



Business School

MÁSTER EN DATA SCIENCE Y BUSINESS
ANALYTICS ONLINE

TRABAJO FIN DE MÁSTER: “MODELOS DE MACHINE LEARNING PARA LA PREDICCIÓN DEL PRECIO DE LA ELECTRICIDAD EN EL MERCADO ESPAÑOL”

TFM elaborado por:
Tutor/a de TFM:

Iñigo Elorza Barea
Abel Ángel Soriano Vázquez

- Madrid a 31 de julio de 2023 -

INDICE

1. SÍNTESIS DEL TRABAJO	7
2. INTRODUCCIÓN	8
2.1. MODULOS DEL MÁSTER RELACIONADOS CON EL TRABAJO FIN DE MÁSTER	9
3. ANTECEDENTES	10
3.1 IMPORTANCIA DE LA ENERGIA ELECTRICA	10
3.2 MARCO REGULATORIO	11
3.3 SISTEMA ELÉCTRICO ESPAÑOL	12
3.4 ETAPAS DEL SISTEMA ELÉCTRICO ESPAÑOL.....	14
3.5 ESTRUCTURA DE GENERACIÓN DE LA EE EN ESPAÑA	15
3.5.1 Tecnologías de producción de ee	15
3.5.2 Evolución de la producción de ee.....	18
3.6 AGENTES DEL MERCADO	23
3.7 MERCADO ELECTRICO ESPAÑOL	25
3.8 FORMACION DEL PRECIO DE LA ENERGIA ELECTRICA.....	26
3.8.1. Formación del precio en el mercado mayorista a plazo	26
3.8.2. Formación del precio en el mercado mayorista diario	27
3.9 ESTADO DEL CONOCIMIENTO	28
3.9.1 Introducción.....	29
3.9.2 Horizontes de previsión	30
3.9.3 Descripción general de las metodologías de modelado	30
4. BUSSINES CASE	33
5. OBJETIVOS A ALCANZAR	34
6. DATOS DE PARTIDA.....	36
6.1 INTERVALO TEMPORAL.....	36
6.2 DIA DE LA SEMANA.....	39
6.3 VARIABLES CONSIDERADAS.....	40
6.4 ANALISIS DE LOS DATOS.....	45
6.5 EXPLORACION DE LAS VARIABLES.....	51
6.5.1 VARIABLE: Precio medio diario energía eléctrica (Pre_elec)	52

6.5.2 VARIABLE: Demanda media diaria (Dem)	53
6.5.3 VARIABLE: Precio del petróleo (Prec_petr)	53
6.5.4 VARIABLE: Precio del gas natural (prec_gas)	54
6.5.5 VARIABLE: Precio medio diario del carbón (Prec_car).....	55
6.5.6 VARIABLES: Producción por tecnología	55
6.5.7 VARIABLE: Temperatura (Temp_).....	58
6.5.8 VARIABLES: Velocidad del viento.....	64
6.5.9 VARIABLE: Reservas hidráulicas (Res_hidr).....	68
6.5.10 VARIABLE: Precio derechos de emisión CO2 (Der_CO2)	69
6.5.11 VARIABLE: Situación socio-económica del país (Ibex)	70
6.5.12 VARIABLE: Intercambio de EE con otros países (int_ee)	71
6.6 TRATAMIENTO DE LOS DATOS Y FORMACION DEL DATASET	71
6.6.1 Union de las variables en un dataframe	72
6.6.2 Tratamiento de los valores faltantes	73
6.6.3 Correlacion entre variables	75
6.6.4 Dataset final.....	79
6.6.5 Tratamiento de los valores outliers.....	81
6.7 ANALISIS DE LAS VARIABLES	83
6.7.1 Variable de respuesta (Pre_elec)	83
6.7.2 Variables de entrada	85
7. METODOS Y TECNICAS EMPLEADAS	91
7.1 DIVISION DE LOS DATOS PARA LOS MODELOS.....	92
7.2 ESTANDARIZACIÓN DE LOS DATOS	93
7.3 METRICAS DE VALIDACION DE LOS MODELOS.....	94
7.4 MODELO DE REGRESION LINEAL MULTIPLE	95
7.5 MODELO K VECINOS MÁS CERCANOS (KNN).....	98
7.6 MODELO ARBOL DE DECISION.....	101
7.7 MODELO RANDOM FOREST	103
7.8 MODELO XG BOOST	105
7.9 MODELO RED NEURONAL	107

8. ANALISIS DE LOS RESULTADOS OBTENIDOS.....	110
REGRESION LINEAL MULTIPLE	110
K VECINOS MÁS CERCANOS.....	111
ARBOLES DE DECISION.....	111
RANDOM FOREST.....	112
XG BOOST	113
RED NEURONAL.....	114
9. CONCLUSIONES.....	115
10. REFERENCIAS BIBLIOGRAFICAS	122
11. ANEXOS	124
ANEXO I: ARCHIVOS CSV	124
ANEXO II: EXPLORACION DE LAS VARIABLES	126
ANEXO III: UNION Y TRATAMIENTO DE LOS DATOS	177
ANEXO IV: CORRELACION ENTRE VARIABLES	209
ANEXO V: PROCESADO OUTLIERS (ELIM)	218
ANEXO VI: PROCESADO OUTLIERS (LIMT)	278
ANEXO VII: ESTUDIO VARIABLES	359
ANEXO VIII: REGRESION LINEAL (OUT-ELIM)	370
ANEXO IX: REGRESION LINEAL (OUT-LIM)	376
ANEXO X: KNN (OUT_ELIM)	383
ANEXO XI: KNN (OUT-LIM)	389
ANEXO XII: ARBOL DE DECISION	395
ANEXO XIII: RANDOM FOREST	401
ANEXO XIV: XG BOOST	410
ANEXO XV: RED NEURONAL 80-20	419
ANEXO XVI: RED NEURONAL 70-30	466

1. SÍNTESIS DEL TRABAJO

El sector eléctrico en España es un mercado altamente complejo debido a los numerosos factores que le afectan. Entre estos están:

- **El equilibrio que debe existir entre la generación y la demanda** en cada instante, ya que la energía eléctrica no puede ser almacenada en grandes cantidades.
- **Las fuentes de producción de energía eléctrica son muy diversas** (hidráulica, eólica, nuclear, solar, térmica, etc.).
- **La transición** que se está produciendo en los últimos años, **de fuentes de energía altamente contaminantes a fuentes muy poco contaminantes o totalmente limpias para la atmósfera**.
- **Cada tecnología** de producción **tiene un coste de producción** asociado, ya sea debido a las infraestructuras necesarias o para la propia producción de la energía.
- También existe un **coste ambiental** como es la cantidad de **emisiones de CO₂**.
- Las fuentes de **energía renovables dependen de las condiciones meteorológicas**.
- **La legislación del sector** también aumenta la complejidad en el sector.

Todo esto hace que **la predicción del precio de la electricidad se convierta en un elemento esencial en la toma de decisiones** en las compañías del sector, no sólo para desarrollar estrategias de generación para las **Productoras**, sino también para otros actores del mercado como **Comercializadoras**, que adquieren energía para su venta al **Consumidor final** o para los **Consumidores Directos en Mercado**, que son grandes consumidores (industria).

Gracias la capacidad computacional actual y al desarrollo de software capaz de manejar con efectividad grandes volúmenes de datos, se pueden emplear herramientas **de Deep Learnin para predecir el precio de la electricidad a largo plazo**.

Con este trabajo **lo que se pretende es analizar diferentes modelos predictivos para pronosticar el precio de la energía eléctrica en nuestro país y analizar cual de ellos son los mejores**.

En la documentación consultada se han encontrado **pocas referencias a estudios para la predicción de precios de la electricidad a más de tres meses**. Por este motivo, la **predicción del precio para el mercado a plazo es un territorio desaprovechado** y supone una oportunidad para realizar un estudio con el **objetivo de aportar nuevos mecanismos en la toma de decisiones** para las compañías del sector.

2. INTRODUCCIÓN

El objetivo final de este trabajo es desarrollar y comparar varios modelos predictivos para pronosticar el precio de la energía eléctrica (en adelante EE) en nuestro país y analizar cual de ellos es el mejor.

Para ello este documento se ha estructurado de la siguiente manera:

Para comenzar en el capítulo **1. SINTESIS DEL TRABAJO**, donde se realiza un breve resumen de la razón que me ha llevado a realizar este Trabajado Fin de Máster.

En el capítulo **2. INTRODUCCIÓN** se hace una descripción de la estructura empleada en la Memoria.

En el capítulo **3. ANTECEDENTES** se expondrá brevemente como se ha llegado a la situación actual del mercado eléctrico en España.

A continuación, en el capítulo **4. BUSINESS CASE**, se desarrollará el valor de negocio que puede suponer el presente estudio, así como las posibilidades de ampliación que pueden existir.

En cuanto a el capítulo **5. OBJETIVOS A ALCANZA** se expondrán los objetivos que se quieren alcanzar con el presente estudio.

El siguiente capítulo, **6. DATOS DE PARTIDA**, es fundamental para obtener unos buenos resultados en los modelos, y en el se indican las distintas fuentes de datos consultadas, las tablas obtenidas, el análisis de los datos, la exploración de las variables y el tratamiento de los datos que se han realizado para obtener el dataset de datos final que se empleará en los diferentes algoritmos.

Una vez se tenga el dataset de datos que se va a emplear en los modelos, en el capítulo **7. MÉTODOS Y TÉCNICAS EMPLEADOS** se explicarán los métodos elegidos, las razones por las que se han elegido para predecir el precio de la EE y se desarrollarán para obtener los resultados finales.

Cuando se finalice el estudio de los diferentes modelos se expondrán en el capítulo **8. ANÁLISIS DE LOS RESULTADOS OBTENIDOS** los resultados obtenidos y se analizarán, para que en la sección **9. CONCLUSIONES** se establezcan las conclusiones finales obtenidas de este estudio.

Los dos últimos capítulos, **10. REFERENCIAS** y **11. ANEXOS**, se emplearán para señalar las referencias de los documentos que se han tenido en cuenta para este trabajo, y para incluir el código empleado para la exploración y análisis de los datos, así como para desarrollar los distintos modelos predictivos.

2.1. MODULOS DEL MÁSTER RELACIONADOS CON EL TRABAJO FIN DE MÁSTER

A continuación, se relacionarán los distintos Módulos impartidos en el **Máster en Data Science y Business Analytics** con el contenido del **Trabajo Fin de Máster** (a partir de ahora **TFM**).

De esta manera el **TFM** se basará principalmente en los siguientes módulos impartidos:

- **MODULO I: Las Herramientas del Científico de Datos.**
- **MODULO II: Impacto y valor del Big Data.**
- **MODULO III: La Ciencia de Datos. Técnicas de análisis, minería y visualización.**
- **MODULO V: Estadística para Científicos de Datos.**
- **MODULO VII: Aprendizaje automático.**
- **MODULO VIII: Inteligencia artificial para la empresa.**

Como es lógico e ineludible, para desarrollar el **TFM**, hay que emplear los conocimientos adquiridos en el **MODULO I**, ya que es donde se desarrollan los conocimientos sobre los dos lenguajes de programación (**Python** y **R**) a través de los cuales se realizará la

recolección, la limpieza y el análisis de los datos y se desarrollarán los modelos de predicción.

Emplearemos los dos lenguajes de programación, R para realizar los primeros trabajos para el procesado de los datos, y Python para desarrollar los modelos de Machine Learning.

Tanto el **MODULO II: Impacto y valor del Big Data**, como el **MODULO III: La Ciencia de Datos. Técnicas de análisis, minería y visualización**, son básicos e imprescindibles para poder entender los módulos posteriores y para poseer los conceptos necesarios para poder actuar correctamente con los datos (obtención, limpieza, transformación y visualización).

El **MODULO V: Estadística para Científicos de Datos** es imprescindible para la comprensión del **análisis estadístico de los datos en el que se basan los algoritmos de Machine Learning**.

El **MODULO VI** es necesario ya que en él se exponen y desarrollan dos temáticas fundamentales en el **TFM** como son, los **algoritmos de Aprendizaje Automático**, más concretamente los empleados para resolver los **problemas de regresión**, y las técnicas de **Deep Learning**, en concreto las **redes neuronales**. Estas dos técnicas serán las empleadas **para realizar la previsión de los precios**.

Finalmente, el **MODULO VII** será también necesario porque en él se abordan las **técnicas para la toma de decisiones**, y más concretamente lo referente al **Aprendizaje Supervisado** en el que se tratan los **algoritmos para problemas de regresión y las redes neuronales**.

3. ANTECEDENTES

3.1 IMPORTANCIA DE LA ENERGIA ELECTRICA

La electricidad se puede utilizar para generar movimiento, calor o frío, luz, así como poner en marcha dispositivos electrónicos, sistemas de telecomunicaciones, sistemas de procesamiento de información, etc. Todas estas aplicaciones, utilizadas en la industria, el sector terciario, los hogares, hospitalares, medios de transporte, etc. funcionan como consecuencia de la circulación de una corriente eléctrica.

Por lo tanto, se puede decir que **la electrificación iniciada en el siglo XIX no sólo fue un proceso técnico, sino un verdadero cambio social de implicaciones extraordinarias**, por lo

que, **además de ser un servicio, es una necesidad básica** para poder realizar una gran cantidad de actividades **en el mundo actual**.

Hoy en día empleamos la energía eléctrica para iluminar las ciudades, los centros de trabajo y las viviendas, la utilizamos también en los procesos productivos tanto alimenticios como de materias primas, se emplea cada vez más para el transporte de personas y mercancías, es imprescindible para las comunicaciones (móviles, radios, televisiones, etc), se emplea también para producir calor o frío en edificios y viviendas, es necesario para todos los aparatos eléctricos existentes en las casas y en las oficinas, es decir, el ser humano no podría seguir viviendo en la actualidad sin el empleo de esta forma de energía.

No hay más que analizar las consecuencias catastróficas que provocaría una interrupción de energía eléctrica, para evidenciar con precisión la dependencia de nuestra sociedad: las fábricas tendrían que parar sus procesos productivos; no funcionarían los teléfonos, ordenadores e internet, las ciudades serían un caos sin semáforos y luz nocturna, no tendríamos abastecimiento de agua potable, no podríamos mantener los alimentos frescos como lo hacemos con los congeladores y neveras, los hospitales se paralizarían y la gran mayoría de pruebas médicas no se podrían realizar, y así un largo etcétera.

3.2 MARCO REGULATORIO

Red Eléctrica es la sociedad que ejerce como **transportista único y operador del sistema eléctrico español (TSO)**. Su misión consiste en garantizar en todo momento la seguridad y continuidad del suministro eléctrico y gestionar el transporte de energía en alta tensión.

El marco regulatorio de las actividades de negocio reguladas realizadas por Red Eléctrica lo establece tanto normativa europea como nacional.

A nivel nacional la **Ley 24/2013**, de 26 de diciembre, del Sector Eléctrico, es la **principal norma reguladora de las actividades de Red Eléctrica**, atribuyéndole el ejercicio de las actividades de transporte y operación del sistema, así como la función de gestor de la red de transporte.

La **retribución de las actividades de transporte de energía eléctrica y operación del sistema** está sujeta al **Real Decreto-ley 1/2019**, de 11 de enero, para adecuar las

competencias de la Comisión Nacional de los Mercados y la Competencia a las exigencias derivadas del derecho comunitario en relación a las Directivas 2009/72/CE y 2009/73/CE. Esta ley atribuye a la **Comisión Nacional de los Mercados y la Competencia (CNMC)**, entre otras, las competencias para aprobar la metodología, los parámetros retributivos, la base regulatoria de activos y la remuneración anual de la actividad del transporte, así como de la operación del sistema.

También relevante para las actividades de negocio de Red Eléctrica es la **Ley 17/2013**, de 29 de octubre, para la **garantía de suministro e incremento de la competencia en los sistemas insulares y extra-peninsulares**, en la que se establece que Red Eléctrica, en su calidad de operador del sistema de estos sistemas eléctricos, sea el titular de todas las nuevas instalaciones de bombeo.

A **nivel europeo** existen cuatro reglamentos y cuatro directivas, que conforman el conjunto de normas que rigen el sector eléctrico europeo. Dentro de él, tienen especial relevancia por su contenido:

- **El Reglamento (UE) 2019/943**, del Parlamento Europeo y del Consejo, de 5 de junio de 2019, relativo al mercado interior de la electricidad.
- **La Directiva (UE) 2019/944**, del Parlamento Europeo y del Consejo, de 5 de junio de 2019, sobre normas comunes para el mercado interior de la electricidad y por la que se modifica la Directiva 2012/27/UE.

Además de toda esta normativa de carácter general, existe una amplia **normativa de carácter técnico** que tiene por objeto regular las medidas necesarias para una adecuada gestión técnica del sistema eléctrico peninsular y de los sistemas eléctricos no peninsulares.

3.3 SISTEMA ELÉCTRICO ESPAÑOL

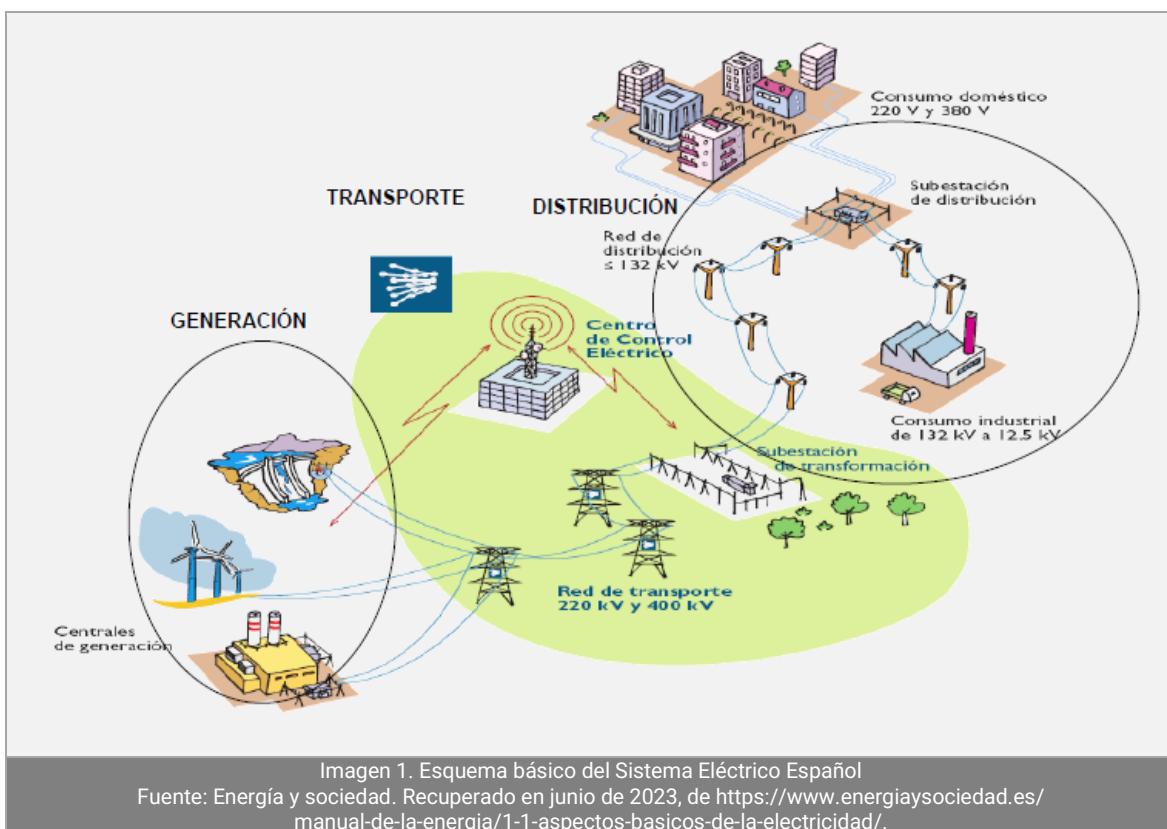
La **energía eléctrica** se puede obtener fácilmente de diferentes tipos de **energías primarias** (combustibles fósiles, nuclear, biomasa, agua, viento, sol, etc.) a partir de un proceso de **transformación** y, posteriormente, **transportarla** desde los lugares de generación hasta los **centros de consumo** a través de **líneas eléctricas** y cables subterráneos.

La **electricidad** se puede utilizar para generar movimiento, calor o frío, luz, así como poner en marcha dispositivos electrónicos, sistemas de telecomunicaciones, sistemas de procesamiento de información, etc. Todas estas aplicaciones, utilizadas en la industria, el sector terciario, los hogares, hospitales, medios de transporte, etc. funcionan como consecuencia de la circulación de una **corriente eléctrica**.

Sin embargo, también hay que tener en cuenta que la **energía eléctrica no se puede almacenar económicamente en grandes cantidades** (obligando a generarla al mismo ritmo que se consume en cada instante) y necesita que exista una continuidad eléctrica para su existencia. Esta continuidad es lo que define el circuito eléctrico y, si se interrumpe dicha continuidad, la circulación de la corriente eléctrica se interrumpe.

Estas dos características hacen que la disponibilidad de esta energía, necesaria en nuestra sociedad, se consiga en base a un **sistema muy complejo que integra un número muy elevado de componentes**, abarcando: fuentes de generación de electricidad, transformación, líneas eléctricas de transporte y distribución, maquinas eléctricas, sistemas de protección, control y gestión, circuitos eléctricos dentro de las viviendas, comercios e industrias, etc., todos ellos interconectados entre sí, conformando lo que se ha denominado como “**el Sistema Eléctrico**” o también “**la máquina más grande jamás construida por el hombre**”.

En el punto siguiente de este capítulo se presenta una breve descripción de cada una de las etapas que forman parte del sistema eléctrico español, que se pueden ver gráficamente en la Imagen 1.



3.4 ETAPAS DEL SISTEMA ELÉCTRICO ESPAÑOL

El sistema eléctrico español está compuesto por las siguientes etapas:

- **CENTRALES ELÉCTRICAS GENERADORAS:** la energía eléctrica se obtiene a partir de diferentes tipos de energías primarias como pueden ser el carbón, el gas, el agua, el viento, el sol, el combustible nuclear, etc., mediante un proceso de transformación que da lugar a diferentes tipos de **plantas productoras**, tales como las centrales hidroeléctricas, térmicas, nucleares, eólicas, solares, etc. Cada tecnología cuenta con ventajas y desventajas.
- **ESTACIONES TRANSFORMADORAS ELEVADORAS:** se ubican a la salida de las **centrales generadoras** y su misión es elevar la tensión de salida de dichas centrales, a un valor de tensión adecuado para el **transporte de la energía eléctrica a alta tensión**.
- **REDES DE TRANSPORTE:** son las líneas que unen las **estaciones transformadoras elevadoras** de las **centrales eléctricas** con las **subestaciones transformadoras reductoras**. Es decir, son las encargadas de realizar el transporte de energía a larga distancia y alta tensión. El desarrollo de la conectividad de las redes de transporte, tanto en el interior de

los países como en las interconexiones entre los mismos, ha permitido el planteamiento de mercados eléctricos de dimensión regional o internacional.

- **SUBESTACIONES TRANSFORMADORAS REDUCTORAS.** Cumplen tres funciones principales: son los **centros de interconexión** de todas las líneas entre sí, son los centros de transformación desde donde se alimentan las **líneas de distribución** que llegan hasta el consumo y son los centros en donde se instalan los **elementos de protección y maniobra del sistema**.
- **REDES DE DISTRIBUCIÓN:** son las líneas eléctricas de aproximación a los grandes centros de consumo (ciudades o instalaciones industriales de cierta importancia). En la mayoría de las ocasiones, estas redes suelen ser aéreas, aunque una vez que llegan a los núcleos urbanos, se utilizan líneas subterráneas.
- **CENTROS DE TRANSFORMACIÓN:** transforman los valores de media tensión de la red de distribución a valores aptos para el consumo en baja tensión. Este consumo puede ser extremadamente variable dependiendo de la hora del día, del día de la semana, época del año, país, etc.

3.5 ESTRUCTURA DE GENERACIÓN DE LA EE EN ESPAÑA

3.5.1 TECNOLOGÍAS DE PRODUCCIÓN DE EE

En España se emplean más de 20 tecnologías diferentes para la producción de EE. A continuación, pasamos a enumerarlas y a definirlas brevemente.

GENERACIÓN RENOVABLE		GENERACIÓN NO RENOVABLE	
Hidráulica		Turbinación bombeo	Residuos no renovables
Eólica		Nuclear	Motores diésel
Solar fotovoltaica		Ciclo combinado	Turbina de vapor
Solar térmica		Carbón	Turbina de gas
Otras renovables		Cogeneración	Fuel + Gas
Residuos renovables			
Hidroeólica			

Las **Energías renovables** son aquellas obtenidas de los recursos naturales y desechos renovables. Incluyen:

- **Hidráulica:** Producción debida a la circulación de un caudal de agua por un circuito hidráulico que salva un desnivel entre dos puntos, lo que se conoce comúnmente como salto, y en el que el agua va adquiriendo velocidad a medida que la energía potencial se va transformando parcialmente en energía cinética. La turbina es la encargada de transformar esa energía cinética en energía mecánica, para que el generador la transforme a su vez en energía eléctrica.
- **Eólica:** Esta fuente de EE se obtiene a partir de la fuerza del viento. A través de un aerogenerador que transforma la energía cinética de las corrientes de aire en energía eléctrica.
- **Hidroeólica:** Producción de energía eléctrica a través de la integración de un parque eólico, un grupo de bombeo y una central hidroeléctrica. El funcionamiento permite al parque eólico suministrar energía eléctrica directamente a la red y, simultáneamente, alimentar a un grupo de bombeo que embalse agua en un depósito elevado, como sistema de almacenamiento energético. La central hidroeléctrica aprovecha la energía potencial almacenada.
- **Solar fotovoltaica:** Luz solar convertida en electricidad mediante el uso de células solares, generalmente de material semiconductor que, expuesto a la luz, genera electricidad.
- **Solar térmica:** Se produce a partir del calentamiento de un fluido (normalmente agua) mediante radiación solar. Su uso en un ciclo termodinámico convencional que produce la potencia necesaria para mover un alternador para la generación de energía eléctrica.
- **Otras renovables:** Incluye la combustión de Biogás o biomasa, y también la hidráulica marina y geotérmica.
- **Residuos renovables:** Material orgánico no fósil de origen biológico resultante de los desechos sólidos urbanos y algunos desechos comerciales, e industriales no peligrosos. Se consideran renovables el 50% de los residuos sólidos urbanos (RSU).

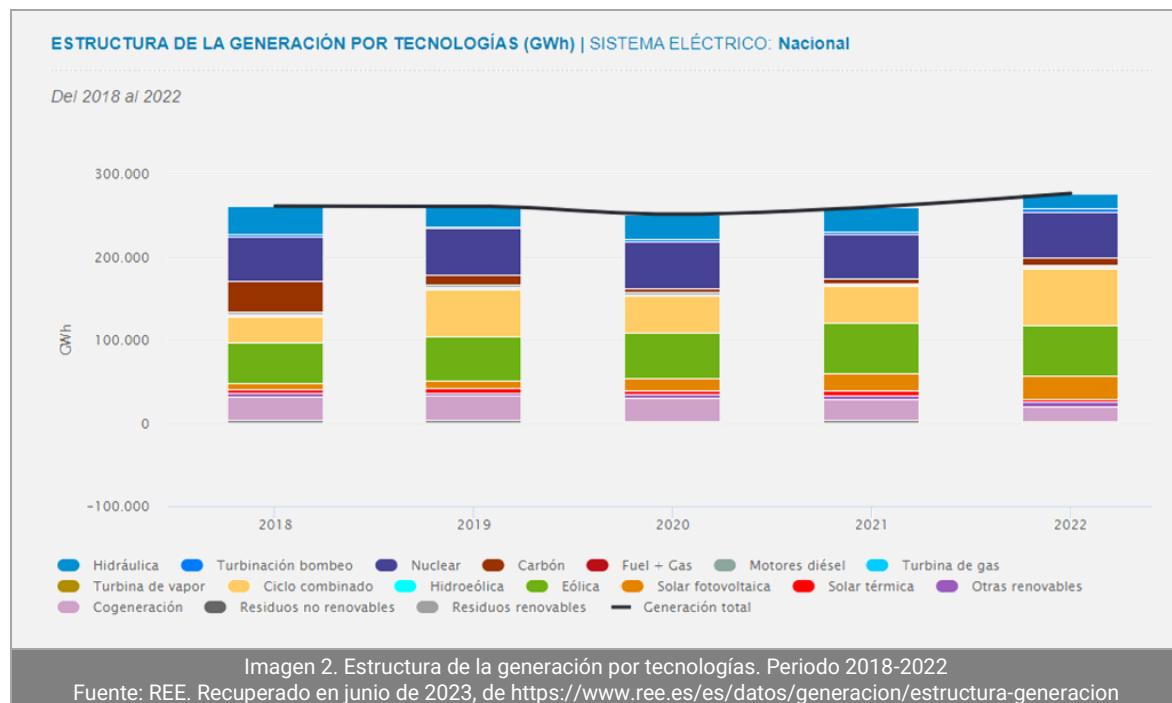
Las **Energías no renovables** son aquellas obtenidas a partir de combustibles fósiles (líquidos o sólidos) y sus derivados. Incluyen:

- **Turbinación bombeo:** Las centrales hidráulicas de bombeo, elevan agua desde un vaso inferior hasta otro situado a un nivel superior para su posterior turbinación.
- **Nuclear:** Se caracteriza por el empleo de combustible nuclear fisionable que mediante reacciones nucleares proporciona calor que a su vez es empleado, a través de un ciclo termodinámico convencional, para producir el movimiento de alternadores que transforman el trabajo mecánico en energía eléctrica.
- **Ciclo combinado:** Tecnología de generación de energía eléctrica en la que coexisten dos ciclos termodinámicos en un sistema con dos circuitos independientes: uno, cuyo fluido de trabajo es el vapor de agua, y otro, cuyo fluido de trabajo es un gas. En una central eléctrica el ciclo de gas genera energía eléctrica mediante una turbina de gas y el ciclo de vapor de agua lo hace mediante una o varias turbinas de vapor. El calor generado en la combustión de la turbina de gas se lleva a una caldera convencional o a un elemento recuperador del calor y se emplea para mover una o varias turbinas de vapor, incrementando el rendimiento del proceso. A ambas turbinas, de gas y vapor, van acoplados generadores eléctricos.
- **Carbón:** el carbón se emplea en las centrales térmicas para alimentar un ciclo termodinámico convencional y así mover un alternador que produce la EE.
- **Cogeneración:** Proceso mediante el cual se obtiene simultáneamente energía eléctrica y energía térmica y/o mecánica útil.
- **Residuos no renovables:** Esta materia procede de los residuos sólidos urbanos (RSU) y se emplea para la combustión en las centrales térmicas
- **Motores diésel:** Es otro tipo de central térmica cuya materia prima empleada para la generación de calor para su transformación en EE es combustible diésel.
- **Turbina de vapor:** El calor generado en la combustión se emplea para mover una o varias turbinas de vapor. Esta turbina va acoplada a un generador eléctrico.

- **Turbina de gas:** la turbina es un dispositivo de conversión de energía, que transforma la energía almacenada en el combustible en energía mecánica útil en forma de potencia de rotación, la cual se transforma en EE.
- **Fuel + Gas:** Consiste en el empleo de estas dos materias primas para la producción de energía mecánica para su posterior transformación en EE.
- otro tipo de central térmica cuya materia prima empleada para la generación de calor para su transformación en EE es combustible diésel.

3.5.2 EVOLUCIÓN DE LA PRODUCCIÓN DE EE

Se muestra a continuación un desglose gráfico de la **generación por tecnologías entre los años 2018 y 2022**. La fuente de estos datos es la web de Red Eléctrica Española. Recuperado en junio 2023, de <https://www.ree.es/es/datos/generacion/estructura-generacion> (la API de REE no permite obtener consultas mayores de 5 años).



Se incluye la misma consulta anterior, pero en forma de datos.

ESTRUCTURA DE LA GENERACIÓN POR TECNOLOGÍAS (GWh) SISTEMA ELÉCTRICO: Nacional					
	2018	2019	2020	2021	2022
Hidráulica	34.117	24.719	30.632	29.626	17.907
Turbinación bombeo	1.994	1.646	2.751	2.649	3.776
Nuclear	53.198	55.824	55.758	54.041	55.984
Carbón	37.277	12.671	5.021	4.983	7.765
Fuel + Gas	0	0	-	0	-
Motores diésel	3.178	2.836	2.399	2.517	2.548
Turbina de gas	1.049	671	407	424	657
Turbina de vapor	2.455	2.189	1.388	1.108	1.207
Ciclo combinado	30.044	55.242	44.023	44.500	68.137
Hidroeólica	24	23	20	23	23
Eólica	49.581	54.245	54.906	60.526	61.194
Solar fotovoltaica	7.766	9.252	15.302	20.981	27.902
Solar térmica	4.424	5.166	4.538	4.706	4.123
Otras renovables	3.557	3.618	4.482	4.720	4.657
Cogeneración	29.007	29.615	27.030	26.091	17.754
Residuos no renovables	2.435	2.222	2.016	2.239	1.900
Residuos renovables	874	890	726	878	878
Generación total	260.982	260.829	251.399	260.011	276.413

Imagen 3. Estructura de la generación por tecnologías. Periodo 2018-2022
Fuente: REE. Recuperado en junio de 2023, de <https://www.ree.es/es/datos/generacion/estructura-generacion>

Como se puede observar en la Imagen 3, en España se generaron aproximadamente 276.000 GWh en el año 2022. En la Imagen 2 se comprueba que la producción en estos últimos cinco años ha sido muy constante, con un leve crecimiento, exceptuando el año 2020 donde hubo un descenso debido al efecto de la pandemia por la Covid-19 (dichos efectos se detallan más adelante en el apartado **6.1 Intervalo temporal**).

En nuestro país vemos que existen más de **20 tecnologías diferentes de producción de electricidad**, aunque hay una gran diferencia entre las seis mayores productores y el resto.

MODELOS PARA LA PREDICCIÓN DEL PRECIO DE LA ELECTRICIDAD EN EL MERCADO ESPAÑOL

Analizamos a continuación las más importantes.

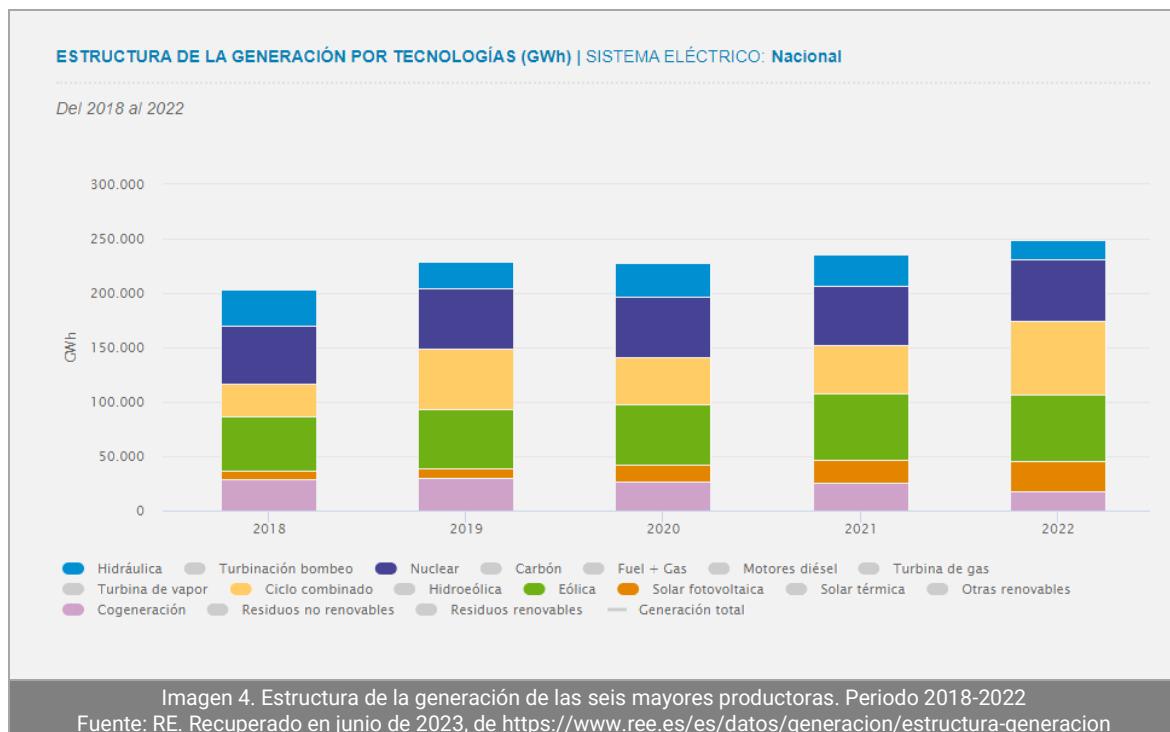


Imagen 4. Estructura de la generación de las seis mayores productoras. Período 2018-2022
Fuente: RE. Recuperado en junio de 2023, de <https://www.ree.es/es/datos/generacion/estructura-generacion>

Estas seis fuentes de producción alcanzan el 90% de la producción total nacional. La fuente de estos datos es la web de Red Eléctrica Española. Recuperado en junio 2023, de <https://www.ree.es/es/datos/generacion/estructura-generacion> (la API de REE no permite obtener consultas mayores de 5 años).

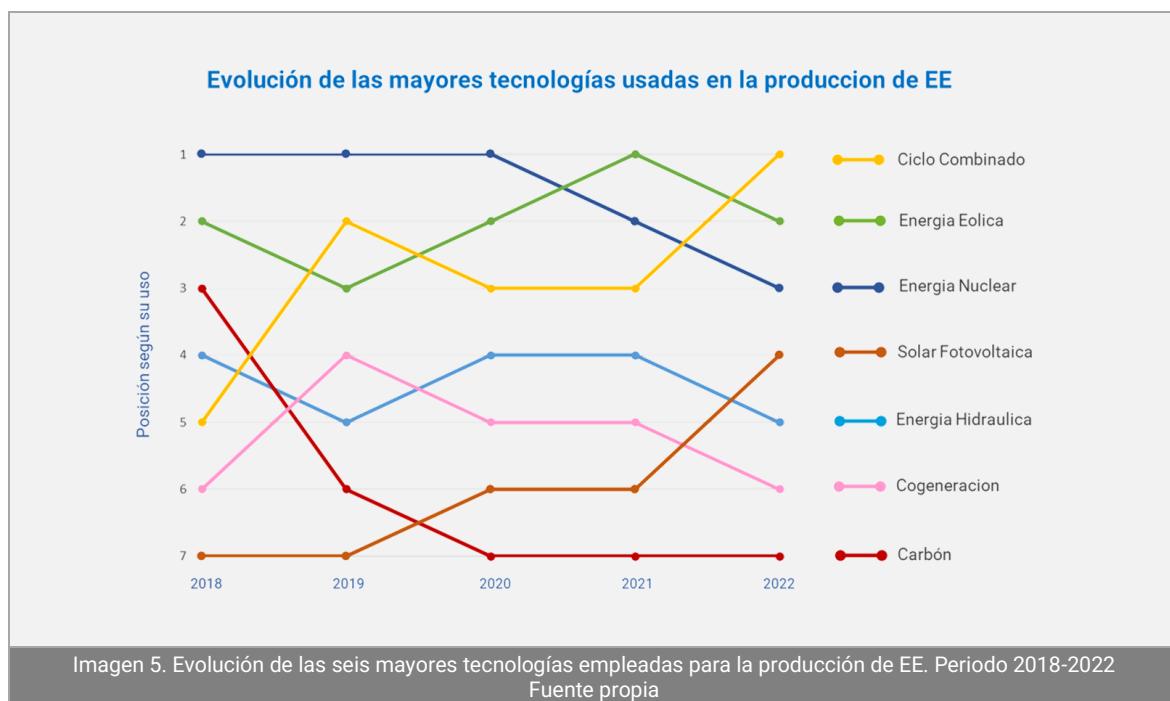


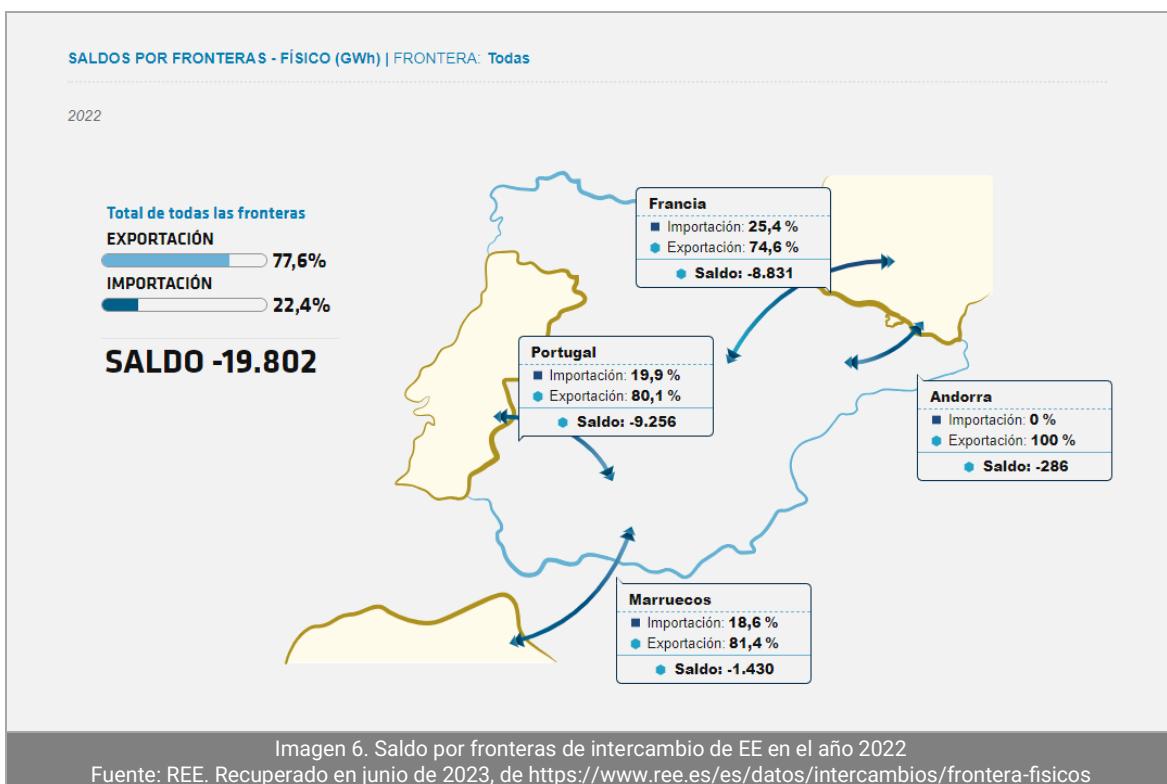
Imagen 5. Evolución de las seis mayores tecnologías empleadas para la producción de EE. Período 2018-2022
Fuente propia

Analizando una por una cada tecnología se observa lo siguiente:

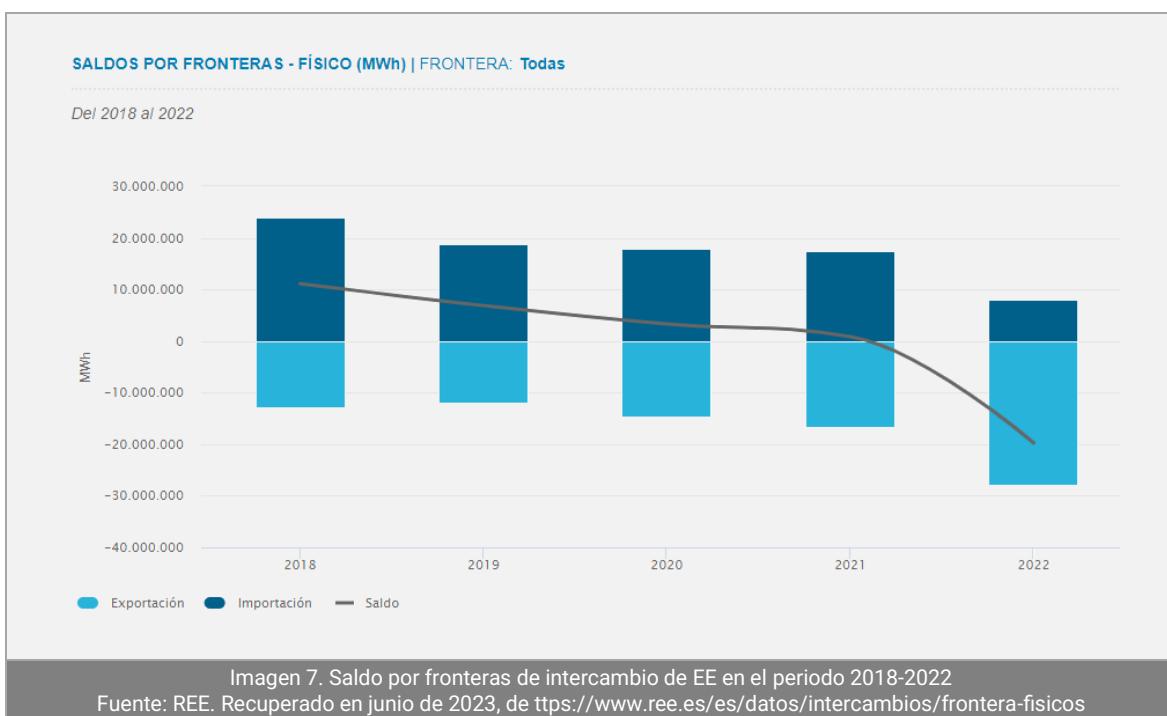
- La producción por **ciclo combinado** tiene una tendencia al alza, aunque mantienen oscilaciones entre el 11% y el 24%.
- La producción de **energía eólica** aumenta ligeramente cada año, desde el 19% al 22%.
- La **energía nuclear** tiene un peso casi constante de alrededor al 21% cada año.
- La **tecnología solar fotovoltaica** ha experimentado una crecida considerable pasando de un 3% a un 10%.
- La **energía hidráulica** es una tecnología que tiene grandes fluctuaciones, debido a que dependencia de las precipitaciones para acumular agua en las presas. Entre el 2018 y el 2022 ha oscilado entre 6,5% y el 13%.
- La **cogeneración** se ha mantenido en torno al 10%, aunque en el ultimo año ha descendido al 6,4% debido al crecimiento de otras tecnologías.
- EL empleo del **carbón** para la producción de EE actualmente está en séptimo lugar, aunque se puede comprobar en las imágenes anteriores que ha pasado de tener una gran importancia en el año 2018 siendo la tercera tecnología más empleada con un 14% a un residual 2,8% en el 2022.

Otro aspecto del sector eléctrico que hay que tener en cuenta para la producción de EE es el **intercambio con otros países**. Este intercambio es necesario no sólo por cuestiones de capacidad de generación, sino por mantener el equilibrio en cada momento entre producción y consumo.

MODELOS PARA LA PREDICCIÓN DEL PRECIO DE LA ELECTRICIDAD
EN EL MERCADO ESPAÑOL



Se observa que **en los últimos años España ha pasado de ser un país netamente importador a ser exportador**. El saldo neto tiene una tendencia a la exportación. Se ha pasado de un saldo neto de importación en el 2018 de 11 millones de MWh, a una exportación neta en el 2022 de casi 20 millones de MWh.



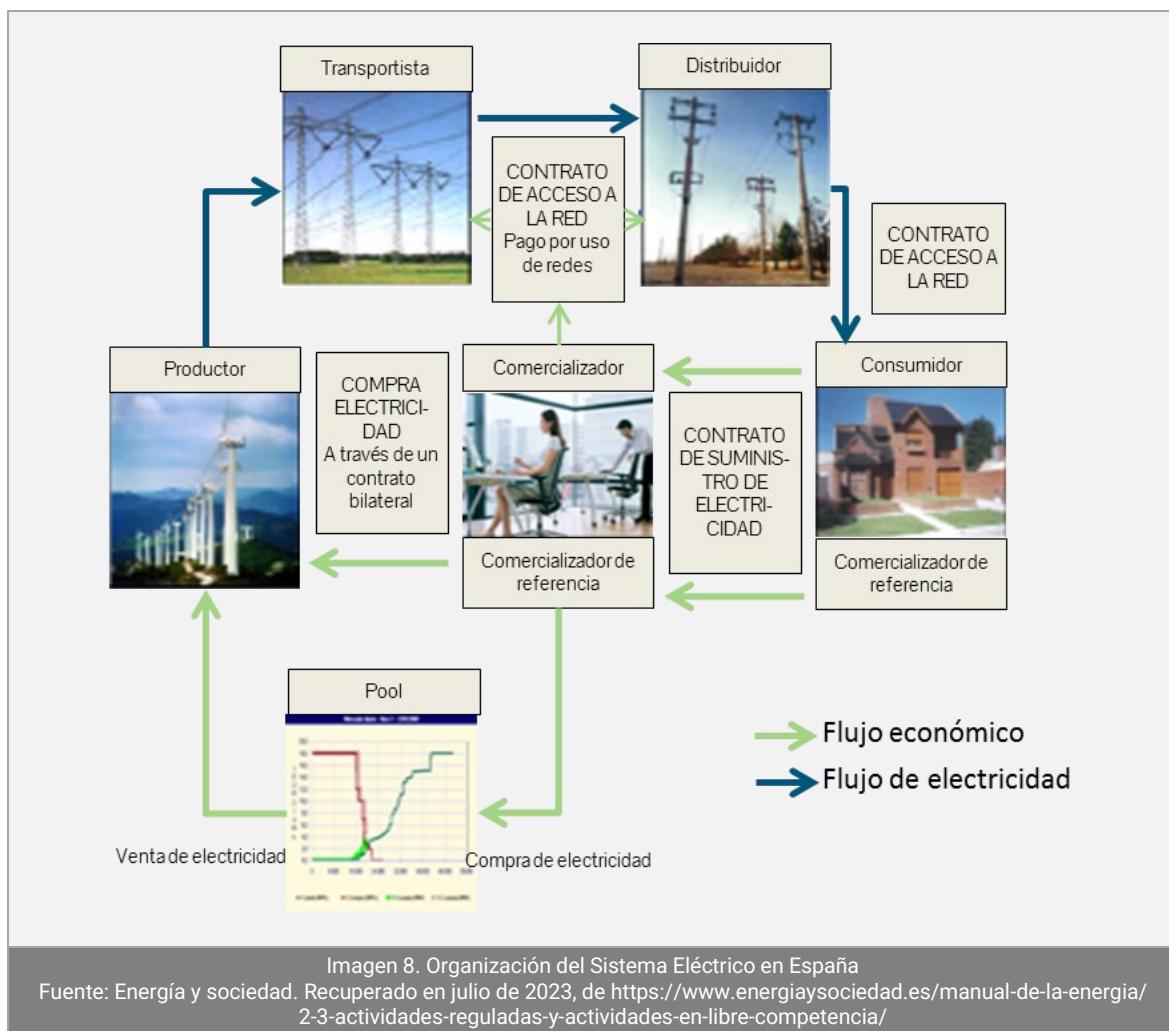
La fuente de estos datos de intercambio entre países es la web de Red Eléctrica Española. Recuperado en junio 2023, de <https://www.ree.es/es/datos/generacion/estructura-generacion> (la API de REE no permite obtener consultas mayores de 5 años).

3.6 AGENTES DEL MERCADO

La actual Ley del Sector Eléctrico (LSE) establece los siguientes sujetos participantes en el sector eléctrico:

- **Productores:** Cuya función es generar energía eléctrica, así como construir, operar y mantener las centrales de producción.
- **Transportista:** Tiene la función de transportar energía eléctrica, así como construir, mantener y maniobrar las instalaciones de transporte. Como se ha indicado anteriormente, Red Eléctrica de España (REE) es la encargada de estas labores.
- **Distribuidores:** Tienen la función de distribuir la energía eléctrica, así como construir, mantener y operar las instalaciones de distribución destinadas a situar la energía en los puntos de consumo.
- **Comercializadores:** Adquieren energía para su venta a los consumidores, a otros sujetos del sistema o para realizar operaciones de intercambio internacional.
- **Consumidores:** Son los compradores de la energía para su propio consumo. Si adquieren energía directamente en el mercado de producción se denominan Consumidores Directos en Mercado.
- **Gestores de cargas del sistema:** son aquellos que, siendo consumidores, están habilitados para la reventa de energía eléctrica para servicios de recarga energética, es decir, desarrollan la actividad destinada al suministro de energía eléctrica para la recarga de vehículos eléctricos.
- **Operador del Mercado Ibérico (OMI):** Gestiona el mercado ibérico de electricidad (MIBEL). Esta gestión distingue:
 - Gestión del mercado ibérico al contado (mercado spot), que está encomendada a OMI-Polo Español, S.A. (OMIE).

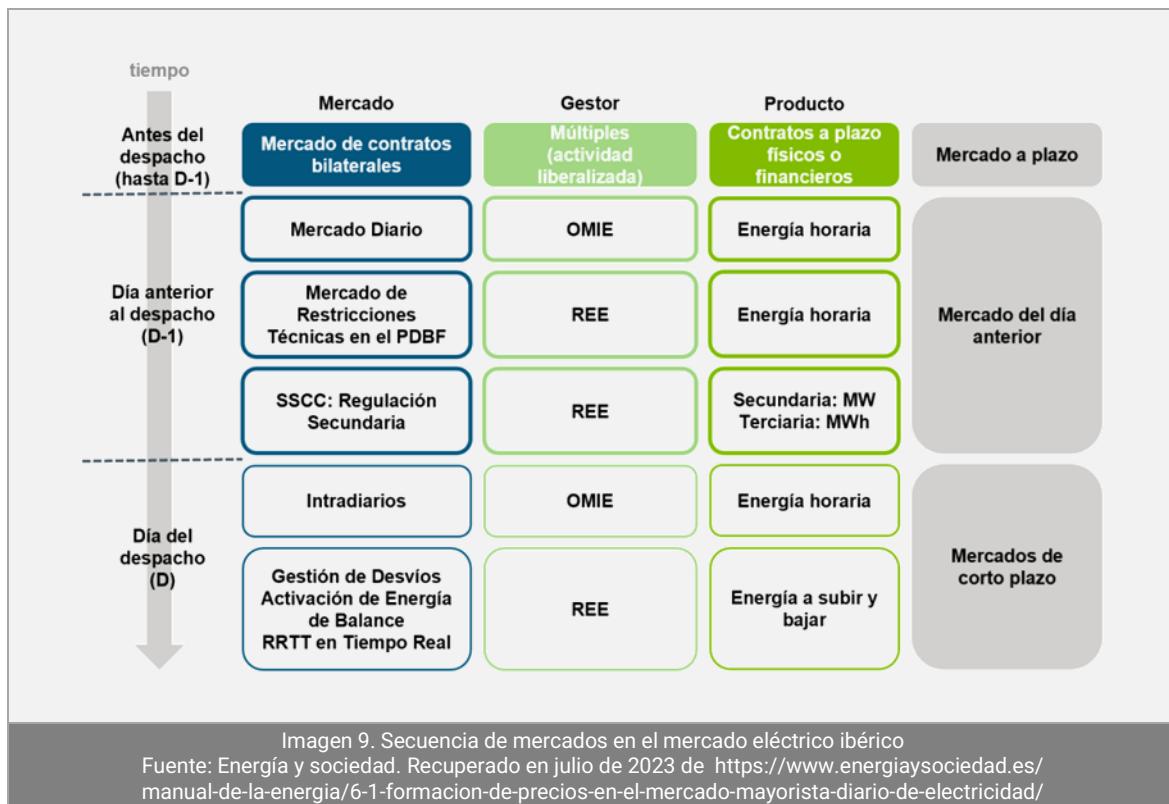
- Gestión de la Plataforma de Negociación del mercado de derivados (mercado a plazo), que es responsabilidad de OMIPolo Portugués, S.A. (OMIP).
- **Operador del sistema:** Es Red Eléctrica de España (REE), que tiene como función principal llevar a cabo las actividades asociadas a la operación técnica del Sistema Eléctrico, garantizando la continuidad y seguridad del suministro eléctrico y la correcta coordinación de los sistemas de producción y transporte.



3.7 MERCADO ELECTRICO ESPAÑOL

El mercado eléctrico es el conjunto de plataformas de negociación en las que se contrata energía eléctrica para su entrega en diferentes horizontes temporales, que pueden ser **a plazo** (para las próximas semanas, meses, trimestres o años) o **al contado** (para el día siguiente o las horas siguientes).

El mercado de electricidad en España, al igual que en otros países, se organiza en una secuencia de mercados en los que generación y demanda intercambian energía y reservas para distintos plazos como se puede observar en la imagen siguiente.



En los **mercados a plazo** los agentes intercambian contratos con períodos de entrega de distinta duración (anual, trimestral, mensual, etc.) con una antelación de días, semanas, meses e incluso años antes del momento en que la energía sea generada y consumida.

En el caso de los **mercados diario**, al llegar al día D-1 (un día antes de que la energía sea generada y consumida), los agentes intercambian energía para cada una de las horas del

día D en el mercado diario organizado por el Operador del Mercado Ibérico-Polo Español (OMIE). Además, dentro de las 24 horas anteriores al momento de generación y consumo, los agentes pueden ajustar sus posiciones contractuales comprando y vendiendo energía en los mercados intradiarios, también gestionados por el OMIE.

En el muy corto plazo, **mercados intradiarios** (desde unas pocas horas hasta unos pocos minutos antes de la generación y consumo) los generadores, y en algunos casos también la demanda, ofrecen una serie de servicios al Sistema en varios mercados organizados por el Operador del Sistema (REE). Estos servicios son necesarios para que la generación se iguale exactamente a la demanda en todo momento, manteniendo así al Sistema en equilibrio físico y con un nivel de seguridad y calidad de suministro adecuado.

3.8 FORMACION DEL PRECIO DE LA ENERGIA ELECTRICA _____

3.8.1. FORMACIÓN DEL PRECIO EN EL MERCADO MAYORISTA A PLAZO _____

Los **mercados a plazo** de electricidad son un conjunto de mercados en los que **se negocian contratos de compra-venta de electricidad con plazos de entrega de la energía superiores a 24 horas**.

En el largo y medio plazo, los agentes negocian diferentes tipos de contratos, con periodos de entrega de distinta duración (año, trimestre, mes, semana, etc.).

Estos mercados a plazo cumplen un papel crucial en un mercado liberalizado desarrollado. Cuando son suficientemente profundos (la oferta/demanda es lo suficientemente amplia como para que los agentes que acuden a él no se encuentren con limitaciones significativas respecto a la cantidad que pueden comprar/vender) y líquidos (cuando un agente puede comprar o vender cantidades significativas sin alterar el precio del mismo), permiten a los agentes compradores y vendedores gestionar sus riesgos, al tiempo que facilitan la competencia en los mercados mayorista y minorista.

A modo de ejemplo, un comercializador que deba adquirir energía para abastecer a sus clientes, si no existiesen los **mercados a plazo profundos y líquidos**, tendría que adquirir esta energía en el **mercado diario**, cuyo precio es desconocido en el momento de ofertar a sus clientes, así, estaría expuesto al riesgo de que el precio en el mercado diario resulte

más elevado que el que consideró a la hora de ofertar a sus clientes, corriendo por tanto el riesgo de incurrir en pérdidas.

Los mercados a plazo que negocian electricidad con referencia al mercado español son los siguientes:

- El **mercado no organizado de contratos bilaterales** (conocido como OTC), en el que se negocian contratos físicos y financieros.
- Los **mercados organizados de futuros eléctricos** gestionados por OMIP con sede en Portugal, EEX-ECC con sede en Alemania y BME con sede en España.

Como en cualquier otro mercado, **el precio se determina por el cruce entre la curva de oferta** (integrada por todas las ofertas que realizan los vendedores) **y la curva de demanda** (integrada por todas las ofertas que realizan los compradores).

De acuerdo a la teoría económica, el precio esperado del mercado diario es el coste de oportunidad de los contratos a plazo, por lo que el **precio del mercado a plazo refleja el precio del mercado diario esperado a futuro**.

3.8.2. FORMACIÓN DEL PRECIO EN EL MERCADO MAYORISTA DIARIO ---

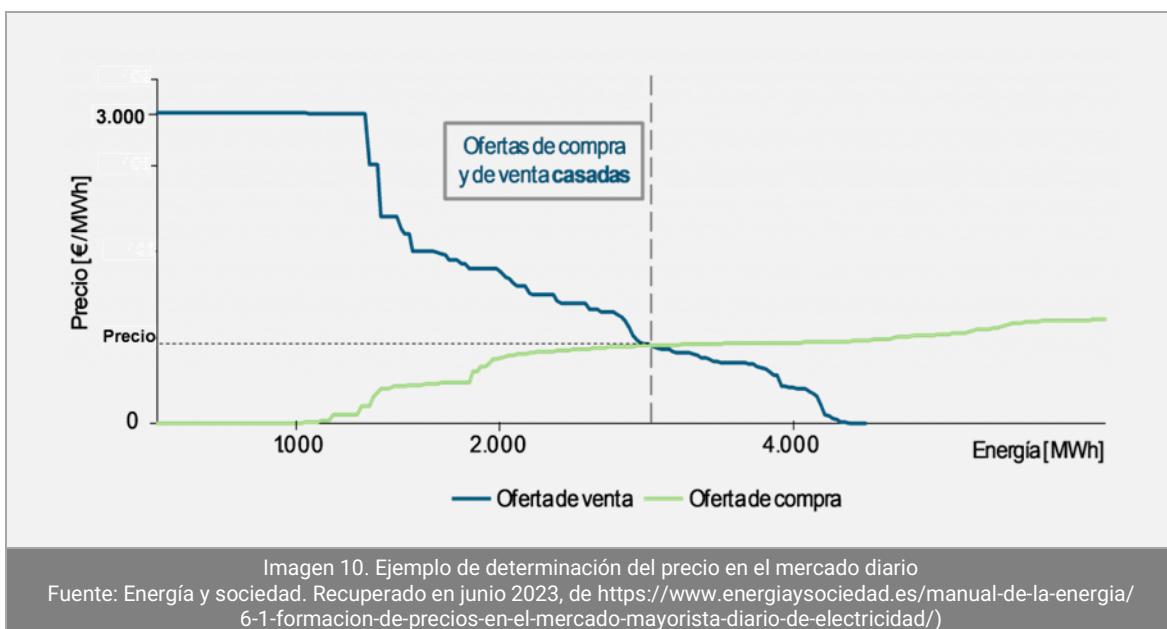
El **mercado diario español** de electricidad está acoplado a los mercados de Portugal y Francia. Para ello, se organiza de acuerdo con lo dispuesto en la normativa europea y española. El mercado está gestionado por OMIE, entidad privada que se encarga de garantizar que la contratación se lleva a cabo en condiciones de transparencia, objetividad e independencia.

El mercado diario se celebra el día anterior al de la entrega de la energía y en él compradores y vendedores intercambian energía para cada una de las horas del día siguiente. Así, en este mercado en realidad hay 24 productos diferentes (energía en cada una de las 24 horas del día siguiente) de la siguiente manera:

- Los **vendedores** (generadores y comercializadores que actúen como importadores) presentan ofertas de venta y los **compradores** (comercializadores que revendan su energía en el mercado minorista o la destinen a la exportación, y consumidores

finales que actúen directamente en el mercado mayorista) presentan **ofertas de compra** a OMIE para cada hora del día siguiente.

- Con estas ofertas, OMIE construye las curvas de oferta y demanda de cada hora del día siguiente.
- **Del cruce de las curvas de las ofertas de venta y de compra, resulta el precio del mercado para cada hora del día siguiente** y se identifican las ofertas “casadas” (ofertas de venta y de compra que se convierten en compromisos firmes de entrega de energía), tal y como se observa en la siguiente imagen



3.9 ESTADO DEL CONOCIMIENTO

Según señala Weron, R. (2014). Electricity price forecasting: A review of the state-of-the-art with a look into the future. *International Journal of Forecasting*, 1030-1081, publicado en la Revista internacional de pronósticos, durante los últimos 23 años se han probado una variedad de métodos e ideas para la previsión del precio de la electricidad (EPF), con diversos grados de éxito.

En este capítulo, sustentado en el artículo anteriormente citado de Rafael Weron, intentaremos explicar de forma breve la complejidad de las soluciones disponibles, sus

fortalezas y debilidades, y las oportunidades y amenazas que ofrecen las herramientas de pronóstico o que se pueden encontrar.

3.9.1 INTRODUCCIÓN

Hasta los principios de la década de los 90 el sector de la energía era monopolios controlados por los gobiernos, pero a partir de esos años el proceso de desregulación y la introducción de mercados competitivos remodelaron ese panorama. Actualmente en muchos países del mundo, la electricidad se negocia bajo las reglas del mercado mediante contratos al contado y derivados.

Sin embargo, la electricidad es un bien muy especial. Es económicamente no almacenable y la estabilidad del sistema de energía requiere un equilibrio constante entre la producción y el consumo. Estas características únicas y específicas conducen a una dinámica de precios que no se observa en ningún otro mercado, exhibiendo estacionalidad a nivel diario, semanal y anual, y picos de precios abruptos, de corta duración y generalmente imprevistos. Esto ha alentado a los investigadores a intensificar sus esfuerzos en el desarrollo de mejores técnicas de pronóstico.

A nivel corporativo, **las previsiones de precios de la electricidad se han convertido en un costo fundamental para los mecanismos de toma de decisiones de las empresas energéticas.**

Las previsiones de precios desde unas pocas horas hasta algunos meses se han vuelto de particular interés para los administradores de carteras de energía. Un generador, una empresa de servicios públicos o un gran consumidor industrial que sea capaz de pronosticar los precios mayoristas volátiles con un nivel razonable de precisión puede ajustar su estrategia de licitación y su propio programa de producción o consumo para reducir el riesgo o maximizar las ganancias en el día siguiente.

En los últimos años se han probado una variedad de métodos e ideas para la previsión del precio de la electricidad (EPF), con diversos grados de éxito.

Según el artículo Weron R. (2014), se pueden clasificar las técnicas de predicción en función de dos factores:

- el **horizonte de planificación**,
- y la **metodología aplicada**.

A continuación, repasamos estos dos enfoques.

3.9.2 HORIZONTES DE PREVISIÓN ---

Es habitual hablar de previsión de precios de la electricidad a corto, medio y largo plazo, pero no hay consenso en la literatura sobre cuáles deberían ser realmente los umbrales.

- El **corto plazo** generalmente comprende entre unos minutos hasta varios días como mucho. Estas predicciones suelen utilizarse en el mercado diario e intradiario.
- El **medio plazo** comprende desde varios días hasta varios meses. El campo de aplicación de estas predicciones es el cálculo de balances, gestión de riesgos y fijación de precios de derivados.
- El **largo plazo** comprende períodos que abarcan desde meses hasta años, y se suelen aplicar estas predicciones para estudios estratégicos de inversión y localización de plantas de generación, entre otros.

3.9.3 DESCRIPCIÓN GENERAL DE LAS METODOLOGÍAS DE MODELADO ---

Existen diversas metodologías que se han desarrollado para analizar y predecir los precios de la electricidad. Algunos de ellos son mejores, algunos son peores, pero todos tienen muchas cosas en común. Para su clasificación tomamos como punto de partida la clasificación de Weron R. (2014) con cinco grupos de modelos.

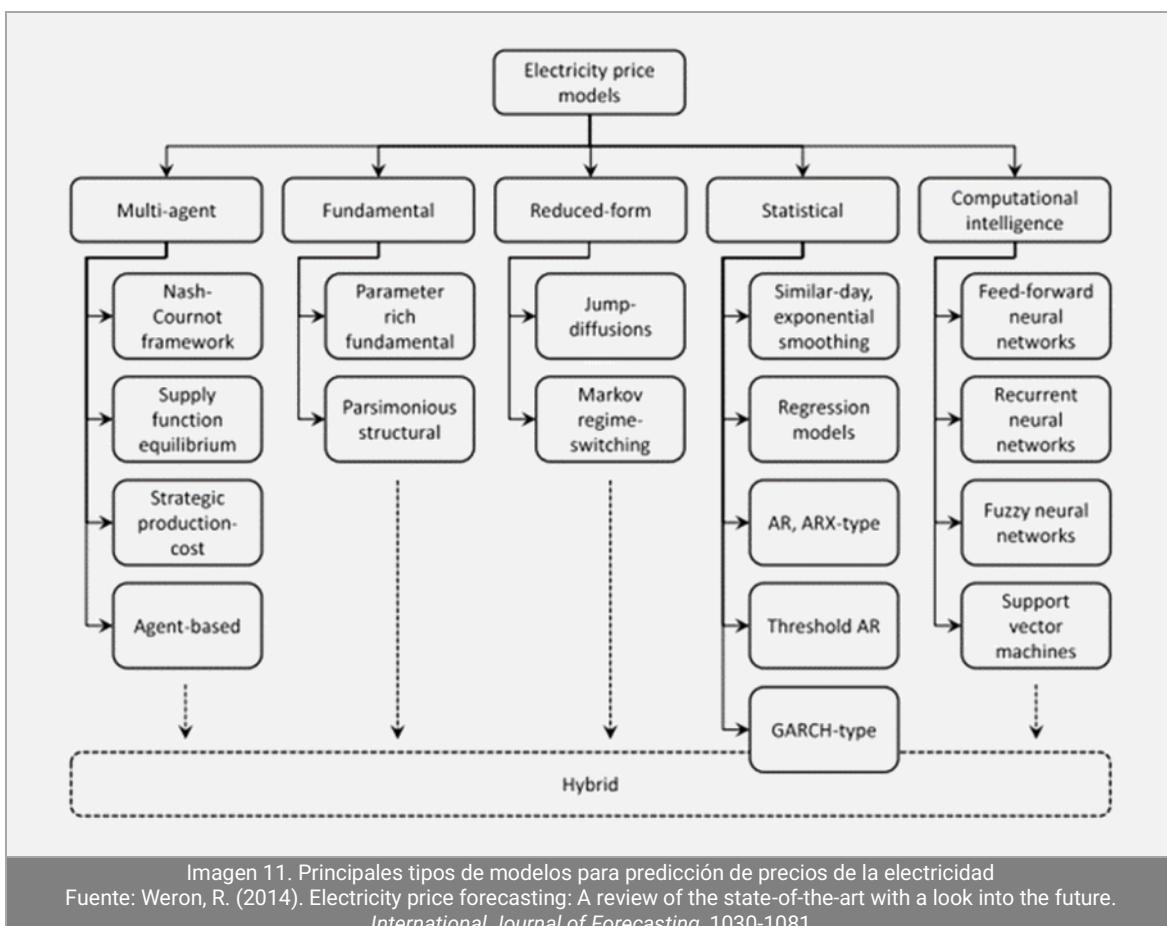


Imagen 11. Principales tipos de modelos para predicción de precios de la electricidad
 Fuente: Weron, R. (2014). Electricity price forecasting: A review of the state-of-the-art with a look into the future.
International Journal of Forecasting, 1030-1081

Debemos mencionar que muchos de los enfoques de modelado y pronóstico de precios son soluciones *híbridas*, que combinan técnicas de dos o más de los grupos enumerados anteriormente.

A continuación, se resume brevemente la orientación de cada grupo de modelos:

- **Multi-agente** (simulación multi-agente, equilibrio y teoría de juegos): son modelos que simulan un sistema de agentes (que serán los generadores de energía) que interaccionan unos con otros, y construyen el proceso de precios haciendo coincidir la oferta y la demanda en el mercado. Son modelos muy flexibles que se aplican para analizar comportamientos estratégicos en el mercado, pero para obtener resultados adecuados requieren a cambio asumir hipótesis acerca de los agentes que intervienen y sus estrategias, por lo que dichas hipótesis deben estar bien fundamentadas y contrastadas para obtener resultados adecuados. Suelen aplicarse para definir aspectos cualitativos más que cuantitativos. Este tipo de

modelos son adecuados para mercados regulados donde hay poca incertidumbre en los precios (Querol, 2019). Desarrollo de un modelo de predicción del precio de la energía eléctrica para el mercado a plazo mediante redes neuronales. Valencia: Universidad Politécnica de Valencia. E.T.S.I. Industriales.

- **Fundamentales o estructurales:** describen la evolución del precio a través de factores físicos y económicos. **Dependen fuertemente de la calidad de los datos de partida y su periodicidad.** Como usualmente estos datos son de carácter semanal o mensual, los modelos que se obtienen son más apropiados para predicciones a medio plazo que a corto plazo.
- **Reducidos** (cuantitativos, estocásticos): utilizan parámetros estadísticos distribuidos en el tiempo. **Este tipo de modelos no buscan tanto predecir el precio con precisión, sino replicar el comportamiento del precio y su correlación con el precio de otros.** Estos modelos se emplean para la evaluación y gestión de riesgos.
- **Estadísticos** (econométricos, análisis técnico): aplican de forma directa técnicas estadísticas para la previsión del mercado o de la demanda de electricidad en base a datos históricos. Como consecuencia de ello, **la calidad de los datos de partida influye fuertemente en la precisión de los resultados.** Algunos autores proponen estos modelos para detectar patrones e indicadores, más que para predecir el precio. Destacar que **estos modelos funcionan mejor cuando el mercado se muestra en valores normales que cuando se producen fuertes fluctuaciones.** Por ello hay numerosos estudios que proponen técnicas de filtrado de datos y ajuste de componentes estocásticas.

Estos modelos son atractivos porque se puede adjuntar alguna interpretación física a sus componentes, lo que permite a los ingenieros y operadores de sistemas comprender su comportamiento. A menudo son criticados por su capacidad limitada para modelar el comportamiento (generalmente) no lineal de los precios de la electricidad y las variables fundamentales relacionadas; sin embargo, en aplicaciones prácticas, su rendimiento es comparable al de sus alternativas no lineales

Los métodos estadísticos se suelen emplear en combinación con modelos más complejos como: ARIMA, ARFIMA, SARIMA, ARMA, GARCH, ARX, ARMAX, ARIMAX, SARIMAX y ARIMA-GARCH (Querol, 2019).

- **Inteligencia computacional** (estadística no lineal): **las técnicas aplicadas en estos modelos combinan la capacidad de aprendizaje, la evolución de los datos y la aparente falta de relaciones de los mismos para generar sistemas complejos capaces de adaptarse.** En este grupo se incluyen los modelos basados en **redes neuronales** como son: redes neuronales de avance, redes neuronales recurrentes, redes neuronales difusas y las máquinas de vectores de soporte. Tienen como **principal ventaja su capacidad para manejar problemas complejos y no lineales, reflejando bien la volatilidad y los picos del mercado.**

Adicionalmente, desde hace pocos años se están desarrollando investigaciones que aplican los modelos tipo árbol (Pórtoles, González, & Moguerza, 2018. Electricity price forecasting with dynamic trees: A benchmark against random forest approach. Energies, 11,1588), (Smarra, F., Jain, A., de Rubeis, T., Ambrosini, D., D'Innocenzo, A., & Mangharam, R. (2018). Data-driven model predictive control using random forests for building energy optimization and climate control. Applied Energy, 1252-1272) y (Ugarte, A. R. (2017). Predicción de precios de energía eléctrica utilizando árboles dinámicos. Madrid: Universidad Politécnica de Madrid. E.T.S.I. Industriales). Predicción de precios de energía eléctrica utilizando árboles dinámicos. Madrid: Universidad Politécnica de Madrid. E.T.S.I. Industriales) entre otros. Este último trabajo utiliza árboles dinámicos, cuya aplicación en el mercado eléctrico español ha sido escasa hasta el momento, y los presenta como una herramienta con gran potencial, para tener en consideración junto con las técnicas basadas en series temporales que se usan en la actualidad (Ugarte A.R., 2017).

4. BUSSINES CASE

La dependencia de la sociedad actual sobre la electricidad, los factores que influyen en su producción y consumo, los costes de las materias primas, los factores climatológicos o la legislación vigente del sector eléctrico, son solo unos pocos factores que afectan en el precio final de la EE.

Estos factores hacen del sector eléctrico un mercado altamente complejo en el que, tal y como indican muchos investigadores (Bunn, D. W. (2004). *Modelling prices in competitive electricity markets*. Chichester: John Wiley), (Eydeland, A., & Wolyniec, K. (2003). *Energy and power risk management*. Hoboken, NJ: Wiley) o (Weron, R. (2006). *Modeling and forecasting electricity loads and prices: a statistical approach*. Chichester: Wiley), una herramienta de **predicción del precio de la electricidad se ha convertido en una pieza clave en la toma de decisiones en las compañías del sector**, no sólo para desarrollar estrategias de generación para las **Productoras**, sino también para otros actores del mercado como **Comercializadoras**, que adquieren energía para su venta al **Consumidor final**.

Las previsiones de precios desde unas pocas horas hasta algunos meses **se han vuelto de particular interés para los administradores de carteras de energía**. Un generador, una empresa de servicios públicos o un gran consumidor industrial, que sea capaz de pronosticar los precios mayoristas volátiles con un nivel razonable de precisión, puede ajustar su estrategia de licitación y su propio programa de producción o consumo para reducir el riesgo o maximizar las ganancias en el día siguiente.

En la documentación consultada se han encontrado **pocas referencias a estudios para la predicción de precios de la electricidad a más de tres meses**. Por este motivo, la **predicción del precio para el mercado a plazo es un territorio desaprovechado** y supone una oportunidad para realizar un estudio con el **objetivo de aportar nuevos mecanismos en la toma de decisiones** para las compañías del sector.

La **aplicación principal** de este trabajo es el de **facilitar diferentes herramientas de predicción de precios** que pueden ser útiles para **Productores, Comercializadores y Consumidores**, aunque también puede ser de interés para otros agentes como **Gestores de carga del Sistema, Distribuidores, entidades financieras y organizaciones de consumidores**, entre otros.

5. OBJETIVOS A ALCANZAR ---

El **objetivo principal** de este Trabajo Fin de Máster es analizar diferentes modelos predictivos para pronosticar el precio de la energía eléctrica en nuestro país, y analizar cual de ellos es el que ofrece mejores resultados.

Se pretende alcanzar este objetivo partiendo de una serie de **variables**, como son los precios de las materias primas que se emplean para la producción de electricidad (gasóleo, gas natural y carbón), la meteorología (temperatura, viento y agua embalsamada), producción diaria por cada tipo de tecnología, la demanda eléctrica de los consumidores y otros factores (derechos de emisión de CO₂ y situación económica del país), y empleando técnicas de **análisis de datos**, de **aprendizaje automático** y de **Deep Learning** (redes neuronales).

Del artículo de investigación realizado por González C., Mira-McWilliams J. y Juárez I. (2015) Evaluación de variables importantes y pronóstico del precio de la electricidad basado en modelos de árboles de regresión: clasificación y árboles de regresión, Bagging y Random Forests. IET Generation, Transmission & Distribution, 9(11), 1120–1128, se pueden extraer los siguientes **errores medios absolutos en porcentaje (MAPE) en las los precios de la electricidad de diversos algoritmos estudiados.**

MÉTODO o MODELO/S EMPLEADO/S	TIPO DE MERCADO	DATOS EMPLEADOS	ERROR MEDIO DE PREDICIÓN	FUENTE
Método híbrido de árboles de regresión y red de función de base radial normalizada	A corto plazo (con una hora de antelación)	Mercado de Nueva Inglaterra	10.36%	(1)
Red neuronal	A corto plazo (con una hora de antelación)	Mercado de Nueva York, Australia y España	5.76%	(2)
Técnicas ponderadas de vecinos más cercanos	Diario (con un día de antelación)	Mercado español	9.80%	(3)
Redes neuronales y lógica difusa	A plazo (con una semana de antelación)	Mercado español	9.44%	(4)
Análisis factorial dinámico estacional	Diario (con un día de antelación)	Mercado español	11.00%	(5)
Modelo de factor dinámico	Diario (con un día de antelación)	Mercado español	7.39%	(6)
Análisis factorial dinámico estacional heterocedástico	Diario (con un día de antelación)	Mercado español	5.76%	(7)

(1) Mori, H., Awata, A.: 'Data mining of electricity price forecasting with regression tree and normalized radial basis function network'. Proc. of the IEEE Int. Conf. on Systems, Man and Cybernetics, 2007, ISIC, (doi:10.1109/ICSMC.2007.4414228)

(2) Neupane, B., Perera, K.S., Aung, Z., Woon, W.L.: 'Artificial neural network-based electricity price forecasting for smart grid deployment'. Int. Conf. on Computer Systems and Industrial Informatics (ICCSII), 18–20 December, 2012, (ISBN: 978-1-4673-5155-3. doi: .119/ICCSII.2012.6454392)

- (3) Troncoso, A., Riquelme, J.M., Gómez, A., Martínez, J.L., Riquelme, J.C.: 'Electricity market price forecasting based on weighted nearest neighbors techniques', IEEE Trans. Power Syst., 2007, 22, pp. 1294–1301
- (4) Catalão, J.P.S., Pousinho, H.M.I., Mendes, V.M.F.: 'Short-term electricity prices forecasting in a competitive market by a hybrid intelligent approach', Energy Convers. Manag., 2011, 52, pp. 1061–1065
- (5) Alonso, A., García-Martos, C., Rodríguez, J., Sánchez, M.J.: 'Seasonal dynamic factor analysis and bootstrap inference: application to electricity market forecasting', Technometrics, 2011, 53, pp. 137–151
- (6) García-Martos, C., Rodríguez, J., Sánchez, M.J.: 'Forecasting electricity prices by extracting dynamic common factors: application to the Iberian Market', IET Gener. Transm. Distrib., 2011, 1, pp. 1–10
- (7) García-Martos, C., Rodríguez, J., Sánchez, M.J.: 'Forecasting electricity prices and their volatilities using unobserved components', Energy Econ., 2011, 33, pp. 1227–1239

Imagen12. Errores medios en diferentes estudios de predicción de precios de EE

Fuente: González C., Mira-McWilliams J. y Juárez I. (2015)

Observando los errores medios, podemos afirmar que **las técnicas actuales consultadas obtienen un error que está entre un 5,76 y un 11%**. Por este motivo se considerará **como un buen resultado para los modelos estudiados en este TFM, los valores de MAPE que estén por debajo de ese 11% y posean un porcentaje de precisión alto.**

El presente Trabajo de Fin de Máster se han decidido aplicar las metodologías más extendidas actualmente y que han sido impartidas en el Master como son: **regresión lineal múltiple, k vecinos más cercanos, árboles de decisión, random forest, XG Boost y, por supuesto, las redes neuronales.**

6. DATOS DE PARTIDA

En este capítulo se definen y analizan los datos con los que se han entrenado, probado y validado los modelos.

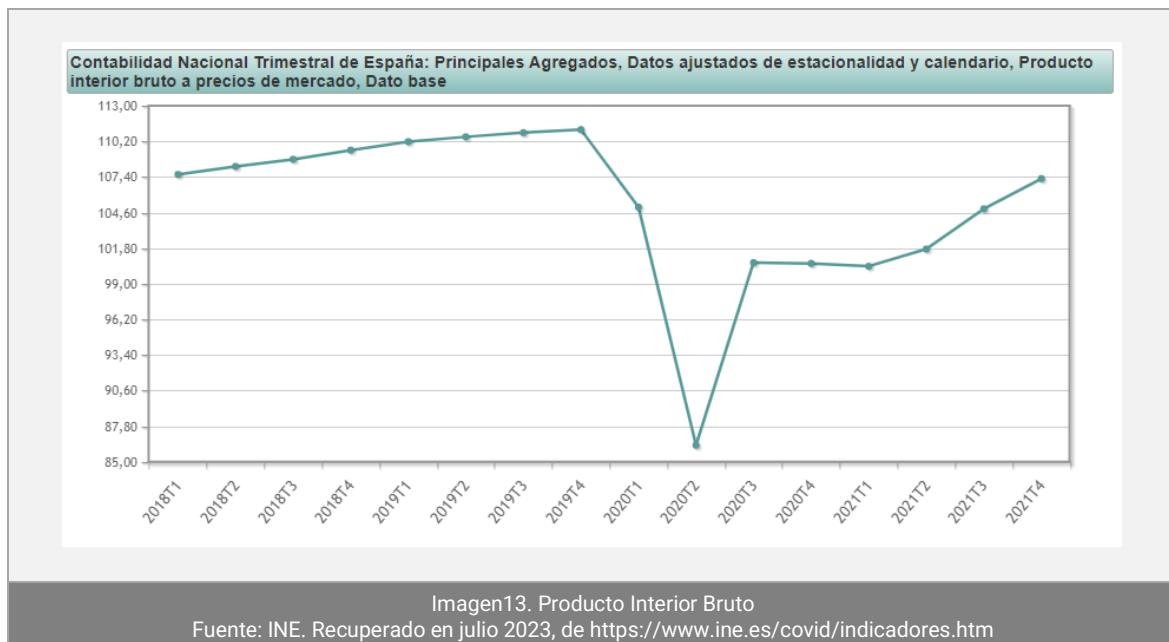
6.1 INTERVALO TEMPORAL

El intervalo de fechas al que corresponden los datos de partida es importante, ya que el conjunto de datos que alimente a los modelos no sólo debe ser suficientemente **extenso**, sino también **representativo**.

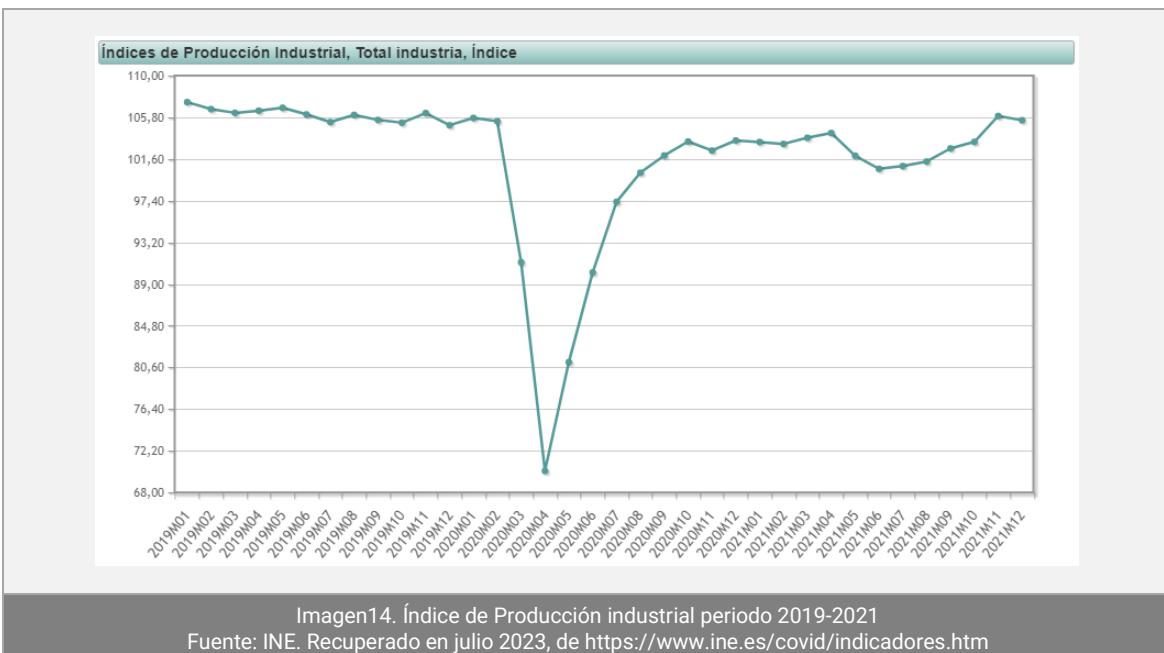
Para cumplir con estas dos premisas se ha establecido como criterio tomar un **horizonte temporal de 10 años**.

De esta forma se han obtenido los datos de las distintas variables consideradas entre los años **2012 y 2022**, pero para que el estudio no se vea distorsionado por la **anomalía que supuso la pandemia del Covid-19**, se ha excluido el año **2020** de dicho intervalo.

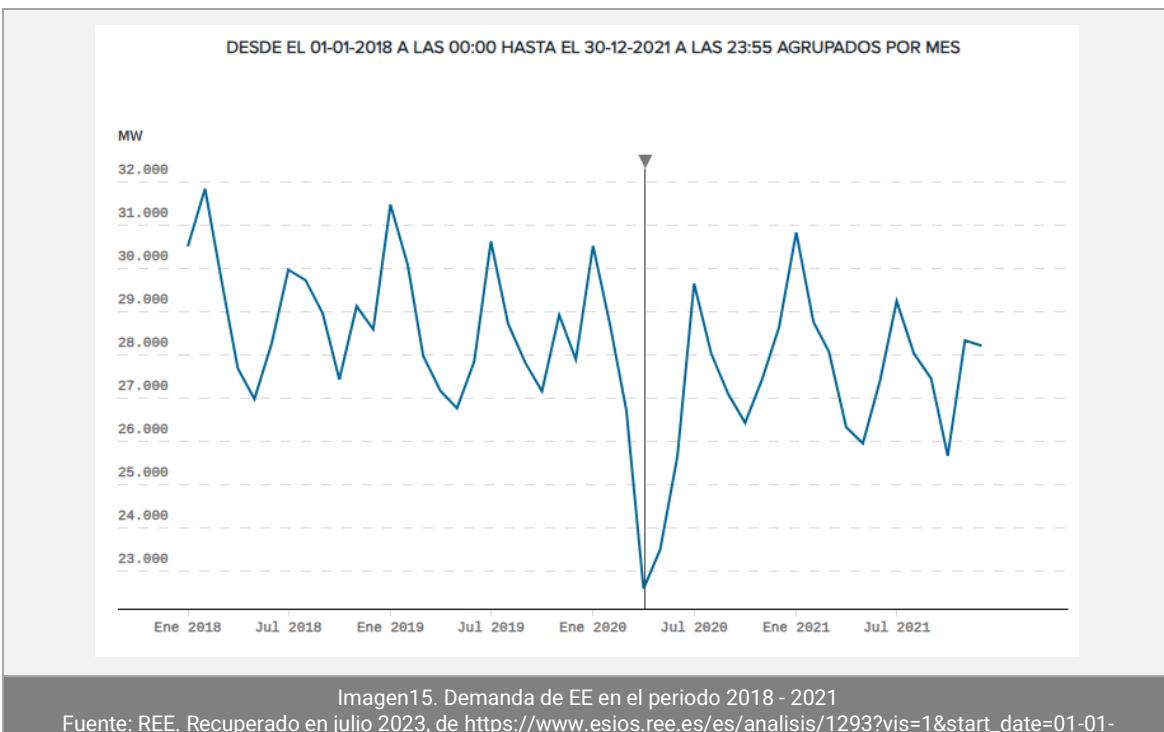
La **pandemia por Covid-19** tuvo un impacto incuestionable en todos los sectores socioeconómicos a nivel mundial, y **dicho impacto se refleja también en el sector eléctrico**. A nivel de la economía española, resultan muy ilustrativos los siguientes gráficos, donde se aprecia la brusca caída de los indicadores con el comienzo de la pandemia, en el primer trimestre de 2020.



MODELOS PARA LA PREDICCIÓN DEL PRECIO DE LA ELECTRICIDAD EN EL MERCADO ESPAÑOL



Centrando el efecto de la pandemia sobre el sector eléctrico, se muestra a continuación la demanda real en España que hubo entre los años 2018 y 2021. Se aprecia claramente el fuerte descenso de la demanda eléctrica que se produjo al inicio del confinamiento y la progresiva recuperación de la misma, aunque sin llegar a los niveles del año anterior.



6.2 DIA DE LA SEMANA

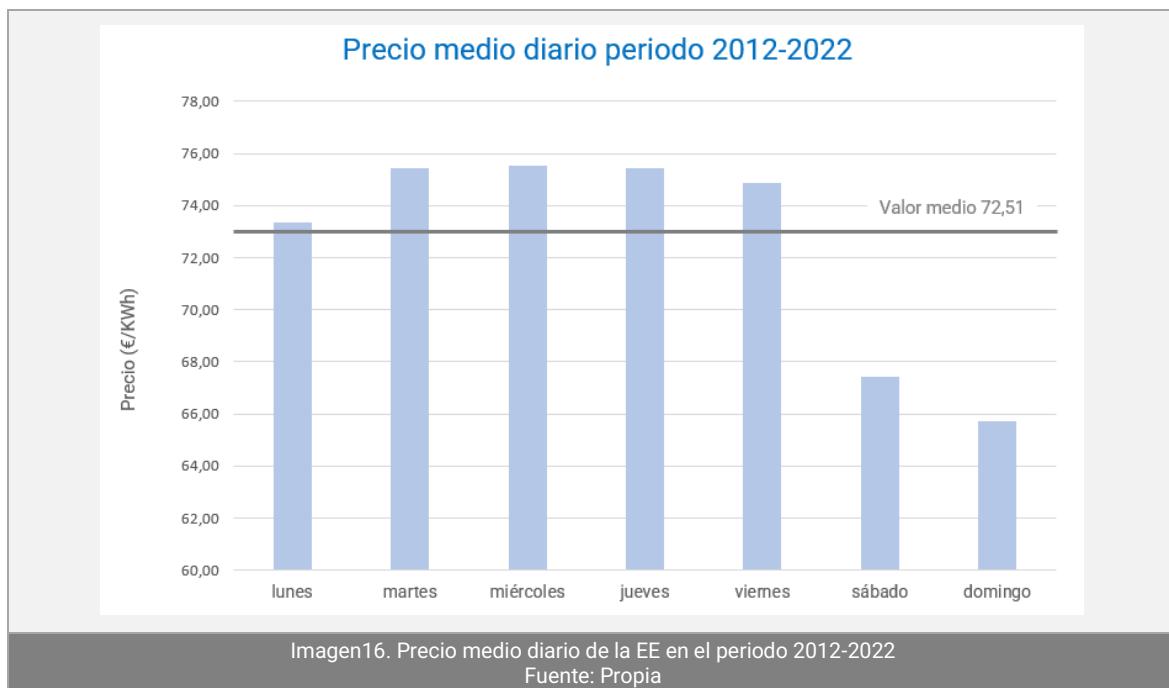
En el **mercado regulado** los precios van alterándose cada día en función de cuánta energía está demandando los consumidores, es decir, **el precio se fija según la oferta y la demanda**. En los picos de demanda, el precio es más caro. En los valles, resulta más barato.

Según esto, en España se han establecido tres períodos diferentes de la siguiente manera:

- **Horas punta:** de 10 a 14 h y de 18 a 22 h, donde el precio es el más caro.
- **Horas llanas:** de 8 a 10 h, de 14 a 18 h y de 22 a 24 h.
- **Horas valle:** de 00 a 8 h de la mañana de lunes a viernes y las 24 horas de los fines de semana y festivos nacionales. Donde el precio es el más barato.

Las horas pico se producen por la gran demanda que existe, de esta manera las de la mañana coinciden con el pleno rendimiento en el que se encuentran las industrias y empresas españolas, y el de la tarde, con la vuelta a casa.

Tomando el **precio medio para cada día de la semana durante todo periodo definido** para el estudio (2012-2022), se obtienen unos valores que reflejan claramente que **el precio de la electricidad los sábados y los domingos es muy inferior al del resto de días de la semana**, teniendo los sábados un precio un 7% inferior a la media de los días de la semana y los domingos hasta un 9% como se puede comprobar en el gráfico siguiente.



Por este motivo, el **día de la semana** se tomará como una variable en los modelos estudiados en este trabajo.

6.3 VARIABLES CONSIDERADAS

La finalidad de este estudio es pronosticar el **precio medio diario de la EE en España**, por lo que queda este objetivo definido como **variable dependiente** del estudio, cuyo código será **Pre_elec**.

Para establecer las **variables independientes** empleadas en los diferentes modelos de cálculo, hay que partir de los **factores que influyen sobre la variable objetivo**. De este modo pasamos a ver dichos factores (asociadas en grupos) y a definir las **variables predictoras**.

- **Factores temporales.** Como ya se ha analizado en el apartado anterior, el día de la semana influye en el precio de la EE, teniendo un valor casi estable entre lunes y viernes, y descendiendo casi un 10% los fines de semana. Esta variable se denominará **día de la semana** y su código será **Dia_sem**.
- **Demandado:** El precio de venta de la EE está ligada directamente a la demanda existente. Esto es muy fácilmente comprobable con la influencia del día de la semana estudiada con anterioridad, en donde los fines de semana se produce una bajada de demanda y por eso los precios son más bajos. Esta variable se denominará **demandado media diaria (Deman)** y se medirá en MW. La fuente de datos de esta variable ha sido Red Eléctrica Española (REE).
- **Precio de las materias primas.** En capítulos anteriores se ha visto que algunas tecnologías de producción de EE necesitan una materia prima para producir EE en centrales térmicas. De esta manera tendremos 3 materias primas que se emplean para este fin:
 - **Petróleo:** la variable para este factor se denominará **precio del petróleo (Prec_petr)** y los valores empleados serán los precios en dólares por barril de Brent futuro a tres meses negociado en el mercado ICE.

- **Gas natural:** su valor se determina en el mercado UK Natural Gas del mercado ICE a 2 meses, medidos en £/MillionBtu, y su nombre será **precio del gas natural (Prec_gas)**.
- **Carbón:** el **precio del carbón (Prec_car)** empleado es este estudio es el de Rotterdam Coal Futures del mercado ICE que se mide en \$/tonelada.
- **Producción de EE según tecnología:** es importante considerar todas las tecnologías empleadas en España porque, como se ha visto, cada una cuenta con sus particularidades y sus costes, lo que repercutirá en el precio final de la electricidad según varie su producción. Estos datos están tomados de Red Eléctrica Española (REE) y están medidos en GWh. Se han clasificado en los siguientes grupos:
 - **Producción en parques eólicos (Prod_eol):** en este grupo están incluidos los parques hidroeólicos.
 - **Producción en parques solares (Prod_sol):** en el que se encuentran las centrales fotovoltaicas y centrales solares térmicas.
 - **Producción hidráulica y turbinación por bombeo (Prod_hidr):** que se refiere a los saltos de agua.
 - **Producción de otras fuentes renovables (Prod_ofr):** que incluye la producción por residuos renovables.
 - **Producción en centrales nucleares (Prod_nucl).**
 - **Producción en centrales convencionales de gasóleo (Prod_pet):** en las que se produce electricidad gracias a un motores diésel.
 - **Producción en centrales con turbina de gas (Prod_gas).**
 - **Producción en centrales convencionales de carbón (Prod_carb).**
 - **Producción en centrales de ciclo combinado (Prod_conv).**
 - **Producción en centrales de cogeneración (Prod_cog).**
 - **Producción de otras fuentes no renovables (Prod_noren):** que engloba a la producción empleando fuel y gas, turbina de vapor y residuos no renovables.

- **Factores climáticos:**

- **Temperatura (Temp):** Para tener en cuenta el posible efecto que producen las altas y bajas temperaturas en la demanda del consumo eléctrico, se introducirá las temperaturas medias diarias de las 5 ciudades más pobladas de España. Estas ciudades, ordenadas de mayor a menor número de habitantes, son: Madrid, Barcelona, Valencia, Sevilla y Zaragoza. Todas estas grandes urbes, con excepción de Zaragoza, están rodeadas de localidades que también soportan una gran población, por lo que se han tenido también en cuenta.

Esta variable estará medida en grados centígrados.

- **Velocidad del viento (Vel_vien):** Teniendo en cuenta la necesidad de considerar el tipo de generación eólica, se ha decidido incluir el posible efecto del viento. Estudiando la potencia instalada de las Comunidades Autónomas, se han seleccionado las cinco más importantes, que son Castilla y León, Castilla La Mancha, Aragón, Galicia y Andalucía, que representan alrededor del 80% de la potencia eólica instalada en el país.

Se han obtenido las velocidades medias diarias del viento para las ciudades más representativas de las CCAA mencionadas anteriormente, en lo que a cercanía de parques eólicos se refiere, es decir, Valladolid, Albacete, Zaragoza, La Coruña y Huelva.

- **Reservas hidráulicas (Res_hidr):** Este dato depende de la pluviometría, pues recoge la cantidad de agua almacenada para producción de energía hidroeléctrica. Los datos se han obtenido a partir de los boletines hidrológicos que publica el Ministerio para la Transición Ecológica del Gobierno de España. Para estimar el posible efecto de la producción hidráulica en los precios diarios, se ha empleado como aproximación el nivel de los embalses.

- **Otros factores:**

- **Precio por Derechos de emisión de CO₂ (Der_CO₂):** Las empresas están obligadas a gestionar un número de bonos (también conocidos como derechos o créditos), que representan el derecho a emitir una cantidad

determinada de CO₂. Las compañías que necesiten aumentar las emisiones por encima de su límite deberán comprar créditos a otras compañías que contaminen por debajo del límite que marca el número de créditos que le ha sido concedido. Su valor se determina en el mercado Carbon Emissions Futures ICE y están expresados en €/tonelada CO₂.

- **Situación socio-económica del país:** para conocer la situación de España se puede consultar los datos del índice bursátil IBEX-35 que mide el comportamiento conjunto de las 35 empresas con mayor liquidez. La variable la llamaremos **Índice IBEX-35** y su código será **Ibex**. Este valor se mide en puntos bursátiles.
- **Intercambio de EE con otros países (Int_ee):** Este intercambio es necesario no sólo por cuestiones de capacidad de generación, sino para mantener el equilibrio en cada momento entre producción y consumo, y por lo tanto influye en el precio final de la EE. Este valor se expresa en GWh.

FACTOR	CODIGO	NOMBRE	UNIDADES DE MEDIDA
VARIABLE DEPENDIENTE			
	Pre_elec	Precio medio diario de la EE en España	€/MWh
VARIABLES PREDICTORAS			
Temporal	Dia_sem	Día de la semana	
Demanda	Dem	Demand media diaria	MWh
Precios materias primas	Prec_petr	Precio del petroleo	\$/barrel de Brent
	Prec_gas	Precio del gas natural	£/MillionBtu
	Prec_carb	Precio del carbon	\$/tonelada
Producción EE según tecnología	Prod_eol	Producción en parques eólicos	GWh
	Prod_sol	Producción en parques solares	GWh
	Prod_hidr	Producción hidráulica y turbinación por bombeo	GWh
	Prod_ofr	Producción de otras fuentes renovables	GWh
	Prod_nucl	Producción en centrales nucleares	GWh
	Prod_pet	Producción en centrales convencionales de gasóle	GWh
	Prod_gas	Producción en centrales con turbina de gas	GWh
	Prod_carb	Producción en centrales convencionales de carbór	GWh
	Prod_comb	Producción en centrales ciclo combinado	GWh
	Prod_cog	Producción en centrales de cogeneración	GWh
	Prod_noren	Producción de otras fuentes no renovables	GWh

MODELOS PARA LA PREDICCIÓN DEL PRECIO DE LA ELECTRICIDAD

EN EL MERCADO ESPAÑOL

	Temp	Temperatura	
Meteorología	Vel_Vien	Velocidad del viento	°C
	Res_hidr	Reservas hidráulicas	m/s
Otros factores	Der_CO2	Precio Derechos de emisión de CO₂	%
	Ibex	Situación socio-económica del país	€/tonelada CO ₂
	Int_ee	Intercambio de EE con otros países	Puntos bursátiles
			GWh

Imagen17. Resumen de las variables consideradas en el estudio
Fuente: Propia

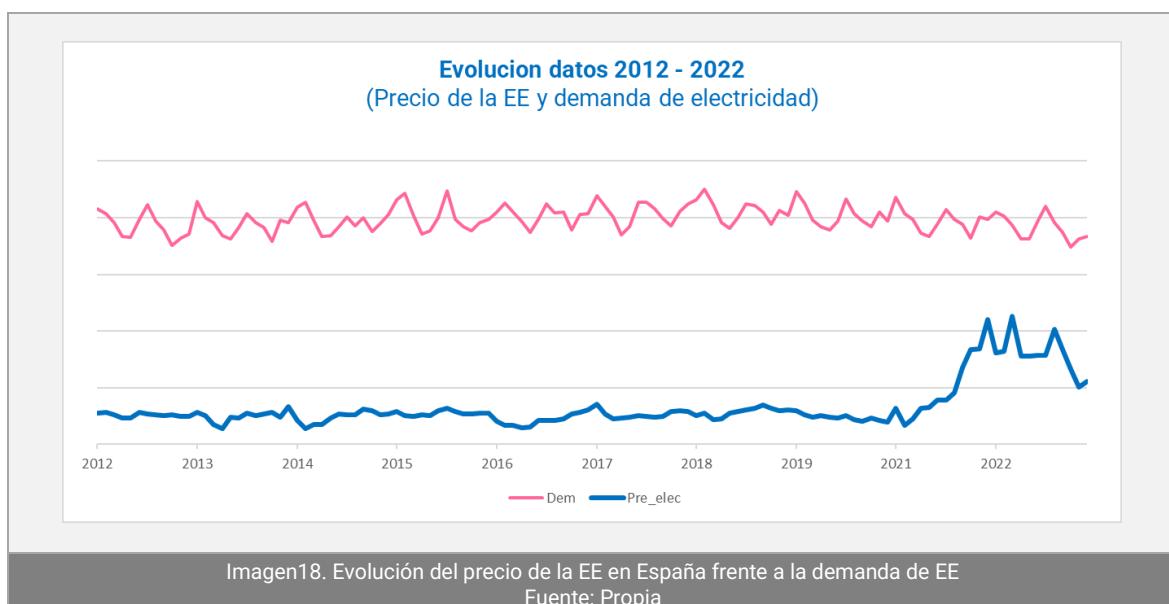
En cuanto a las **fuentes** de las que se han obtenido los datos las indicamos a continuación:

VARIABLE	FUENTE Y ENLACE WEB
Precio medio diario de la EE en España	REE (https://www.esios.ree.es/es/analisis/1293?compare_indicators=&start_date=01-01-2012T00%3A00&geoids=&vis=1&end_date=31-12-2022T23%3A55&compare_start_date=31-12-2011&groupby=day)
Demanda media diaria	INVESTING (https://es.investing.com/commodities/brent-oil-historical-data?cid=1184864)
Precio del petróleo	INVESTING (https://es.investing.com/commodities/natural-gas-historical-data?cid=1057002)
Precio del gas natural	INVESTING (https://es.investing.com/commodities/rotterdam-coal-futures-historical-data)
Producciones según tecnologías	REE (https://www.ree.es/es/datos/generacion/estructura-generacion)
Temperatura	AEMET (https://opendata.aemet.es/centrodedescargas/productosAEMET?)
Velocidad del viento	AEMET (https://opendata.aemet.es/centrodedescargas/productosAEMET?)
Reservas hidráulicas	MINISTERIO PARA LA TRANSICIÓN ECOLÓGICA (https://www.miteco.gob.es/es/agua/temás/evaluacion-de-los-recursos-hídricos/boletín-hidrológico/default.aspx)
Precio Derechos de emisión de CO₂	INVESTING (https://es.investing.com/commodities/carbon-emissions-historical-data)
Situación socio-económica del país	INVESTING (https://es.investing.com/indices/spain-35-historical-data?cid=26491)
Intercambio de EE con otros países	REE (https://www.ree.es/es/datos/intercambios)

6.4 ANALISIS DE LOS DATOS

En este capítulo se realizará un análisis temporal de las variables de entrada con la variable de salida. Esto se realiza para conocer la evolución que han tenido los datos y poder interpretar de mejor manera los resultados de los modelos predictivos.

La primera variable que analizaremos será la de variable de salida “**Pre_elec**”, que muestra la evolución del precio de la EE que buscamos predecir en este estudio. Como se puede observar en el gráfico inferior, durante la etapa temporal empleada en el estudio (recordemos que el año 2020 se ha suprimido) se ven dos períodos claramente diferenciados. El primero que va desde el inicio hasta finales del año 2019, y la segunda que va desde principios del 2021 hasta el final del 2022.



En la primera etapa el precio de la EE se mantuvo en un tramo entre los 103 y los 110 €/MWh, alguna caída rápida, recuperándose posteriormente lentamente. En la segunda etapa, correspondiente a la época posterior al Covid-19, nos encontramos con otro escenario completamente diferente donde el precio creció rápidamente alcanzando máximos de 556 €/MWh. A finales del año 2022 se observa como el precio se ha ido recuperando hasta recuperar la mitad de lo que había subido.

Con este análisis podemos concluir que el precio habitual de la EE en España corresponde a los que tenemos en la primera etapa y no en la segunda, en la que se han producido

algunas anomalías, como la pandemia del Covid-19 y la guerra en Ucrania, que han provocado un comportamiento anómalo en este precio.

En cuanto a la **demand**a, en el gráfico observamos que no es constante, produciéndose puntas de consumo generalmente todos los años en los meses de enero, julio y marzo, y picos de menor demanda los meses de mayo y septiembre. Estas variaciones se han movido durante los 10 años del estudio entre los 35.300 y los 19.000 MWh. A simple vista no se puede observar una influencia de la demanda en el precio de la EE.

Las próximas variables que analizaremos serán las del precio de las **materias primas empleadas para producir electricidad**.

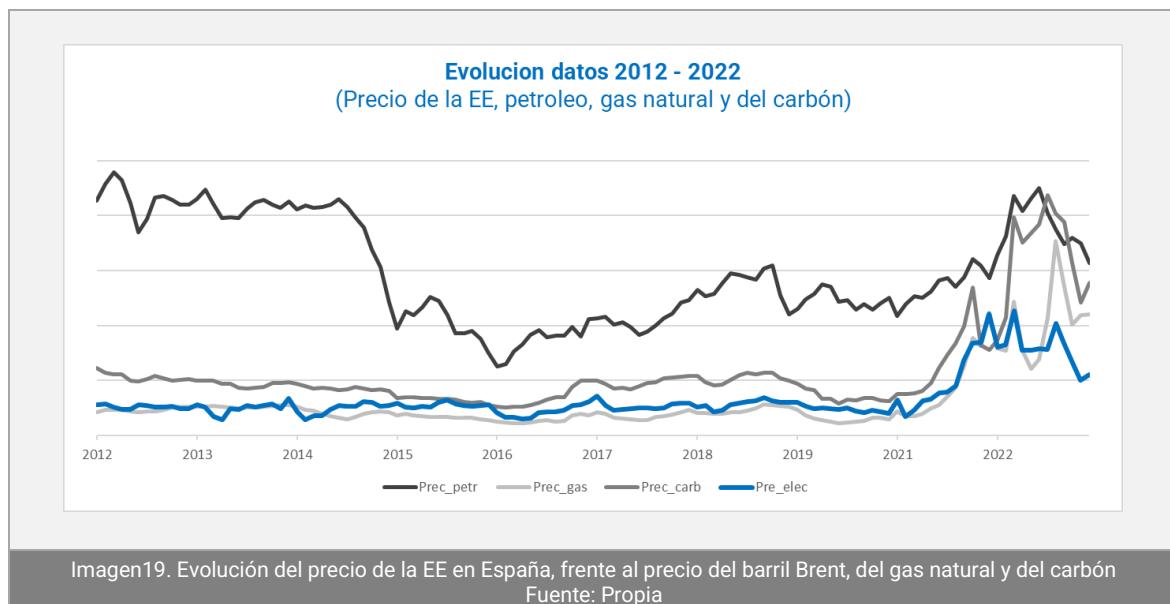


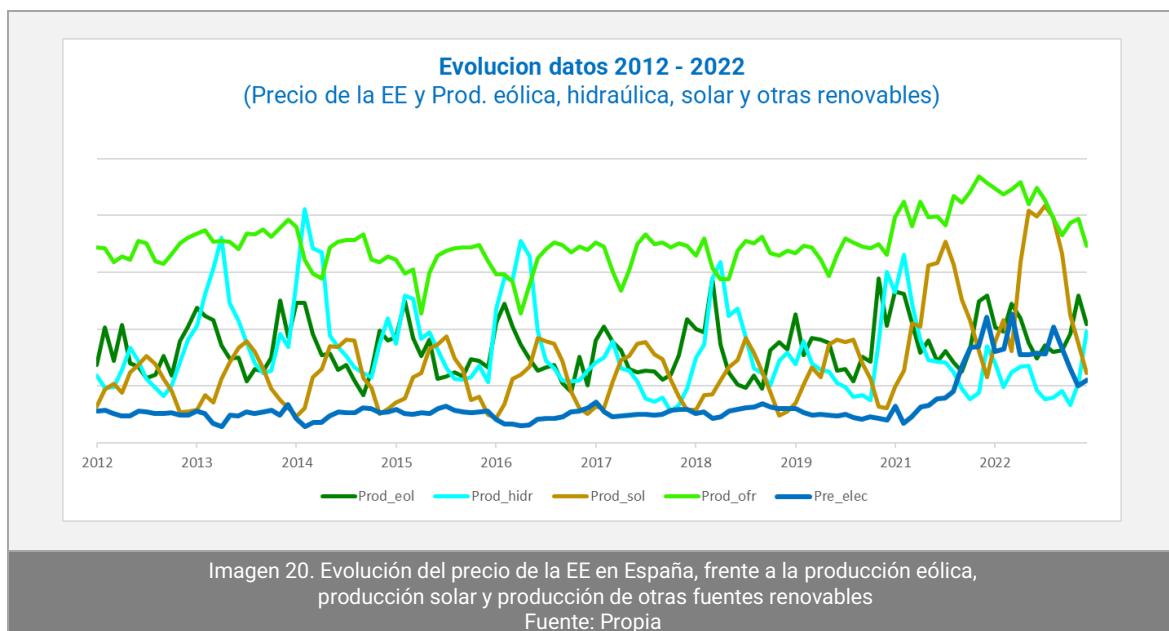
Imagen19. Evolución del precio de la EE en España, frente al precio del barril Brent, del gas natural y del carbón
Fuente: Propia

En general se comprueba que estos precios han sufrido también las anomalías experimentadas por el precio de la EE a partir del año 2021, ya que sufrieron unas subidas muy acusadas en sus valores.

En particular el precio del barril de Brent mantiene muchos altibajos en su cotización, siendo el más acusado el sufrido en el año 2014, con una brusca bajada. Se comprueba que el precio de la EE está desligado de esta materia prima.

En cuanto al precio del carbón y del Gas Natural se observa que, exceptuando los años 2021 y 2022, no sufren variaciones bruscas, y que el precio de la EE sigue de forma muy parecida la evolución del precio de estas materias primas.

Analizando las **técnicas de producción** denominadas **renovables** tendremos estas dos graficas.



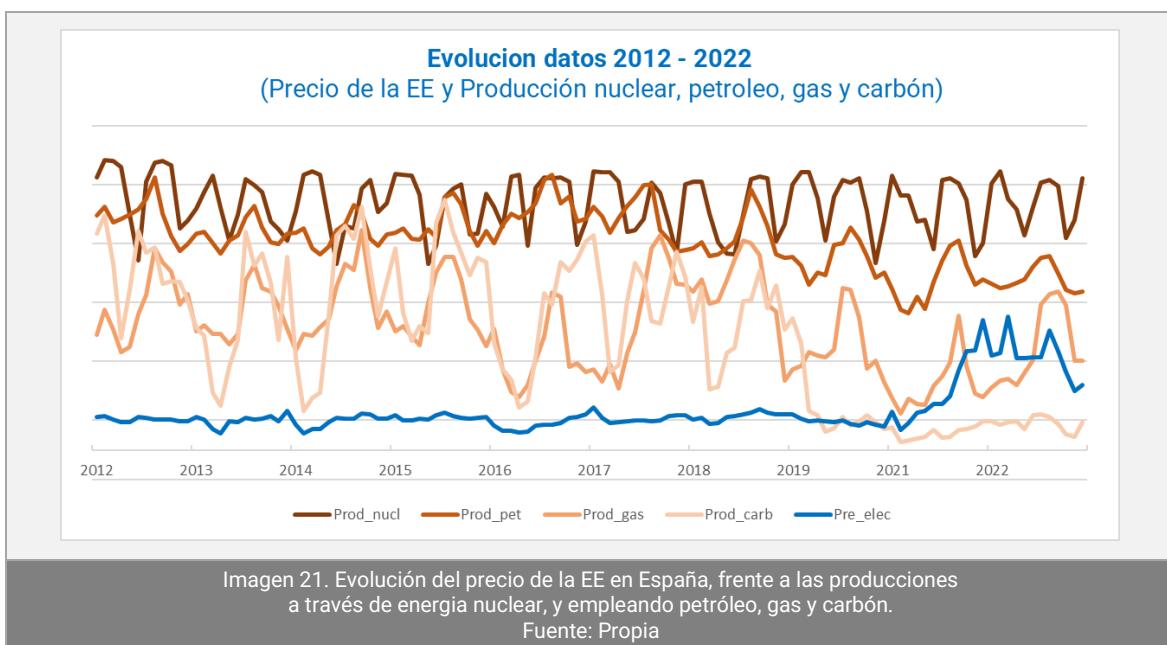
En este grafico se puede constatar que los tipos de producción eólica e hidráulica siguen una misma distribución teniendo los picos máximos en los mismos momentos, al contrario que la producción solar, que los máximos los alcanza cuando se encuentran las otras dos tecnologías en sus producciones mínimas.

En cuanto al resto de tecnologías renovables poseen una cuota de producción más o menos constante sufriendo bruscas bajadas los meses de marzo.

También podemos comprobar que en los dos últimos años la producción solar y otras tecnologías han subido su producción.

No se ve una relación clara entre estas variables y la variable objetivo.

A continuación, analizamos empleando dos graficas las **tecnologías de producción no renovables**.

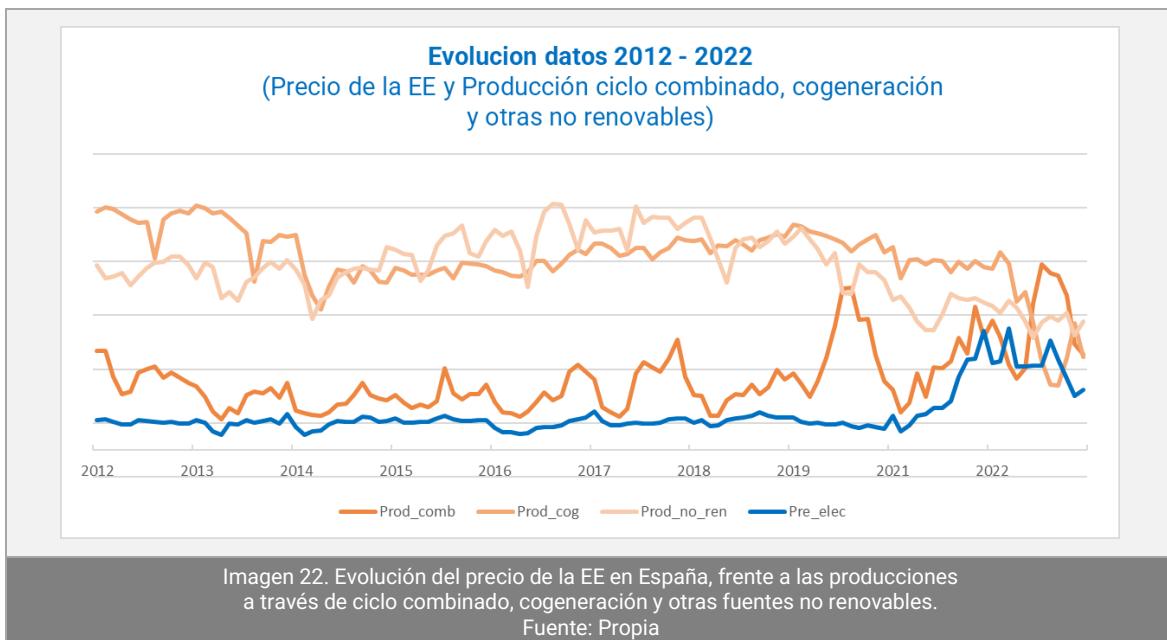


Se observa claramente que el comportamiento de estas variables va en parejas, es decir, la producción a través de centrales nucleares y las que emplean el petróleo como materia prima para producir EE tienen un comportamiento bastante similar, al igual que sucede con la producción a través de Gas Natural y carbón.

En el primer caso, la producción se mantiene bastante elevada respecto a su potencial total, produciéndose subidas y bajadas a lo largo del año.

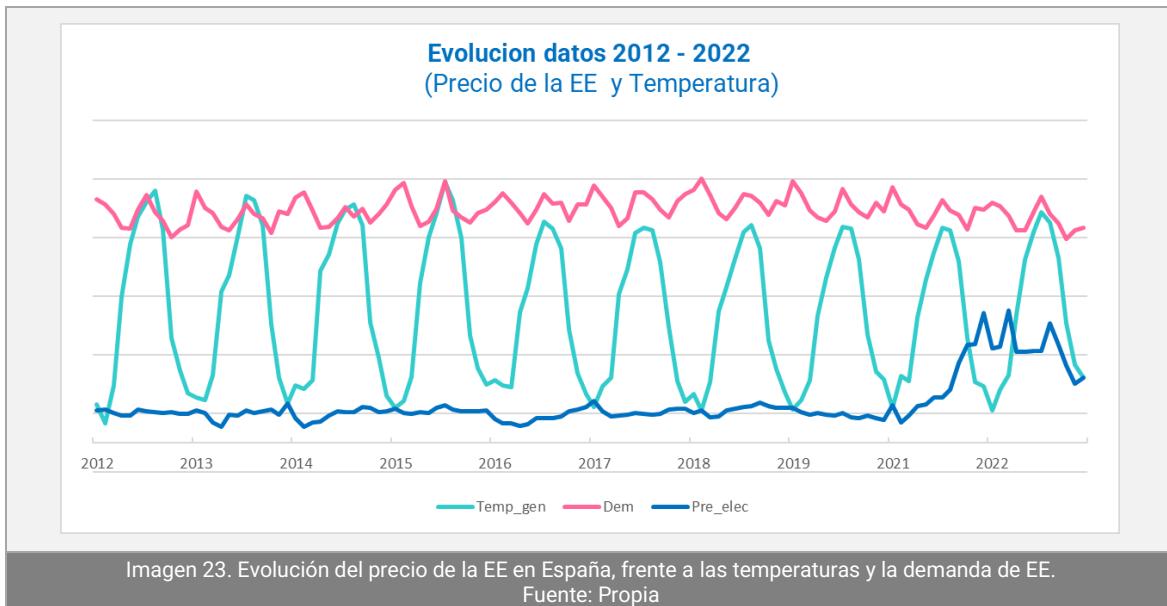
En cuanto a las producciones con Gas Natural y carbón, las variaciones que presentan son mucho más marcadas entre los puntos máximos y mínimos.

A simple vista no se aprecia relación entre todos estos tipos de producción y el precio de la EE.



En cuanto a la producción mediante Ciclo Combinado se comprueba que tiene mucha relación con el precio de la EE, cosa que no sucede con los otros dos tipos de producción.

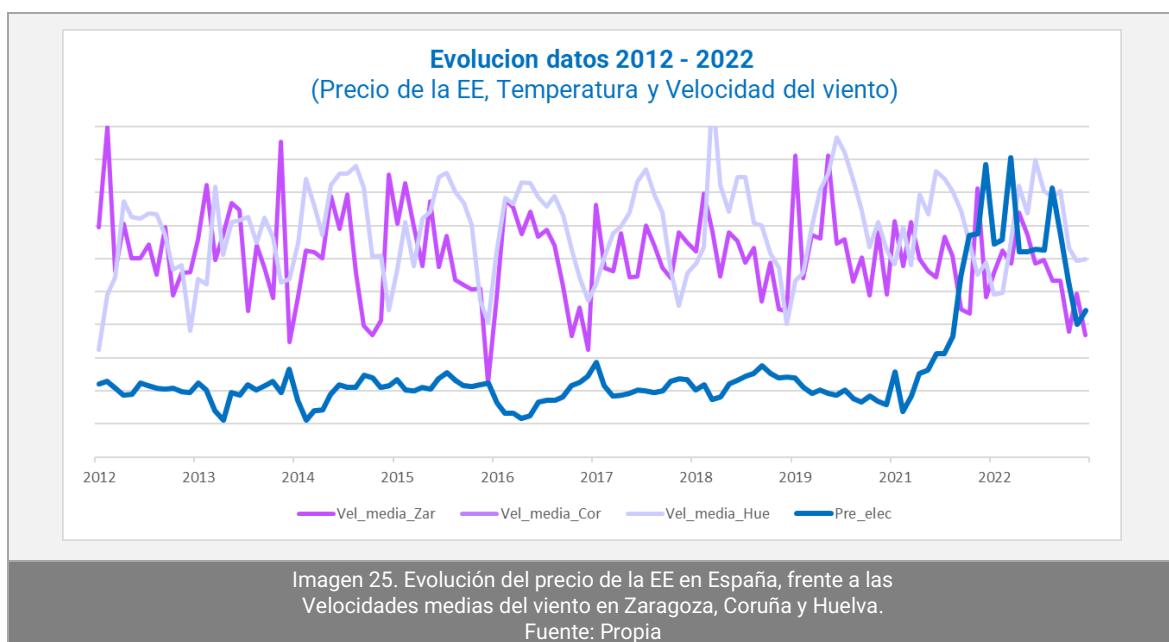
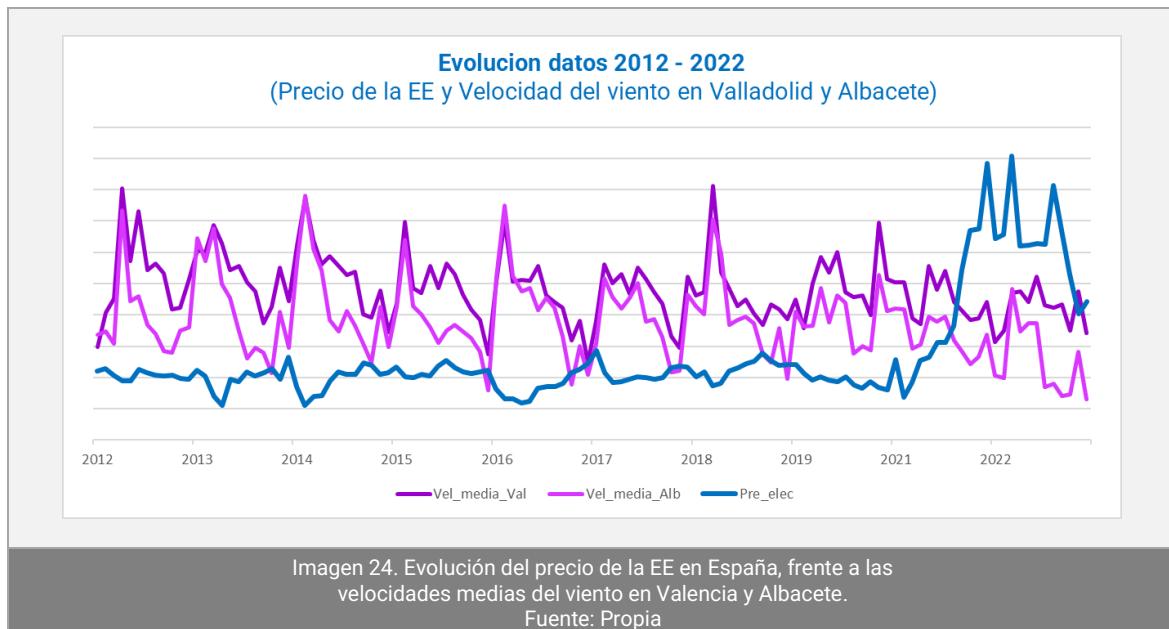
A continuación, entramos a analizar las variables de factores climáticos como son la **velocidad del viento** y la **temperatura**.



MODELOS PARA LA PREDICCION DEL PRECIO DE LA ELECTRICIDAD

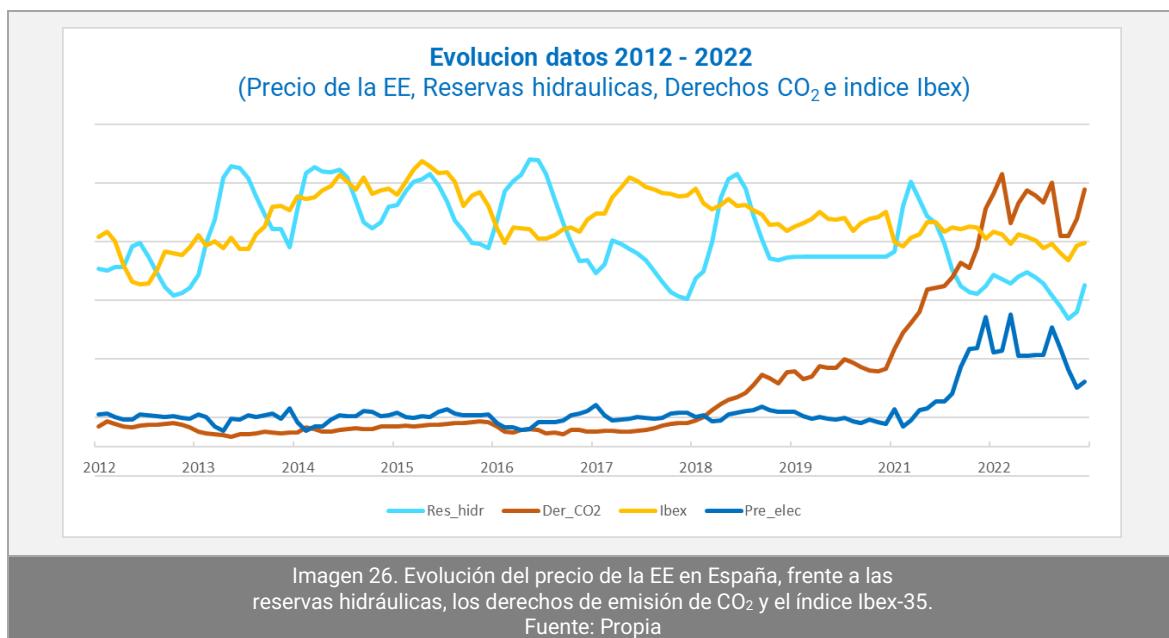
EN EL MERCADO ESPAÑOL

En cuanto a la temperatura se observa claramente que no tiene ninguna relación con el precio de la EE. Se comprueba que la demanda de EE posee sus picos máximos cuando las temperaturas toman sus valores más altos, pero también más bajos.



Aunque las diferentes variables de la velocidad del viento no siguen la misma distribución que el precio de la EE, si se puede observar que cuando se producen picos altos de velocidad el precio de la EE tiende a bajar.

Para terminar en la siguiente grafica constatamos que ni las **reservas hidráulicas** ni el **índice Ibex-35** mantienen una relación clara con el precio de la EE.



En cambio, el **precio de los derechos de emisión de CO₂** si se comprueba que guarda una clara relación con el precio de la EE.

6.5 EXPLORACION DE LAS VARIABLES

Una vez establecidas las variables que se van a considerar en los modelos estudiados para predecir el precio de la EE, en este capítulo pasamos a presentar los datos obtenidos de las diferentes fuentes y hacer un primer análisis de estos.

Para realizar este trabajo se ha recurrido al lenguaje de programación **R**, a través del IDE (Entorno de Desarrollo Integrado) **RStudio**. Se ha optado por este lenguaje ya que es uno de los lenguajes más destacados en estadística, posee todo lo necesario para analizar los datos con eficacia y es un lenguaje permite crear un código limpio que facilita la gestión de los datos.

Las explicaciones de las variables, presentadas a continuación, se basan en el código desarrollado con anterioridad en RStudio, cuyo código puede ser consultado en el **ANEXO II. ANÁLISIS DE LAS VARIABLES CONSIDERADAS.**

Los datos originales obtenidos de diferentes fuentes y empleados para el análisis de las variables, se han guardado en archivos con extensión csv. En el **ANEXO II. ARCHIVOS CSV EMPLEADOS** se relacionan todos estos archivos.

6.5.1 VARIABLE: Precio medio diario energía eléctrica (Pre_elec)

El dataframe de los datos donde se encuentra esta variable tiene 6 columnas de las que solo nos interesan dos, “**value**”, que es la de la variable dependiente y “**datetime**”. Existen **3652 registros**, correspondientes al número de días de los años del periodo del estudio. Ninguna de las dos columnas posee valores perdidos.

Posteriormente habrá que transformar los datos de la columna “**datetime**” para que sean del tipo “fecha”.

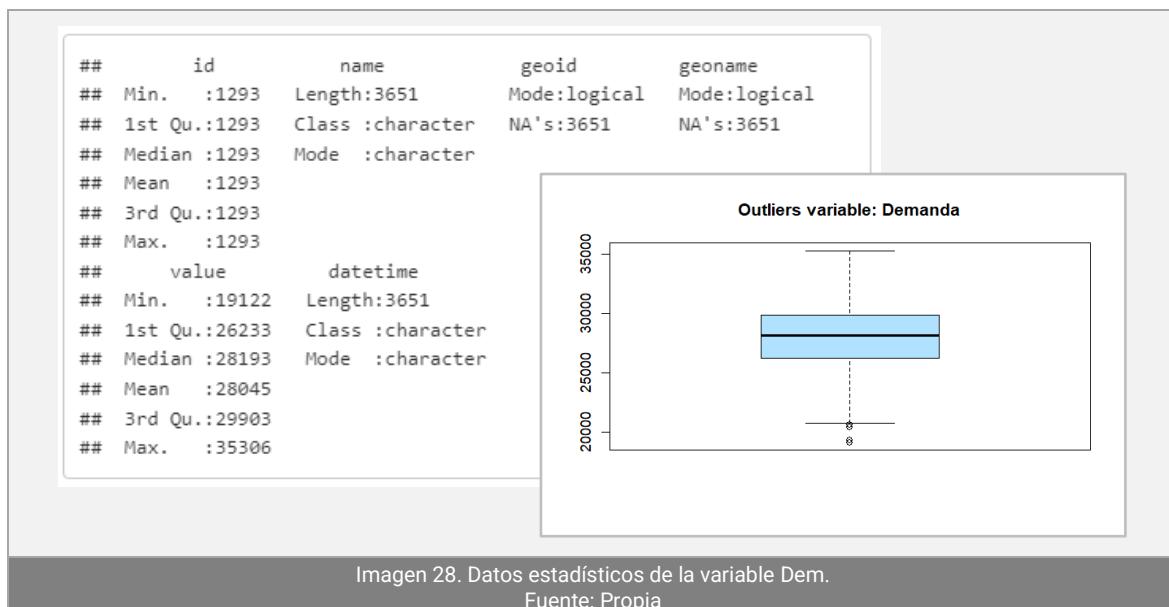
```
##      id          name        geoid      geoname
##  Min.   :10211  Length:3652      Mode:logical  Mode:logical
##  1st Qu.:10211  Class :character  NA's:3652    NA's:3652
##  Median :10211  Mode  :character
##  Mean   :10211
##  3rd Qu.:10211
##  Max.   :10211
##      value      datetime
##  Min.   : 10.04  Length:3652
##  1st Qu.: 51.10  Class :character
##  Median : 59.18  Mode  :character
##  Mean   : 73.89
##  3rd Qu.: 69.06
##  Max.   :556.15
```

Imagen 27. Datos estadísticos de la variable Pre_elec.
Fuente: Propia

6.5.2 VARIABLE: Demanda media diaria (Dem)

El dataframe que incluye los datos de esta variable tiene la misma estructura que la anterior. De la misma forma las columnas que nos interesan son “**value**” y “**datetime**”.

Ninguna de las dos columnas presenta **valores perdidos (NA's)** y en cuanto a los **outliers**, esta variable solo presenta 3 valores anormales, por lo que no tienen ningún peso en el estudio.



6.5.3 VARIABLE: Precio del petróleo (Prec_petr)

En este dataframe existen **2889 registros**, correspondientes a 763 días menos. Esto se ha producido porque el precio del barril de petróleo Brent se negocia en los mercados de lunes a sábado. Para solucionar esto, se tomará como valor para los sábados y domingos el precio del viernes anterior. En este dataframe nos interesan las dos primeras columnas “**Fecha**” y “**Último**”, ya que esta ultima indica el valor final que tomó el barril de petróleo ese día.

```

##      fecha         ultimo       apertura       maximo
## Length:2889    Min. : 19.33   Min. : 19.90   Min. : 21.29
## Class :character 1st Qu.: 53.92   1st Qu.: 53.97   1st Qu.: 54.69
## Mode  :character Median : 68.87   Median : 68.81   Median : 69.63
##               Mean : 75.20   Mean : 75.20   Mean : 76.19
##               3rd Qu.:103.41  3rd Qu.:103.32  3rd Qu.:104.60
##                   Max. :130.24  Max. :130.28  Max. :139.13
##
##      minimo       volumen     variacion
## Min. : 15.98   Min. : 0.02 Length:2889
## 1st Qu.: 53.02  1st Qu.:173.08 Class :character
## Median : 67.73  Median :228.19 Mode :character
## Mean : 74.16   Mean :225.16
## 3rd Qu.:102.07  3rd Qu.:285.35
## Max. :125.00   Max. :779.72
## NA's :1

```

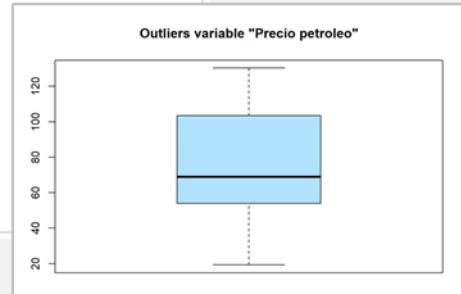


Imagen 29. Datos estadísticos de la variable Prec_petr
Fuente: Propia

Ninguna de las dos columnas que nos interesan para el estudio presenta **valores NA's** y en cuanto a los **outliers**, esta variable no presenta valores anormales.

6.5.4 VARIABLE: Precio del gas natural (prec_gas)

Este caso tiene la misma estructura que el anterior por lo que las columnas que nos interesan serán “**Fecha**” y “**Ultimo**”. Para esta variable tenemos **2541 registros** correspondientes a los días de lunes a viernes. Para solucionar esto, se tomará como valor para los sábados y domingos el valor del viernes anterior.

```

##      fecha         ultimo       apertura       maximo
## Length:2541    Min. : 24.18   Min. : 24.49   Min. : 24.95
## Class :character 1st Qu.: 41.57   1st Qu.: 41.60   1st Qu.: 42.10
## Mode  :character Median : 53.86   Median : 53.95   Median : 54.50
##               Mean : 77.93   Mean : 78.25   Mean : 81.17
##               3rd Qu.: 66.59   3rd Qu.: 66.66   3rd Qu.: 67.21
##                   Max. :640.36   Max. :659.50   Max. :800.00
##
##      minimo       volumen     variacion
## Min. : 23.59   Min. : 1.11   Min. :-0.299500
## 1st Qu.: 41.21  1st Qu.: 6.55   1st Qu.:-0.014200
## Median : 53.40  Median : 9.92   Median :-0.000300
## Mean : 75.52   Mean :10.47   Mean : 0.001295
## 3rd Qu.: 66.11  3rd Qu.:13.41  3rd Qu.: 0.015100
## Max. :559.29   Max. :38.61   Max. : 0.509300
## NA's :15

```

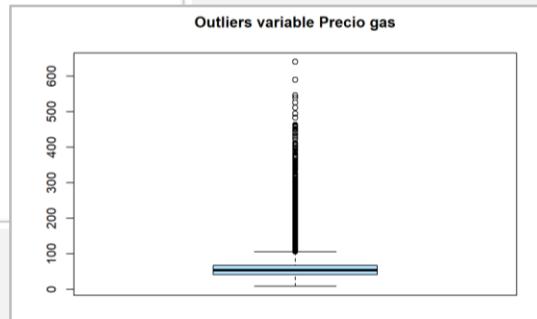
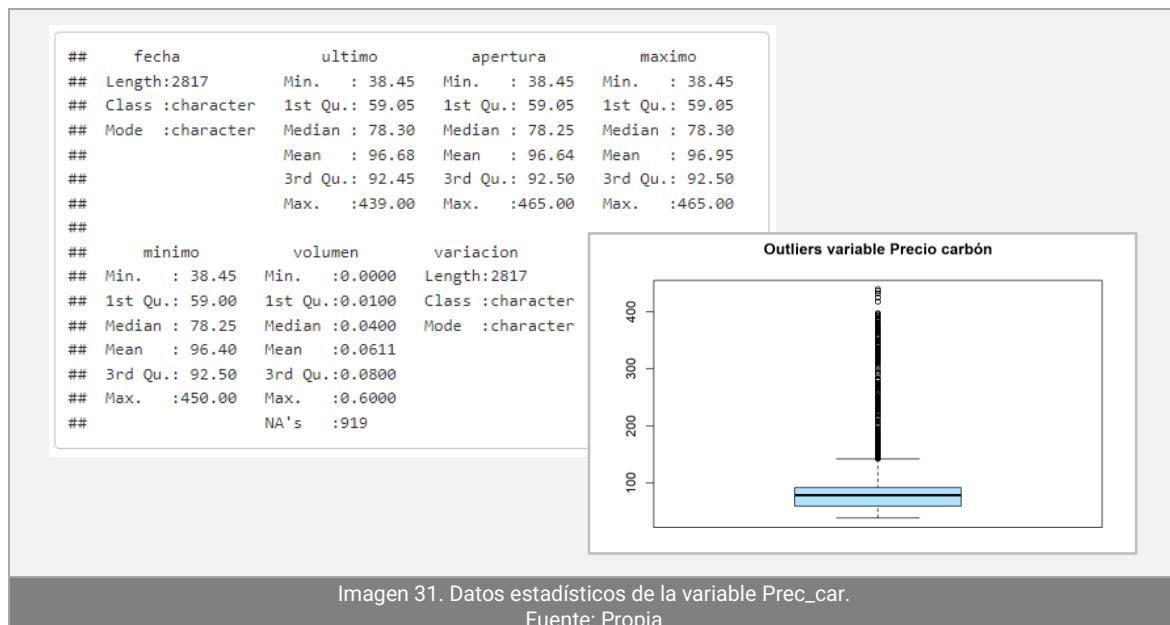


Imagen 30. Datos estadísticos de la variable Prec_gas.
Fuente: Propia

Ninguna de las dos columnas posee **valores perdidos**, y la de los precios del gas presenta **357 valores anormales**, que suponen un **14,05%**, por lo que se tendrá que tomar alguna decisión con estos valores.

6.5.5 VARIABLE: Precio medio diario del carbón (Prec_car)

Es el mismo caso que el anterior, lo único que cambia es que existe mayor número de registros debido a que su cotización se produce de lunes a sábado, y esto hace que más adelante tomemos el valor del día anterior para los domingos.



No presenta **valores perdidos** en las columnas que nos interesan y en cuanto a los **outliers** presenta **328 valores anormales**, que suponen un **11,64%**, por lo que habrá que tomar una decisión con estos valores.

6.5.6 VARIABLES: Producción por tecnología

En este caso el dataframe obtenido de Red Eléctrica Española posee todas las variables de producción según la tecnología y todas las columnas nos interesan en este caso. El **número de registros** es de **3652** correspondientes al total de días de todo el periodo estudiado.

MODELOS PARA LA PREDICCIÓN DEL PRECIO DE LA ELECTRICIDAD

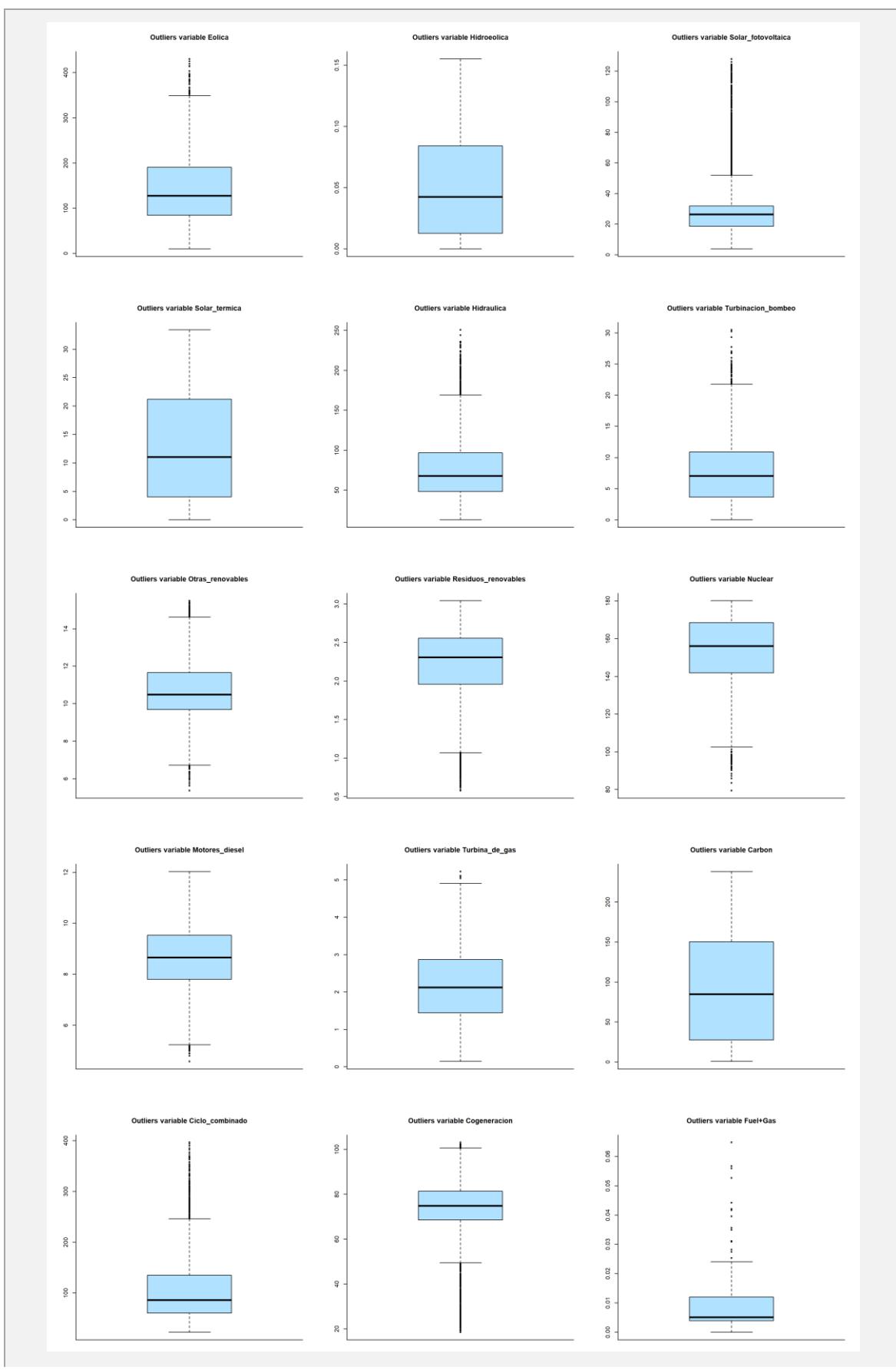
EN EL MERCADO ESPAÑOL

```
##      fecha          Eolica       Hidroelica   Solar_fotovoltaica
##  Length:3652      Min. : 9.688     Min. :0.0000  Min.  : 3.739
##  Class :character 1st Qu.: 84.011    1st Qu.:0.0128  1st Qu.: 18.479
##  Mode  :character Median :127.595   Median :0.0425  Median : 26.327
##                  Mean  :143.356   Mean  :0.0512  Mean   : 31.560
##                  3rd Qu.:190.193   3rd Qu.:0.0842  3rd Qu.: 31.889
##                  Max.  :430.064   Max.  :0.1550  Max.   :127.947
##                  NA's   :911
##      Solar_termica   Hidraulica  Turbinacion_bombeo Otras_renovables
##  Min.  : 0.000  Min.  : 12.74  Min.  : 0.0121  Min.  : 5.363
##  1st Qu.: 4.039 1st Qu.: 48.07  1st Qu.: 3.6524  1st Qu.: 9.684
##  Median :11.027  Median : 67.69  Median : 7.0555  Median :10.484
##  Mean   :12.821  Mean   : 78.46  Mean   : 7.7359  Mean   :10.669
##  3rd Qu.:21.210  3rd Qu.: 96.68  3rd Qu.:10.9008  3rd Qu.:11.666
##  Max.   :33.388  Max.   :250.70  Max.   :30.4844  Max.   :15.495
##  NA's   :4
##      Residuos_renovables Nuclear      Motores_diesel  Turbina_de_gas
##  Min.  :0.5756  Min.  : 79.42  Min.  : 4.579  Min.  :0.1438
##  1st Qu.:1.9600 1st Qu.:141.95  1st Qu.: 7.805  1st Qu.:1.4439
##  Median :2.3065  Median :156.10  Median : 8.655  Median :2.1196
##  Mean   :2.1779  Mean   :151.38  Mean   : 8.600  Mean   :2.1835
##  3rd Qu.:2.5544  3rd Qu.:168.51  3rd Qu.: 9.527  3rd Qu.:2.8689
##  Max.   :3.0419  Max.   :180.17  Max.   :12.026  Max.   :5.2203
##
##      Carbon        Ciclo_combinado Cogeneracion   Fuel.Gas
##  Min.  : 0.4657  Min.  : 22.73  Min.  : 18.50  Min.  :0.0000
##  1st Qu.:27.4792 1st Qu.: 60.56  1st Qu.: 68.57  1st Qu.:0.0040
##  Median :84.5868  Median : 86.07  Median : 74.90  Median :0.0051
##  Mean   :91.4891  Mean   :105.83  Mean   : 73.72  Mean   :0.0076
##  3rd Qu.:150.2429 3rd Qu.:134.89  3rd Qu.: 81.38  3rd Qu.:0.0120
##  Max.   :238.1793 Max.   :396.45  Max.   :103.09  Max.   :0.0647
##  NA's   :2590
##      Turbina_de_vapor Residuos_no_renovables Generacion_total
##  Min.  :0.5636  Min.  :2.541      Min.  :530.8
##  1st Qu.:4.7191 1st Qu.:4.993      1st Qu.:679.6
##  Median :6.2775  Median :6.007      Median :730.4
##  Mean   :5.9172  Mean   :5.932      Mean   :731.9
##  3rd Qu.:7.2975 3rd Qu.:6.948      3rd Qu.:781.7
##  Max.   :9.5210  Max.   :8.821      Max.   :997.2
##
```

Imagen 32. Datos estadísticos de las variables Producción por tecnología.
Fuente: Propia

MODELOS PARA LA PREDICCIÓN DEL PRECIO DE LA ELECTRICIDAD

EN EL MERCADO ESPAÑOL



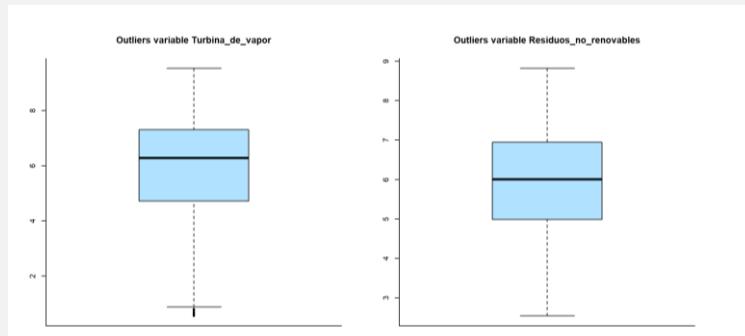


Imagen 33. Boxplots de las variables Producción por tecnología.

Fuente: Propia

Solo 3 variables contienen **valores faltantes**. En el caso de la “**Solar_termica**” son despreciables ya que son solo **4** casos. Para la variable “**Hidroeolica**” existen **911** casos debido a que hasta junio del 2014 no entró en funcionamiento la primera instalación de estas características en España. Y finalmente la variable “**Fuel+Gas**” posee **2590**, es decir, la gran mayoría de los días registrados. Esto se debe a que, a partir de octubre del 2014, se ha convertido en una tecnología muy residual.

En cuanto a los **valores anormales** vemos que hay una variable que presenta gran cantidad de outliers que es “**Solar_fotovoltaica**” (13,34%). Posteriormente hay una serie de ellas que poseen un porcentaje entre el 5 y el 1% como son “**Cogeneración**”, “**Residuos_renovables**”, “**Hidraulica**”, “**Ciclo_combinado**”, “**Otras_renovables**”, “**Nuclear**” y “**Turbinacion_bombeo**”. El resto, o bien, no poseen valores anormales, o son inferiores al 1%. Más adelante habrá que tomar una decisión con estos valores en las variables que existen en gran cantidad.

6.5.7 VARIABLE: Temperatura (Temp_)

Es este caso **tendremos un dataframe para cada una de las cinco ciudades que se han tomado para ver la influencia de este factor** (Madrid, Barcelona, Valencia, Sevilla y Zaragoza).

Para todos los dataframes tienen la misma estructura con 15 columnas de las cuales solo nos interesan tres, “fecha”, “tmin” y “tmax”. Estas dos ultimas correspondientes a las **temperaturas mínima y máxima diarias** respectivamente.

El **número de registros** de los dataframes es de **3651**, menos en el caso de la ciudad de **Barcelona**, en el que existen **3441**. Esto se debe a que los datos proporcionados por la fuente (AEMET) no proporciona los datos desde julio a diciembre del 2020. Más adelante habrá que tomar una decisión para solucionar este problema.

- Madrid

```
##   fecha      indicativo     nombre     provincia
## Length:3651      Min.    :3195  Length:3651      Length:3651
## Class :character 1st Qu.:3195  Class :character  Class :character
## Mode  :character Median :3195   Mode :character  Mode :character
##                  Mean   :3195
##                  3rd Qu.:3195
##                  Max.   :3195
##
##   altitud      tmed      prec      tmin
## Min.    :667  Min.   :-3.40  Length:3651  Min.   :-7.40
## 1st Qu.:667  1st Qu.: 9.40  Class :character 1st Qu.: 5.30
## Median :667  Median :15.00  Mode  :character  Median :10.30
## Mean   :667  Mean   :16.04  NA's    :14        Mean   :10.96
## 3rd Qu.:667  3rd Qu.:22.80  NA's    :14        3rd Qu.:16.60
## Max.   :667  Max.   :33.40  NA's    :14        Max.   :26.20
## NA's   :14
##
##   horatmin      tmax      horatmax      dir
## Length:3651  Min.   : 0.30  Length:3651  Length:3651
## Class :character 1st Qu.:13.10  Class :character  Class :character
## Mode  :character Median :19.80  Mode  :character  Mode :character
## Mean   :21.11
## 3rd Qu.:29.00
## Max.   :40.70
## NA's   :14
##
##   velmedia      racha      horaracha
## Length:3651  Min.   : 1.9  Length:3651
## Class :character 1st Qu.: 7.2  Class :character
## Mode  :character Median : 9.2  Mode :character
## Mean   :105.0
## 3rd Qu.:12.2
## Max.   :951.1
## NA's   :143
```

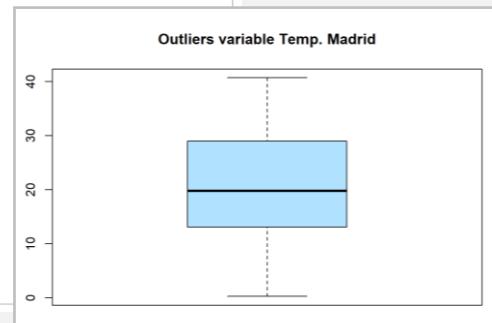


Imagen 34. Datos estadísticos de la variable Tem_Mad.
Fuente: Propia

MODELOS PARA LA PREDICCIÓN DEL PRECIO DE LA ELECTRICIDAD
EN EL MERCADO ESPAÑOL

- Barcelona

```
##      fecha      indicativo      nombre      provincia
##  Length:3441      Length:3441      Length:3441      Length:3441
##  Class :character  Class :character  Class :character  Class :character
##  Mode  :character  Mode  :character  Mode  :character  Mode  :character
##
##
##
##      altitud      tmed      prec      tmin
##  Min.   :6   Min.   : 2.70   Min.   : 0.000   Min.   :-0.20
##  1st Qu.:6   1st Qu.:12.90   1st Qu.: 0.000   1st Qu.: 9.80
##  Median :6   Median :17.10   Median : 0.000   Median :14.20
##  Mean   :6   Mean   :17.64   Mean   : 1.164   Mean   :14.69
##  3rd Qu.:6   3rd Qu.:22.70   3rd Qu.: 0.000   3rd Qu.:20.00
##  Max.   :6   Max.   :30.00   Max.   :83.900   Max.   :27.10
##  NA's    :215  NA's    :22     NA's    :215
##      horatmin      tmax      horatmax      dir
##  Length:3441      Min.   : 4.60   Length:3441      Min.   : 1.0
##  Class :character  1st Qu.:15.90   Class :character  1st Qu.:11.0
##  Mode  :character  Median :20.10   Mode  :character  Median :20.0
##  Mean   :20.58
##  3rd Qu.:25.40
##  Max.   :35.20
##  NA's    :214
##      velmedia      racha      horaracha
##  Min.   : 0.0   Min.   : 3.900  Length:3441
##  1st Qu.: 2.5   1st Qu.: 7.200  Class :character
##  Median : 3.3   Median : 8.900  Mode  :character
##  Mean   : 3.5   Mean   : 9.652
##  3rd Qu.: 4.2   3rd Qu.:11.400
##  Max.   :13.9   Max.   :26.100
##  NA's    :14    NA's    :25
```

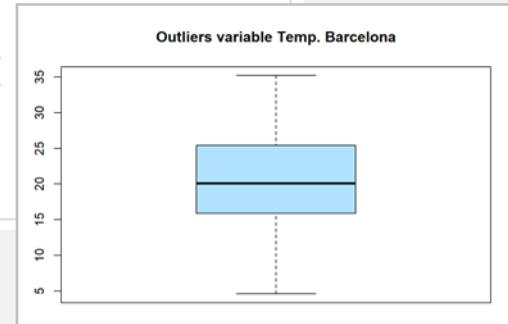


Imagen 35. Datos estadísticos de la variable Tem_Bar.
Fuente: Propia

- Valencia

```

##    fecha         nombre      provincia      altitud
##  Length:3652    Length:3652    Length:3652    Min.   :11
##  Class :character Class :character Class :character  1st Qu.:11
##  Mode  :character Mode  :character Mode  :character Median  :11
##                                         Mean   :11
##                                         3rd Qu.:11
##                                         Max.   :11
##
##    tmed        prec       tmin      horatmin
##  Min.   : 5.30  Length:3652  Min.   : 0.00  Length:3652
##  1st Qu.:14.20  Class :character  1st Qu.: 9.70  Class :character
##  Median :18.85  Mode  :character  Median :14.50  Mode  :character
##  Mean   :19.04                           Mean   :14.72
##  3rd Qu.:23.90                           3rd Qu.:19.80
##  Max.   :32.80                           Max.   :27.50
##
##    tmax      horatmax      dir      velmedia
##  Min.   : 0.00  Length:3652  Min.   : 1.00  Min.   :0.600
##  1st Qu.: 7.70  Class :character  1st Qu.:11.00  1st Qu.:1.100
##  Median :11.10  Mode  :character  Median :22.00  Median :1.550
##  Mean   :13.86                           Mean   :20.78  Mean   :1.702
##  3rd Qu.:21.20                           3rd Qu.:30.00  3rd Qu.:1.975
##  Max.   :42.00                           Max.   :35.00  Max.   :3.900
##  NA's    :8                             NA's   :3592  NA's   :3592
##
##    racha      horaracha
##  Min.   : 3.600  Length:3652
##  1st Qu.: 5.225  Class :character
##  Median : 6.800  Mode  :character
##  Mean   : 7.920
##  3rd Qu.: 8.675
##  Max.   :19.700
##  NA's   :3592

```

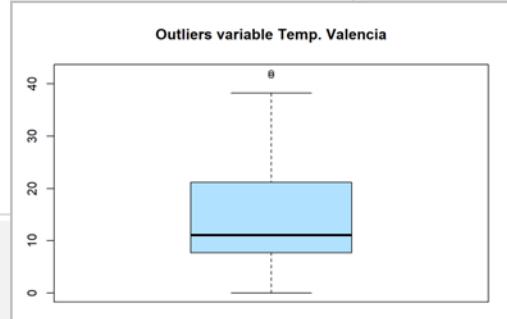


Imagen 36. Datos estadísticos de la variable Tem_Val.

Fuente: Propia

- Sevilla

```

##    fecha      indicativo     nombre     provincia
##  Length:3652      Min. :5783  Length:3652      Length:3652
##  Class :character 1st Qu.:5783  Class :character  Class :character
##  Mode  :character Median :5783   Mode :character  Mode :character
##                Mean  :5783
##                3rd Qu.:5783
##                Max. :5783
##
##    altitud      tmed      prec      tmin
##  Min.  :34  Min.  : 4.80  Length:3652  Min.  :-2.00
##  1st Qu.:34  1st Qu.:13.80  Class :character  1st Qu.: 8.20
##  Median :34  Median :19.10  Mode :character  Median :13.50
##  Mean   :34  Mean   :19.58          NA's       Mean   :13.16
##  3rd Qu.:34  3rd Qu.:25.40          NA's       3rd Qu.:18.10
##  Max.   :34  Max.   :36.30          NA's       Max.   :28.10
##                NA's   :15          NA's       NA's   :15
##
##    horatmin      tmax      horatmax      dir
##  Length:3652      Min.  : 7.6  Length:3652  Min.  : 1.00
##  Class :character 1st Qu.:18.8  Class :character  1st Qu.:21.00
##  Mode  :character Median :25.2  Mode :character  Median :24.00
##                Mean  :26.0          NA's       Mean  :45.38
##                3rd Qu.:32.9          NA's       3rd Qu.:99.00
##                Max.  :45.9          NA's       Max.  :99.00
##                NA's   :13          NA's       NA's   :32
##
##    velmedia      racha      horaracha
##  Min.  : 0.000  Min.  : 3.600  Length:3652
##  1st Qu.: 1.900 1st Qu.: 7.800  Class :character
##  Median : 3.100  Median : 9.700  Mode :character
##  Mean   : 3.196  Mean   : 9.798
##  3rd Qu.: 4.200 3rd Qu.:11.400
##  Max.   :12.800  Max.   :28.900
##  NA's   :10      NA's   :32

```

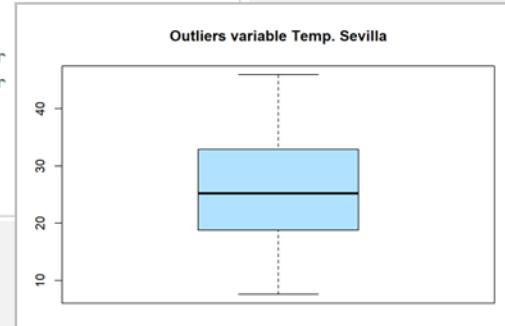


Imagen 37. Datos estadísticos de la variable Tem_Sev.

Fuente: Propia

- Zaragoza

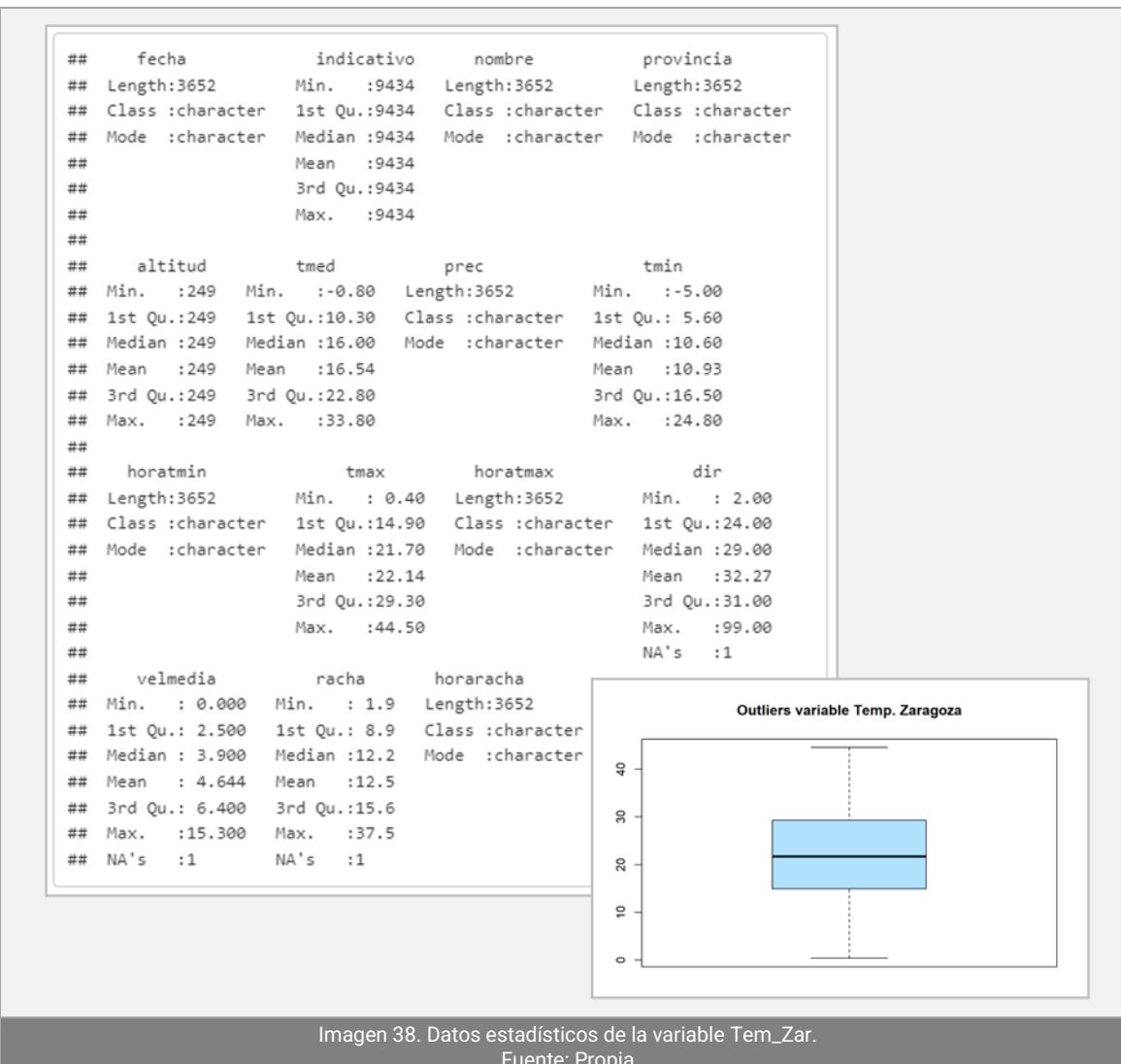


Imagen 38. Datos estadísticos de la variable Tem_Zar.

Fuente: Propia

En cuanto a los **datos faltantes** Madrid presenta 14 en las dos variables, Valencia 8 en la temperatura máxima y Sevilla 15 y 13 en cada una. Esto no supone ningún inconveniente ya que al lado del numero de valores existentes son despreciables. De todas formas, para el estudio se dará el valor del día anterior. Otro caso es el de **Barcelona**, donde el número de datos faltantes es elevado (215 y 214). La causa ya se ha visto al principio del apartado. Más adelante habrá que tomar una decisión de que hacer con estos datos para el estudio. En el caso de **Zaragoza** no existen valores faltantes.

En cuanto a los **outliers** no tienen ninguna influencia en ninguna de las ciudades que solo en el caso de Valencia existen dos.

6.5.8 VARIABLES: Velocidad del viento

En este caso ocurre lo mismo que en el anterior, **existe un dataframe para cada una de las cinco ciudades que se han tomado para ver la influencia de este factor** (Valladolid, Albacete, Zaragoza, La Coruña y Huelva).

Estos dataframe tienen la misma estructura que los de las variables anteriores, aunque en esta ocasión las columnas que nos interesan serán “**fecha**” y “**velmedia**”, correspondiente esta ultima a la **velocidad media diaria del viento**.

El **número de registros** para cada ciudad es de **3652**.

-Valladolid

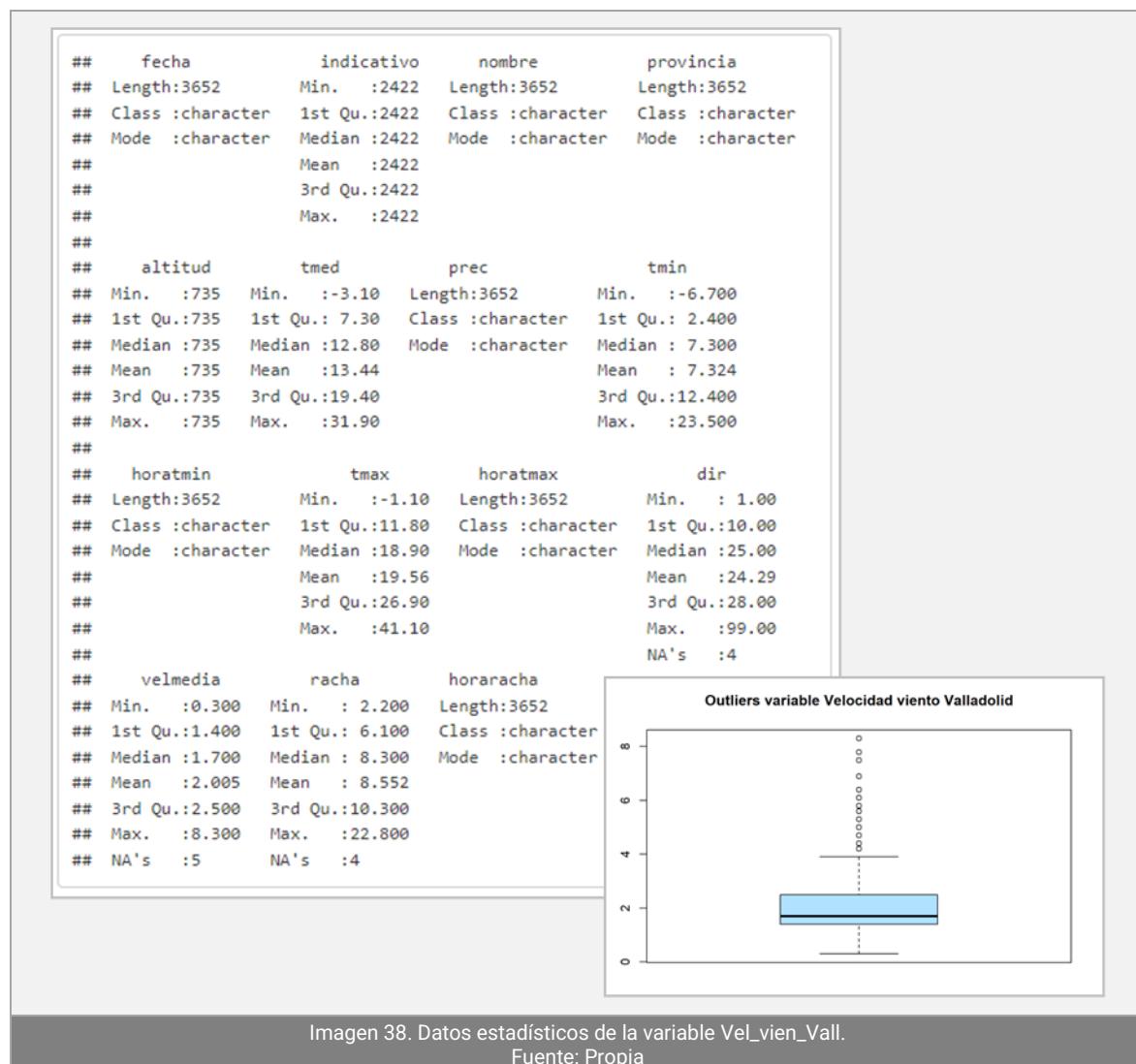


Imagen 38. Datos estadísticos de la variable Vel_vien_Vall.

Fuente: Propia

MODELOS PARA LA PREDICCIÓN DEL PRECIO DE LA ELECTRICIDAD
EN EL MERCADO ESPAÑOL

- Albacete

```
##      fecha      indicativo      nombre      provincia
##  Length:3652    Length:3652    Length:3652    Length:3652
##  Class :character  Class :character  Class :character  Class :character
##  Mode  :character  Mode  :character  Mode  :character  Mode  :character
##
##
##
##      altitud      tmed      prec      tmin
##  Min.   :674   Min.  :-4.00  Length:3652   Min.   :-11.300
##  1st Qu.:674   1st Qu.: 9.40  Class :character  1st Qu.:  3.800
##  Median :674   Median :15.20  Mode  :character  Median :  9.300
##  Mean   :674   Mean   :15.99                    Mean   :  9.555
##  3rd Qu.:674   3rd Qu.:22.80                    3rd Qu.: 15.400
##  Max.   :674   Max.   :34.10                    Max.   : 26.100
##
##      horatmin      tmax      horatmax      dir
##  Length:3652   Min.   : 0.40  Length:3652   Min.   : 0.00
##  Class :character  1st Qu.:14.70  Class :character  1st Qu.:14.00
##  Mode  :character  Median :21.90  Mode  :character  Median :26.00
##                                Mean   :22.42                    Mean   :25.87
##                                3rd Qu.:30.20                    3rd Qu.:28.00
##                                Max.   :42.70                    Max.   :99.00
##                                NA's   :19
##
##      velmedia      racha      horaracha
##  Min.   :0.000   Min.   : 0.000  Length:3652
##  1st Qu.:0.600   1st Qu.: 6.700  Class :character
##  Median :1.100   Median : 8.300  Mode  :character
##  Mean   :1.359   Mean   : 8.592
##  3rd Qu.:1.900   3rd Qu.:10.000
##  Max.   :7.200   Max.   :26.400
##  NA's   :16     NA's   :19
```

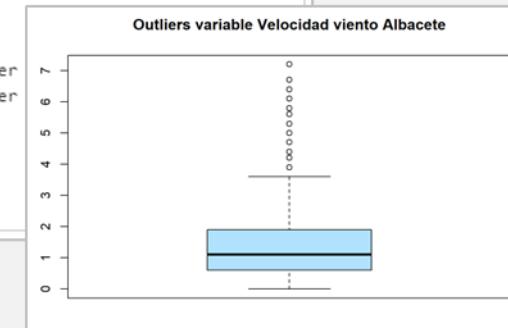


Imagen 40. Datos estadísticos de la variable Vel_vien_Alb.
Fuente: Propia

MODELOS PARA LA PREDICCIÓN DEL PRECIO DE LA ELECTRICIDAD
EN EL MERCADO ESPAÑOL

- Zaragoza

```
##      fecha      indicativo      nombre      provincia
##  Length:3652      Min. :9434      Length:3652      Length:3652
##  Class :character 1st Qu.:9434      Class :character  Class :character
##  Mode  :character Median :9434      Mode  :character Mode  :character
##                                         Mean  :9434
##                                         3rd Qu.:9434
##                                         Max. :9434
##
##      altitud      tmed      prec      tmin
##  Min.  :249  Min.  :-0.80  Length:3652  Min.  :-5.00
##  1st Qu.:249 1st Qu.:10.30  Class :character 1st Qu.: 5.60
##  Median :249  Median :16.00  Mode  :character Median :10.60
##  Mean   :249  Mean   :16.54                    Mean   :10.93
##  3rd Qu.:249 3rd Qu.:22.80                    3rd Qu.:16.50
##  Max.   :249  Max.   :33.80                    Max.   :24.80
##
##      horatmin      tmax      horatmax      dir
##  Length:3652      Min.  : 0.40  Length:3652      Min.  : 2.00
##  Class :character 1st Qu.:14.90  Class :character 1st Qu.:24.00
##  Mode  :character Median :21.70  Mode  :character Median :29.00
##                                         Mean  :22.14
##                                         3rd Qu.:29.30
##                                         Max.  :44.50
##                                         NA's  :1
##
##      velmedia      racha      horaracha
##  Min.  : 0.000  Min.  : 1.9  Length:3652
##  1st Qu.: 2.500 1st Qu.: 8.9  Class :character
##  Median : 3.900  Median :12.2  Mode  :character
##  Mean   : 4.644  Mean   :12.5
##  3rd Qu.: 6.400  3rd Qu.:15.6
##  Max.   :15.300  Max.   :37.5
##  NA's   :1       NA's   :1
```

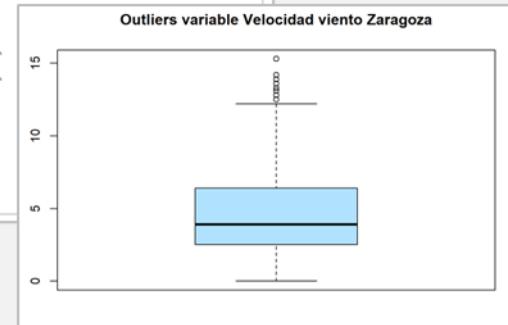


Imagen 41. Datos estadísticos de la variable Vel_vien_Zar.
Fuente: Propia

- La Coruña

```

##    fecha      indicativo      nombre      provincia
##  Length:3652  Length:3652  Length:3652  Length:3652
##  Class :character  Class :character  Class :character  Class :character
##  Mode  :character  Mode  :character  Mode  :character  Mode  :character
##
##
##
##    altitud      tmed      prec      tmin
##  Min.   :674  Min.  :-4.00  Length:3652  Min.   :-11.300
##  1st Qu.:674  1st Qu.: 9.40  Class :character  1st Qu.: 3.800
##  Median :674  Median :15.20  Mode  :character  Median : 9.300
##  Mean   :674  Mean   :15.99                    Mean   : 9.555
##  3rd Qu.:674  3rd Qu.:22.80                    3rd Qu.: 15.400
##  Max.   :674  Max.   :34.10                    Max.   : 26.100
##
##    horatmin      tmax      horatmax      dir
##  Length:3652  Min.   : 0.40  Length:3652  Min.   : 0.00
##  Class :character  1st Qu.:14.70  Class :character  1st Qu.:14.00
##  Mode  :character  Median :21.90  Mode  :character  Median :26.00
##  Mean   :22.42                    Mean   :25.87
##  3rd Qu.:30.20                    3rd Qu.:28.00
##  Max.   :42.70                    Max.   :99.00
##  NA's   :19
##
##    velmedia      racha      horaracha
##  Min.   :0.000  Min.   : 0.000  Length:3652
##  1st Qu.:0.600  1st Qu.: 6.700  Class :character
##  Median :1.100  Median : 8.300  Mode  :character
##  Mean   :1.359  Mean   : 8.592
##  3rd Qu.:1.900  3rd Qu.:10.000
##  Max.   :7.200  Max.   :26.400
##  NA's   :16    NA's   :19

```

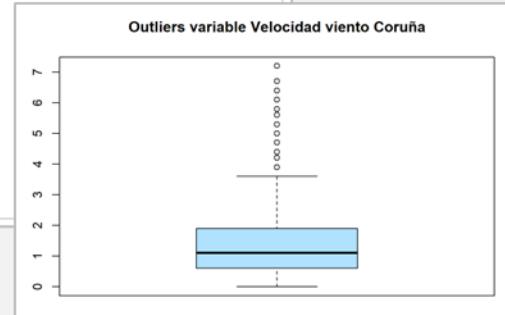


Imagen 42. Datos estadísticos de la variable Vel_vien_Cor.

Fuente: Propia

- Huelva

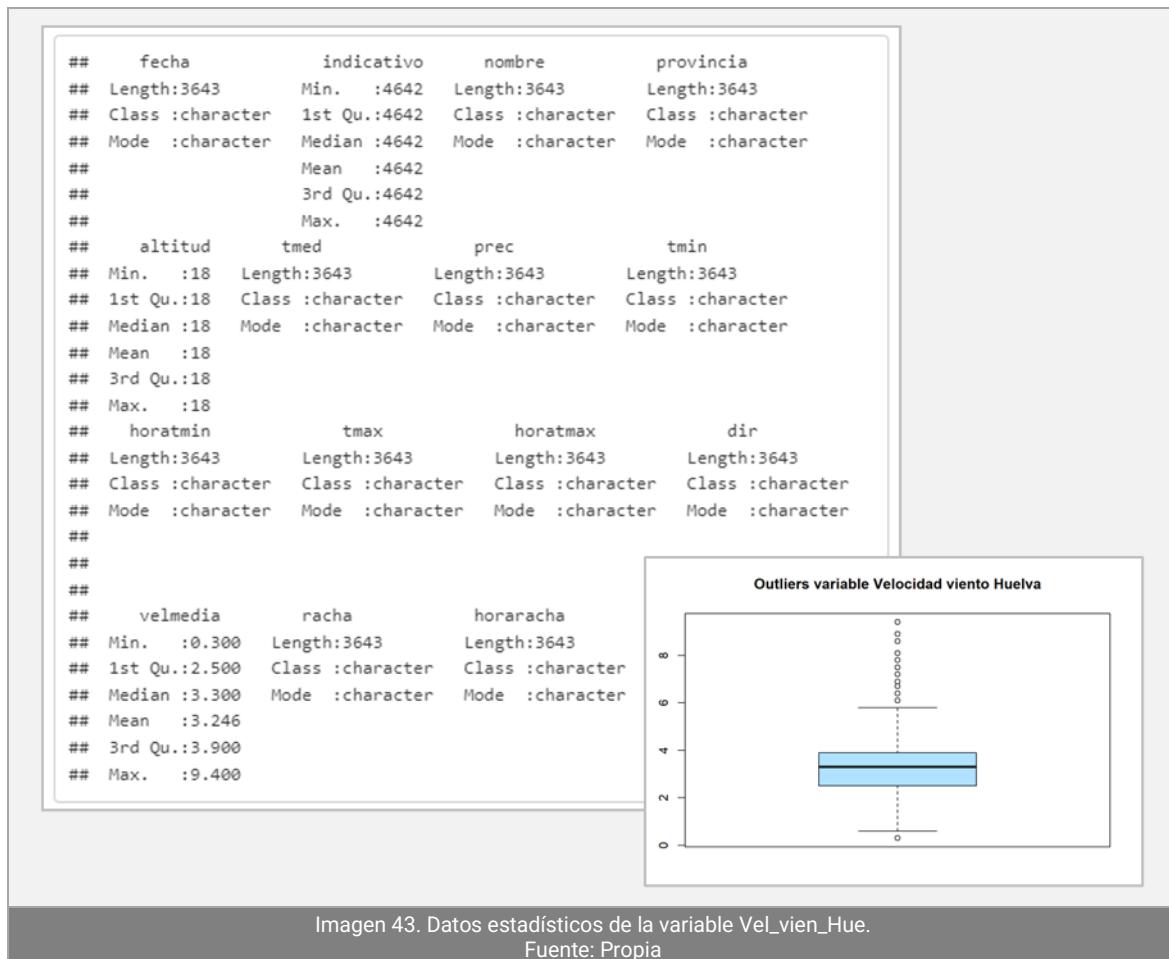


Imagen 43. Datos estadísticos de la variable Vel_vien_Hue.

Fuente: Propia

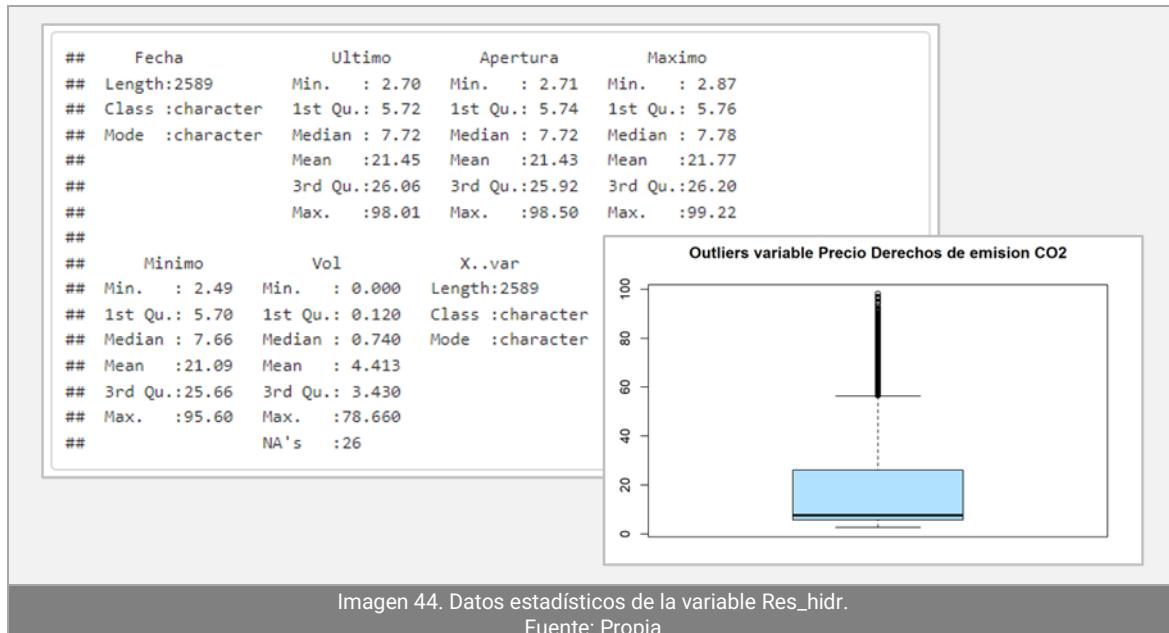
En cuanto a los **datos faltantes** en ninguna de las 5 ciudades serán un problema, ya que la que peor resultado tiene solo posee 42 NA's, sobre un total de 3652.

En cuanto a los **outliers** todas las ciudades están entre un 5,37%, de Valladolid que es la que más posee y un 0,6% las que menos.

6.5.9 VARIABLE: Reservas hidráulicas (Res_hidr)

En este dataframe existen **48.205 registros**. Aunque es una variable que está medida de forma temporal cada semana, el número de registros es muy elevado debido a que existe un gran número de embalses productores de EE en nuestro país. Para los modelos habrá que trabajar con este dataframe para reducir a un solo dato por registro por día, y no uno

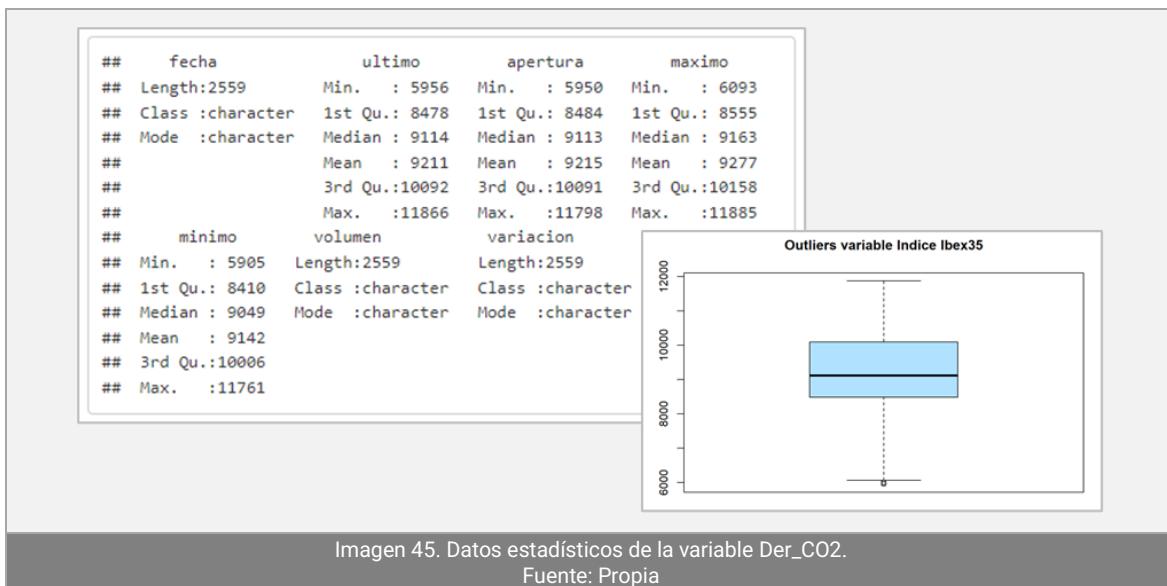
por embalse y día. Aunque tenemos ya el porcentaje de llenado de cada embalse, lo que nos interesaría para el estudio será el porcentaje general de todos ellos, por lo que deberemos calcularlo a partir de las columnas de “Agua_total” y “Agua_actual”.



Este dataframe no presenta **valores faltantes**, y los **outliers** son despreciables porque representan un 0.22% del total.

6.5.10 VARIABLE: Precio derechos de emisión CO₂ (Der_CO2)

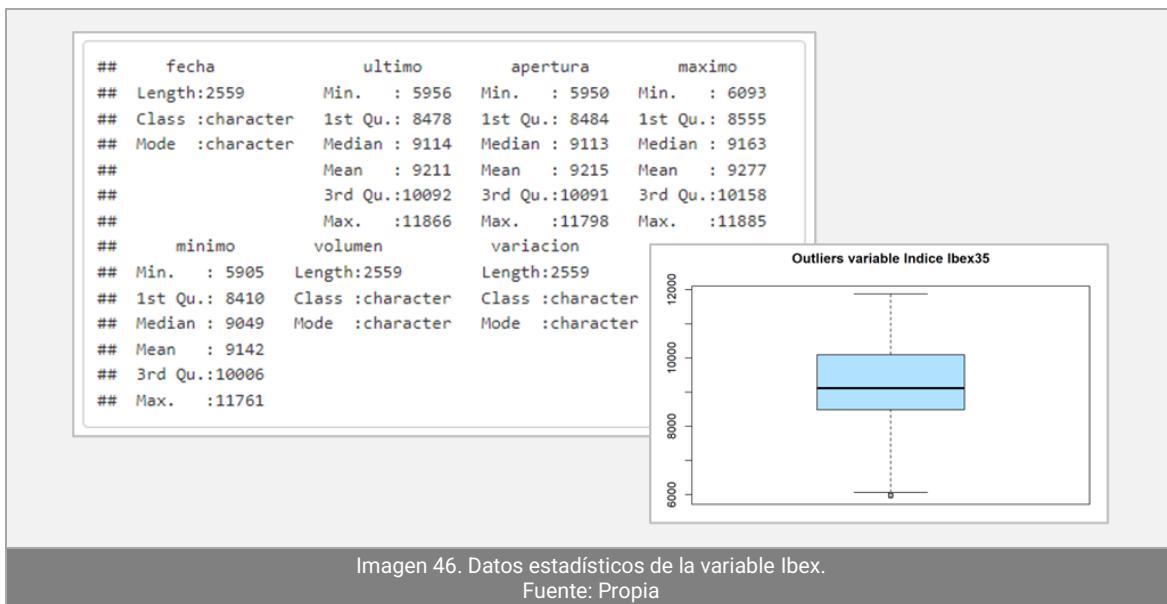
El dataframe obtenido de la fuente presenta varias columnas representando diferentes precios, pero solo nos interesa el “ultimo”, y como no, la “fecha”. Según el número de registros vemos que la cotización de estos derechos se negocia de lunes a viernes. Para solucionar esto último, se dará a los sábados y domingos el mismo valor el del viernes anterior.



Como podemos observar para la columna “Ultimo” no tenemos valores faltantes. En cuanto a los valores anormales la variable presenta **357 valores** que representan el **13,79%**, por lo que se tendrá que tomar alguna decisión de cómo actuar con estos valores.

6.5.11 VARIABLE: Situación socio-económica del país (Ibex)

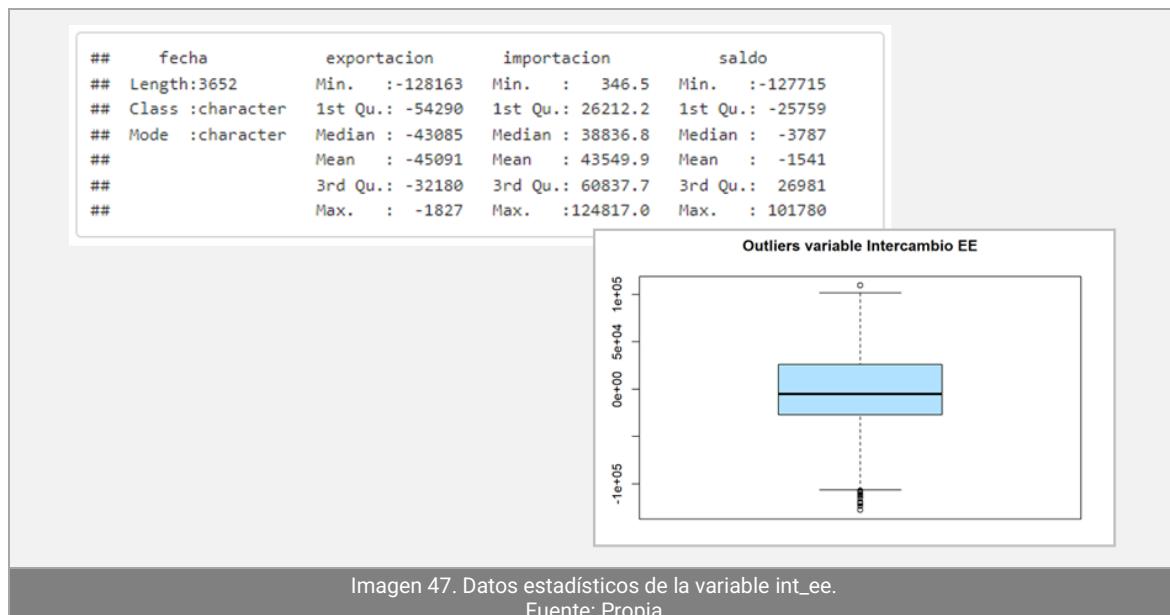
El dataframe que incluye los datos de esta variable tiene la misma estructura que el caso anterior por lo que las dos columnas que nos interesan son también “ultimo” y “fecha”. El número de registros corresponde también de lunes a viernes, por lo que se asignará para los sábados y domingos el mismo valor que el del viernes anterior.



En esta ocasión ni tenemos valores faltantes, ni valores anormales.

6.5.12 VARIABLE: Intercambio de EE con otros países (int_ee) _____

Esta dataframe solo contiene 4 columnas, de las que nos interesan la de “fecha” y la de “saldo”. Al tener **3653 registros** podemos comprobar que tenemos registros todos los días de la semana.



No presenta **valores faltantes** y los outliers son despreciables en esta variable, ya que solo suponen un **0,52%**.

6.6 TRATAMIENTO DE LOS DATOS Y FORMACION DEL DATASET

En este capítulo, partiendo de la exploración realizado en el apartado anterior, se explicarán las **acciones que hay que llevar a cabo sobre los datos originales** obtenidos de las diversas fuentes, para poder obtener un **dataset** con el que trabajar con los algoritmos.

Para trabajar de forma más efectiva, antes de realizar las transformaciones, se ha creado el **dataset** con todas las **variables** indicadas en el capítulo anterior, y a partir de él, se han realizado las acciones necesarias, para obtener el **dataset definitivo** con el que trabajar en los **modelos de predicción**.

6.6.1 UNION DE LAS VARIABLES EN UN DATAFRAME

Para realizar la unión de las variables se ha partido del archivo de la **variable dependiente** (**Pre_elec**), ya que a parte de que es la variable más importante al ser la que se quiere predecir, la columna “**fecha**” del archivo donde se encuentra, no le falta ningún registro, es decir, están todos los días de los 10 años que se han tomado como serie temporal para el estudio.

Del dataframe de este archivo se han borrado todas las columnas exceptuando “**fecha**” y “**Prec_elec**”, transformándose al mismo tiempo los datos de la columna “**fecha**” a datos “**Date**”, ya que al cargarlos **RStudio** los consideras “**character**” y no podríamos trabajar adecuadamente con ellos.

Esta ultima operación con la columna “**fecha**” se ha realizado en todos los archivos al cargar las diferentes variables.

Sobre el dataframe creado, se ha establecido otra columna (variable) llamada “**dia_semana**” donde se ha incluido los días de la semana en números según la columna “**fecha**”.

Posteriormente, una vez cargado el archivo de la siguiente variable (**Dem**), se han borrado las columnas que no nos interesan, para quedarnos solo con “**fecha**” y “**Dem**”.

Este ultimo proceso expuesto, se repetirá con todos los archivos restantes en los que se encuentran las distintas variables que debemos ir uniendo al **dataframe principal**.

Una especial mención requiere la variable “**Res_hidr**”. Esta variable, como ya se ha indicado anteriormente, posee 48.205 registros debido a que existen 100 embalses productores de EE y cada uno posee aproximadamente un registro por semana. Introducir esta variable así no seria viable por lo que tenemos que transformarla.

Lo que se ha hecho es sumar todas las capacidades para cada registro semanal y se ha creado una nueva columna “**Res_hidr**” para obtener el porcentaje de llenado total de todos los embalses para cada semana. Este ultimo dato es el que se ha introducido en el dataframe principal.

El **dataset obtenido después de estas labores** se ha descargado en el archivo:

- 13.Dataset de datos (df_3).csv
-

6.6.2 TRATAMIENTO DE LOS VALORES FALTANTES

A continuación, incluimos el resumen obtenido de RStudio de la **existencia de datos faltantes (NA's) en las variables del dataset principal.**

##	fecha	dia_sem	Pre_elec	Dem	Prec_petr
##	0	0	0	1	1022
##	Prec_gas	Prec_carb	Prod_eol	Prod_sol	Prod_hidr
##	1111	1094	0	0	0
##	Prod_ofr	Prod_nucl	Prod_pet	Prod_gas	Prod_carb
##	0	0	0	0	0
##	Prod_comb	Prod_cog	Prod_no_ren	Temp_min_Mad	Temp_max_Mad
##	0	0	0	15	15
##	Temp_min_Bar	Temp_max_Bar	Temp_min_Val	Temp_max_Val	Temp_min_Sev
##	426	425	0	8	15
##	Temp_max_Sev	Temp_min_Zar	Temp_max_Zar	Vel_media_Val	Vel_media_Alb
##	13	0	0	5	16
##	Vel_media_Zar	Vel_media_Cor	Vel_media_Hue	Res_hidr	Der_CO2
##	1	42	9	3183	1063
##	Ibex	Int_ee			
##	1093	0			

Imagen 52. Valores faltantes en las variables
Fuente: Propia

El número puede ser diferente al indicado en el capítulo de “Exploración de datos” debido a que en él se estudiaba las variables independientemente, es decir, según se han obtenido y en ellas solo se muestran los valores NA's de los registros anotados. Al unir todas las variables en un solo dataframe que posee un registro por cada día de los 10 años, se ha dado el caso en algunas variables que no poseían registros para todos esos días y por lo tanto han aparecido valores NA's en esos días. A continuación, señalamos la solución adoptada para esos valores.

Para las variables que presentan pocos valores faltantes se ha adoptado el criterio general de asignarles el valor del día anterior. Este es el caso de:

- “**Dem**” con un solo **un valor** faltante.
- “**Temp_min_Mad**” y la “**Temp_max_Mad**” que presentan **15** cada una.
- “**Temp_max_Val**” solo **8**.

- "**Temp_min_Sev**" y "**Temp_max_Sev**" existen **15** y **13** respectivamente.
- "**Vel_media_Val**" posee **5**.
- "**Vel_media_Alb**" tiene **16**.
- "**Vel_media_Zar**" presenta solo **1**.
- "**Vel_media_Cor**" posee **42**.
- "**Vel_media_Hue**" presenta **9**.

Para "**Prec_petr**" que posee **1.022** correspondientes a los **sábados** y los **domingos** en el periodo de los 10 años, ya que el mercado de negociación de esta materia prima no se efectúa estos días, se ha adoptado la decisión de asignar el valor del viernes anterior.

El caso de las variables "**Prec_gas**" y "**Prec_carb**" que presentan **1.111** y **1.094** valores NA, es el mismo caso que el anterior y se ha seguido el mismo criterio.

En el caso de la "**Temp_min_Bar**" y "**Temp_max_Bar**" presentan **426** y **425** respectivamente cada una. El motivo de esto es que la fuente de datos no los proporciona al hacer la consulta, faltando sobre todo en tres periodos claramente marcados. El primero corresponden a los 11 primeros valores del año 2012, el segundo entre el 1 de enero y el 13 junio de 2017, y el segundo entre el 1 de julio y el 31 de diciembre del 2020. La solución adoptada ha sido la de tomar para el primer tramo la media de los valores de los dos años posteriores, para el segundo tramo la media de los valores del año anterior y posterior, y para el ultimo tramo la media de los valores de los dos años anteriores.

no influirán en el resultado. Se ha tomado la decisión de tomar el valor del día anterior.

La variable "**Res_hidr**" posee **3.183**. En este caso existen tantos datos faltantes debido a que los registros obtenidos son semanales, por lo que solo se tiene 1 valor de cada 7 de cada semana. Ya que la variación de los recursos hídricos varía lentamente, la solución adoptada es la de dar el mismo valor a todos los días de la semana.

En cuanto a "**Der_CO2**" que posee **1063** y "**Ibex**" que posee **1.093** datos faltantes, sucede lo mismo que las variables de precio del petróleo, gas natural y carbón, por lo que la solución tomada ha sido la misma.

El resto de variables no presentan ningún dato NA.

6.6.3 CORRELACION ENTRE VARIABLES

El escenario óptimo es aquel en que todas las variables predictoras se correlacionen con la variable de salida, pero no entre sí. Esto en la práctica es una situación muy improbable.

Una solución a este problema sería excluir aquellas variables que presenten una correlación notablemente alta, aunque también se puede incurrir en el error de prescindir de factores importantes en la predicción a realizar.

Poseer en el dataframe variables correlacionadas, supone agregar complejidad al modelo, es decir, sería más complicado de entender, supondría un modelo menos explicable, sería también un modelo menos preciso, y no podemos olvidar que normalmente llevaría más tiempo a la hora de los cálculos y obtener los resultados.

Para cuantificar la correlación entre dos variables se emplea el **coeficiente de correlación**. El resultado del coeficiente puede tomar valores entre **-1** y **+1**.

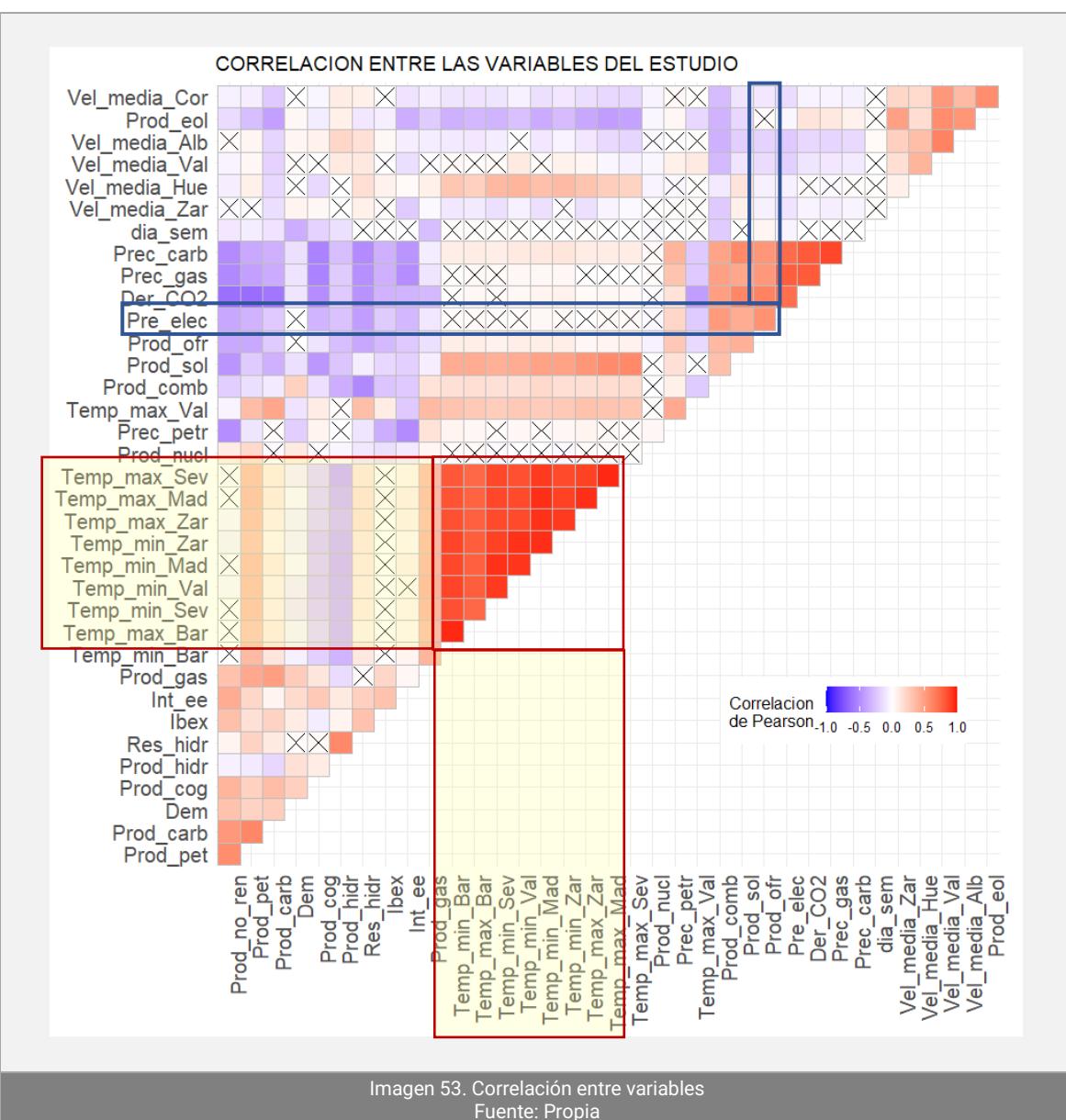
El signo indica el tipo de correlación entre las dos variables, un signo positivo indica que existe una relación directa entre las dos variables, es decir, que la subida de una de las variables supondrá la subida de la otra, y un signo negativo indica que existe una relación indirecta, si el valor de una variable sube, la de la otra baja. Si el coeficiente de correlación es cero dos variables son independientes. La fuerza de la relación incrementa a medida que el coeficiente de correlación se aproxima a +1 o a -1.

El método empleado para estudiar la correlación entre variables ha sido el cálculo del **Coeficiente de correlación de Pearson**. Si el valor de correlación entre dos variables es superior a 0,95 puntos se puede considerar que esas dos variables están correlacionadas.

A continuación, se incluye el gráfico de Correlación si como los colores de Coeficiente de Pearson.

MODELOS PARA LA PREDICCION DEL PRECIO DE LA ELECTRICIDAD

EN EL MERCADO ESPAÑOL



MODELOS PARA LA PREDICCIÓN DEL PRECIO DE LA ELECTRICIDAD
EN EL MERCADO ESPAÑOL

dia_sem	1											
Pre_elec	-0.05	1										
Dem	-0.35	0	1									
Prec_petr	0	0.21	-0.21	1								
Prec_gas	0	0.79	-0.12	0.31	1							
Prec_carb	0	0.73	-0.13	0.39	0.87	1						
Prod_eol	0.01	-0.04	0.04	0.04	0.12	0.07	1					
Prod_sol	0	0.43	-0.13	0.16	0.46	0.6	-0.17	1				
Prod_hidr	-0.13	-0.26	0.16	0	-0.22	-0.28	0.08	-0.25	1			
Prod_ofr	0.05	0.56	0	0.29	0.53	0.54	0.02	0.41	-0.28	1		
Prod_nucl	-0.02	-0.03	0.11	0.04	0.02	0.03	-0.04	0	-0.04	-0.06	1	
Prod_pet	-0.08	-0.33	0.23	-0.11	-0.41	-0.37	-0.28	-0.22	-0.09	-0.37	0.2	1
Prod_gas	-0.28	-0.1	0.27	0.15	-0.11	-0.08	-0.22	-0.07	-0.15	-0.13	0.05	0.45
Prod_carb	-0.12	-0.23	0.26	0.02	-0.36	-0.34	-0.41	-0.34	-0.17	-0.22	-0.02	0.61
Prod_comb	-0.22	0.52	0.24	0.08	0.49	0.47	-0.33	0.33	-0.35	0.37	-0.01	-0.14
Prod_cog	-0.19	-0.32	0.25	0.06	-0.54	-0.54	-0.08	-0.45	0.1	-0.12	-0.02	0.24
Prod_no_ren	-0.1	-0.38	0.32	-0.49	-0.51	-0.5	-0.15	-0.46	-0.06	-0.37	0.1	0.58
Temp_min_Mad	0	0.04	-0.04	0.02	0.04	0.12	-0.36	0.48	-0.34	0.09	0	0.36
Temp_max_Mad	0	0.03	-0.04	0.03	0	0.1	-0.42	0.59	-0.32	0.06	0.02	0.37
Temp_min_Bar	0.01	0.02	-0.06	0.06	0.01	0.09	-0.34	0.43	-0.32	0.09	0.02	0.36
Temp_max_Bar	0.01	0.03	-0.05	0.04	0.03	0.1	-0.31	0.43	-0.32	0.09	0.03	0.32
Temp_min_Val	0.01	0.02	-0.08	0.04	0.04	0.11	-0.27	0.43	-0.33	0.09	-0.01	0.36
Temp_max_Val	0.01	-0.22	-0.12	0.44	-0.23	-0.23	-0.17	-0.01	0.02	-0.16	0.02	0.35
Temp_min_Sev	0.01	0.02	-0.1	0.02	0.03	0.09	-0.32	0.41	-0.32	0.07	-0.02	0.35
Temp_max_Sev	0	0.03	-0.07	0.02	0	0.1	-0.4	0.6	-0.36	0.07	0.01	0.35
Temp_min_Zar	0	0.03	-0.06	0.04	0.04	0.13	-0.29	0.48	-0.33	0.09	-0.03	0.33
Temp_max_Zar	0.01	0	-0.06	0.06	-0.01	0.1	-0.38	0.54	-0.29	0.04	0.02	0.35
Vel_media_Val	0	-0.2	0.02	0.08	-0.1	-0.09	0.58	-0.11	0.11	-0.15	-0.04	0.04
Vel_media_Alb	0.03	-0.23	0.05	0	-0.18	-0.16	0.54	-0.21	0.21	-0.19	-0.03	0.05
Vel_media_Zar	0.01	-0.12	0.06	0.02	-0.07	-0.05	0.49	0.08	0	-0.06	0	-0.01
Vel_media_Cor	0	-0.15	0	0.03	-0.05	-0.06	0.58	-0.11	0.1	-0.1	-0.05	-0.07
Vel_media_Hue	0.01	-0.06	-0.01	-0.03	-0.03	0.02	0.18	0.14	0	-0.06	-0.05	0.08
Res_hidr	0	-0.42	0.02	-0.11	-0.48	-0.5	-0.1	-0.07	0.6	-0.38	-0.1	0.23
Der_CO2	0	0.72	-0.11	0.15	0.8	0.8	0.14	0.61	-0.23	0.63	-0.02	-0.63
Ibex	0	-0.22	0.1	-0.35	-0.34	-0.34	-0.07	-0.17	0.06	-0.3	-0.13	0.16
Int_ee	-0.01	-0.31	0.19	-0.5	-0.49	-0.46	-0.35	-0.2	0.11	-0.23	-0.09	0.21
	dia_sem	Pre_elec	Dem	Prec_petr	Prec_gas	Prec_carb	Prod_eol	Prod_sol	Prod_hidr	Prod_ofr	Prod_nucl	Prod_pet

MODELOS PARA LA PREDICCIÓN DEL PRECIO DE LA ELECTRICIDAD EN EL MERCADO ESPAÑOL

Temp_max_Sev	1													
Temp_min_Zar	0.86	1												
Temp_max_Zar	0.89	0.91	1											
Vel_media_Val	-0.06	0.07	0.05	1										
Vel_media_AlB	-0.18	-0.05	-0.1	0.63	1									
Vel_media_Zar	-0.03	-0.02	-0.13	0.17	0.25	1								
Vel_media_Cor	-0.17	-0.1	-0.14	0.53	0.36	0.21	1							
Vel_media_Hue	0.23	0.36	0.31	0.39	0.34	0.08	0.19	1						
Res_hidr	0.14	0.13	0.17	0.12	0.16	0.08	0.08	0.17	1					
Der_CO2	0.04	0.07	0.03	-0.11	-0.17	-0.05	-0.05	0.02	-0.47	1				
Ibex	0.01	0	0	-0.01	0.04	0.02	0	0.07	0.32	-0.33	1			
Int_eee	0.07	0.05	0.08	-0.14	-0.07	-0.22	-0.12	0.03	0.25	-0.32	0.32	1		
	Temp_max_Sev	Temp_min_Zar	Temp_max_Zar	Vel_media_Val	Vel_media_AlB	Vel_media_Zar	Vel_media_Cor	Vel_media_Hue	Res_hidr	Der_CO2	Ibex	Int_eee		

Imagen 54. Coeficiente de correlación de Pearson
Fuente: Propia

Como se puede comprobar en el grafico **existe una gran correlación entre las variables de temperaturas máximas y mínimas** de las 5 ciudades tomadas.

Como se indicó anteriormente, tener en el dataframe variables con alta correlación supone agregar complejidad al modelo. Por este motivo **se ha decidido reducir las variables de temperatura máxima y mínima en una sola**, que llamaremos "**Temp_gen**".

Centrando nuestra atención en los valores de correlación de la variable de entrada con las de salida, podemos comprobar que los valores no son altos, excepto para la relación con la variable "**Prec_gas**", aunque no alcanza un valor suficiente para considerar que estas dos variables presentan una relación directa.

6.6.4 DATASET FINAL

Para obtener el dataset definitivo con el que trabajar con los modelos de predicción **se unificarán las variables de temperaturas mínimas y máximas de las cinco ciudades seleccionadas**.

Para la nueva variable que unifique las anteriores, se asignará en los meses más cálidos, entre abril y septiembre, la temperatura media de las temperaturas máximas de esos meses de las 5 ciudades, y para los otros 6 meses del año, la temperatura media de las temperaturas mínimas.

Esto se realiza así porque con estas variables lo que se busca es conocer la influencia de las temperaturas en el precio de la electricidad, y en verano la influencia es por las altas temperaturas y en invierno por las bajas temperaturas. Para realizar esta labor se ha empleado el programa **Microsoft Excel**.

Una vez realizado este trabajo, se sustituyen las variables de temperaturas de las 5 ciudades, por la obtenida en esta labor, y de esta forma obtenemos un **dataframe** formado por **28 variables y 3652 registros**, con el que se va a trabajar en los modelos de predicción.

El **dataset final obtenido después de estas labores** y con el que se va a trabajar en los modelos de predicción, se encuentra en el archivo:

- **15.Dataset final de datos.csv**

MODELOS PARA LA PREDICCIÓN DEL PRECIO DE LA ELECTRICIDAD

EN EL MERCADO ESPAÑOL

A continuación, adjuntamos captura de **RStudio** del resumen estadístico de este dataframe.

```
##   fecha          dia_sem     Pre_elec      Dem
## Length:3652      Min.   :1       Min.   : 10.04  Min.   :19122
## Class :character  1st Qu.:2      1st Qu.: 51.10  1st Qu.:26233
## Mode  :character  Median :4      Median : 59.18  Median :28195
##                  Mean   :4      Mean   : 73.89  Mean   :28045
##                  3rd Qu.:6      3rd Qu.: 69.06  3rd Qu.:29903
##                  Max.   :7      Max.   :556.15  Max.   :35306
##   Prec_petr      Prec_gas      Prec_carb      Prod_eol
## Min.   :27.88    Min.   :24.18    Min.   :43.40    Min.   : 9.688
## 1st Qu.:56.98   1st Qu.:41.62   1st Qu.:63.04   1st Qu.: 84.028
## Median :71.89   Median :53.89   Median :82.05   Median :127.643
## Mean   :77.94   Mean   :77.80   Mean   :101.30   Mean   :143.395
## 3rd Qu.:105.68  3rd Qu.:66.66  3rd Qu.:94.80   3rd Qu.:190.209
## Max.   :130.24  Max.   :640.36  Max.   :439.00  Max.   :430.148
##   Prod_sol      Prod_hidr      Prod_ofr      Prod_nucl
## Min.   : 3.795   Min.   :17.72    Min.   : 6.781   Min.   : 79.42
## 1st Qu.:23.712  1st Qu.:54.43   1st Qu.:11.915  1st Qu.:141.95
## Median :40.571  Median :74.95   Median :12.780  Median :156.10
## Mean   :44.367  Mean   :86.20   Mean   :12.847  Mean   :151.38
## 3rd Qu.:56.877  3rd Qu.:105.64  3rd Qu.:13.610  3rd Qu.:168.51
## Max.   :157.109  Max.   :266.07  Max.   :18.309  Max.   :180.17
##   Prod_pet      Prod_gas      Prod_carb      Prod_comb
## Min.   : 4.579   Min.   :0.1438   Min.   : 0.4657  Min.   : 22.73
## 1st Qu.: 7.805  1st Qu.:1.4439  1st Qu.:27.4792 1st Qu.: 60.56
## Median : 8.655  Median :2.1196  Median :84.5868  Median : 86.07
## Mean   : 8.600  Mean   :2.1835  Mean   :91.4891  Mean   :105.83
## 3rd Qu.: 9.527  3rd Qu.:2.8689  3rd Qu.:150.2429 3rd Qu.:134.89
## Max.   :12.026  Max.   :5.2203  Max.   :238.1793  Max.   :396.45
##   Prod_cog      Prod_no_ren    Temp_gen      Vel_media_Val
## Min.   :18.50    Min.   : 5.669   Min.   :-1.140   Min.   : 0.300
## 1st Qu.:68.57    1st Qu.:10.069  1st Qu.: 7.827   1st Qu.:1.400
## Median :74.90    Median :12.012   Median :16.200  Median : 1.700
## Mean   :73.72    Mean   :11.851   Mean   :17.079  Mean   : 2.006
## 3rd Qu.:81.38    3rd Qu.:13.593  3rd Qu.:26.330  3rd Qu.:2.500
## Max.   :103.09   Max.   :17.348   Max.   :38.280  Max.   : 8.300
##   Vel_media_Alb  Vel_media_Zar  Vel_media_Cor  Vel_media_Hue
## Min.   : 0.00    Min.   : 0.000   Min.   : 0.60   Min.   : 0.300
## 1st Qu.: 0.60    1st Qu.: 2.500   1st Qu.: 2.20   1st Qu.: 2.500
## Median : 1.10    Median : 3.900   Median : 3.30   Median : 3.300
## Mean   : 1.36    Mean   : 4.644   Mean   : 3.61   Mean   : 3.243
## 3rd Qu.: 1.90    3rd Qu.: 6.400   3rd Qu.: 4.70   3rd Qu.: 3.900
## Max.   : 7.20    Max.   :15.300   Max.   :11.10   Max.   : 9.400
##   Res_hidr      Der_CO2       Ibex        Int_ee
## Min.   :40.27    Min.   : 2.700   Min.   : 5956   Min.   :-127715
## 1st Qu.:56.26    1st Qu.: 5.720   1st Qu.: 8486   1st Qu.: -25759
## Median :62.82    Median : 7.745   Median : 9125   Median : -3787
## Mean   :66.22    Mean   :21.555   Mean   : 9215   Mean   : -1541
## 3rd Qu.:77.59    3rd Qu.:26.282  3rd Qu.:10093  3rd Qu.: 26981
## Max.   :92.90    Max.   :98.010   Max.   :11866  Max.   : 101780
```

Imagen 55. Resumen estadístico de las variables del dataset final
Fuente: Propia

6.6.5 TRATAMIENTO DE LOS VALORES OUTLIERS

Los **valores extremos** que nos encontramos en nuestro dataframe final, no se han generado en el proceso de la formación de éste, ni tampoco en el proceso de la toma de datos de las variables, por lo tanto, su eliminación o sustitución por otro valor puede no ser una buena solución, ya que introduciría un sesgo artificial en las variables que puede afectar al resultado de las predicciones.

Para conocer que solución tomar con este tipo de valores debemos fijarnos en los modelos predictivos que vamos a emplear. Según que modelos habrá que actuar o no sobre los valores anormales. De esta forma tendremos que:

- El **modelo de Regresión lineal múltiple** es muy sensible a los valores extremos, por lo que antes de aplicarlo, deberemos eliminar o limitar estos valores (esto se hará al aplicar el modelo al dataset final).
- El **modelo de K-vecinos más cercanos (KNN)** puede ofrecer buenos resultados, aunque existen outliers, aunque su la cantidad debe ser pequeña.
- En cuanto al resto de modelos que se emplearán, **Árbol de decisión, Random Forest, XG Boost y Red Neuronal**, no se ven afectados por este tipo de valores, así que no habrá que hacer un pretratamiento con ellos.

A continuación, presentamos el resumen de los valores extremos tomados de una captura de **RStudio**.

```
## [1] "El número de Outliers existentes en la columna 'Dem' es: 5"
## [2] "El número de Outliers existentes en la columna 'Prec_petr' es: 0"
## [3] "El número de Outliers existentes en la columna 'Prec_gas' es: 516"
## [4] "El número de Outliers existentes en la columna 'Prec_carb' es: 468"
## [5] "El número de Outliers existentes en la columna 'Prod_eol' es: 33"
## [6] "El número de Outliers existentes en la columna 'Prod_sol' es: 207"
## [7] "El número de Outliers existentes en la columna 'Prod_hidr' es: 172"
## [8] "El número de Outliers existentes en la columna 'Prod_ofr' es: 323"
## [9] "El número de Outliers existentes en la columna 'Prod_nucl' es: 67"
## [10] "El número de Outliers existentes en la columna 'Prod_pet' es: 20"
## [11] "El número de Outliers existentes en la columna 'Prod_gas' es: 4"
## [12] "El número de Outliers existentes en la columna 'Prod_carb' es: 0"
## [13] "El número de Outliers existentes en la columna 'Prod_comb' es: 160"
## [14] "El número de Outliers existentes en la columna 'Prod_cog' es: 248"
## [15] "El número de Outliers existentes en la columna 'Prod_no_ren' es: 0"
## [16] "El número de Outliers existentes en la columna 'Temp_gen' es: 0"
## [17] "El número de Outliers existentes en la columna 'Vel_media_Val' es: 197"
```

MODELOS PARA LA PREDICCIÓN DEL PRECIO DE LA ELECTRICIDAD

EN EL MERCADO ESPAÑOL

```
## [18] "El número de Outliers existentes en la columna 'Vel_media_Alb' es: 101"
## [19] "El número de Outliers existentes en la columna 'Vel_media_Zar' es: 22"
## [20] "El número de Outliers existentes en la columna 'Vel_media_Cor' es: 22"
## [21] "El número de Outliers existentes en la columna 'Vel_media_Hue' es: 76"
## [22] "El número de Outliers existentes en la columna 'Res_hidr' es: 0"
## [23] "El número de Outliers existentes en la columna 'Der_CO2' es: 505"
## [24] "El número de Outliers existentes en la columna 'Ibex' es: 5"
## [25] "El número de Outliers existentes en la columna 'Int_ee' es: 24"
```

Imagen 56. Número de valores anormales en las variables predictoras

Fuente: Propia

Estos valores en porcentaje sobre la cantidad de datos totales son:

```
## [1] "Y el porcentaje en la columna 'Dem' es: 0.137 %"
## [2] "Y el porcentaje en la columna 'Prec_petr' es: 0 %"
## [3] "Y el porcentaje en la columna 'Prec_gas' es: 14.129 %"
## [4] "Y el porcentaje en la columna 'Prec_carb' es: 12.815 %"
## [5] "Y el porcentaje en la columna 'Prod_eol' es: 0.904 %"
## [6] "Y el porcentaje en la columna 'Prod_sol' es: 5.668 %"
## [7] "Y el porcentaje en la columna 'Prod_hidr' es: 4.71 %"
## [8] "Y el porcentaje en la columna 'Prod_ofr' es: 8.844 %"
## [9] "Y el porcentaje en la columna 'Prod_nucl' es: 1.835 %"
## [10] "Y el porcentaje en la columna 'Prod_pet' es: 0.548 %"
## [11] "Y el porcentaje en la columna 'Prod_gas' es: 0.11 %"
## [12] "Y el porcentaje en la columna 'Prod_carb' es: 0 %"
## [13] "Y el porcentaje en la columna 'Prod_comb' es: 4.381 %"
## [14] "Y el porcentaje en la columna 'Prod_cog' es: 6.791 %"
## [15] "Y el porcentaje en la columna 'Prod_no_ren' es: 0 %"
## [16] "Y el porcentaje en la columna 'Temp_gen' es: 0 %"
## [17] "Y el porcentaje en la columna 'Vel_media_Val' es: 5.394 %"
## [18] "Y el porcentaje en la columna 'Vel_media_Alb' es: 2.766 %"
## [19] "Y el porcentaje en la columna 'Vel_media_Zar' es: 0.602 %"
## [20] "Y el porcentaje en la columna 'Vel_media_Cor' es: 0.602 %"
## [21] "Y el porcentaje en la columna 'Vel_media_Hue' es: 2.081 %"
## [22] "Y el porcentaje en la columna 'Res_hidr' es: 0 %"
## [23] "Y el porcentaje en la columna 'Der_CO2' es: 13.828 %"
## [24] "Y el porcentaje en la columna 'Ibex' es: 0.137 %"
## [25] "Y el porcentaje en la columna 'Int_ee' es: 0.657 %"
```

Imagen 57. Porcentaje de valores anormales en las variables predictoras

Fuente: Propia

Para evitar los valores fuera de rango vamos a emplear el **método de los percentiles**, para lo cual se ha empleado en **R** la función `boxplot()`, la cual detecta los outliers como los valores que está más allá de los bigotes. Los bigotes son las líneas que se determinan como el **tercer cuartil + 1,5 veces el rango intercuartílico** (tercer cuartil menos el primer cuartil) y el **primer cuartil - 1,5 veces el rango intercuartílico**.

Si una variable presenta valores fuera de estos límites intercuartílicos, estos pueden generar "ruido" en nuestro análisis, por lo que debemos actuar sobre ellos en los modelos que se ven afectados.

Las dos **actuaciones que se han llevado a cabo** para conocer cual es la más efectiva han sido:

- **Eliminar los outliers del dataset de datos.**
- **Sustituir los outliers por los valores máximos y mínimos calculados por el método intercuartílico.**

Estos tratamientos se han llevado a cabo con **Python** en los archivos señalados a continuación, que por su poco interés técnico no se van a desarrollar más aquí, trasladando al lector para su consulta a dichos archivos:

- 6.6.5 Procesado Outliers v06 (elim).ipynb
- 6.6.5 Procesado Outliers v07 (sust).ipynb

6.7 ANALISIS DE LAS VARIABLES

Para el análisis de las variables se ha empleado el lenguaje de programación **Python**, ya que se obtienen fácilmente las graficas de distribución y de relación de las variables.

6.7.1 VARIABLE DE RESPUESTA (Pre_elec)

Cuando se emplean algoritmos de Machine Learning, es importante estudiar la distribución de la variable respuesta, ya que, es lo que interesa predecir. La variable “**Pre_elec**” tiene una distribución asimétrica con una cola positiva debido a que, algunas fechas, tienen un precio muy superior a la media.

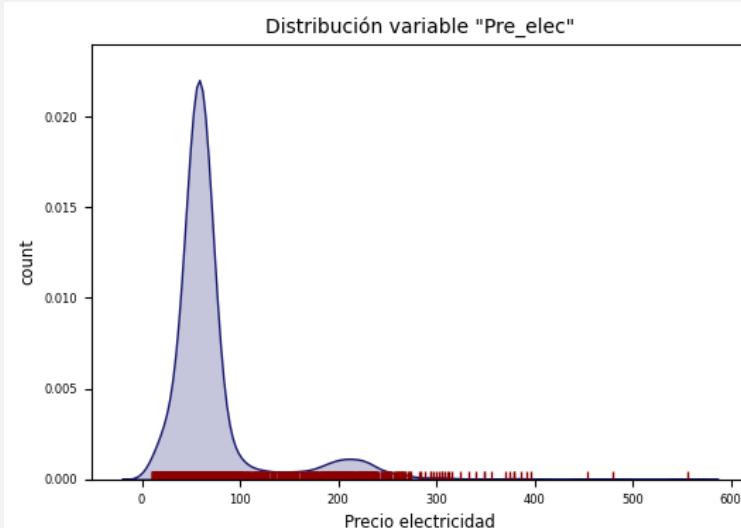


Imagen 48. Distribución de la variable "Pre_elec"

Fuente: Propia

Para conocer mejor, si la variable sigue una distribución Normal o Gaussiana, se ha recurrido a comparar su distribución, con la normal, a través de la representación de los cuartiles teóricos (Q-Q plot) que exponemos a continuación.

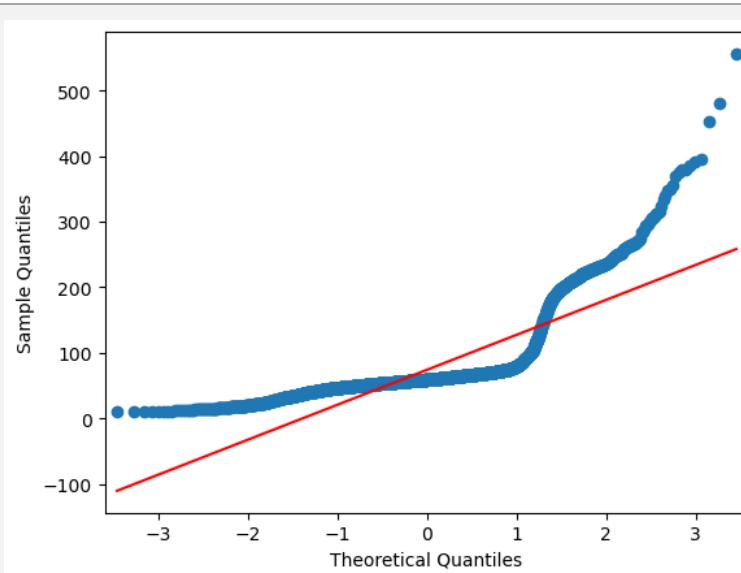


Imagen 49. Comparación de la distribución de la variable "Pre_elec" con la normal a través de los cuartiles teóricos (Q-Q plot)

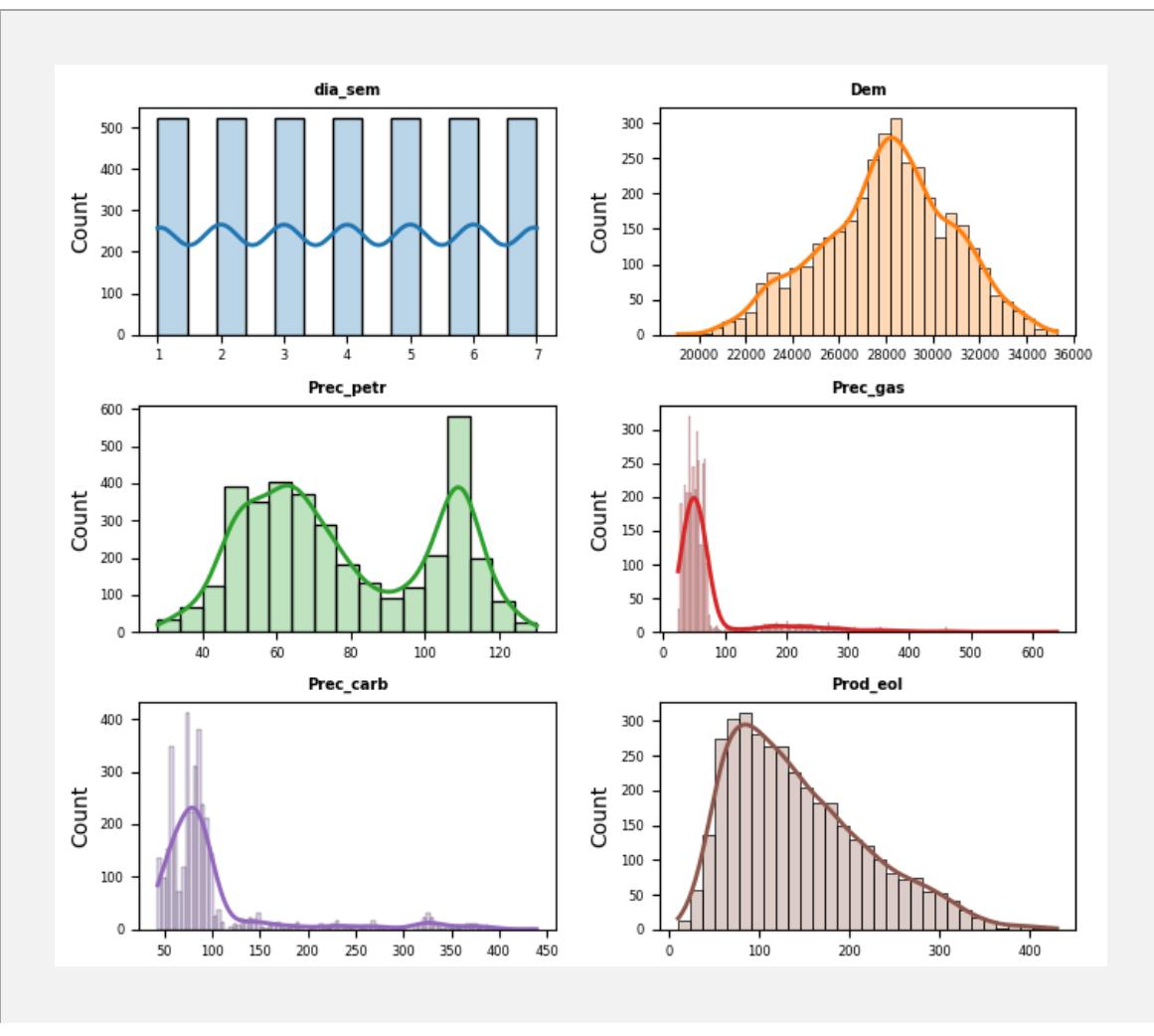
Fuente: Propia

Como se puede observar la distribución de nuestra variable casi no coincide con la línea roja, la cual marcaría la distribución normal.

Algunos modelos requieren que la **variable respuesta** se distribuya de forma normal, como son los **modelos de Regresión Lineal**. En nuestro caso al no poseer la variable una distribución normal, podemos intuir que el modelo de Regresión Lineal no tendrá unos buenos resultados, y existirán mejores algoritmos.

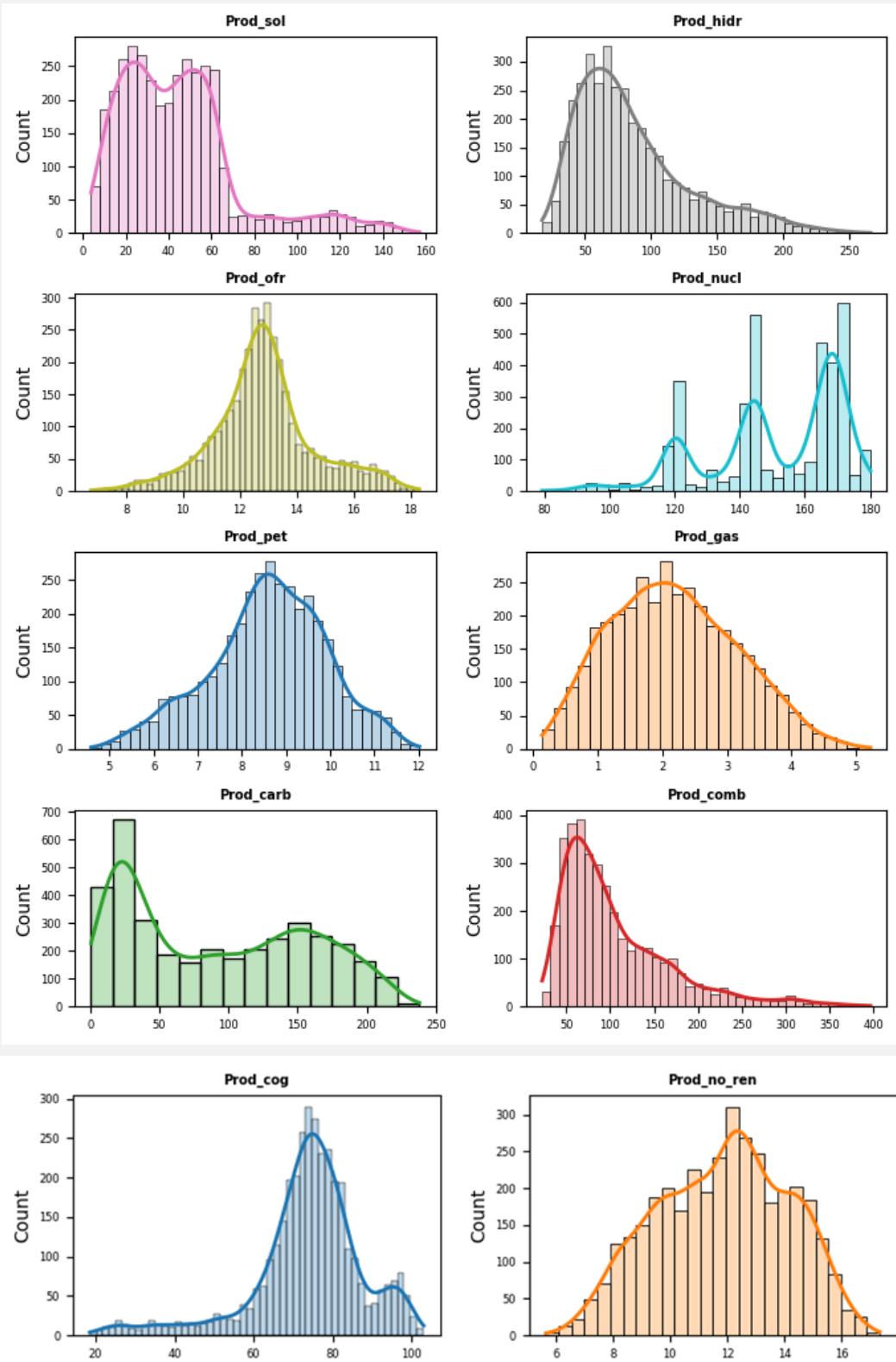
6.7.2 VARIABLES DE ENTRADA

A continuación, mostraremos las graficas de las variables de entrada para conocer su distribución.



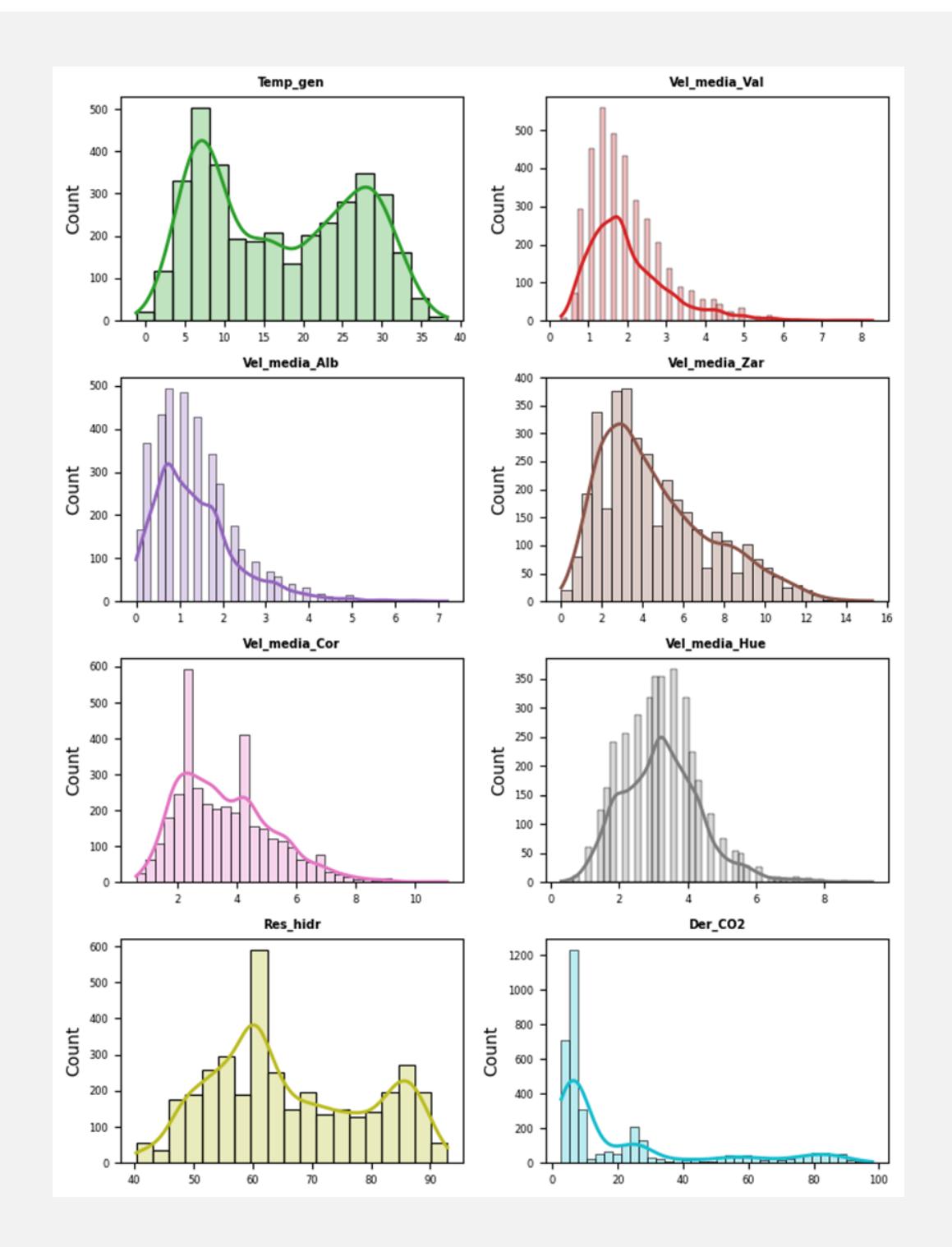
MODELOS PARA LA PREDICCIÓN DEL PRECIO DE LA ELECTRICIDAD

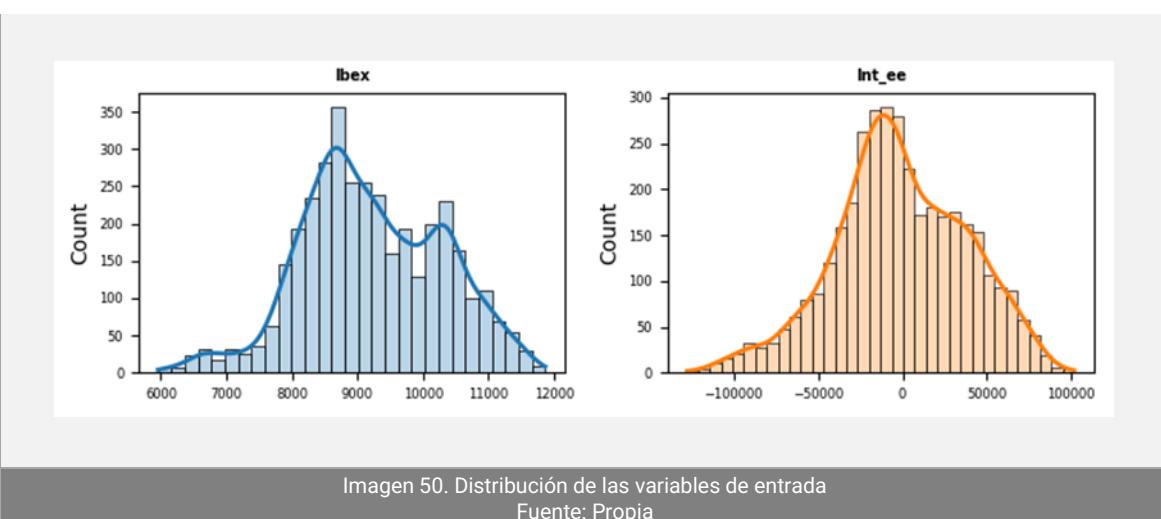
EN EL MERCADO ESPAÑOL



MODELOS PARA LA PREDICCIÓN DEL PRECIO DE LA ELECTRICIDAD

EN EL MERCADO ESPAÑOL

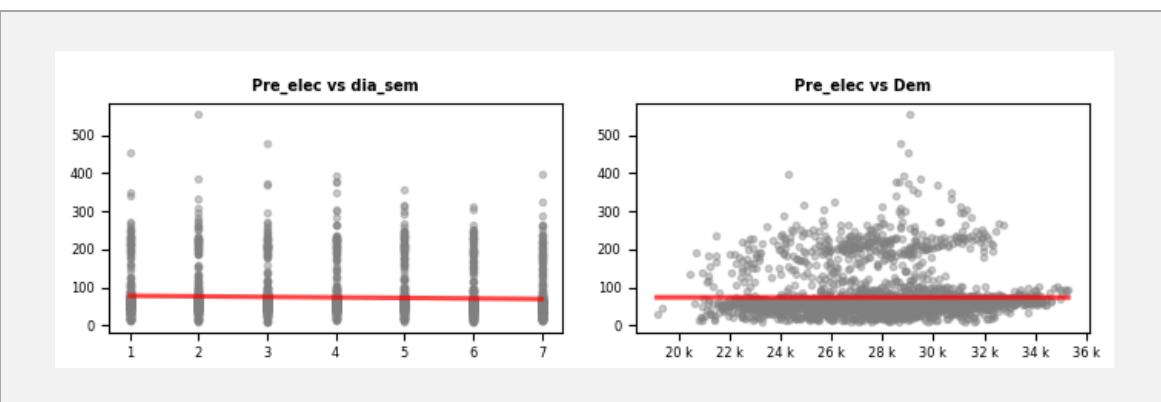




La variable “**dia_sem**”, aunque es de tipo numérico, toma pocos valores, del 1 al 7, ya que representan a los días de la semana. En estos casos se podría tratar esta variable como categórica, pero como todos los valores tienen más o menos el mismo numero de registros se ha optado por dejarla como está.

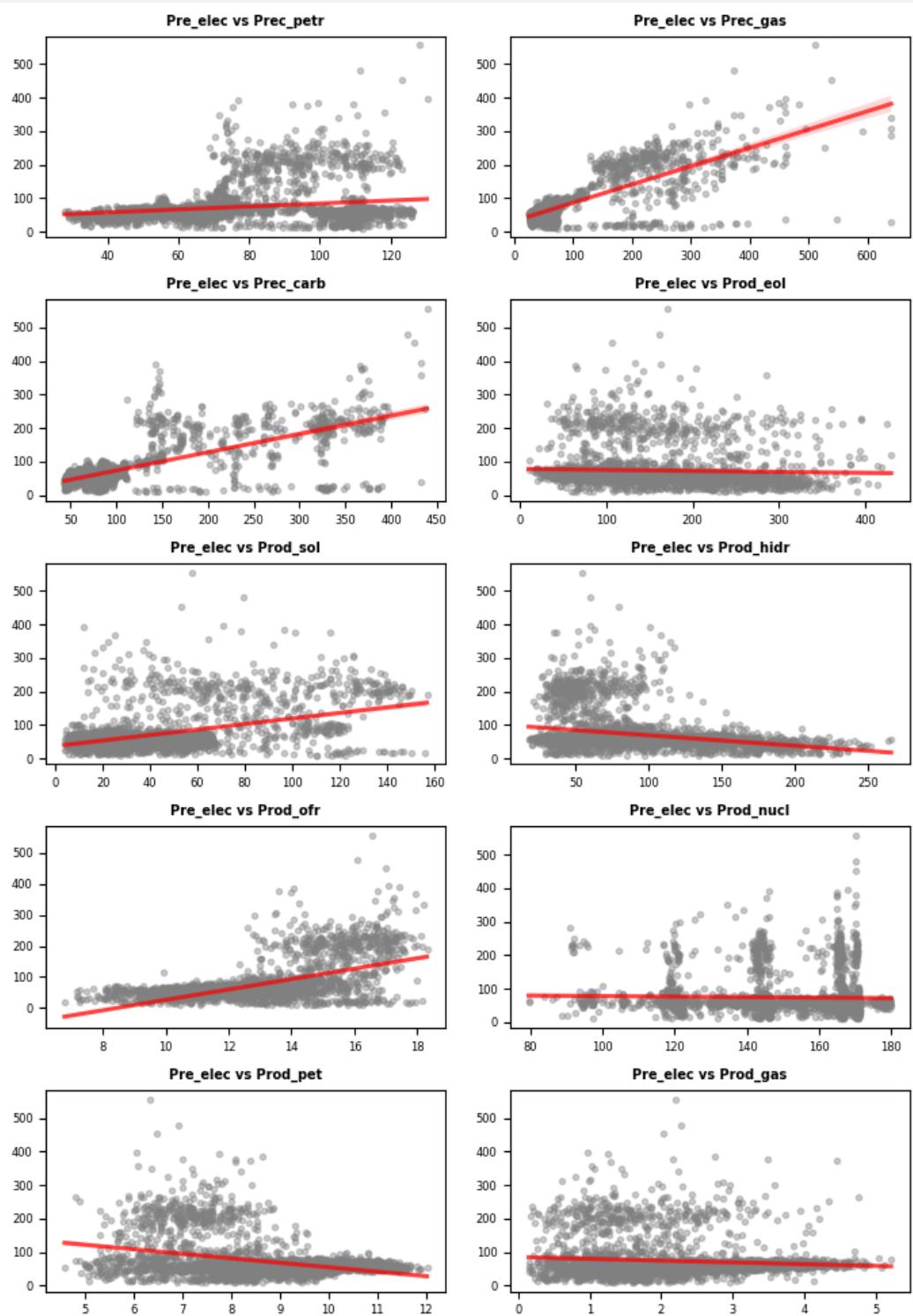
Tambien, con estas graficas podemos percibir que hay una serie de variables como son “**Prec_gas**”, “**Prec_carb**”, “**Prod_sol**”, “**Prod_ofr**”, “**Prod_comb**”, “**Prod_cog**”, las variables correspondientes a las **velocidades medias del viento**, asi como a la variable “**Der_CO2**”, que tendrán valores anormales y necesitarán de algun tratamiento para ciertos modelos.

Como el objetivo del estudio es predecir el precio de la EE, se realizará a continuación el análisis de cada variable en relación a la variable respuesta precio. Analizando los datos de esta forma, se pueden extraer ideas sobre qué variables están más relacionadas con el precio y de qué forma.



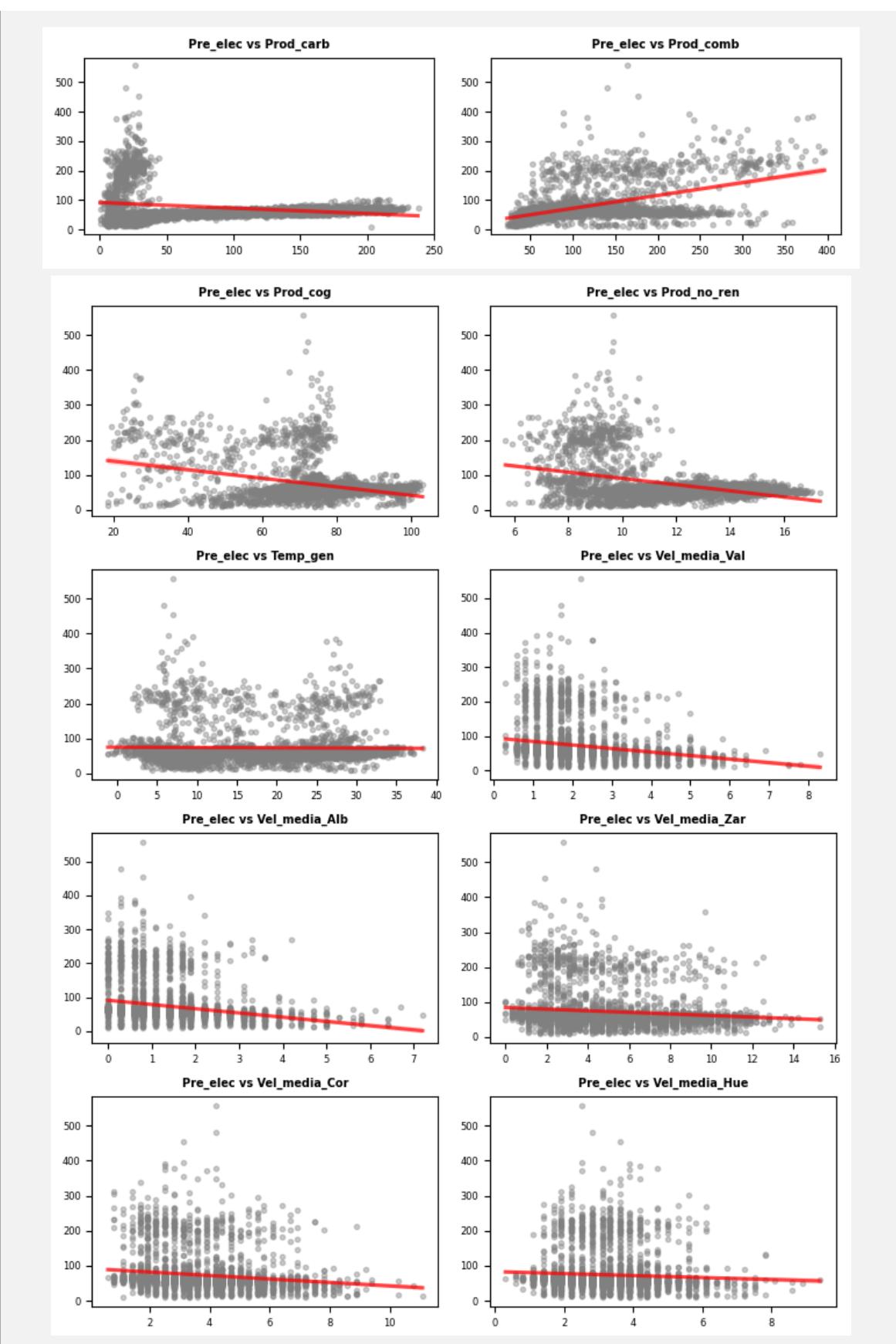
MODELOS PARA LA PREDICCIÓN DEL PRECIO DE LA ELECTRICIDAD

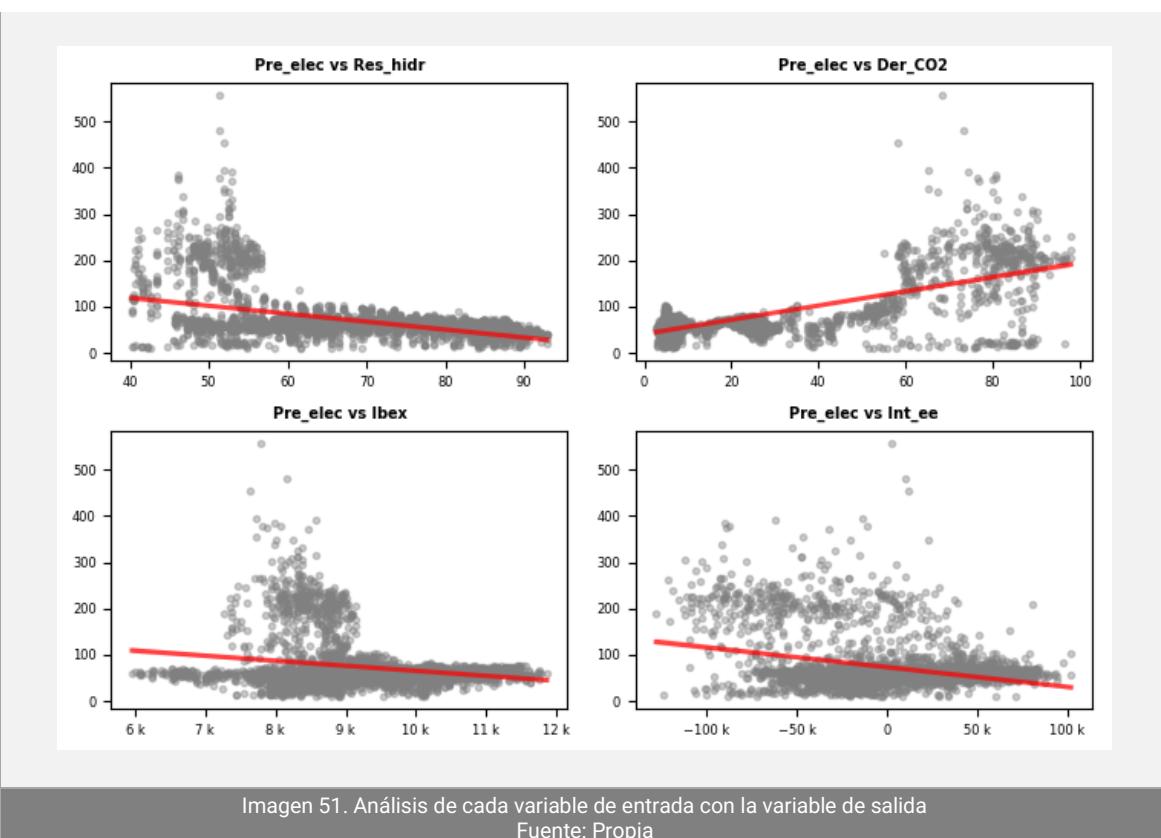
EN EL MERCADO ESPAÑOL



MODELOS PARA LA PREDICCIÓN DEL PRECIO DE LA ELECTRICIDAD

EN EL MERCADO ESPAÑOL





7. METODOS Y TECNICAS EMPLEADAS

Como se ha señalado en capítulos anteriores, el objetivo de este estudio es el de analizar distintos **modelos de regresión** para comprobar cual es el mejor a la hora de predecir el precio de la EE en España.

Para esto se partirá de una serie de variables, 26 en total, pertenecientes a diferentes grupos de factores que influyen en la que será la variable objetivo “**Pre_elec**”.

Existen muchas técnicas de regresión en función del tipo de variables y de la forma interactuar entre ellas. Los modelos estudiados en este trabajo son:

- Modelo de regresión lineal múltiple.
- Modelo k-vecinos más cercanos (KNN).
- Modelo árbol de decisión.
- Modelo Random Forest.

- Modelo XG Boost.
- Modelo red neuronal.

Antes de meternos en el estudio de los modelos, explicaremos las acciones que se deben realizar anteriormente sobre los datos, bien para que se puedan desarrollar con normalidad los modelos, o bien para que los resultados obtenidos podamos emplearlos para obtener una buena comparativa. Estas acciones son:

- Dividir el conjunto de datos en **grupo de entrenamiento y grupo de prueba**, para que todos los modelos usen las mismas divisiones de datos.
- Para los modelos de Regresión Lineal, K-vecinos más cercanos y las Redes Neuronales se debe realizar una **estandarización de los datos**, para que todos tengan la misma importancia.
- Obtener las mismas **métricas de error y precisión**, para que los modelos se puedan comparar.

Todo el estudio de modelado de métodos se ha realizado con el lenguaje de programación **Python**. De esta forma complementamos el TFM, empleando los dos lenguajes de programación que el Máster ha dedicado especial atención.

7.1 DIVISIÓN DE LOS DATOS PARA LOS MODELOS

Para evaluar la capacidad predictiva de un modelo debemos comprobar lo próximas que están sus predicciones a los verdaderos valores de la variable respuesta. Para poder cuantificar esto de forma correcta, es necesario disponer de un conjunto de observaciones que no hayan sido empleadas anteriormente en el entrenamiento del modelo.

Para poder lograr esto se dividen los datos del dataset en dos conjuntos, uno llamado de “entrenamiento” o “**train**” y otro de “prueba” o “**test**”.

En este estudio se ha recurrido a realizar dos divisiones diferentes, para estudiar más factores que pueden afectar a los precios. Según esto se han establecido las dos siguientes divisiones:

- 80% para el conjunto “train” y el 20% para el de “test”.
- 70% para el conjunto “train” y el 30% para el de “test”.

Para asegurarse que la distribución de las variables es similar en los dos conjuntos se ha empleado el método `train_test_split()` de **scikit-learn**. Este reparto estratificado asegura que el conjunto de entrenamiento y de test son similares en cuanto a la variable respuesta, aunque, no garantiza que sea así con las variables de predicción.

7.2 ESTANDARIZACIÓN DE LOS DATOS

Cuando las **variables predictoras** son numéricas, la escala en la que se miden puede influir en el algoritmo. Muchos algoritmos de Machine Learning, como son la **regresión lineal**, **k-vecinos más cercanos** o las **redes neuronales**, se ven afectado por esta circunstancia, de forma que, si no se ajustan de alguna forma los datos de estas variables, aquellas que posean una escala mayor, influirán más en el modelo, aunque no tienen por qué ser las que más relación tengan con la variable respuesta.

Para evitar esto existen dos técnicas:

- **Centrado:** consiste en restarle a cada valor la media del predictor al que pertenece. Como resultado de esta transformación, todos los predictores pasan a tener una media de cero, es decir, los valores se centran en torno al origen.
- **Normalización (estandarización):** consiste en transformar los datos de forma que todos los predictores estén aproximadamente en la misma escala. Hay dos formas de lograrlo:
 - **Normalización Z-score:** divide cada variable predictor entre su desviación típica después de haber sido centrada, de esta forma, los datos pasan a tener una distribución normal.
 - **Estandarización max-min:** transforma los datos de forma que estén dentro del rango 0-1.

En este estudio se ha optado por normalizar, en los modelos que lo necesitan, empleando el método `stats.z-score ()` de la biblioteca **scipy**.

7.3 METRICAS DE VALIDACION DE LOS MODELOS

La **evaluación de los modelos** es muy importante para el desarrollo de los algoritmos de aprendizaje automático. La evaluación ayuda a medir el **rendimiento del modelo**, es decir, **cuantificar la calidad de las predicciones que efectúa**.

Para realizar esta cuantificación empleamos las llamadas **métricas de validación**, y que según del tipo de algoritmo que utilicemos deberemos optar por unas o por otras.

En los modelos de regresión es casi imposible predecir el valor exacto, por lo que lo que hay que buscar es estar lo más cerca posible del valor real, y es por eso que la mayoría de las métricas se centran en medir eso: lo cerca (o lejos) que están las predicciones de los valores reales.

Las métricas de evaluación más comunes para los modelos de regresión son la siguientes:

- **Error medio absoluto (MAE)**: que corresponde a la media de las diferencias absolutas entre el valor real y el predicho. Mide el promedio de los residuos. No penaliza los errores grandes al no elevarse al cuadrado, lo que la hace no muy sensible a valores anómalos. Por este motivo no es una métrica recomendable en modelos en los que se deba prestar atención a los outliers. Representa el error en la misma escala que los valores reales y lo deseable es que su valor sea cercano a cero.
- **Error cuadrático medio (MSE)**: es una de las métricas más utilizadas en regresión. Corresponde a la media de las diferencias entre el valor real y el predicho al cuadrado. Mide la varianza de los residuos. Al elevar al cuadrado los errores, magnifica los errores grandes, por lo que hay que utilizarlo con cuidado cuando tenemos valores extremos. Puede tomar valores entre 0 e infinito. Cuanto más cerca de cero esté la métrica, mejor.
- **Raíz cuadrada del error cuadrático medio (RMSE)**: corresponde a la raíz cuadrada de la métrica anterior y mide la desviación estándar de los residuos. La ventaja de esta métrica es que presenta el error en las mismas unidades que la variable objetivo, lo que la hace más fácil de entender.
- **R cuadrado o coeficiente de determinación (R^2)**: esta métrica determina la calidad del modelo para replicar los resultados, y la proporción de los resultados que puede

explicarse por el modelo. Los valores que puede tomar esta métrica van desde menos infinito a 1. Cuanto más cercano a 1 sea el valor de esta métrica, mejor será el modelo.

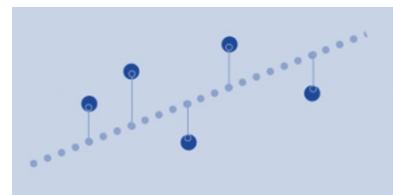
- **Error en porcentaje medio absoluto (MAPE):** es una medida relativa que escala esencialmente el MAE para que se muestre en unidades de porcentaje en lugar de en unidades de la variable.

Aunque las métricas anteriores son más comunes a la hora de comparar modelos, como se ha tomado como base para valorar la calidad de nuestros algoritmos los datos contenidos en el artículo de investigación realizado por González C., Mira-McWilliams J. y Juárez I. (2015), debemos obtener este error que es el de referencia en dicho artículo.

En todos los modelos de este estudio se ha obtenido estas cinco métricas, para poder compararlos y poder conocer cuales son los mejores.

7.4 MODELO DE REGRESIÓN LINEAL MÚLTIPLE

Los modelos de **regresión lineal múltiple** son **algoritmos de aprendizaje supervisado**. Son de los modelos más sencillos que existen, y tratan de ajustar modelos lineales, o linealizables, entre una variable dependiente y más de una variable independiente, utilizando una media de error que minimiza en un proceso iterativo.



El modelo es lineal porque consiste en que cada variable es un predictor, que se multiplica por un coeficiente estimado, para modelar la ecuación de una recta, la cual podrá estimar un valor, que será el valor a predecir.

Para estimar los parámetros de la recta el algoritmo emplea el método de los mínimos cuadrados ordinarios

Para trabajar con este tipo de algoritmos ha de tenerse en cuenta una serie de consideraciones como:

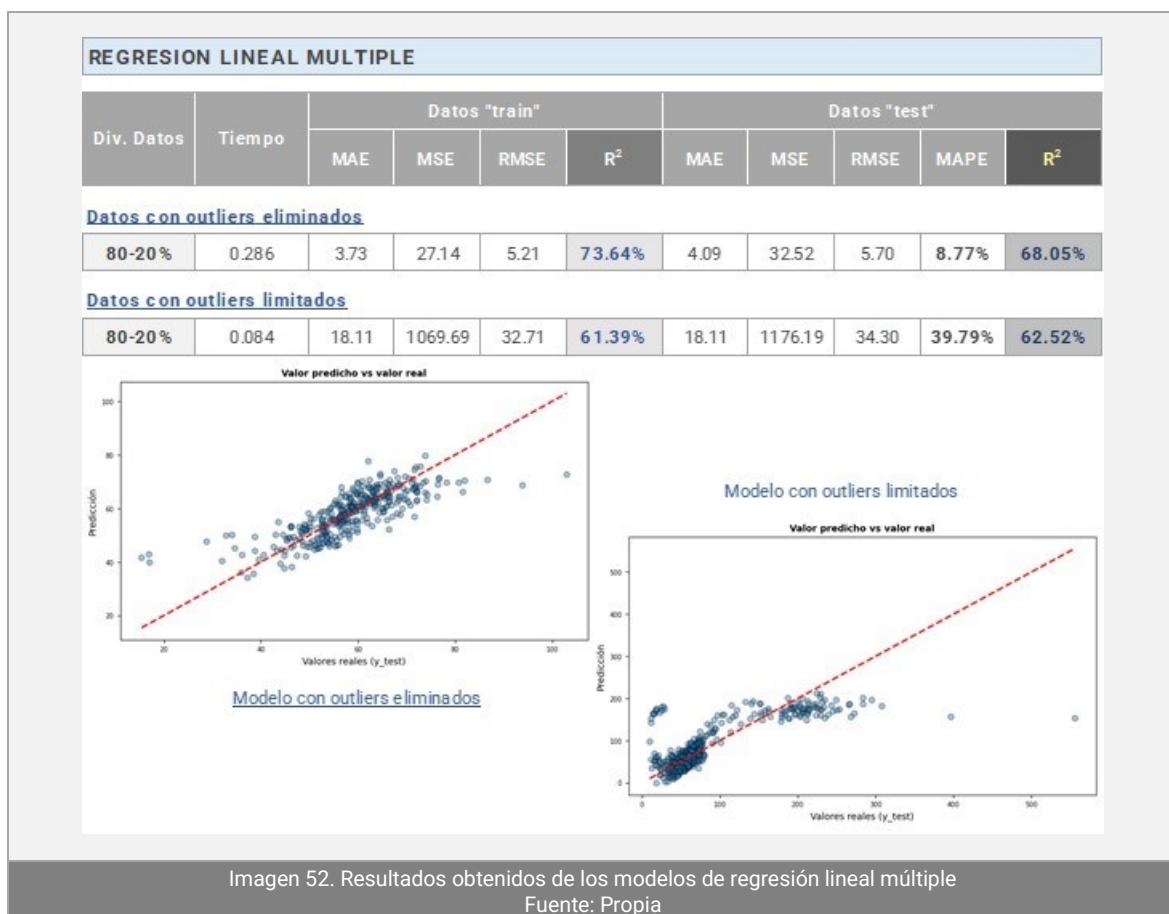
- Debe existir relación lineal entre las variables de entrada y la de salida. Esta relación se ha estudiado en el apartado **6.6.2 Variables de entrada**.

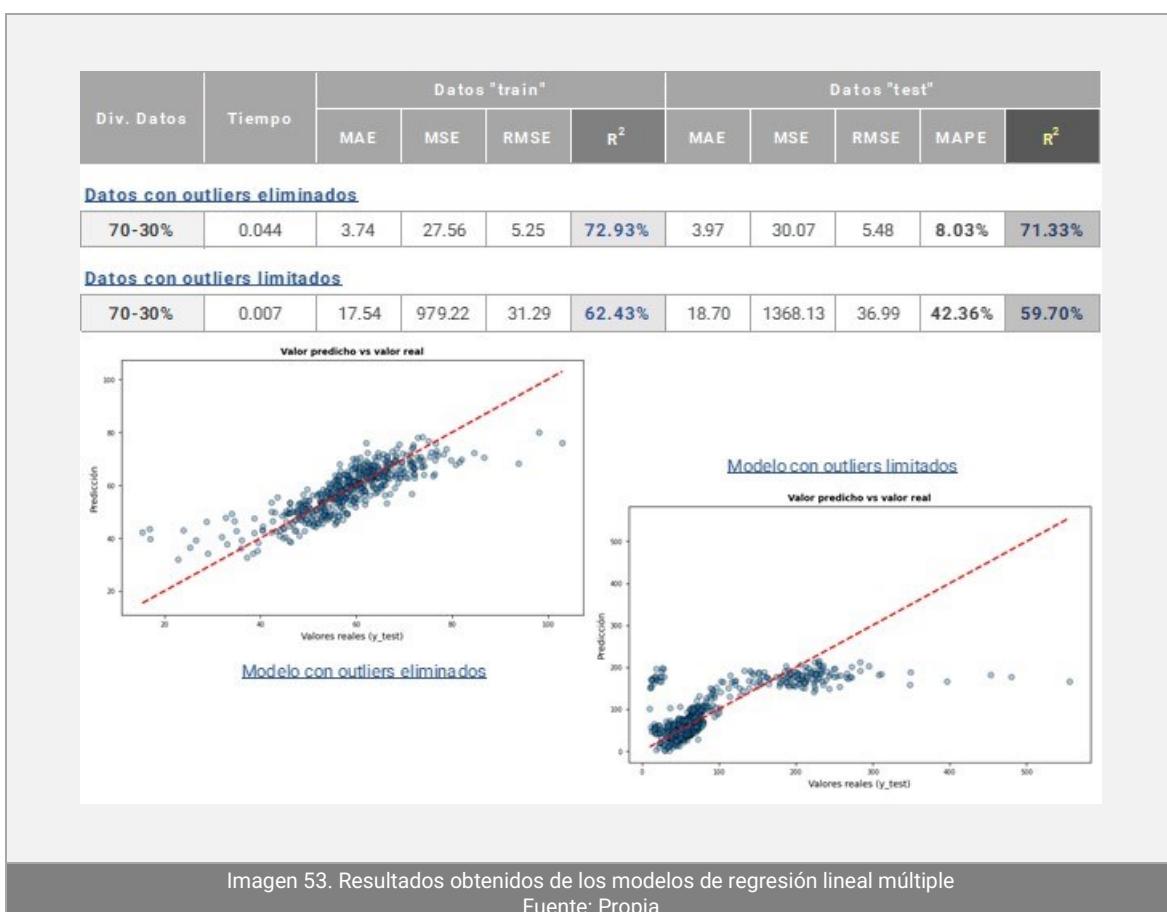
- Los residuos u errores deben seguir una distribución normal y para cumplir esto es necesario que las variables predictoras sean normalizadas. Como se ha explicado en el apartado **7.2 Estandarización de los datos**, se ha empleado el método de Normalización Z-score para realizarlo.
- Las variables predictoras deben ser linealmente independientes, cosa que se ha estudiado en el capítulo **6.7.3 Correlación entre variables**, en donde se unieron una serie de variables ya que presentaban dependencia entre ellas.
- Son modelos muy sensibles a los valores atípicos u outliers, debido a esto han sido eliminados o limitados a un valor máximo. En el capítulo **6.7.5 Tratamiento de los valores outliers** se expone la forma de tratarlos y como han sido utilizarlos en los algoritmos de regresión lineal múltiple.

Remitimos a los **Anexos VII y VIII** para la consulta del código desarrollado para estos modelos.

A continuación, mostramos el resumen de los resultados obtenidos con este algoritmo diferenciados según el porcentaje de división de los datos.

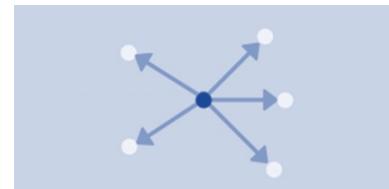
MODELOS PARA LA PREDICCIÓN DEL PRECIO DE LA ELECTRICIDAD
EN EL MERCADO ESPAÑOL





7.5 MODELO K VECINOS MÁS CERCANOS (KNN)

Los modelos **k vecinos más cercanos** son algoritmos de **aprendizaje supervisado** no paramétrico basado en instancias, que utiliza la proximidad para hacer predicciones o clasificaciones sobre la agrupación de un punto de datos individual, partiendo de la suposición de que se pueden encontrar puntos similares cerca uno del otro. Forman parte de una familia de modelos de "aprendizaje perezoso", lo que significa que solo almacena un conjunto de datos de entrenamiento en lugar de pasar por una etapa de entrenamiento.



Estos modelos asignan un valor sobre la base de una distancia, es decir, se utiliza el promedio de los k vecinos más cercanos para hacer una predicción sobre punto de datos determinado. La distancia euclídea es la más utilizada.

Para trabajar con este tipo de algoritmos ha de tenerse en cuenta una serie de consideraciones como:

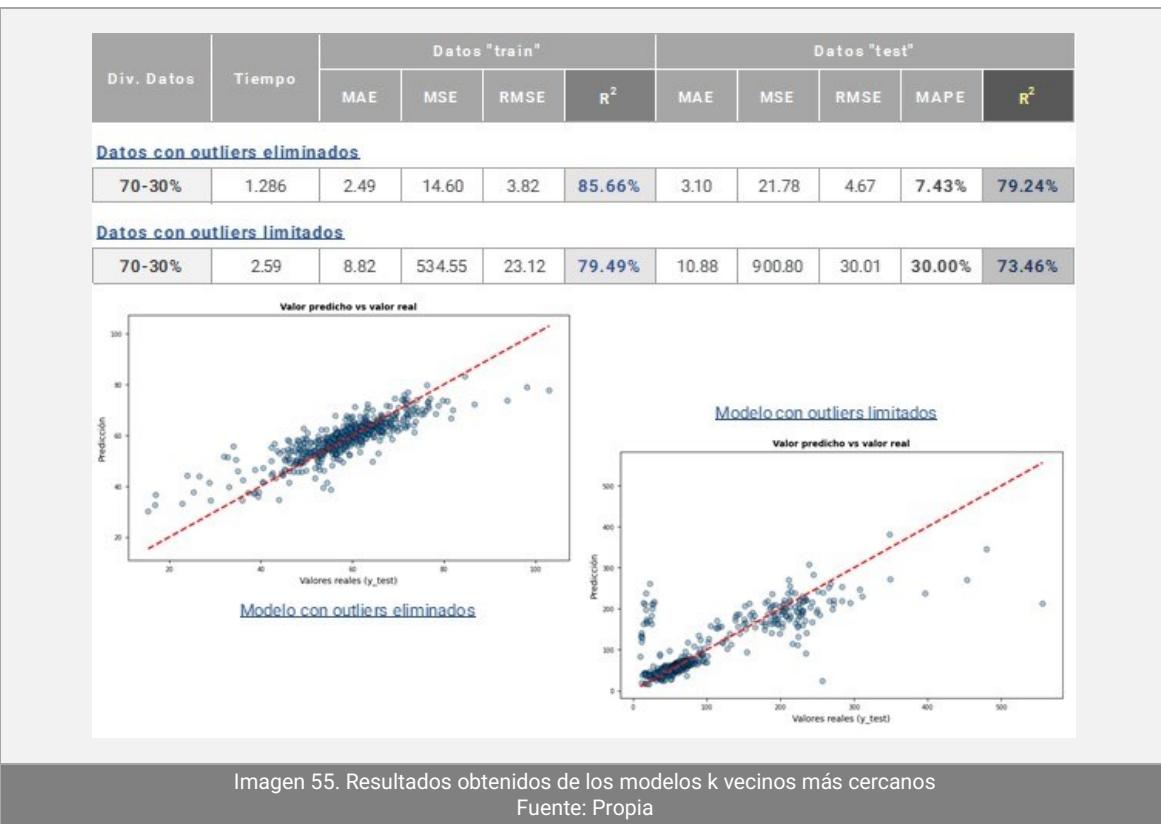
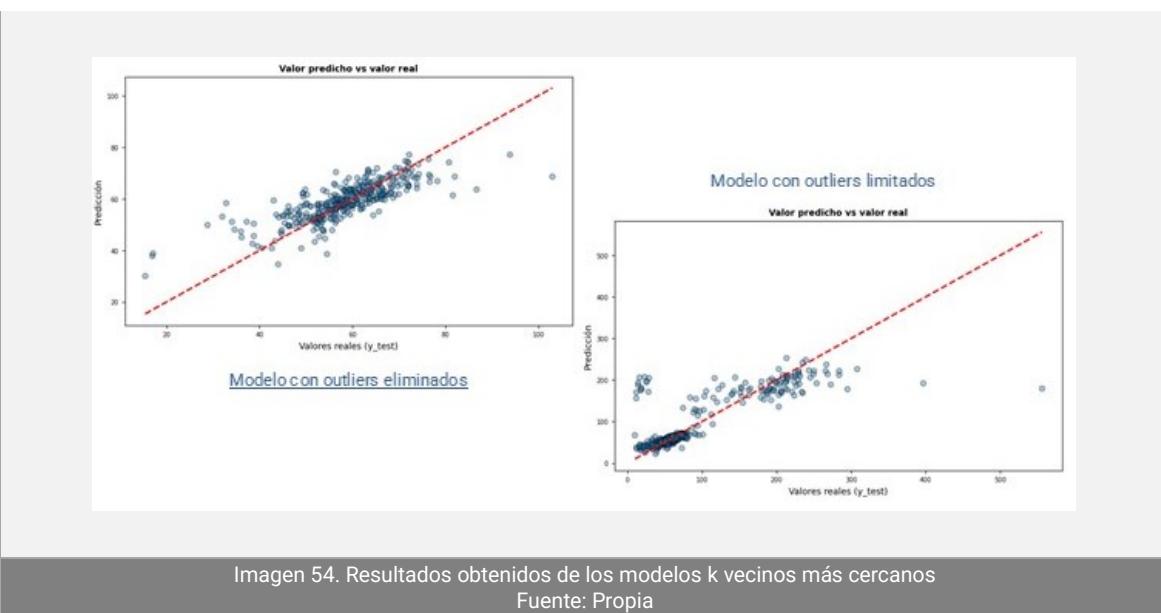
- La selección del número de vecinos se realiza mediante validación cruzada dentro del desarrollo del algoritmo.
- La distancia empleada en el estudio ha sido la distancia euclídea que es la más utilizada en este tipo de modelos.
- Los datos empleados deben ser homogéneos y limpios, por lo que se ha realizado un tratamiento de los datos para lograr estas características, expuestos a lo largo del capítulo **6.7 Tratamiento de los datos y formación del dataset**.
- Los datos utilizados no deben contener muchos valores anómalos, por lo que se han tratado según lo expuesto en el capítulo **6.7.5 Tratamiento de los valores outliers**.
- No es aconsejable emplear este método cuando el conjunto de datos es pequeño, cosa que no sucede con nuestro dataset.

Remitimos a los **Anexos IX y X** para la consulta del código desarrollado para estos modelos.

A continuación, mostramos el resumen de los resultados obtenidos con este algoritmo diferenciados según el porcentaje de división de los datos.

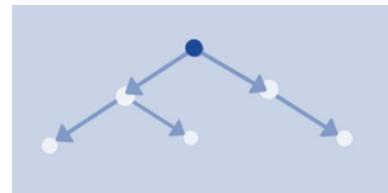
K-VECINOS MAS CERCANOS										
Div. Datos	Tiempo	Datos "train"				Datos "test"				
		MAE	MSE	RMSE	R ²	MAE	MSE	RMSE	MAPE	R ²
Datos con outliers eliminados										
80-20 %	1.312	2.43	13.41	3.66	86.97%	3.18	22.57	4.75	8.68%	77.82%
Datos con outliers limitados										
80-20 %	3.71	9.70	647.35	25.44	76.63%	10.21	773.79	27.82	30.67%	75.35%

MODELOS PARA LA PREDICCIÓN DEL PRECIO DE LA ELECTRICIDAD EN EL MERCADO ESPAÑOL



7.6 MODELO ARBOL DE DECISION

Los modelos de **árbol de decisión** son **algoritmos de aprendizaje supervisado** no paramétrico basado en una estructura de árbol jerárquica que divide el espacio de los predictores agrupando observaciones con valores similares para la variable respuesta o dependiente.



La estructura resultante, cuando se visualiza tiene forma de un árbol, que consta de un nodo raíz, ramas, nodos internos y nodos hoja. Este tipo de estructura de diagrama de flujo puede crear una representación fácil de comprender para emplearla en la toma de decisiones

En el caso de los árboles aplicados a la regresión, el criterio estándar para decidir los cortes es el error cuadrático medio.

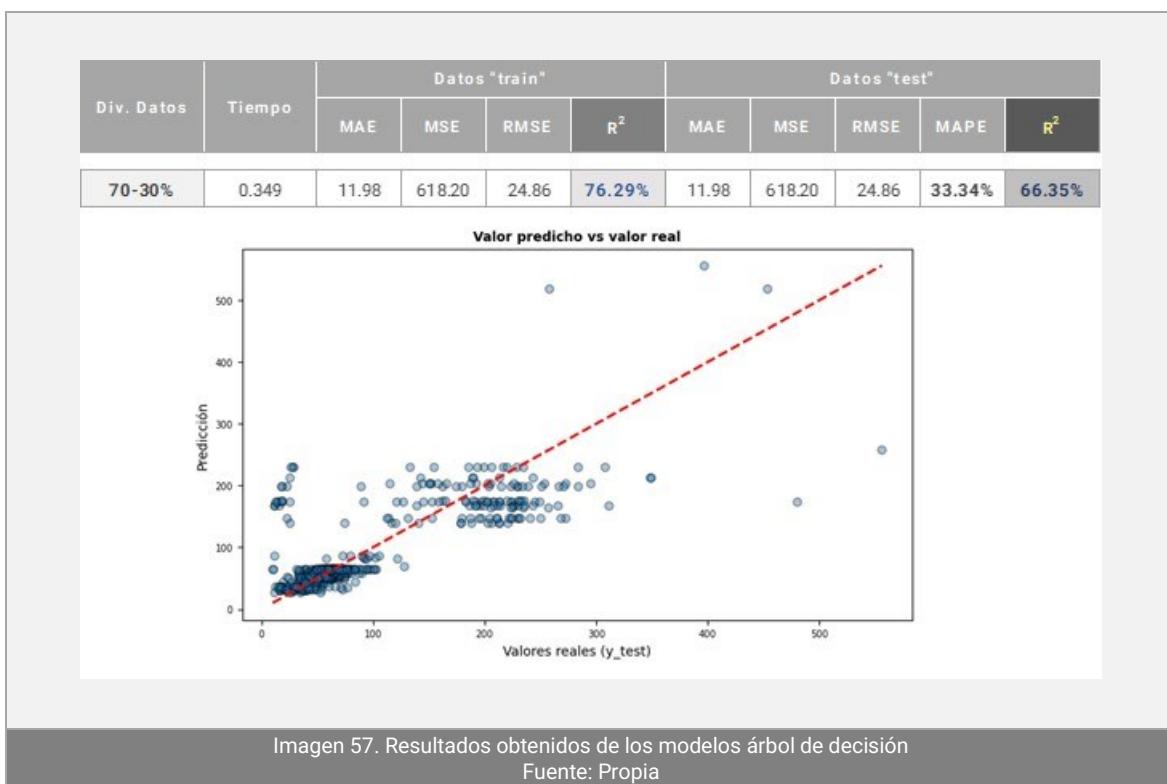
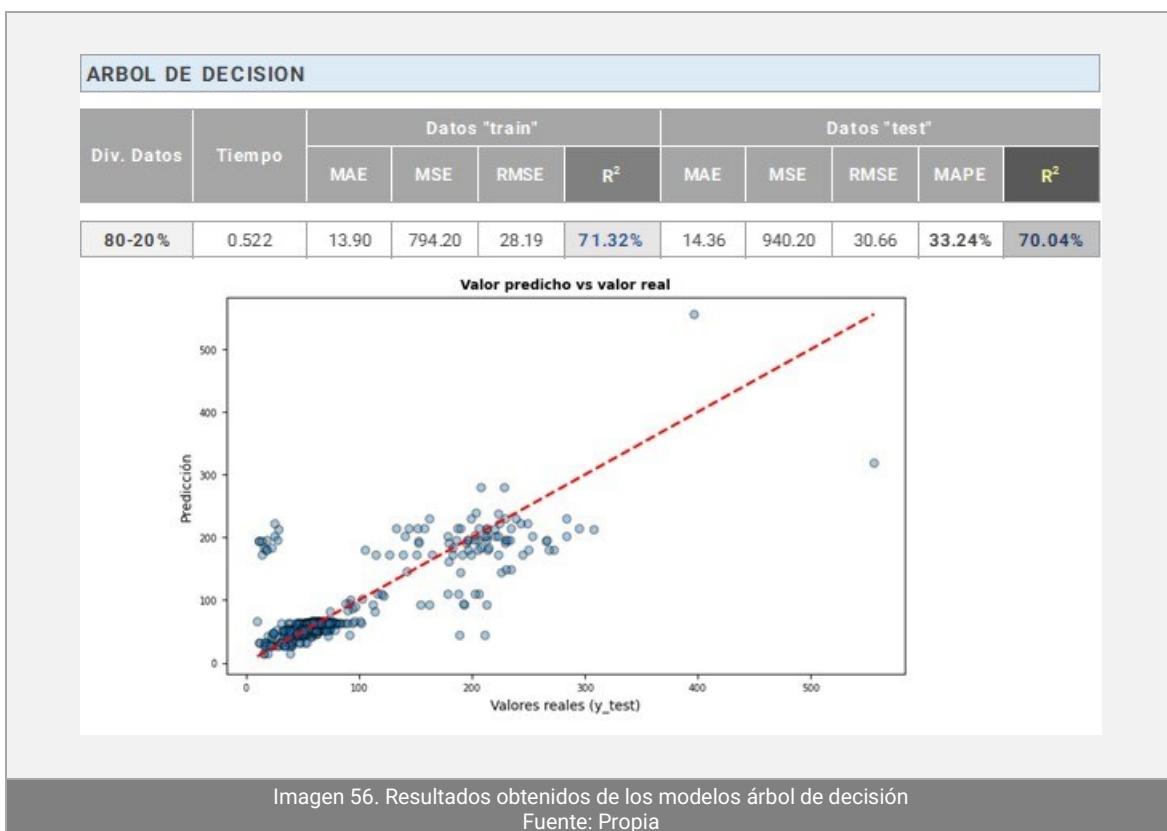
Ha de tenerse en cuenta una serie de consideraciones para emplear este tipo de algoritmos:

- Para una buena selección de los parámetros del algoritmo se ha empleado validación cruzada.
- No son modelos sensibles a los valores atípicos u outliers, por lo que el dataset empleado en estos modelos no ha sufrido ninguna trasformación en este sentido y se han mantenido los valores anormales de las variables.
- Si el modelo forma una estructura muy grande y compleja, no generalizará bien los nuevos datos y producirá sobreajustes.

Remitimos al **Anexo XI** para la consulta del código desarrollado para estos modelos.

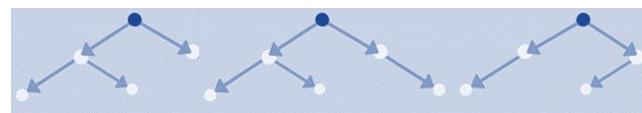
A continuación, mostramos el resumen de los resultados obtenidos con este algoritmo diferenciados según el porcentaje de división de los datos.

MODELOS PARA LA PREDICCIÓN DEL PRECIO DE LA ELECTRICIDAD
EN EL MERCADO ESPAÑOL



7.7 MODELO RANDOM FOREST

El algoritmo **Random Forest** es una técnica de **aprendizaje supervisado** que genera múltiples árboles de decisión



que se combinan a fin de obtener un modelo único más robusto, en comparación con los resultados de cada árbol por separado, cuyos resultados se combinan en una sola salida conjunta final, obtenida, para los casos de regresión, generalmente mediante el promedio de dichas salidas.

Cada árbol generado por el algoritmo contiene un grupo de observaciones aleatorias del dataset de datos, empleando las no utilizadas para validar el modelo.

Ha de tenerse en cuenta una serie de consideraciones para emplear este tipo de algoritmos:

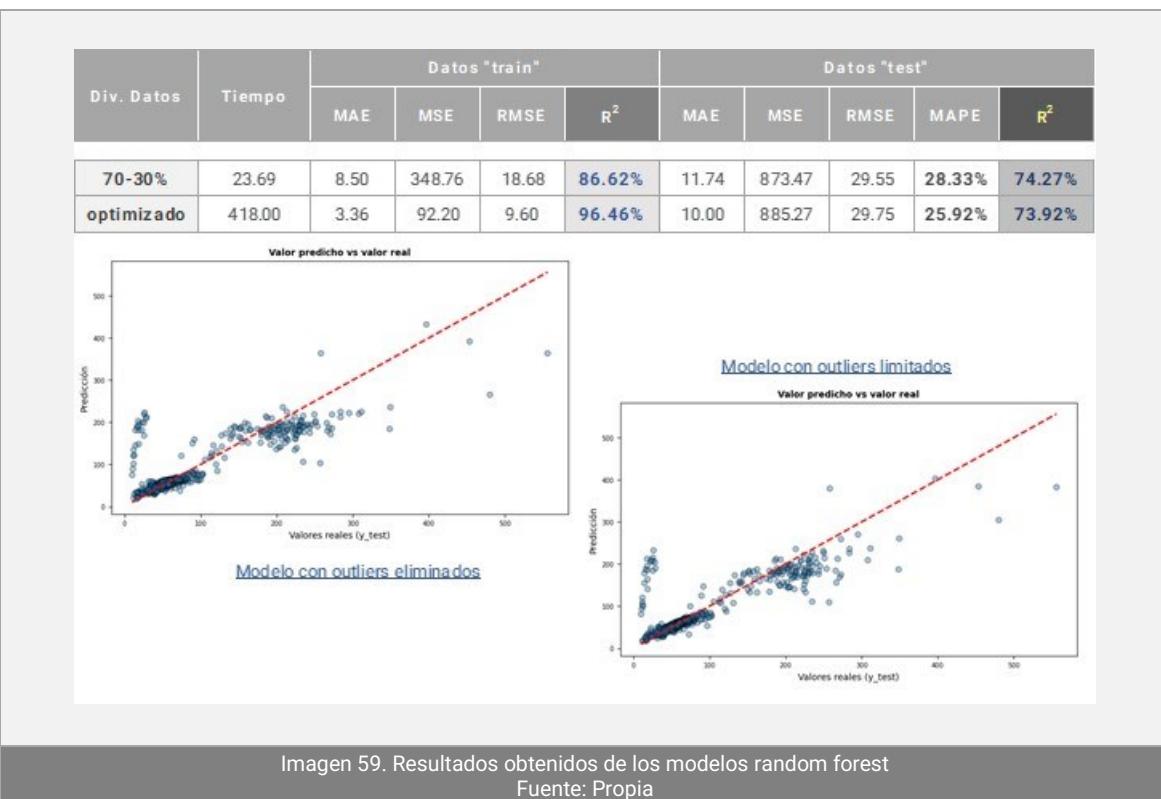
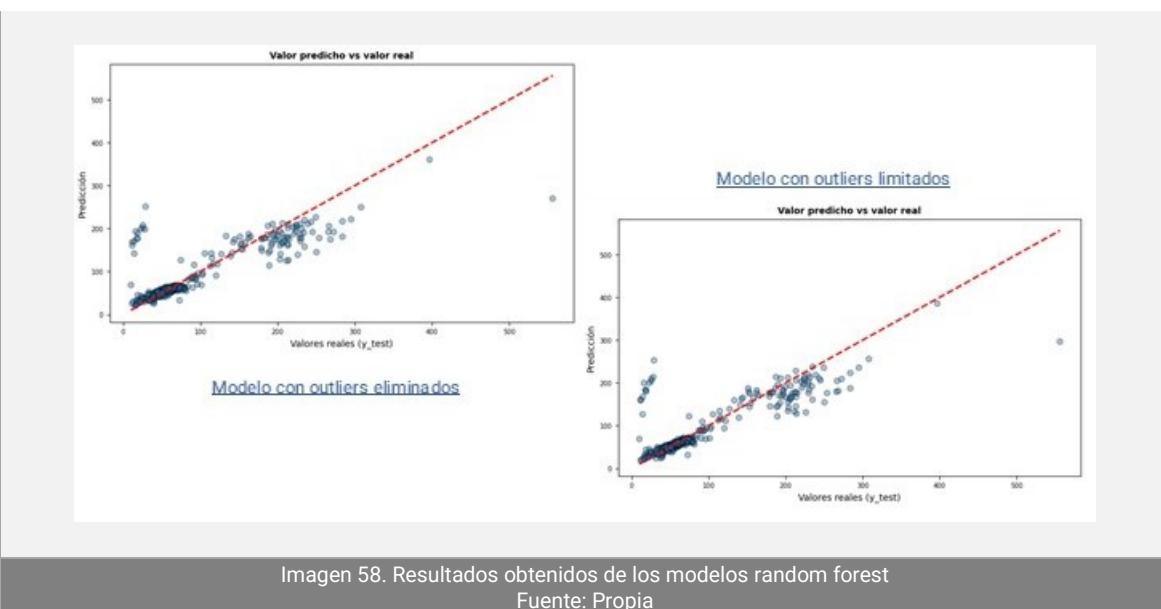
- Para una buena selección de los parámetros del algoritmo se ha empleado validación cruzada.
- No son modelos sensibles a los valores atípicos u outliers, por lo que el dataset empleado en estos modelos no ha sufrido ninguna transformación en este sentido y se han mantenido los valores anormales de las variables.
- Es muy común que se formen grandes estructuras complejas que no generalizan bien los nuevos datos y produciendo sobreajustes.

Remitimos al **Anexo XII** para la consulta del código desarrollado para estos modelos.

A continuación, mostramos el resumen de los resultados obtenidos con este algoritmo diferenciados según el porcentaje de división de los datos.

RANDOM FOREST											
Div. Datos	Tiempo	Datos "train"				Datos "test"					
		MAE	MSE	RMSE	R ²	MAE	MSE	RMSE	MAPE	R ²	
80-20 %	24.59	8.55	352.87	18.78	87.26%	11.34	791.17	28.13	28.00%	74.79%	
optimizado	488.70	3.35	94.96	9.74	96.57%	9.78	809.44	28.45	26.20%	74.21%	

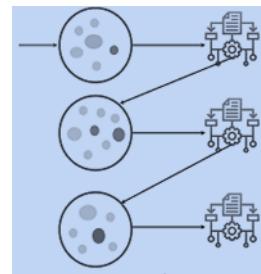
MODELOS PARA LA PREDICCIÓN DEL PRECIO DE LA ELECTRICIDAD EN EL MERCADO ESPAÑOL



7.8 MODELO XG BOOST

El modelo **XGBoost** es un método de **aprendizaje automático supervisado**, que se basa en árboles de decisión y supone una mejora sobre los métodos bosque aleatorio y refuerzo de gradientes.

Los **algoritmos basados en Boosting** se basan en la combinación de modelos simples, a los cuales se le llama **algoritmos ensamblados**, que son empleados en serie.



El rendimiento de un modelo simple puede ser mejorado, haciendo que otro modelo simple posterior, dé más importancia a los errores cometidos por el modelo previo. Es como si el segundo algoritmo al intentar resolver un problema se aprovechase del conocimiento de los errores del anterior.

Las predicciones de cada modelo de regresión simple se combinan, por medio de una suma ponderada, para producir la predicción final.

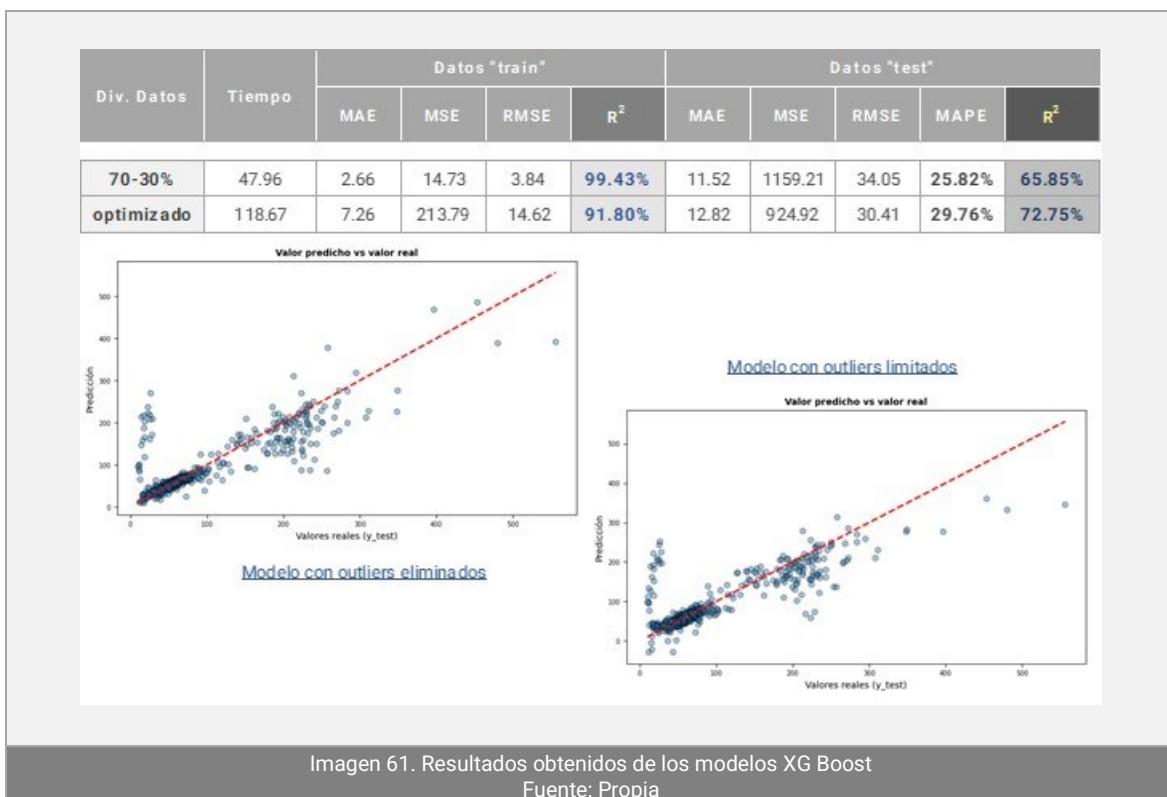
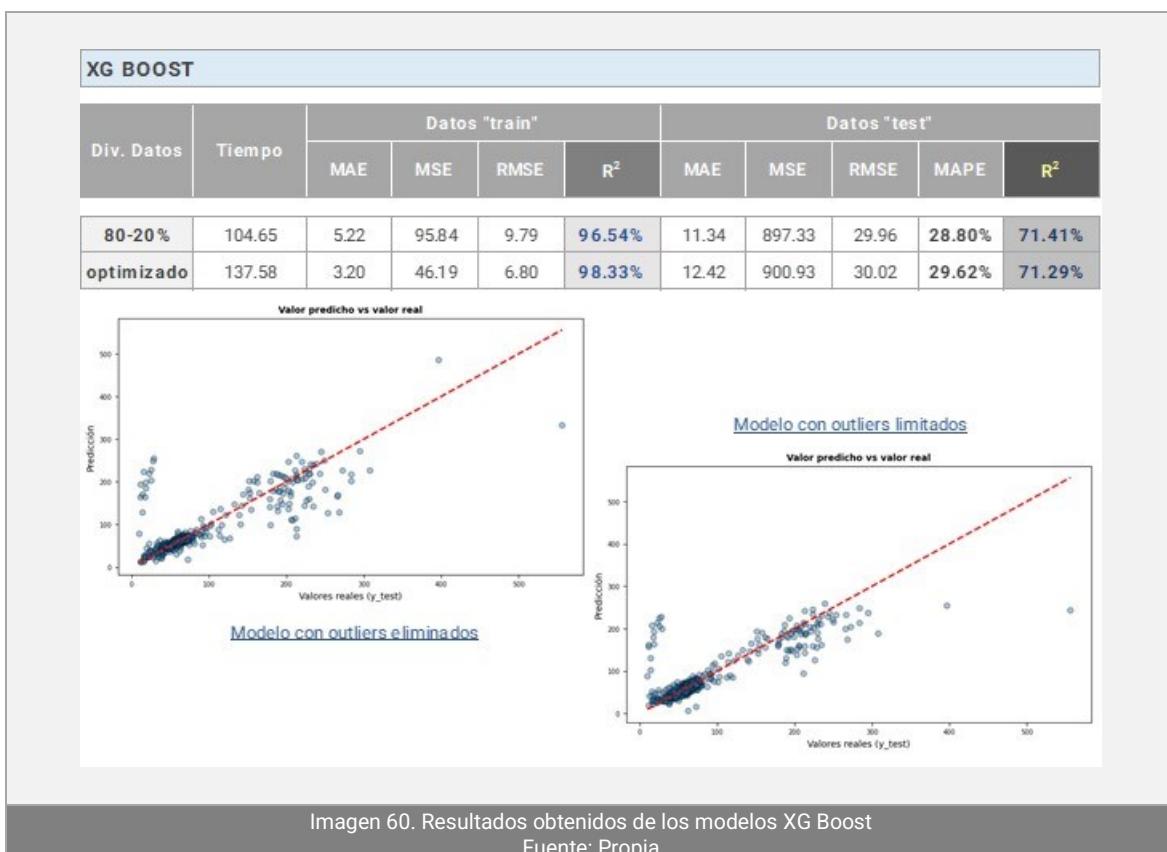
Ha de tenerse en cuenta una serie de consideraciones para emplear este tipo de algoritmos:

- Aunque este algoritmo posee un método que trata automáticamente los valores faltantes, en el capítulo **6.7.2 Tratamiento de los valores faltantes** se explica como se ha procedido en nuestro dataset de trabajo.
- Aunque se obtienen buenos resultados con los parámetros que el algoritmo tiene por defecto, se han buscado mejores resultados realizando una selección de parámetros empleando validación cruzada.
- Es un algoritmo que trata bien los valores atípicos u outliers, por lo que el dataset empleado en este modelo no ha sufrido ninguna trasformación en este sentido y se han mantenido los valores anormales de las variables.

Remitimos a los **Anexos XIII** para la consulta del código desarrollado para estos modelos.

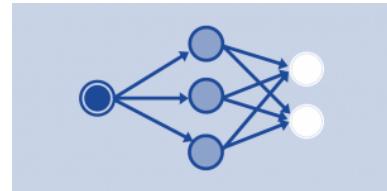
A continuación, mostramos el resumen de los resultados obtenidos con este algoritmo diferenciados según el porcentaje de división de los datos.

MODELOS PARA LA PREDICCIÓN DEL PRECIO DE LA ELECTRICIDAD
EN EL MERCADO ESPAÑOL



7.9 MODELO RED NEURONAL

Los **modelos de redes neuronales** son **algoritmos de Deep Learning de tipo no paramétrico**, y no asumen ningún tipo de distribución de la variable respuesta, por lo tanto, no es necesario que esta siga ninguna distribución concreta.



Estos modelos se crean ordenando operaciones matemáticas siguiendo una determinada estructura. Esta estructura estará formada por **capas** consecutivas, que contienen las **neuronas**. Cada neurona, realiza una operación sencilla y está conectada a las neuronas de las capas anterior y posterior mediante **pesos**, cuya función es regular la información que se propaga de una neurona a otra.

En el aprendizaje supervisado se presenta a la red neuronal un conjunto de patrones, junto con la salida deseada u objetivo, e iterativamente la red ajusta los pesos hasta que la salida tiende a ser la deseada, utilizando para ello información detallada del error que se comete en cada paso. De este modo, la red es capaz de estimar relaciones entrada/salida sin necesidad de proponer una cierta forma funcional de partida.

Ha de tenerse en cuenta una serie de consideraciones para emplear este tipo de algoritmos:

- Es un algoritmo que necesita que las variables predictoras sean normalizadas ya que la escala en la que se miden puede afectar a los resultados, influyendo en mayor medida aquellas variables que posean escalas mayores, aunque puede que no sean las que mayor relación tengan con la variable de respuesta. Esta normalización se ha realizado según se ha expuesto en el apartado **7.2 Estandarización de los datos**.
- Buscando la mejor respuesta para nuestro estudio para este tipo de algoritmos, se han definido diferentes estructuras de la red neuronal combinando múltiples capas ocultas y con diferentes números de neuronas en estas capas.
- No es un algoritmo al que le afecten negativamente los valores atípicos u outliers, por lo que el dataset empleado en este modelo no ha sufrido ninguna transformación en este sentido y se han mantenido los valores anormales de las variables.

Remitimos a los **Anexos XIV y XV** para la consulta del código desarrollado para estos modelos.

MODELOS PARA LA PREDICCIÓN DEL PRECIO DE LA ELECTRICIDAD

EN EL MERCADO ESPAÑOL

A continuación, mostramos el resumen de los resultados obtenidos con este algoritmo diferenciados según el porcentaje de división de los datos.

ESTRUCTURAS REDES NEURONALES ESTUDIADAS														
Nº capas	Nº neuronas / capa													Nº neur. total
	Entrada	1º C.O.	2º C.O.	3º C.O.	4º C.O.	5º C.O.	6º C.O.	7º C.O.	8º C.O.	9º C.O.	10º C.O.	...	27º C.O.	
3	26	5												1 32
	26	10												1 37
	26	25												1 52
	26	50												1 77
6	26	5	10	20	40									1 102
	26	40	20	10	5									1 102
	26	10	10	10	10									1 67
	26	20	20	20	20									1 107
	26	40	40	40	40									1 187
12	26	20	20	20	20	20	20	20	20	20	20			1 227
	26	25	50	25	50	25	50	25	50	25	50			1 402
	26	25	10	10	10	10	10	10	10	10	10			1 142
30	26	20	30	10	20	30	10	20	30	10	20	...	10	1 237

C.O.: capa oculta

Imagen 62. Estructuras de las redes neuronales

Fuente: Propia

RED NEURONAL												
Div. Datos	Nº capas	Nº neur. total	Tiempo	Datos "train"				Datos "test"				
				MAE	MSE	RMSE	R ²	MAE	MSE	RMSE	MAPE	R ²
80-20%	3	32	302.27	9.19	697.04	26.40	94.84%	9.62	736.51	27.14	26.52%	76.53%
		37	75.03	10.16	612.39	24.75	77.90%	11.68	765.27	27.66	26.33%	75.62%
		52	70.81	10.10	592.03	24.33	78.63%	11.80	791.64	28.14	25.86%	74.78%
		77	45.12	9.32	559.58	23.66	79.80%	11.15	778.53	27.90	25.55%	75.19%
	6	102	26.13	10.68	594.66	24.39	78.53%	12.48	849.82	29.15	27.87%	72.92%
		102	32.36	8.56	437.05	20.91	84.22%	11.66	828.58	28.79	27.31%	73.60%
		67	47.65	9.82	506.07	22.50	81.73%	11.93	781.40	27.95	28.11%	75.10%
		107	34.30	9.06	441.05	21.00	84.08%	11.61	813.93	28.53	25.92%	74.07%
		187	24.59	8.44	327.74	18.10	88.17%	13.01	992.06	31.50	27.31%	68.39%
	12	227	28.27	10.30	382.69	19.56	86.19%	15.20	1118.64	33.45	31.01%	64.36%
		402	23.83	8.06	247.95	15.75	91.05%	14.34	1138.28	33.74	27.16%	63.73%
		142	35.54	11.45	556.89	23.60	79.90%	13.29	884.25	29.74	32.23%	71.83%
	30	237	52.79	6.81	229.61	15.15	91.71%	11.87	824.83	28.72	26.38%	73.72%

MODELOS PARA LA PREDICCIÓN DEL PRECIO DE LA ELECTRICIDAD
EN EL MERCADO ESPAÑOL

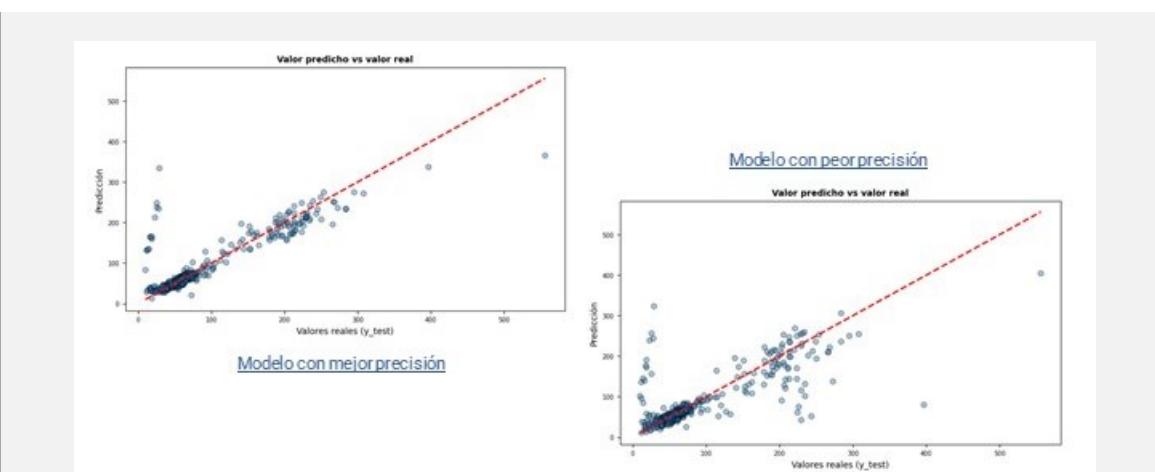


Imagen 63. Resultados obtenidos de los modelos de redes neuronales
Fuente: Propia

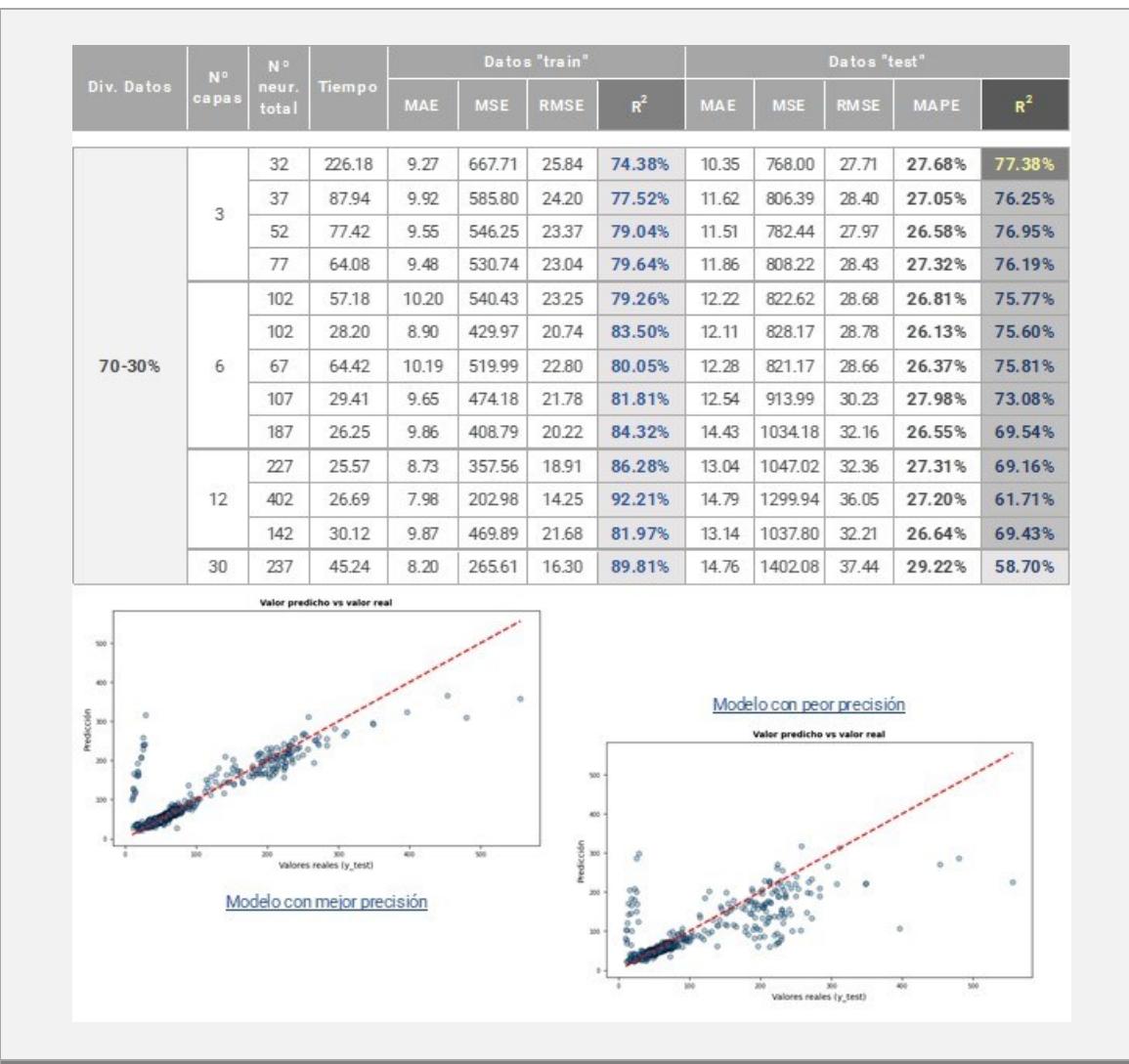


Imagen 64. Resultados obtenidos de los modelos de redes neuronales
Fuente: Propia

8. ANALISIS DE LOS RESULTADOS OBTENIDOS

En Este capitulo analizaremos los resultados obtenidos por cada tipo de algoritmo estudiado.

REGRESION LINEAL MULTIPLE

Para este tipo de algoritmo se han establecido **4 variantes**, que han sido obtenidas al emplear dos divisiones con diferentes proporciones para los datos de “train” y de “test” y como resultado de aplicar el dataset con los valores anormales eliminados o con ellos limitados. Los resultados obtenidos con el conjunto de datos “test” han sido los siguientes:

	Div. Datos	MAE	MSE	RMSE	MAPE	R ²	Tratamiento de outliers
RLM_2	70-30%	3.97	30.07	5.48	8.03%	71.33%	Eliminados
RLM_1	80-20%	4.09	32.52	5.70	8.77%	68.05%	Eliminados
RLM_3	80-20%	18.11	1176.19	34.30	39.79%	62.52%	Limitados
RLM_4	70-30%	18.70	1368.13	36.99	42.36%	59.70%	Limitados

Imagen 65. Resultados obtenidos de los modelos de regresión lineal múltiple
Fuente: Propia

Se comprueba que las dos primeras variantes presentan unos valores muy buenos de error, sea cual sea la métrica elegida, y unos valores aceptables de precisión para un algoritmo de predicción de precios.

En los gráficos que relacionan los valores reales con los valores de predicción, incluidos en el capitulo anterior, se observa que las dos modalidades en que los outliers han sido eliminados, los puntos se acumulan entre 20 y 100, mientras que, en los otros dos casos, las posiciones se diseminan entre los 20 y los 300 puntos. Esto quiere decir que aun siendo limitados los valores extremos a un valor máximo, todavía están influyendo negativamente en el logaritmo. Esto también puede comprobarse en la raíz del error cuadrático medio (**RMSE**) que se incrementa notablemente.

Se puede concluir que los modelos de regresión lineal múltiple pueden funcionar bien si se eliminan los valores anormales, y parece que la división del dataset para datos de “entrenamiento” y “test” no tienen mucha influencia.

K VECINOS MÁS CERCAOS

Este caso es igual que el anterior obteniéndose **4 variantes** para el modelo, derivadas por las mismas razones de división de datos y tratamiento de los outliers. Los resultados obtenidos con el conjunto de datos “test” han sido los siguientes:

	Div. Datos	MAE	MSE	RMSE	MAPE	R ²	Tratamiento de outliers
KNN_2	70-30%	3.10	21.78	4.67	7.43%	79.24%	Eliminados
KNN_1	80-20%	3.18	22.57	4.75	8.68%	77.82%	Eliminados
KNN_3	80-20%	10.21	773.79	27.82	30.67%	75.35%	Limitados
KNN_4	70-30%	10.88	900.80	30.01	30.00%	73.46%	Limitados

Imagen 66. Resultados obtenidos de los modelos k vecinos más cercanos
Fuente: Propia

Sucede lo mismo que en el caso anterior, en que **las dos primeras variantes presentan unos valores muy buenos de error**, sea cual sea la métrica elegida, y unos **valores mejores de precisión**, alcanzando valores muy buenos para la predicción de precios.

En cuanto los gráficos incluidos en el capítulo anterior correspondientes a la relación entre los valores reales y los predichos, se ha obtenido la misma situación que en los modelos de regresión lineal múltiple, **los valores anormales que han sido limitados, siguen influyendo negativamente en el logaritmo**.

Teniendo todo esto en cuenta, **se puede concluir que los modelos k vecinos más cercanos tienen una buena precisión para predecir valores y funcionan mucho mejor eliminando los valores anormales**, pero no se puede confirmar que las proporciones en la división del dataset para datos de “entrenamiento” y “test” tengan influencia.

ARBOLES DE DECISION

En este tipo de algoritmos los outliers no tienen influencia por lo que solo tendremos 2 variantes dependiendo de la proporción en la división de los datos.

	Div. Datos	MAE	MSE	RMSE	MAPE	R ²
AD_1	80-20%	14.36	940.20	30.66	33.24%	70.04%
AD_2	70-30%	11.98	618.20	24.86	33.34%	66.35%

Imagen 67. Resultados obtenidos de los modelos árbol de decisión

Fuente: Propia

El modelo que mayor precisión ha alcanzado ha sido en el que la **división corresponde a un 80-20%** alcanzando una **precisión** del **70.04%** y un **RMSE** de **30.66**. Aunque la raíz cuadrada del error cuadrático medio es superior que los mejores modelos analizados anteriormente, es una cifra aceptable y la precisión es buena.

RANDOM FOREST

Como en todos los modelos, tenemos 2 variantes derivadas al aplicar diferentes proporciones en la división de los datos para formar los conjuntos de “entrenamiento” y de “test” y a parte de esto, se ha intentado optimizar los algoritmos a través de validación cruzada, por lo que finalmente tenemos **4 variantes** para este modelo.

	Div. Datos	MAE	MSE	RMSE	MAPE	R ²
RF_1	80-20%	11.34	791.17	28.13	28.00%	74.79%
RF_2	70-30%	11.74	873.47	29.55	28.33%	74.27%
RF_1 optimizado	80-20%	9.78	809.44	28.45	26.20%	74.21%
RF_2 optimizado	70-30%	10.00	885.27	29.75	25.92%	73.92%

Imagen 68. Resultados obtenidos de los modelos de regresión random forest

Fuente: Propia

Los cuatro modelos creados han mostrado una gran igualdad, habiendo unas diferencias ínfimas en la precisión y para la RMSE. **Los valores obtenidos por estos cuatro modelos son aceptables en un caso de predicción de precios.**

Se puede comprobar en los gráficos que, aun teniendo valores anormales, ya que en este modelo no se han eliminado ni limitado, funciona bien, estando los puntos muy próximos a la línea roja.

Se puede concluir que los modelos random forest trabajan bien en predicción, aun teniendo outliers en los datos. Pero no se puede llegar a una conclusión firme en cuanto a las proporciones en la división del dataset.

XG BOOST

Este caso es igual que el anterior obteniéndose **4 variantes** para el modelo, derivadas por las mismas razones de división de datos y optimización de los modelos. Los resultados obtenidos con el conjunto de datos “test” han sido los siguientes:

	Div. Datos	MAE	MSE	RMSE	MAPE	R ²
XGB_2 optimizado	70-30%	12.82	924.92	30.41	29.76%	72.75%
XGB_1	80-20%	11.34	897.33	29.96	28.80%	71.41%
XGB_1 optimizado	80-20%	12.42	900.93	30.02	29.62%	71.29%
XGB_2	70-30%	11.52	1159.21	34.05	25.82%	65.85%

Imagen 69. Resultados obtenidos de los modelos de XG Boost
Fuente: Propia

De los cuatro modelos creados los 3 mejores han mostrado una gran igualdad, habiendo una pequeña diferencia en la precisión y en la RMSE. El otro modelo creado se descuelga bastante de los buenos resultados obtenidos por el resto.

Los valores obtenidos por los tres mejores modelos son aceptables en un caso de predicción de precios.

Se puede comprobar en los gráficos que, aun teniendo valores anormales, ya que en este modelo tampoco se han eliminado ni limitado, funciona bien, estando los puntos muy próximos a la línea roja.

Por tanto, **se concluye que los modelos XG Boost trabajan bien en predicción, aun teniendo outliers en los datos.** Pero no se puede llegar a una conclusión firme en cuanto a las proporciones en la división del dataset.

RED NEURONAL

En el caso de las **redes neuronales** se ha optado por crear numerosas variantes, modificando tanto el número de capas ocultas dentro de los modelos, como el número de neuronas que forman en cada una de esas capas, para ver como funcionan diferentes arquitecturas. Todas estas variantes están duplicadas al haber empleado las dos proporciones diferentes, señaladas en los modelos anteriores, para dividir los datos de "entrenamiento" y "test".

Los resultados obtenidos con el conjunto de datos "test" han sido los siguientes:

	Div. Datos	Nº capas	Nº neuronas	MAE	MSE	RMSE	MAPE	R ²
RN_14	70-30%	3	32	10.35	768.00	27.71	27.68%	77.38%
RN_16	70-30%	3	52	11.51	782.44	27.97	26.58%	76.95%
RN_1	80-20%	3	32	9.62	736.51	27.14	26.52%	76.53%
RN_15	70-30%	3	37	11.62	806.39	28.40	27.05%	76.25%
RN_17	70-30%	3	77	11.86	808.22	28.43	27.32%	76.19%
RN_20	70-30%	6	67	12.28	821.17	28.66	26.37%	75.81%
RN_18	70-30%	6	102	12.22	822.62	28.68	26.81%	75.77%
RN_2	80-20%	3	37	11.68	765.27	27.66	26.33%	75.62%
RN_19	70-30%	6	102	12.11	828.17	28.78	26.13%	75.60%
RN_4	80-20%	3	77	11.15	778.53	27.90	25.55%	75.19%
RN_7	80-20%	6	67	11.93	781.40	27.95	28.11%	75.10%
RN_3	80-20%	3	52	11.80	791.64	28.14	25.86%	74.78%
RN_8	80-20%	6	107	11.61	813.93	28.53	25.92%	74.07%
RN_13	80-20%	30	237	11.87	824.83	28.72	26.38%	73.72%
RN_6	80-20%	6	102	11.66	828.58	28.79	27.31%	73.60%
RN_21	70-30%	6	107	12.54	913.99	30.23	27.98%	73.08%
RN_5	80-20%	6	102	12.48	849.82	29.15	27.87%	72.92%
RN_12	80-20%	12	142	13.29	884.25	29.74	32.23%	71.83%
RN_22	70-30%	6	187	14.43	1034.18	32.16	26.55%	69.54%
RN_25	70-30%	12	142	13.14	1037.80	32.21	26.64%	69.43%
RN_23	70-30%	12	227	13.04	1047.02	32.36	27.31%	69.16%
RN_9	80-20%	6	187	13.01	992.06	31.50	27.31%	68.39%

RN_10	80-20%	12	227	15.20	1118.64	33.45	31.01%	64.36%
RN_11	80-20%	12	402	14.34	1138.28	33.74	27.16%	63.73%
RN_24	70-30%	12	402	14.79	1299.94	36.05	27.20%	61.71%
RN_26	70-30%	30	237	14.76	1402.08	37.44	29.22%	58.70%

Imagen 70. Resultados obtenidos de los modelos de redes neuronales
Fuente: Propia

De todas estas variantes creadas las **5 que mejor resultados han logrado están formadas por 3 capas** (1 de entrada, 1 oculta y 1 de salida), esto es, **las más simples de todas las configuraciones**. También se puede observar que, las redes con mayor número de capas y neuronas son las que peor se han comportado, presentando una gran diferencia en la precisión con los mejores modelos.

También se puede comprobar que los RMSE van aumentando en los modelos, a medida que disminuye su precisión.

En cuanto a la división de los datos, claramente el **porcentaje 70-30% obtiene mejores resultados**, habiendo 6 de sus modelos, en las 7 primeras posiciones.

De los datos expuestos en la tabla anterior, **se puede concluir que los modelos de redes neuronales varían mucho según la estructura adoptada de capas y neuronas**, pero que, una vez encontrada una estructura adecuada para los datos, **obtienen buenos resultados con alta precisión y un valor no muy alto de RMSE**.

9. CONCLUSIONES

El **objetivo principal de este Trabajo Fin de Máster** que nos habíamos marcado, suponía analizar diferentes modelos predictivos para pronosticar el precio de la energía eléctrica en España, y analizar cual de ellos es el que ofrece mejores resultados.

Para lograrlo se ha creado, a partir de diferentes fuentes, un set de datos con más de 100.000 reseñas, pertenecientes a 27 variables, medidas durante un periodo temporal de 10 años.

Los modelos analizados han sido:

- Regresión lineal múltiple.
- k vecinos más cercanos.
- Árbol de decisión.
- Random forest.
- XG Boost.
- Red neuronal.

No nos hemos limitado a desarrollar un solo ejemplo por modelo, sino que **se han realizado variantes de cada uno de ellos, tratando de obtener más evidencias, a la hora de seleccionar los mejores modelos para predecir la variable objetivo, como para analizar la influencia que ejercen diferentes factores en el desarrollo de ellos.**

A modo de resumen, incluimos a continuación una tabla con los algoritmos y variantes estudiadas ordenadas de mayor a menor precisión obtenida.

	Div. Datos	MAE	MSE	RMSE	MAPE	R ²
KNN_2	70-30%	3.10	21.78	4.67	7.43%	79.24%
KNN_1	80-20%	3.18	22.57	4.75	8.68%	77.82%
RN_14	70-30%	10.35	768.00	27.71	27.68%	77.38%
RN_16	70-30%	11.51	782.44	27.97	26.58%	76.95%
RN_1	80-20%	9.62	736.51	27.14	26.52%	76.53%
RN_15	70-30%	11.62	806.39	28.40	27.05%	76.25%
RN_17	70-30%	11.86	808.22	28.43	27.32%	76.19%
RN_20	70-30%	12.28	821.17	28.66	26.37%	75.81%
RN_18	70-30%	12.22	822.62	28.68	26.81%	75.77%
RN_2	80-20%	11.68	765.27	27.66	26.33%	75.62%
RN_19	70-30%	12.11	828.17	28.78	26.13%	75.60%
KNN_3	80-20%	10.21	773.79	27.82	30.67%	75.35%
RN_4	80-20%	11.15	778.53	27.90	25.55%	75.19%
RN_7	80-20%	11.93	781.40	27.95	28.11%	75.10%

MODELOS PARA LA PREDICCIÓN DEL PRECIO DE LA ELECTRICIDAD

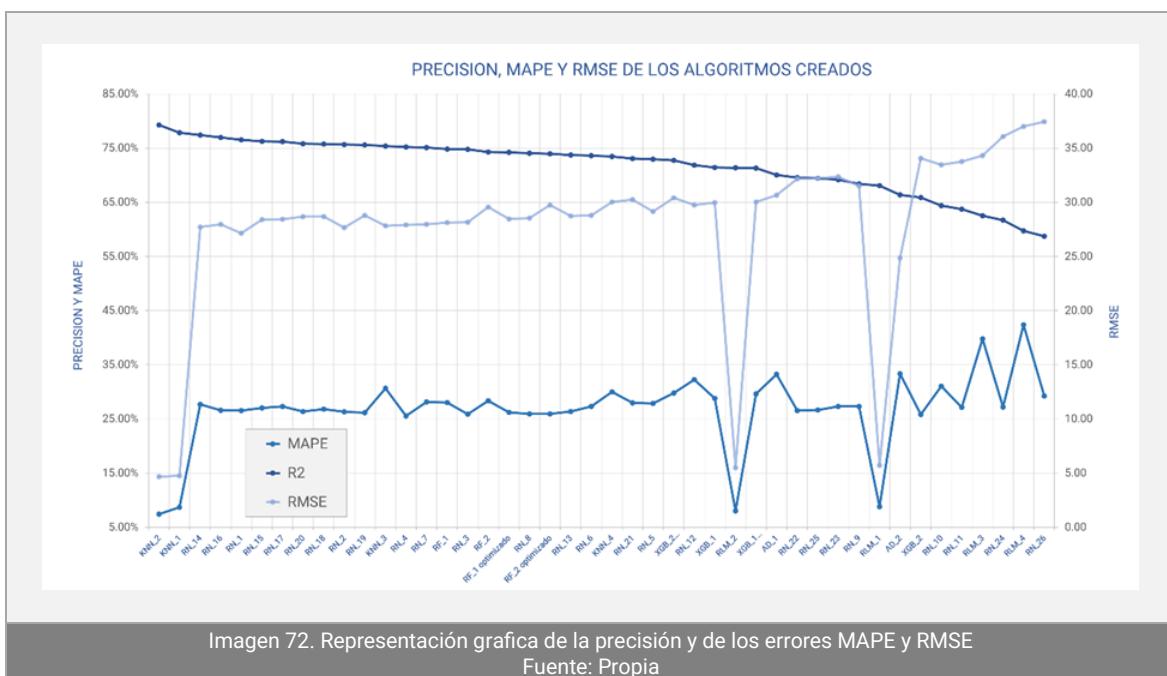
EN EL MERCADO ESPAÑOL

RF_1	80-20%	11.34	791.17	28.13	28.00%	74.79%
RN_3	80-20%	11.80	791.64	28.14	25.86%	74.78%
RF_2	70-30%	11.74	873.47	29.55	28.33%	74.27%
RF_1 optimizado	80-20%	9.78	809.44	28.45	26.20%	74.21%
RN_8	80-20%	11.61	813.93	28.53	25.92%	74.07%
RF_2 optimizado	70-30%	10.00	885.27	29.75	25.92%	73.92%
RN_13	80-20%	11.87	824.83	28.72	26.38%	73.72%
RN_6	80-20%	11.66	828.58	28.79	27.31%	73.60%
KNN_4	70-30%	10.88	900.80	30.01	30.00%	73.46%
RN_21	70-30%	12.54	913.99	30.23	27.98%	73.08%
RN_5	80-20%	12.48	849.82	29.15	27.87%	72.92%
XGB_2 optimizado	70-30%	12.82	924.92	30.41	29.76%	72.75%
RN_12	80-20%	13.29	884.25	29.74	32.23%	71.83%
XGB_1	80-20%	11.34	897.33	29.96	28.80%	71.41%
RLM_2	70-30%	3.97	30.07	5.48	8.03%	71.33%
XGB_1 optimizado	80-20%	12.42	900.93	30.02	29.62%	71.29%
AD_1	80-20%	14.36	940.20	30.66	33.24%	70.04%
RN_22	70-30%	14.43	1034.18	32.16	26.55%	69.54%
RN_25	70-30%	13.14	1037.80	32.21	26.64%	69.43%
RN_23	70-30%	13.04	1047.02	32.36	27.31%	69.16%
RN_9	80-20%	13.01	992.06	31.50	27.31%	68.39%
RLM_1	80-20%	4.09	32.52	5.70	8.77%	68.05%
AD_2	70-30%	11.98	618.20	24.86	33.34%	66.35%
XGB_2	70-30%	11.52	1159.21	34.05	25.82%	65.85%
RN_10	80-20%	15.20	1118.64	33.45	31.01%	64.36%
RN_11	80-20%	14.34	1138.28	33.74	27.16%	63.73%
RLM_3	80-20%	18.11	1176.19	34.30	39.79%	62.52%
RN_24	70-30%	14.79	1299.94	36.05	27.20%	61.71%
RLM_4	70-30%	18.70	1368.13	36.99	42.36%	59.70%
RN_26	70-30%	14.76	1402.08	37.44	29.22%	58.70%

Imagen 71. Clasificación de los algoritmos según su precisión

Fuente: Propia

También incluimos una grafica para la misma clasificación exponiendo, tanto la **precisión** de los modelos, como su **RMSE** y su **MAPE**.



Observando la grafica, la primera conclusión que podemos sacar, es que a medida que la precisión baja en los modelos, el error en ellos, entre el valor real y el predicho, aumenta.

Esta clasificación nos revela que **los mejores modelos para predecir el precio de la electricidad en España** son los **k vecinos más cercanos** en los casos en que los **outliers han sido eliminados**.

El siguiente modelo que mejor se ha comportado con los datos disponibles ha sido claramente las **redes neuronales**, y dentro de ellas **las más simples**, es decir, las que menos capas y neuronas tienen. Estas han tenido muy buenos valores de predicción, aunque a lo que se refiere a las métricas de errores, poseen valores más elevados que las dos mejores variantes de los modelos de KNN.

Las variantes de los algoritmos **random forest** también **han presentado buenos precisiones**, aunque ligeramente inferiores a las redes neuronales. Sus niveles en los errores son muy similares a los de estas ultimas.

En cuanto a los modelos **XG Boost** los resultados obtenidos son peores a los anteriores modelos, manteniendo los mismos valores en los errores, pero bajando su precisión. Aun así, la precisión de las tres primeras variantes es aceptables para algoritmos de predicción de precios.

Los modelos de **regresión lineal múltiple**, según sean tratados los valores anormales, tienen un comportamiento muy diferente, ya que si se eliminan se obtienen unos valores en las métricas de errores muy bajos, pero si se limitan a un valor máximo estos errores aumentan considerablemente.

Los peores algoritmos de todos los que se han estudiado han sido los **árboles de decisión**, aunque la diferencia con los modelos de regresión lineal múltiple y XG Boost no es muy elevada.

Según la **métrica MAPE** (métrica que nos sirve para comparar con los modelos señalados en el artículo de investigación realizado por González C., Mira-McWilliams J. y Juárez I. (2015), cuyos valores se han señalado en el capítulo **V. Objetivos a alcanzar** y que oscilan entre los **5,76%**, obtenidos en modelos de redes neuronales y análisis factorial dinámico estacional heterocedástico, y los **11%** de modelos de análisis factorial dinámico estacional, podemos comprobar que solo **4 de los 44 modelos y variantes** creadas en este estudio **poseen un MAPE acorde a esos valores, siendo estas las 2 mejores variantes de KNN y las 2 mejores de regresión lineal múltiple**.

	Div. Datos	MAE	MSE	RMSE	MAPE	R ²
KNN_2	70-30%	3.10	21.78	4.67	7.43%	79.24%
KNN_1	80-20%	3.18	22.57	4.75	8.68%	77.82%
RLM_2	70-30%	3.97	30.07	5.48	8.03%	71.33%
RLM_1	80-20%	4.09	32.52	5.70	8.77%	68.05%

Imagen 73. Algoritmos que llegan al objetivo marcado en el TFM
Fuente: Propia

De entre estos 4 modelos claramente los mejores son los KNN ya que poseen una precisión más alta que los RLM.

Otro factor, que en algunas circunstancias puede ser fundamental a la hora de emplear un algoritmo u otro, es el **coste computacional** que requiere. En la tabla adjunta a continuación se pueden ver los **tiempos que tarda el ordenador en desarrollar y calcular los distintos algoritmos** creados.

MODELOS PARA LA PREDICCIÓN DEL PRECIO DE LA ELECTRICIDAD

EN EL MERCADO ESPAÑOL

	Div. Datos	Nº capas	Nº neur. total	Tiempo (sg)	Tiempo (min)
RLM_4	70-30%	--	--	0.007	0.000
RLM_2	70-30%	--	--	0.044	0.001
RLM_3	80-20%	--	--	0.084	0.001
RLM_1	80-20%	--	--	0.286	0.005
AD_2	70-30%	--	--	0.349	0.006
AD_1	80-20%	--	--	0.522	0.009
KNN_2	70-30%	--	--	1.286	0.021
KNN_1	80-20%	--	--	1.312	0.022
KNN_4	70-30%	--	--	2.59	0.043
KNN_3	80-20%	--	--	3.71	0.062
RF_2	70-30%	--	--	23.69	0.395
RN_11	80-20%	12	402	23.83	0.397
RF_1	80-20%	--	--	24.59	0.410
RN_9	80-20%	6	187	24.59	0.410
RN_23	70-30%	12	227	25.57	0.426
RN_5	80-20%	6	102	26.13	0.436
RN_22	70-30%	6	187	26.25	0.438
RN_24	70-30%	12	402	26.69	0.445
RN_19	70-30%	6	102	28.20	0.470
RN_10	80-20%	12	227	28.27	0.471
RN_21	70-30%	6	107	29.41	0.490
RN_25	70-30%	12	142	30.12	0.502
RN_6	80-20%	6	102	32.36	0.539
RN_8	80-20%	6	107	34.30	0.572
RN_12	80-20%	12	142	35.54	0.592
RN_4	80-20%	3	77	45.12	0.752
RN_26	70-30%	30	237	45.24	0.754
RN_7	80-20%	6	67	47.65	0.794
XGB_2	70-30%	--	--	47.96	0.799
RN_13	80-20%	30	237	52.79	0.880
RN_18	70-30%	6	102	57.18	0.953
RN_17	70-30%	3	77	64.08	1.07
RN_20	70-30%	6	67	64.42	1.07
RN_3	80-20%	3	52	70.81	1.18
RN_2	80-20%	3	37	75.03	1.25
RN_16	70-30%	3	52	77.42	1.29
RN_15	70-30%	3	37	87.94	1.47
XGB_1	80-20%	--	--	104.65	1.74
XGB_2 optimizado	optimizado	--	--	118.67	1.98
XGB_1 optimizado	optimizado	--	--	137.58	2.29
RN_14	70-30%	3	32	226.18	3.77
RN_1	80-20%	3	32	302.27	5.04
RF_2 optimizado	optimizado	--	--	418.00	6.97
RF_1 optimizado	optimizado	--	--	488.70	8.15

Imagen 74. Tiempo de cálculo de cada modelo

Fuente: Propia

Se observa que los más rápidos son los algoritmos de regresión lineal múltiple, seguidos de los arboles de decisión, que tardan en los dos casos menos de medio segundo en realizar los cálculos.

Los siguientes, en términos de rapidez de cálculo, son los k vecinos más cercanos que se mueven en una horquilla de entre 1 y 4 segundos.

Las redes neuronales aumentan considerablemente el tiempo tardando entre los 24 segundos y los 5 minutos.

En cuanto los modelos random forest, los simples tardan apenas 25 segundos, pero los optimizados llegan a tener un coste de 8 minutos.

Los modelos basados en XG Boost poseen un coste no muy excesivo presentando una media de 1,7 minutos.

9.1 CONCLUSION FINAL

Se ha logrado el objetivo marcado de desarrollar algoritmos capaces de predecir los precios de la energía eléctrica en el mercado español, con valores de error (MAPE) dentro de los varemos establecidos en estudios especializados en la materia.

Estos modelos han sido 2 algoritmos basados en k vecinos más cercanos con los outliers eliminados y otros 2 basados en regresión lineal múltiple, presentando los primeros un nivel de precisión muy alto y los segundos, aceptable.

9.2 DESARROLLO FUTURO

Tras los buenos resultados de este proyecto, y debido al interés que puede generar en los distintos agentes intervenientes en el mercado de la electricidad en España, se han definido una serie de posibles actuaciones a realizar:

- Incorporar nuevas variables que afecten en el precio de la electricidad.
- Aplicar nuevos modelos y comprobar si los resultados obtenidos mejoran los presentados en este documento.

- En cuanto a las redes neuronales, se podría seguir investigando en nuevas arquitecturas.
- Combinar algoritmos buscando mejores resultados.

10. REFERENCIAS BIBLIOGRAFICAS

A continuación, enumeramos las reseñas de la documentación que ha sido consultada para la realización de este TFM.

- Alonso, A., García-Martos, C., Rodríguez, J., Sánchez, M.J.: 'Seasonal dynamic factor analysis and bootstrap inference: application to electricity market forecasting', *Technometrics*, 2011, 53, pp. 137–151
- Bunn, D. W. (2004). *Modelling prices in competitive electricity markets*. Chichester: John Wiley.
- Catalão, J.P.S., Pousinho, H.M.I., Mendes, V.M.F.: 'Short-term electricity prices forecasting in a competitive market by a hybrid intelligent approach', *Energy Convers. Manag.*, 2011, 52, pp. 1061–1065
- Eydeland, A., & Wolyniec, K. (2003). *Energy and power risk management*. Hoboken, NJ: Wiley.
- García-Martos, C., Rodríguez, J., Sánchez, M.J.: 'Forecasting electricity prices by extracting dynamic common factors: application to the Iberian Market', *IET Gener. Transm. Distrib.*, 2011, 1, pp. 1–10
- García-Martos, C., Rodríguez, J., Sánchez, M.J.: 'Forecasting electricity prices and their volatilities using unobserved components', *Energy Econ.*, 2011, 33, pp. 1227–1239
- González, C., Mira-McWilliams, J., & Juárez, I. (2015). Importance variable assessment and electricity price forecasting based on regression tree models: classification and regression trees, Bagging and Random Forests. *IET Generation, Transmission & Distribution*, 9, 1120-1128.

- Mori, H., Awata, A.: 'Data mining of electricity price forecasting with regression tree and normalized radial basis function network'. Proc. of the IEEE Int. Conf. on Systems, Man and Cybernetics, 2007, ISIC, (doi:10.1109/ICSMC.2007.4414228)
- Neupane, B., Perera, K.S., Aung, Z., Woon, W.L.: 'Artificial neural network-based electricity price forecasting for smart grid deployment'. Int. Conf. on Computer Systems and Industrial Informatics (ICCSII), 18–20 December, 2012, (ISBN: 978-1-4673-5155-3. doi: .119/ICCSII.2012.6454392)
- Pórtoles, J., González, C., & Moguerza, J. (2018). Electricity price forecasting with dynamic trees: A benchmark against random forest approach. *Energies*, 11,1588.
- Querol, J. (2019). Desarrollo de un modelo de predicción del precio de la energía eléctrica para el mercado a plazo mediante redes neuronales. Valencia: Universidad Politécnica de Valencia. E.T.S.I. Industriales.
- Smarra, F., Jain, A., de Rubeis, T., Ambrosini, D., D'Innocenzo, A., & Mangharam, R. (2018). Data-driven model predictive control using random forests for building energy optimization and climate control. *Applied Energy*, 1252-1272.
- Troncoso, A., Riquelme, J.M., Gómez, A., Martínez, J.L., Riquelme, J.C.: 'Electricity market price forecasting based on weighted nearest neighbors techniques', *IEEE Trans. Power Syst.*, 2007, 22, pp. 1294–1301
- Ugarte, A. R. (2017). Predicción de precios de energía eléctrica utilizando árboles dinámicos. Madrid: Universidad Politécnica de Madrid. E.T.S.I. Industriales.
- Weron, R. (2006). Modeling and forecasting electricity loads and prices: a statistical approach. Chichester: Wiley.
- Weron, R. (2014). Electricity price forecasting: A review of the state-of-the-art with a look into the future. *International Journal of Forecasting*, 1030-1081.