



ESCUELA TÉCNICA SUPERIOR DE INGENIERÍA (ICAI)

MÁSTER EN BIG DATA: TECNOLOGÍA Y ANALÍTICA  
AVANZADA

# MODELO DE PREDICCIÓN DE SOBRECOSTES DEL SISTEMA ELÉCTRICO ESPAÑOL

Autor: Sergio Rincón Simón

Director: María Olivares

**Madrid**

Junio 2020



Declaro, bajo mi responsabilidad, que el Proyecto presentado con el título

**MODELO DE PREDICCIÓN DE SOBRECOSTES DEL SISTEMA ELÉCTRICO ESPAÑOL** en la ETS de Ingeniería - ICAI de la Universidad Pontificia Comillas en el curso académico 2020/21 es de mi autoría, original e inédito y no ha sido presentado con anterioridad a otros efectos.

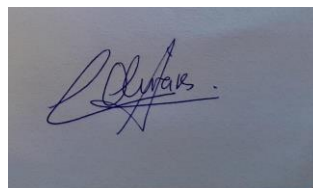
El Proyecto no es plagio de otro, ni total ni parcialmente y la información que ha sido tomada de otros documentos está debidamente referenciada.

Fdo.: Sergio Rincón Simón                      Fecha: 30/ 06/ 2021

Autorizada la entrega del proyecto

EL DIRECTOR DEL PROYECTO

Fdo.: María Olivares                      Fecha: 28/ 06/ 2021

A rectangular box containing a handwritten signature in blue ink, which appears to be 'M. Olivares'.

Vº Bº del Coordinador de Proyectos

Fdo.: Carlos Morrás Ruiz-Falcó                      Fecha: 28/ 06/ 2021



**Agradecimientos:**

*Quiero dedicar este TFM a mi abuelo, que no pudo verme terminar mis estudios, pero desde dónde esté, sé que estará muy orgulloso.*

## Índice de la memoria

**Resumen:** 7

<b>1. Introducción.....</b>	<b>8</b>
1.1 Contextualización del proyecto.....	8
1.2 Estado de la cuestión.....	9
1.3 Motivación del proyecto.....	9
1.4 Objetivos del proyecto.....	10
1.5 Metodología del trabajo.....	11
1.6 Recursos para emplear.....	12
<b>2. Contextualización del mercado eléctrico español.....</b>	<b>14</b>
2.1 Marco general del mercado eléctrico español.....	14
2.1.1 Aspectos básicos de la electricidad.....	14
2.1.2 Contextualización histórica del mercado eléctrico español.....	15
2.1.3 Marco regulatorio actual.....	18
2.2 Funcionamiento del mercado eléctrico.....	20
2.2.1 Agentes del sistema eléctrico.....	20
2.2.2 Operativa del mercado.....	24
<b>3. Desarrollo del Proyecto I: Caso de negocio.....</b>	<b>29</b>
3.1 Los sobrecostes del mercado eléctrico: Servicios de ajustes del sistema.....	29
3.2 Ventaja competitiva que ofrece la modelización del precio del sobrecoste.....	31
<b>4. Desarrollo del Proyecto II: Extracción de datos.....</b>	<b>33</b>
4.1 Horizonte temporal para la descarga de series.....	33
4.2 Extracción de datos de fuentes públicas.....	34
4.2.1 Fuentes públicas utilizadas.....	34
4.2.2 Metodología de extracción.....	34
4.2.3 Variables recopiladas.....	35
4.3 Extracción de datos de fuentes internas de engie.....	37
4.3.1 Fuentes internas utilizadas.....	37
4.3.2 Variables recopiladas.....	38

---

4.4 Limpieza y preparación de datos.....	39
4.4.1 Creación de nuevas variables.....	39
4.4.2 Transformación de datos.....	41
<b>5. Desarrollo del proyecto III: Análisis exploratorio de datos.....</b>	<b>42</b>
5.1 Análisis preliminar de las series.....	42
5.2 Análisis de relaciones.....	46
5.3 Análisis del precio de sobrecoste.....	52
5.3.1 Análisis de la serie.....	52
5.3.2 Outliers.....	56
5.3.3 Distribución.....	57
5.3.4 Correlaciones.....	59
5.3.5 Estacionalidad.....	61
<b>6. Desarrollo del proyecto IV: Modelización de sobrecostes.....</b>	<b>64</b>
6.1 Planteamiento inicial.....	64
6.2 Desarrollo de modelos.....	66
6.2.1 Regresión múltiple.....	66
6.2.2 árbol de decisión.....	70
6.2.3 Random forest.....	73
6.2.4 Stacking.....	78
<b>7. Desarrollo del proyecto V: Análisis de resultados.....</b>	<b>81</b>
7.1 Comparativa de modelos y elección de modelo para producción.....	81
7.2 Aplicación del modelo en escenarios.....	88
<b>8. Futuras líneas de desarrollo.....</b>	<b>91</b>
8.1 Puesta en producción.....	91
8.2 Mejora continua del modelo.....	92
<b>9. Conclusiones del proyecto.....</b>	<b>94</b>
<b>10. Bibliografía.....</b>	<b>96</b>
<b>ANEXO A: Diagrama de Gantt.....</b>	<b>99</b>

## *Índice de figuras*

Figura 1: Esquema proyecto de machine learning.....	12
Figura 2: Agentes del mercado eléctrico [6] .....	22
Figura 3: Secuencia mercados eléctricos en España [8].....	24
Figura 4: Agregación de la curva de oferta [8].....	26
Figura 5: Agregación de la curva de demanda [8].....	27
Figura 6: Casación del precio de la energía. [8] .....	28
Figura 7: Mix energético diario, fecha: 2020-10-25.....	42
Figura 8. Comparación diaria mix energético, demanda y spot diario.....	43
Figura 9: Comparación diaria mix energético, demanda y spot diario del mes de diciembre de 2020. ....	44
Figura 10: Comparación de 3 series de precios en el histórico de datos de manera horaria. ....	45
Figura 11: “Zoom” comparativo de series sobre los primeros meses de 2019.....	46
Figura 12. Matriz de correlaciones horarias. ....	47
Figura 13: Matriz de correlaciones diarias. ....	48
Figura 14: Matriz de correlaciones mensuales. ....	49
Figura 15: Relación entre Precio Spot y Generación por Cogeneración (izquierda) y entre CSS y Precio medio horario componente restricciones PBF (derecha). ....	51
Figura 16: Relación entre CSS y Hueco térmico (izquierda) y entre Generación de Ciclo combinado y Precio Spot (derecha).....	52
Figura 17: Representación horaria de la serie de precios de sobrecostes (2018-2020).....	53
Figura 18: Representación diaria de la serie de precios de sobrecostes (2018-2020). ....	54
Figura 19: Representación mensual de la serie de precios de sobrecostes (2018-2020).....	54
Figura 20: Comparativa de las series de Precio de sobrecoste, Precio Spot diario y Demanda (2018-2020). ....	55



---

Figura 21: “Zoom” de la comparativa de las series de Precio de sobrecoste, Precio Spot diario y Demanda (Feb 2020-Junio2020). .....	55
Figura 22: Distribución de outliers de precios de sobrecostes de manera mensual (2018-2020). .....	57
Figura 23: Histograma de distribución del precio de sobrecoste diariamente (2018-2020). .....	58
Figura 24: Diagrama de barras de la distribución del precio de sobrecoste por día de semana (2018-2020). Siendo 0 lunes y 6 domingo. ....	58
Figura 25: Comparación diagrama barras horario de precio de spot y precio de sobrecoste. ....	59
Figura 26: Correlación de las variables con el precio del sobrecoste de manera diaria. ....	60
Figura 27: Autocorrelación del precio de sobrecoste con lag = 24 horas. ....	61
Figura 28: Autocorrelación del precio de sobrecoste con la última semana. ....	62
Figura 29: Autocorrelación parcial del precio de sobrecoste con las últimas 24 horas.....	62
Figura 30: Autocorrelación parcial del precio de sobrecoste con las última semana.....	63
Figura 31:Residuos de modelo de regresión múltiple. ....	67
Figura 32: Comparación actuación de tres modelos de regresión múltiple en los conjuntos de train y test. Rojo-Real, Verde-Predicho.....	69
Figura 33: Barrido del parámetro de complejidad a la hora de crear un modelo de árbol de regresión. ....	70
Figura 34: Ejemplo de estructura de modelo de árbol de regresión (no se muestran el contenido del árbol por motivos de confidencialidad). ....	71
Figura 35: Ejemplo de gráfica que mide la importancia de las variables que se han utilizado para definir el árbol (el contenido del árbol no se muestra por motivos de confidencialidad). ....	71
Figura 36: Comparación actuación de los dos modelos de árboles de regresión en los conjuntos de train y test. Rojo-Real, Verde-Predicho. ....	73
Figura 37: Cálculo del error de predicción frente al número de árboles combinados en random.forest.....	74

Figura 38: Comparación de Out of Bag Error y Test Error en función del hiperparámetro $m$ try.....	75
Figura 39: Comparación actuación de los dos modelos de random forest en el conjunto de test. Rojo-Real, Verde-Predicho.....	77
Figura 40: Comparación actuación del modelo de stacking en los conjuntos de train y test. Rojo-Real, Verde-Predicho. ....	80
Figura 41: Comparación actuación de los modelos de regresión lineal multivariante en el conjunto de datos de febrero a mayo de 2021. Rojo-Real, Verde-Predicho. ....	83
Figura 42: Comparación actuación de los modelos de árboles de regresión en el conjunto de datos de febrero a mayo de 2021. Rojo-Real, Verde-Predicho. ....	84
Figura 43: Comparación actuación de los modelos de random forest en el conjunto de datos de febrero a mayo de 2021. Rojo-Real, Verde-Predicho.....	85
Figura 44: Actuación del modelo de stacking en el conjunto de datos de febrero a mayo de 2021. Rojo-Real, Verde-Predicho. ....	86
Figura 45: Predicción del precio de sobrecoste para tres escenarios de la hipótesis 1 junto a la serie del precio de spot base. ....	89
Figura 46: Predicción del precio de sobrecoste para tres escenarios de la hipótesis 2 junto a la serie del precio de spot base. ....	90
Figura 47: Flujograma puesta en producción del modelo. ....	92

## *Índice de tablas*

Tabla 1: Correlaciones entre variables más relevantes.....	50
Tabla 2:Resultados estadísticos modelos de regresión lineal múltiple.....	68
Tabla 3: Resultados estadísticos modelos de árboles de regresión. ....	72
Tabla 4: Resultados estadísticos modelos de random forest. ....	76
Tabla 5:Pesos de los distintos modelos combinados con Stacking .....	78
Tabla 6: Resultados estadísticos modelo de stacking.....	79
Tabla 7:Resultados estadísticos de la validación de los modelos en los meses de febrero a mayo de 2021. ....	81
Tabla 8: Comparativa error mensual de predicción de los modelos para los meses de febrero y marzo (Verde mejores resultados).....	87
Tabla 9: Comparativa error mensual de predicción de los modelos para los meses de abril y mayo (Verde mejores resultados).....	87

# ***RESUMEN:***

El trabajo realizado trata de cubrir en su totalidad un proyecto de machine learning o analítica de datos:

Se trata de modelar la serie de precios del mercado eléctrico que representa el sobrecoste del sistema eléctrico (compuesto por varias series de precios). El mercado eléctrico mayorista no es un mercado perfecto, toda la energía que se genera no se dedica a cubrir la demanda perfectamente, existen restricciones técnicas y mercados secundarios que tratan de regular estos desajustes en el mercado. Esta serie de precios es la que se trata de predecir, con el objetivo de poder presentar mejores ofertas de precio fijo a clientes y reducir riesgos financieros asociados. Tras la comprensión del caso de negocio se ha procedido a la extracción de datos mediante APIs públicas e internas de la empresa con su posterior tratamiento y limpieza de datos.

Una vez con los datos preparados se trata de resolver un problema de regresión para predecir el precio de sobrecoste. Se trata de resolver el problema inicialmente usando modelos de regresión lineal multivariante, árboles de regresión, random forest y stacking. Finalmente, tras un análisis de resultados, se decide utilizar el primer algoritmo de los enumerados. Para ponerlo en producción se integrará el código desarrollado en python con la herramienta de Power-BI para facilitar la accesibilidad de la solución encontrada al departamento o interesados que vayan a hacer uso de ella.

# **1. INTRODUCCIÓN.**

En este capítulo se hace una introducción de este proyecto despertando el interés del lector por el proyecto y describiendo la motivación del proyecto.

## **1.1 CONTEXTUALIZACIÓN DEL PROYECTO.**

A la hora de desarrollar un proyecto de Ciencia de datos hay que tener en cuenta una triple visión:

1. Un enfoque que trate de comprender el caso de negocio, es decir, se trata de comprender el marco y contexto que rodea al problema al que se quiere dar solución.
2. Un enfoque matemático y estadístico, que se basen en algoritmos conocidos de manera que se puedan proponer distintos caminos para alcanzar una solución al problema.
3. Un enfoque que sea consciente de las limitaciones y posibilidades tecnológicas que pueden ser aplicadas para la resolución del problema.

El problema que se quiere resolver mediante la realización de este proyecto trata de abarcar estos tres enfoques.

La electricidad desempeña un papel básico en la sociedad actual y es de vital importancia garantizar su suministro a todos los niveles. El funcionamiento del sector eléctrico tiene una cierta complejidad por la existencia de multitud de agentes involucrados. Para una mejor comprensión del caso negocio y poder alcanzar una mejor solución es muy importante definir y comprender el funcionamiento del mercado eléctrico.

Construir un modelo capaz de predecir series de precios del mercado eléctrico requiere de conocimientos sobre el sector. También sobre los posibles métodos matemáticos y estadísticos aplicables al caso de uso y de qué herramientas y tecnologías se pueden usar para alcanzar los objetivos.

## ***1.2 ESTADO DE LA CUESTIÓN.***

Entre las actividades a desarrollar durante el proyecto encontramos las siguientes:

1. Análisis de fuentes de datos a utilizar en el modelo (internas, externas).
2. Visualización de la información, cuadros de mando para el seguimiento de las previsiones y contextualización del caso de negocio.
3. Modelos de previsión utilizando algoritmos Machine Learning, Deep learning, etc.
4. Descripción del ecosistema analítico y tecnológico de la compañía para ejecución y puesta en producción del proyecto realizado.

## ***1.3 MOTIVACIÓN DEL PROYECTO.***

El mercado eléctrico español con todos sus órganos reguladores y agentes de acción conforman una compleja red interrelacionada. El funcionamiento del mercado pese a ser desconocida para la mayoría de la sociedad repercute de manera muy directa en el modo de vida contemporáneo. En este mercado es dónde se comercializa con la energía que llega a los hogares, industrias, vías públicas, centros de ocio, etc.

En el mercado eléctrico a grandes rasgos las compañías generadoras ofertan una cantidad de energía en función de la demanda de esta. Por otra parte, las compañías comercializadoras ofertan precios de compra por esa energía generada de manera que según la oferta que haga la comercializadora su precio será casado con la generación según los precios de la competencia. En esta situación, se puede apreciar sin duda una posible oportunidad de negocio para aplicar conocimientos analíticos y de predicción.

Poder desarrollar modelos de predicción apoyados en potentes herramientas de cómputo y en las matemáticas permite estimar precios de la electricidad con cierta precisión. Este tipo de proyectos pueden servir para la mejora de la toma de decisiones a la hora de ofertar precios de compra de la energía y sus procesos asociados. Por otro lado, puede servir de un

conocimiento extra a los distintos traders o responsables encargados de operar con precios y derivados de la energéticos

El proyecto que se presenta trata de servir de apoyo y utilidad a una empresa comercializadora de primera línea como es Engie, de manera que se podrá apreciar y valorar un resultado real del trabajo realizado en un caso de uso real, con el objetivo de crear valor para la empresa.

### ***1.4 OBJETIVOS DEL PROYECTO.***

Los objetivos del proyecto se pueden descomponer a dos niveles. El primer nivel define los objetivos principales y un segundo nivel que de soporte a estos objetivos. Los objetivos principales del proyecto son:

1. El proyecto consistirá en el desarrollo de modelos de predicción de los precios de los servicios complementarios del mercado mayorista eléctrico español y análisis y previsión de los sobrecostes del sistema aplicando modelos de Machine Learning.
2. La aplicación práctica de estos sistemas será la optimización de las ofertas a los clientes para optimizar también el beneficio y disminuir el riesgo de la comercializadora.

Dentro de estos alcances, existen una serie de objetivos a menor nivel que se encuentran estrechamente relacionados:

1. Estudio y comprensión del funcionamiento del mercado eléctrico español: para el desarrollo de una solución consistente es necesario comprender el contexto de aplicación del proyecto. Es por ello necesario realizar un estudio e interiorización del funcionamiento del mercado, de todos los agentes y variables que componen el sector eléctrico.
2. Estudio y comprensión de modelos matemáticos aplicables al caso de uso: se explorará el uso de diferentes algoritmos de Machine Learning y series temporales

aplicables al caso de uso. Es necesario realizar un análisis descriptivo de las variables, definir métricas de error, encontrar un modelo o modelos que satisfagan la predicción de precios con el menor error posible, estudiar los resultados de cada uno de los modelos y hacer uso de esta información para de nuevo realizar una búsqueda de los mejores modelos

3. Estudio de la puesta en producción del proyecto: se trata de un seguimiento y control del correcto funcionamiento de la solución desarrollada en el ámbito operativo de la empresa.

## ***1.5 METODOLOGÍA DEL TRABAJO.***

Se ha decidido estructurar el proyecto en 4 fases de trabajo:

1. Estudio del caso de uso: se definirá el alcance y principales objetivos que abarcar en el proyecto. A su vez, se realizará un estudio y comprensión del mercado eléctrico y los agentes que se encuentran involucrados para comprender de mejor manera el caso de negocio y sus posibles soluciones.
2. Detección, análisis de fuentes de información y selección de los datos: se realizará un estudio de las posibles bases de datos de donde poder extraer datos, tanto externas como internas de la empresa. Se tratará de garantizar la calidad del dato extraído para poder conformar los dataframes que interesen.
3. Modelización y analítica predictiva: se realizará un análisis exploratorio de los datos extraídos viendo relación entre las variables y demás análisis estadísticos preliminares. Posteriormente se aplicarán diferentes métodos de Machine Learning para poder predecir el precio de las series de precios de sobreajustes del mercado eléctrico. Finalmente se valorará la actuación de los distintos modelos probados.
4. Explotación de los resultados: una vez contrastado el correcto funcionamiento y validez de los modelos, se procede a su puesta en producción para que pueda servir



de apoyo a los departamentos de la empresa que lo necesiten. Se realizará un control y seguimiento de los resultados de la implantación en producción para depurar cualquier tipo de error o problema que pudiera surgir.

En definitiva, el proyecto trata de cubrir en la totalidad un proyecto de machine learning como se esquematiza en la siguiente figura:

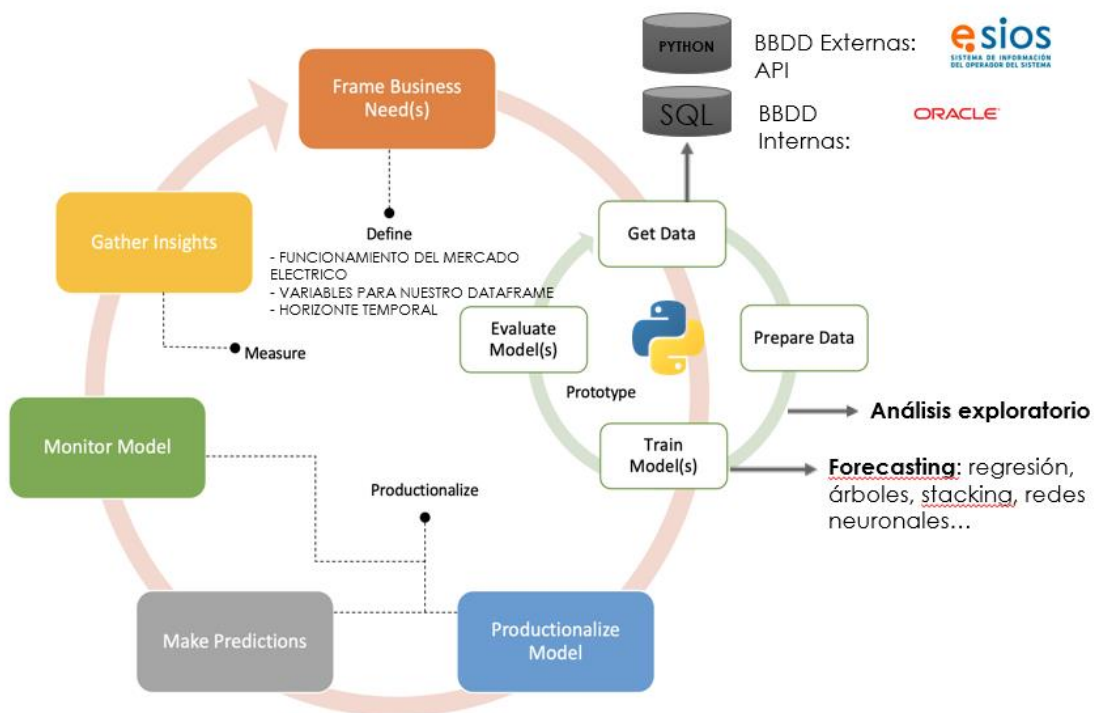


Figura 1: Esquema proyecto de machine learning.

El cronograma del proyecto se detalla en el anexo A.

## 1.6 RECURSOS PARA EMPLEAR.

El trabajo consta de una parte eminentemente analítica, para ello se ha optado por usar las siguientes herramientas para dar soporte:

1. Notebooks de Jupyter (Python): es una aplicación web que sirve a modo de puente constante entre el código y los textos explicativos. De este modo, los usuarios

pueden crear y compartir en tiempo real código, ecuaciones, visualizaciones, etc., junto con los textos explicativos. El programa se ejecuta desde la aplicación web cliente que funciona en cualquier navegador estándar. El lenguaje que se ha decidido para el desarrollo ha sido Python por la gran versatilidad que ofrece. Para la parte de visualización se utilizarán sobre todo las librerías de: Pandas, Matplotlib y Seaborn.

2. R-Studio (R): es un entorno de desarrollo integrado (IDE) para el lenguaje de programación R, dedicado a la computación estadística y gráficos. Se trata de un entorno que está sobradamente preparado para la modelización de algoritmos y su análisis. En este entorno se desarrollarán modelos de Machine Learning para la consecución de los objetivos del proyecto. Las principales librerías que se utilizaran son: ggplot, dplyr, caret...etc.
3. API ESIOS: La nueva web pública de e-sios (Red Eléctrica) pone a disposición de todos los usuarios una API para la descarga de información. Para poder utilizar esta API deberán solicitar un token personal enviando un correo a [consultasios@ree.es](mailto:consultasios@ree.es), ya que el token público que usa la propia página web cambia cada cierto tiempo.
4. Base de datos de Oracle (SQL): se trata de una base de datos interna de la empresa que cuenta con datos y variables del mercado eléctrico de uso concurrido por los empleados de la organización. La manera de consultar esta base de datos es mediante el lenguaje SQL.
5. Herramientas de ofimática de Microsoft: herramientas como PowerBI para visualizaciones, Excel como hoja de cálculo, Word y PowerPoint para documentación del proyecto.

Cabe destacar que el alcance del proyecto se podría verse ampliado en un futuro integrando herramientas de Big Data del ecosistema Hadoop.

## **2. CONTEXTUALIZACIÓN DEL MERCADO ELÉCTRICO ESPAÑOL**

En este apartado se enmarcará el proyecto realizado en el entorno del mercado eléctrico español. Se explicarán conceptos del mercado y su funcionamiento.

### ***2.1 MARCO GENERAL DEL MERCADO ELÉCTRICO ESPAÑOL.***

#### **2.1.1 ASPECTOS BÁSICOS DE LA ELECTRICIDAD.**

En física, la energía se define como la capacidad de un cuerpo o sustancia para realizar un trabajo.

Durante la explicación de este proyecto se usarán cierta terminología relacionada con la energía. Los parámetros que se usan para medir la energía son: la tensión cuya unidad es el voltio (V), la corriente o intensidad cuya unidad es el amperio (A), la potencia eléctrica cuya unidad es el vatio (W) y la energía eléctrica generada o consumida que se mide en n vatios-hora (Wh) y todas sus unidades de medida derivadas.

Atendiendo a aspectos tecnológicos y económicos, la energía se refiere a un recurso natural que adecuadamente manipulado y transformado es capaz de realizar un trabajo. Si la energía es transformada puede tener un uso industrial y este uso puede ser monetizado.

La energía como tal, ni se crea ni se destruye, sólo se transforma. No obstante, el hecho de que las transformaciones empleadas no sean reversibles hace que la energía se degrade y que, a la postre, no sea posible extraer más trabajo del recurso natural del que se parte inicialmente.

Uno de los aspectos más importantes derivado de la anterior afirmación, es que la energía no se puede almacenar en grandes cantidades y necesita que exista una continuidad eléctrica para su existencia. Este es un efecto físico básico para entender más adelante el funcionamiento del mercado y del sistema eléctrico

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. [1]

### **2.1.2 CONTEXTUALIZACIÓN HISTÓRICA DEL MERCADO ELÉCTRICO ESPAÑOL.**

La electricidad llega a España el año 1852 de la mano de un farmacéutico barcelonés que consiguió iluminar su farmacia con unas baterías, es entonces como en base a este hito, se estimula el uso de la electricidad como un bien del que se debe disponer. De hecho y a partir de acontecimientos sucesores al originario, como por ejemplo el alumbrado en las Ramblas, el Castillo de Montjuic y parte del Paseo de Gracia. Se crea en 1881 el primer proveedor eléctrico bajo el nombre de Sociedad Española de Electricidad.

La corriente alterna<sup>1</sup> que apareció a principios del siglo XX marcó un punto de inflexión dentro del sector, ya que se trata de una nueva forma de transportar la energía a largas distancias.

En España, durante los años de la guerra civil y los primeros años de la posguerra se produjo un estancamiento de la capacidad de producción. Además, la sequía de 1944-1945 años en los que la producción eléctrica dependía en gran parte de la generación hidroeléctrica, en torno al 80% del total, impidió atender una demanda creciente. de hasta el 27% anual para los años venideros. Para gestionar esta situación, en 1944 se fundó la empresa Unidad Eléctrica S.A. (UNESA), integrada por las principales compañías del sector. El objetivo principal de este conglomerado empresarial fue el de mejorar las interconexiones entre los diferentes sistemas eléctricos existentes y las centrales eléctricas para poder completar la red de transporte a nivel nacional.

---

<sup>1</sup> Corriente alterna: corriente eléctrica variable en la que las cargas eléctricas cambian el sentido del movimiento de manera periódica. Se transporta a grandes distancias con poca de pérdida de energía

La década de los sesenta, fueron años de bonanza para España ,acompañados de un gran crecimiento económico para el sector eléctrico, ya que, supuso la apertura al exterior y la consecución de las economías de escala con una reducción de los costes debido al aumento de grupo de generadores. Durante estos años, en base a los bajos precios del petróleo, se construyeron plantas generadoras que dependían de este combustible fósil y de esta manera abaratar los costes de la electricidad. En consecuencia, se produjo un descenso del poder de mercado de la energía hidroeléctrica, pues paso de acaparar un 84% en 1960 a un 39% del total en 1973.

En 1973 estalla la llamada “Crisis del petróleo” que supuso un aumento exponencial en el precio de este. Este acontecimiento supuso un gran inconveniente para los grupos de generación que entraron en servicio a mediados de los años setenta, que eran generadores de fuel-oil, que respondían a proyectos contratados con anterioridad a la primera crisis del petróleo. Tras la segunda crisis del petróleo, en 1979, se tomaron medidas para contener la dependencia del petróleo. En consecuencia, se promulgó en 1980 la Ley de Conservación de la Energía, con la idea de alcanzar tres objetivos:

1. Reducir el poder de mercado que tenía el petróleo sobre el sector energético.
2. Fomentar el ahorro de energía.
3. Promover las fuentes de energía renovables.

En línea con esas directrices, en la primera mitad de la década de los años 80 entraron en servicio las centrales de carbón nacional para contrarrestar la tendencia alcista que tenían los precios del petróleo. Simultáneamente, entre 1980 y 1986 entraron en servicio cinco grupos nucleares. También se empezó a apostar por la cogeneración y las energías renovables.

Gracias a estas políticas, a finales de los años ochenta el sector eléctrico español se encontraba en una situación bastante competitiva con una gran capacidad ociosa de generación energética. Este aumento de la capacidad energética de España fue acompañado de un aumento de la demanda que impulsó definitivamente la inversión del sector.

Pese a la mejora económica del sector seguía habiendo muchas empresas todavía lastradas por la crisis del petróleo. Los primeros pasos para estabilizar la situación de estas empresas fueron en 1985 con un intercambio de activos y plantas energéticas. Cabe señalar como el mayor logro en la estabilización del sector fue el establecimiento de un nuevo sistema de cálculo de las tarifas eléctricas. Esta política disminuiría el desequilibrio financiero, el sistema, fue conocido como Marco Legal y Estable.

Durante la década de los noventa, se aprobó la Directiva europea sobre normas comunes para el mercado interior de la electricidad, con el principal objetivo de introducir una liberalización del mercado basada en la competencia.

Entre 1996-2001, la demanda de electricidad se incrementó en más de un 30%, muy por encima de las previsiones. La demanda punta además también se vio incrementada significativamente, casi un 45%, al final este es el aspecto determinante a la hora de determinar las necesidades de infraestructuras energéticas. Durante este período, los precios medios de la electricidad se redujeron en torno a un 30% en términos reales. El conjunto del sector eléctrico se vio sumido en un período desafiante para poder hacer frente a estos acontecimientos, en un entorno de creciente incertidumbre, debido a la liberalización del sector que supuso una sensibilización del precio de la electricidad y a la ausencia de un sistema regulatorio estable sin incentivos para una mejoría de la gestión de la curva de generación eléctrica.

A partir de aquí, el siglo XXI se ha visto abocado a hacer frente a numerosos retos tanto técnicos como económicos. El sistema de producción de energía eléctrica se encuentra inmerso en un proceso de transformación, las fuentes de energía primaria que incluyen las renovables, las tecnologías a utilizar y los requisitos medioambientales. Desde una concepción tradicional en la que grandes centros de generación abastecían a los lugares de consumo por medio de redes de transporte y distribución de gran capacidad y distancia, se está evolucionando hacia lo que se ha denominado “Generación Distribuida” de energía eléctrica. En esta nueva concepción, los puntos de generación y consumo se encuentran más próximos, y las pérdidas de transporte y distribución pueden disminuir significativamente.

En este nuevo contexto de transición ecológica, cobra mayor importancia la adecuada integración de las fuentes de generación renovables (con su carácter intermitente y no gestionable) y el concepto de eficiencia energética. [2]

### **2.1.3 MARCO REGULATORIO ACTUAL.**

#### ***2.1.3.1 Marco Legal y Estable.***

Debido a las importantes medidas llevadas a cabo para contrarrestar y minimizar los efectos negativos debido a la crisis del petróleo, se produjo un fuerte endeudamiento acompañado de elevados tipos de interés en el sector eléctrico.

Ante esta situación de incertidumbre fue necesario la culminación de la reforma del marco regulador con el denominado Marco Legal Estable, aprobado en el Real Decreto 1538/1987 y vigente entre 1988 y 1997. La estabilidad se consiguió paulatinamente con su instauración, puesto que comportaba:

1. Una metodología de amortización y retribución de las inversiones.
2. Una retribución de los costes de producción y distribución en base a valores estándares.
3. Un sistema de compensaciones entre los agentes.
4. Una corrección por desviaciones al finalizar el año.

La instauración del Marco Legal Estable supuso las bases para la estandarización de costes, la retribución de los costes de inversión, producción y distribución e incentivos en la gestión. En definitiva, se reorientó el sector hacia resultados económicamente sostenibles. [3]

#### ***2.1.3.2 Liberalización del mercado.***

Para poder alcanzar la liberalización y la creación de un mercado único de electricidad en toda Europa se implantaron unas bases de carácter general. El nacimiento de ello viene precedido por la publicación en 1990 de la Directiva 90/547EEC y, seis años más tarde por la Directiva 96/92EC: Directiva europea sobre normas comunes para el mercado interior de la electricidad. Ésta última puede considerarse el punto de partida pues en ella quedaron

establecidas las bases para que los países modificasen su legislación relacionada con el sector eléctrico.

El proceso de liberalización en España comienza el 1997 a partir de la aprobación de la Ley del Sector Eléctrico 54/1997. Dicha ley entró en vigor el 1 de enero de 1998 y comportó los cambios normativos más relevantes del país, entre los que cabe destacar:

1. La liberalización del mercado minorista pudiendo elegir libremente el proveedor.
2. Diferenciación entre actividades reguladas por el Estado (Transporte y Distribución) en consecuencia, no sujetas a competencia y aquellas sujetas a competencia (Generación y Comercialización), fomentando el desarrollo del mercado, beneficiándose así los clientes finales de unos menores precios y la explotación del sistema nacional paso a encargarse la Red Eléctrica España (REE).

Esta ley está ya derogada por la nueva Ley del Sector Eléctrico, Ley 24/2013. [4]

### ***2.1.3.3 Ley actual del mercado eléctrico español.***

Desde su entrada en vigor y hasta el 2013 en especial, la Ley 54/1997 ha sufrido un periodo de transición hacia lo que sería la próxima Ley sobre la que se sustente el sector eléctrico, dicha etapa se caracterizó por la aprobación de una serie de leyes, reglamentos y ordenanzas limitando su alcance de poder normativo. La necesidad de una nueva Ley radica en dos motivos principalmente:

1. Se sitúa en un entorno de crisis económica global donde los países de la periferia europea son los más perjudicados debido a su alto endeudamiento, ante este contexto, el crecimiento de la demanda energética en España se estanca, por consiguiente, no es de extrañar la imposibilidad existente en aquel entonces de garantizar el equilibrio financiero del sistema a largo plazo.
2. Las ayudas para la inserción de energías renovables no fueron provechosas a causa de la distribución no óptima que se hizo, derivando en un aumento del déficit del país.



España estaba sometida a un contexto de incertidumbre donde el déficit tarifario ascendía aproximadamente a 3.600 millones de euros, lo que supuso un entorno insostenible a nivel económico y técnico para el sistema eléctrico del país.

En diciembre del año 2013 la Ley 24/2013 con el objetivo de retornar la estabilidad que gozaba el sector, garantizando el suministro de energía eléctrica bajo los principios de sostenibilidad económica y financiera de la producción eléctrica. Se abandonan los conceptos de régimen ordinario y especial, dando lugar a que todas las instalaciones respondan bajo una misma normativa.

En la actualidad y con el fin de fomentar la transición ecológica, se aprobó el Real Decreto 244/2019 el 5 de abril, con carácter administrativo, económico y técnico del autoconsumo de energía eléctrica (BOE-A-2019-5089.pdf,). Entre los principales aspectos cabe destacar la normalización del autoconsumo colectivo, permisividad que consumidor y propietario de la instalación sean diferentes y una tramitación más accesible para el consumidor final. [5]

## ***2.2 FUNCIONAMIENTO DEL MERCADO ELÉCTRICO.***

### **2.2.1 AGENTES DEL SISTEMA ELÉCTRICO.**

El Sistema Eléctrico Español se caracteriza por la existencia de un Mercado Mayorista de generación de electricidad (Polo Español o “Pool”) donde cada consumidor es libre para elegir a su empresa comercializadora desde enero de 2003 y donde es posible firmar contratos bilaterales entre productores y comercializadoras desde 2006 para fijar el precio de venta de la energía (PPAs).

Entre los agentes que intervienen en el suministro de electricidad a los consumidores en el mercado liberalizado, figuran los siguientes:

- Generadores. Producen electricidad y deben construir, operar y mantener las centrales de generación.

*CONTEXTUALIZACIÓN DEL MERCADO ELÉCTRICO ESPAÑOL*

---

- Productores en régimen especial. Son empresas productoras que tienen un tratamiento económico especial al mejorar la eficiencia energética y reducir el impacto medioambiental.
- Transportistas. Llevan la electricidad desde los centros de producción hasta la red de distribución, haciendo las tareas de construcción y mantenimiento de la red eléctrica de transporte.
- Distribuidores. Tienen que llevar la energía hasta el punto de consumo y venderla. Además, deben construir, mantener y operar las instalaciones de la red de distribución.
- Comercializadores. Todas las personas jurídicas que tienen como función la venta de energía eléctrica a los consumidores.
- Consumidores calificados. Son los consumidores que tienen un nivel de consumo anual que opera sobre unos valores determinados.
- Reguladores. La Administración del Estado y la Comisión Nacional de Energía.
- Operadores. El operador del mercado y el operador del sistema.

En la siguiente imagen se aprecia de manera esquemática las funciones de los distintos actores que forman el sistema eléctrico:

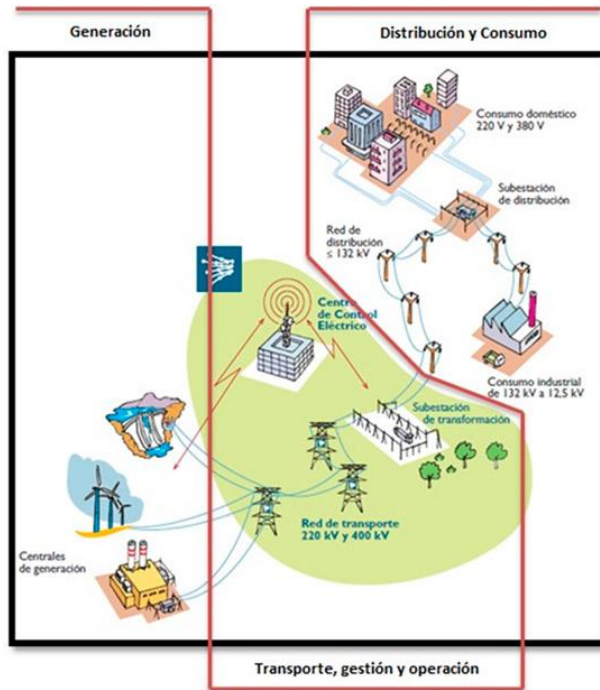


Figura 2: Agentes del mercado eléctrico [6]

### **2.2.1.1 Los organismos de gestión del mercado de la electricidad**

El operador del mercado y el operador del sistema son los organismos encargados de la gestión económica y técnica del sistema. La Compañía Operadora del Mercado Español de Electricidad (OMEL) es la responsable de la gestión económica de la generación y Red Eléctrica de España (REE) es la encargada de la gestión técnica del sistema.

#### **2.2.1.2 El operador del mercado**

El operador del mercado es el organismo que determina los precios finales de la electricidad. Lo hace gestionando las casaciones de las ofertas y las demandas en el mercado de producción.

En primer lugar, recibe las ofertas de venta de energía que hacen los productores de electricidad y también recibe las demandas de energía que hacen los agentes autorizados para realizarlos.

Una vez recibida esta información, el OMEL selecciona para cada hora del día siguiente la entrada en funcionamiento de las unidades de generación, empezando por las que han comunicado las ofertas de energía más baratas, hasta cubrir la totalidad de la demanda.

Deben publicar en los medios de difusión nacional toda la información de carácter público sobre el mercado de producción, asesorar al OMEL y proponer medidas regulatorias para mejorar su funcionamiento.

### ***2.2.1.3 El operador del sistema***

Red Eléctrica de España (REE) es responsable de la red de transporte de alta tensión y el operador del sistema. Por lo tanto, es el organismo encargado de garantizar una correcta coordinación del sistema de producción y transporte de electricidad, con la finalidad de asegurar la calidad y seguridad en el suministro de energía.

Dentro de las funciones del operador del sistema existe la de gestionar los programas de intercambios internacionales de electricidad con otros países, que son necesarios para mantener o incrementar la seguridad y la calidad en el suministro de electricidad.

También debe:

Informar de la capacidad de transporte y de interconexión del sistema eléctrico, así como las necesidades de interconexión con otras redes eléctricas.

Analizar todas las nuevas solicitudes de conexión a la red y limitar el acceso a ella cuando no disponga de capacidad suficiente o existan riesgos para la seguridad del suministro.

Establecer, en coordinación con los agentes del sistema, planes de maniobra para garantizar la reposición del servicio en caso de interrupciones.

El operador del mercado y el operador del sistema deben funcionar con un elevado grado de coordinación para enfrentarse de forma adecuada a situaciones excepcionales que puedan producirse en las redes de transporte o en el sistema de generación de electricidad. [7]

## 2.2.2 OPERATIVA DEL MERCADO.

### 2.2.2.1 Secuencia de mercados.

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.

En la siguiente imagen se muestra de manera esquemática la secuencia de mercados en el mercado ibérico de electricidad:



Figura 3: Secuencia mercados eléctricos en España [8]

Días, semanas, meses e incluso años antes del momento en que la energía sea generada y consumida, los agentes intercambian contratos con períodos de entrega de distinta duración

(anual, trimestral, mensual, etc.). Estas transacciones se realizan en los llamados mercados a plazo.

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 Eléctrico (OMIE). Además, ya 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 (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 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. [8]

### ***2.2.2.2 Formación del precio de la electricidad.***

En general, existen tres tipos de mercados en función de cómo se forma el precio en los mismos y que cubren la demanda energética total:

1. Mercados “pay as bid”, en los que un generador recibe exactamente el precio que él ha ofertado.
2. PPAs (Power Purchase Agreement) o acuerdos bilaterales: es un acuerdo de compraventa de energía limpia a largo plazo desde un activo concreto y a un precio prefijado entre un desarrollador renovable y un consumidor (por lo general, empresas que necesitan grandes cantidades de electricidad) o entre un desarrollador y un comercializador que revenderá la energía. [9]
3. Mercados marginalistas, en los que todos los generadores casados reciben un mismo precio, el cual se determina por el cruce de las curvas de oferta y demanda.

A pesar de las diferencias en cuanto a cómo se forma el precio, la teoría económica muestra que en ambos tipos de mercados (“pay as bid” y marginalistas) se obtienen los mismos resultados (es decir, mismos precios y cantidades) siempre que funcionen correctamente.

En España, el mercado diario pertenece al tipo marginalista. En este tipo de mercados, la oferta de un generador representa la cantidad de energía que está dispuesto a vender a partir de un cierto precio mínimo. De esta manera el precio de la energía se fija de la siguiente manera:

- 1) Oferta: Una vez que los vendedores han presentado sus ofertas al mercado para cada una de las horas del día siguiente, el OMIE las agrega y ordena por precio ascendente, resultando así la curva de oferta del mercado para cada hora. En la siguiente imagen se refleja los tramos o escalones que corresponden a ofertas de centrales de la misma tecnología. A la vista de ella, es importante resaltar nuevamente que las ofertas de los vendedores reflejan sus costes de oportunidad<sup>2</sup>:

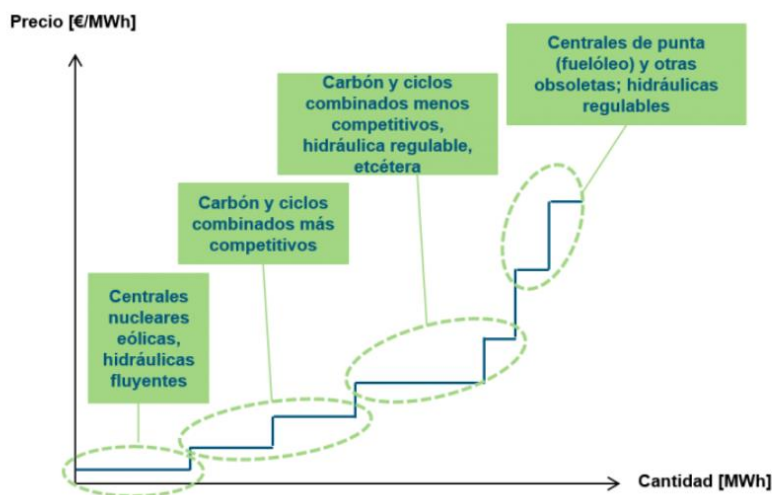


Figura 4: Agregación de la curva de oferta [8]

<sup>2</sup> Coste de oportunidad: es el precio ofertado que supone producir electricidad al generador. No es lo mismo que el coste variable. Reflejan costes en los que evitaría incurrir de optar por no producir e ingresos a los que renuncia por el hecho de producir.

- 2) Demanda: Los consumidores finales suelen clasificarse en función de la magnitud de su consumo y del fin para el que utilizan la energía. Se suele distinguir entre grandes consumidores, consumidores de tamaño medio en sectores industriales y de servicios y, finalmente, pequeños consumidores conectados a las redes de baja tensión (como los domésticos y los pequeños negocios). En la siguiente imagen se aprecia de manera esquemática como se forma la curva agregada de demanda:

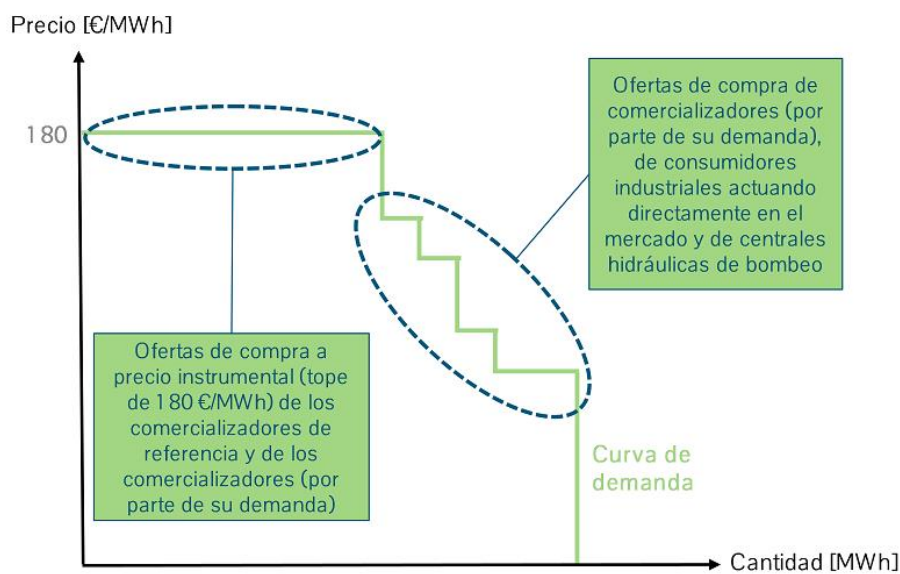


Figura 5: Agregación de la curva de demanda [8]

- 3) Casación de la curva de oferta y demanda: el precio del mercado para la hora  $h$  del día  $D$  se determina por la intersección de la curva de oferta y demanda de electricidad del mercado para esa hora. Este precio determina las ofertas de compra y de venta que resultan casadas (es decir, la energía que se intercambiará finalmente al precio del mercado). En cada hora, todas las ofertas de venta (compra) que resulten casadas reciben (pagan) el precio del mercado. La siguiente figura muestra un ejemplo de las casaciones de oferta y demanda que lleva a cabo diariamente el OMIE para cada hora del día siguiente:



CONTEXTUALIZACIÓN DEL MERCADO ELÉCTRICO ESPAÑOL

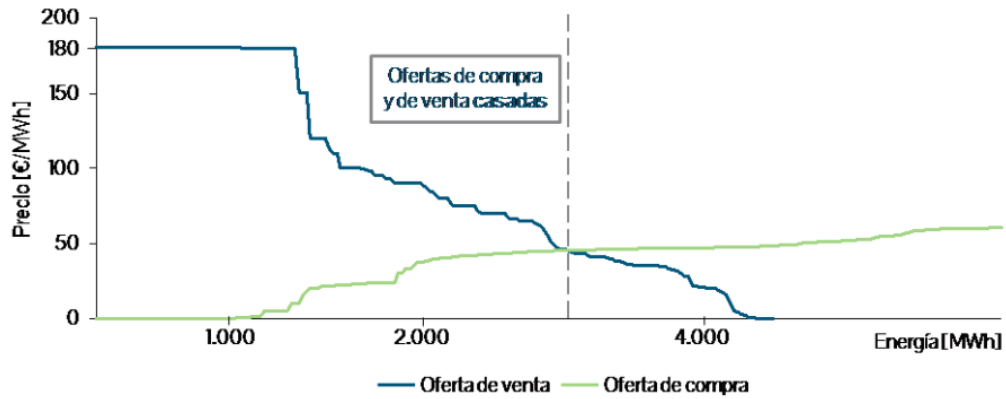


Figura 6: Casación del precio de la energía. [8]

### **3. DESARROLLO DEL PROYECTO I: CASO DE NEGOCIO.**

En este apartado se pretende contextualizar el caso de negocio que se quiere resolver y qué ventaja competitiva reporta a la empresa.

#### ***3.1 LOS SOBRECOSTES DEL MERCADO ELÉCTRICO: SERVICIOS DE AJUSTES DEL SISTEMA.***

Una de las características de la energía eléctrica es que no puede almacenarse en grandes cantidades. Esto supone que, para el correcto funcionamiento del sistema eléctrico, la producción de las centrales de generación debe igualarse al consumo de forma precisa e instantánea.

La función de Red Eléctrica, como operador del sistema, consiste en garantizar ese equilibrio en el sistema eléctrico español. Para ello, realiza las previsiones de la demanda de energía eléctrica y gestiona en tiempo real las instalaciones de generación y transporte eléctrico, logrando que la producción programada en las centrales eléctricas coincida en cada instante con la demanda de los consumidores. En el caso de que difiera, envía las órdenes oportunas a las centrales para que ajusten sus producciones aumentando o disminuyendo la generación de energía de manera que se mantengan márgenes de generación suficientes para hacer frente a posibles pérdidas sobrevenidas de generación o cambios en el consumo previsto.

Además, en el sistema eléctrico peninsular, Red Eléctrica gestiona los denominados mercados de servicios de ajuste, mercados mediante los que se adecuan los programas de producción, libremente establecidos por los sujetos en el mercado diario y mediante contratación bilateral, y posteriormente en el mercado intradiario, a los requisitos de calidad, fiabilidad y seguridad del sistema eléctrico. Se entiende por servicios de ajuste o mercados de ajuste: la solución de restricciones técnicas, la asignación de los servicios complementarios y la gestión de desvíos.

Según la normativa actual los servicios de ajuste son:

- 1) Solución de restricciones técnicas: se trata de un servicio de ajuste cuyo objetivo es resolver las restricciones técnicas que se dan en el sistema mediante la limitación y modificación de los programas de producción, generación de energía y consumo por bombeo que se resuelven las restricciones técnicas identificadas con un menor coste para el sistema. Tras la identificación de estas restricciones y modificación de los programas, se realiza un reequilibrio de la generación y demanda.
- 2) Servicios de balance gestionados por mecanismos de mercado:
  - I. Regulación secundaria: Servicio de carácter potestativo cuya finalidad es el mantenimiento del equilibrio generación -demanda, corrigiendo de manera automática los desvíos generados de las frecuencias del sistema y del programa de intercambio de energía con el bloque del sistema energético español. Este servicio es retribuido mediante mecanismos de mercado por dos conceptos: disponibilidad (banda de regulación) y utilización (energía).
  - II. Regulación terciaria. Servicio de carácter potestativo y oferta obligatoria gestionado y retribuido mediante mecanismos de mercado que tiene por objeto resolver los desvíos entre generación y consumo y restituir la reserva de regulación secundaria utilizada.
  - III. Gestión de desvíos. El mecanismo de gestión de desvíos es un servicio de carácter potestativo gestionado y retribuido mediante mecanismos de mercado que tiene por objeto resolver los desvíos entre generación y consumo que pudieran identificarse con posterioridad al cierre de cada sesión del mercado intradiario y hasta el inicio del horizonte de efectividad de la siguiente sesión. [10]

### ***3.2 VENTAJA COMPETITIVA QUE OFRECE LA MODELIZACIÓN DEL PRECIO DEL SOBRECOSTE.***

Como se ha explicado en el anterior punto debe existir un equilibrio entre generación eléctrica y demanda en tiempo real para evitar desequilibrios en el sistema, que se traducen en desvíos. Para evitar que estos desvíos interfirieran en el suministro de energía eléctrica existen mercados de desvíos regulados por el TSO (Transmission System Operator), Red Eléctrica de España donde se fija el precio de estos desvíos.

Desde el punto de vista de negocio el poder modelizar y predecir el precio de sobrecoste eléctrico supone una ventaja competitiva. Tener una idea del precio de sobrecoste permite a la empresa elaborar una estrategia de negocio más sólida; reduciendo el riesgo de pérdidas económicas, anticipándose a situaciones de incertidumbre y permitiendo elaborar un plan comercial más fiable.

La empresa en cuestión tiene una gran cartera de clientes a los que se les ofrece suministro eléctrico a precios fijos o variables en forma de contrato bilateral con un horizonte temporal a medio y largo plazo. De esta manera es de crucial importancia tener unas buenas predicciones a medio y largo plazo con un doble objetivo:

1. Ofrecer contratos más ajustados a los clientes: Los contratos a medio o largo plazo que se ofrecen a los clientes se cierran de antemano con un precio cerrado o variable que se calcula en función de una previsión del precio de la electricidad del mercado. En este precio se incluye el precio del sobrecoste, que se calcula en base a un histórico y aplicando un coeficiente en función de la creencia o intuición de la tendencia del mercado. De manera que si la empresa es conocedora con una mayor precisión del precio de la energía el margen de beneficio que se puede estimar para un cliente puede ser más fiable.
2. Mejorar la estrategia de negocios. El saber con mayor precisión a qué precio va a estar la energía permite a la empresa no solamente reducir los riesgos financieros, si no también estratégicos. Este hecho se traduce en poder diferenciarte de otras

---

*DESARROLLO DEL PROYECTO I: CASO DE NEGOCIO.*

---

compañías y poder desarrollar estrategias de negocio más seguras, por ejemplo, para planificar la cartera de clientes, estrategias de mercado en la compraventa de energía o definición de contratos bilaterales con clientes.

## **4. DESARROLLO DEL PROYECTO II: EXTRACCIÓN DE DATOS.**

En este apartado se explican que fuentes de datos se utilizaron para la extracción de información. Así mismo, se aclara que procedimiento se realizó para la limpieza y preparación de los datos extraídos para la posterior modelización del precio del sobrecoste.

### ***4.1 HORIZONTE TEMPORAL PARA LA DESCARGA DE SERIES.***

A la hora de descargar los datos se ha elegido un horizonte temporal de 3 años, comprendiendo los años 2018, 2019, 2020 y hasta mayo de 2021.

Se elige este horizonte debido a nuevas regulaciones y tendencias que se han observado desde el área de negocio:

Desde el año 2018 se puede apreciar una mayor cantidad de generación de renovables en el mix de energético<sup>3</sup>, acompañado de nuevas regulaciones ha producido que la marcada correlación negativa<sup>4</sup> que existía anteriormente entre los sobrecostes y el precio del spot diario se haya diluido.

Este nuevo escenario compromete el desafío de modular los sobrecostes en un nuevo contexto, de manera que el cálculo que se realizaba hasta ahora del precio del sobrecoste no es fiable actualmente. Es por ello por lo que surge la necesidad de modelar los sobrecostes de una manera distinta y en un contexto distinto.

---

<sup>3</sup> Mix energético: alude a la combinación de las diferentes fuentes de energía que cubren el suministro eléctrico de un país

<sup>4</sup> Correlación negativa: relación entre dos variables que muestra que una variable disminuye conforme otra aumenta.

## **4.2 EXTRACCIÓN DE DATOS DE FUENTES PÚBLICAS.**

### **4.2.1 FUENTES PÚBLICAS UTILIZADAS.**

Para la extracción de datos de origen público se utiliza la API<sup>5</sup> de ESIOS.

Red Eléctrica de España (REE) tiene como misión asegurar el funcionamiento global del sistema eléctrico español mediante dos actividades esenciales: la operación del sistema eléctrico y el transporte de electricidad en alta tensión.

Para lograr estos objetivos como OS, REE ha desarrollado un sistema de información que denomina Sistema de Información del Operador del Sistema (E-SIOS), diseñado especialmente para ejecutar los procesos que permiten la explotación segura y económica del sistema eléctrico español en tiempo real.

Para el desarrollo del proyecto a API se utiliza para recopilar los resultados de los distintos mercados y programaciones, con el fin de construir un dataframe<sup>6</sup> con las variables necesarias para poder predecir el precio del sobrecoste.

### **4.2.2 METODOLOGÍA DE EXTRACCIÓN.**

Para la extracción de datos vía API se ha desarrollado un código de Python en Jupyter Notebooks. Se han utilizado librerías como Pandas, Request, Datetime, Json... entre otras.

Para la extracción de datos vía API ESIOS es necesario indicar los siguientes parámetros:

1. Token de autorización: este token se debe pedir previamente vía email a ESIOS y es personal.

---

<sup>5</sup> API: Una API es un conjunto de definiciones y protocolos que se utiliza para desarrollar e integrar el software de las aplicaciones. API significa interfaz de programación de aplicaciones. Las API permiten que sus productos y servicios se comuniquen con otros, sin necesidad de saber cómo están implementados.

<sup>6</sup> Dataframe: es una estructura de datos con dos dimensiones en la cual se puede guardar datos de distintos tipos (como caracteres, enteros, valores de punto flotante, factores y más) en columnas.

*DESARROLLO DEL PROYECTO II: EXTRACCIÓN DE DATOS.*

---

2. Horizonte temporal: se debe apuntar fecha de inicio y fin para descargar las series de precios y variables que se quieran tratar.
3. Formato de descarga: se debe indicar en que formato se descargan los datos. En el proyecto desarrollado se descargan en formato JSON. Una vez descargada las series en formato JSON, se realizan una serie de transformaciones para concatenar todas las series descargadas y aunarlas en un único dataframe.

#### **4.2.3 VARIABLES RECOPIADAS.**

Las variables descargadas mediante la API de ESIOS son principalmente las que se encuentran en dos programas de servicio:

1. Programa P48: Es el programa operativo de unidades de programación correspondientes a ventas y adquisiciones de energía en el sistema eléctrico peninsular español que el OS establece en cada periodo de programación hasta el final del horizonte diario de programación. El programa horario operativo incorporará todas las asignaciones y redespachos de programa aplicados por el OS hasta su publicación, 15 minutos antes del inicio de cada hora.
2. Programa PBF: Es el programa de energía diario, con desglose por periodos de programación, de las diferentes unidades de programación correspondientes a ventas y adquisiciones de energía en el sistema eléctrico peninsular español. Este programa es establecido por el OS a partir del programa resultante de la casación del mercado diario comunicado por el OM, y la información de ejecución de contratos bilaterales con entrega física comunicada de acuerdo con lo establecido en el presente procedimiento de operación.

Estos programas son los más representativos a la hora de poder calcular y predecir el precio del sobrecoste.

Las series descargadas mediante la API de ESIOS de manera horaria son:

- Generación programada p48 total (MW): energía correspondiente al programa operativo que el OS establece en cada período hasta el final del horizonte de



*DESARROLLO DEL PROYECTO II: EXTRACCIÓN DE DATOS.*

---

programación diario. Su desglose muestra la energía programada para los distintos tipos de producción de generación.

- Generación programada P48 otras renovables (MW).
- Generación programada P48 Ciclo combinado (MW).
- Generación programada P48 Eólica (MW).
- Generación programada P48 Carbón (MW).
- Generación programada P48 Hidráulica UGH: este indicador se refiere a las unidades de programación con tipo de producción de energía hidráulica regulada (MW).
- Generación programada P48 Hidráulica UGH: este indicador se refiere a las unidades de programación con tipo de producción de energía hidráulica no regulada (MW).
- Generación programada P48 Nuclear (MW).
- Generación programada P48 Residuos no renovables (MW).
- Generación programada P48 Solar fotovoltaica (MW).
- Generación programada P48 Solar térmica (MW).
- Generación programada P48 Turbinación bombeo (MW).
- Ajuste de programas p48 (MW): El desglose de este indicador muestra la energía programada para los conceptos Corrección eólica y Corrección solar fotovoltaica.
- Demanda programada P48 total (MW): Su desglose muestra la energía programada de todos los tipos de oferta de demanda.
- Generación programada P48 Consumo bombeo (MW).
- Generación programada P48 Enlace Baleares (MW).
- Saldo total interconexiones programa p48 (MW): su desglose muestra la energía programada en exportaciones e importaciones de todas las interconexiones con España.
- Precio mercado SPOT<sup>7</sup> Diario (€/MWh): El Mercado Diario es el mecanismo integrante del mercado de producción de energía eléctrica que tiene por objeto llevar

---

<sup>7</sup> Precio Spot: El precio spot o precio corriente de un producto, de un bono o de una divisa es el precio que es pactado para transacciones (compras o ventas) de manera inmediata.

---

*DESARROLLO DEL PROYECTO II: EXTRACCIÓN DE DATOS.*

---

a cabo las transacciones de energía eléctrica para el día siguiente, mediante la presentación de ofertas de venta y adquisición de energía eléctrica por parte de los sujetos del mercado diario.

- Precio medio horario componente restricciones PBF comercializadores de referencia (€/MWh).
- Precio medio horario componente restricciones tiempo real comercializadores de referencia (€/MWh).
- Precio medio horario componente reserva de potencia adicional a subir comercializadores de referencia (€/MWh).
- Precio medio horario componente banda secundaria comercializadores de referencia (€/MWh).
- Precio medio horario componente desvíos medidos comercializadores de referencia (€/MWh).
- Precio medio horario componente incumplimiento energía de balance (€/MWh).
- Generación programada PBF Ciclo combinado (MW).
- Generación programada PBF Hidráulica UGH (MW).
- Requerimientos Banda de regulación secundaria a subir (MW).
- Requerimientos Banda de regulación secundaria a bajar (MW).

### **4.3 EXTRACCIÓN DE DATOS DE FUENTES INTERNAS DE ENGIE.**

#### **4.3.1 FUENTES INTERNAS UTILIZADAS.**

Para la extracción de series específicas del mercado eléctrico que no se podían obtener de fuentes públicas se utilizaron fuentes internas de la empresa.

Los datos que se podían obtener de esta fuente son proveídos por empresas externas y hay varias maneras de interactuar con la aplicación que recoge todas las series. Además, para acceder a la aplicación debes tener los permisos internos pertinentes.

---

*DESARROLLO DEL PROYECTO II: EXTRACCIÓN DE DATOS.*

---

La forma de interacción con la aplicación que se ha realizado es mediante una integración de la aplicación mediante un API a la que te conectas con un código de Python. Este código no se puede mostrar por motivos de confidencialidad con la empresa.

### **4.3.2 VARIABLES RECOPIADAS.**

Las variables recogidas mediante la API interna de la empresa son tres y son descargadas de manera horaria, diaria y mensual. Las variables son las siguientes:

- Precio Mercado SPOT de Francia (€/MWh). Esta variable se recopila de manera horaria.
- Precio del CO2 (€/t): El precio de emitir gases de efecto invernadero es una herramienta económica destinada a integrar en los precios de mercado los costes ocultos de los daños causados por las emisiones de estos gases, con el fin de orientar las decisiones de los agentes económicos hacia soluciones de bajas emisiones. Esta variable se recopila de manera mensual.
- Precio D+1<sup>8</sup> del gas (€/MWh): se trata de un precio a futuro del gas natural. El regulador del precio es MIBGAS, que es el responsable de la gestión del Mercado Organizado de Gas, tanto de su correcto y adecuado funcionamiento como de la gestión económica de los servicios que oferta. Esta variable se recopila de manera diaria.

---

<sup>8</sup> Precio Forward: un forward, como instrumento financiero derivado, es un contrato a largo plazo entre dos partes para comprar o vender un activo a precio fijado y en una fecha determinada.

## **4.4 LIMPIEZA Y PREPARACIÓN DE DATOS.**

### **4.4.1 CREACIÓN DE NUEVAS VARIABLES.**

A la hora de modelizar el precio de los sobrecostes algunas variables se pueden simplificar como combinación lineal de otras y se pueden crear otras variables temporales que ayuden a la modelización. Las variables temporales creadas son las siguientes:

- Datetime: se trata de la fecha, más la hora, más la zona horaria que se descarga con el siguiente formato desde la API; 2021-05-31 00:00:00+0000 (ejemplo).
- Fecha: con el formato %Y-%m-%d.
- Año: con el formato %Y.
- Mes: con el formato %m
- Día: con un valor entero, siendo 0 el lunes, 1 el martes, [...], 6 el domingo.
- Fin de semana: diferenciando entre días entre semana con un valor entero; 0 los días de lunes a viernes, 1 el sábado y 2 el domingo.
- Hora: con el formato %H
- Business Hour: diferenciando entre horas con luz y sin luz; 0 de la hora 7 a las 21 y 1 de la hora 22 a las 6.
- Quarter: se trata de una variable por cuatrimestres.

Las variables combinación lineal de otras series descargadas son las siguientes:

- Precio sobrecoste (€/MWh): se trata de la variable que se trata de predecir, la variable respuesta. La formación del precio del sobrecoste está compuesta de la suma de las siguientes variables con la siguiente fórmula;

*Precio Sobrecoste*

- = Precio componente restricciones PBF
- + Precio medio horario componente restricciones tiempo real
- + Precio medio horario componente reserva de potencia adicional a subir
- + Precio medio horario componente banda secundaria
- + Precio medio horario componente desvíos medidos
- + Precio medio horario componente incumplimiento energía de balance

- Hueco térmico (MWh): es la parte de la demanda que no queda cubierta por estas tecnologías y que ha de ser cubierta con la generación de electricidad a partir de térmicas convencionales y ciclos combinados. La fórmula que se ha utilizado para representar el hueco térmico es:

*Hueco Térmico*

- = Demanda programada P48 total
- (Generación programada P48 Eólica
- + Generación programada P48 Hidráulica UGH
- + Generación programada P48 Hidráulica no UGH
- + Generación programada P48 Nuclear
- + Generación programada P48 Solar fotovoltaica
- + Generación programada P48 Solar térmica)

- Clean Spark Spread: representa el margen unitario de un generador de ciclo combinado en base al precio del spot diario, precio del gas y precio por derechos de emisión de carbono. Las constantes que aparecen en la fórmula genérica dependen de diferentes factores particulares de cada planta como pueden ser eficiencia, costes de mantenimiento, estructuras de peaje contratada, impuestos repercutidos, ... La fórmula usada no se puede mostrar por motivos de confidencialidad con la empresa.

$$CSS = a * Precio Spot + b * Precio gas + c * Precio CO_2 + K$$

#### **4.4.2 TRANSFORMACIÓN DE DATOS.**

A la hora de crear el dataframe se han tenido en cuenta varios factores derivados de la descarga de datos para poder trabajar en la posteriormente en la modelización. Uno de los problemas más habituales en la limpieza de datos se debe a los datos nulos:

- Los datos nulos encontrados en la descarga de las series vía API ESIOS son en las series de generación de solar, carbón, bombeo y ciclo combinado. Estos valores nulos corresponden a valores en los que no se ha producido generación de estas tecnologías, por ello esos valores nulos den ser sustituidos por ceros.

El descargar los datos con formatos temporales distintos puede suponer un problema a la hora de conformar el dataframe, por lo que los datos extraídos tendrán que ser tratados previamente a introducirlos en un modelo:

- Huso horario: esta variable también nos indica si estamos en horario de verano o de invierno. El día que se realiza cambio de hora se debe tener un especial cuidado. De manera que el día que hay cambio de hora de verano a invierno debe haber 25 horas y el día que se cambia la hora de invierno a verano habrá 23 horas.
- Series horarias, diarias, mensuales: el join<sup>9</sup> entre las distintas series debe ser con un formato estandarizado, en nuestro caso las series diarias y mensuales las replicaremos de manera horaria para poder conformar un dataframe común con todos los datos horarios.
- Dataframe diario: el tener un histórico de tantos datos permite crear un dataframe de series diarias, que puede ser de ayuda a la hora de modelizar el precio de los sobrecostes.

---

<sup>9</sup> Join: Los JOINS son una función (SQL) sirven para combinar filas de dos o más tablas basándose en un campo común entre ellas, devolviendo por tanto datos de diferentes tablas.

## 5. DESARROLLO DEL PROYECTO III: ANÁLISIS EXPLORATORIO DE DATOS.

En este apartado se detalla un análisis exploratorio preliminar de los datos extraídos de las fuentes con el objetivo de garantizar la calidad del dato, su fiabilidad y mostrar una contextualización real del problema basada en estadísticas básicas.

### 5.1 ANÁLISIS PRELIMINAR DE LAS SERIES.

Para el análisis preliminar de las series descargadas y que componen el dataframe creado, se han realizado distintas gráficas de interés para el área de negocio con ayuda de la librería Plotly de Python. Este análisis preliminar servirá para analizar días concretos o tendencias de determinadas series de precios o generaciones.

Se ha realizado un gráfica de tarta del mix energético para días concretos, la siguiente figura muestra un ejemplo:

Mix de Energia

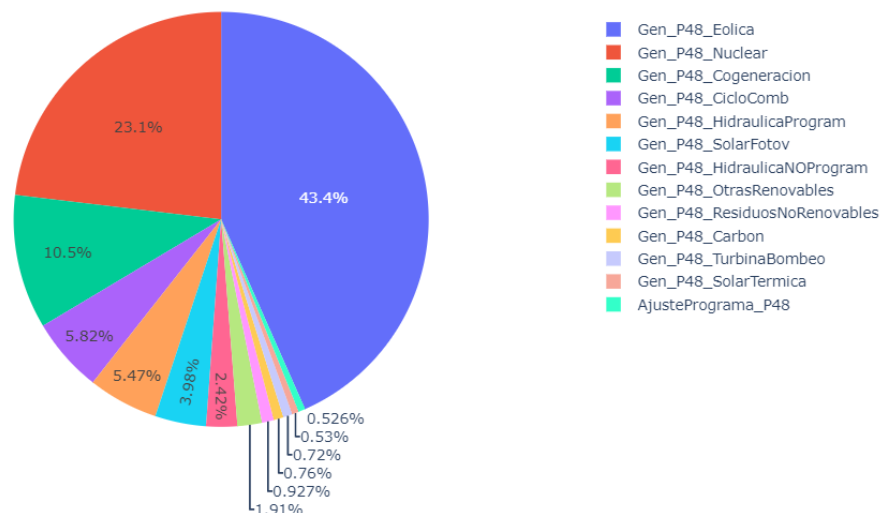
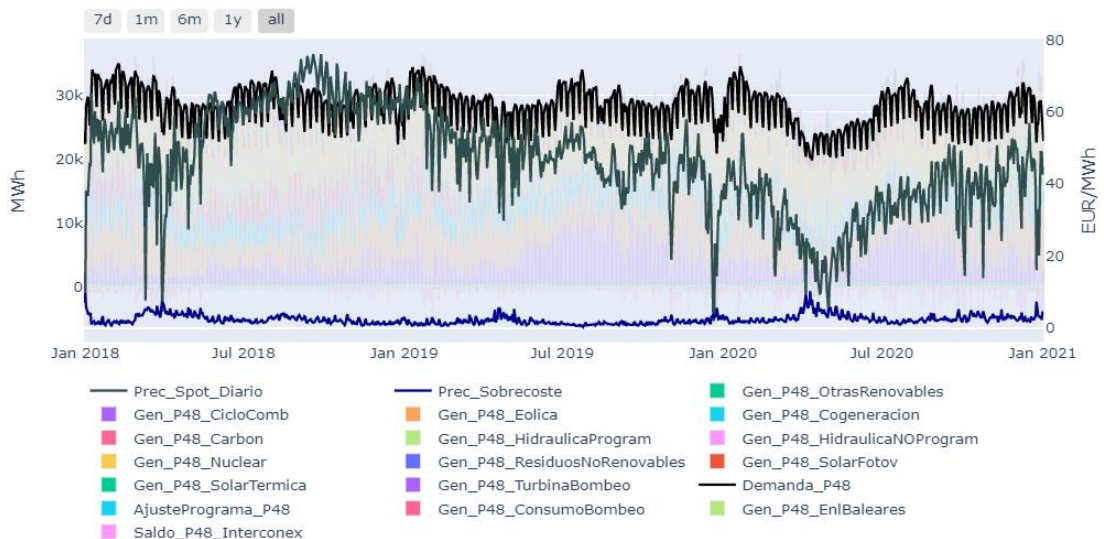


Figura 7: Mix energético diario, fecha: 2020-10-25.

*DESARROLLO DEL PROYECTO III: ANÁLISIS EXPLORATORIO DE DATOS.*

Gracias a esta gráfica se podrá tener conocimiento del mix energético hubo un día en el que una serie de precio se comporte de manera anómala y poder comprender mejor el origen de la anomalía.

Se ha analizado el mix energético comparado con la demanda y el spot diario. De esta manera se podrá apreciar si la generación total coincide con la demanda y en tendencia que influencia tiene con el precio del spot diario. En la siguiente imagen se aprecia de manera general la gráfica realizada con la librería Plotly, en el eje X1 (MWh) se muestran las generaciones y en el eje X2 las series de precios:



*Figura 8. Comparación diaria mix energético, demanda y spot diario.*

Si nos fijamos más en detalle en la figura 7 (como se mostrará en la siguiente figura 8), el nivel de demanda es siempre muy similar al nivel agregado de todas las generaciones que deben cubrir la demanda. De esta manera, se asegura que estamos representando correctamente el contexto del mercado eléctrico para la modelización de la serie del sobre coste.

Gracias a las funciones de la librería plotly se puede desplazar a lo largo del histórico de la gráfica para obtener información sobre determinados intervalos de tiempo (últimos 7 días,



*DESARROLLO DEL PROYECTO III: ANÁLISIS EXPLORATORIO DE DATOS.*

último mes, últimos 6 meses, último año). En la siguiente figura se muestra de manera detallada la comparativa para el último mes:

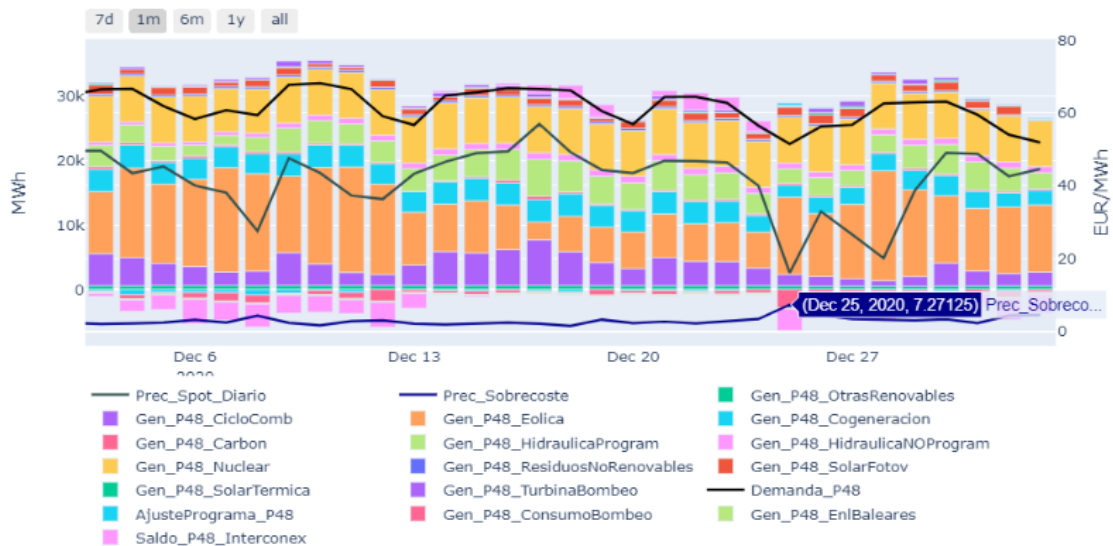
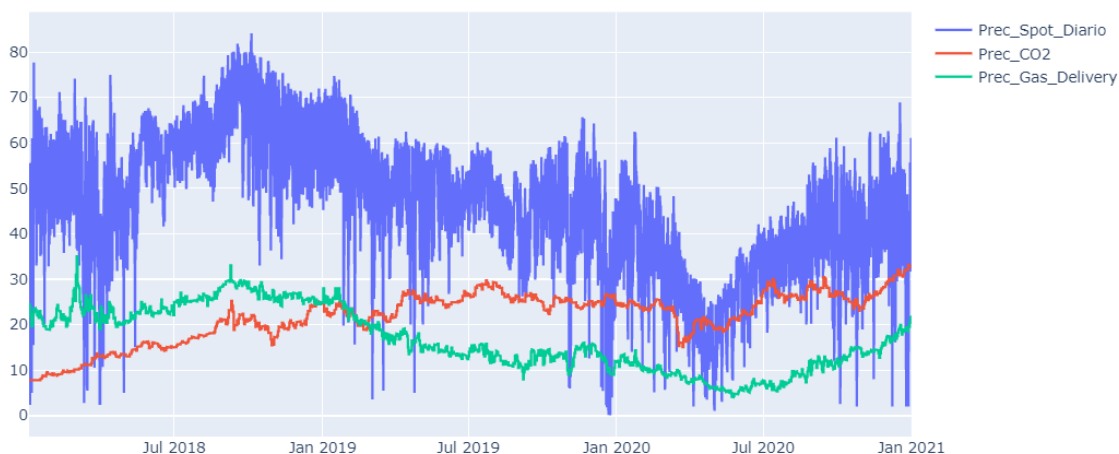


Figura 9: Comparación diaria mix energético, demanda y spot diario del mes de diciembre de 2020.

Por ejemplo, un análisis que se puede extraer de la gráfica representada en la figura 8 es el siguiente: se puede apreciar que el día 25 de diciembre de 2020 hubo un máximo local en los sobrecostes con un precio medio asociado de 7,27 euros. Este día podemos apreciar que la generación eólica cubrió una gran parte de la demanda. Además, se aprecia que hay un mínimo local en la serie del spot medio diario de 16,03 euros, un mínimo local en la demanda media con 22.52 MW. Añadir, que fue un día en el que hubo bastante exportación de energía y se consumió bastante energía por bombeo, es decir, se generó más energía de la que se demandaba.

Por último, para el análisis personalizado de las series se ha desarrollado un gráfico horario de Plotly que permite comparar 3 variables a la vez (para facilitar la distinción entre series). En la siguiente figura se muestra una comparación entre los precios del spot diario, el precio de emisiones de CO2 y el precio del gas:

Hourly Serie



*Figura 10: Comparación de 3 series de precios en el histórico de datos de manera horaria.*

Un primer análisis llamativo que se puede apreciar de la comparación de las tres series en la figura 9 es el descenso repentino del precio de emisiones del CO2 entorno al mes de marzo de 2020, coincidiendo con una bajada acusada del del precio del spot producido por una bajada de la demanda energética debido al confinamiento por la pandemia del COVID-19.

Otro de los análisis que se pueden apreciar de la comparación de estas tres series es la bajada acusada de precios que se produce de manera puntual los primeros meses del año 2019 que coincide con el sobrepaso del precio de emisiones de CO2 al precio del gas, debido a las nuevas regulaciones que favorecen a las generaciones energéticas “limpias”. En la siguiente figura derivada de hacer “zoom” en la anterior gráfica se aprecia con mayor detalle:

Hourly Serie

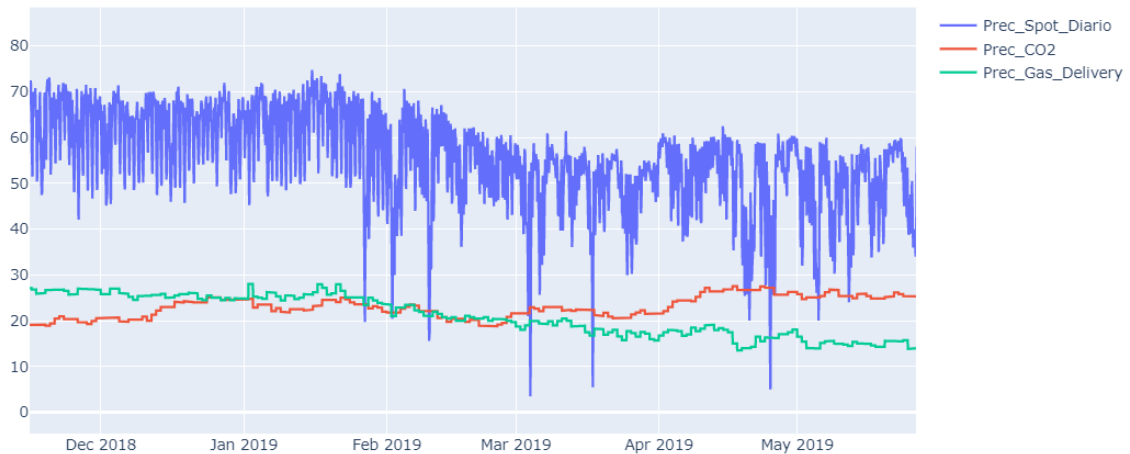


Figura 11: “Zoom” comparativo de series sobre los primeros meses de 2019.

## 5.2 ANÁLISIS DE RELACIONES.

Para el análisis de relaciones entre variables se han realizado 3 gráficas de correlaciones que nos dieran información de las variables que estaban altamente correlacionadas más del 70% tanto positiva como negativamente. De entre estas variables correlacionadas se elegirá la que se considere que puede aportar más información a la hora de predecir los sobrecostos con ayuda y opinión de expertos en el negocio.

Las gráficas de correlación se han definido de manera horaria, diaria y mensual. En el conjunto de datos horarios es donde menos correlaciones significativas se han encontrado y en el conjunto de datos mensual es dónde más correlaciones significativas se han encontrado. Las siguientes figuras muestran las tres matrices de correlaciones como mapas de calor:

DESARROLLO DEL PROYECTO III: ANÁLISIS EXPLORATORIO DE DATOS.

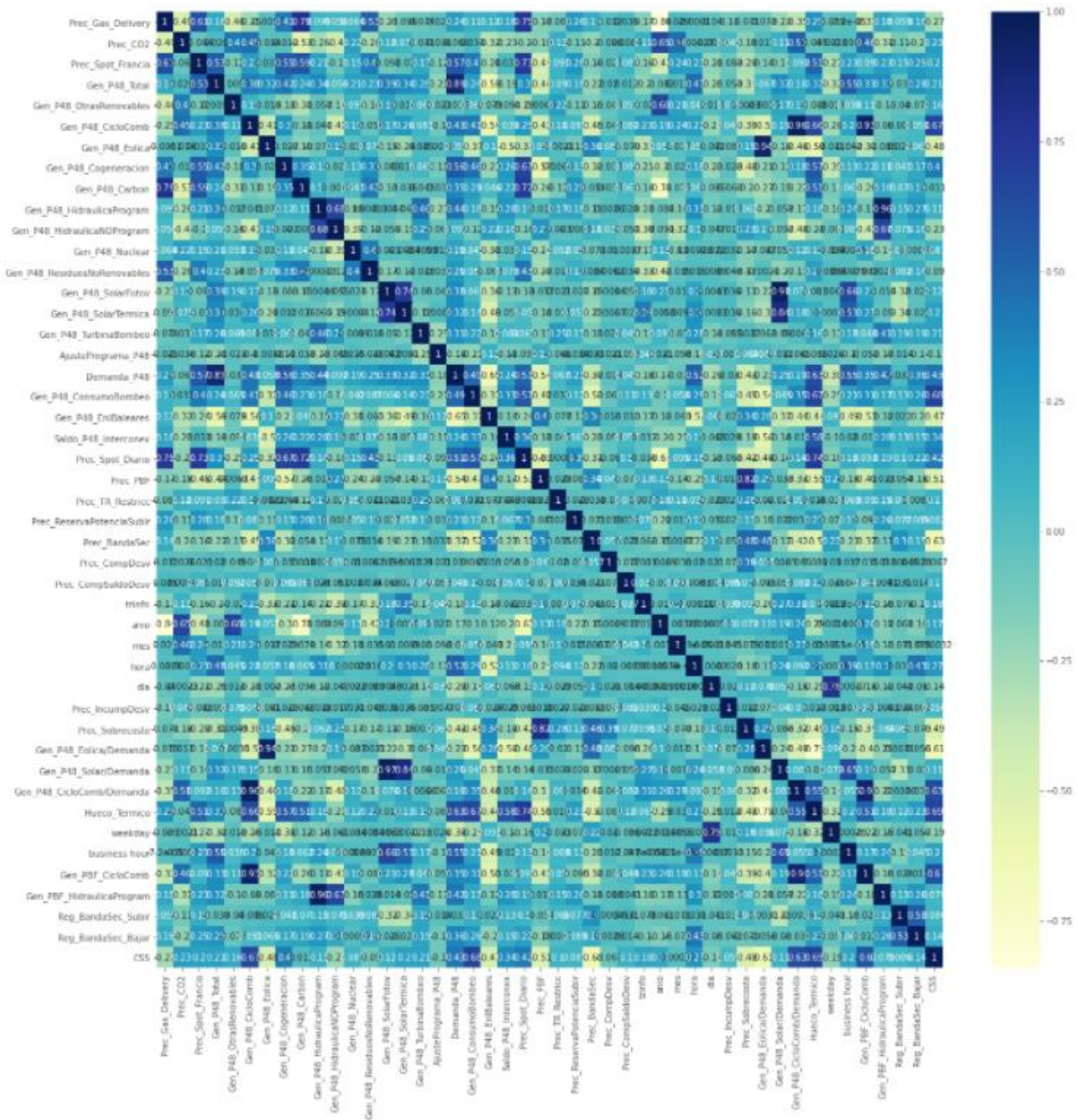


Figura 12. Matriz de correlaciones horarias.

DESARROLLO DEL PROYECTO III: ANÁLISIS EXPLORATORIO DE DATOS.

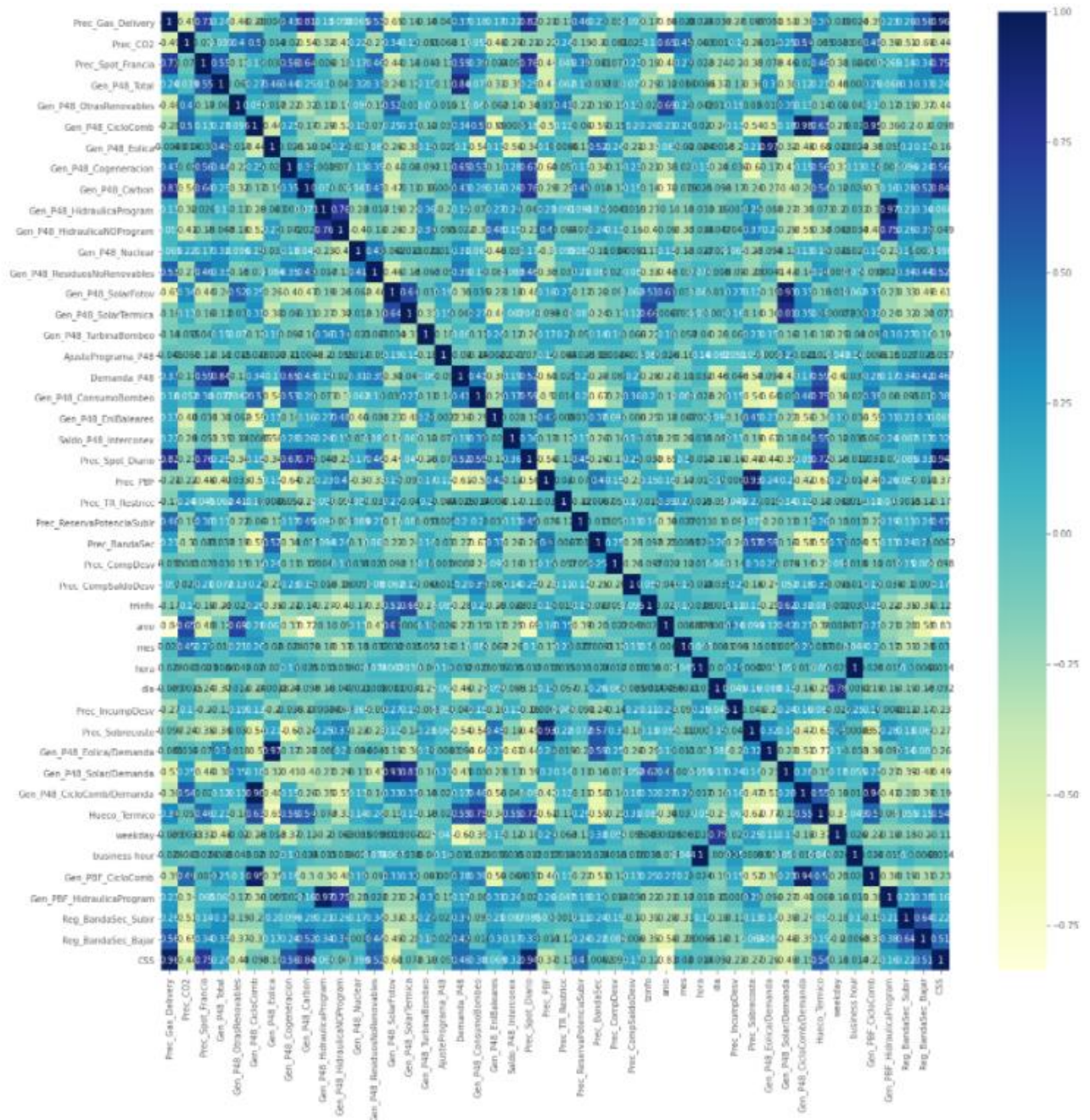


Figura 13: Matriz de correlaciones diarias.

DESARROLLO DEL PROYECTO III: ANÁLISIS EXPLORATORIO DE DATOS.

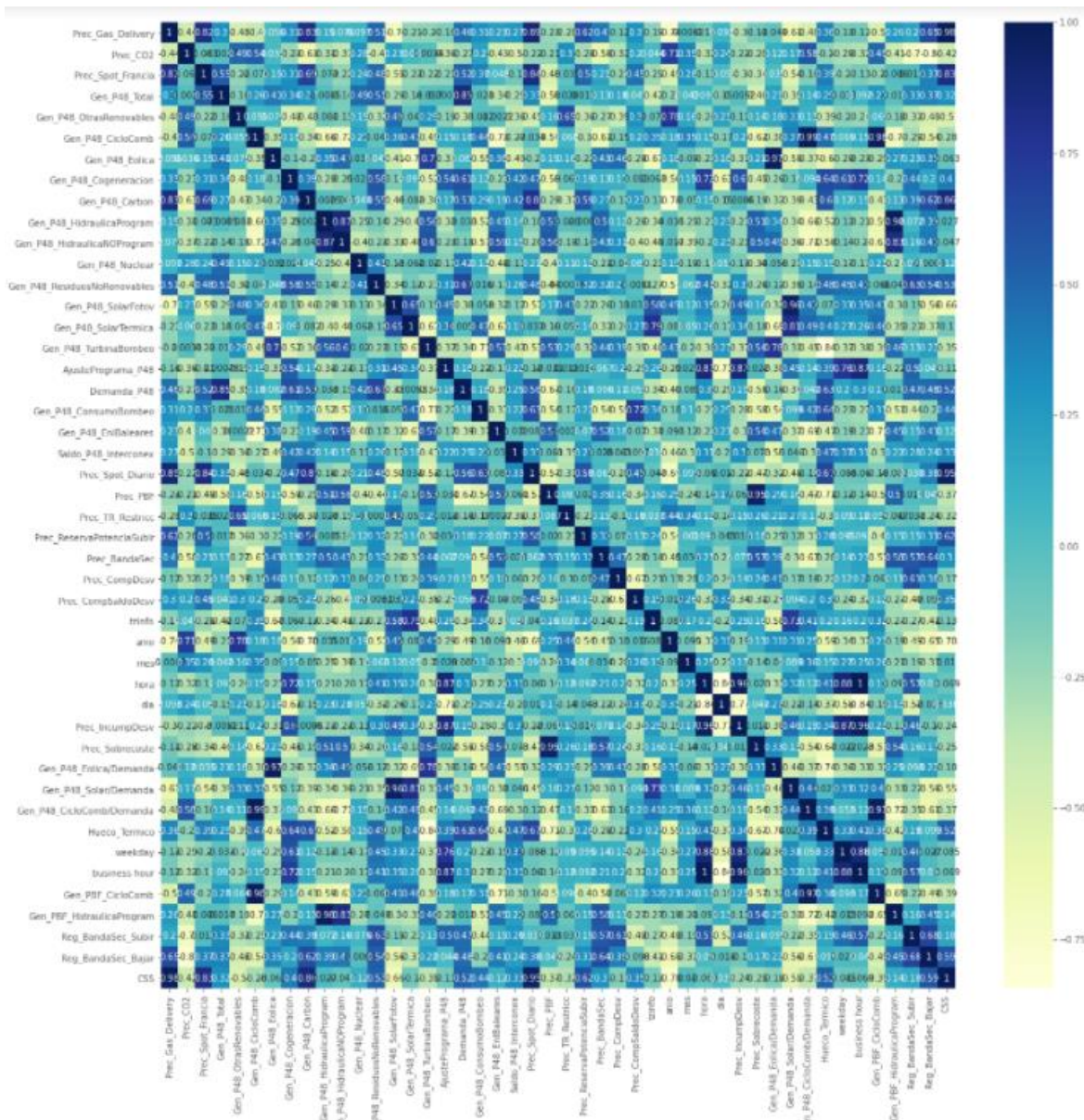


Figura 14: Matriz de correlaciones mensuales.

Debido al gran número de variables que tenemos para analizar, estudiar visualmente las variables correlacionadas obteniendo las matrices de correlación es complejo. Por ello se decide filtrar las variables más correlacionadas obteniendo los siguientes resultados en forma de tabla:

*DESARROLLO DEL PROYECTO III: ANÁLISIS EXPLORATORIO DE DATOS.*

*Tabla 1: Correlaciones entre variables más relevantes.*

Variable 1	Variable 2	Correlación
Precio del gas	Año	-80%
Generación de carbón	Año	-70%
Precio del gas	Precio Spot de Francia	70%
Hueco térmico	Precio Spot diario	70%
Hueco térmico	Generación de Consumo de bombeo	75%
Generación Hidráulica regulable	Generación no Hidráulica regulable	75%
Generación de carbón	Precio Spot diario	75%
Precio Spot diario	Precio Spot de Francia	75%
Generación de carbón	Precio del gas	80%
Precio Spot diario	Precio del gas	80%
Demanda	Generación total	85%

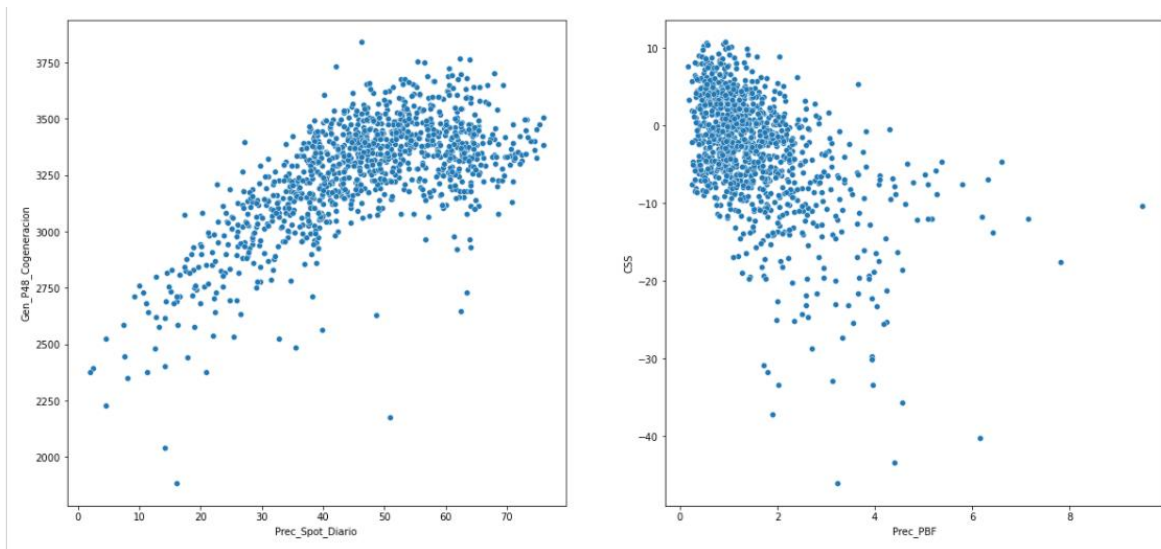
Las conclusiones que se han extraído con ayuda de opiniones de expertos son las siguientes:

- Las variables de Precio de gas, generación de carbón, precio spot diario, generación de consumo por bombeo y generación total son variables que pueden ser suprimidas como posibles variables significativas para la modelización por su alta correlación con otras variables.

*DESARROLLO DEL PROYECTO III: ANÁLISIS EXPLORATORIO DE DATOS.*

- En su defecto, las variables de año, precio de spot de Francia, Demanda total y hueco térmico son variables que se estudiarán como posibles entradas al modelo.
- Se creará una variable de generación hidráulica total como suma de la generación hidráulica regulable y no regulable.

Por otro lado, se ha realizado un análisis de las relaciones en formato diario entre pares de variables con un pairplot. Las relaciones que se han considerado más significantes por el grupo de expertos se muestran a continuación:

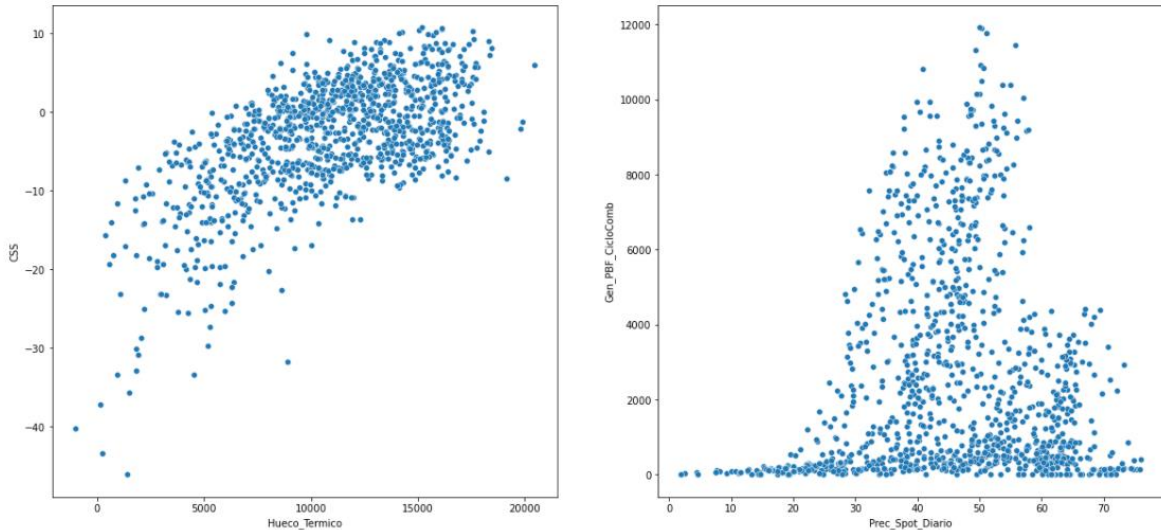


*Figura 15: Relación entre Precio Spot y Generación por Cogeneración (izquierda) y entre CSS y Precio medio horario componente restricciones PBF (derecha).*

Se puede apreciar una cierta relación cuadrática entre la cogeneración y el precio del spot diario. En cuanto al valor del CSS parece que se encuentra casi siempre en valores entre 10 y -10 y que estos valores se dispersan muy negativamente en pocas ocasiones siempre con valores de precio de restricciones de PBF mayores de 2 euros, el valor del CSS parece difícil de explicar con el valor del precio de restricciones PBF (principal componente del precio de sobrecoste).

Otras relaciones que se han considerado relevantes para el estudio preliminar de los datos son las que se muestran en la siguiente figura:





*Figura 16: Relación entre CSS y Hueco térmico (izquierda) y entre Generación de Ciclo combinado y Precio Spot (derecha).*

Se puede apreciar una cierta correlación entre el CSS y el hueco térmico, debido a que cuanto mayor será el hueco térmico más demanda queda sin cubrir por parte de las tecnologías renovables el CSS aumenta lo que significa que es más rentable producir para los ciclos combinados. Por otro lado, podemos apreciar una cierta relación entre la generación de ciclo combinado y el precio spot diario; la generación de ciclo combinado solo produce de manera relevante cuando el precio spot está entre los 30 y 70 euros. Dentro de esta franja se diferencian dos más entre los 30 y 50 euros la generación de ciclo varía entre los 0 y los 12 MWh, mientras que en la horquilla entre los 50 y 70 euros la generación de ciclo varía entre los 0 y los 4 MWh.

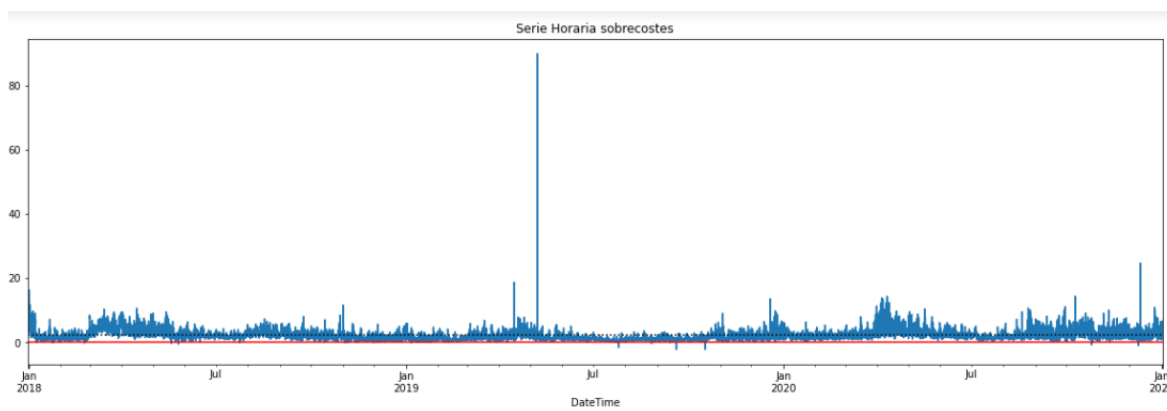
### **5.3 ANÁLISIS DEL PRECIO DE SOBRECOSTE.**

#### **5.3.1 ANÁLISIS DE LA SERIE.**

Se considera que para realizar una buena predicción del precio del sobrecoste hay que entender bien el perfilado de la serie y tratar de encontrar relaciones o estadísticas básicas que ayuden a la comprensión de la serie. En las siguientes imágenes observamos la forma

*DESARROLLO DEL PROYECTO III: ANÁLISIS EXPLORATORIO DE DATOS.*

de la serie del precio del sobrecoste de manera horaria, diaria y mensual para los últimos tres años (2018,2019 y 2020).



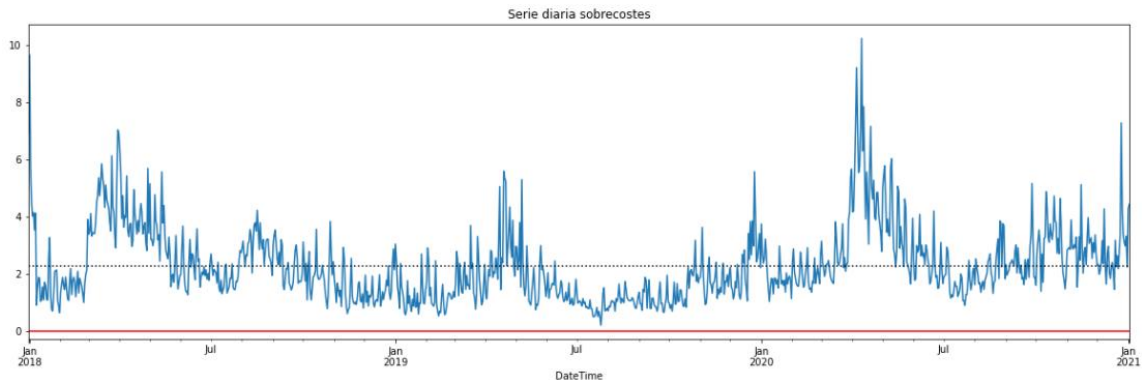
*Figura 17: Representación horaria de la serie de precios de sobrecostos (2018-2020).*

De manera horaria llama la atención dos aspectos a destacar:

1. Valores negativos: esto se debe a que a ciertas horas las variables de *Precio medio horario componente desvíos* y *Precio medio horario componente incumplimiento energía de balance* pueden tener valores muy negativos en comparación con las otras variables que componen el precio.
2. Valores radicalmente altos: es el caso del 7 de mayo de 2019 entre las 20 y 21 horas, cuando se dio un precio de sobrecoste de casi 90 euros/MWh. La razón fue la coincidencia del fallo de un generador térmico con el desvío de producción eólica respecto de lo previsto y un adelanto de la entrada de la punta de consumo nocturna, lo que requirió que se utilizase la práctica totalidad de la reserva de potencia rápida del sistema [11]. Esto se traduce en una subida de precio anómala de la variable *Precio medio horario componente desvíos*.

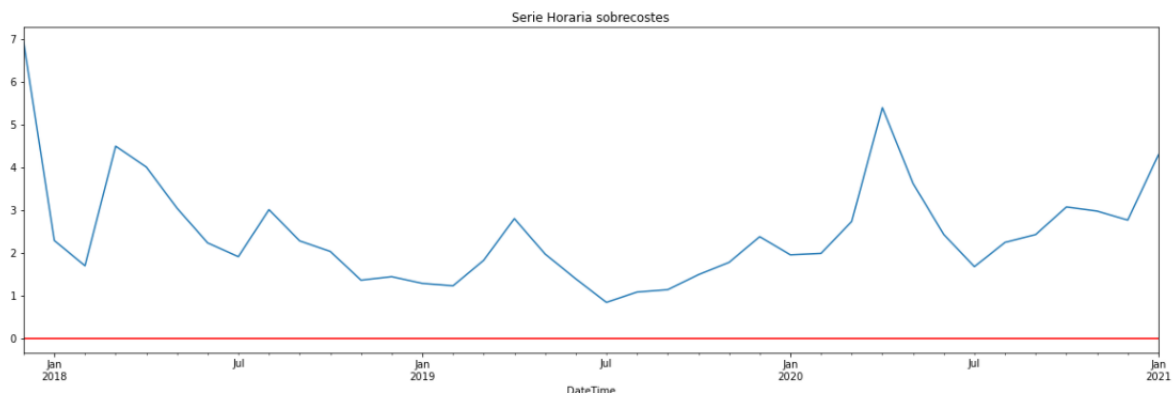
De manera diaria (realizando un agregado diario) como se puede apreciar en la siguiente figura no tenemos ningún valor negativo y la media de precios se encuentra en torno a los 2,5 euros/MWh:

*DESARROLLO DEL PROYECTO III: ANÁLISIS EXPLORATORIO DE DATOS.*



*Figura 18: Representación diaria de la serie de precios de sobrecostes (2018-2020).*

De manera mensual (figura 18) se puede apreciar que son los meses de marzo y abril los que suelen tener sobrecostes más altos. Esto se explica porque es durante estos meses dónde suele haber una mayor generación de tecnologías renovables, es decir, existe un menor hueco térmico. Por este motivo los ciclos y tecnologías más flexibles para arrancar y cubrir esa demanda lo hacen a precios más altos. Es por ello por lo que los servicios de ajuste durante esos meses son más caros elevando el precio de los sobrecostes.



*Figura 19: Representación mensual de la serie de precios de sobrecostes (2018-2020).*

Cabe señalar el efecto que tuvo la pandemia del COVID-19 en los sobrecostes. Para ello, es de gran ayuda la siguiente gráfica; que muestra una comparativa entre la demanda, precio del spot y precio del sobrecoste:

*DESARROLLO DEL PROYECTO III: ANÁLISIS EXPLORATORIO DE DATOS.*

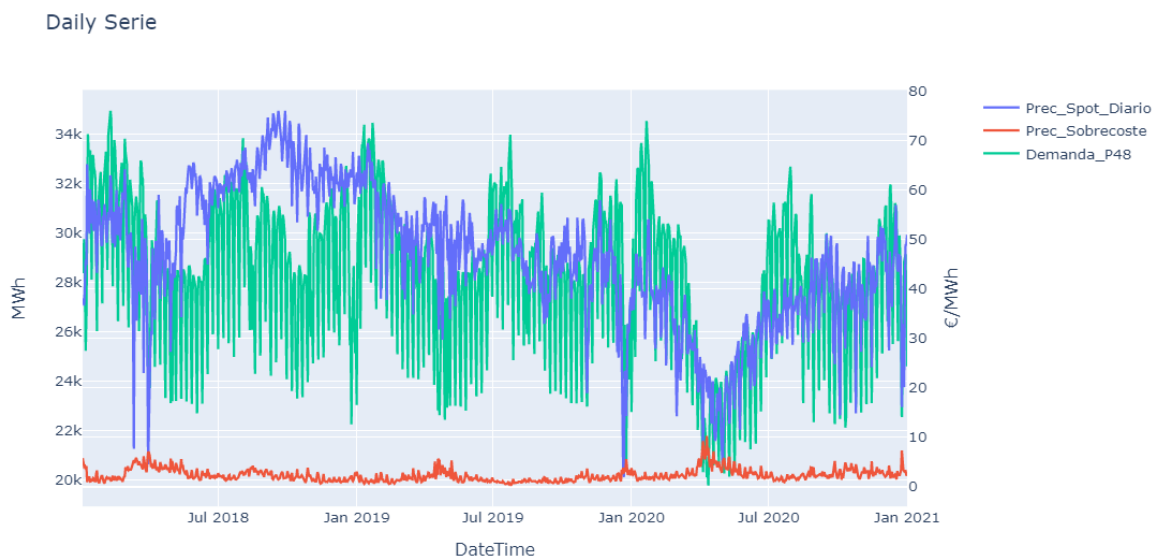


Figura 20: Comparativa de las series de Precio de sobrecoste, Precio Spot diario y Demanda (2018-2020).

De manera global podemos apreciar que es en los meses de marzo y abril de 2020 cuando hay un efecto notorio de bajada de demanda y de spot diario y una subida del precio del sobrecoste. En la siguiente imagen se hace “zoom” sobre las fechas en cuestión.

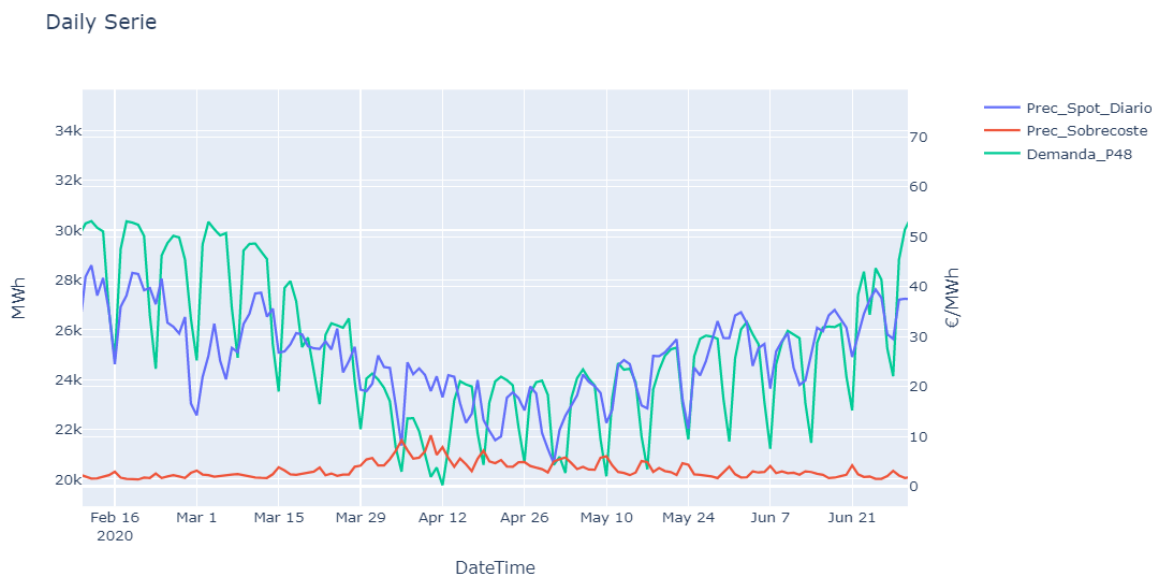


Figura 21: “Zoom” de la comparativa de las series de Precio de sobrecoste, Precio Spot diario y Demanda (Feb 2020-Junio2020).

*DESARROLLO DEL PROYECTO III: ANÁLISIS EXPLORATORIO DE DATOS.*

---

Se observa que durante el confinamiento derivado por la pandemia del COVID la acusada bajada de la demanda y del precio del spot influye en la subida del precio del sobreajuste. Esto se debe a que como se ha explicado con anterioridad si la demanda baja y las generaciones de tecnologías renovables se mantienen en valores normales para esos meses el hueco térmico disminuirá. De esta manera será más caro para las centrales térmicas (ciclos combinados) arrancar y esto supondrá un aumento del precio en el mercado de ajustes.

### **5.3.2 OUTLIERS.**

Un análisis básico previo es el de outliers o valores atípicos de la variable respuesta del futuro modelo.

Se define como outlier aquellos valores que se encuentran fuera del rango intercuartílico, de manera estadística expresado como:

$$q < Q_1 - 1.5 * IQR$$

*o*

$$q > Q_3 + 1.5 * IQR$$

*siendo*

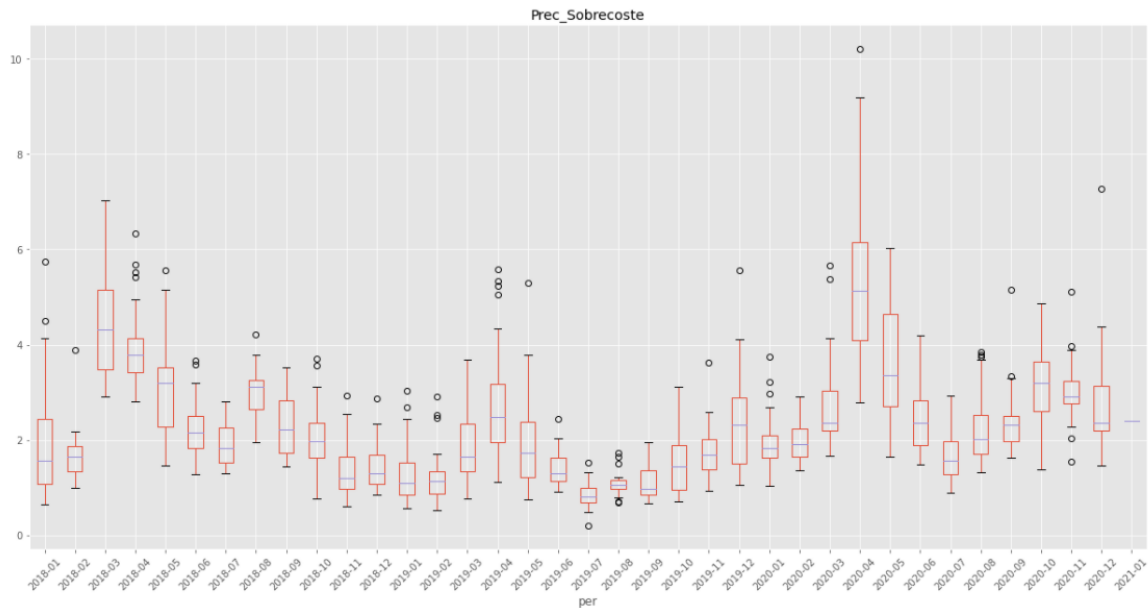
$$IQR = Q_3 - Q_1$$

*siendo*

$$Q_1 = 1 * \frac{N + 1}{4} \text{ y } Q_3 = 3 * \frac{N + 1}{4}$$

En la siguiente imagen se puede apreciar la aparición de valores atípicos de precios de sobrecostos de manera mensual durante el histórico de tres años desde 2018 hasta 2020.

*DESARROLLO DEL PROYECTO III: ANÁLISIS EXPLORATORIO DE DATOS.*



*Figura 22: Distribución de outliers de precios de sobrecostes de manera mensual (2018-2020).*

Se observa que mensualmente no hay muchos valores atípicos (1 o 2 por mes de media) y prácticamente todos son valores atípicos superiores. El mes con un rango intercuartílico mayor y que muestra una mayor variabilidad de datos y con valores bastante elevados en medio respecto al resto de meses es marzo de 2020 (primer mes de confinamiento por la COVID-19).

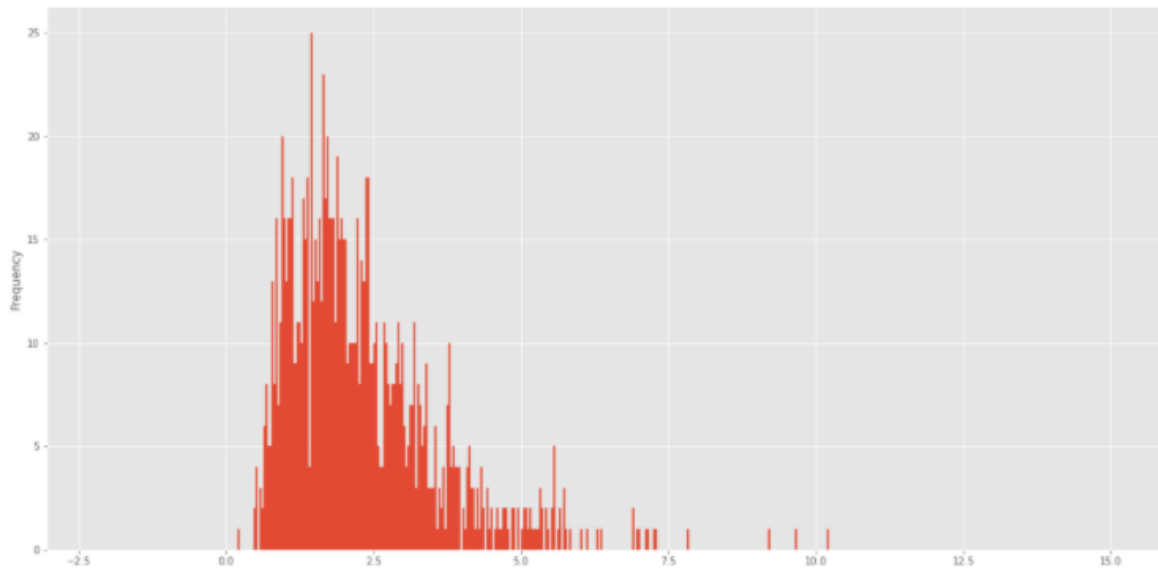
Uno de los retos que presenta la modelización de los sobrecostes es la predicción de estos valores atípicos, por lo que no se decide sacar estos datos atípicos del dataframe original de datos.

### **5.3.3 DISTRIBUCIÓN.**

Para entender mejor la variable respuesta del futuro se realiza un análisis estadístico básico de la distribución del precio de sobrecoste, en aras de crear nuevas variables explicativas que ayuden a la predicción del precio del sobrecoste.

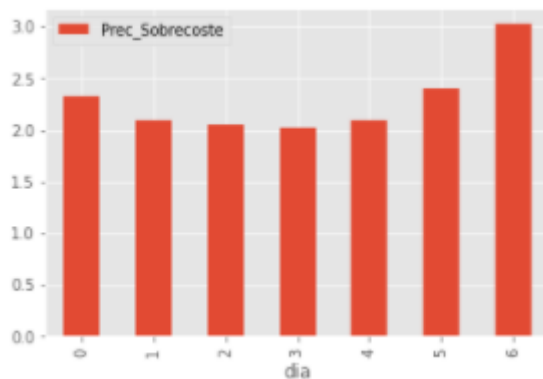
En la siguiente figura se aprecia que la distribución del precio del sobrecoste está sesgada a la izquierda, es decir, la mayoría de los datos se concentran entre los valores de 0 y 2.5 (diariamente):

*DESARROLLO DEL PROYECTO III: ANÁLISIS EXPLORATORIO DE DATOS.*



*Figura 23: Histograma de distribución del precio de sobrecoste diariamente (2018-2020).*

De manera diaria podemos apreciar unos distintos precios de media para cada día de la semana como muestra la siguiente figura:

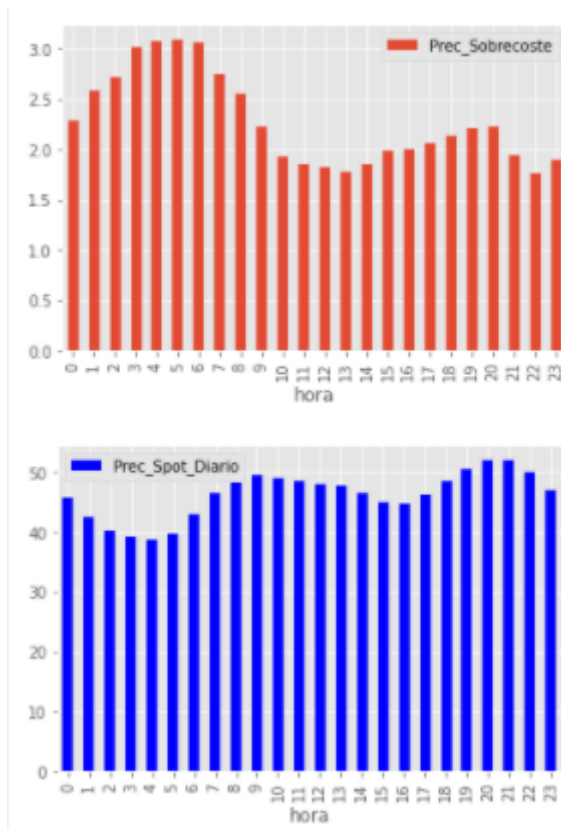


*Figura 24: Diagrama de barras de la distribución del precio de sobrecoste por día de semana (2018-2020). Siendo 0 lunes y 6 domingo.*

Se puede inferir que los días correspondientes a los fines de semana tienen precios más altos que los días laborales. Especialmente el domingo que es el día con sobrecostes más altos.

*DESARROLLO DEL PROYECTO III: ANÁLISIS EXPLORATORIO DE DATOS.*

Finalmente, en la siguiente figura se muestra un diagrama de barras que compare el precio del spot con el precio del sobrecoste de manera horaria para ver algún posible efecto del precio spot en el sobrecoste:



*Figura 25: Comparación diagrama barras horario de precio de spot y precio de sobrecoste.*

Se observa que la relación alto sobrecoste – bajo precio spot en el histórico recogido no se cumple de manera muy precisa como se adelantaba desde negocio. Pese a ello se aprecia que las horas con un spot más bajo son las de la madrugada que coinciden con las horas de mayor precio de sobrecoste, pero esto se debe a un efecto de baja demanda.

### 5.3.4 CORRELACIONES.

Para garantizar que las variables explicativas que se usen en el modelo no den información redundante a la hora de predecir la variable respuesta del precio de sobrecoste, realizamos



*DESARROLLO DEL PROYECTO III: ANÁLISIS EXPLORATORIO DE DATOS.*

un análisis de las correlaciones que existe con las variables explicativas como se puede apreciar en la siguiente imagen:

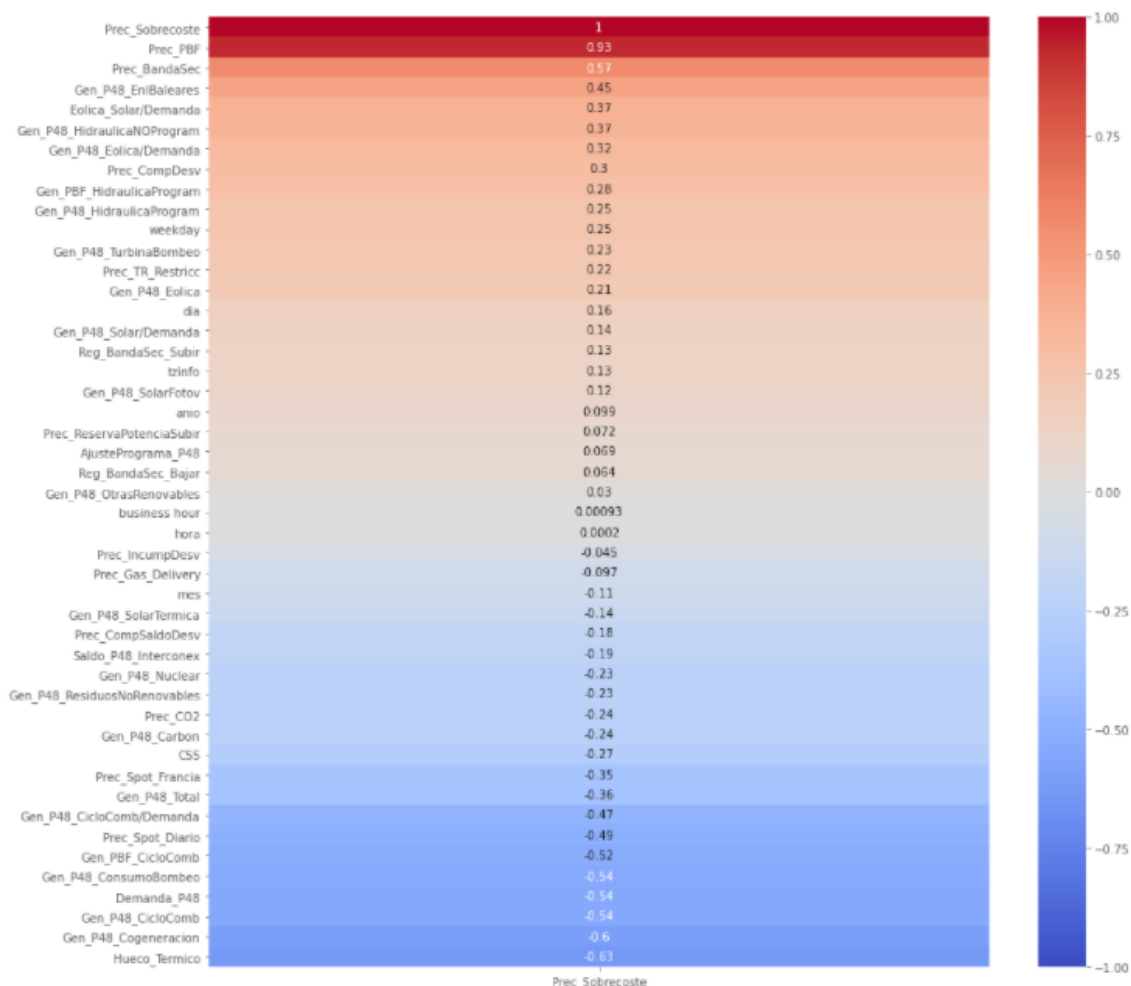


Figura 26: Correlación de las variables con el precio del sobrecoste de manera diaria.

Se observa que no hay ninguna correlación realmente significativa, mayor del 70% (solamente *Precio medio horario componente restricciones PBF*). Aunque sí podemos apreciar que hay una mayor cantidad de variables ligeramente correlacionadas negativamente que posiblemente vayan a tener cierto peso o influencia a la hora de definir nuestro modelo.

### 5.3.5 ESTACIONALIDAD.

La estacionalidad es un comportamiento o patrón que a veces observamos en una serie de tiempo. Consiste en subidas y bajadas periódicas que se presentan en forma regular en la serie de tiempo (si hay tendencia no es estacionario). Si un modelo es estacionario va a ser bueno para poder realizar predicciones.

Para comprobar la estacionariedad se realiza el test de Dicker-Fuller Aumentado. Este test tiene la hipótesis nula de que la serie no estacionaria. Tras aplicar la prueba se consigue un p-valor de  $1.45e-15$  y valor negativo de  $\text{Test}=-9.25$ , indica que rechazamos la hipótesis nula de no estacionariedad [12]. Es decir, la serie del precio de sobrecoste es estacionaria.

De manera gráfica se busca extraer nueva información acerca de la serie de precio de sobrecostes analizando la función de autocorrelación (ACF) y autocorrelación parcial (PACF).

En las siguientes gráficas se analizará la función ACF, primeramente, con un lag<sup>10</sup> de 24 horas y posteriormente con una semana:

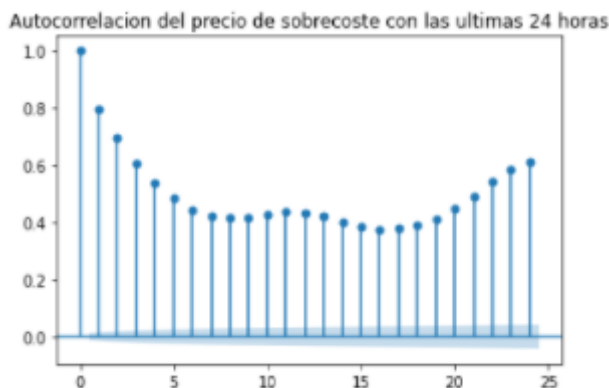


Figura 27: Autocorrelación del precio de sobrecoste con lag = 24 horas.

<sup>10</sup> Lag: es una cantidad fija de tiempo transcurrido

*DESARROLLO DEL PROYECTO III: ANÁLISIS EXPLORATORIO DE DATOS.*

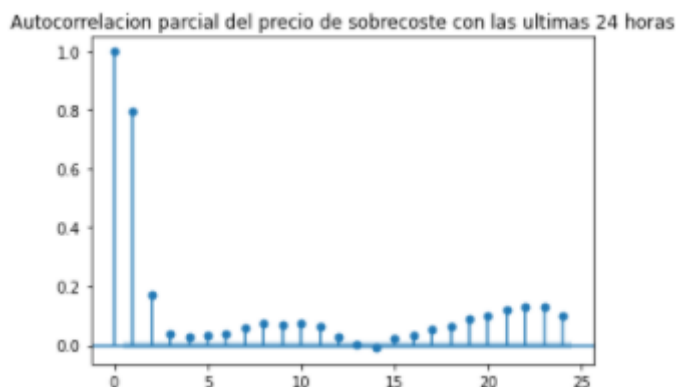
Se puede ver que las horas inmediatamente anteriores a la hora en cuestión y a medida que nos acercamos al día anterior muestran la mayor autocorrelación, esto está en línea con lo que esperaríamos dado que nuestro análisis muestra que los precios fluctúan significativamente a lo largo del día.



*Figura 28: Autocorrelación del precio de sobrecoste con la última semana.*

Esto muestra que la tendencia general de cada 24 horas está bastante correlacionada y mirando hacia atrás una semana completa, vemos que una semana en el pasado está más auto correlacionada que en cualquier otro momento que no sea el par de horas anteriores. Es probable que se desee tener en cuenta las tendencias semanales, ya sea mediante la inclusión de los precios de la semana anterior o mediante la inclusión del día de la semana en el análisis.

En las siguientes gráficas se analizará la función PACF, primeramente, con un lag de 24 horas y posteriormente con una semana:



*Figura 29: Autocorrelación parcial del precio de sobrecoste con las últimas 24 horas.*

*DESARROLLO DEL PROYECTO III: ANÁLISIS EXPLORATORIO DE DATOS.*

---

Al mostrar la autocorrelación parcial, parece que todas las horas anteriores tienen alguna importancia, aunque 1 y 2 horas en el pasado muestran la mayor importancia.



*Figura 30: Autocorrelación parcial del precio de sobrecoste con la última semana.*

Al observar la autocorrelación parcial, todavía vemos que cada día del pasado tiene un significado, pero retroceder una semana no parece tan importante.

## **6. DESARROLLO DEL PROYECTO IV: MODELIZACIÓN DE SOBRECOSTES.**

En este apartado se detalla cómo se plantea la modelización de los sobrecostos, qué técnicas de Machine Learning se han aplicado. Hay que recalcar que en el desarrollo de este punto se explicarán que procedimientos que se han seguido a la hora de desarrollar los modelos, pero no se mostrarán datos relevantes que puedan comprometer los intereses de la empresa como las variables usadas en los modelos o el código desarrollado para implementar los modelos. Por otro lado, sí que se mostrarán resultados estadísticos que validen la eficacia y fiabilidad de los modelos entrenados.

### **6.1 PLANTEAMIENTO INICIAL.**

Hasta ahora todo el progreso del proyecto se había desarrollado en código Python, la parte de modelización se decidió desarrollar con lenguaje R.

Llegados a este punto se tienen en cuenta dos necesidades del área de negocio para elegir qué tipo de modelo es el más apropiado para predecir los sobrecostos:

1. Predicciones mensuales: el uso principal que va a tener el modelo es el de una predicción media mensual del precio del sobrecoste por la forma en que se les hacen las ofertas a los clientes. Es por ello por lo que se decide agregar de manera diaria los datos recopilados de manera horaria para simplificar el número de variables que puedan afectar a la predicción. El objetivo será realizar predicciones diarias que posteriormente se agregarán de manera mensual y que podrán ser aplicadas a las ofertas de clientes.
2. ¿Cómo hacer las predicciones a futuro?: para hacer las predicciones, el modelo que se define a base de datos reales tendrá como inputs variables a futuro, extraídas de un modelo interno de la empresa.
3. Modelo explicativo: la necesidad de un modelo explicativo es en aras de una mayor comprensión de las variables y las relaciones que ayudarán a predecir los sobrecostos.

*DESARROLLO DEL PROYECTO IV: MODELIZACIÓN DE SOBRECOSTES.*

---

Esto permitirá al área de negocio desarrollar estrategias para hedgearse<sup>11</sup> en caso de valores anómalos de los sobrecostes que no se puedan predecir, como es el caso de fallos en tecnologías de generación, desastres naturales como el reciente temporal de “Filomena” o el colapso del canal de Suez. Es por ello, que se decide empezar a probar modelos de mayor simplicidad para ver su comportamiento antes de probar modelos de mayor complejidad.

A la hora de entrenar los modelos se han seguido las siguientes condiciones:

1. Declaración de las variables discretas como factor.
2. Entrenamiento con el 75% de los datos reales extraídos mediante APIs públicas e internas.
3. Test con el 25% de los datos de los datos reales extraídos mediante APIs públicas e internas.
4. Se ha utilizado la técnica de K-Folds Cross Validation: El método K-Fold Cross-Validation es también un proceso iterativo. Consiste en dividir los datos de forma aleatoria en k grupos de aproximadamente el mismo tamaño, k-1 grupos se emplean para entrenar el modelo y uno de los grupos se emplea como validación. Este proceso se repite k veces utilizando un grupo distinto como validación en cada iteración. El proceso genera k estimaciones del error cuyo promedio se emplea como estimación final [13].
4. Para validar la eficacia y performance de los modelos se han utilizado tres métricas estadísticas para el conjunto de train y test:
  - a.  $R^2$ : El coeficiente determina la calidad del modelo para replicar los resultados, y la proporción de variación de los resultados que puede explicarse por el modelo a la hora de predecir una variable respuesta.

---

<sup>11</sup> Hedging: o cobertura es la realización de una actividad financiera para reducir o eliminar las posibles pérdidas que pueden causar las inversiones financieras [17].

- b. Error absoluto medio (MAE): muestra las diferencias absolutas entre la predicción y la observación real, donde todas las diferencias individuales tienen el mismo peso.
- c. Error cuadrático medio (RMSE): es una regla de puntuación cuadrática que también mide la magnitud media del error. Es la raíz cuadrada del promedio de diferencias cuadradas entre la predicción y la observación real [14].

## **6.2 DESARROLLO DE MODELOS.**

### **6.2.1 REGRESIÓN MÚLTIPLE.**

#### **6.2.1.1 Descripción del modelo.**

Los tres modelos de regresión que se han probado con distintas variables han tratado de minimizar la función de pérdida del error absoluto medio, que no es tan sensible a los valores atípicos como el error cuadrático medio [14].

En cada modelo entrenado se ha obtenido una serie de coeficientes para la definición de la fórmula de regresión de la siguiente manera:

$$y_j = b_0 + b_1x_{1j} + b_2x_{2j} + \dots + b_kx_{kj} + u_j$$

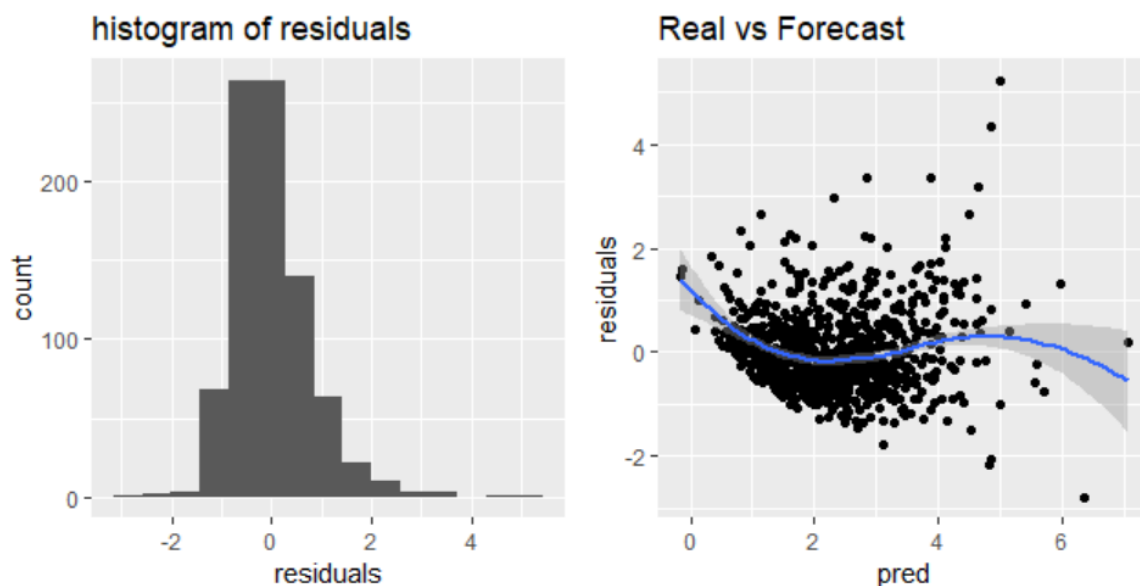
Donde  $y$  es la variable endógena,  $x$  las variables exógenas,  $u$  los residuos y  $b$  los coeficientes estimados del efecto marginal entre cada  $x$  e  $y$ .

Toda variable con un p-valor asociado mayor de 0.05 no se consideraba influyente para el modelo y esa variable era eliminada.

*DESARROLLO DEL PROYECTO IV: MODELIZACIÓN DE SOBRECOSTES.*

También se ha realizado un análisis de colinealidad de las variables mediante a métrica VIF<sup>12</sup>. Si esta métrica era mayor de 10 para una variable existía multicolinealidad<sup>13</sup> y esa variable era eliminada del modelo.

Tras el entrenamiento del modelo y comprobación de los resultados estadísticos que se muestran en el siguiente apartado se ha realizado un análisis de los residuos. Estos residuos deben seguir una distribución normal y estar centrados en cero en media. Si no ocurren estos dos factores quiere decir que el modelo es mejorable y que la interacción de alguna variable explicativa con la variable respuesta no sigue una relación lineal con la variable salida; si no que puede existir alguna relación cuadrática o no lineal y es preciso revisar el modelo para mejorar su fiabilidad y precisión. En la siguiente imagen se muestra un ejemplo de los residuos generados en uno de los modelos entrenados:



*Figura 31:Residuos de modelo de regresión múltiple.*

<sup>12</sup> Factor de inflación de la varianza VIF(j): se interpreta como el incremento en la varianza del coeficiente  $b_j$ , debido a la multicolinealidad de  $X_j$  con las restantes variables explicativas

<sup>13</sup> Multicolinealidad: es la relación de dependencia lineal fuerte entre más de dos variables explicativas en una regresión múltiple



*DESARROLLO DEL PROYECTO IV: MODELIZACIÓN DE SOBRECOSTES.*

En este ejemplo se puede apreciar que el histograma de residuos parece seguir una distribución normal. Pese a ello, el valor medio entre la relación de residuos y valores predichos no sigue una media en cero. Esto demuestra que el modelo de regresión se puede mejorar y se debe a dos factores principalmente:

1. No eliminación de valores atípicos debido al interés que reporta el tratar de modelar estos outliers.
2. Existencia de relaciones cuadráticas o no lineales entre la variable salida y las variables de entrada. Se ha tratado de mejorar este aspecto, aplicando transformaciones a los inputs e incluyendo interacciones entre variables, se consigue mejorar la distribución de los residuos y el modelo.

**6.2.1.2 Análisis de resultados.**

Para el análisis de los modelos y valoración de la fiabilidad del modelo nos fijaremos en tres métricas estadísticas de los 3 modelos de regresión múltiple desarrollados. Los resultados se resumen en la siguiente tabla:

*Tabla 2: Resultados estadísticos modelos de regresión lineal múltiple.*

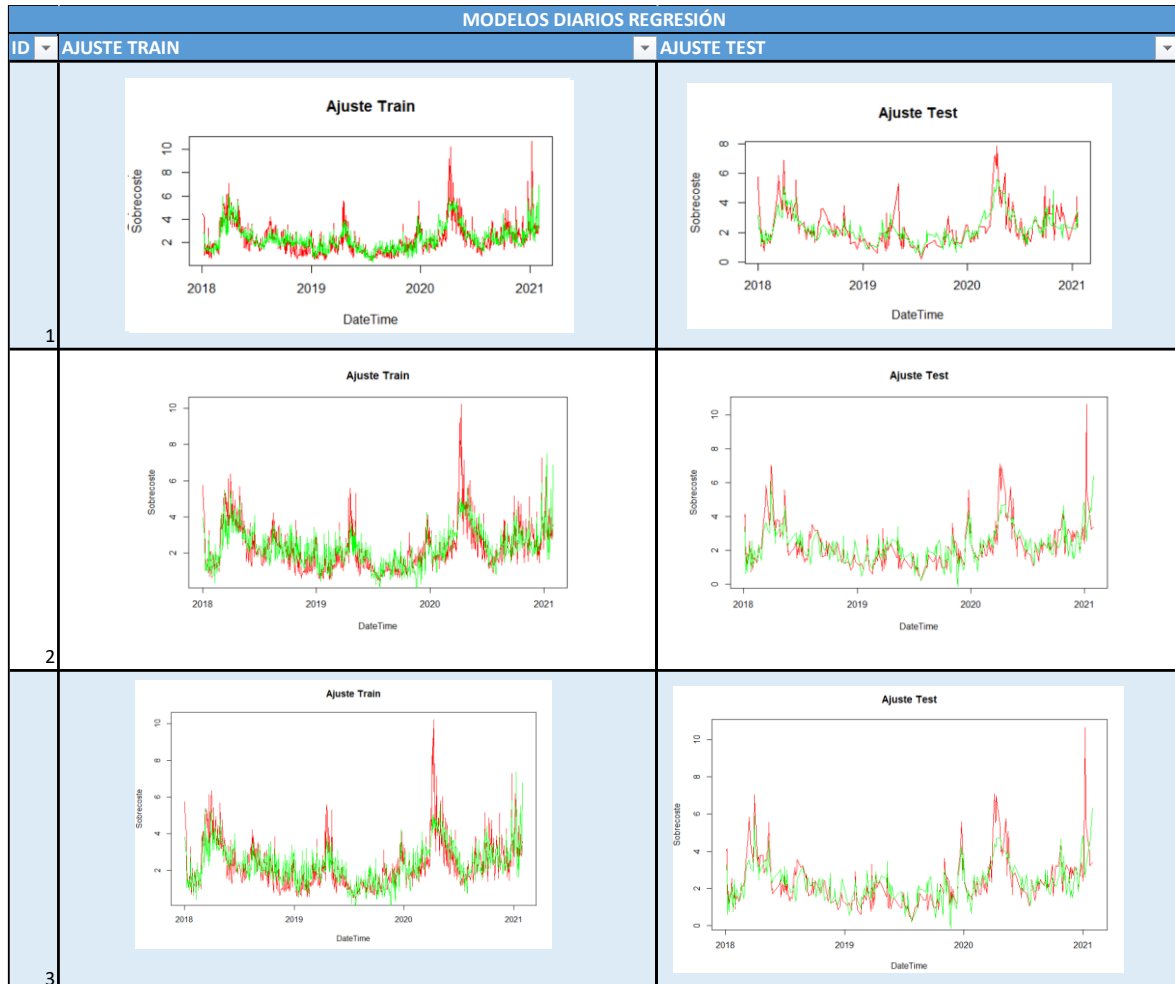
MODELOS DIARIOS REGRESIÓN						
ID	R2 TRAIN	R2 TEST	MAE TRAIN	MAE TEST	RMSE TRAIN	RMSE TEST
1	0.67	0.58	0.49	0.59	0.73	0.78
2	0.7	0.65	0.45	0.52	0.69	0.75
3	0.66	0.56	0.55	0.59	0.74	0.88

Generalizando los tres modelos tienen algo de overfitting, es decir, que el modelo se ha aprendido bastante bien los datos de entrenamiento y a la hora de exportar este modelo a datos con los que no ha entrenado (test) su actuación es peor. Esto se puede apreciar en el valor de  $R^2$ , donde el valor de train es mejor que en test.

Pese a ello, el segundo modelo entrenado parece que no tiene un sobre aprendizaje tan acusado como los otros modelos. Este mismo modelo (ID2) parece actuar ligeramente mejor que los otros dos en cuanto a MAE y RMSE.

*DESARROLLO DEL PROYECTO IV: MODELIZACIÓN DE SOBRECOSTES.*

En la siguiente tabla con imágenes se aprecia una comparativa entre lo real (rojo) y lo predicho (verde) de los tres modelos probados:



*Figura 32: Comparación actuación de tres modelos de regresión múltiple en los conjuntos de train y test. Rojo-Real, Verde-Predicho.*

Observamos que los tres modelos de manera visual tienen una performance bastante similar, confirmando lo que nos decían las métricas de valoración de modelos ( $R^2$ , MAE y RMSE). Los tres modelos son capaces de explicar bastante bien la tendencia e incluso los valores más atípicos y elevados son capaces de predecirlos relativamente bien.

Desde el punto de vista estadístico y visual viendo lo real y predicho parece que el modelo ID2 es el más fiable y robusto de los tres.

## 6.2.2 ÁRBOL DE DECISIÓN.

### 6.2.2.1 Descripción del modelo.

Para desarrollar los modelos de árboles de regresión se han utilizado entre otras librerías de R: Cart y Rpart.

Se han desarrollado dos modelos de árbol de regresión para predecir el precio del sobrecoste. Con estos modelos se han tenido en cuenta los siguientes parámetros e hiperparámetros para construir el mejor árbol de regresión para el problema:

1. Elección de variables explicativas.
2. Barrido del parámetro de complejidad del árbol: El parámetro de complejidad en rpart es la mejora mínima en la modelo necesaria en cada nodo. En la siguiente gráfica se observa como en los modelos desarrollados nuestro cp óptimo va a ser 0:

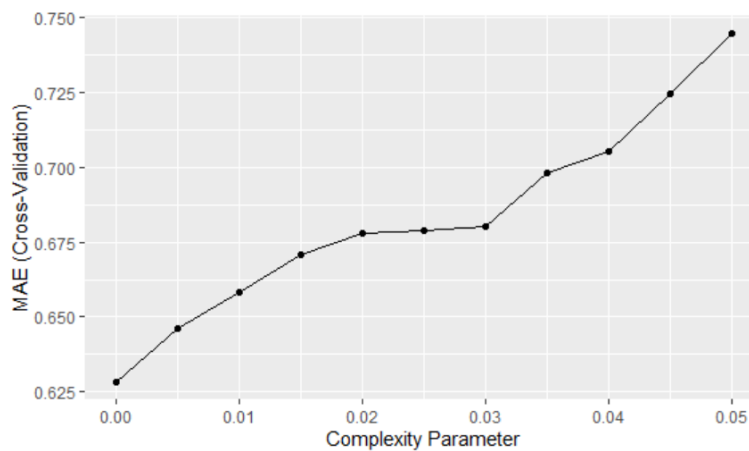


Figura 33: Barrido del parámetro de complejidad a la hora de crear un modelo de árbol de regresión.

3. Barrido en parámetro de minsplit: número de observaciones mínimas en un nodo para seguir podando.
4. Barrido en el parámetro de minbucket: número mínimo de observaciones en un nodo terminal.

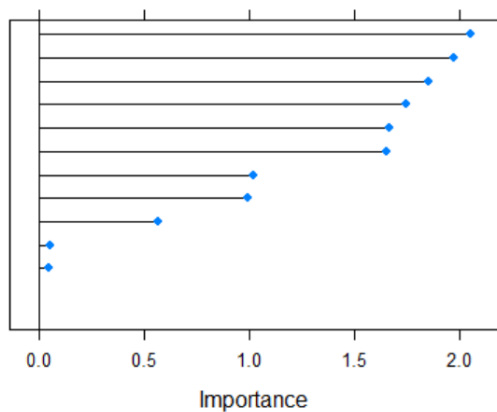
*DESARROLLO DEL PROYECTO IV: MODELIZACIÓN DE SOBRECOSTES.*

Los dos árboles de regresión creados siguen una estructura similar a la mostrada en la siguiente imagen. En cada nodo hay una condición que lleva a un nodo terminal con el valor predico para cada entrada del modelo.



*Figura 34: Ejemplo de estructura de modelo de árbol de regresión (no se muestran el contenido del árbol por motivos de confidencialidad).*

También se muestra la importancia de las variables escogidas a la hora de crear el árbol, como se muestra en la siguiente figura:



*Figura 35: Ejemplo de gráfica que mide la importancia de las variables que se han utilizado para definir el árbol (el contenido del árbol no se muestra por motivos de confidencialidad).*

Este tipo de gráficas ayudan a refinar y mejorar el modelo en la cuestión de la elección de variables.

### 6.2.2.2 *Análisis de resultados.*

La valoración de la calidad de los modelos de árboles de regresión se mide con las métricas anteriormente comentadas:  $R^2$ , MAE y RMSE. Los resultados se resumen en la siguiente tabla:

*Tabla 3: Resultados estadísticos modelos de árboles de regresión.*

MODELOS DIARIOS REGRESSION TREES						
ID	R2 TRAIN	R2 TEST	MAE TRAIN	MAE TEST	RMSE TRAIN	RMSE TEST
1	0.66	0.54	0.51	0.68	0.74	0.93
2	0.68	0.60	0.47	0.6	0.72	0.86

SE puede apreciar que en términos de  $R^2$  los modelos entrenados con árboles de regresión tienden ligeramente al sobre aprendizaje. En cuestión de MAE y RMSE consiguen resultados parecidos a los modelos de regresión lineal multivariante.

En la siguiente tabla con imágenes se aprecia una comparativa entre lo real (rojo) y lo predicho (verde) de los tres modelos probados:

*DESARROLLO DEL PROYECTO IV: MODELIZACIÓN DE SOBRECOSTES.*

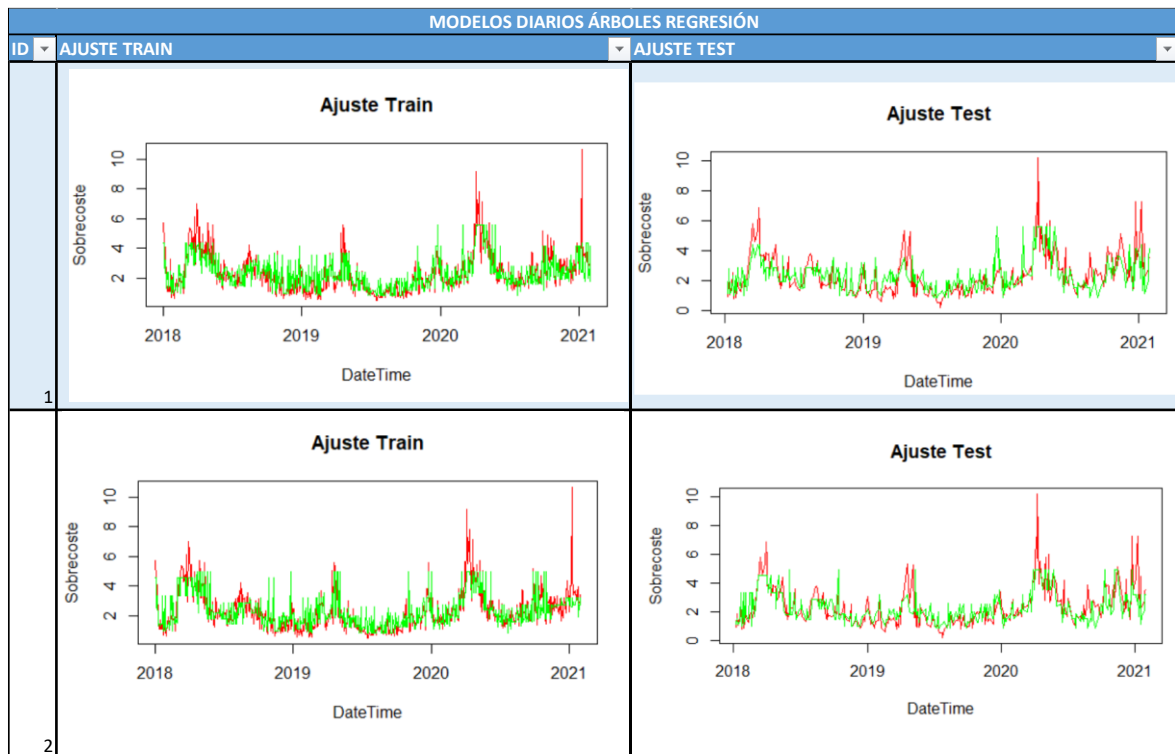


Figura 36: Comparación actuación de los dos modelos de árboles de regresión en los conjuntos de train y test. Rojo-Real, Verde-Predicho.

Si nos fijamos en los resultados gráficamente, podemos apreciar que el performance de los árboles de regresión es algo peor a la hora de predecir valores de máximos relativos de la gráfica. Esto es debido a que debido a sus nodos terminales los valores predichos están en un rango de valores con un máximo y un mínimo. Con nuestros modelos no se ha sido capaz desarrollar un árbol que no esté muy overfiteado y que sea capaz de predecir correctamente estos valores máximo de precios.

### 6.2.3 RANDOM FOREST.

#### 6.2.3.1 Descripción del modelo.

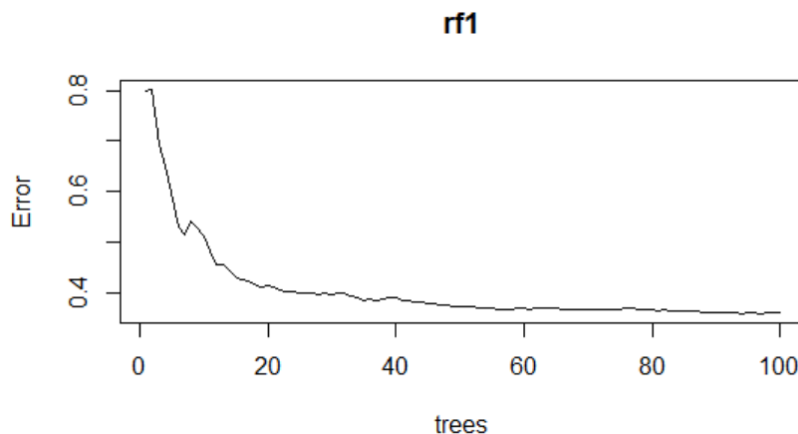
Para desarrollar los modelos de random forest se han utilizado entre otras librerías de R: Random Forest y Ranger. Se trata de un modelo con mayor complejidad que los dos anteriores y que por tanto va a ser menos explicativo. Pese a ello se ha probado para comprobar resultados y una posible utilidad.

*DESARROLLO DEL PROYECTO IV: MODELIZACIÓN DE SOBRECOSTES.*

Un algoritmo de random forest es un conjunto (ensemble) de árboles de decisión combinados con bagging (Bootstrap<sup>14</sup> aggregation). Al usar bagging, lo que en realidad está pasando, es que distintos árboles ven distintas porciones de los datos. Ningún árbol ve todos los datos de entrenamiento. Esto hace que cada árbol se entrene con distintas muestras de datos para un mismo problema. De esta forma, al combinar sus resultados, unos errores se compensan con otros y tenemos una predicción que generaliza mejor [15]. Los hiperparámetros del modelo se han tratado de ajustar para este modelo son los siguientes:

1. Mtry: el número de variables seleccionadas aleatoriamente en cada ramificación para que no haya correlación entre los árboles combinados.
2. N° de árboles: el número de árboles que combinaremos. Se han entrenado los dos modelos desarrollados con un total de 100 árboles. Se ha considerado este un número aceptable, no muy alto, debido a que el conjunto de datos entrenados no es demasiado grande.

En la siguiente imagen podemos apreciar el resultado de combinar 100 árboles en uno de los modelos desarrollados:

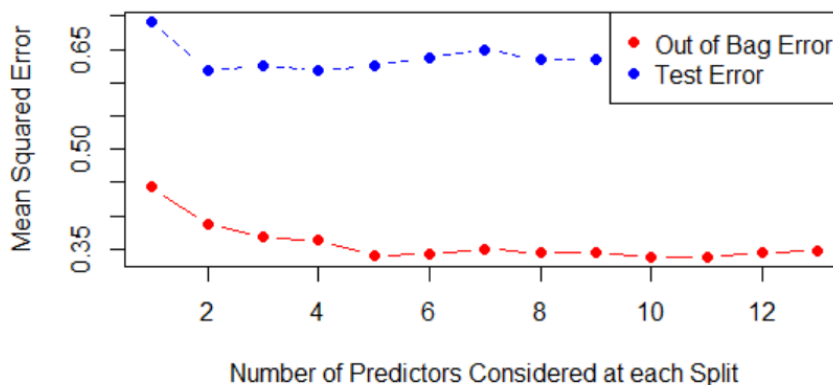


*Figura 37: Cálculo del error de predicción frente al número de árboles combinados en random forest.*

<sup>14</sup> Muestra Bootstrap: La principal utilidad del empleo del bootstrap es reducir el sesgo en el performance de un modelo o, en otras palabras, aproximar la varianza gracias a la realización de remuestreos aleatorios de la muestra inicial. [15]

En la gráfica podemos observar que a medida que aumentamos el número de árboles combinados el erro de predicción se disminuye hasta estabilizarse en torno a un error de 0,36.

Cabe señalar que para refinar el modelo se ha tenido en cuenta el Out of Bag Error (OOB). El OOB es el error promedio que se comete usando predicciones de los árboles que no contienen en su respectiva muestra bootstrap [15]. En la siguiente gráfica se observa la relación del OOB con el error de test real, MSE (error cuadrático medio) que comete el algoritmo random forest en el testeo dependiendo del valor del hiperparámetro *mtry* asignado:



*Figura 38: Comparación de Out of Bag Error y Test Error en función del hiperparámetro *mtry*.*

En este caso se aprecia que el valor óptimo de *mtry*, es decir, el número de variables seleccionadas aleatoriamente en cada ramificación es cinco; con este valor el error OOB es mínimo mientras que el valor de test (MSE) parece que se estabiliza en torno a 0,55.

### **6.2.3.2 Análisis de resultados.**

La valoración de la calidad de los modelos de random forest se mide con las métricas anteriormente comentadas:  $R^2$ , MAE y RMSE. Los resultados se resumen en la siguiente tabla:



*DESARROLLO DEL PROYECTO IV: MODELIZACIÓN DE SOBRECOSTES.*

*Tabla 4: Resultados estadísticos modelos de random forest.*

MODELOS DIARIOS RANDOM FOREST						
ID	R2 TRAIN	R2 TEST	MAE TRAIN	MAE TEST	RMSE TRAIN	RMSE TEST
1	0.97	0.65	0.17	0.47	0.26	0.78
2	0.96	0.66	0.20	0.48	0.27	0.79

Se puede apreciar un claro caso de overfitting, ya que el modelo es capaz de predecir prácticamente a la perfección los datos con los que entrena. Los dos modelos entrenados con distintas variables y distintos hiperparámetros consiguen un  $R^2$  cercano al 100%, es decir, son capaces de explicar casi el 100% de la variabilidad de los datos. Este efecto se replica en las métricas de MAE y RMSE, donde el error obtenido con la muestra de entrenamiento es realmente pequeño. Sin embargo, cuando vemos la actuación del modelo con el conjunto de testeo, vemos que el performance del modelo decae drásticamente frente al performance del modelo con los datos de entrenamiento con todas las métricas:  $R^2$ , MAE y RMSE.

Visualmente podemos ver los resultados de aplicar la combinación de árboles a los datos de validación en la siguiente imagen:

*DESARROLLO DEL PROYECTO IV: MODELIZACIÓN DE SOBRECOSTES.*

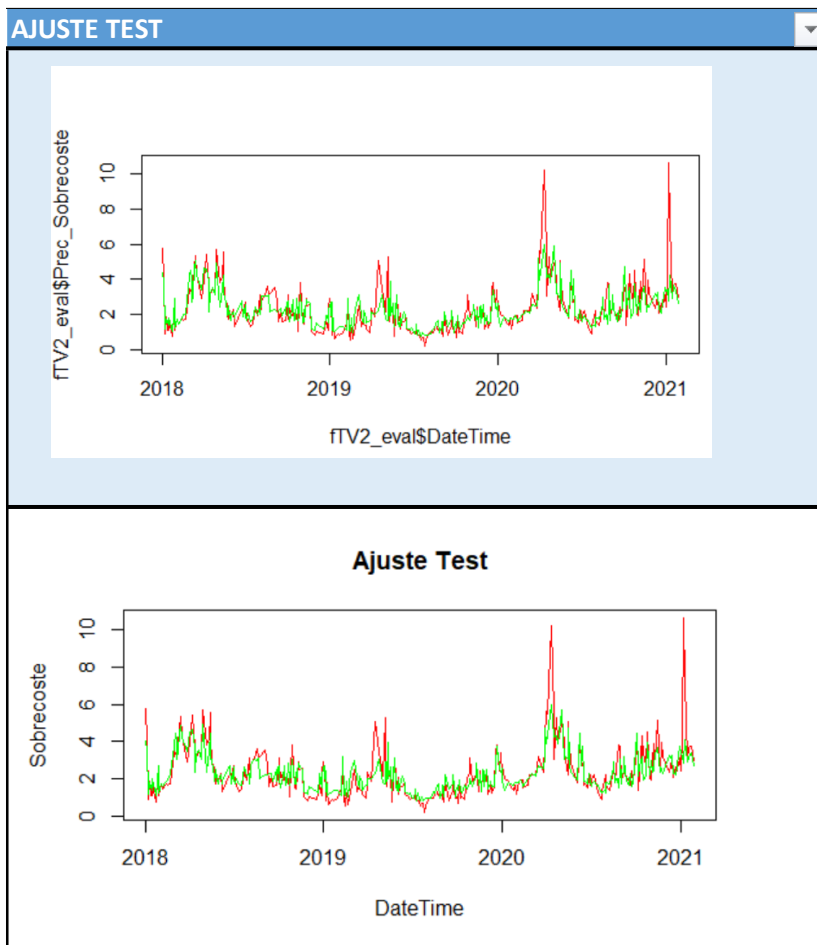


Figura 39: Comparación actuación de los dos modelos de random forest en el conjunto de test. Rojo-Real, Verde-Predicho.

Visualmente se observa que para el rango de valores más comunes de precio de sobrecoste la actuación del modelo es relativamente similar a los anteriores modelos entrenados, más concretamente a los modelos de regresión. Pese a ello parece que no es capaz de predecir igual de bien rangos de valores más extremos, si se refiere a valores máximos locales.

En conclusión, se trata de un modelo muy poco fiable y que no garantiza una buena predicción y robustez a la hora de probarlo con datos nuevos.

## 6.2.4 STACKING.

### 6.2.4.1 Descripción del modelo.

Stacking es un algoritmo definido en el conjunto de algoritmos de “ensamble methods”.

Utiliza un algoritmo de meta aprendizaje para aprender cómo combinar mejor las predicciones de dos o más algoritmos básicos de aprendizaje automático.

El beneficio del apilamiento es que puede aprovechar las capacidades de una variedad de modelos con buen desempeño en una tarea de clasificación o regresión, en el caso se presenta el segundo, y hacer predicciones que tienen un mejor desempeño que cualquier modelo individual en el conjunto [16].

Los parámetros se han tratado definido para la ejecución de este modelo son los siguientes:

1. Variables elegidas para los modelos que se entrenen conjuntamente.
2. Modelos que se van a combinar: en este caso se probarán los 3 modelos entrenados y explicados anteriormente: regresión lineal multivariante, árboles de regresión y random forest.

### 6.2.4.2 Análisis de resultados.

Una vez se entrena el modelo como resultado a cada modelo se le asigna un determinado peso, resumido en la siguiente tabla:

*Tabla 5: Pesos de los distintos modelos combinados con Stacking*

Regresión lineal multivariante	Árbol de regresión	Random forest
0.4394	-0.0889	0.71

Se puede apreciar que el algoritmo de stacking da mayor peso e importancia al modelo de random forest, seguidamente al de regresión lineal y por último al árbol de regresión, lo que

*DESARROLLO DEL PROYECTO IV: MODELIZACIÓN DE SOBRECOSTES.*

quiere decir que este último modelo no realiza un buen trabajo de predicción combinado con los otros dos modelos.

La valoración de la calidad del modelo de stacking se mide con las métricas anteriormente comentadas:  $R^2$ , MAE y RMSE. Los resultados se resumen en la siguiente tabla:

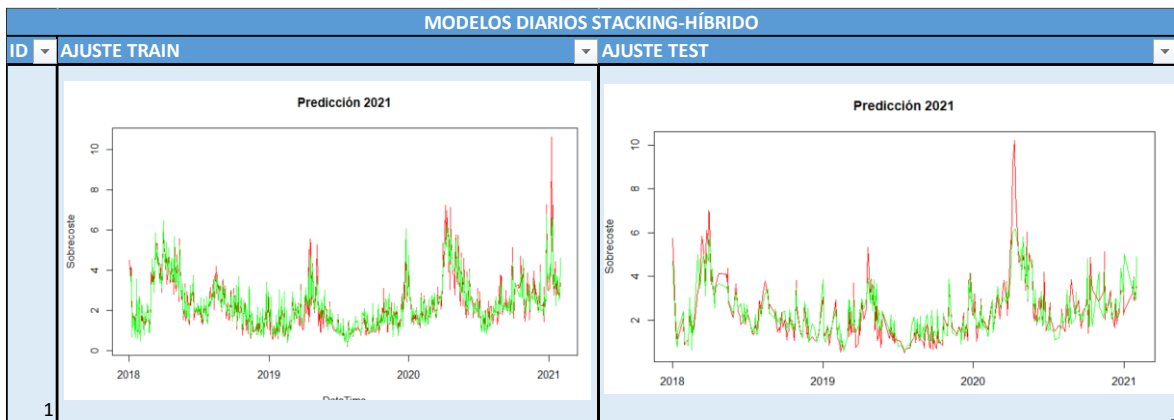
*Tabla 6: Resultados estadísticos modelo de stacking.*

MODELOS DIARIOS STACKING-HÍBRIDO						
ID	R2 TRAIN	R2 TEST	MAE TRAIN	MAE TEST	RMSE TRAIN	RMSE TEST
1	0.905	0.858	0.27	0.37	0.39	0.53

Se puede apreciar que el modelo pese a tener unos resultados que relativamente puede considerarse que están sobreentrenados si nos fijamos en el  $R^2$  (mayor en el training que en el test), la realidad es que se puede considerar que el modelo hace un gran trabajo e incluso llegando confirmar que la combinación de modelos es la técnica que mejor ha funcionado de las probadas. SE puede observar igualmente que en términos de MAE y RMSE no hay diferencias abismales entre los errores del modelo con el conjunto de entrenamiento y con el de prueba.

Visualmente podemos ver los resultados de aplicar la combinación de modelos a los datos de entrenamiento y prueba en la siguiente imagen:

*DESARROLLO DEL PROYECTO IV: MODELIZACIÓN DE SOBRECOSTES.*



*Figura 40: Comparación actuación del modelo de stacking en los conjuntos de train y test. Rojo-Real, Verde-Predicho.*

Se puede apreciar que el modelo en el rango de valores medios para la predicción del sobrecoste hace un trabajo bastante bueno. Además, el modelo también es capaz de predecir con bastante buena exactitud los picos de valores en rangos más elevados.

## **7. DESARROLLO DEL PROYECTO V: ANÁLISIS DE RESULTADOS.**

En este apartado se detalla una comparativa de los resultados de los modelos entrenados para decidir que modelo se pone en producción. Como en apartados anteriores no se detalle ningún dato relevante que pueda comprometer a la empresa, ni se aporta código desarrollado.

### ***7.1 COMPARATIVA DE MODELOS Y ELECCIÓN DE MODELO PARA PRODUCCIÓN.***

Para la comparación del funcionamiento de los modelos se han llevado a cabo una prueba de validación de los modelos que complemente y justifique los resultados estadísticos obtenidos y explicados en el capítulo anterior.

Para la validación del correcto funcionamiento de los modelos definidos se han probado con un set de datos desde febrero a mayo de 2021. El mes de enero debido a su complejidad asociada al temporal Filomena que azotó la península ese mes derivando en comportamientos anómalos en el mercado mayorista de la energía. Este mes posteriormente se decidió incluirlo en los datos de entrenamiento de los modelos.

Los resultados estadísticos de los modelos se resumen en la siguiente tabla:

*Tabla 7: Resultados estadísticos de la validación de los modelos en los meses de febrero a mayo de 2021.*

Modelo	R2 VALIDATION	MAE VALIDATION	RMSE VALIDATION
1-RegLin	0.52	0.83	1.03
2-RegLin	0.60	0.63	0.84
3-RegLin	0.59	0.83	0.98
1-Tree	0.5	0.81	1.06
2-Tree	0.48	0.93	1.24
1-RF	0.59	0.74	1.0
2-RF	0.63	0.74	1.0
1-Hib	0.346	0.86	1.07

---

*DESARROLLO DEL PROYECTO V: ANÁLISIS DE RESULTADOS.*

---

A la vista de los resultados estadísticos que arroja la tabla podemos confirmar que los dos modelos más fiables y que mejores resultados obtienen son los modelos 2 y 3 de regresión lineal multivariante. Si se comparan los resultados obtenidos de los modelos con los conjuntos de train y test anteriormente explicados, los resultados no varían en gran medida de los conseguidos con el nuevo conjunto de datos. Esto confirma la robustez del modelo para poder exportarlo a nuevos conjuntos de datos. Este fenómeno también ocurre en los modelos de árboles de regresión, aunque los resultados obtenidos son peores, es decir, son modelos robustos pero que predicen en este caso de estudio peor que la regresión lineal multivariante.

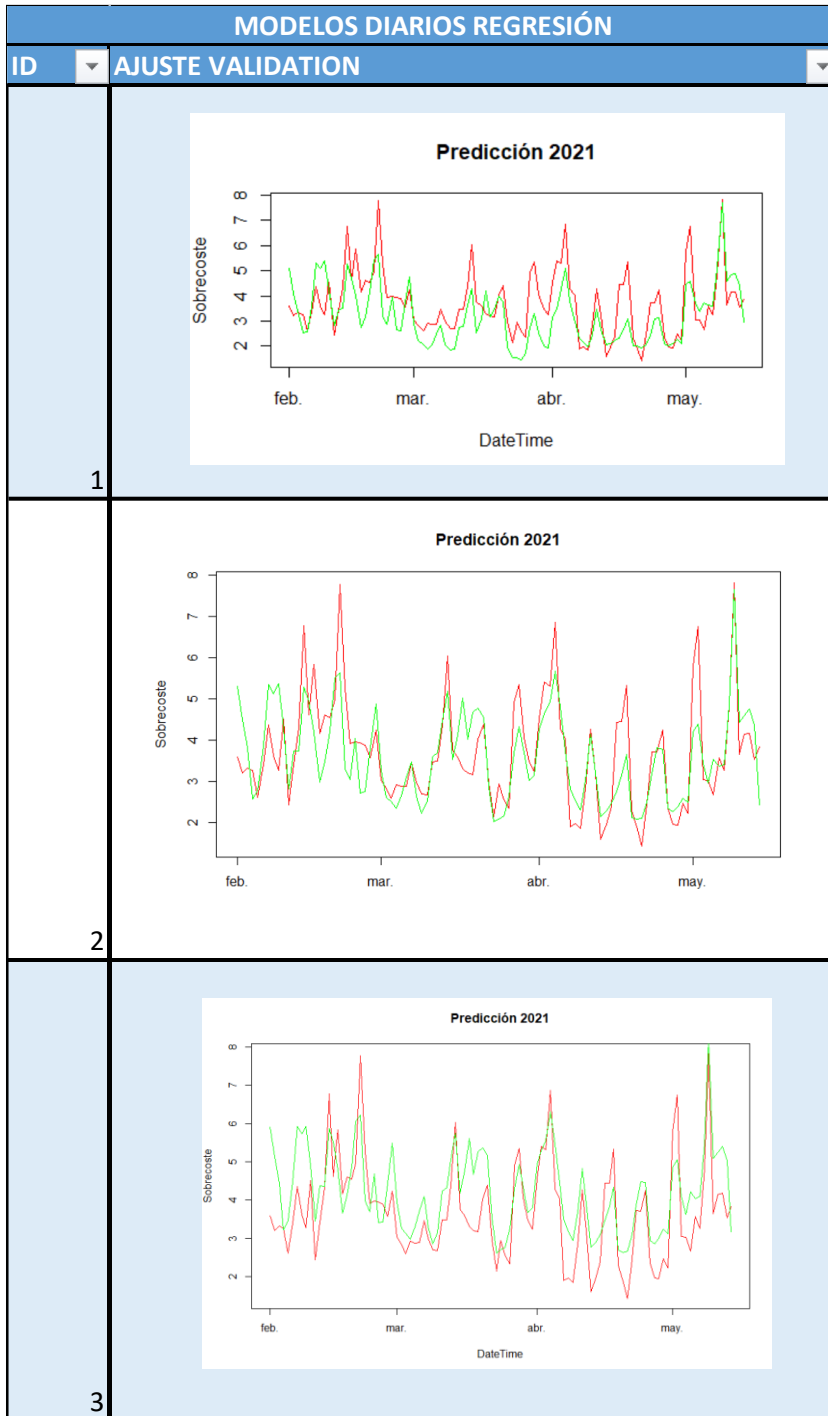
Respecto a los modelos más complejos, los modelos de random forest y stacking, se puede apreciar que su complejidad no permite exportar tan bien como los modelos simples sus resultados a nuevos conjuntos de datos, es decir, son menos fiables. Los resultados de las métricas distan bastante de los resultados obtenidos cuando se entrenaba el modelo, es decir, podemos confirmar estos modelos como un caso de sobre aprendizaje.

Los modelos de random forest muestran resultados similares en el nuevo conjunto de datos y en el testeo, pero la diferencia abismal con los resultados obtenidos con el conjunto de train hace que se descarte este modelo.

Por otro lado, el modelo de stacking que tan buenos resultados parecía obtener en los conjuntos de train y test (sin considerar que hubiera un overfitting muy relevante), cuando se exporta a nuevos datos no parece actuar muy bien. Los resultados obtenidos con el nuevo conjunto de datos son radicalmente peores, confirmado el sobre entrenamiento del modelo. Por su poca fiabilidad se decide descartar para su puesta en producción.

Para apreciar visualmente los resultados de los modelos con el nuevo conjunto de datos se pintan los sobrecostos de manera real frente a lo predicho. Las siguientes figuras muestran los resultados del modelo de regresión lineal múltiple:

*DESARROLLO DEL PROYECTO V: ANÁLISIS DE RESULTADOS.*



*Figura 41: Comparación actuación de los modelos de regresión lineal multivariante en el conjunto de datos de febrero a mayo de 2021. Rojo-Real, Verde-Predicho.*



*DESARROLLO DEL PROYECTO V: ANÁLISIS DE RESULTADOS.*

Según las gráficas parece que el modelo que mejor acierta la tendencia y mejor ajusta las predicciones a los datos reales es el modelo 2. Los otros dos modelos parecen coger bien la tendencia, pero no llegan a ajustarse del todo a los valores más acusados.

Las siguientes figuras muestran los resultados del modelo de árbol de regresión:

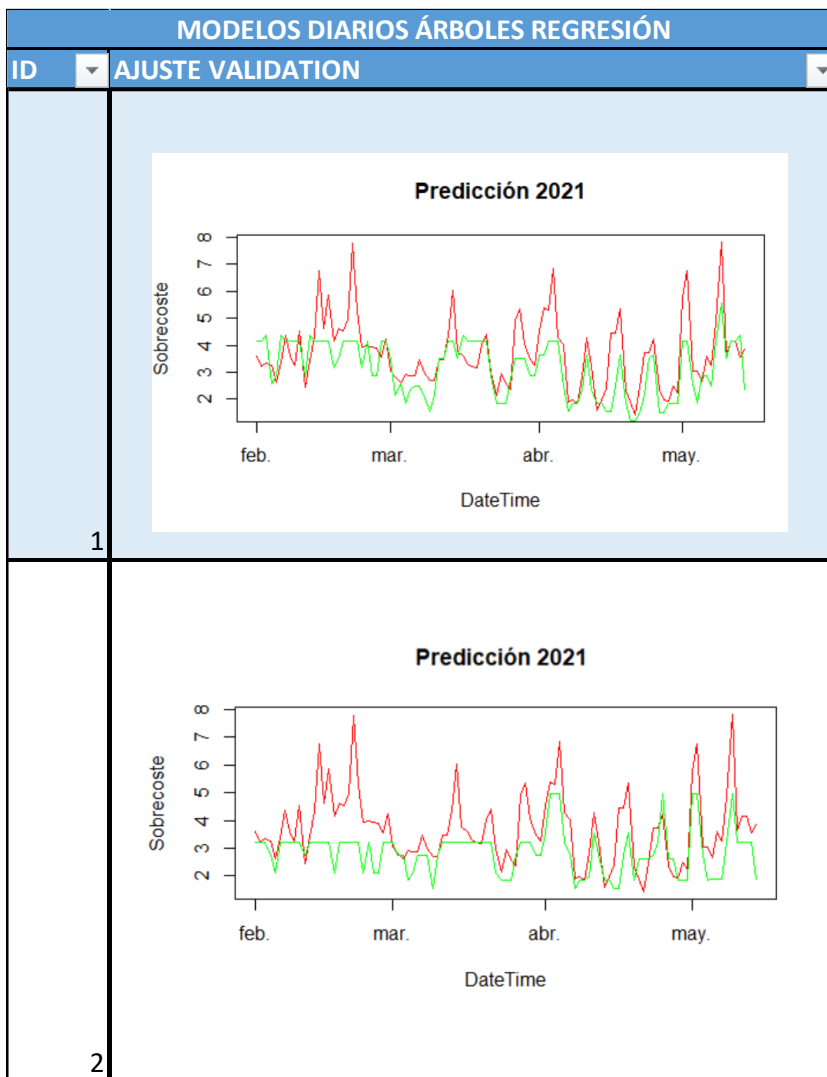


Figura 42: Comparación actuación de los modelos de árboles de regresión en el conjunto de datos de febrero a mayo de 2021. Rojo-Real, Verde-Predicho.

Si nos fijamos en como actúan los modelos, parece que en ciertos puntos dan precios que se estabilizan dando lugar a mesetas y no ajustándose del todo bien a los valores reales, sobre

*DESARROLLO DEL PROYECTO V: ANÁLISIS DE RESULTADOS.*

todo en los meses de febrero y marzo. Esto puede deberse a que el árbol no tiene suficientes nodos terminales y se podría haber hecho más grande. Aunque las pruebas realizadas confirmaban que a medida que se era menos restrictivo con los parámetros de poda del árbol, los modelos tendían a sobrentrenarse. A partir del mes de abril sus predicciones empiezan a mejorar, aunque no lo hacen mejor que los modelos de regresión lineal multivariante.

Las siguientes figuras muestran los resultados del modelo de random forest:

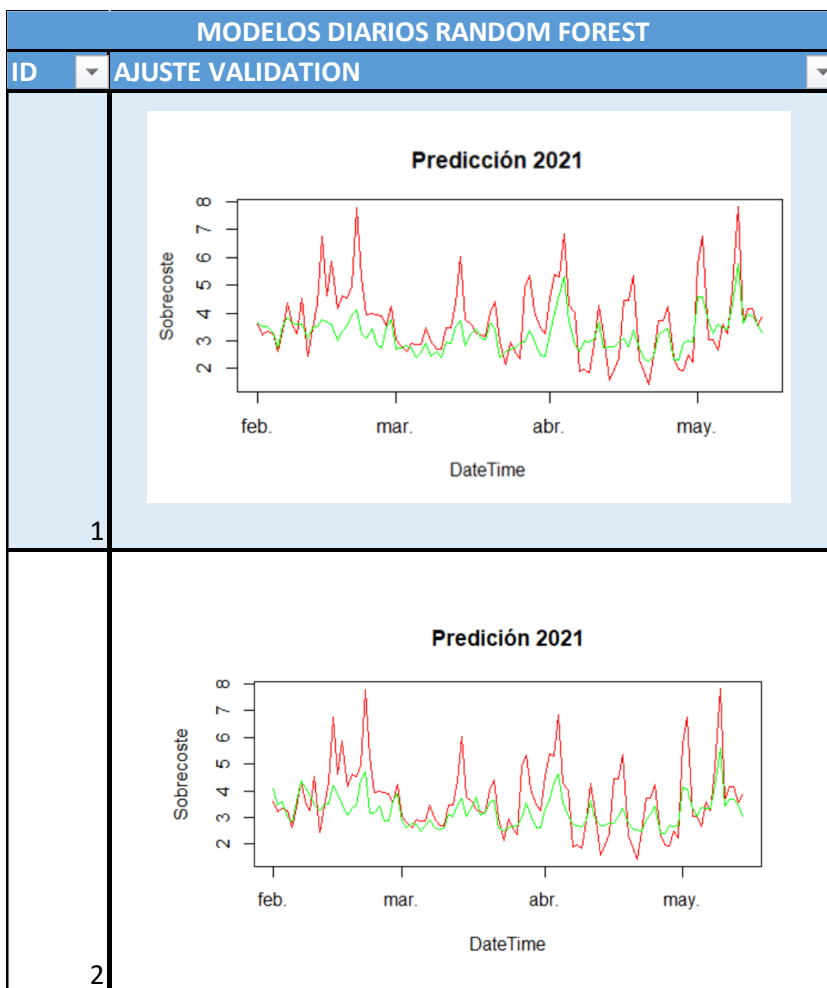
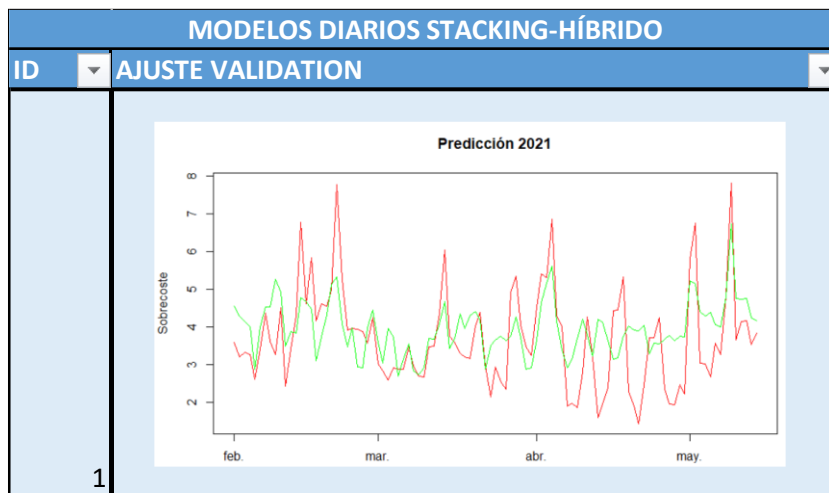


Figura 43: Comparación actuación de los modelos de random forest en el conjunto de datos de febrero a mayo de 2021. Rojo-Real, Verde-Predicho.

*DESARROLLO DEL PROYECTO V: ANÁLISIS DE RESULTADOS.*

Los modelos de random forest parecen que cogen bien la tendencia, pero no son capaces de predecir bien valores reales de precios de manera diaria, no se ajustan bien a los precios reales. Por eso se descartan.

Las siguientes figuras muestran los resultados del modelo de stacking:



*Figura 44: Actuación del modelo de stacking en el conjunto de datos de febrero a mayo de 2021. Rojo-Real, Verde-Predicho.*

Se puede apreciar que el modelo solamente predice de manera aceptable en los meses de febrero y mitad de marzo, a partir de entonces los precios se convierten en más extremos y es capaz de coger la tendencia, pero no de predecir bien los valores. El modelo de stacking en nuestro caso no es un modelo fiable y por eso se descarta.

La siguiente tabla muestra los resultados en forma de agregado mensual (valor que interesa al área de negocio) del valor real medio del precio del sobrecoste y su predicción:

*DESARROLLO DEL PROYECTO V: ANÁLISIS DE RESULTADOS.*

*Tabla 8: Comparativa error mensual de predicción de los modelos para los meses de febrero y marzo (Verde mejores resultados).*

Modelo	Febrero Real	Febrero predicho	Delta-Febrero	Marzo Real	Marzo Predicho	Delta-Marzo
1-RegLin	4,19	4,13	0,06	3,41	2,73	0,68
2-RegLin	4,19	4,07	0,12	3,41	3,38	0,03
3-RegLin	4,19	4,69	-0,50	3,41	4	-0,59
1-Tree	4,19	3,84	0,35	3,41	3,05	0,36
2-Tree	4,19	2,97	1,22	3,41	2,76	0,65
1-RF	4,19	3,44	0,75	3,41	2,97	0,44
2-RF	4,19	3,5	0,69	3,41	2,94	0,47
1-Hib	4,19	4,13	0,06	3,41	3,6	-0,19

*Tabla 9: Comparativa error mensual de predicción de los modelos para los meses de abril y mayo (Verde mejores resultados).*

Modelo	Abril Real	Abril predicho	Delta-Abril	Mayo(15 días) Real	Mayo(15 días) predicho	Delta-Mayo
1-RegLin	3,25	2,75	0,5	4,29	4,51	-0,22
2-RegLin	3,25	3,14	0,11	4,29	4,17	0,12
3-RegLin	3,25	3,78	-0,53	4,29	4,82	-0,53
1-Tree	3,25	2,41	0,84	4,29	3,51	0,78
2-Tree	3,25	2,74	0,51	4,29	3,07	1,22
1-RF	3,25	3,2	0,05	4,29	3,9	0,39
2-RF	3,25	3,07	0,18	4,29	3,7	0,59
1-Hib	3,25	3,82	-0,57	4,29	4,7	-0,41

Esta tabla es bastante reveladora para los intereses del área de negocio, resume de una manera clara y concisa el error real que se tiene al predecir el precio de sobrecoste medio de manera mensual. Observamos con una celda verde los modelos que mejor predicen en media mensual y se confirma que el modelo que menos error produce de manera agregada mensual es el segundo modelo de regresión definido.

Cumpliendo con los requisitos exigidos desde el área de negocio, combinado con la visión estadística extraída del funcionamiento de los modelos con su posterior validación, se concluyen los siguientes puntos:

- Los modelos más fiables son aquellos más simples, es decir, los algoritmos de regresión lineal múltiple y árboles de regresión.
- Los modelos que mejor se ajustan y predicen los datos reales son los modelos de regresión lineal múltiple.

- El mejor modelo que mejor se ajusta a los datos reales, que se exporta de manera más fiable a nuevos datos es el modelo ID-2 de regresión lineal múltiple. Este modelo ha demostrado robustez y que cumple los requisitos exigidos desde el área de negocio: un modelo fácilmente explicable y que prediga de manera ajustada los sobrecostes del sistema de manera media mensual.

## **7.2 APLICACIÓN DEL MODELO EN ESCENARIOS.**

Para realizar la predicción de los precios de sobrecostes a largo plazo mediante una regresión se necesita tener datos a largo plazo que alimenten al modelo. Estos datos se obtienen mediante un modelo interno de la empresa ya validado y fiable.

El modelo interno que alimentará de datos futuros al modelo desarrollado genera tres escenarios distintos en base a unas determinadas hipótesis. Las hipótesis en las que se basan los escenarios pueden ser variadas, según la visión de negocio; puede variar la generación renovable, el precio de spot, generación no renovable... Los inputs del modelo muestran bajo una misma hipótesis tres escenarios donde el precio del sobrecoste puede salir alto, más bajo o un escenario base.

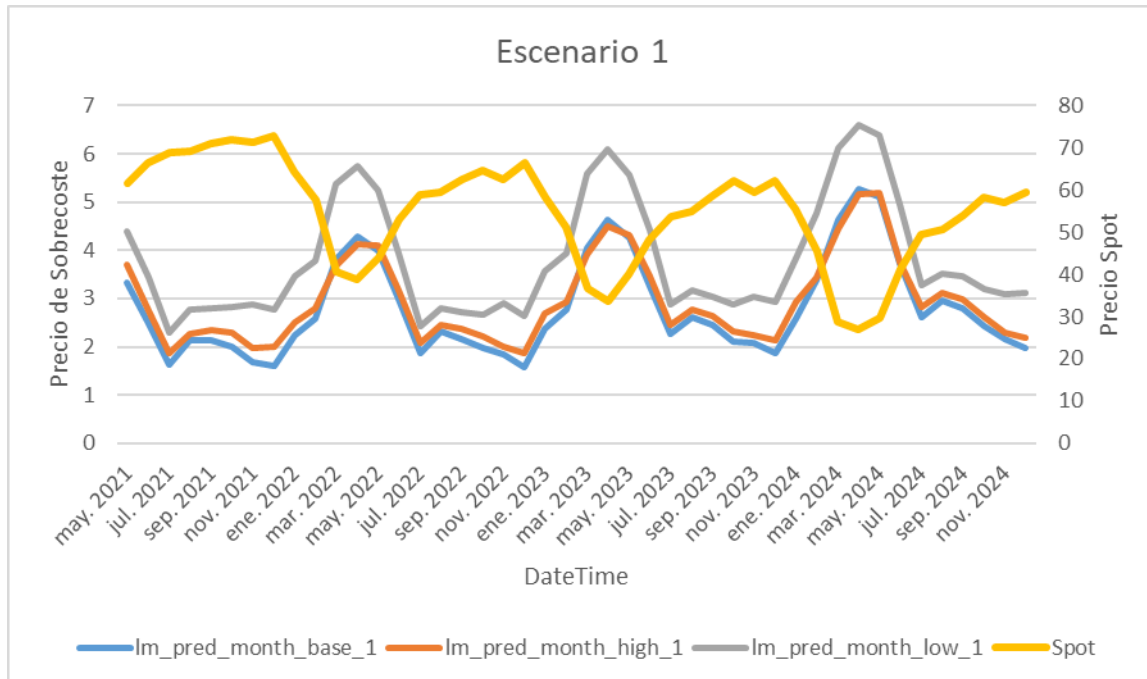
Hay que señalar que los datos no se obtienen de manera directa con el modelo y hay que realizar una preparación, transformación y limpieza de datos antes de introducir las variables como inputs al modelo. También hay algunas series de precios como la del gas o la del CO<sub>2</sub> que se deben obtener como series de precio a futuro mediante bases de datos internas.

Una vez tratados los datos se introducen al modelo (código desarrollado en R) y se predicen los precios de sobrecoste para un horizonte temporal definido. Se ha decidido establecer un horizonte temporal hasta diciembre de 2024 y diciembre de 2025 con cada uno de los escenarios respectivamente.

La primera hipótesis planteada para estimar el precio de sobrecoste a largo plazo es modelar tres escenarios donde la generación de energía hidráulica varía. Es decir, se modelan tres escenarios, uno en el que se considera climatológicamente seco (escenario Low), otro

*DESARROLLO DEL PROYECTO V: ANÁLISIS DE RESULTADOS.*

húmedo (escenario High) y un escenario intermedio (escenario Base). Los resultados de los escenarios derivados de la primera hipótesis se muestran en la siguiente gráfica:

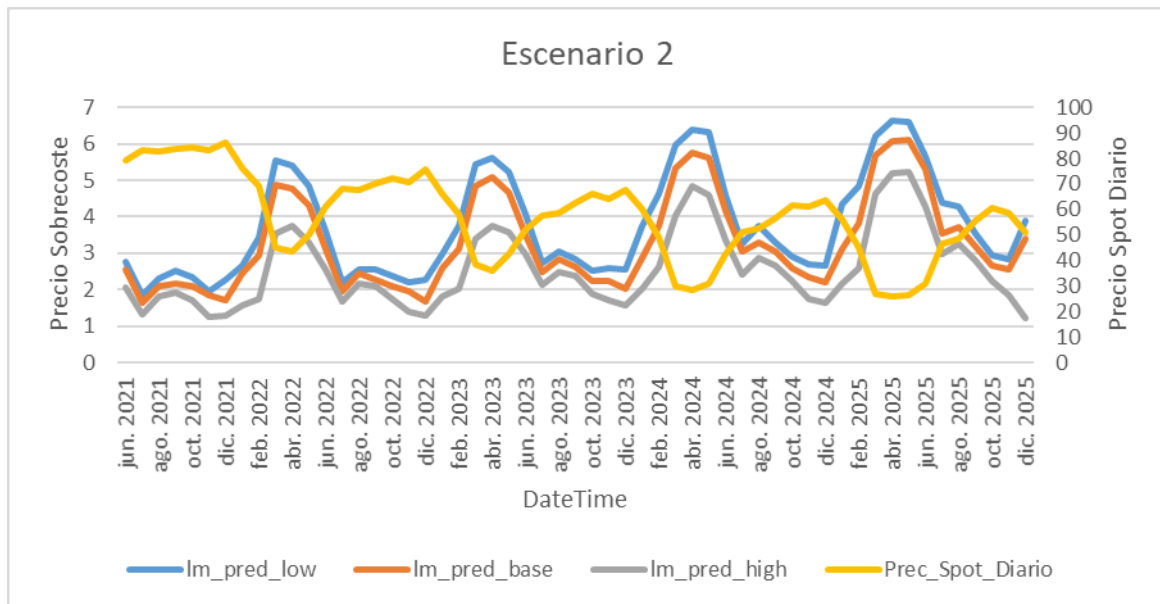


*Figura 45: Predicción del precio de sobrecoste para tres escenarios de la hipótesis 1 junto a la serie del precio de spot base.*

Se puede apreciar que existe una tendencia al alza de los precios del sobrecoste de manera anual, comparando con el precio del spot que tiene una tendencia más a la baja, esto último debido a la posibilidad de que aumenten la generación de tecnologías renovables. Existen unos precios valle que se dan las épocas de verano y otoño. Mientras que los valores más altos se empiezan a dar en la estación invernal llegando a máximos en la estación primaveral. Hay que señalar que, bajo esta hipótesis, el escenario high y base se comportan de manera muy similar.

La segunda hipótesis planteada trata de modelar tres escenarios en los que el precio del spot diario varía. Es decir, se modelan tres escenarios, uno en el que se considera que el precio del Spot diario es bajo (escenario Low), otro alto (escenario High) y un escenario intermedio (escenario Base). Los resultados de los escenarios derivados de la primera hipótesis se muestran en la siguiente gráfica:

*DESARROLLO DEL PROYECTO V: ANÁLISIS DE RESULTADOS.*



*Figura 46: Predicción del precio de sobrecoste para tres escenarios de la hipótesis 2 junto a la serie del precio de spot base.*

Se puede observar que el escenario que da una serie de precio de sobrecoste más alto es el escenario con una estimación de los precios del spot más bajos (escenario low). Esto nos indica que sigue existiendo una cierta correlación negativa entre ambas series, aunque no sea tan acusada como en el histórico de años atrás que se corrobora desde el área de negocio. En esta hipótesis los escenarios están más equiespaciados estando el escenario base entremedias de los escenarios low y high.

## **8. FUTURAS LÍNEAS DE DESARROLLO.**

### **8.1 PUESTA EN PRODUCCIÓN.**

Para la puesta en producción de este proyecto se tratará de crear un cuadro de mando y de un informe detallado para la monitorización del precio del sobrecoste a medio y largo plazo, según las necesidades de negocio:

1. Informe: este documento explicaría de manera detallada las hipótesis que se tienen en cuenta para la ejecución del modelo, el horizonte temporal elegido, las variables usadas en el modelo y un análisis exhaustivo de las tres salidas del modelo según los distintos escenarios. Este informe serviría de apoyo al área de negocio para realizar el “pricing”<sup>15</sup>, coberturas y las ofertas a los clientes.
2. Cuadro de mando: se trataría de integrar el código desarrollado en Python con la herramienta de visualización de Power-BI. En esta herramienta se puede realizar distintas gráficas que representen de manera histórica o detallada las predicciones del precio del sobrecoste realizadas por el modelo. Estas gráficas pueden ir acompañadas de gráficas de apoyo que representen las distintas series o variables consideradas en los escenarios o en el modelo. De esta manera, se busca dar una visión más completa del problema y de la solución propuesta.

Respecto a la periodicidad con la que se ejecutará este procedimiento se valora dos posibles casuísticas:

1. De manera mensual o semanal: bajo una hipótesis definida mostrar las predicciones del precio del sobrecoste de manera mensual o semanal para que los posibles interesados se puedan hacer una idea del posible precio de los sobrecostes a futuro.
2. Bajo demanda de negocio: siempre que se realiza una oferta a un nuevo cliente se tiene en cuenta el valor del precio del sobrecoste. De manera que, si desde negocio

---

<sup>15</sup> Pricing: el método para determinar lo que una compañía va a recibir por sus productos.



se realiza una oferta a un nuevo cliente, es interesante lanzar el proceso para la predicción del precio del sobrecoste bajo ciertas hipótesis definidas juntamente con el equipo de negocio en aras de hacer la mejor oferta al nuevo cliente.

La siguiente imagen muestra un flujograma de la puesta en producción del proyecto:



*Figura 47: Flujograma puesta en producción del modelo.*

## 8.2 MEJORA CONTINUA DEL MODELO.

Se proponen las siguientes mejoras:

1. Recalibrado del modelo en producción: el modelo inicial que se ha puesto en producción es un modelo de regresión lineal multivariante. Las mejoras que se pueden realizar sobre este modelo y que por motivos de tiempo no se han podido desarrollar de manera más exhaustiva son los siguientes:
  - Actualización de datos: se entrenaría el modelo con nuevos datos actuales que no hayan sido recogidos en el histórico de la muestra de entrenamiento. Se recuerda que los modelos probados se han entrenado con un histórico de

tres años, desde 2018 hasta 2020. Se incluirían los meses hasta junio en este caso de 2021 y así sucesivamente.

- Prueba con nuevas variables análisis exhaustivo de residuos: se trataría de realizar un análisis más detallado de los residuos obtenidos en el modelo de regresión lineal multivariante, en aras de mejorar el modelo y encontrar nuevas relaciones, o variables influyentes para la predicción del precio del sobrecoste.
2. Nuevo histórico de datos: se estudiaría la posibilidad de ampliar el horizonte de datos recopilados para el entrenamiento de los modelos. En vez de recoger un histórico de 3 años ampliarlo a 4 o cinco años (desde 2015).
  3. Prueba de modelos con series horarias: los modelos se han entrenado con series diarias, porque en un inicio se considero que los resultados podrían ser más fiables y robustos con una serie diaria que con una serie horaria con más valores atípicos, pese a ello de manera horaria se tiene más datos con los que poder entrenar, lo que podría suponer una ventaja.
  4. Prueba de nuevos modelos: existe una gran variedad de modelos de machine learning. En este proyecto se han probado dos tipos de modelos, más simples: regresión lineal multivariante y árboles de regresión, los que en un principio se ha pensado que mejor podían cubrir las necesidades que requerían desde negocio. Y dos modelos más complejos pertenecientes al grupo de algoritmos de “ensamble methods”: random, forest y stacking. Llegados a este punto se propondría probar nuevos modelos de machine learning como pudieran ser: gradient boosting (“ensamble methods”), técnicas de reducción de dimensionalidad como PCA (Análisis de componentes principales) debido a la gran cantidad de variables que tenemos en el dataframe o implementar algún algoritmo de red neuronal.

## **9. CONCLUSIONES DEL PROYECTO.**

El proyecto realizado ha tratado de abarcar las etapas principales de un proyecto de machine learning: entendimiento del caso de negocio, extracción y preparación de datos, aplicación de modelos, interpretación de resultados y puesta en producción.

La parte de entender el caso de negocio, aunque sea la parte menos técnica es la parte quizá más importante, ya que es la que marca el primer paso y estructura el proyecto.

Sin duda el proceso más tedioso ha sido el referido a la extracción, limpieza y preparación de datos, es el proceso que más tiempo ha llevado y en el que más problemática se ha encontrado.

La parte de modelización e interpretar los resultados se podría considerar el grueso del proyecto. Es la parte que da sentido al proyecto y determina la viabilidad de su puesta en producción para que pueda ser usado o no en función de si responde correctamente o no a las necesidades de negocio. En el caso de este proyecto si se ha podido encontrar una solución satisfactoria al problema que se quería resolver, aunque como se indica en apartados anterior la mejora continua del proyecto es posible.

La puesta en producción ratifica el buen trabajo realizado. Permite una monitorización del problema a resolver y facilita el acceso a la solución.

Cabe señalar la problemática que surge a la hora de convencer y vender un proyecto técnicamente ambicioso que trata de mejorar u optimizar un proceso ya arraigado en la empresa. Se considera un problema común, como se ha comentado sobre ello durante ciertas asignaturas del máster, pero reconfortante una vez la convicción a través de datos se consolida.

El proyecto realizado ha sido un proyecto ambicioso que ha resultado reconfortante realizar y observar que se ha llegado a una buena solución aceptable por parte del área de negocio. Cubrir en la totalidad lo considerado como un proyecto de machine learning da una visión global del trabajo de un data scientist y motiva a seguir en proyectos en esta línea.

---

*CONCLUSIONES DEL PROYECTO.*

El proyecto realizado se estima que tenga una buena repercusión tanto a nivel de resultados tanto monetarios como de ahorro de horas de trabajo. Se trata de una solución que servirá de soporte para el área de negocio y que se considera de gran utilidad.

## **10. BIBLIOGRAFÍA.**

- [1] «[www.energiaysociedad.es](http://www.energiaysociedad.es),» Escuela Técnica Superior de Ingenieros Industriales. Universidad Politécnica de Madrid., [En línea]. Available: <http://www.energiaysociedad.es/manenergia/1-1-aspectos-basicos-de-la-electricidad/>.
- [2] «[www.energiaysociedad.es](http://www.energiaysociedad.es),» Escuela Técnica Superior de Ingenieros Industriales. Universidad Politécnica de Madrid, [En línea]. Available: <http://www.energiaysociedad.es/manenergia/1-2-historia-de-la-electricidad-en-espana/>.
- [3] «[www.energiaysociedad.es](http://www.energiaysociedad.es),» Escuela Técnica Superior de Ingenieros Industriales. Universidad Politécnica de Madrid, [En línea]. Available: <http://www.energiaysociedad.es/manenergia/2-2-el-marco-normativo-espanol/>.
- [4] «[www.energiaysociedad.es](http://www.energiaysociedad.es),» Escuela Técnica Superior de Ingenieros Industriales. Universidad Politécnica de Madrid, [En línea]. Available: <http://www.energiaysociedad.es/manenergia/2-3-actividades-reguladas-y-actividades-en-libre-competencia/>.
- [5] Escuela Técnica Superior de Ingenieros Industriales. Universidad Politécnica de Madrid, «[www.energiaysociedad.es](http://www.energiaysociedad.es),» [En línea]. Available: [http://www.energiaysociedad.es/wp-content/uploads/pdf/documentos/regulacion\\_tarifas/regulacion\\_nacional/PPTLey24\\_2013.pdf](http://www.energiaysociedad.es/wp-content/uploads/pdf/documentos/regulacion_tarifas/regulacion_nacional/PPTLey24_2013.pdf).
- [6] «[www.atalaya.eu](http://www.atalaya.eu),» Atalaya Generación, [En línea]. Available: <https://www.atalaya.eu/renovables/sistema->

electrico.php#:~:text=El%20Sistema%20El%C3%A9ctrico%20Espa%C3%B1ol%20se,bilaterales%20entre%20productores%20y%20comercializadoras.

- [7] «[www.fundacionendesa.org](http://www.fundacionendesa.org),» Endesa Fundación, [En línea]. Available: <https://www.fundacionendesa.org/es/recursos/a201908-el-mercado-electrico>.
- [8] «[www.energiaysociedad.es](http://www.energiaysociedad.es),» Escuela Técnica Superior de Ingenieros Industriales. Universidad Politécnica de Madrid., [En línea]. Available: <http://www.energiaysociedad.es/manenergia/6-1-formacion-de-precios-en-el-mercado-mayorista-diario-de-electricidad/>.
- [9] «[www.iberdrola.com](http://www.iberdrola.com),» Iberdrola, [En línea]. Available: <https://www.iberdrola.com/conocenos/contrato-ppa-energia>.
- [10] «[www.ree.es](http://www.ree.es),» Red Eléctrica de España, [En línea]. Available: <https://www.ree.es/es/actividades/operacion-del-sistema-electrico>.
- [11] «La CNMC investiga la fuerte subida del precio de la luz en la noche de este martes,» *EL INDEPENDIENTE*, 9 mayo 2019.
- [12] «[www.finanzaszone.com](http://www.finanzaszone.com),» [En línea]. Available: <https://finanzaszone.com/analisis-y-prediccion-de-series-temporales-con-r-iii-autocorrelacion/>.
- [13] «[www.cienciadedatos.net](http://www.cienciadedatos.net),» [En línea]. Available: [https://www.cienciadedatos.net/documentos/30\\_cross-validation\\_oneleaveout\\_bootstrap](https://www.cienciadedatos.net/documentos/30_cross-validation_oneleaveout_bootstrap).
- [14] «[sitiobigdata.com](http://sitiobigdata.com),» [En línea]. Available: <https://sitiobigdata.com/2018/08/27/machine-learning-metricas-regresion-mse/>.

- [15] «Ensemble methods: Random Forest,» Universidad Pontificia de Comillas ICAI: Documentación Máster Big Data.
- [16] J. Brownlee, «[www.machinelearningmastery.com](http://www.machinelearningmastery.com),» [En línea]. Available: <https://machinelearningmastery.com/stacking-ensemble-machine-learning-with-python/>.
- [17] «[www.energiaysociedad.es](http://www.energiaysociedad.es),» Escuela Técnica Superior de Ingenieros Industriales. Universidad Politécnica de Madrid, [En línea]. Available: <http://www.energiaysociedad.es/manenergia/6-1-formacion-de-precios-en-el-mercado-mayorista-diario-de-electricidad/>.
- [18] «[www.energiaysociedad.es](http://www.energiaysociedad.es),» Escuela Técnica Superior de Ingenieros Industriales. Universidad Politécnica de Madrid., [En línea]. Available: <http://www.energiaysociedad.es/manenergia/6-1-formacion-de-precios-en-el-mercado-mayorista-diario-de-electricidad/>.
- [19] «[www.bbva.com](https://www.bbva.com),» [En línea]. Available: <https://www.bbva.com/es/que-es-el-hedging-o-cobertura/>.

