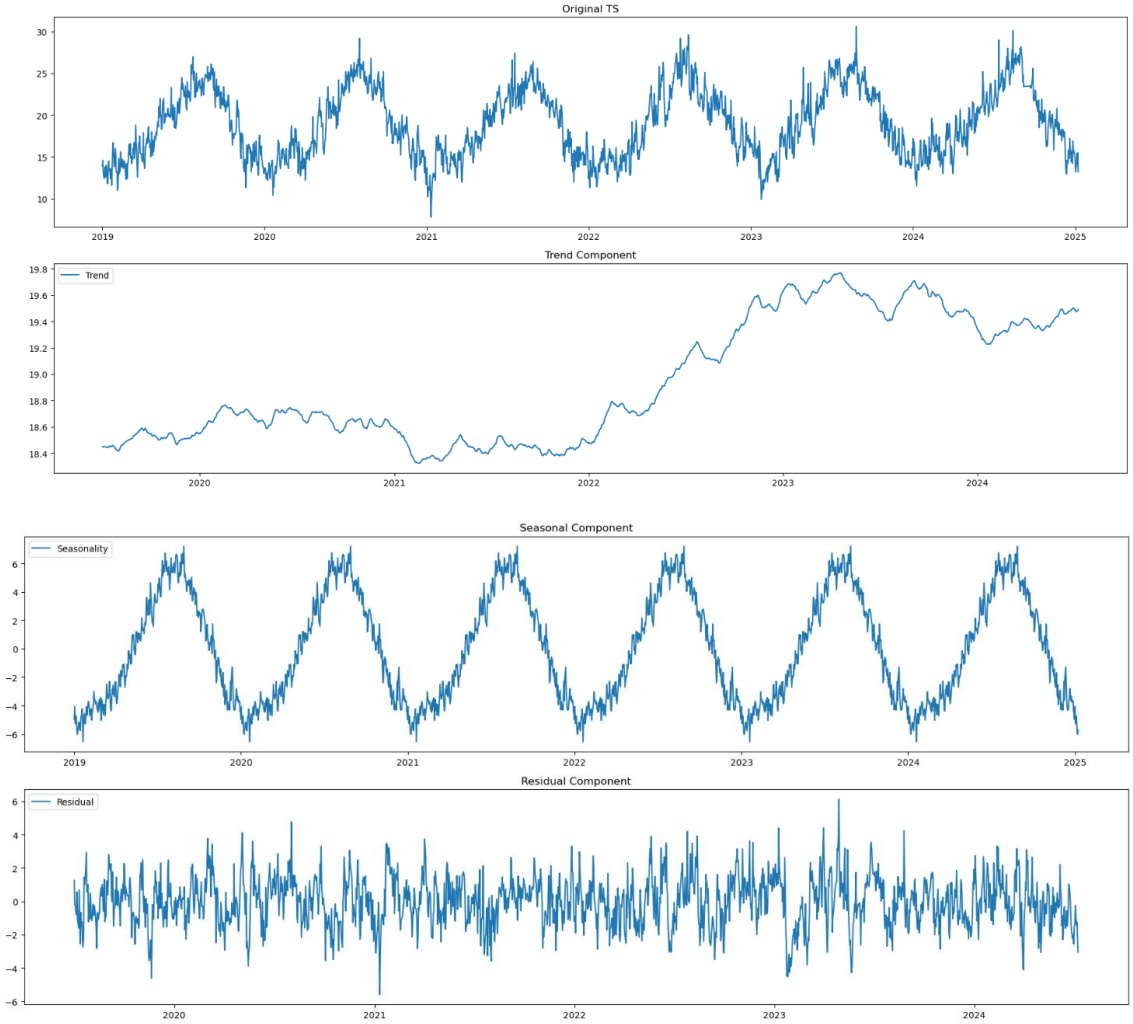
Para eliminar los outliers lo que se hace es pasar dicho valor al del filtro en dicha fecha, es decir, al valor de la linea gris en ese punto.

Datos diarios Marbella: para entender los datos, hemos realizado un análisis de serie temporal, para todas las columnas de variables a predecir, en este caso, se va a predecir la temperatura media (tmed), temperatura máxima (tmax), temperatura mínima (tmin) y el valor de la humedad relativa (hrMedia). Tomando como ejemplo exponemos el AST para la temperatura media, mostramos una gráfica por cada año:

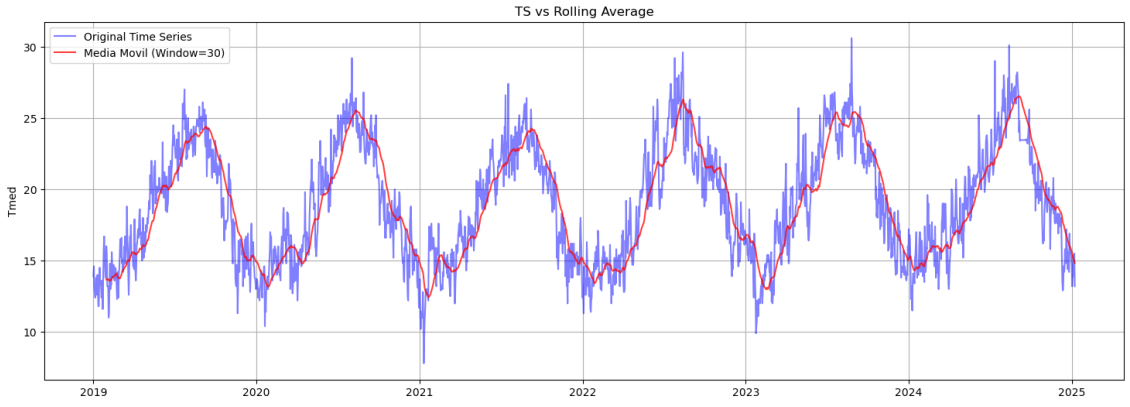




Y la descomposición de la serie temporal:

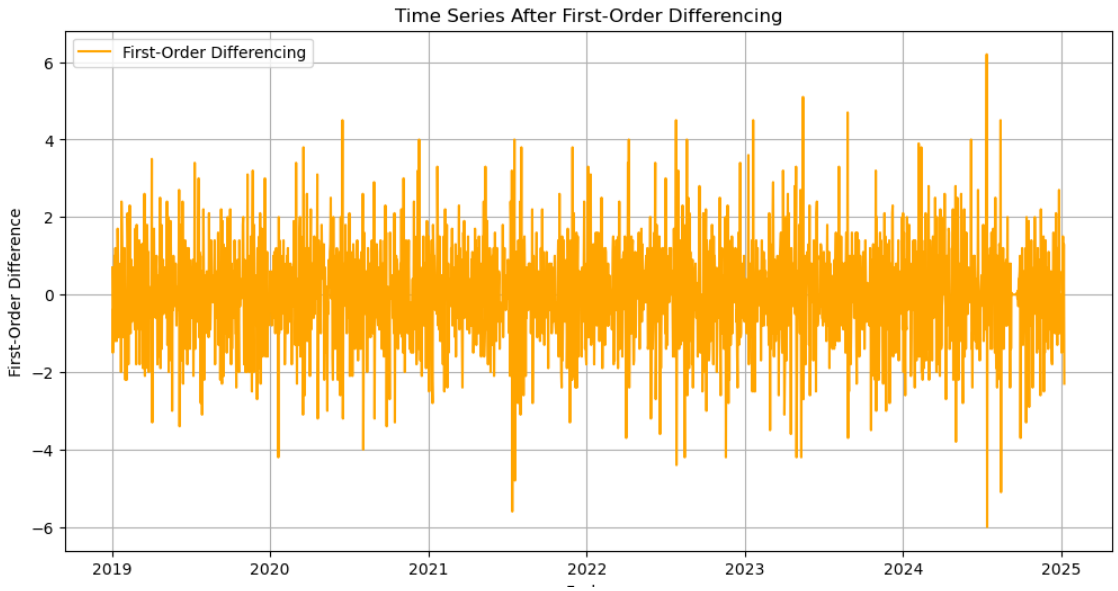


Después revisamos la media movil

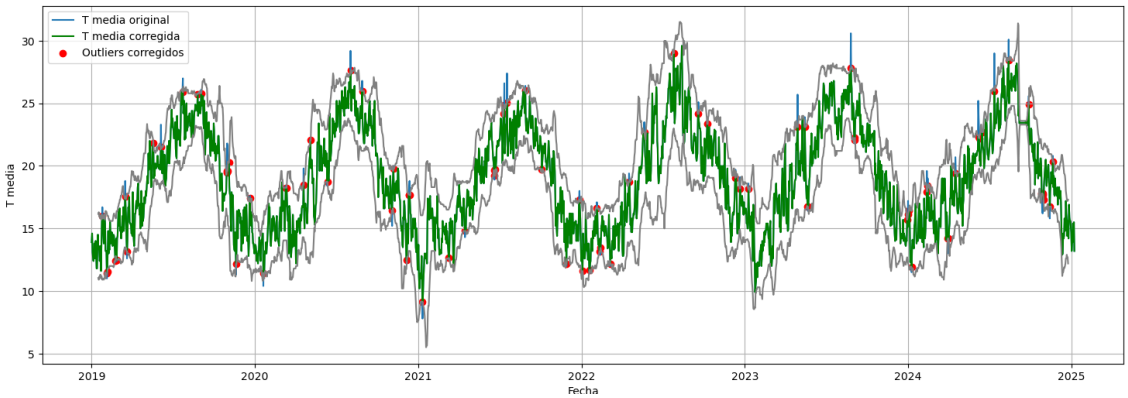


Y el estudio de si la serie es o no estacionaria, para lo que hemos decidido utilizar el test de Dickey-Fuller, el resultado nos muestra que la serie no es estacionaria, y como método para ajustar la serie realizamos el método de diferenciación, que como resultado

finalmente después de un diff si se transforma en una serie estacionaria



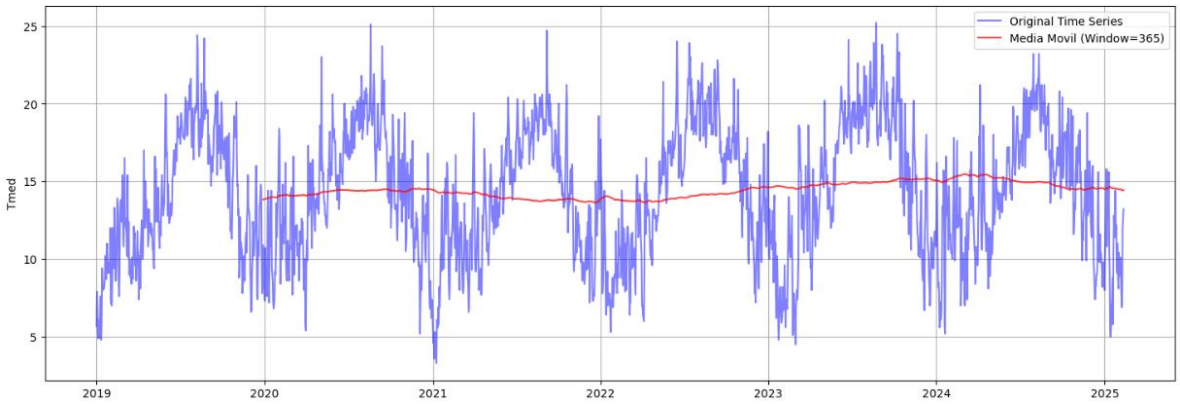
Para la detección de los outliers, he concluido en utilizar el Filtro Hampel, un método muy robusto basado en la variación de sigmas que calcula la mediana de una ventana que incluye la muestra y sus 2 vecinos (superior e inferior). Y sustituir el valor de estos datos con sus vecinos más próximos para los valores atípicos, resultando en:



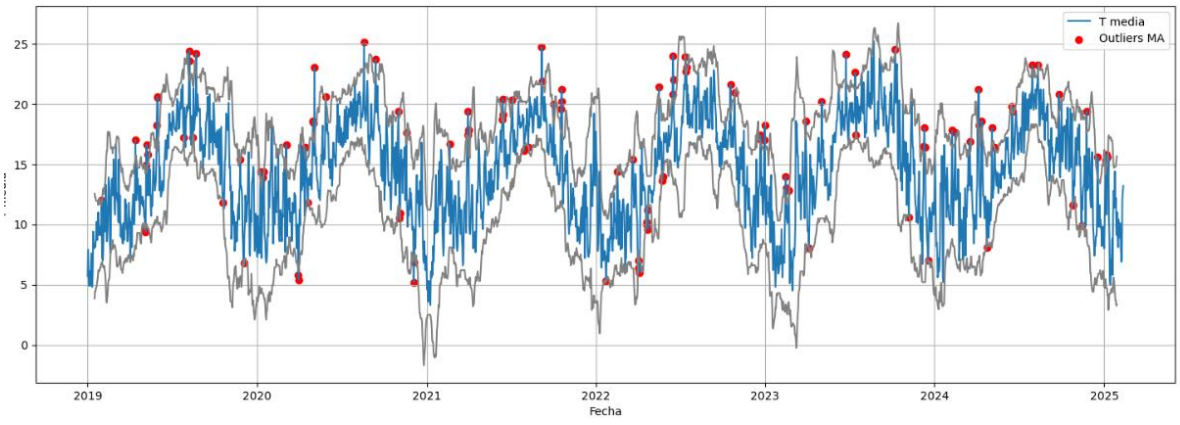
Para la ampliación de variables que aporten a la predicción del modelo, se han creado finalmente mediante el uso de OHE(OneHotEncoding) columnas para conocer en que estación del año se encuentra y valores medios máximos y mínimos por cada día de cada año para el conjunto de datos (por ejemplo para el día 1 de enero se ha hecho la media, máximo y mínimo de todos los 1 de enero y se han dispuesto en esta casilla)

Datos diarios Asturias: No se observa una tendencia de descenso significativo, pero es posible notar una cierta caída que coincide con

principios de año concretamente en enero de 2021 con el fenómeno Filomena. También se pueden apreciar algunas ligeras caídas tanto en 2023, como en 2024 por ciertas rachas de temporales. Se puede observar como en el 2020 a causa de las restricciones de movilidad y la disminución de la actividad industrial durante la pandemia podría haber afectado por eso hay tantos altibajos. Lo mostramos lo explicado en la siguiente imagen:

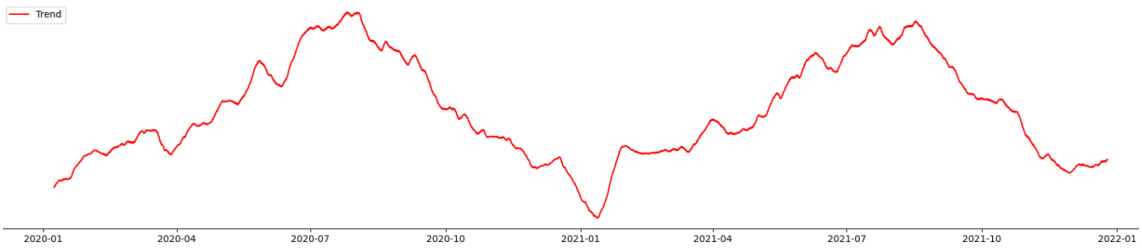


Respecto a los valores atípicos, se ha empleado el procedimiento de detección mediante el filtro de Hampel, el cual identifica anomalías dentro de un intervalo utilizando la mediana, dado que esta métrica es poco sensible a valores extremos. A modo de ilustración, se presenta un ejemplo de visualización de valores atípicos en la humedad relativa, donde los puntos en rojo representan dichos valores, ya que exceden los límites marcados en gris.

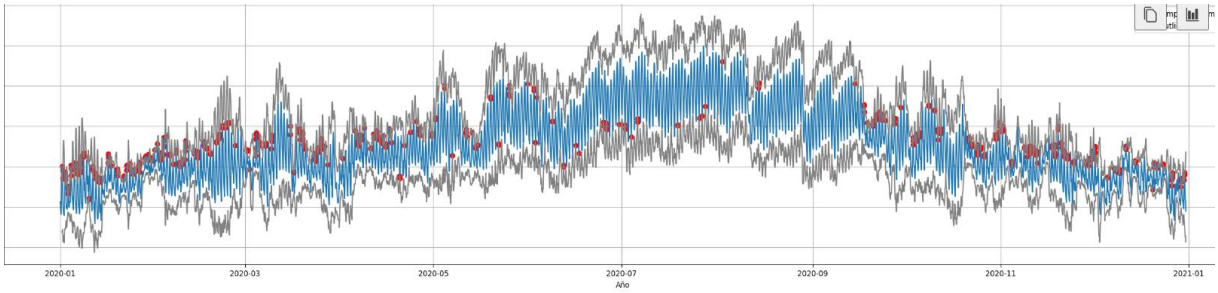


Datos horarios Madrid: por cada variable se ha hecho una descomposición estacional para ver sus componentes principales

como su tendencia, estacionalidad y residuo. Se observa que la tendencia baja en invierno y sube en verano como muestra la imagen, el pico en enero de 2021 fue debido al fenómeno meteorológico de Filomena:



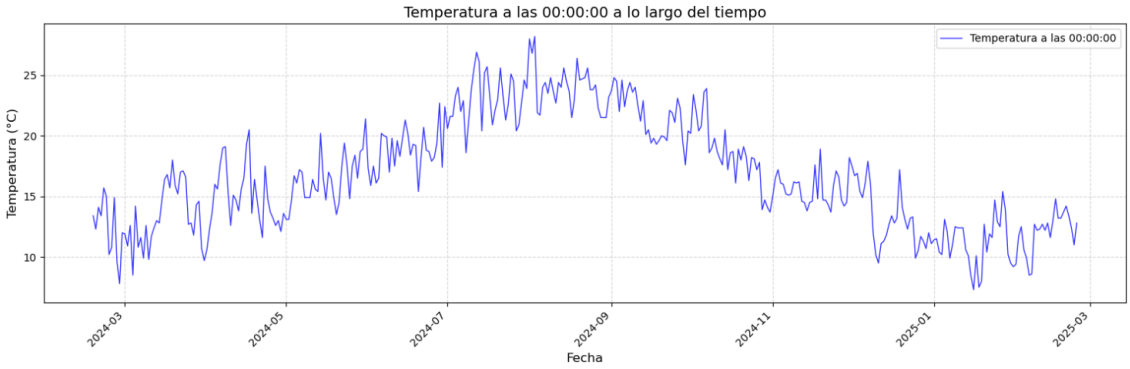
En relación con los outliers, para detectarlos se ha utilizado el método de filtro de hampel, que detecta los outliers dentro de una ventana utilizando la mediana, ya que la mediana no suele ser muy afectada por outliers. Un ejemplo de visualización de outliers de año 2020 de la temperatura es el siguiente en el que los puntos rojos son los outliers ya que sobrepasan las líneas grises:



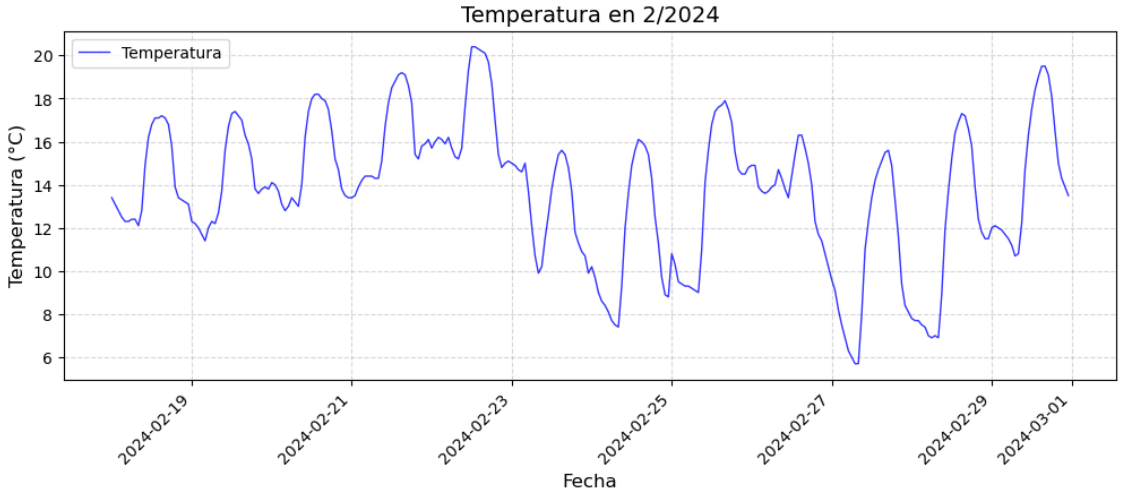
Para eliminar los outliers lo que se hace es pasar dicho valor al del filtro en dicha fecha, es decir, al valor de la linea gris en ese punto.

Datos horarios Marbella: para el caso de datos horarios, se ha procedido de manera similar a los datos diarios de Marbella, se ha preparado una gráfica que muestre como varia la temperatura por

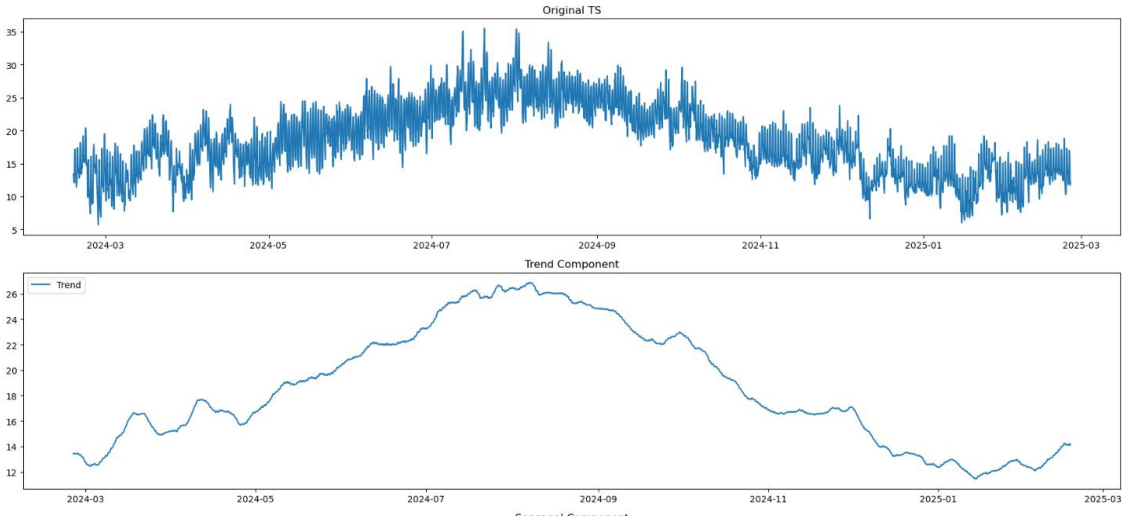
cada hora a lo largo del año:

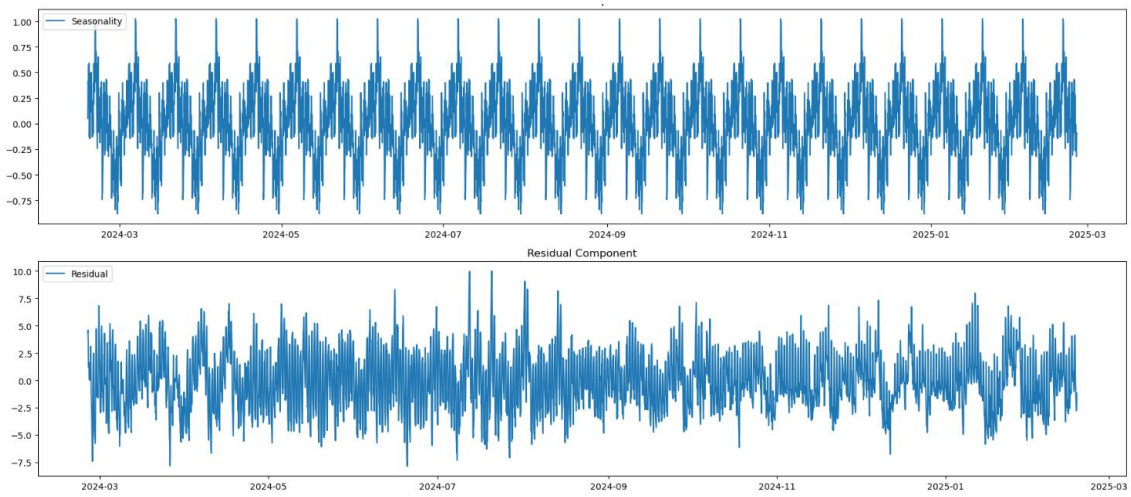


Y además como varían las temperaturas por mes:

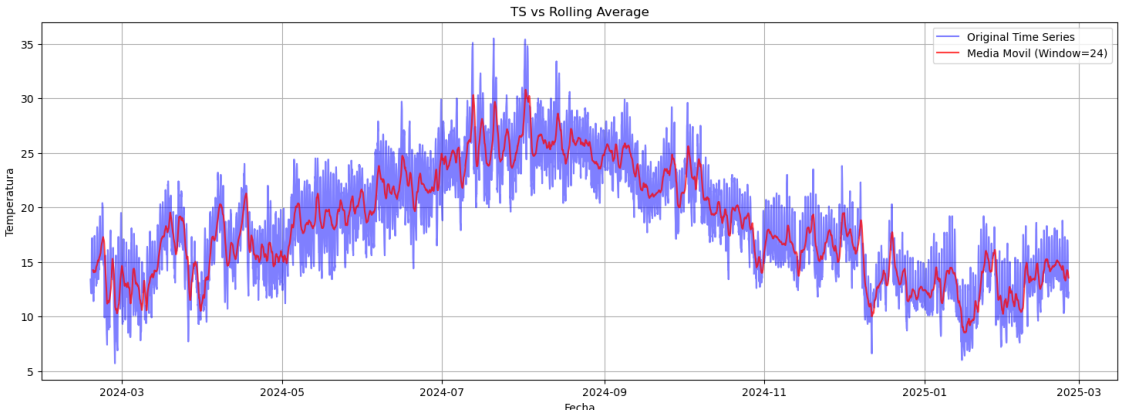


Como resultado de la descomposición en componentes de la serie temporal:

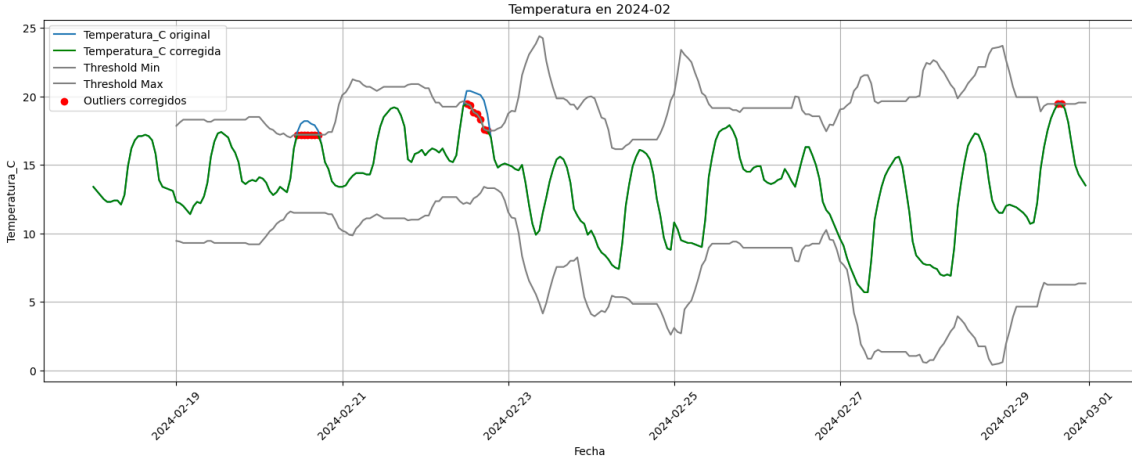




Análisis de la media movil



Y la comprobación para conocer si la serie temporal es o no estacionaria, para este caso ha resultado ser estacionaria por lo que no requiere de ningúna transformación de diferencia. Para trabajar con los outliers se ha procedido de la misma que con los datos diarios, se ha detectado outliers utilizando el Filtro Hampel, y la imputación sobre estos valores se ha realizado de acuerdo a los vecinos superior e inferior según convenga.

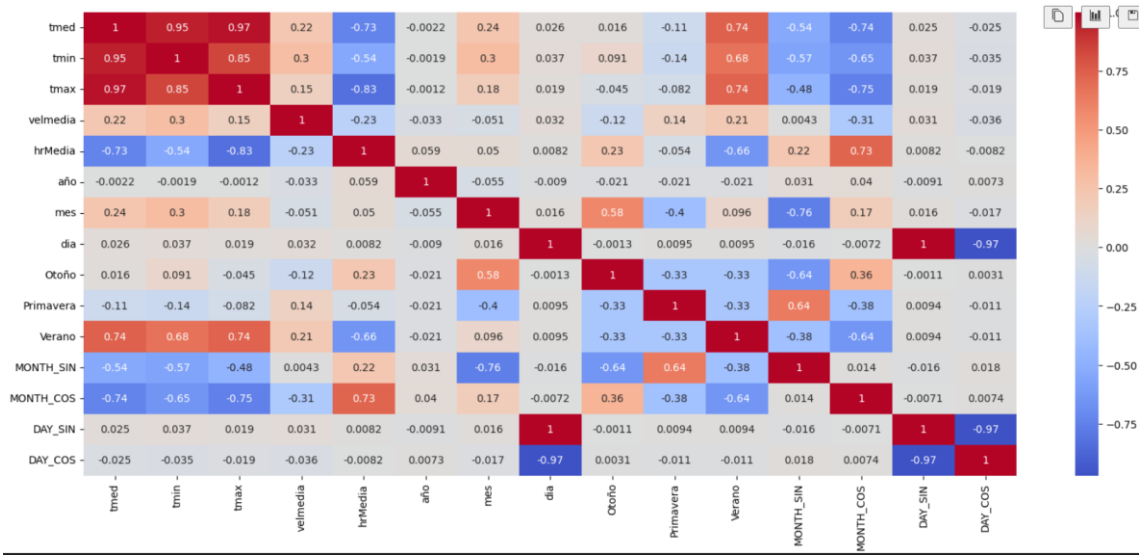


De manera complementaria se han creado las columnas para la estación en la que se encuentre, según la hora si es de día o de noche, y valores de año, mes y día.

5. MODELING

Con el paso anterior realizado, procedemos aplicar los modelos que mejor se adaptan a nuestros datos:

Predicción datos diarios Alcalá de Henares: antes de realizar ninguna predicción se crearon varias columnas exógenas como la temperatura media, temperatura máxima media y temperatura mínima media por día del mes, una columna que represente la estación en la que se encuentra dicha fecha a la que luego se le aplicará One Hot Encoding para convertir la columna categórica en numérica y por último varias columnas para representar de forma cíclica los meses y días. Hechas las columnas exógenas se muestra el mapa de calor obtenido:



Se ve que hay mucha relación entre las variables de temperatura y las variables de seno y coseno de las fechas.

Ya realizados los anteriores pasos se eligieron las columnas a predecir: en el caso de predicción de un día se eligieron temperatura media, temperatura máxima, temperatura mínima y humedad relativa y para la predicción de 7 días temperatura media y temperatura máxima. Los modelos utilizados para predecir dichas variables han sido:

Regresión Lineal Ridge: es una variante de la regresión lineal para manejar los problemas que conlleva predecir variables con columnas exógenas muy correlacionadas entre sí. Además, introduce un índice de determinación que reduce que las predicciones tomen valores extremos por lo que reduce el overfitting.

XGBoost Regressor: modelo de machine learning que combina modelos débiles de árboles de regresión para crear un modelo más potente, dichos modelos intentan mejorar respecto sus antecesores.

Random Forest Regressor: modelo de machine learning que combina múltiples árboles de regresión para crear un modelo más robusto. El resultado es la media de todos los resultados de cada árbol.

SARIMAX (Seasonal AutoRegressive Integrated Moving Average with eXogenous variables): modelo estadístico que es una combinación de varias componentes: *AR* relaciona el valor actual con los pasados mediante lags, *MA* coge los errores pasados, *I* indica que los valores de la serie original han sido reemplazados por la diferencia entre valores consecutivos.

Librería SKForecast Recursive: esta librería facilita las predicciones de series temporales haciendo predicciones multipaso en el que cada predicción se usa como entrada para predecir el siguiente valor. Algunos algoritmos descritos anteriormente se han utilizado con esta librería.

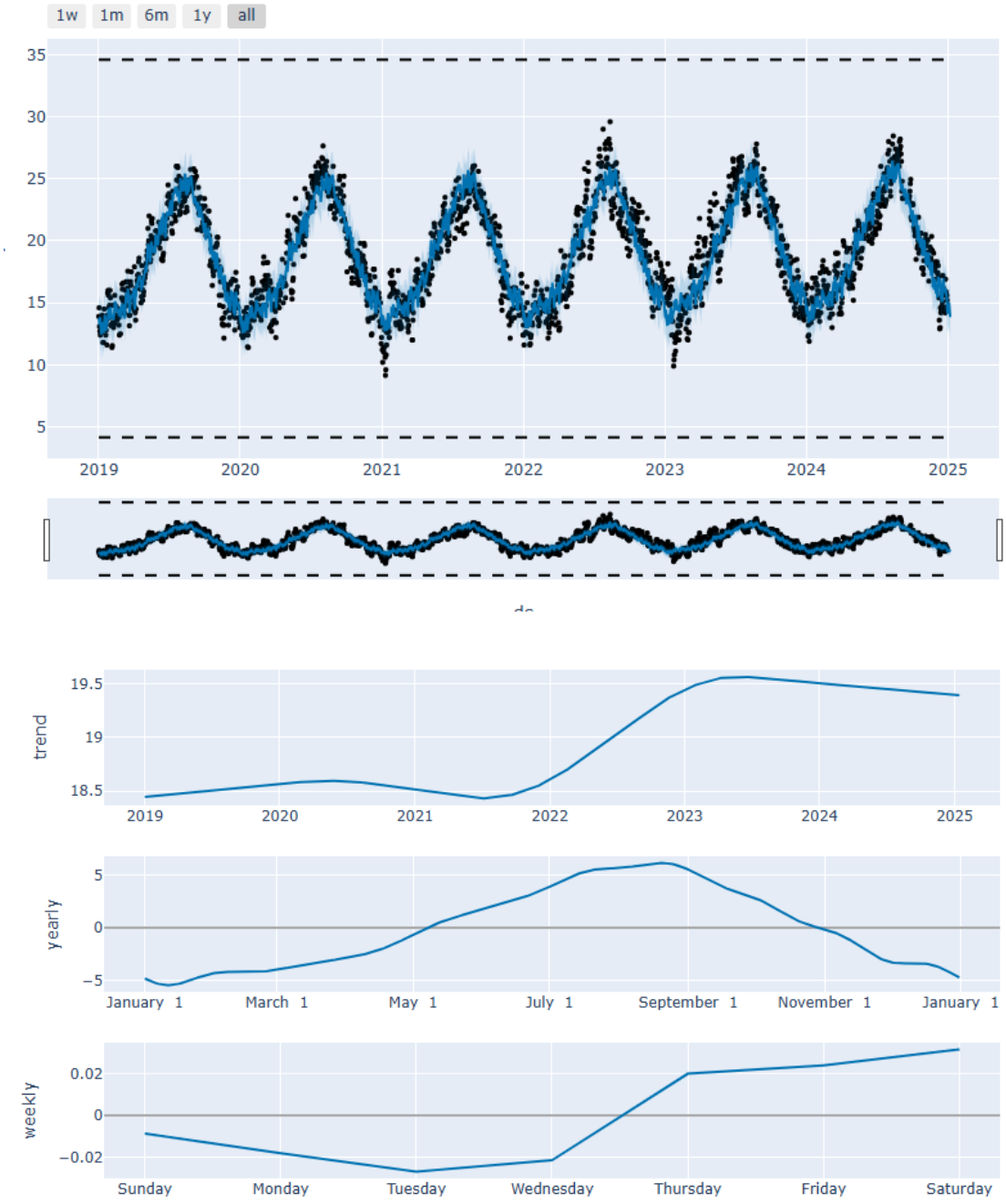
Para aplicar los distintos algoritmos, primero ha habido una etapa de pruebas con un 90% de train y 10 % de test en las que se demuestra la eficacia de los modelos junto a las variables exógenas. Se observa que para esta etapa hay un poco de data leakage con las variables exógenas ya que al hacer la media de día y mes le estoy dando pequeña información de lo que va a pasar, aunque no es muy relevante ya que la librería skforecast toma más en consideración los valores anteriores, pero para las predicciones a futuro no debería de haber data leakage ya que los valores exógenos solo cogen los valores que se tienen en el dataframe original.

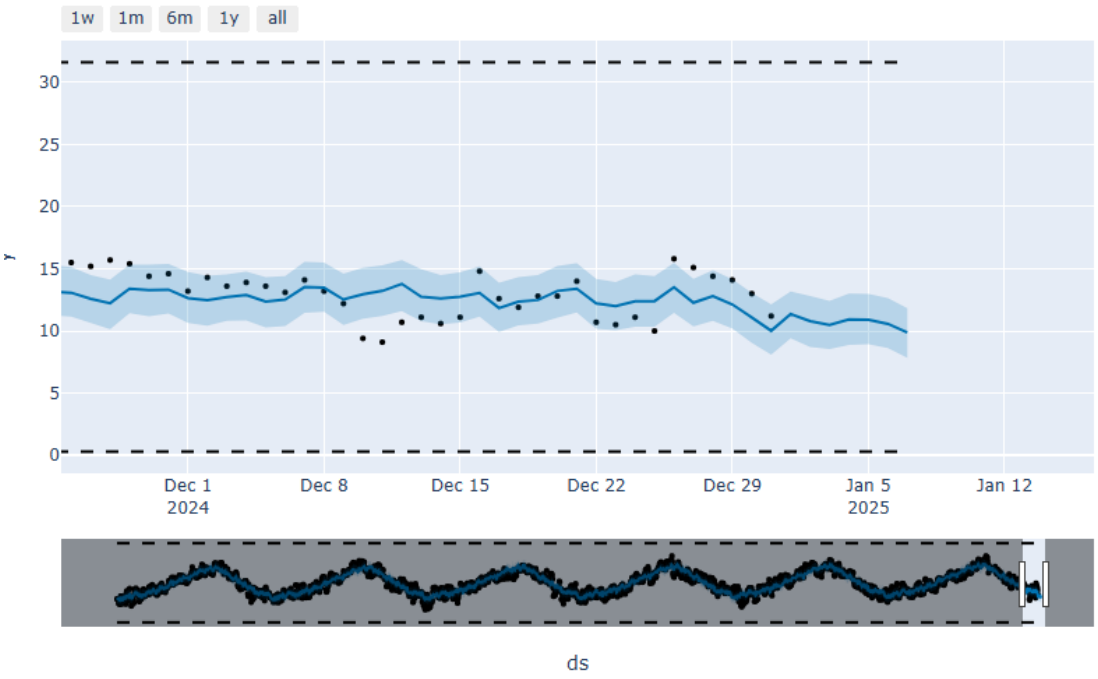
Para medir el error de cada modelo se ha utilizado la métrica RMSE que calcula la diferencia de los valores predichos y los reales, los eleva al cuadrado para dar más peso a outliers, hace la media de esos errores y por último hace la raíz cuadrada para volver a la escala original. Se ha utilizado dicha métrica ya que está en la misma escala que los datos reales y es fácil de entender.

Al aplicar los modelos se observa que se han obtenido ligeramente peores resultados para predecir la temperatura mínima ya que hay mucha varianza por los valores negativos en invierno y los valores de verano y en las predicciones de 7 días en el futuro ya que al utilizar valores anteriores para predecir el futuro, el modelo tiende a predecir valores muy parecidos mientras más nos alejemos de los valores reales debido al sumatorio de errores. Por ejemplo, en el caso de la temperatura media los resultados de test han tenido un error que varía entre 0.25 en el caso de Ridge (error más bajo) y 1.51 en SARIMAX (error más alto) lo que significa que en el peor de los casos se equivoca como máximo 1.51 unidades por arriba o por abajo.

Predicción datos horarios Marbella: Para realizar la predicción de datos diarios de Marbella se han utilizado los siguientes algoritmos:

Prophet: es una librería proporcionada por Facebook centrada en la predicción según series temporales, para este caso hemos realizado un primer entrenamiento con los datos de fecha y variable a predecir, para obtener una primera impresión al respecto para poder compararlo con otros modelos, en este caso ha sido el que mejores resultados y menor error ha demostrado en su predicción.


1w
1m
6m
1y
all
35
30
25
20
15
10
5
2019
2020
2021
2022
2023
2024
2025
trend
19.5
19
18.5
yearly
5
0
−5
January 1
March 1
May 1
July 1
September 1
November 1
January 1
weekly
0.02
0
−0.02
Sunday
Monday
Tuesday
Wednesday
Thursday
Friday
Saturday



Para la predicción tanto del día siguiente como de los próximos 7 días se ha dispuesto de variables exógenas como ['tmed_max', 'tmed_min', 'tmed_mean', 'estacion_Invierno', 'estacion_Otoño', 'estacion_Primavera', 'estacion_Verano']. Estos valores mediante itertools se ha encontrado los mejores hiperparametros para reducir los errores dispuestos de mae y rmse:

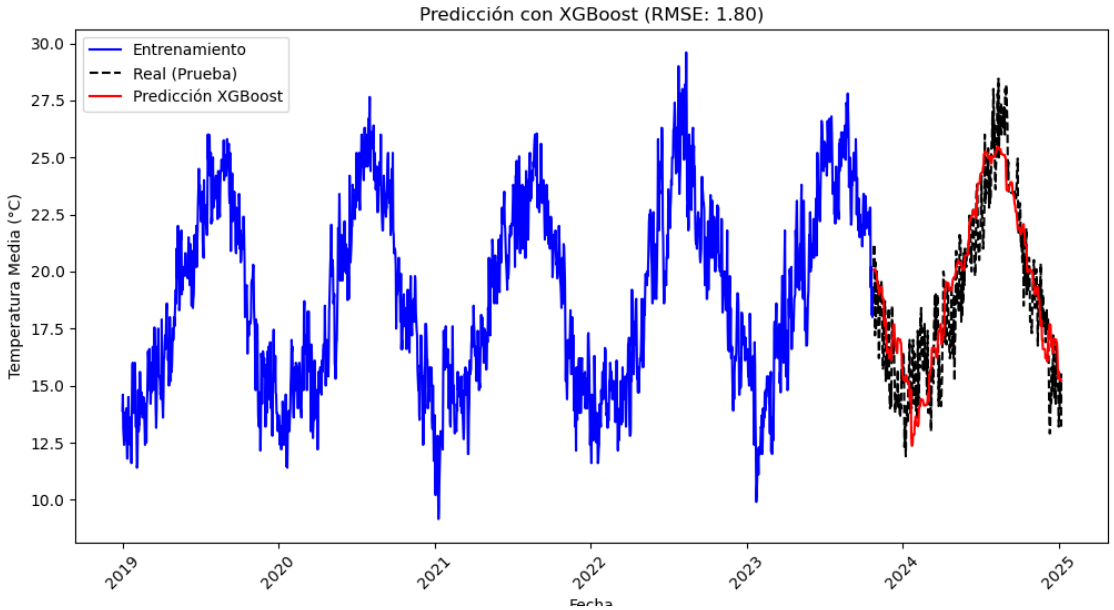
```
Error MSE: 0.9319529906695896
Error RMSE: 0.9653771235478856
Error MAE: 0.787282492090071
```

Se ha procedido de la misma manera para las 4 variables a predecir.

SARIMAX: este modelo de AST mostró peor rendimiento en las predicciones. Para este modelo se dispuso de hiperparámetros iniciales deducidos del análisis previo para la predicción de datos teniendo en cuenta que funciona de la siguiente manera: parámetros de entrada: (p, d, q) (P, D, Q, S)

Donde p es el número de términos autorregresivos (AR), del número de diferenciación para convertir la serie en estacionaria, q el número de términos de media movil (MA), P el número de términos AR estacionales, D diferenciación estacional, Q número de términos MA estacional, S periodicidad de la estacionalidad.

XGBoost: modelo basado en arboles de decisión, tiene mejor rendimiento en las predicciones que SARIMAX pero peor que prophet.



Predicción datos horarios Asturias: Para realizar la predicción de datos meteorológicos en Asturias, se utilizó el modelo **Prophet**, una herramienta especializada en series temporales que permite descomponer la tendencia, la estacionalidad y los efectos residuales de los datos. Antes de entrenar el modelo, se realizó una división del conjunto de datos, reservando los últimos 7 días para la fase de prueba y empleando el resto para el entrenamiento.

Con el objetivo de mejorar la precisión del modelo, se llevó a cabo un ajuste de hiperparámetros, evaluando distintas combinaciones para optimizar los valores de **changepoint_prior_scale** y **seasonality_prior_scale**. Estos parámetros permiten regular la sensibilidad del modelo a los cambios bruscos en la tendencia y la influencia de los componentes estacionales en la predicción.

Para cada combinación de valores, se entrenó Prophet y se compararon las predicciones con los datos reales mediante el cálculo del **error cuadrático medio (RMSE)**, seleccionando finalmente la configuración que minimizaba este error.

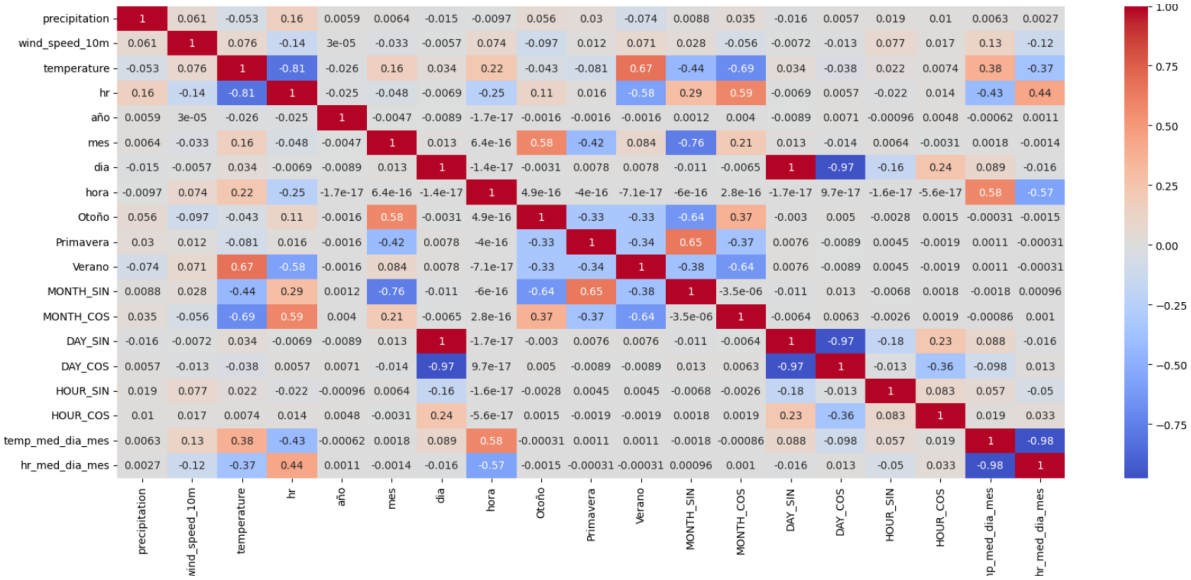
Tras el entrenamiento, se generó un dataframe de predicción a 7 días, permitiendo evaluar la capacidad del modelo para anticipar la evolución de las variables meteorológicas. Se observó que Prophet consigue capturar correctamente la tendencia general de los datos,

aunque en predicciones a mayor plazo tiende a suavizar los valores debido a la acumulación de errores. Este comportamiento es habitual en modelos de series temporales cuando dependen exclusivamente de datos históricos sin variables exógenas adicionales.

Para la interpretación de los resultados, se utilizaron herramientas de visualización como **plotly**, lo que permitió analizar la relación entre los valores reales y los predichos de manera interactiva. A través de estas representaciones gráficas, se pudo identificar patrones en la evolución de las variables meteorológicas y evaluar la efectividad del modelo en distintos periodos.

En conclusión, el uso de **Prophet** demostró ser una estrategia efectiva para modelar las series temporales meteorológicas en Asturias, proporcionando predicciones razonablemente precisas. Aunque los resultados pueden verse afectados por la acumulación de errores en predicciones a largo plazo, el modelo logra un equilibrio entre simplicidad y precisión, permitiendo generar estimaciones útiles para el análisis y la toma de decisiones.

Predicción datos horarios Madrid: antes de realizar ninguna predicción se crearon varias columnas exógenas como la temperatura media y la humedad media por hora del día, una columna que represente la estación en la que se encuentra dicha fecha a la que luego se le aplicará One Hot Encoding para convertir la columna categórica en numérica y por último varias columnas para representar de forma cíclica los meses y días. Hechas las columnas exógenas se muestra el mapa de calor obtenido:



Se ve que hay mucha relación entre las variables de temperatura y las variables de seno y coseno de las fechas.

Ya realizados los anteriores pasos se eligieron las columnas a predecir: tanto como el caso de predicción de una hora y el de 24 horas se eligieron la temperatura y humedad relativa. Los modelos utilizados para predecir dichas variables han sido:

Regresión Lineal Ridge: es una variante de la regresión lineal para manejar los problemas que conlleva predecir variables con columnas exógenas muy correlacionadas entre sí. Además, introduce un índice de determinación que reduce que las predicciones tomen valores extremos por lo que reduce el overfitting.

XGBoost Regressor: modelo de machine learning que combina modelos débiles de árboles de regresión para crear un modelo más potente, dichos modelos intentan mejorar respecto sus antecesores.

Random Forest Regressor: modelo de machine learning que combina múltiples árboles de regresión para crear un modelo más robusto. El resultado es la media de todos los resultados de cada árbol.

Librería SKForecast Recursive: esta librería facilita las predicciones de series temporales haciendo predicciones multipaso en el que cada predicción se usa como entrada para predecir el

siguiente valor. Algunos algoritmos descritos anteriormente se han utilizado con esta librería.

Para aplicar los distintos algoritmos, primero ha habido una etapa de pruebas con un 90% de train y 10 % de test en las que se demuestra la eficacia de los modelos junto a las variables exógenas. Se observa que para esta etapa hay un poco de data leakage con las variables exógenas ya que al hacer la media de hora y día le estoy dando pequeña información de lo que va a pasar, aunque no es muy relevante ya que la librería skforecast toma más en consideración los valores anteriores, pero para las predicciones a futuro no debería de haber data leakage ya que los valores exógenos solo cogen los valores que se tienen en el dataframe original.

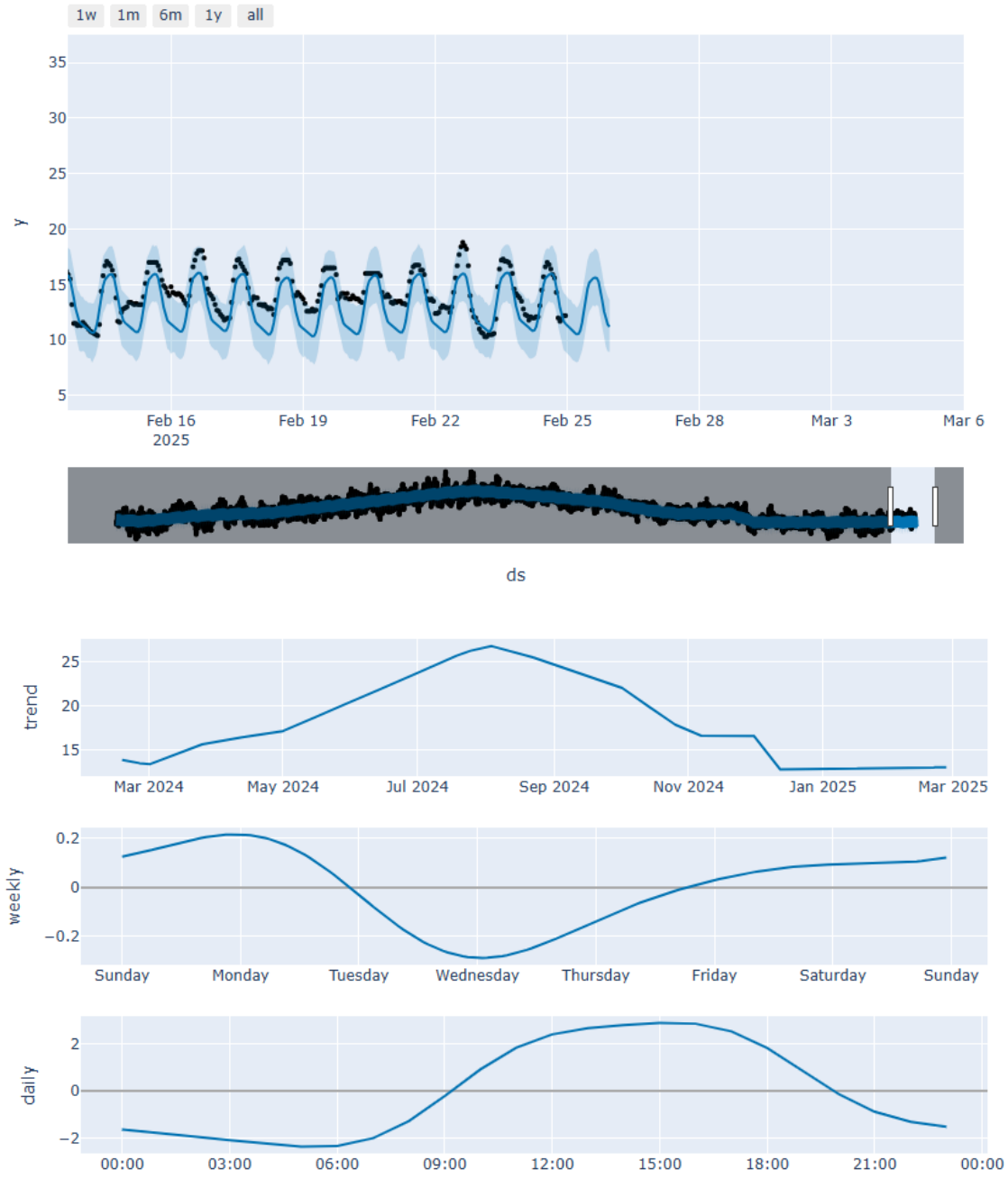
Para medir el error de cada modelo se ha utilizado la métrica RMSE que calcula la diferencia de los valores predichos y los reales, los eleva al cuadrado para dar más peso a outliers, hace la media de esos errores y por último hace la raíz cuadrada para volver a la escala original. Se ha utilizado dicha métrica ya que está en la misma escala que los datos reales y es fácil de entender.

Al aplicar los modelos se observa que se han obtenido ligeramente peores resultados para predecir las variables en la ventana de 24 horas ya que al utilizar la librería recursive toma los valores anteriores para predecir y mientras más se alejen las predicciones de los datos peor va a predecir por el sumatorio de errores.

Predicción datos horarios Marbella: para las predicciones de los datos horarios para una hora y 24 horas de marbella, basándonos en la experiencia previa, se ha realizado con el modelo que mejores predicciones dio en los datos diarios de Marbella, la librería de Facebook prophet.

Prophet: para este caso se ha procedido de manera similar a la utilización de prophet para la predicción de 7 dias, primero con el objeto de mejor entendimiento de la serie se realiza una predicción únicamente sobre las fechas y la variable a predecir, sobre ello se

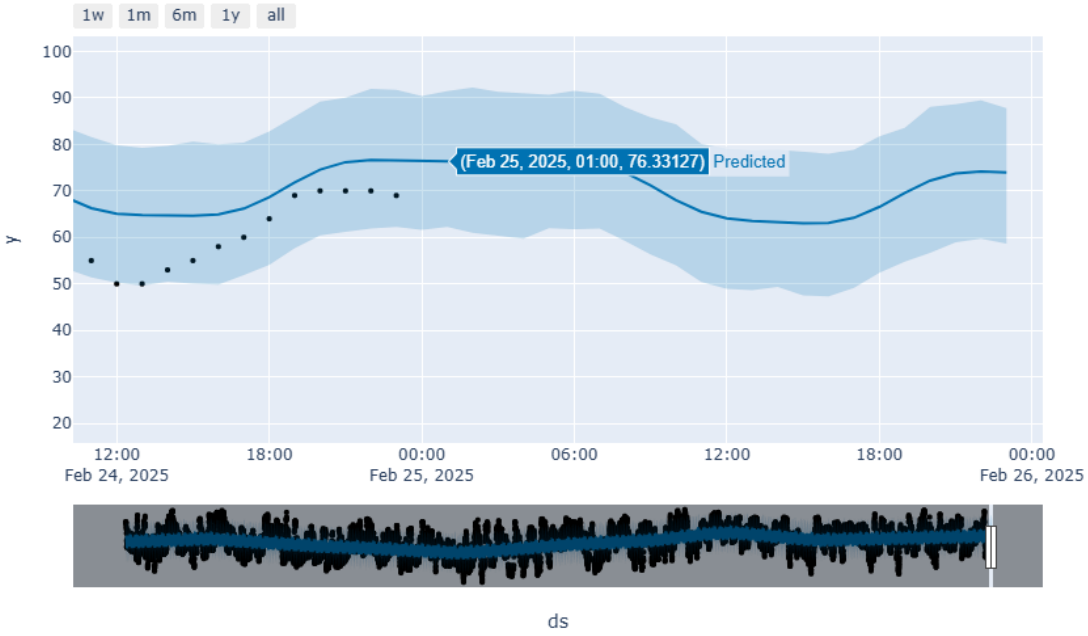
analiza sus componentes para seguir adelante:



Sobre estos datos disponemos de la variables exógenas a nuestra disposición, en este caso consiste en ['is_day', 'estacion_Invierno', 'estacion_Otoño', 'estacion_Primavera', 'estacion_Verano']

El modelo se vuelve a entrenar recibiendo estos parámetros de variables exógenas como valores regresivos y se procede a realizar la predicción de las próximas 24 horas (ejemplo se utiliza la

temperatura):



6. CUADRO DE MANDOS

Una vez aplicados todos los modelos predictivos y obtenidos los resultados esperados se procede a la visualización de dichos datos junto a su histórico para una mayor comprensión del trabajo realizado. Para esta tarea se optado por utilizar PowerBI, que es una herramienta que permite analizar y visualizar datos además de facilitar la conexión con varias fuentes de datos y crear informes interactivos.

Antes de crear ninguna visualización se han tenido que realizar una serie de transformaciones sobre los datos para su correcto funcionamiento:

- Primero se han cargado los datos y se han realizado transformaciones sencillas como cambio de tipo de columna, eliminar columnas innecesarias y se ha añadido una columna nueva llamada tipo que diferenciará los datos históricos de las predicciones.

- Acto seguido se ha realizado un append de las dos tablas para juntar los datos, de ahí el crear la columna anterior.

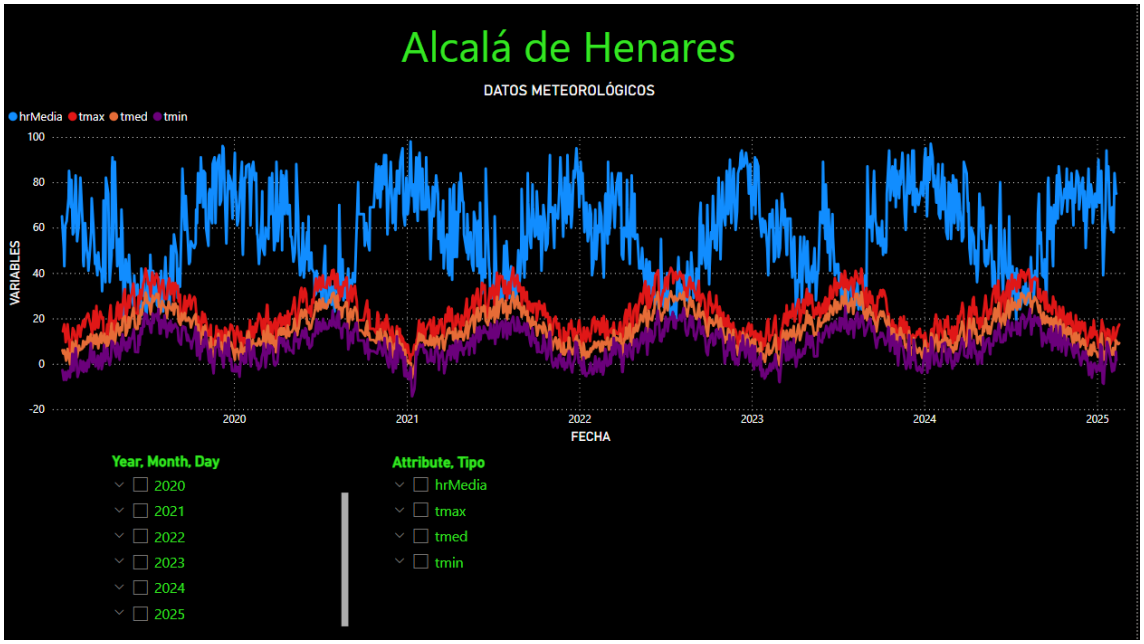
- Para poder aplicar filtros a nuestras futuras visualizaciones hay que convertir nuestras columnas de datos meteorológicos en un formato atributo-valor como muestra la imagen:

1.2 tmed	1.2 tmin	1.2 tmax	1.2 hrMedia
5,3	-4,1	14,7	62
4,6	-5,5	14,6	65
6	-2,8	14,7	63

ABC Attribute	1.2 Value
tmed	5,3
tmin	-4,1
tmax	14,7
hrMedia	62

Como se ve en la imagen hemos agrupado todas las columnas en una y hemos creado otra para el valor de cada una.

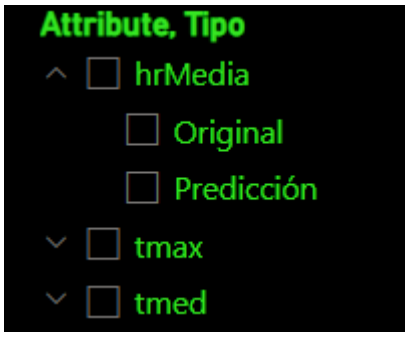
Ya realizadas las transformaciones pertinentes se ha creado la siguiente visualización, cuyo formato será el mismo para todos los municipios:



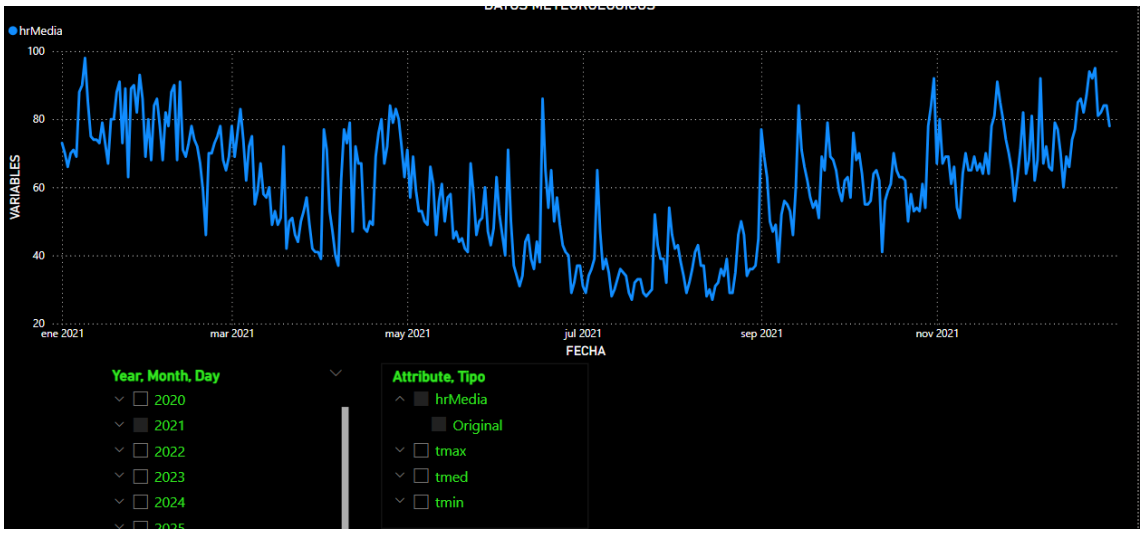
La visualización principal es una gráfica de líneas a la que en el eje X se le ha puesto la fecha, en el eje Y la columna Value y como leyenda se ha puesto la columna Atributte.

Las dos cajas que están debajo de la gráfica son filtros, el de la izquierda corresponde al filtro de fechas en el que podemos filtrar por año, mes y día en el caso de datos diarios. Para el caso de datos horarios también se podrá filtrar por hora.

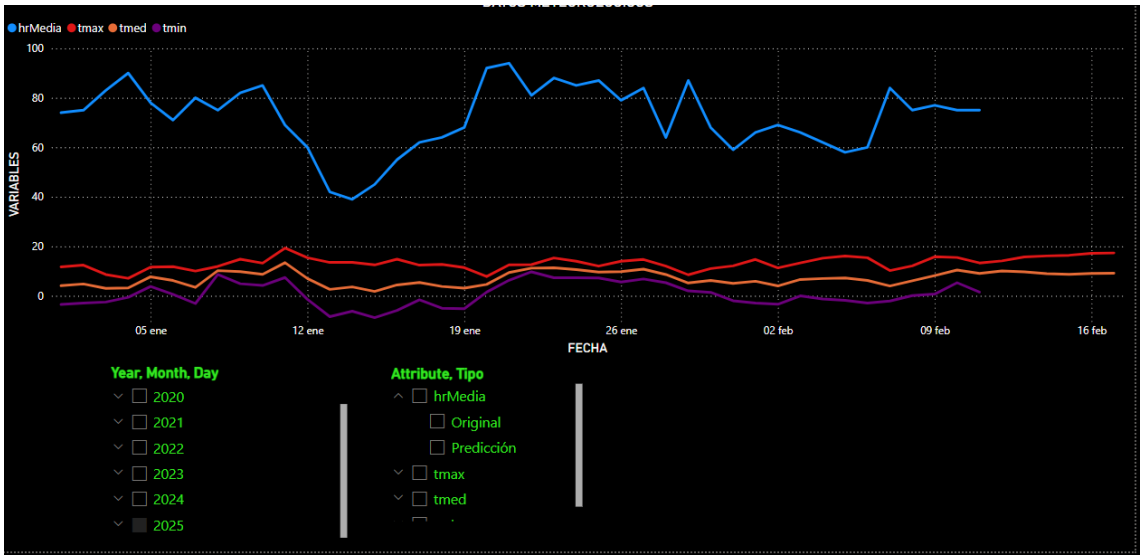
La caja de la derecha es el filtro de variable en el que podemos elegir que variable mostrar además de mostrar los datos históricos y las predicciones ya que hay un desplegable como muestra la imagen:



Se muestra un ejemplo de una gráfica con filtros aplicados, en este caso, muestra los datos de la temperatura media del año 2021 en formato mes:

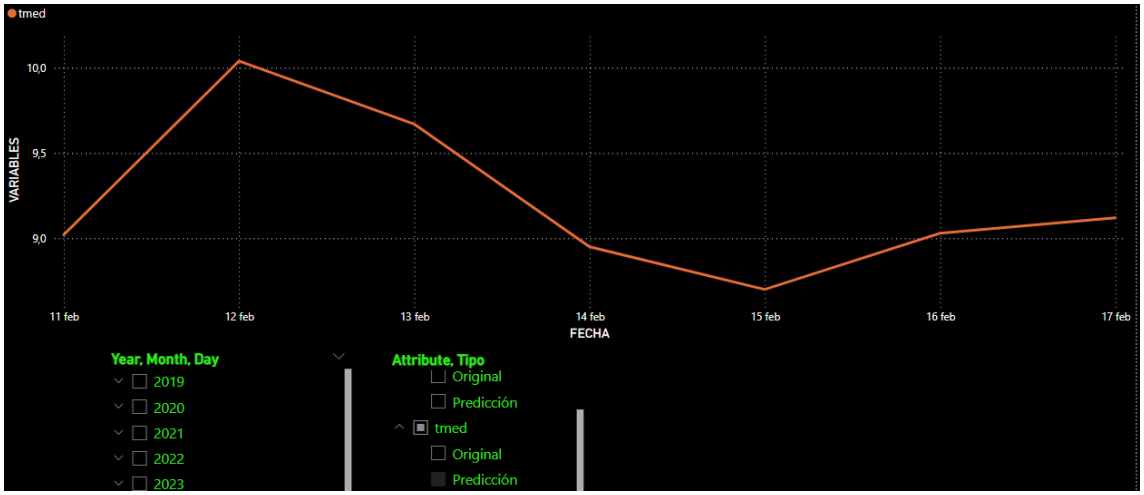


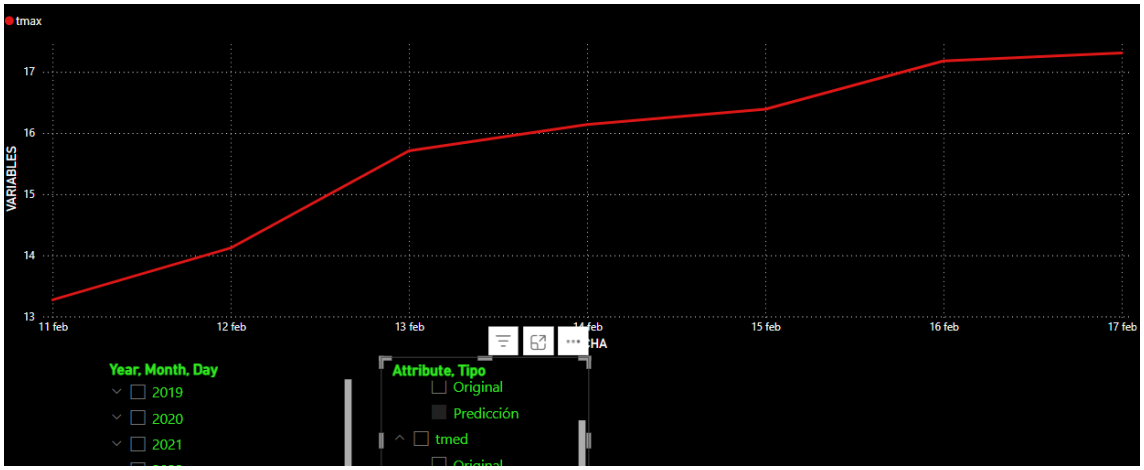
En el siguiente ejemplo se muestra una gráfica de todas las variables en el año 2025 de datos históricos junto con las predicciones:



Se ve que las variables tmax y tmed se han predicho para 7 días y las otras dos solo para un día.

Los resultados de la predicción se muestran en estas imagen:




tmax
17
16
15
14
13
VARIABLES
11 feb
12 feb
13 feb
15 feb
16 feb
17 feb
HA
Year, Month, Day
2019
2020
2021
Attribute, Tipo
Original
Predicción
tmed