



UNIVERSIDAD NACIONAL  
DE EDUCACIÓN A DISTANCIA

Escuela Técnica Superior de Ingeniería Informática

# Técnicas de Deep Learning para la Predicción de Precios de Electricidad del Mercado Diario Español

Hugo Alberto Pedrero Lozoya

Director: Rafael Vargas Pastor  
Co-director: Wolfram Rozas Rodríguez

Trabajo de Fin de Máster

Máster Universitario en  
Ingeniería y Ciencia de Datos

Febrero 2025



## RESUMEN

La estimación de los precios de electricidad del mercado diario es una tarea de gran importancia e interés para todos los agentes involucrados en el sector, de cara a realizar las ofertas de compra y venta de electricidad antes del cierre del mercado.

En este proyecto se van a desarrollar métodos de *deep learning* para predecir los precios de la electricidad del mercado diario español (“day-ahead”).

En concreto, se pretende implementar técnicas de deep learning basadas en Redes Neuronales Recurrentes (RNN) tales como LSTM (Long Short-Term Memory) y GRU (Gated Recurrent Unit), y también modelos más novedosos como NBEATS (Neural Basis Expansion Analysis for Time Series) o un *transformer* especializado en series temporales llamado TimeGPT.

Tomando como ventana de datos de entrada los 168 valores previos (1 semana), se realizarán predicciones para los precios de la electricidad de las 24 horas del día siguiente. Se emplean datos históricos que abarcan el periodo comprendido entre 2015 y 2023, incluyendo variables fundamentales como la demanda eléctrica, la generación de fuentes renovables y el precio del gas natural.

Además, se introduce el concepto de "hueco hidrotérmico", una nueva variable creada a partir de la combinación de la demanda eléctrica y la generación de energía eólica y solar. Esta variable corresponde a la energía que habrá que generar con fuentes térmicas o hidráulicas para cubrir la demanda, una vez sustraída la generación eólica y solar. La adición de esta variable, altamente correlacionada con el precio de la electricidad, mejora notablemente el comportamiento de los modelos.



## Tabla de contenido

<b>RESUMEN .....</b>	<b>3</b>
<b>ÍNDICE DE FIGURAS .....</b>	<b>7</b>
<b>ÍNDICE DE TABLAS .....</b>	<b>9</b>
<b>1. INTRODUCCIÓN .....</b>	<b>11</b>
1.1 Contexto.....	11
1.2 Estado del Arte .....	12
1.3. Objetivos .....	12
<b>2. ANÁLISIS Y PREPARACIÓN DE LOS DATOS .....</b>	<b>14</b>
2.1. Análisis exploratorio de las variables .....	16
2.1.1. Demanda .....	17
2.1.2. Precio de la electricidad y del gas .....	18
2.1.3. Generación de energía renovable no gestionable: solar fotovoltaica y eólica .....	21
2.1.4. Generación y bombeo hidráulico .....	23
2.1.5. Cantidad de agua embalsada .....	24
2.1.6. Generación de Ciclos Combinados .....	25
2.2. Nuevas variables.....	26
2.2.1. Hueco Térmico .....	26
2.2.2. Resumen de variables: cobertura de la demanda.....	27
2.2.3. Variables relativas al calendario.....	30
2.3. Correlación entre variables .....	31
2.4. Descomposición de la variable objetivo y estacionariedad .....	33
<b>3. METODOLOGÍA Y MODELOS.....</b>	<b>38</b>
3.1. Modelos de Machine Learning .....	39
3.2. LSTM.....	41
3.3. GRU .....	43
3.4. NBEATSx .....	44
3.5. NHITS.....	47
3.6. TimeGPT .....	49
<b>4. RESULTADOS .....</b>	<b>51</b>
<b>5. CONCLUSIONES Y LÍNEAS FUTURAS.....</b>	<b>60</b>
<b>BIBLIOGRAFÍA.....</b>	<b>62</b>



# ÍNDICE DE FIGURAS

Figura 1: Evolución de la demanda eléctrica anual.....	17
Figura 2: Evolución de la demanda eléctrica de 2023 PONER LEYENDA.....	18
Figura 3: Evolución temporal de los precios .....	19
Figura 4: Precios medios anuales .....	20
Figura 5: Día promedio de invierno (enero 2023).....	20
Figura 6: Día promedio de verano (julio 2023).....	21
Figura 7:Evolución temporal de la generación de energía solar fotovoltaica .....	21
Figura 8: Generación FV de un día promedio de verano y otro de invierno (2023) .....	22
Figura 9: Generación eólica horaria por meses (media de todos los años) .....	22
Figura 10: Generación total hidráulica anual .....	23
Figura 11: Generación y bombeo hidráulico de un día promedio de enero y agosto (media de todos los años) .....	24
Figura 12: Nivel medio anual de los embalses .....	24
Figura 13: Distribución mensual del nivel de los embalses y la generación hidráulica media .....	25
Figura 14: Generación diaria de un día medio para tres meses distintos (media de todos los años).....	26
Figura 15: Histograma de la generación de ciclos combinados .....	26
Figura 16: Cobertura de la demanda en una semana de invierno eje Y en MW! .....	28
Figura 17: Cobertura de la demanda en una semana de verano .....	28
Figura 18: Cobertura de la demanda en una semana de invierno (hueco térmico) .....	29
Figura 19: Cobertura de la demanda en una semana de verano (hueco térmico).....	30
Figura 20: Codificación cíclica de la hora del día.....	31
Figura 21: Matriz de correlaciones de las variables para todo el periodo considerado..	32
Figura 22: Correlaciones con la variable precio para el perfil diario (24h) promedio ...	32
Figura 23: Autocorrelación y autocorrelación parcial de la variable objetivo .....	33
Figura 24: Descomposición de la serie temporal “precio de la electricidad”.....	34
Figura 25: Descomposición de la serie temporal “precio de la electricidad” en enero de 2023 .....	35
Figura 26: Funcionamiento de la validación cruzada .....	39
Figura 27: Esquema de la arquitectura de una célula LSTM (Antonio Gulli & Sujit Pal, 2017).....	42
Figura 28: Esquema de la arquitectura de una célula GRU (Antonio Gulli & Sujit Pal, 2017).....	44
Figura 29:Esquema de la arquitectura de NBEATS(Oreshkin et al., 2020).....	46
Figura 30: Esquema de la arquitectura de NHITS (Challu et al., 2022).....	48
Figura 31: Arquitectura del transformer TimeGPT .....	49
Figura 32: Valores reales vs predichos por LGBM de los 10 primeros días.....	52
Figura 33: Valores reales vs predichos por LSTM de los 10 primeros días.....	53
Figura 34: Valores reales vs predichos por GRU de los 10 primeros días .....	54

Figura 35: Predicciones del modelo NBEATSx para los 10 primeros días del conjunto de test.....	55
Figura 36: Predicciones del modelo NHITS para los 10 primeros días del conjunto de test.....	56
Figura 37: Predicciones del transformer TimeGPT para los 10 primeros días del conjunto de test.....	57
Figura 38:Predicciones de los tres mejores modelos para los 10 primeros días del conjunto de test.....	58

## ÍNDICE DE TABLAS

Tabla 1: Conjunto de datos .....	16
Tabla 2: Principales estadísticos de las variables .....	16
Tabla 3: Resultados de la validación cruzada con modelos de machine learning .....	51
Tabla 4: Resultados de los dos mejores modelos de machine learning tras el afinado de hiperparámetros .....	51
Tabla 5: Influencia del hueco térmico con el modelo LGBM .....	52
Tabla 6: Resultados de la validación cruzada con el modelo basado en LSTM .....	52
Tabla 7: Resultados de la validación cruzada con el modelo basado en GRU .....	53
Tabla 8: Resultados de la validación cruzada con el modelo NBEATSx .....	54
Tabla 9: Comparación de resultados de las pruebas con NBEATSx de la influencia del hueco térmico .....	55
Tabla 10: Resultados de la validación cruzada con el modelo NHITS .....	56
Tabla 11: Resultados de la validación cruzada con el transformer TimeGPT .....	56
Tabla 12: Resumen comparativo de los resultados de todos los modelos .....	57



## 1. INTRODUCCIÓN

### 1.1 Contexto

En el mercado diario (o “day-ahead”) español los diferentes agentes realizan sus ofertas de compra y venta de energía para las 24 horas del día siguiente. Se trata de un mercado marginalista, es decir, el precio de cada hora queda determinado por el coste de la tecnología necesaria para suministrar el último megavatio de la demanda.

Todos los días, los diferentes agentes del sector energético realizan sus ofertas de compra y venta de electricidad del día siguiente a través del operador del mercado (OMIE). A las 12:00 “cierra” el mercado y se llevan a cabo las casaciones entre oferta y demanda a nivel europeo, quedando determinados los precios y volúmenes de electricidad para cada hora del día siguiente (OMIE).

La predicción de los precios de la electricidad del mercado diario es fundamental para los diversos actores implicados en el sector energético. Por un lado, los generadores necesitan disponer de una estimación precisa de los precios de cara a planificar su producción y a realizar ofertas de venta de energía que maximicen sus beneficios. Por otro lado, los consumidores y comercializadores necesitan conocer de antemano una estimación de los precios del día siguiente de cara a optimizar su estrategia de compra y minimizar costes.

Estas necesidades se han acentuado en los últimos años debido a la creciente participación de las energías renovables en el mix energético español. Este rápido despliegue se ha visto motivado por la necesidad de reducir las emisiones de efecto invernadero que contribuyen al cambio climático. Todos los países europeos, en mayor o menor medida, han adquirido el compromiso de descarbonizar su economía y apostar por energías limpias. Esto ha quedado reflejado en las sucesivas cumbres del clima de los últimos años, así como en las diversas medidas adoptadas por la Comisión Europea, como por ejemplo la propuesta Fit for 55, aprobada en 2021 y por la se pretende “adaptar la normativa comunitaria en materia de energía y cambio climático al nuevo objetivo de la UE para 2030: reducir al menos en un 55% la emisión de gases de efecto invernadero (GEI) respecto a 1990” [link](#).

En concreto, en España se han impulsado gran cantidad de proyectos de parques solares fotovoltaicos y eólicos en los últimos años, pasando de los 4,8 GW de potencia instalada de fotovoltaica que había a finales de 2018 (y que se había mantenido invariable años atrás) a los 25,7 GW a finales de 2023. Mientras que la eólica ha pasado de 23,7 a 30 GW en este mismo periodo.

Estas fuentes de energía son variables, dependientes de las condiciones meteorológicas y no son regulables. Además, tienen costes variables cercanos a cero y son las primeras en el orden de mérito, por lo que van a introducir una gran variabilidad y volatilidad en los precios, que van a tender a la baja en los momentos en los que se genere una gran cantidad de energía renovable y al alza cuando haya falta de recurso.

En este contexto, se hace imprescindible disponer de modelos que puedan aproximar los precios de la electricidad en cada hora del día siguiente de cara a la realización de ofertas de venta o compra. Para ello, en este proyecto se van a implementar métodos de aprendizaje automático profundo que permitan predecir esos precios de la electricidad antes del cierre del mercado diario.

### 1.2 Estado del Arte

En la bibliografía consultada existen trabajos previos en los que se han implementado métodos de Deep Learning para predecir los precios de la electricidad del “day-ahead”. Estos estudios suelen desarrollar, además de modelos estadísticos y técnicas clásicas de *machine learning*, modelos basados en RNN. Además, suelen predecir precios de mercados eléctricos de otros países, y periodos de tiempos anteriores a 2020.

En (Poggi et al., 2023) llevan a cabo un estudio para predecir el precio de la electricidad del mercado spot de Alemania en 2020 utilizando métodos estadísticos y de aprendizaje profundo. Introducen la novedad de dividir la serie temporal del precio en una componente de tendencia estacional y otra estocástica. Concluyen el estudio afirmando que el modelo basado en LSTM obtiene los mejores resultados, por encima de otros métodos como ARIMA o XGBoost. Además, mencionan como aspectos de mejora la adición de la demanda o la generación del día siguiente como variables exógenas y su utilización en modelos como NBEATSx.

En (Lago et al., 2018) se lleva a cabo una extensa revisión bibliográfica del estado del arte hasta la fecha del artículo, identificando hasta 23 modelos, y se proponen 4 arquitecturas basadas en DNN, CNN, LSTM y GRU para predecir los precios de la electricidad del mercado eléctrico belga. El conjunto de datos seleccionado abarca de 2010 a 2017. Se concluye afirmando que todos los modelos menos CNN obtienen mejores resultados que el resto de los métodos usados en la bibliografía analizada, tanto estadísticos como de machine learning. Además, el modelo basado en DNN obtiene ligeramente mejor resultado, en términos de la métrica sMAPE, que LSTM o GRU.

En (Yousefi et al.) se implementan modelos estadísticos y de aprendizaje automático para predecir el precio de la electricidad a largo plazo en el estado de California, con un conjunto de datos que abarca desde 2001 hasta 2017. Se utilizan una serie de características adicionales y se muestran sus correlaciones con el precio de la electricidad.

Finalmente, (García et al., 2024) es un estudio publicado en el Ministerio de Industria el primer trimestre de 2024 y centrado en predecir los precios de la electricidad del mercado diario español. En él, se desarrollan diversos modelos para diferentes períodos temporales (el más reciente de 2021), utilizando técnicas de deep learning y machine learning tradicional. Los mejores resultados los obtiene con el algoritmo *StreamWNN* (basado en KNN) seguido de LSTM (para el periodo más reciente). Sin embargo, se echa en falta un modelo centrado en la predicción de precios en la actualidad, así como la incorporación de variables explicativas.

### 1.3. Objetivos

En este proyecto, los modelos que se van a presentar abordan concretamente la predicción de los precios del mercado diario eléctrico español, abarcando hasta 2023, lo cuál incluye el periodo de confinamiento debido al COVID-19, el periodo de precios de gas muy altos debido a la guerra en Ucrania y la variabilidad introducida en los precios de la electricidad por el aumento de la penetración de fuentes renovables en el mix energético en los últimos años.

Se establecen los siguientes objetivos en este proyecto:

1. Desarrollo de modelos que sean capaces de predecir los precios del mercado diario eléctrico español en la actualidad, capturando la variabilidad aportada por un mix energético con gran participación de energía renovable, así como en un contexto de elevado precio de gas.
2. Incorporación de variables exógenas, tales como la demanda o generación de energía renovable para predecir de forma más precisa los precios.
3. Implementación de arquitecturas más novedosas que las ya utilizadas ampliamente en la bibliografía y potencialmente propicias, o que han demostrado buen comportamiento, para la predicción de series temporales, tales como NBEATS o TimeGPT.
4. Adición de una variable nueva, creada a partir de las conocidas, que mejore la predicción de precios al proporcionar al modelo información sobre la energía hidrotérmica que se ha de generar para cubrir la demanda en cada hora.

En el siguiente capítulo se presentan las variables exógenas que se consideran más importantes para la predicción de precios de la electricidad, para finalmente preparar el conjunto de datos de cara al entrenamiento.

## 2. ANÁLISIS Y PREPARACIÓN DE LOS DATOS

Las variables que se han considerado como más importantes o que pueden tener más influencia en el precio de la electricidad del mercado diario son:

- Demanda eléctrica
- Generación de energía renovable no gestionable (solar y eólica)
- Generación de energía renovable gestionable (hidráulica)
- Nivel de los embalses
- Consumo eléctrico de centrales hidráulicas reversibles o de bombeo
- Generación de energía térmica
- Precio del gas
- Relativas al calendario (hora del día, día de la semana, mes, etc)

Cabe resaltar que la generación solar y eólica no son gestionables o regulables y dependen de las condiciones meteorológicas (radiación solar, temperatura, velocidad del viento, humedad...). Es decir, en función de estas variables, se generará una determinada cantidad de electricidad en cada momento que se inyectará directamente a la red, sin tener ningún tipo de control sobre ella (salvo en el caso de que la generación eléctrica sea mayor que la demanda, en cuyo caso se verterá el exceso de electricidad).

También hay que señalar que en el corto plazo (para el día siguiente o incluso a varios días vista), las previsiones meteorológicas son bastante precisas y por tanto también lo son las estimaciones de generación solar y eólica. Es decir, se puede aproximar con gran precisión la electricidad que las plantas de este tipo van a generar al día siguiente.

Del mismo modo, la demanda eléctrica se puede estimar con bastante precisión si se conocen las previsiones meteorológicas, pues la curva diaria de consumo eléctrico suele seguir unos patrones conocidos que son función de variables tales como la temperatura y la humedad, así como del mes y día de la semana.

Por otro lado, la generación hidráulica tiene una parte no gestionable (“fluyente”) y otra gestionable o regulable. Ambas dependen también de las condiciones climatológicas (de la cantidad de lluvia principalmente). Tanto la generación hidráulica regulable como el consumo de las centrales de bombeo hidráulico van a tener una fuerte correlación con el precio de la electricidad, ya que cuando el precio sea alto, el consumo del bombeo será bajo o nulo, mientras que, si el precio es bajo, la generación hidráulica será baja y el consumo del bombeo alto. Sin embargo, no es posible conocer con exactitud la cantidad de energía que estas tecnologías generarán (o consumirán) a futuro, pues va a depender de las casaciones de mercado. Por tanto, sólo se podrán usar los valores históricos de estas variables.

Es por ello, que se utiliza otra variable con el fin de tener una idea de la capacidad disponible de generación hidráulica. Esta variable no es otra que el nivel o la capacidad en la que se encuentran los embalses en España, medido en  $\text{hm}^3$ .

Finalmente, el precio del gas natural va a tener una gran influencia en el precio de la electricidad en España, al ser este un mercado marginalista en el que el precio de la

electricidad lo marca el coste de la última tecnología en generar electricidad. Por tanto, las centrales de ciclo combinado, que utilizan gas natural y van a ser necesarias en muchas ocasiones para cubrir la demanda, serán las que marquen el precio en estas horas.

También se recogen los datos de generación térmica o de ciclos combinados, pero una vez más, esta variable dependerá de las casaciones de mercado y es desconocida a futuro. Por ello, se crea una nueva variable a partir de otras conocidas, llamada “hueco hidrotérmico”, con el fin de dar información sobre esa posible cantidad de energía térmica. Esta variable se explicará más adelante.

Los datos que se van a utilizar son todos de libre acceso y provienen de fuentes públicas. La mayor parte provienen del portal de transparencia (esios) del operador del sistema eléctrico español (Red Eléctrica Española). En concreto, se descargan las variables con los siguientes nombres:

- Precio medio horario componente mercado diario
- Generación medida solar fotovoltaica
- Generación medida eólica
- Generación medida consumo bombeo
- Generación medida hidráulica
- Generación medida ciclo combinado
- Generación medida energía nuclear (sólo se utilizará para la construcción de la gráfica de cobertura de la demanda)
- Demanda real

Por otro lado, el precio del gas se obtiene de la página web de MIBGAS (Mercado Ibérico del Gas), donde se tienen ficheros anuales de los resultados de compras de gas con su precio y volumen.

Finalmente, el nivel de los embalses se descarga de la web EpData (Europa Press). En esta web se recogen datos de esta variable desde 1988 de forma semanal, tal y como es publicado por el Ministerio de Agricultura y Pesca, Alimentación y Medio Ambiente

Se descargan los datos de todas estas variables desde que se tienen registros en el portal de Red Eléctrica. Se obtienen, por tanto, datos horarios desde 2015 hasta 2023, 9 años completos, lo que hacen un total de 78840 filas y 9 columnas. En el caso del precio del gas, los datos tienen una frecuencia diaria, por lo que simplemente se rellenan todas las horas de un día con el mismo valor de ese día para obtener frecuencia horaria. Del mismo modo, los datos del nivel de los embalses tienen frecuencia semanal, por lo que se hace una interpolación a valores horarios.

Estos datos se descargan en formato “csv” y se cargan en un DataFrame, al que se le pone la fecha como índice. En la Tabla 1 se puede apreciar el aspecto del DataFrame con los datos recopilados. Cabe indicar que las unidades de las variables de generación y demanda, así como el precio de la electricidad, son MW<sub>e</sub>, mientras que el precio del gas es en €/MWh<sup>1</sup>.

---

<sup>1</sup> El subíndice “t” hace referencia a MWh térmicos, mientras que se utiliza el subíndice “e” o simplemente sin subíndice para referirse a MWh eléctricos

## 2. ANÁLISIS Y PREPARACIÓN DE LOS DATOS

---

	demand	gener_ccgt	gener_solar	gener_eolica	agua_emb	gener_hidraulica	consumo_bombeo	precio_gas	precio
2015-01-01 00:00:00	25435.000000	3416.260010	0.047	6283.000000	39291.714286	2955.143799	-833.476990	23.30	50.099998
2015-01-01 01:00:00	24511.500000	3775.267090	0.050	5788.224121	39292.523810	2701.602539	-891.435974	23.30	48.099998
2015-01-01 02:00:00	22866.166016	3452.924072	0.045	5368.730957	39293.333333	2370.163574	-1138.496948	23.30	47.330002
2015-01-01 03:00:00	21392.833984	2907.243896	0.051	5150.655762	39294.142857	1791.629395	-1477.296021	23.30	42.270000
2015-01-01 04:00:00	20319.666016	2721.058105	0.043	4835.700195	39294.952381	1716.074951	-1804.074951	23.30	38.410000
...	...	...	...	...	...	...	...	...	...
2023-12-31 19:00:00	29135.666016	4254.813965	5.924	6268.881836	25803.517857	8320.655273	-1.405000	31.08	101.300003
2023-12-31 20:00:00	28703.500000	4173.980957	5.850	6062.973145	25802.071429	8363.883789	-1.467000	31.08	92.970001
2023-12-31 21:00:00	27423.166016	4304.809082	5.718	5589.799805	25800.625000	7587.621582	-1.455000	31.08	85.489998
2023-12-31 22:00:00	24613.916016	3340.175049	1.260	5140.062012	25799.178571	6828.814453	-1.398000	31.08	74.739998
2023-12-31 23:00:00	23074.500000	3504.464111	1.050	5155.777832	25797.732143	5553.238281	-136.904999	31.08	71.959999

78888 rows x 9 columns

Tabla 1: Conjunto de datos

Hay que señalar que no se tiene ningún valor nulo, y que en el proceso de carga se ha impuesto que todas las variables sean del tipo ‘float32’.

En la siguiente sección se lleva a cabo la exploración del conjunto de datos, para posteriormente desarrollar nuevas variables, como el “hueco térmico”, que se explicará posteriormente, y las relativas al calendario: mes del año, día de la semana, hora y día laborable.

### 2.1. Análisis exploratorio de las variables

A continuación, se va a realizar el análisis y exploración de los datos. En la Tabla 2 se puede apreciar un resumen de los principales estadísticos de las variables, lo cuál sirve para hacerse una primera idea sobre los valores que pueden tomar.

De esta tabla se aprecia, por ejemplo, que el consumo del bombeo toma valores negativos. Al ser un consumo en lugar de una generación se representa con valores negativos. Sin embargo, lo que llama la atención es que el valor máximo es positivo, lo cuál no tendría sentido físico. También se puede observar que los valores máximos que del precio de la electricidad y del gas están muy alejados de los valores medios.

	demand	gener_ccgt	gener_solar	gener_eolica	agua_emb	gener_hidraulica	consumo_bombeo	precio_gas	precio
count	78888.0	78888.0	78888.0	78888.0	78888.0	78888.0	78888.0	78888.0	78888.0
mean	27923.5	4278.0	1762.5	6067.4	30033.2	3442.2	-537.4	32.8	71.9
std	4570.8	2951.4	2911.4	3639.4	6218.8	2034.6	873.6	30.9	55.9
min	16116.8	244.2	-0.6	127.8	17599.0	470.0	-4538.0	4.2	0.0
25%	24183.8	2051.6	0.3	3205.5	25328.8	1836.8	-751.9	16.2	41.2
50%	27898.7	3332.6	59.9	5353.1	29656.8	2950.7	-86.8	22.1	53.7
75%	31386.8	5885.8	2573.2	8279.9	33948.2	4639.5	-1.8	34.1	77.9
max	41828.8	17579.4	16062.1	20886.6	44192.0	11657.5	80.6	225.0	700.0

Tabla 2: Principales estadísticos de las variables

En los siguientes apartados se van a analizar cada una de las variables por separado, para posteriormente estudiar sus correlaciones.

### 2.1.1. Demanda

Se comienza visualizando cuál ha sido la demanda eléctrica anual en España en los últimos años (Figura 1). Obsérvese que el eje varía entre los 200 y los 260 TWh.

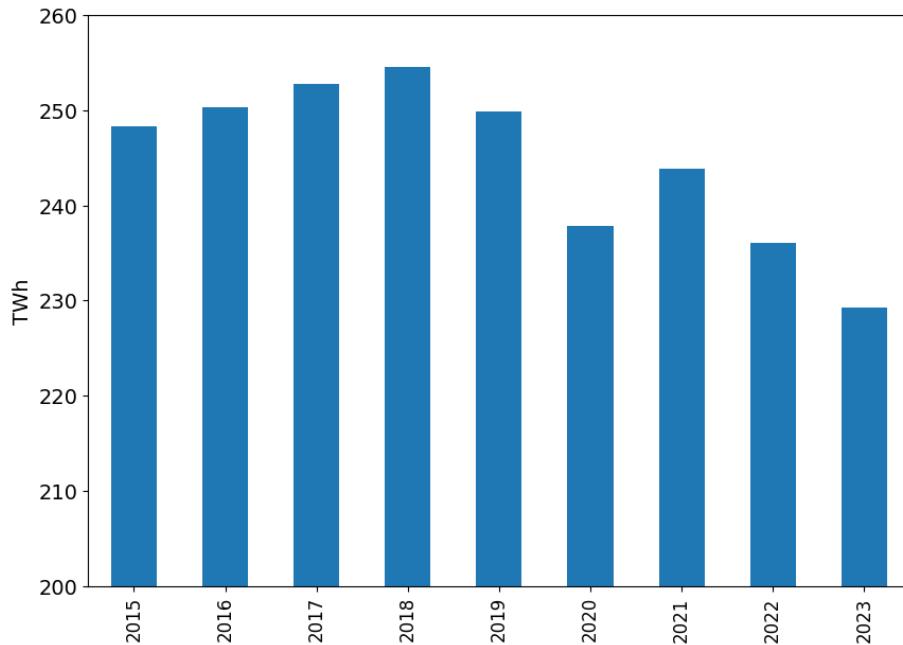


Figura 1: Evolución de la demanda eléctrica anual

Se aprecia que el máximo de demanda anual se encuentra en el año 2018, con unos 254 TWh y el mínimo en 2023, con 229 TWh. Además, de este gráfico se sacan varias conclusiones:

- El consumo anual de electricidad tiene una tendencia al alza en torno al 0.8% hasta 2018
- A partir del año 2018 la tendencia se revierte y se empieza a reducir la demanda eléctrica, en torno a un -3% los últimos años.
- En 2020 se aprecia una clara reducción de la demanda como consecuencia del cese de actividad y confinamiento debido al COVID-19

Por otro lado, en la Figura 2 se muestra la evolución temporal horaria de la demanda durante el año 2023 junto con su media móvil semanal (“rolling mean”).

El máximo de demanda se encuentra en febrero, alcanzándose un pico de unos 38 GW, mientras que el mínimo tiene lugar en abril, con unos 17 GW.

## 2. ANÁLISIS Y PREPARACIÓN DE LOS DATOS

---

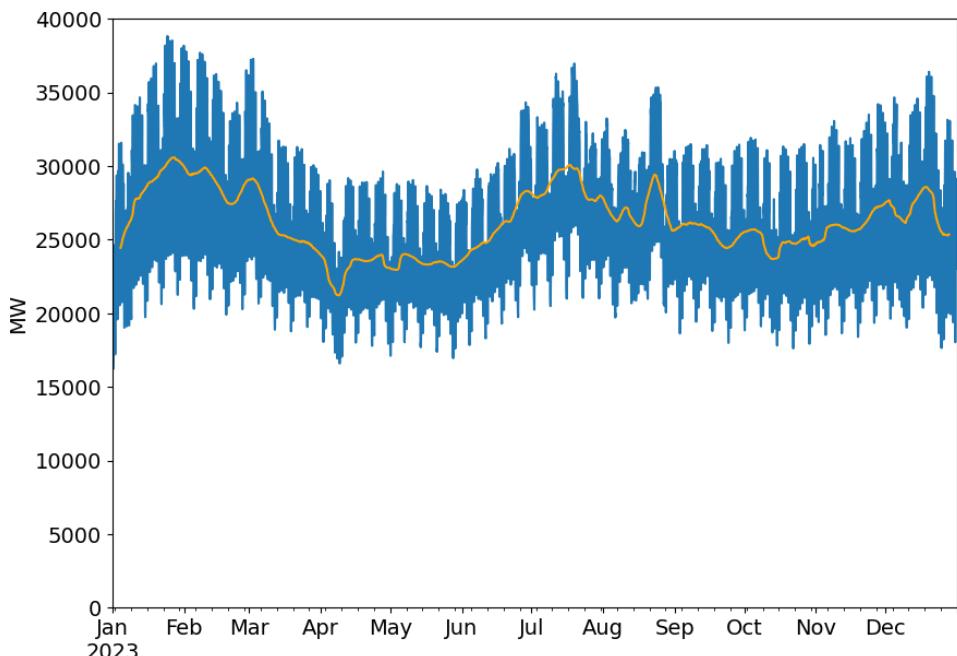


Figura 2: Evolución de la demanda eléctrica de 2023

Se aprecia una clara estacionalidad en la demanda. En los meses de invierno y de verano, debido a las temperaturas se produce mayor consumo de electricidad. Mientras que, en primavera y otoño, que las temperaturas son más agradables, se tienen menores consumos. También se observan, por este mismo efecto de las temperaturas, variaciones repentinas en determinadas semanas en las que probablemente haya habido una desviación respecto de la temperatura habitual de la época del año.

Además, se puede observar una gran variación entre valores para cualquier momento del año. Esto es debido a que la demanda eléctrica es muy dependiente de la hora del día, como se verá en el siguiente apartado, concretamente en las Figuras 5 y 6.

### 2.1.2. Precio de la electricidad y del gas

Estas dos variables se van a analizar de forma conjunta, ya que como se verá a continuación, están muy correlacionadas.

En la Figura 3 se muestra la evolución temporal de ambos precios, y en el caso de la electricidad se muestra también la media móvil semanal. Nótese que la escala y las unidades son las mismas ( $\text{€}/\text{MWh}$ ), aunque en el caso del gas se refiere a  $\text{MWh}$  térmicos.

Se observan precios muy estables desde 2015 hasta 2020, oscilando entre 0 y 100  $\text{€}/\text{MWh}$  y una media de unos 50  $\text{€}/\text{MWh}$  en el caso de la electricidad, mientras que el precio del gas natural se muestra muy estable en torno a los 20  $\text{€}/\text{MWh}$ . Estos valores medios se aprecian más claramente en la Figura 4, donde se muestran las medias anuales.

En 2020 se aprecia un descenso de los precios, por la disminución en la demanda a causa de la crisis del COVID-19, mientras que a partir del 2021 comienza un aumento paulatino que termina disparándose a principios de 2022 como consecuencia de la invasión de Ucrania por parte de Rusia. Este suceso desencadenó el aumento desorbitado del precio

del gas natural, que a su vez hizo que el precio de la electricidad en el mercado mayorista aumentara drásticamente, llegando a alcanzar los 700 €/MWh.

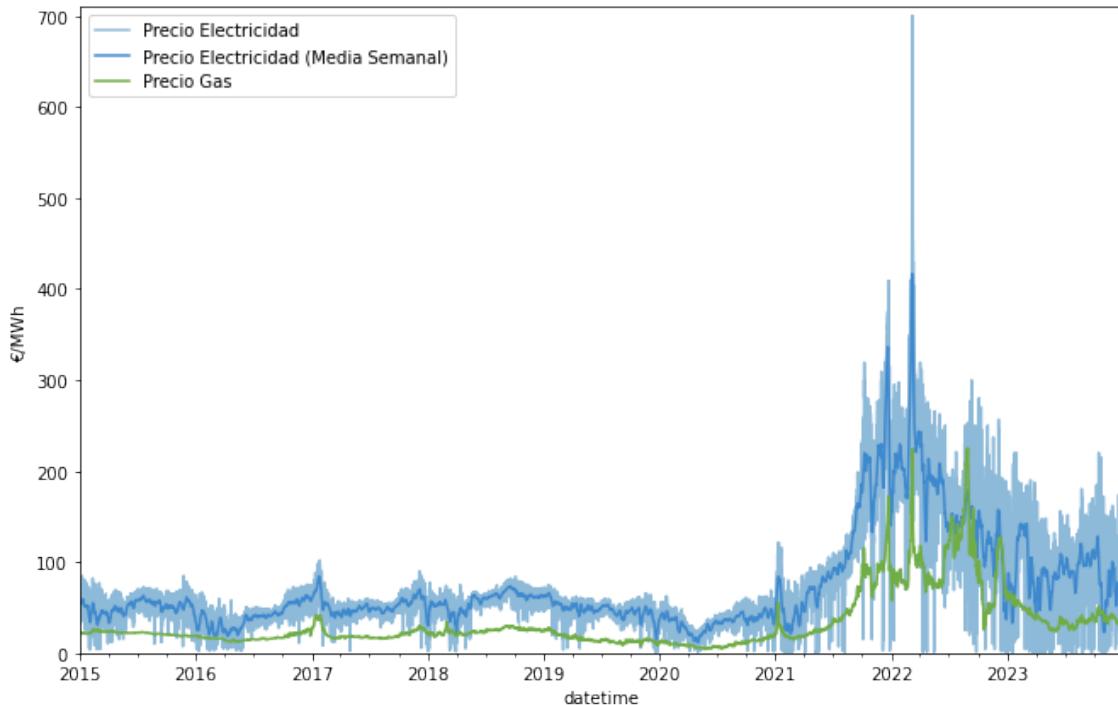


Figura 3: Evolución temporal de los precios

Hay que recordar una vez más, que las centrales de ciclo combinado queman gas natural y que, al ser un mercado marginalista, si con las otras tecnologías no se puede cubrir la demanda, los ciclos combinados tendrán que generar energía para llegar a cubrirla y serán los que marquen el precio. Cuando el coste del combustible es muy elevado, también lo serán, por tanto, sus costes variables, y tendrán que ofertar la electricidad muy cara para, como mínimo, poder cubrir esos costes.

Por ello, se observa siempre una clara correlación entre gas y electricidad: el precio de la electricidad, siempre que el ciclo combinado es la tecnología marginal, es, como mínimo, el doble que el precio del gas, ya que el rendimiento de un ciclo combinado es aproximadamente un 50%. A parte del coste del combustible, se tendrían que añadir otros costes variables (emisión CO<sub>2</sub>, costes de arranque, O&M...).

La excepción se produce cuando en la segunda mitad de 2022 el Gobierno implanta el mecanismo llamado “Excepción Ibérica”, a través del cuál se desacoplaba el precio de la electricidad con el precio del gas limitando este último. Esta medida finalizó en febrero de 2023, debido a la estabilización de los precios.

## 2. ANÁLISIS Y PREPARACIÓN DE LOS DATOS

---

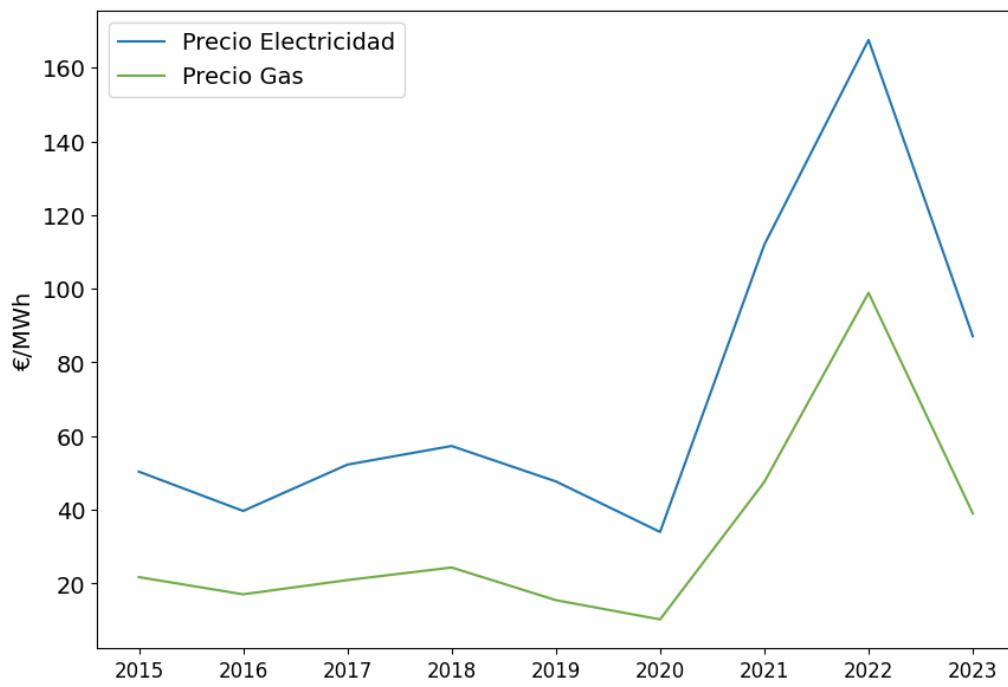


Figura 4: Precios medios anuales

Por otra parte, para comprender la relación entre precio y demanda con las horas del día, se analiza un día promedio de invierno y otro de verano (Figuras 5 y 6). En invierno se observa claramente cómo el precio varía acorde con la demanda. Además, en las horas centrales del día, a causa de la producción de energía solar, el precio disminuye considerablemente. De hecho, en verano, aunque las mayores demandas se encuentran en estas horas centrales, debido a la gran producción solar, el precio disminuye igualmente.

Sin embargo, esta curva de precios de un día medio en verano es algo distinta entre 2015 y 2019, ya que la potencia instalada de energía solar fotovoltaica era mucho más pequeña.

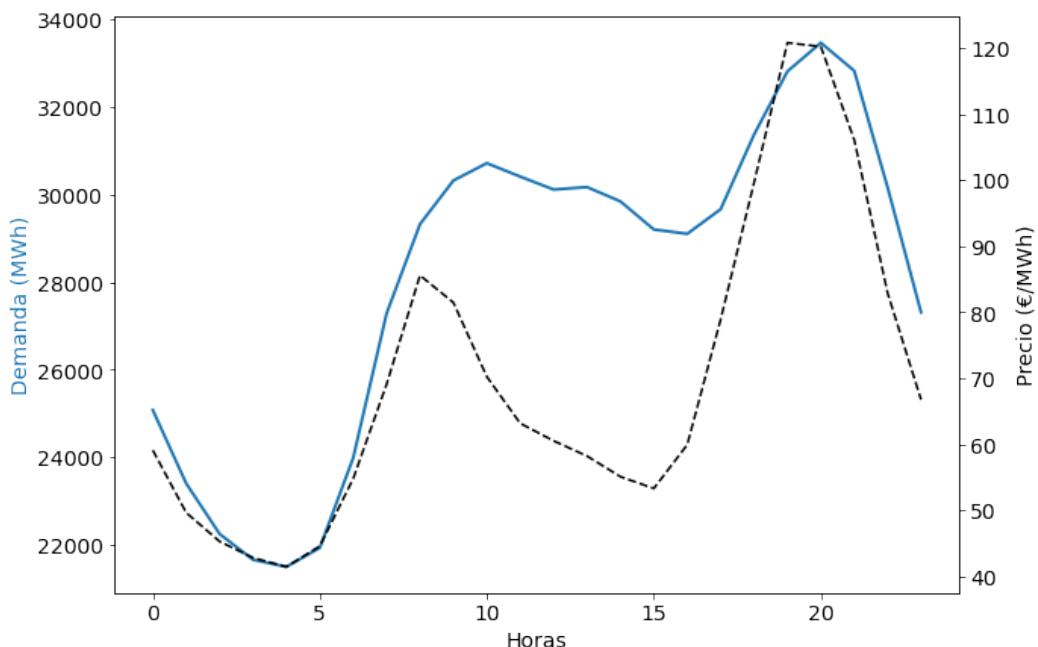


Figura 5: Día promedio de invierno (enero 2023)

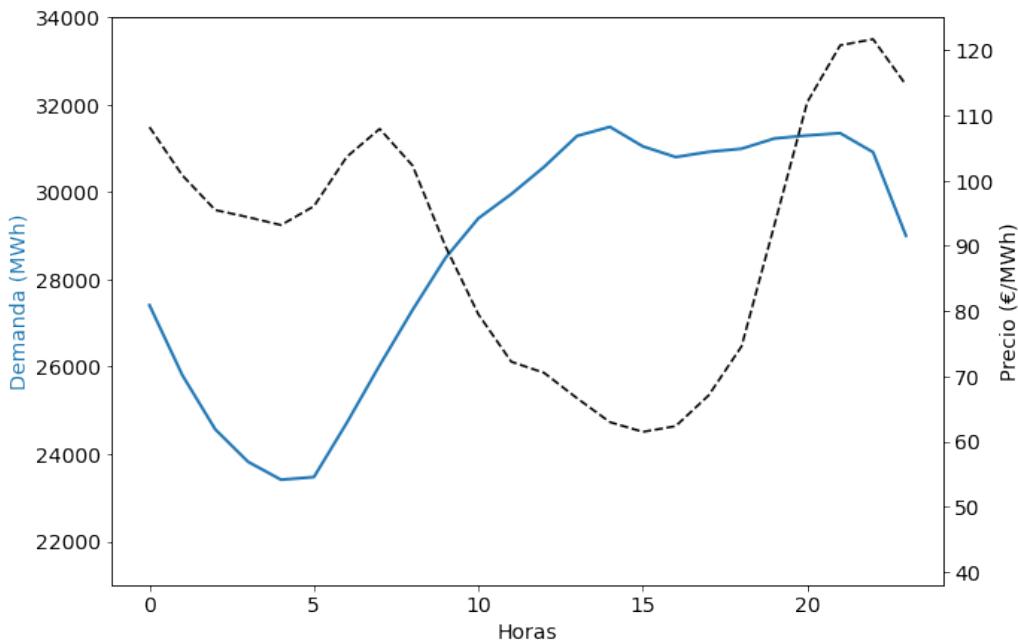


Figura 6: Día promedio de verano (julio 2023)

### 2.1.3. Generación de energía renovable no gestionable: solar fotovoltaica y eólica

Para ilustrar lo comentado en la Introducción sobre al gran aumento en la potencia instalada de las energías renovables, se muestra en la Figura 7 la media móvil mensual de generación solar fotovoltaica, donde se aprecia claramente este incremento a partir de 2019. También se puede observar las diferencias de generación que hay entre los meses de invierno y de verano.

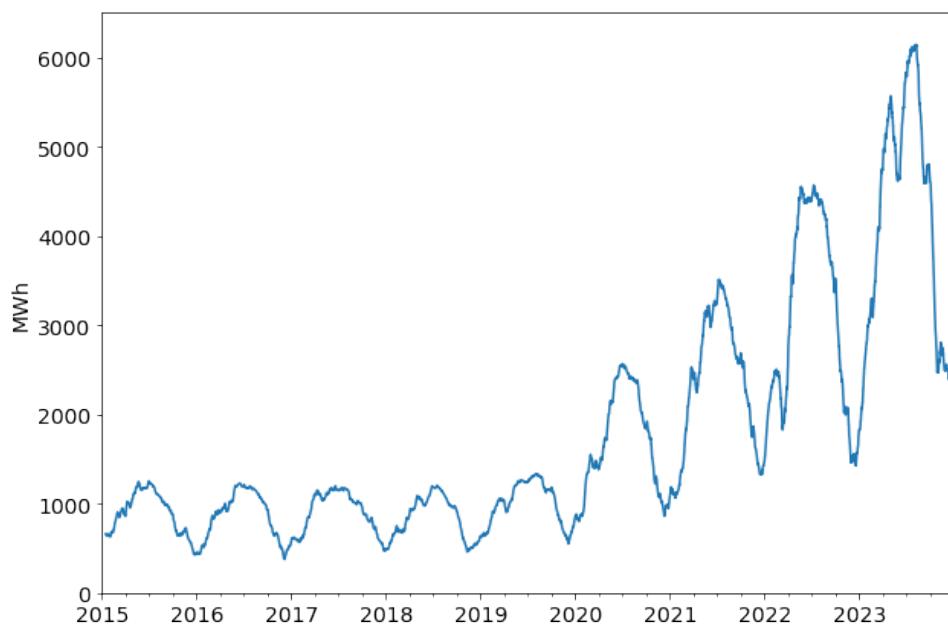


Figura 7: Evolución temporal de la generación de energía solar fotovoltaica

## 2. ANÁLISIS Y PREPARACIÓN DE LOS DATOS

Estas diferencias también se aprecian en la Figura 8, donde se muestra el perfil diario promedio de generación en verano y otro en invierno para el año 2023. Nótese que la curva de generación en julio no sólo es más alta sino también más ancha, es decir hay más horas de sol y con mayor radiación.

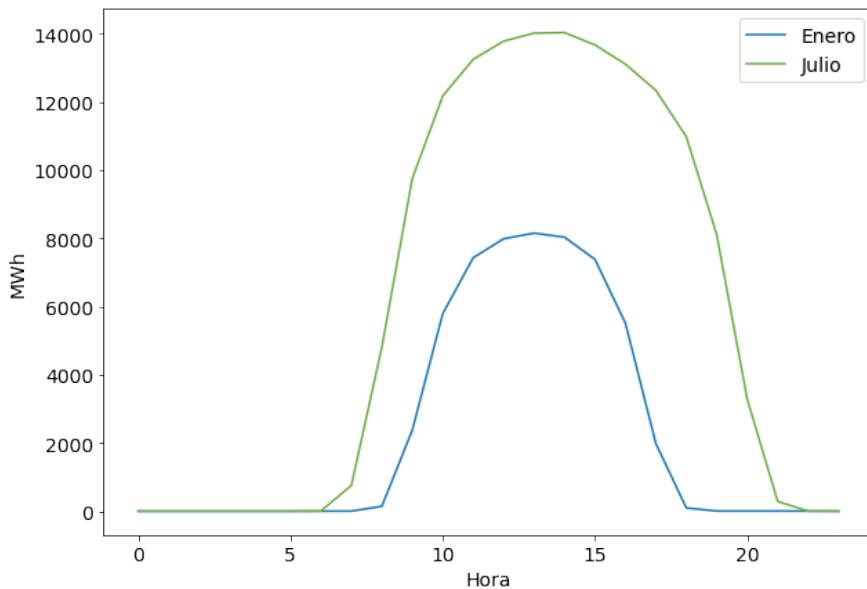


Figura 8: Generación FV de un día promedio de verano y otro de invierno (2023)

Por otro lado, la energía eólica también tiene estacionalidad, aunque no tan marcada como en el caso de la energía fotovoltaica (ver Figura 9). Se observa una mayor generación en los meses de invierno y menor en los de verano.

Además, a diferencia de la solar, la generación eólica apenas depende de la hora del día.

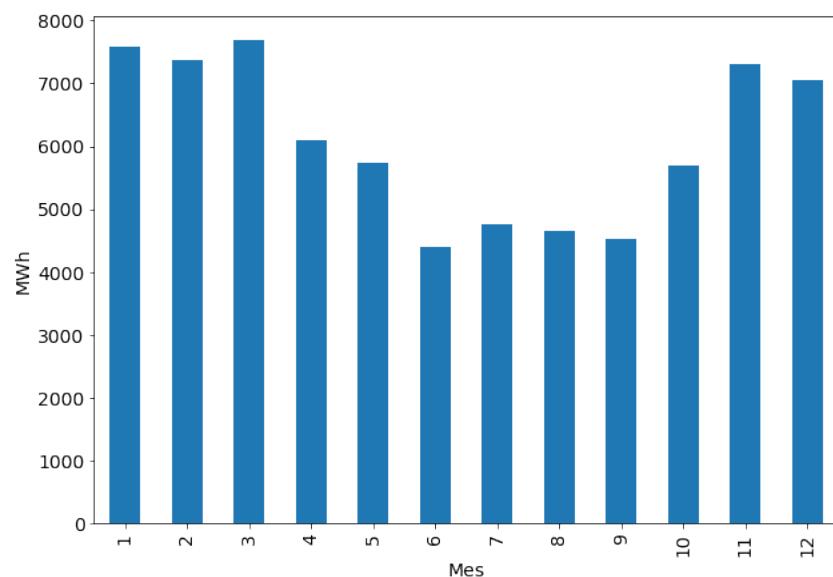


Figura 9: Generación eólica horaria por meses (media de todos los años)

#### 2.1.4. Generación y bombeo hidráulico

La energía hidráulica es otro tipo de energía renovable que depende de las condiciones climatológicas, principalmente de la cantidad de agua de lluvia. Se diferencia de la solar y la eólica en que puede almacenarse en embalses y, por tanto, se puede gestionar (no toda, ya que hay una parte del agua que no se puede regular, la denominada hidráulica fluyente, que depende entre otras cosas de los caudales mínimos que han de llevar los ríos, o restricciones de los embalses).

Por tanto, la generación hidráulica va a variar a lo largo del año en función del agua almacenada en los embalses, que a su vez depende de la cantidad de lluvia. Si se analiza la cantidad total anual de generación hidráulica se puede determinar si ha sido un año “seco” o “húmedo”. Observando la Figura 10, se puede establecer como años hidráulicos medios aquellos que se encuentran entre los 30 y 35 TWh, es decir 2015, 2020, 2021 y 2023

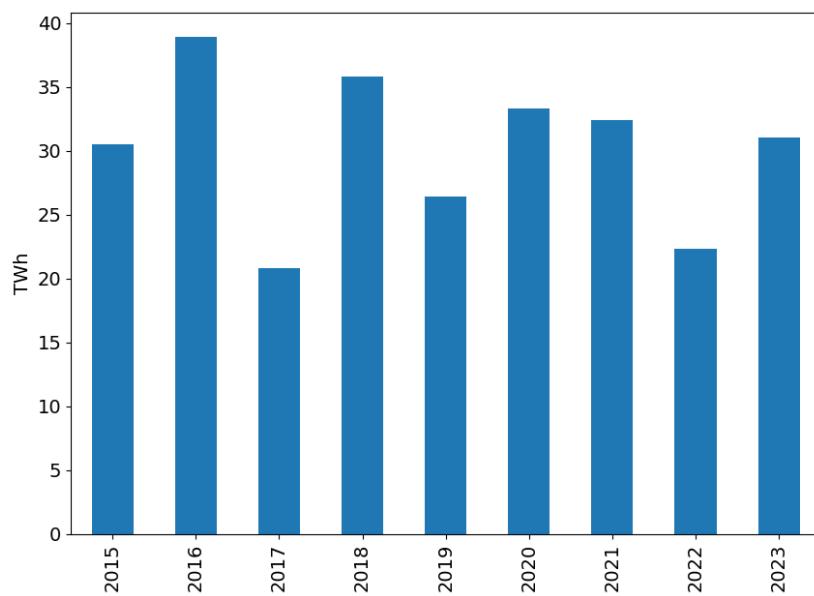


Figura 10: Generación total hidráulica anual

La diferencia de generación hidráulica entre estaciones del año se puede contemplar en la Figura 11, donde se ha representado esta generación diaria de un mes de verano y otro de invierno. Analizando todos los meses del año, se puede concluir que en invierno y primavera la generación será mayor, mientras que el mínimo se alcanza hacia el final del verano.

Por otro lado, algunas de las centrales de generación hidráulica también tienen la capacidad de operar en sentido contrario, bombeando agua de un lago o reserva de agua a una altura inferior a otra reserva superior para su posterior utilización en la generación de energía. De este modo, se bombeará agua en horas de precios bajos, habitualmente de baja demanda, y se turbinará en horas de precios altos (véase de nuevo la Figura 11).

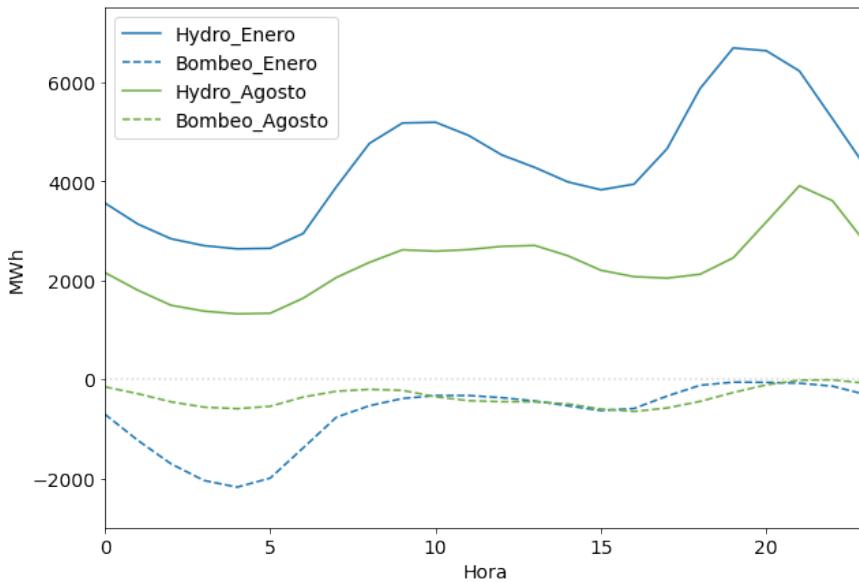


Figura 11: Generación y bombeo hidráulico de un día promedio de enero y agosto (media de todos los años)

Sin embargo, como ya se mencionó anteriormente, estas dos variables son desconocidas a futuro, por lo que sólo se podrán utilizar sus valores históricos.

### 2.1.5. Cantidad de agua embalsada

Para proveer al modelo de información sobre la capacidad hidráulica disponible para generar electricidad, se extraen datos del nivel o capacidad de los embalses en  $\text{hm}^3$ , agregado para todo el país.

En la Figura 12 se representa la capacidad disponible media anual de los embalses. Se observa, comparando con la Figura 10, que los años con mayor generación hidráulica concuerdan aproximadamente con los años con el nivel de los embalses más elevado.

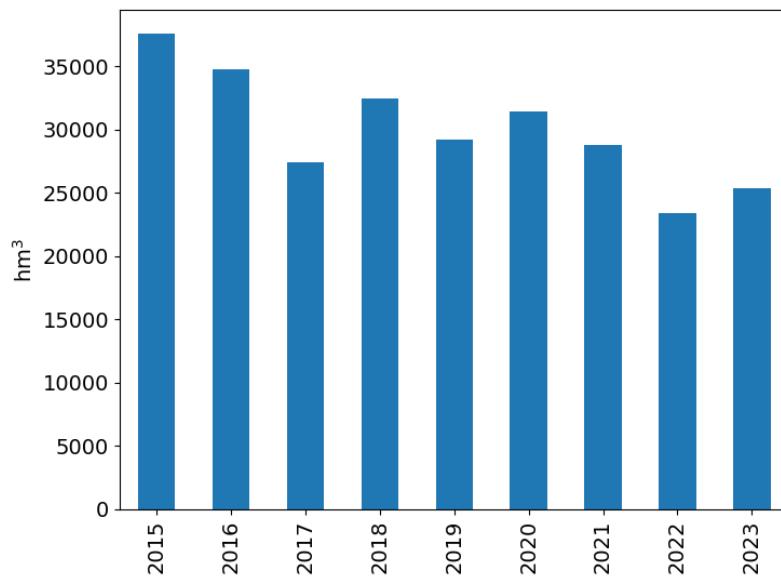


Figura 12: Nivel medio anual de los embalses

Por otro lado, en la Figura 13, se muestra una comparativa entre la distribución mensual promedio del nivel de los embalses y de la generación hidráulica.

Se puede apreciar una clara correlación, aunque con un cierto desfase temporal, del nivel de los embalses respecto de la generación hidráulica, si bien el mínimo de ambas variables se alcanza en el mes de octubre.

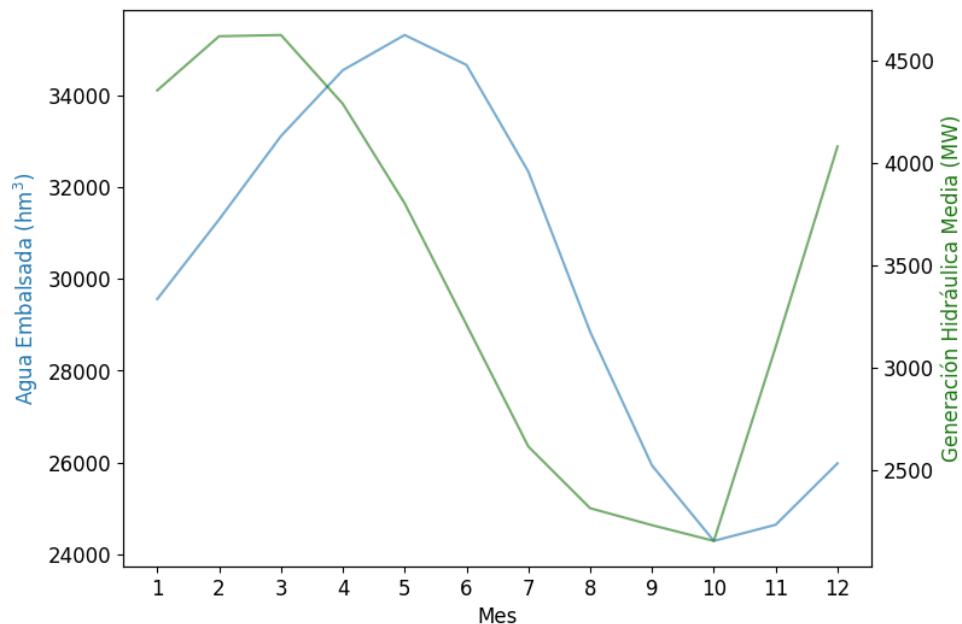


Figura 13: Distribución mensual del nivel de los embalses y la generación hidráulica media

#### 2.1.6. Generación de Ciclos Combinados

También se ha introducido en el *dataset* la generación de energía de las centrales de ciclo combinado, que utilizan gas natural. Como se comentó anteriormente, sus costes variables van a estar muy influenciados por el coste del combustible. Por tanto, cuando sea necesario generar energía mediante estas centrales para cubrir la demanda y debido a sus mayores costes variables respecto a otras tecnologías, serán las que marquen el precio de la electricidad en muchos casos.

En la Figura 14 se puede observar como varía la generación de ciclo combinado a lo largo de un día medio para tres meses diferentes del año. En los meses de baja demanda y elevada generación de energía renovable, como en primavera, la generación media de ciclo combinado es menor, mientras que en verano e invierno aumenta.

Se aprecia también que la forma de las curvas es muy similar en todos los casos, con mayores producciones de energía coincidentes con los dos picos de demanda del día (mañana y noche).

## 2. ANÁLISIS Y PREPARACIÓN DE LOS DATOS

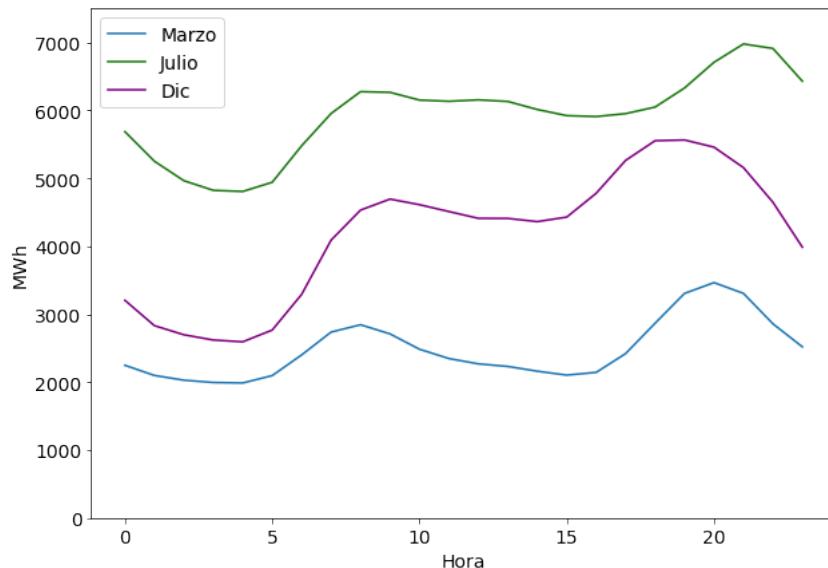


Figura 14: Generación diaria de un día medio para tres meses distintos (media de todos los años)

Por otro lado, aunque en España hay 24.5 GW de potencia instalada de ciclos combinados, a partir del histograma de la Figura 15 se observa que su generación suele ser mucho menor, situándose el valor más frecuente en torno a los 2500MW.

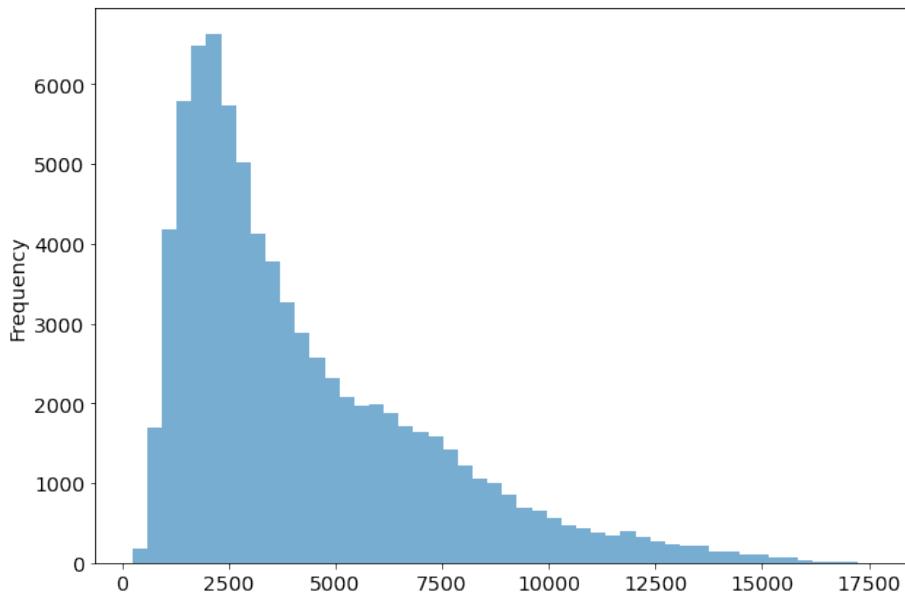


Figura 15: Histograma de la generación de ciclos combinados

### 2.2. Nuevas variables

#### 2.2.1. Hueco Térmico

Si bien la generación de energía de los ciclos combinados es una variable importante de cara a predecir el precio de la electricidad del día siguiente, sus valores futuros son

desconocidos pues dependerán de las casaciones de mercado. Por tanto, sólo se podrá hacer uso de valores históricos.

Por esta razón se crea una variable adicional llamada “hueco hidrotérmico” (o por simplicidad, simplemente “hueco térmico”), que no es más que la demanda menos la generación de fuentes no gestionables (solar y eólica). Se obtendría así la parte de la demanda que ha de cubrirse con la generación de ciclos combinados, energía nuclear y generación hidráulica principalmente.

Dado que la energía nuclear es prácticamente constante y muy inflexible, no va a tener mucha influencia en el precio. Esto se comprueba a posteriori introduciendo esta variable en los modelos. Por otro lado, la energía hidráulica, como ya se comentó anteriormente, tiene una parte no gestionable (“fluyente”) y otra regulable, que se “colocará” en las puntas de demanda, donde el precio será mayor, sustituyendo parte de la generación de ciclos combinados.

Por tanto, la variable hueco térmico servirá muy bien para capturar esa variabilidad de la energía térmica e hidráulica regulable, como se verá a continuación.

### 2.2.2. Resumen de variables: cobertura de la demanda

Para tener una visión global de los valores que toman las variables de generación descritas anteriormente y su relación con la demanda, se van a graficar las generaciones de todas las tecnologías en un gráfico de áreas apiladas junto con la demanda, para una semana de verano y una de invierno (Figuras 16 y 17). De esta forma, se puede observar cómo la generación de las diversas tecnologías satisface la demanda. También se añade el precio en el eje derecho de la gráfica.

Nótese que se ha añadido la generación de energía nuclear, que supone unos 7000MW prácticamente constantes. También hay que indicar que la suma de la generación considerada no llega a cubrir la demanda en la mayoría de las horas, mientras que en algunas otras la sobrepasa. Esto es debido a que no se han incluido otras tecnologías residuales como cogeneración, carbón o biomasa (que también generan en “carga base”, o son prácticamente constantes), así como los intercambios de energía con Francia, Portugal y Marruecos.

## 2. ANÁLISIS Y PREPARACIÓN DE LOS DATOS

---

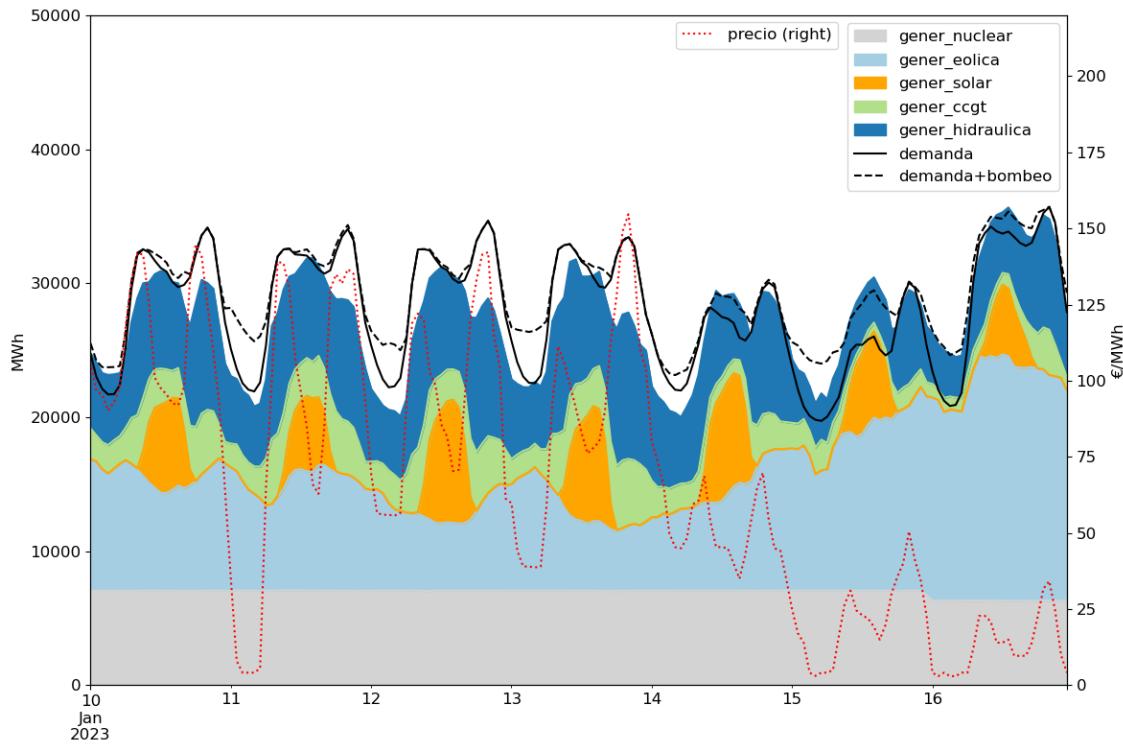


Figura 16: Cobertura de la demanda en una semana de invierno

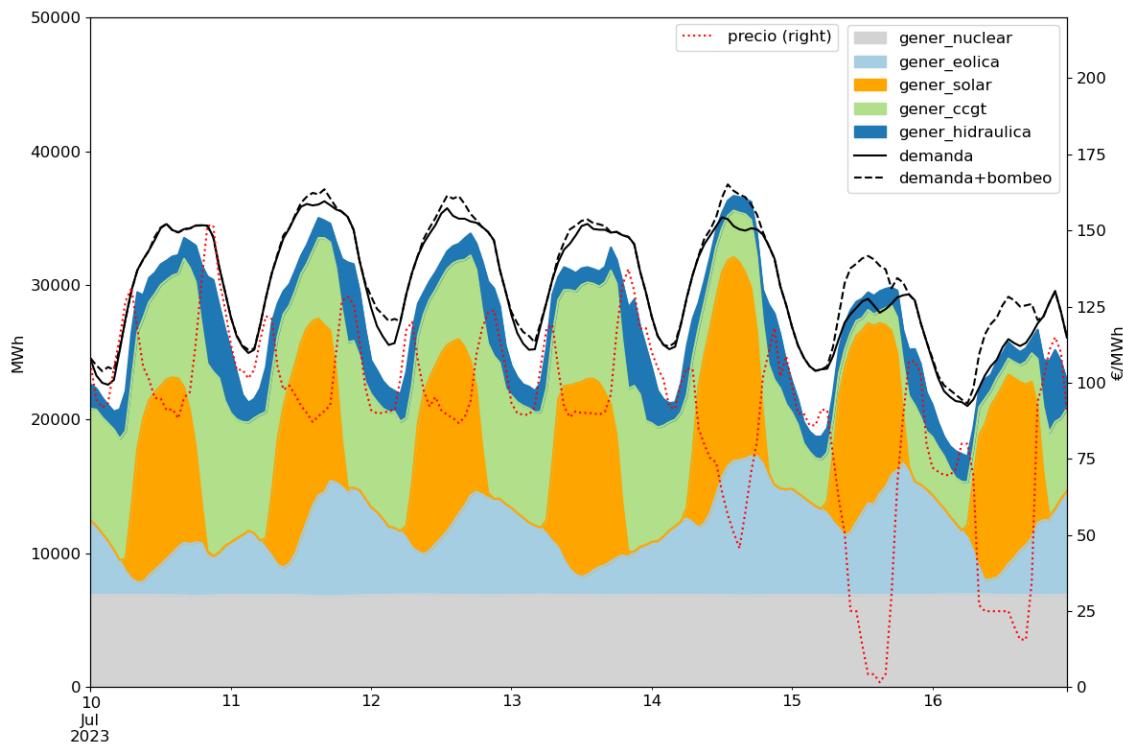


Figura 17: Cobertura de la demanda en una semana de verano

De las gráficas se observa que:

- La generación eólica e hidráulica es mayor en invierno, mientras que la solar es mayor en verano. Además, tienen una elevada variabilidad de un día a otro.

- La forma de la demanda es diferente en verano debido al elevado consumo eléctrico de los sistemas de aire acondicionado en las horas centrales del día.
- En verano hay una mayor generación de energía térmica debido a la menor generación eólica e hidráulica.
- En las horas en las que es necesario generar con ciclos combinados e hidráulica para cubrir la demanda, el precio es mayor.
- Por otro lado, cuando sólo con la generación solar, eólica, hidráulica fluyente y nuclear se cubre prácticamente la demanda, el precio cae hasta casi los 0 €/MWh.
- En estos casos de gran generación de energía renovable respecto de la demanda, el consumo de las centrales de bombeo es mayor.

Sin embargo, como ya se comentó anteriormente, los valores tanto de generación de ciclos combinados y de hidráulica como de consumo de bombeo, son desconocidos a futuro y sólo se podrán utilizar sus históricos. Es por esto que se crea la variable “hueco térmico” como diferencia entre la demanda y la generación solar y eólica (Figuras 18 y 19).

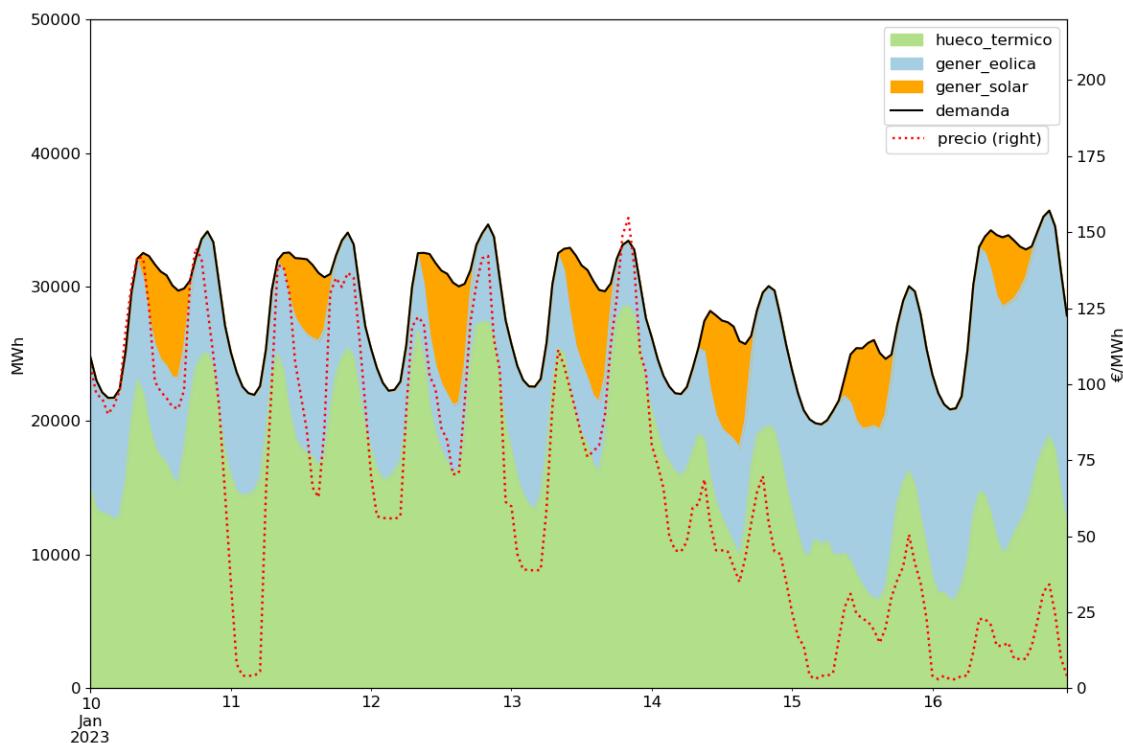


Figura 18: Cobertura de la demanda en una semana de invierno (hueco térmico)

## 2. ANÁLISIS Y PREPARACIÓN DE LOS DATOS

---

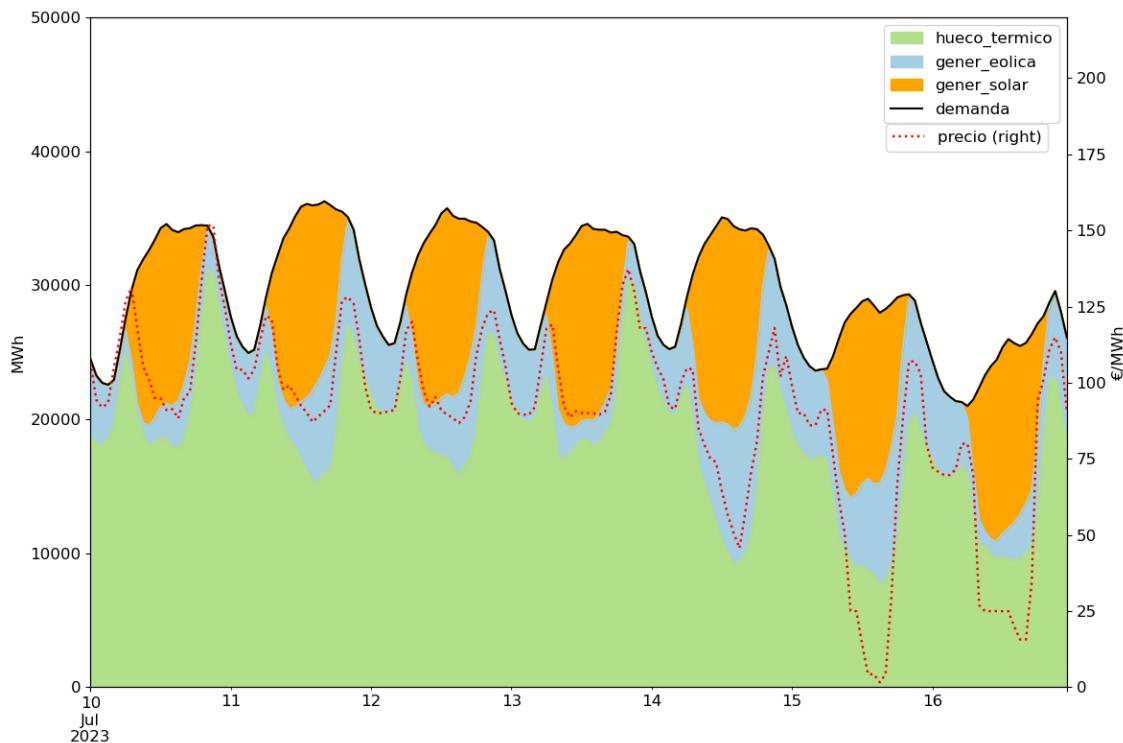


Figura 19: Cobertura de la demanda en una semana de verano (hueco térmico)

Se observa una clara relación entre el precio y el “hueco térmico”, por lo que esta variable probablemente vaya a ser útil en el modelo de cara a predecir los precios.

### 2.2.3. Variables relativas al calendario

Por último, también se van a incluir en el conjunto de datos las variables relativas al calendario, más concretamente:

- El mes del año: toma valores del 1 al 12, que se transforman a valores cílicos.
- El día de la semana: del 0 al 6
- La hora del día: de 0 a 23h, que se transforman a valores cílicos.
- Si el día es laborable o no: valores 0 o 1.

Tanto el mes como la hora se codifican como variables cílicas utilizando el seno y el coseno. De este modo, queda recogido que tras el último valor que toma la variable le sigue el primero (p.ej.: después de la hora 23 viene la 0). Por tanto, los valores que toma la variable “hora” serían los mostrados en la Figura 20.

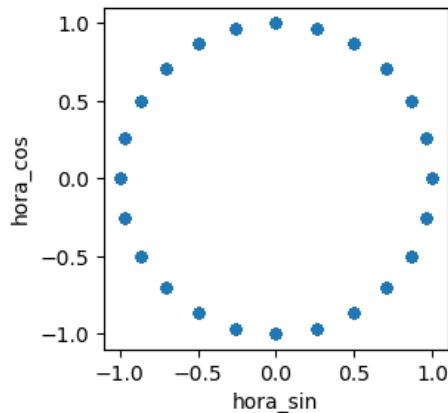


Figura 20: Codificación cíclica de la hora del día

Y del mismo modo se tendría la codificación cíclica para la variable “mes”.

### 2.3. Correlación entre variables

A continuación, se van a obtener las correlaciones entre el precio de la electricidad y el resto de las variables.

La matriz de correlaciones se representa en un mapa de calor para una mejor visualización (Fig. 21). Nótese que estas correlaciones son a nivel global, es decir para todos los valores de las variables desde 2015 a 2023.

Se aprecia una gran correlación del precio de la electricidad con el precio del gas y la generación de los ciclos combinados. Esto es bastante lógico porque esta tecnología suele ser en muchas ocasiones marginal, y el precio al que oferten la electricidad va a depender del coste del combustible (gas).

El resto de las variables a priori no parece tener mucha correlación con la variable objetivo al utilizar todos los valores del año (los 8760h), para todo el periodo considerado. Sin embargo, si se pone el foco en cómo se correlacionan las variables a lo largo del día, sí que se pueden apreciar valores mucho más elevados que indican una gran correlación entre el precio y el resto de las variables, tal y como se muestra en la Figura 22. Nótese que sólo se muestran las variables que evolucionan a lo largo del día.

Se puede apreciar una gran correlación del precio con la generación de ciclos combinados, la generación hidráulica, el consumo de bombeo y el hueco térmico.

De las anteriores, el resultado más relevante es el hueco térmico, ya que como se comentó anteriormente, las otras tres variables son desconocidas a futuro y sólo se podrán utilizar sus valores históricos. Por tanto, es de gran importancia el hecho de que la variable hueco térmico, construida a partir de la demanda y la generación de renovables, esté altamente correlacionada con el precio, ya que será muy útil de cara a predecir sus valores futuros.

También cabe destacar la correlación negativa del nivel de agua de los embalses con el precio. Es decir, cuando menor es el nivel de agua embalsada, y por tanto menor es la capacidad disponible para generar electricidad, mayor es el precio.

## 2. ANÁLISIS Y PREPARACIÓN DE LOS DATOS

---

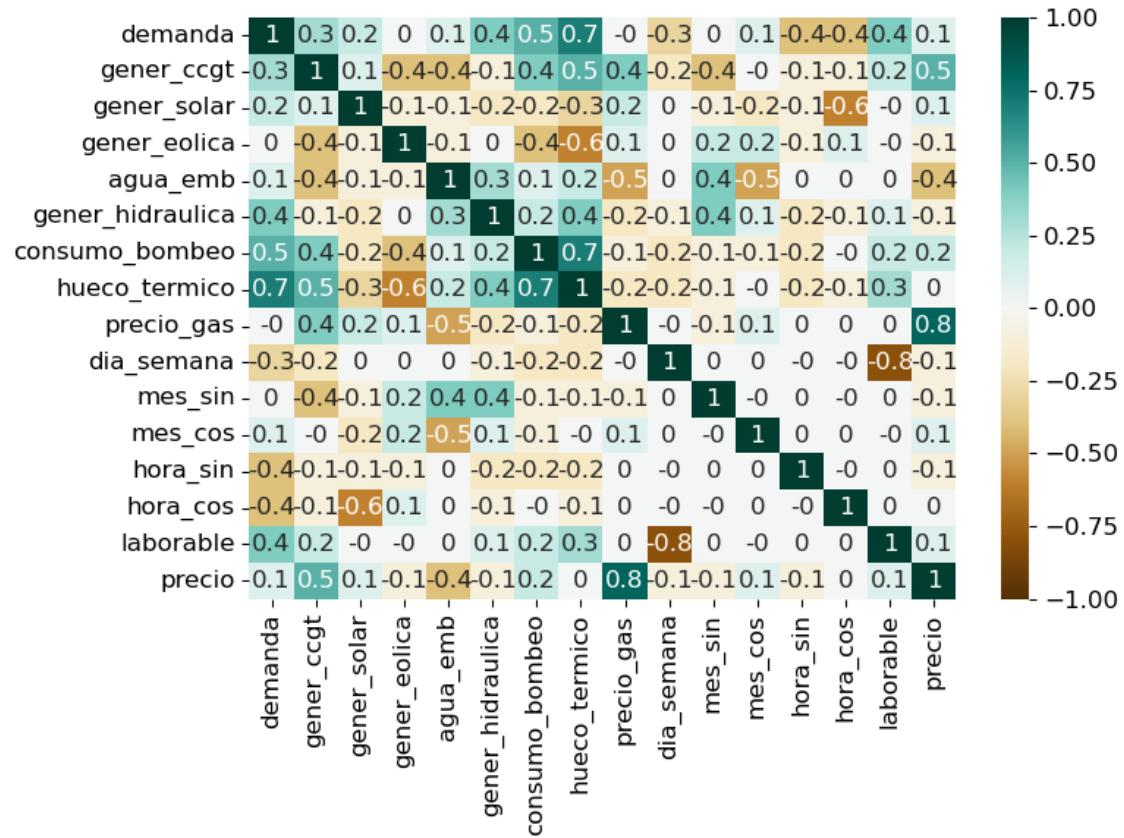


Figura 21: Matriz de correlaciones de las variables para todo el periodo considerado

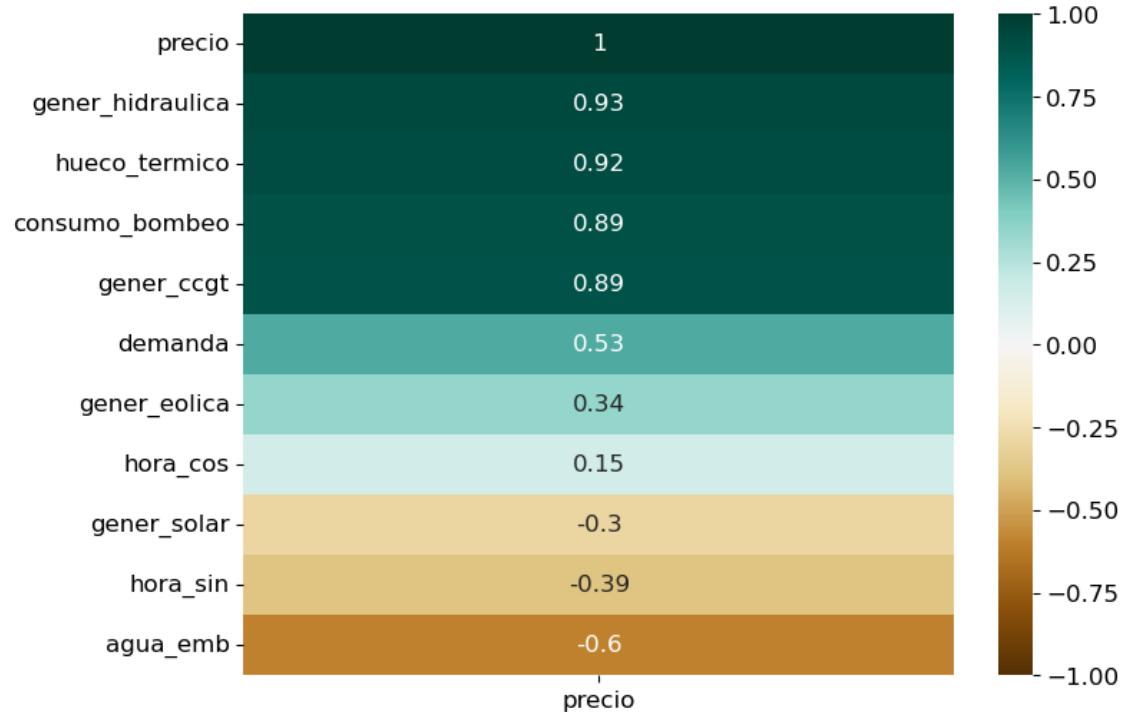


Figura 22: Correlaciones con la variable precio para el perfil diario (24h) promedio

Por otro lado, se muestra en la Figura 23 la autocorrelación de la serie temporal del precio de la electricidad. Se observa que el precio en una hora determinada ( $t$ ) está altamente correlacionada con los valores en  $t-1$ ,  $t-2$ ,  $t-24$  y  $t-25$ . En general, para una observación determinada, se tiene una cierta correlación con las 25 horas previas, y en menor medida, con las observaciones múltiplos de 24 (es decir 48, 72, 96, etc).

Por tanto, de cara al entrenamiento de los modelos, se considerarán los retardos múltiplos de 24 y el siguiente (24, 25; 48, 49; ...; 168, 169).

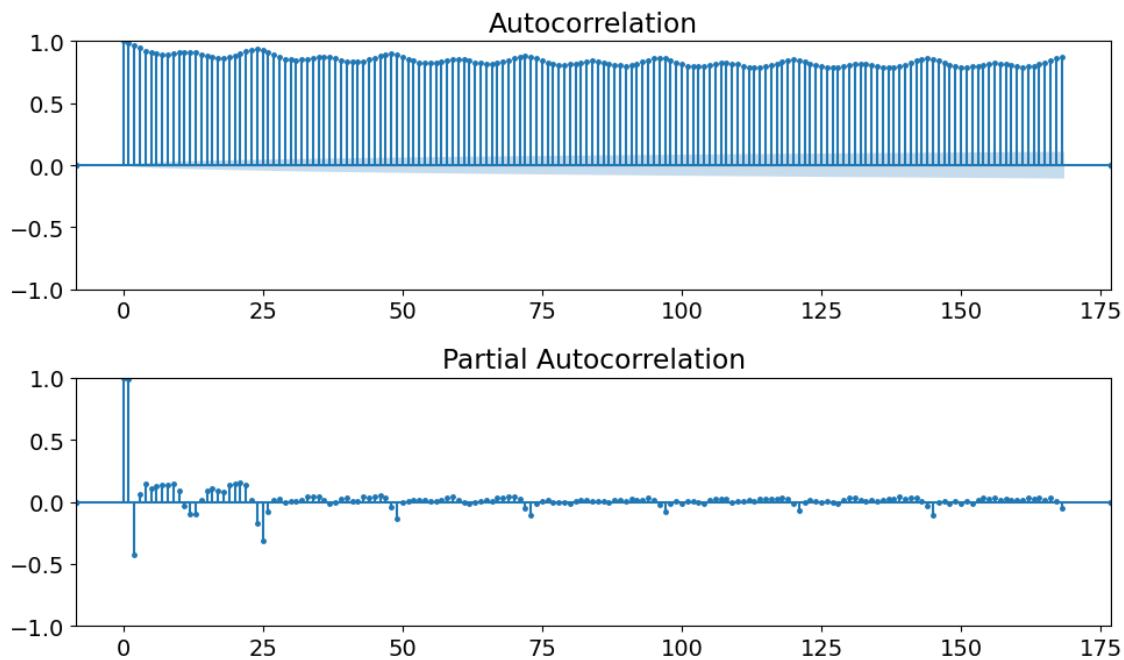


Figura 23: Autocorrelación y autocorrelación parcial de la variable objetivo

## 2.4. Descomposición de la variable objetivo y estacionariedad

A continuación, se muestra la descomposición de la serie temporal “precio” en sus diferentes componentes (Fig. 24)

En este gráfico se puede apreciar la gran variabilidad en el precio a partir de la segunda mitad de 2021.

## 2. ANÁLISIS Y PREPARACIÓN DE LOS DATOS

---

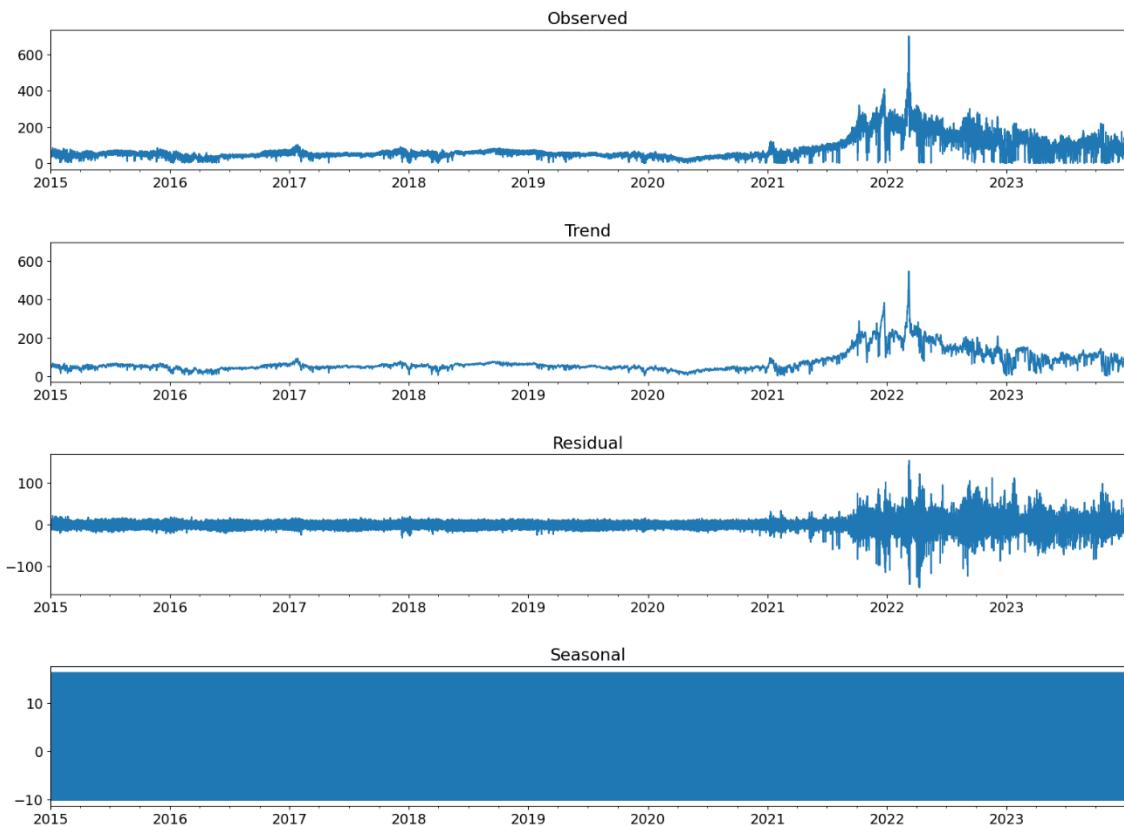


Figura 24: Descomposición de la serie temporal “precio de la electricidad”

Para poder apreciar la estacionalidad (“seasonality”) se muestra el mismo gráfico, pero centrado en un solo mes, en concreto en enero de 2023 (Figura 25). Aquí se pueden observar los patrones en el precio cada 24h y cada semana.

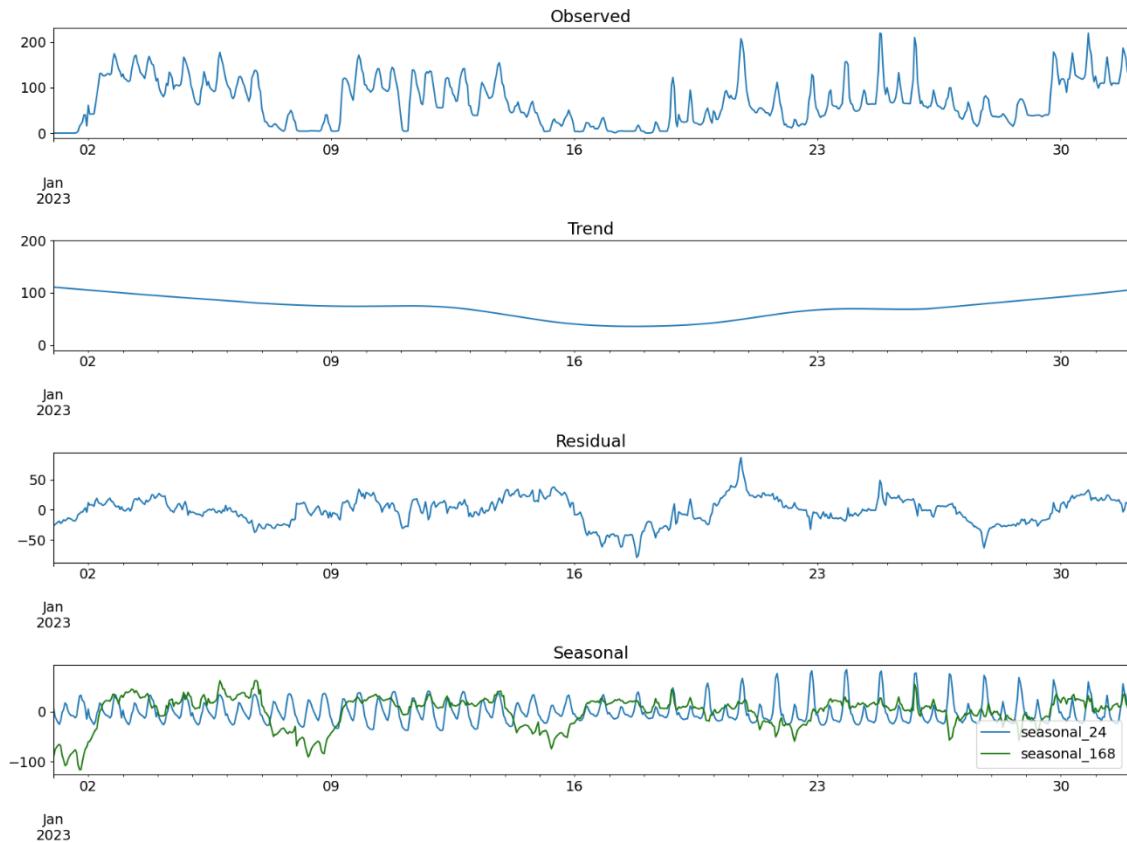


Figura 25: Descomposición de la serie temporal “precio de la electricidad” en enero de 2023

Por otro lado, se quiere comprobar si el precio de la electricidad es una serie temporal estacionaria, es decir si sus propiedades estadísticas (valor esperado, varianza, autocorrelación) no varían con el tiempo. Para ello, se van a implementar los test de estacionariedad ADF (Augmented Dickey-Fuller) y KPSS (Kwiatkowski-Phillips-Schmidt-Shin).

El test ADF es un tipo de test de raíz unitaria que determina si una serie temporal está definida por una tendencia. Las hipótesis son las siguientes:

- Hipótesis nula  $H_0$ : existe una raíz unitaria, la serie temporal está autocorrelada y no es estacionaria.
- Hipótesis alternativa  $H_1$ : la serie temporal no tiene una raíz unitaria y es estacionaria o estacionaria en la diferencia.

Tras realizar esta prueba sobre la serie temporal ‘precio’ se obtienen los siguientes resultados:

```

ADF Statistic: -6.603
p-value: 0.000
Critical Value (1%): -3.430
Critical Value (5%): -2.862
Critical Value (10%): -2.567

```

## 2. ANÁLISIS Y PREPARACIÓN DE LOS DATOS

El estadístico ADF es menor que el valor crítico al 1% (-3.43), por lo que se rechaza la hipótesis nula: la serie temporal no tiene una raíz unitaria y es estacionaria (o estacionaria en torno a una constante).

Por otro lado, el test KPSS sigue un criterio opuesto al del ADF, de modo que busca comprobar la estacionariedad. Sus hipótesis son las siguientes:

- Hipótesis nula  $H_0$ : la serie temporal es estacionaria o estacionaria en torno a una constante.
- Hipótesis alternativa  $H_1$ : la serie temporal tiene una raíz unitaria y por tanto no es estacionaria.

Tras su implementación se obtienen los siguientes resultados:

```
KPSS Statistic: 39.49
p-value: 0.01
Critical Value (10%): 0.35
Critical Value (5%): 0.46
Critical Value (2.5%): 0.57
Critical Value (1%): 0.74
```

El valor del estadístico KPSS (39.49) es muy superior al del valor crítico al 1% (0.74) por lo que no se puede rechazar la hipótesis nula  $H_0$  y, por tanto, la serie temporal es estacionaria o estacionaria en torno a una constante.

Por tanto, a la vista de los resultados de ambas pruebas se concluye que el precio de la electricidad del mercado diario español es una serie temporal estacionaria.



### 3. METODOLOGÍA Y MODELOS

El objetivo principal de este trabajo es entrenar diferentes modelos de aprendizaje automático profundo y evaluar su rendimiento en la predicción de los precios de la electricidad del mercado diario del día siguiente.

Tal y como se comentó anteriormente, se dispone de datos horarios de 2015 a 2023. Como conjunto de entrenamiento se van a utilizar los años de 2015 a 2022, mientras que la primera mitad del 2023 se utilizará como conjunto de validación y la segunda mitad como conjunto de test. Hay que destacar que es fundamental que el año 2022 se encuentre dentro del conjunto de entrenamiento, para que el modelo sea capaz de captar la gran variabilidad de los precios de ese año.

Para predecir los precios de las 24 horas del día siguiente, se van a utilizar los 168 valores previos de las variables. Es decir, para predecir los precios del día siguiente se tienen en cuenta los de la semana anterior. Adicionalmente, también se dispondrá de los valores a futuro (de las siguientes 24h) de las variables “precio del gas”, “demanda”, “generación solar” y “generación eólica”, así como la variable “hueco térmico”, creada a partir de las anteriores, y las variables relativas al calendario.

En cuanto a los modelos que se van a implementar, como ya se comentó en el Estado del Arte, las redes neuronales recurrentes (LSTM y GRU) han demostrado obtener muy buenos resultados cuando se trata de series temporales y en concreto en la predicción de precios de la electricidad. Adicionalmente, se van a utilizar otros modelos basados en redes neuronales tipo MLP que han sido desarrollados recientemente y que han mostrado un muy buen comportamiento con series temporales, como NBEATS o NHITS. Por último, también se utilizará un transformer especializado en la predicción de series temporales.

Como base sobre la que comparar todos estos modelos de *deep learning*, se van a entrenar también modelos de machine learning tales como *linear regression*, *random forests* y *gradient boosting* entre otros.

Para implementar todos los modelos de redes neuronales se ha decidido utilizar la librería de alto nivel *NeuralForecast*. Esto es debido principalmente a que ya tiene implementadas clases para los modelos más recientes en predicción de series temporales, facilitando así el poder utilizarlos sobre nuestro conjunto de datos.

Esta librería forma parte del ecosistema Nixtla (Nixtla), que se compone de 5 librerías orientadas a la predicción de series temporales. Además de *NeuralForecast*, otra de ellas es *MLForecast*, la cual se va a utilizar para implementar los modelos de machine learning. También consta de un transformer llamado TimeGPT especializado en series temporales que se utilizará para predecir el precio de la electricidad.

Para evaluar el rendimiento de estos modelos sobre el conjunto de test, se implementa una validación cruzada (*cross validation*). Su funcionamiento consiste en deslizar una ventana de datos de entrada (en este caso 168 valores) y predecir el periodo posterior (en este caso 24 valores) sobre el conjunto de test, tal y como se ilustra en la Figura 26.

Además, para optimizar los modelos se implementa una búsqueda de los hiperparámetros de los modelos. Para ello, se utilizan las versiones “Auto” de los modelos (por ejemplo, “AutoLSTM”), que permiten realizar una búsqueda de hiperparámetros utilizando algún algoritmo de búsqueda con el objetivo de obtener los que hacen que el modelo obtenga un menor error (MSE).

Por otro lado, las métricas que se van a utilizar son RMSE, MSE y el coeficiente R<sup>2</sup>. Al tomar la variable objetivo valores iguales a cero, se descarta utilizar MAPE por los errores tan grandes que podría dar.

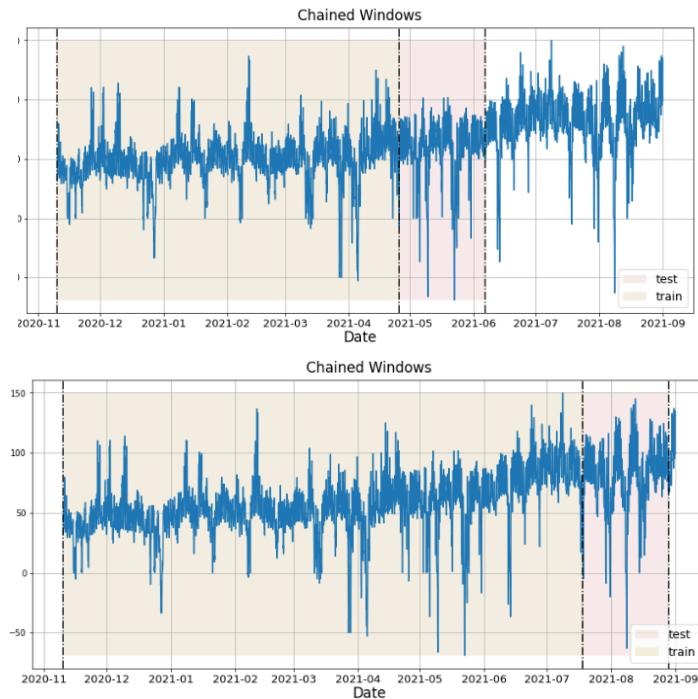


Figura 26: Funcionamiento de la validación cruzada

A continuación, se va a realizar una breve descripción de los modelos que se van a utilizar, para posteriormente pasar a su evaluación.

### 3.1. Modelos de Machine Learning

Aunque el objetivo de este proyecto es la implementación de modelos de aprendizaje profundo basados en redes neuronales recurrentes y métodos más novedosos como NBEATS o transformers, se decide también entrenar modelos clásicos de machine learning que sirvan como punto de partida. Algunos de estos algoritmos que se van a utilizar son los siguientes:

- Regresión lineal
- Regresión Lasso y Ridge
- Random forest
- Gradient boosting
- K-Nearest Neighbors

Aunque hay otros algoritmos más que podrían implementarse, por ejemplo SVM, como ya se comentó, no es el objetivo de este proyecto implementar y optimizar todas las técnicas de machine learning existentes, sino desarrollar métodos eficaces de deep learning y transformers, que ya han demostrado un mejor comportamiento en la predicción de series temporales. Por tanto, se proponen los mencionados anteriormente como punto de partida y para verificar a posteriori que no pueden competir con estos otros métodos.

Por otro lado, tal y como se mencionó anteriormente, para implementar estos modelos se va a utilizar MLForecast. Esta librería es rápida y sencilla de utilizar y admite cualquier modelo de *sklearn*.

Para evaluar los modelos se va a llevar a cabo una validación cruzada sobre el conjunto de test (segundo semestre de 2023).

El primer método que se utiliza es una regresión lineal, utilizando el regresor `LinearRegression()` de *sklearn*. El modelo de regresión lineal realiza una predicción simplemente computando la suma ponderada de las variables de entrada(Aurélien Géron, 2019):

$$\hat{y} = \theta_0 + \theta_1 y_1 + \theta_2 y_2 + \cdots + \theta_n y_n = h_{\theta}(\mathbf{x}) = \boldsymbol{\theta} \mathbf{x}$$

Donde  $\boldsymbol{\theta}$  es el vector de parámetros del modelo (incluyendo el término de bias) y  $x$  es el vector de variables de entrada. Al entrenar un modelo de regresión lineal, se halla el valor de  $\boldsymbol{\theta}$  que minimiza la función de coste “mean squared error” (MSE):

$$MSE(X, h_{\theta}) = \frac{1}{m} \sum_{i=1}^m (\theta^T x^{(i)} - y^{(i)})^2$$

También se van a desarrollar los métodos Lasso y Ridge, que aplican una regularización  $l_1$  y  $l_2$  respectivamente, modificando las funciones de coste de la siguiente manera

$$\begin{aligned} \text{Lasso: } J(\theta) &= MSE(\theta) + \alpha \sum_{i=1}^n \theta_i \\ \text{Ridge: } J(\theta) &= MSE(\theta) + \alpha \frac{1}{2} \sum_{i=1}^n \theta_i^2 \end{aligned}$$

El siguiente método que se utiliza es Random Forest, un método de aprendizaje conjunto (“ensemble”) basado en Árboles de Decisión que utiliza la técnica de *bagging* para el entrenamiento (Aurélien Géron, 2019).

El *bagging* es una técnica que utiliza un conjunto de estimadores en paralelo (en este caso árboles de decisión), cada uno de los cuales sobreajusta los datos, y promedia los resultados para obtener un mejor ajuste(VanderPlas, 2016). Un random forest no es más que un conjunto aleatorio de árboles de decisión que utiliza esta técnica.

Otro algoritmo de machine learning que se utiliza es LightGBM (Microsoft, 2016), que se basa en árboles de decisión y utiliza la técnica de *gradient boosting*. Como características de este gradient boosting en particular se encuentran: mayor rapidez de entrenamiento, menor uso de memoria, mejor precisión, posibilidad de entrenamiento distribuido y apto para grandes conjuntos de datos.

El *boosting* es una técnica de aprendizaje conjunto que combina una serie de estimadores “débiles” en un estimador “fuerte”. La idea detrás de la mayoría de los métodos de boosting es la combinación de los predictores de forma secuencial, de forma que cada uno intente corregir la predicción de su predecesor. Tal es el caso del gradient boosting, que además tiene la particularidad de intentar ajustar el nuevo predictor a los errores residuales del anterior (Aurélien Géron, 2019).

Por último, se utiliza el algoritmo K-Nearest Neighbors (*sklearn - nearest neighbors.*), que se fundamenta en el principio de encontrar un número predefinido ( $k$ ) de instancias de entrenamiento más cercanas a la nueva instancia y predecir su valor en base a esta distancia

Una vez entrenados los modelos con los parámetros por defecto, se lleva a cabo un afinado de los hiperparámetros para los modelos que mejor resultados han obtenido. Estos se exponen en el siguiente capítulo.

### 3.2. LSTM

LSTM (Long Short Term Memory) (Hasim Sak, 2014) es una variante de red neuronal recurrente (RNN) capaz de aprender dependencias a largo plazo, resistiendo el efecto de desvanecimiento del gradiente (“vanishing gradient”).

Su arquitectura se muestra en la Figura 27. A diferencia de la RNN simple, donde sólo se tenía una capa con *tanh*, ahora se tienen cuatro interactuando de una manera específica.

Las puertas  $i$ ,  $f$  y  $o$  hacen referencia a *input*, *forget* y *output*. La función sigmoide modula su salida entre 0 y 1. La puerta  $f$  define qué cantidad del estado previo  $h_{t-1}$  se quiere dejar pasar, la puerta  $i$  define qué cantidad del nuevo estado computado a partir de la entrada se quiere dejar pasar y la puerta  $o$  determina qué cantidad del estado interno se quiere pasar a la siguiente capa.

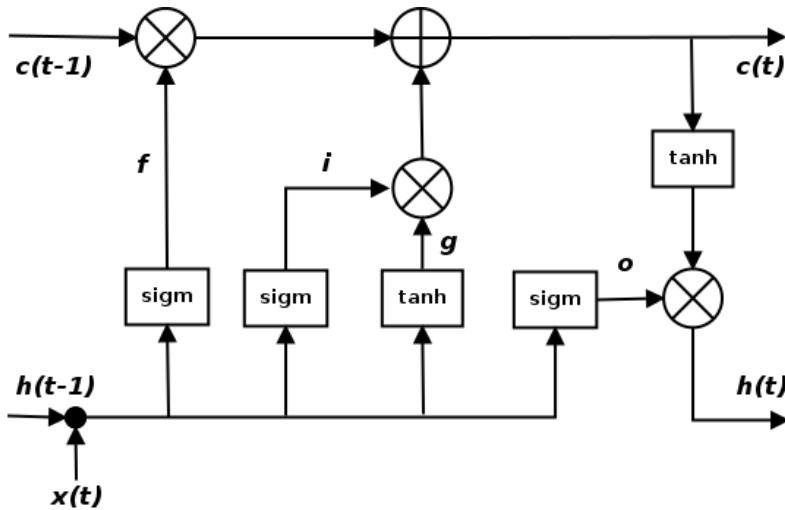


Figura 27: Esquema de la arquitectura de una célula LSTM (Antonio Gulli & Sujit Pal, 2017)

El estado interno  $g$  se calcula de forma idéntica a una RNN simple, teniendo en cuenta la entrada actual y el estado oculto previo, salvo que ahora su salida se modula con la salida de la puerta  $i$ .

Dados  $i, f, g$  y  $o$  se puede calcular el estado de la celda  $c_t$ , teniendo en cuenta el estado anterior  $c_{t-1}$ . Finalmente, el estado oculto  $h_t$  se obtiene multiplicando  $c_t$  por la puerta de salida  $o$ .

Las ecuaciones son fácilmente deducibles a la vista de la arquitectura y pueden encontrarse en el *paper* donde se presentó (Hasim Sak, 2014).

### 3.2.2. Entrenamiento

Como ya se comentó, para implementar esta arquitectura se utiliza la librería neuralforecast, que permite en unas pocas líneas de código entrenar el conjunto de datos con LSTM.

En concreto se utiliza la clase AutoLSTM que permite realizar varios entrenamientos iterando entre diversos parámetros para obtener los valores que obtengan un mejor rendimiento.

Para evaluar este rendimiento se implementa una validación cruzada sobre datos históricos, en concreto sobre el conjunto de test (segundo semestre de 2023).

El conjunto de parámetros que se establecen para la búsqueda es el siguiente:

```
config_lstm = {
    "context_size": horizon,
    "futr_exog_list": futr_exog_list,
    "hist_exog_list": hist_exog_list,
    "encoder_hidden_size": 32,
    "encoder_n_layers": 2,
```

```
"decoder_layers": 1,  
"decoder_hidden_size": 16,  
"learning_rate": tune.loguniform(1e-4, 6e-3),  
"scaler_type": 'minmax',  
"val_check_steps": 25,  
"max_steps": tune.choice([500, 750]),  
"batch_size": tune.choice([4, 16, 32]),  
"random_seed": tune.randint(10, 17),  
}
```

Como puede observarse, la arquitectura desarrollada se compone de dos capas de 128 celdas LSTM y una capa de 50 celdas totalmente conectadas. Estos parámetros se han fijado tras haber realizado búsquedas previas iterando sólo entre los parámetros específicos de la arquitectura y haber comprobado que se obtienen buenos resultados con los mencionados.

Además, cabe resaltar que se selecciona una ventana de datos de entrada de 24h, igual al horizonte, ya que se obtienen ligeramente mejores resultados que con 168 valores.

Otros parámetros que aparecen en la configuración anterior son la lista de nombres de las variables exógenas futuras e históricas o el tipo de escalado (“MinMax”).

Los parámetros que pueden tomar diferentes valores y que se introducen con el objetivo de obtener la mejor combinación de estos tras las iteraciones son la tasa de aprendizaje, el tamaño del lote, el número de ‘steps’ del entrenamiento y la ‘random seed’.

Los resultados de la validación cruzada, así como los hiperparámetros obtenidos se presentan en el siguiente capítulo (Capítulo “Resultados”)

### 3.3. GRU

#### 3.3.1. Descripción

GRU (Gated Recurrent Unit) es una variante de LSTM, que mantiene sus ventajas mientras simplifica su arquitectura, siendo más simple y rápida de entrenar.

Su arquitectura se muestra en la Figura 28. Ahora sólo se tiene dos puertas (*gates*): z (*actualización*) y r (*reset*). La primera define qué cantidad de memoria previa quedarse y la segunda define cómo combinar la nueva entrada con la memoria previa.

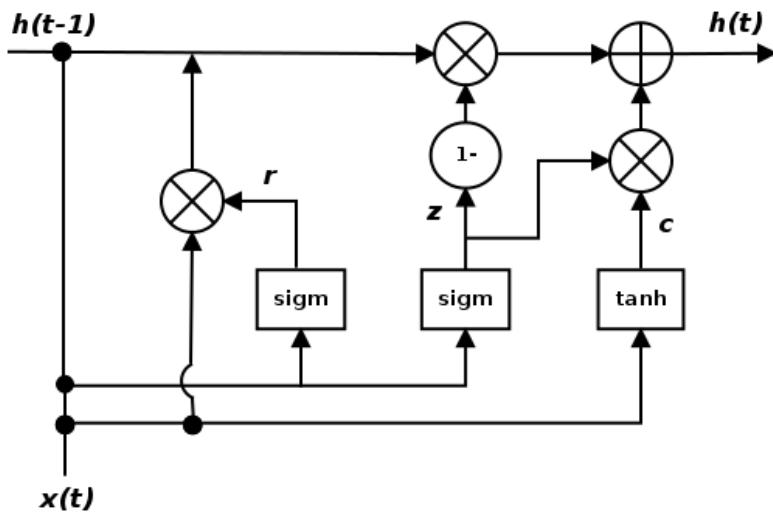


Figura 28: Esquema de la arquitectura de una célula GRU (Antonio Gulli & Sujit Pal, 2017)

De nuevo, no se muestran las ecuaciones por simplicidad, pero son fácilmente deducibles y se pueden encontrar en el *paper* original(Chung et al., 2014).

### 3.3.2. Entrenamiento

De forma muy similar a LSTM, se implementa GRU utilizando la clase AutoGRU de neuralforecast, que permite variar los hiperparámetros para obtener la mejor combinación de estos.

La configuración de esta arquitectura y la combinación de los posibles hiperparámetros se establece idéntica a la utilizada en LSTM:

```
config_gru = {
    "context_size": horizon,
    "futr_exog_list": futr_exog_list,
    "hist_exog_list": hist_exog_list,
    "encoder_hidden_size": 32,
    "encoder_n_layers": 2,
    "decoder_layers": 1,
    "decoder_hidden_size": 16,
    "learning_rate": tune.loguniform(1e-4, 6e-3),
    "scaler_type": 'minmax',
    "val_check_steps": 25,
    "max_steps": tune.choice([500, 750]),
    "batch_size": tune.choice([4, 16, 32]),
    "random_seed": tune.randint(10, 17),
}
```

## 3.4. NBEATsX

### 3.4.1. Descripción

NBEATS (Neural Basis Expansion Analysis)(Oreshkin et al., 2020) es una arquitectura de deep learning basada en MLP con enlaces residuales hacia delante (*forward*) y hacia atrás (*backward*). Está compuesto por pilas o stacks profundos con capas totalmente conectadas.

Según el artículo en el que se presenta este modelo, tiene como características que es interpretable, rápido de entrenar y es aplicable a un amplio dominio de conjuntos de datos sin necesidad de aplicarles ninguna modificación. Además, utilizarlo sobre conjuntos de datos ampliamente conocidos obtienen muy buenos resultados.

La arquitectura se muestra en la Figura 29. Los bloques están conectados entre sí de forma que la entrada de un bloque es la salida residual del anterior. La entrada del primer bloque es el input del modelo, que no es más que una ventana de datos históricos, de longitud típicamente de entre 2 y 7 veces el horizonte de predicción, que termina con la última observación medida.

Cada uno de los stacks está formado por bloques que, a su vez, están compuestos por una red de varias capas totalmente conectadas (FC: fully connected) que produce los coeficientes de expansión hacia delante ( $\theta^f$ ) y hacia atrás ( $\theta^b$ ). Estos coeficientes pasan finalmente por una capa base que genera las predicciones a futuro (forecast), para el horizonte de predicción, y a pasado (backcast), para la ventana de datos de entrada. Las ecuaciones para un bloque cualquiera  $i$  son las siguientes:

$$h_{i,1} = FC_{i,1}(x_i), \quad h_{i,2} = FC_{i,2}(h_{i,1}), \quad h_{i,3} = FC_{i,3}(h_{i,2}), \quad h_{i,4} = FC_{i,3}(h_{i,3})$$

$$\theta_i^b = Linear_i^b(h_{i,4}), \quad \theta_i^f = Linear_i^f(h_{i,4})$$

La linealización de  $\theta_i$  no es más que una capa que proyecta linealmente  $h$ , es decir:  $\theta_i^f = W_i^f(h_{i,4})$ . Por otro lado, FC hace referencia una capa totalmente conectada con función de activación ReLU. De este modo,  $h$  se podría escribir como:  $h_{i,4} = ReLU_i(W_{i,4}h_{i,3} + b_{i,4})$ .

Finalmente, en las capas base se pasa de los coeficientes de expansión a las predicciones forward y backward, es decir  $\hat{y}_i = g_i^f(\theta_i^f)$ ,  $\hat{x}_i = g_i^b(\theta_i^b)$

Dependiendo de la función  $g$  que se utilice, se tendrán los siguientes tipos de stacks: tendencia, estacionalidad o genérico. Esto viene detallado en el artículo al que se ha aludido anteriormente, donde además se indica que los mejores resultados se obtienen mediante el aprendizaje conjunto (ensembling).

Otra característica de esta arquitectura es la existencia de enlaces residuales dobles a nivel de la pila o stack. Por un lado, el input del bloque se añade al output de la predicción backcast antes de pasarlo como entrada del siguiente bloque, y por otro, las predicciones forecast se agregan para dar lugar a la predicción de la pila. Las ecuaciones para cada bloque  $i$  y para cada stack serían las siguientes:

$$x_i = x_{i-1} - \hat{x}_{i-1}, \quad \hat{y} = \sum_i \hat{y}_i$$

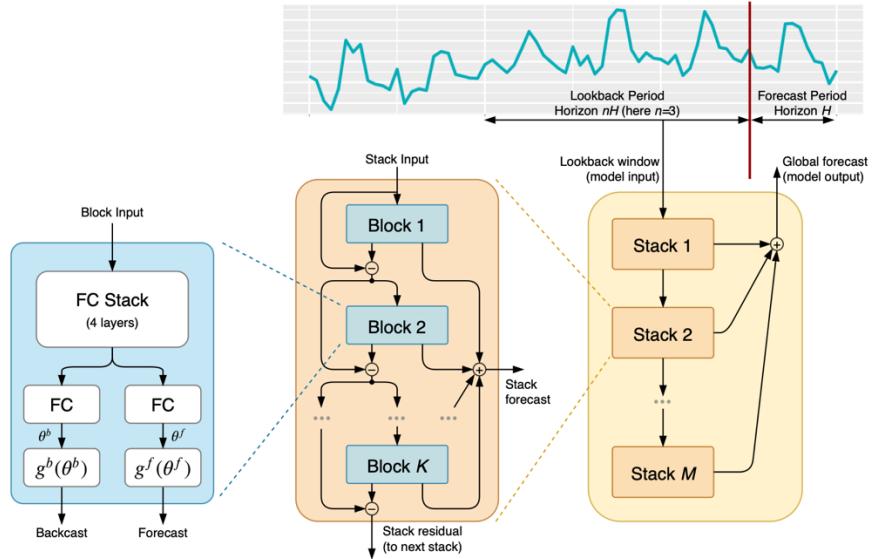


Figura 29: Esquema de la arquitectura de NBEATS (Oreshkin et al., 2020)

#### 3.4.2. Entrenamiento

Nuevamente, para implementar esta arquitectura se utiliza de nuevo la librería neuralforecast, que permite en unas pocas líneas de código entrenar el conjunto de datos con NBEATS.

Al igual que con LSTM y GRU, vamos a utilizar variables adicionales o exógenas, tanto históricas como futuras, para predecir el precio de la electricidad. De este modo, se utiliza el modelo NBEATSx(Olivares et al., 2022) que permite incluir este tipo de variables.

En concreto, se utiliza la clase AutoNBEATSx que permite iterar el entrenamiento del modelo variando los parámetros para seleccionar el que mejor puntuación obtenga sobre el conjunto de validación.

Los parámetros que se utilizan son los siguientes:

```
config_nbeatsx = {
    "input_size": horizon*7,
    "futr_exog_list": futr_exog_list,
    "n_blocks": [3*[1]],
    "mlp_units": [3 * [[512, 512]]],
    "learning_rate": tune.loguniform(2e-4, 1e-3),
    "val_check_steps": 50,
    "max_steps": 1000,
```

```
"scaler_type": "minmax",
"batch_size": tune.choice([4, 12, 24, 36]),
"random_seed": tune.randint(10, 17),
}
```

Obsérvese que se fijan los siguientes parámetros: horizonte, tamaño de la ventana de entrada, tipo de bloques (trend, seasonality y identity), número de bloques de cada tipo, el número máximo de steps, el número steps cada el que se evalúa sobre el conjunto de validación y el tipo de scaler. Mientras que se deja iterar la tasa de aprendizaje, el tamaño del lote y la random seed.

Hay que indicar que algunos de los parámetros, como “max\_steps”, “mlp\_units” o “n\_blocks”, se han dejado fijos tras pruebas iniciales donde se ha verificado que claramente eran los valores más adecuados. De hecho, tanto “mlp\_units” como “n\_blocks” se quedan con los valores por defecto, que parecen ser los más idóneos.

Con esta configuración, y utilizando como función de pérdidas MSE, se entrena el modelo y se lleva a cabo una validación cruzada para evaluar el modelo sobre el conjunto de validación, correspondientes a los primeros seis meses de 2023, para 10 combinaciones de parámetros.

### 3.5. NHITS

#### 3.5.1. Descripción

NHITS (Neural Hierarchical Interpolation for Time Series) (Challu et al., 2022) es una arquitectura basada en NBEATS, especializada en la predicción a largo plazo gracias a su interpolación jerárquica y su procesamiento de datos de diferentes frecuencias.

Tal y como se muestra en la Figura 30, la arquitectura es muy similar a la de NBEATS. La principal diferencia reside en la existencia de diferentes stacks, que se focalizan en predecir las diferentes frecuencias de la serie temporal de forma jerárquica. Además, en cada bloque, antes de las capas MLP, se introducen capas de *subsampling*, tales como Max Pooling, que reducen las necesidades computacionales.

### 3. METODOLOGÍA Y MODELOS

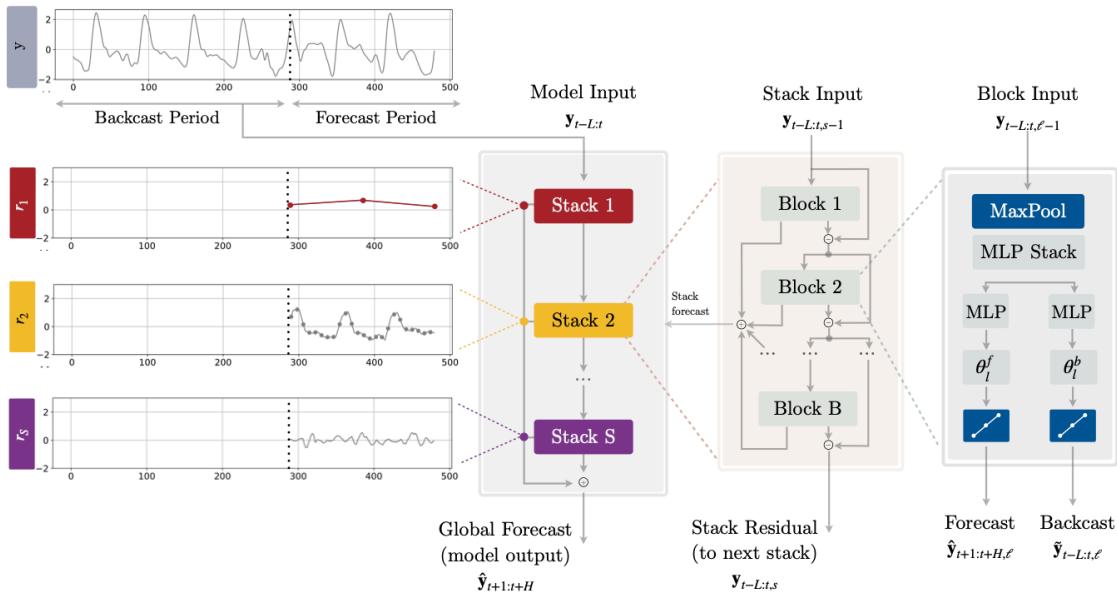


Figura 30: Esquema de la arquitectura de NHITS (Challu et al., 2022)

#### 3.5.2. Entrenamiento

De nuevo se utiliza la clase AutoNHITS de *neuralforecast* para iterar sobre algunos de los parámetros, mientras que otros se dejarán fijos:

```
config_nhits = {
    "input_size": horizon*7,
    "futr_exog_list": futr_exog_list,
    "n_blocks": 3*[1],
    "mlp_units": 3 * [[512, 512]],
    "n_freq_downsample": tune.choice([[4, 2, 1], [8, 4, 1]]),
    "n_pool_kernel_size": tune.choice([[4, 2, 1], [8, 4, 1]]),
    "learning_rate": tune.loguniform(1e-4, 1e-3),
    "val_check_steps": 100,
    "max_steps": tune.choice([500, 1000]),
    "scaler_type": "minmax",
    "batch_size": tune.choice([4, 12, 32]),
    "random_seed": tune.randint(10, 17),
}
```

Además de los parámetros que se indicaban para su iteración en NBEATSx, aquí también se añaden dos parámetros específicos de esta arquitectura: “`n_freq_downsample`”, y “`n_pool_kernel_size`”.

Con esta configuración, y utilizando como función de pérdidas MSE, se entrena el modelo y se lleva a cabo una validación cruzada para evaluar el modelo sobre el conjunto de

validación, correspondientes a los primeros seis meses de 2023, para 20 combinaciones de parámetros.

### 3.6. TimeGPT

#### 3.6.1. Descripción

TimeGPT es un transformer entrenado sobre enormes cantidades de datos de series temporales. Está basado en mecanismos de atención, que son capaces de capturar la diversidad de los datos pasados y extrapolarlos al futuro (Vaswani et al., 2017).

Su arquitectura consiste en una estructura de encoder-decoder con múltiples capas, cada una con enlaces residuales y capa de normalización. Finalmente, una capa lineal asigna la salida del decoder con la dimensión de la ventana de datos que se predice. Un esquema de su arquitectura y funcionamiento se muestra en la Figura 31.

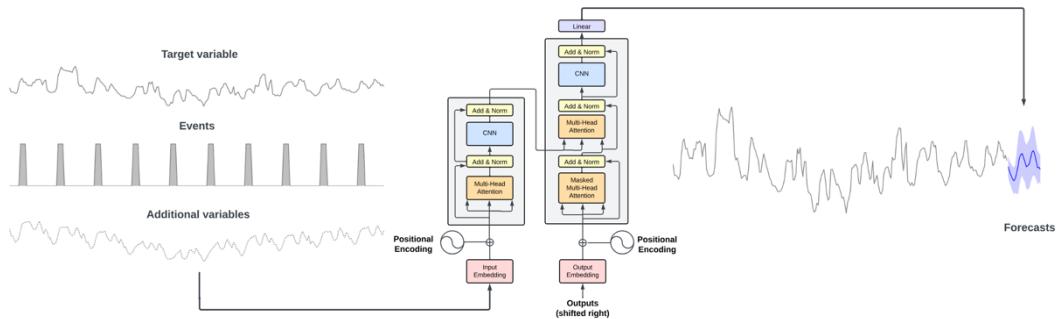


Figura 31: Arquitectura del transformer TimeGPT (Nixtla)

#### 3.6.2. Entrenamiento

Para poder utilizar TimeGPT es necesario utilizar un token proporcionado al registrarse en Nixtla. Una vez introducido el token personal, simplemente se escala el conjunto de datos utilizando ‘MinMaxScaler’ y, de manera muy sencilla y sin parámetros, se implementa una validación cruzada con este Transformer. El único parámetro que se le indica es que utilice el modelo “timegpt-1-long-horizon”, enfocado en predicciones a largo plazo, ya que se comprueba que obtiene mejores resultados.



## 4. RESULTADOS

Comenzando por los modelos de machine learning, la validación cruzada sobre el conjunto de test con los parámetros por defecto arroja los resultados de la Tabla 3.

	<b>Naive</b>	<b>LGBM</b>	<b>Lasso</b>	<b>Lin. Reg.</b>	<b>Ridge</b>	<b>KNN</b>	<b>RF</b>
RMSE	30.3	<b>17.2</b>	40.4	<b>20.7</b>	20.9	31.9	23.9
MAE	21.9	<b>12.4</b>	34.2	<b>14.9</b>	15.1	24.4	16.7
R2	0.35	0.79	-0.15	0.70	0.69	0.28	0.60

Tabla 3: Resultados de la validación cruzada con modelos de machine learning

El mejor resultado lo obtiene el gradient boosting LightGBM, seguido por la regresión lineal, mientras que la regresión Lasso y KNN obtienen peores resultados que el modelo ‘Naive’.

Tras el afinado de hiperparámetros de los modelos LGBM y Ridge, se obtiene una ligera mejora en el caso de LGBM (Tabla 4):

	<b>LGBM</b>	<b>Ridge</b>
RMSE	16.8	20.7
MAE	12.0	14.9
R2	0.80	0.70

Tabla 4: Resultados de los dos mejores modelos de machine learning tras el afinado de hiperparámetros

En la Figura 32 se muestran los resultados de las predicciones del modelo **LGBM** para los 10 primeros días del conjunto de test, junto con los precios reales. Se puede observar que el modelo es capaz de captar razonablemente bien las variaciones en el precio. Las mayores diferencias entre los valores reales y predichos parecen encontrarse cuando el precio llega a los 0 €/kWh.

Para comprobar que la adición de la variable hueco térmico es de gran utilidad para la predicción del precio de la electricidad y que proporciona parte de la información contenida en la generación de ciclos combinados y la generación hidráulica, se realizan dos pruebas de validación cruzada con el modelo LGBM: una quitándole la variable hueco térmico y otra incluyendo la generación de ciclos combinados, la generación hidráulica y el consumo de bombeo. Los resultados de estas pruebas se recogen en la Tabla 5.

En el primer caso, sin el hueco térmico, el modelo empeora considerablemente en términos de RMSE, mientras que en el segundo caso, al añadir las tres variables mencionadas anteriormente, el modelo mejora contenidamente.

### 3. RESULTADOS

---

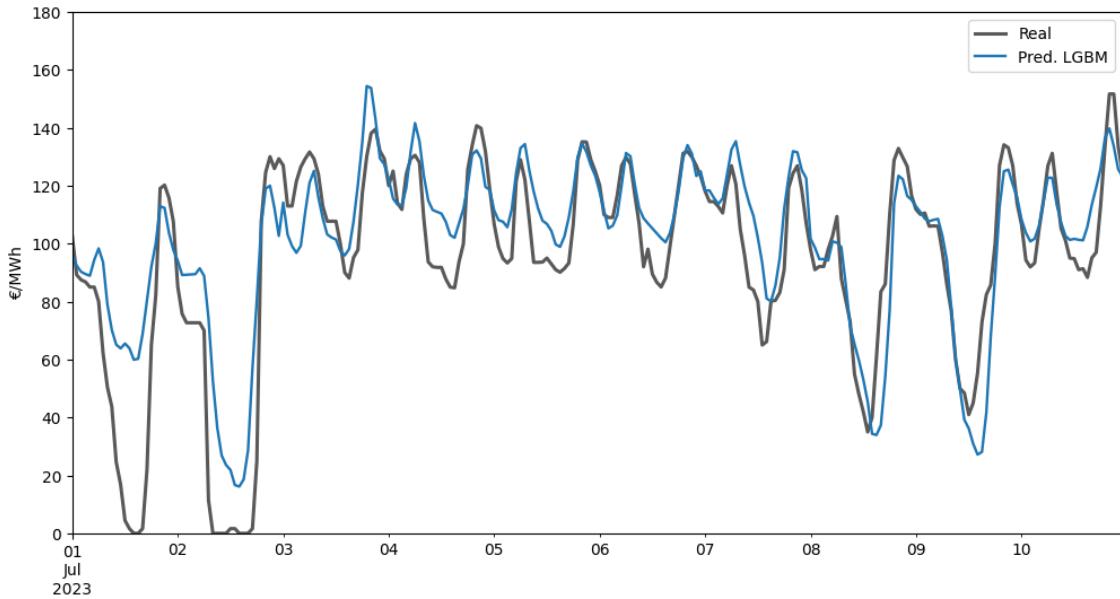


Figura 32: Valores reales vs predichos por LGBM de los 10 primeros días

LGBM	LGBM sin Hueco Termico	LGBM con todas las variables
RMSE	17.2	18.3
MAE	12.4	12.6
R2	0.79	0.76

Tabla 5: Influencia del hueco térmico con el modelo LGBM

Por otro lado, los mejores resultados en la validación cruzada con el modelo de **LSTM** se muestran en la Tabla 6 y se han obtenido con los siguientes hiperparámetros:

```
learning_rate = 3.3e-3
max_steps = 350
batch_size = 4
random_seed = 12
```

LSTM	
RMSE	17.8
MAE	12.9
R2	0.68

Tabla 6: Resultados de la validación cruzada con el modelo basado en LSTM

En la Figura 33 se muestran los resultados de las predicciones del modelo LSTM para los 10 primeros días del conjunto de test, junto con los precios reales. A la vista de la gráfica, parece que el modelo esté sobreentrenado y una mejor búsqueda de los parámetros que determinan la arquitectura sea necesario.

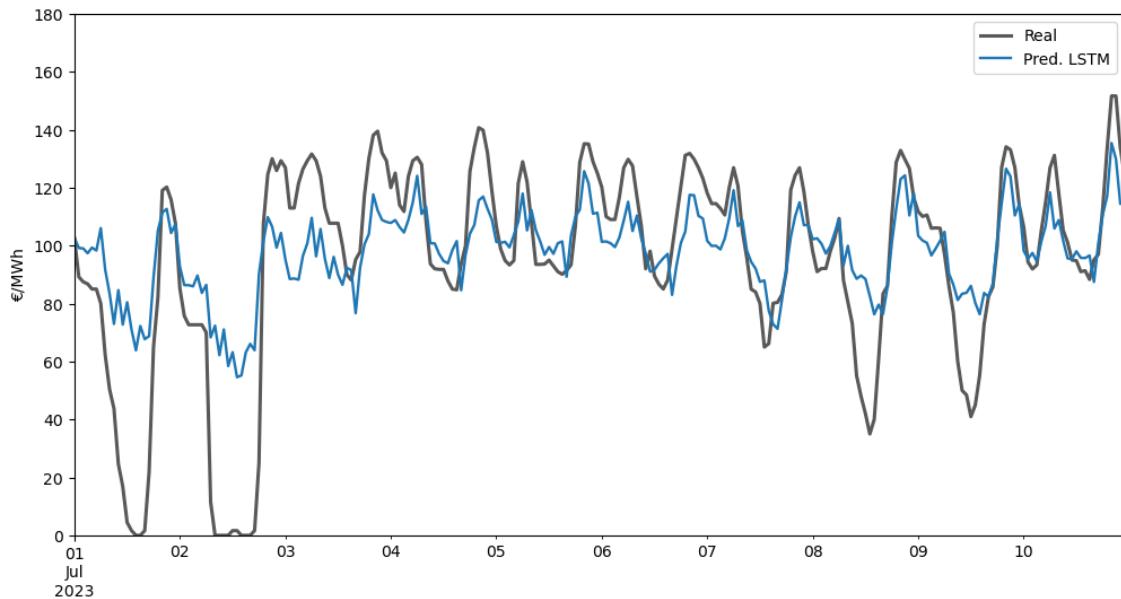


Figura 33: Valores reales vs predichos por LSTM de los 10 primeros días

En el caso de **GRU**, tras la validación cruzada, el mejor modelo obtiene los resultados mostrados en la Tabla 7, con los siguientes hiperparámetros:

```
learning_rate = 3.3e-3
max_steps = 350
batch_size = 4
random_seed = 12
```

GRU	
RMSE	17.2
MAE	11.9
R2	0.67

Tabla 7: Resultados de la validación cruzada con el modelo basado en GRU

En la Figura 34 se muestran los resultados de las predicciones del modelo GRU para los 10 primeros días del conjunto de test. A la vista de la gráfica, parece que el modelo esté sobreentrenado y una mejor búsqueda de los parámetros que determinan la arquitectura sea necesario.

### 3. RESULTADOS

---

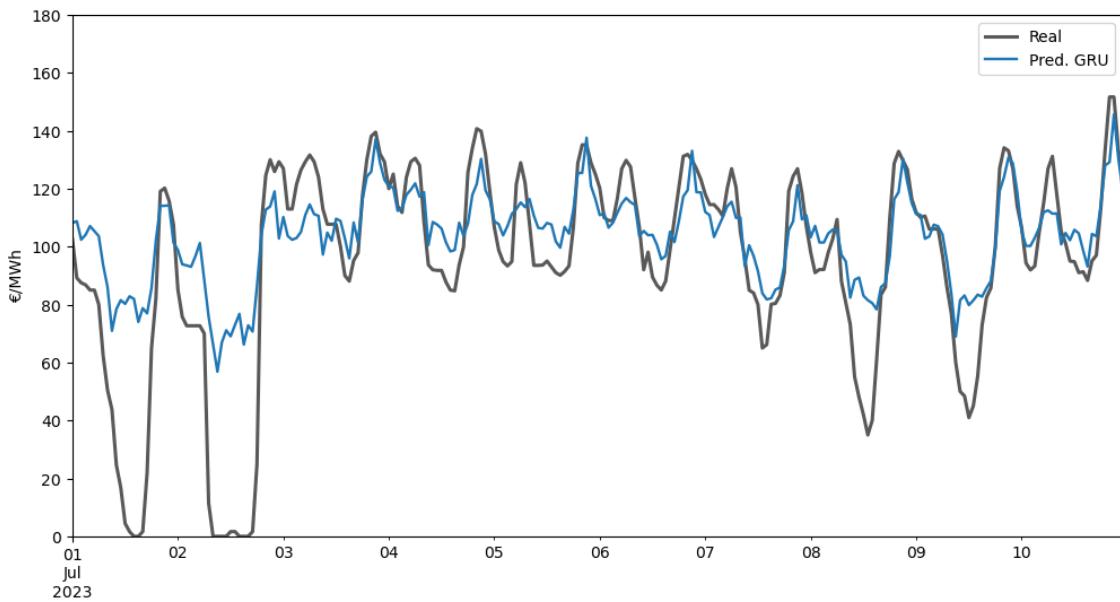


Figura 34: Valores reales vs predichos por GRU de los 10 primeros días

Para el caso de **NBEATSx**, los mejores resultados se muestran en la Tabla 8 y se obtienen con los siguientes hiperparámetros:

```
learning_rate = 5.2e-4  
batch_size = 12  
random_seed = 15
```

NBEATSx	
RMSE	14.6
MAE	10.6
R2	0.83

Tabla 8: Resultados de la validación cruzada con el modelo NBEATSx

En la Figura 35 se muestran los resultados de las predicciones del modelo NBEATSx para los 10 primeros días del conjunto de test. Se puede observar que captura muy bien las variaciones en el precio. Las mayores diferencias entre los valores reales y predichos parecen encontrarse una vez más cuando el precio llega a los 0 €/kWh. En general, también parece observarse una mayor diferencia en las puntas de precio o en los valles.

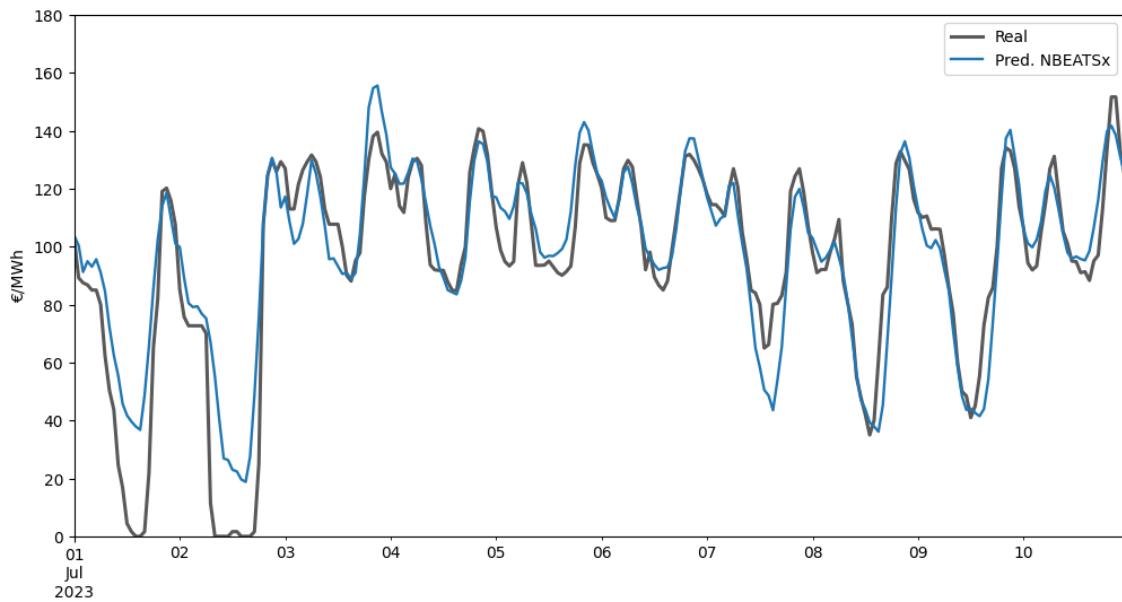


Figura 35: Predicciones del modelo NBEATSx para los 10 primeros días del conjunto de test

Al igual que con el modelo de gradient boosting, se realizan pruebas para ver la influencia del hueco térmico en la predicción de los precios de la electricidad con el modelo NBEATSx, obteniendo los resultados de la Tabla 9.

	NBEATSx	NBEATSx sin Hueco Termico	NBEATSx con todas las variables
RMSE	14.6	16.1	14.4
MAE	10.6	11.5	10.5
R2	0.84	0.80	0.85

Tabla 9: Comparación de resultados de las pruebas con NBEATSx de la influencia del hueco térmico

Se observa una gran mejora del modelo al añadir la variable hueco térmico, pasando del 16.1 al 14.6 de RMSE y del 11.5 al 10.6 de MSE. Por otro lado, si se añaden las variables de generación de energía térmica, generación hidráulica y consumo de bombeo, apenas se produce una mejora en el modelo, disminuyendo tan sólo entre 1 y 2 décimas los valores de las métricas.

Por otra parte, en el caso de **NHITS**, los mejores resultados se muestran en la Tabla 10 y se obtienen con los siguientes hiperparámetros tras la búsqueda:

```
n_freq_downsample = (4, 2, 1)
n_pool_kernel_size = (8, 4, 1)
learning_rate = 1.3e-4
batch_size = 32
random_seed = 12
```

### 3. RESULTADOS

---

NHITS	
RMSE	15.8
MAE	11.7
R2	0.81

Tabla 10: Resultados de la validación cruzada con el modelo NHITS

De nuevo, en la Figura 36 se muestran los resultados de las predicciones del modelo NHITS para los 10 primeros días del conjunto de test. Tanto por las métricas como a la vista de la gráfica, se observa un peor comportamiento del modelo respecto a NBEATSx.

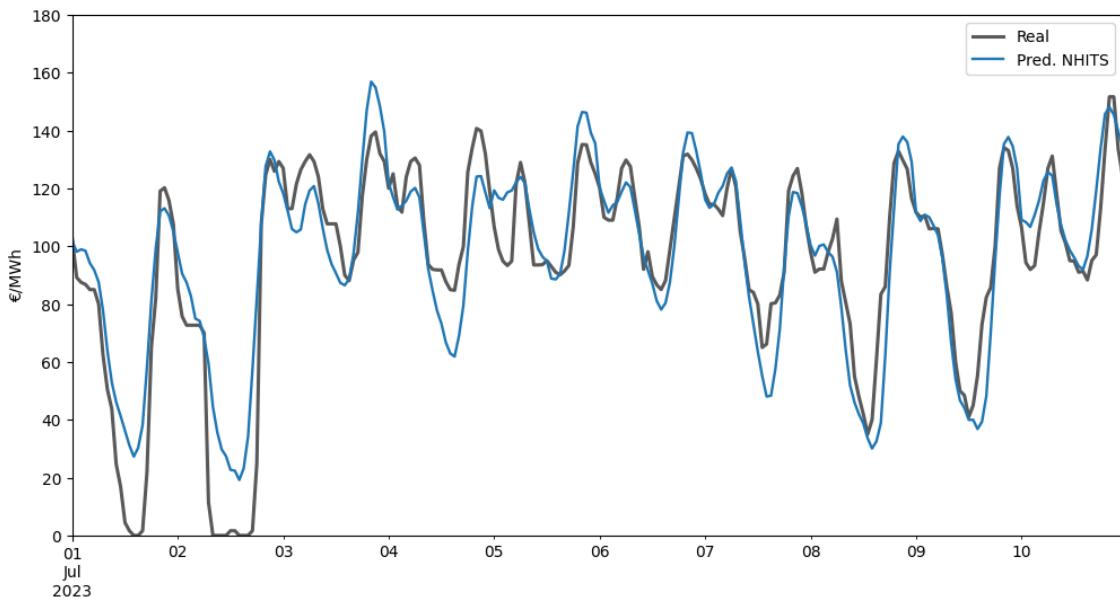


Figura 36: Predicciones del modelo NHITS para los 10 primeros días del conjunto de test

Por último, para el caso de **TimeGPT**, los resultados obtenidos se muestran en la Tabla 11, mientras que en la Figura 36 se muestran las predicciones del precio de la electricidad para los 10 primeros días del conjunto de test.

TimeGPT	
RMSE	15.1
MAE	11.0
R2	0.79

Tabla 11: Resultados de la validación cruzada con el transformer TimeGPT

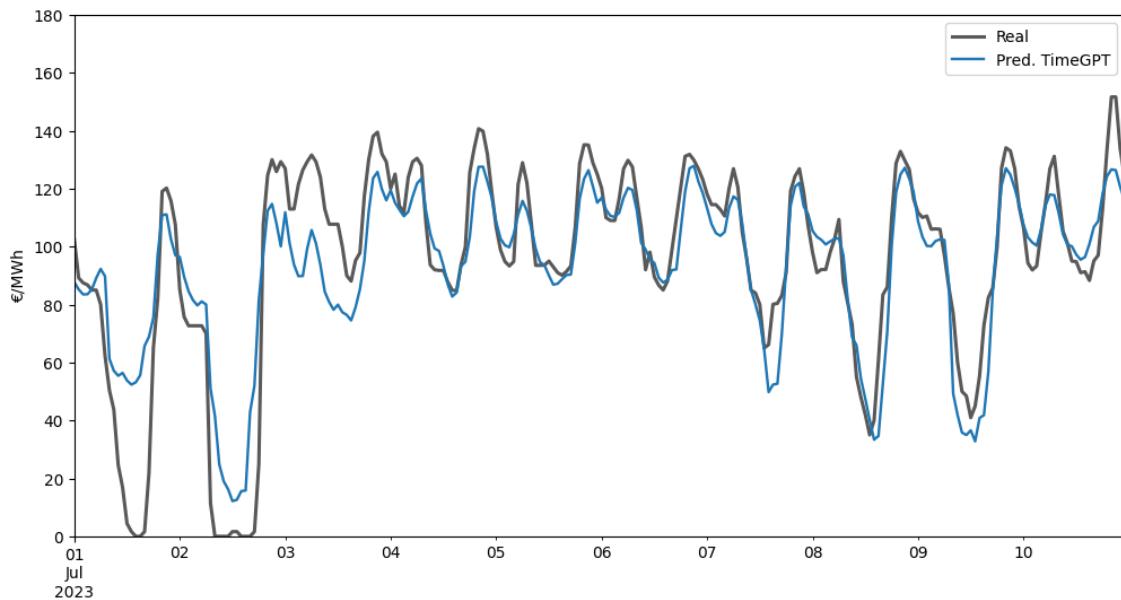


Figura 37: Predicciones del transformador TimeGPT para los 10 primeros días del conjunto de test

A modo de resumen y de comparativa, se recoge en la Tabla 12 los resultados de la validación cruzada de todos los modelos.

	<b>LGBM</b>	<b>LSTM</b>	<b>GRU</b>	<b>NBEATSx</b>	<b>NHITS</b>	<b>TimeGPT</b>
RMSE	16.8	17.8	17.2	<b>14.6</b>	15.9	<b>15.1</b>
MAE	12.0	12.9	11.9	<b>10.6</b>	11.8	<b>11.0</b>
R2	0.80	0.68	0.67	<b>0.83</b>	0.81	0.79

Tabla 12: Resumen comparativo de los resultados de todos los modelos

También se muestran en una misma figura las curvas de precios de la electricidad de los primeros 10 días del conjunto de test, correspondientes a las fechas del 1 de julio al 9 de julio de 2023 (inclusive), predichas por los tres mejores modelos junto con la curva de precios real (Fig. 38)

### 3. RESULTADOS

---

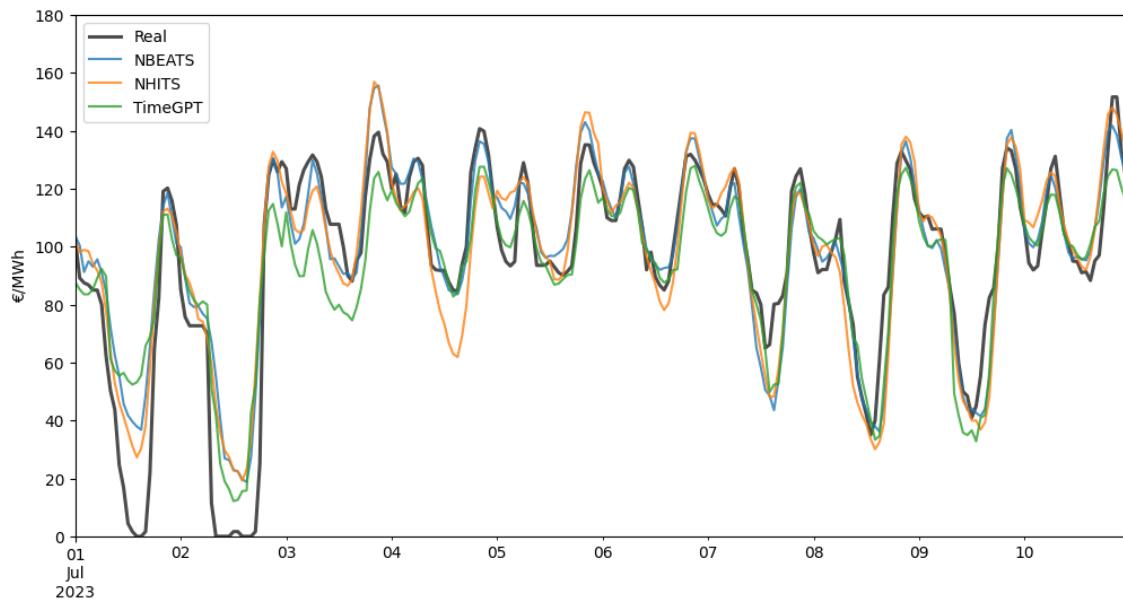


Figura 38: Predicciones de los tres mejores modelos para los 10 primeros días del conjunto de test



### 5. CONCLUSIONES Y LÍNEAS FUTURAS

En este proyecto se han implementado técnicas de deep learning para predecir los precios del mercado eléctrico diario español en la actualidad de forma satisfactoria, ya que los modelos desplegados han sido capaces, en gran medida, de captar la gran variabilidad que en la actualidad presentan las curvas de precio diarias, debido en buena parte al gran incremento de la generación de energía renovable en comparación con años atrás (véase Figura 38).

Se han añadido variables exógenas con gran influencia en la formación de los precios de la electricidad, como son el precio del gas, la demanda eléctrica y la generación de energía eólica y solar. Estas variables, que son conocidas o se pueden estimar para el día siguiente, ayudan notablemente a predecir los precios de la electricidad.

Se han implementado modelos de machine learning, como gradient boosting, modelos basados en RNN como LSTM y GRU, ampliamente utilizados en la literatura para tareas similares, y métodos más novedosos como NBEATSx y TimeGPT, que han obtenido los mejores resultados (véase Tabla 12).

En concreto, NBEATSx ha obtenido el mejor resultado en todas las métricas con cierta diferencia sobre el siguiente modelo, TimeGPT. El tercer mejor modelo ha sido NHITS. Por otra parte, las redes neuronales recurrentes han resultado ser peores que el LGBM.

Con la creación de la variable hueco térmico, combinación de la demanda y la generación eólica y solar, se ha conseguido aproximar la información proporcionada por la generación de energía térmica e hidráulica, variables desconocidas a futuro, ya que se obtienen casi los mismos valores en las tres métricas con el modelo NBEATSx al añadir esta variable nueva que con las otras dos anteriores.

Por otro lado, es necesario mejorar las arquitecturas sobre todo de las redes neuronales recurrentes, haciendo búsquedas de los parámetros más extensas, con el fin de evitar el *overfitting* y mejorar su rendimiento.

Otra línea de mejora sería introducir una variable adicional que indique a los modelos el periodo de desacople del precio del gas y el de la electricidad, la llamada “Excepción Ibérica” mencionada en este trabajo.

También se podría considerar añadir otras variables, tales como las meteorológicas (temperatura, humedad, velocidad del viento, etc), con el fin de mejorar los modelos.

Finalmente, sería interesante implementar estos modelos con una granularidad temporal mayor, por ejemplo 15-minutal.



## BIBLIOGRAFÍA

- Antonio Gulli, &., & Sujit Pal. (2017). *Deep Learning with Keras*. Packt Publishing.
- Aurélien Géron. (2019). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow, 2nd Edition*
- Challu, C., Olivares, K. G., Oreshkin, B. N., Garza, F., Mergenthaler-Canseco, M., & Dubrawski, A. (2022). N-HiTS: Neural Hierarchical Interpolation for Time Series Forecasting.
- Chung, J., Gulcehre, C., Cho, K., & Bengio, Y. (2014). Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling.
- Europa Press. . *EpData*. <https://www.epdata.es/evolucion-agua-embalsada-capacidad-reservas-espana/874be4ee-4011-4908-8e88-26b3702decba/espana/106>
- García, &., L.M., Torres Maldonado, &., J.F., Troncoso, &., A., Riquelme Santos, &., & J.C. (2024). *Técnicas Big Data para la predicción de la demanda y precio eléctrico* .
- Hasim Sak, A. S., Francoise Beaufays. (2014). *LONG SHORT-TERM MEMORY BASED RECURRENT NEURAL NETWORK ARCHITECTURES FOR LARGE VOCABULARY SPEECH RECOGNITION*
- Lago, J., De Ridder, F., & De Schutter, B. (2018). Forecasting spot electricity prices: Deep learning approaches and empirical comparison of traditional algorithms. *Applied Energy*, 221, 386. 10.1016/j.apenergy.2018.02.069
- Mercado Ibérico del Gas. . *Acceso a ficheros*. MIBGAS. <https://www.mibgas.es/es/file-access>
- Microsoft. (2016). *LightGBM*. <https://lightgbm.readthedocs.io/en/stable/>
- Nixtla. . *Nixtla Ecosystem*. <https://nixtlaverse.nixtla.io>
- Olivares, K. G., Challu, C., Marcjasz, G., Weron, R. L., & Dubrawski, A. (2022). Neural basis expansion analysis with exogenous variables: Forecasting electricity prices with NBEATSx.
- OMIE. . *Mercado Spot de Electricidad*. OMIE. <https://www.omie.es/es/mercado-de-electricidad>

Oreshkin, B. N., Carpor, D., Chapados, N., & Bengio, Y. (2020). *N-BEATS: NEURAL BASIS EXPANSION ANALYSIS FOR INTERPRETABLE TIME SERIES FORECASTING* .

Poggi, A., Di Persio, L., & Ehrhardt, M. (2023). Electricity Price Forecasting via Statistical and Deep Learning Approaches: The German Case. *AppliedMath*, 3(2), 316. 10.3390/appliedmath3020018

Red Eléctrica Española. . *Sistema de Información del Operador del Sistema (e-sios)*.  
<https://www.esios.ree.es/es>

*sklearn - nearest neighbors* . . <https://scikit-learn.org/stable/modules/neighbors.html#regression>

VanderPlas, J. (2016). *Python Data Science Handbook*. O'Reilly Media, Inc.

Vaswani, A., Shazeer, N., Brain, G., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., & Kaiser, Ł. (2017). *Attention Is All You Need*

Yousefi, A., Sianaki, O. A., & Sharafi, D. Long-Term Electricity Price Forecast Using Machine Learning Techniques. Paper presented at the