Grado en Ingeniería en Tecnologías Industriales

Trabajo de fin de grado

Predicción del perfil del spread entre mercados usando técnicas de aprendizaje automático. Aplicación a Mercados Energéticos.

Autor
Pedrós Tobaruela, Jaime

Director
del Saz-Orozco Huang, Pablo Carlos

Madrid
Julio 2026

Declaración de originalidad

Declaro bajo mi responsabilidad que el Proyecto presentado con el título **Predicción del perfil del spread entre mercados usando técnicas de aprendizaje automático. Aplicación a Mercados Energéticos** e la ETS de Ingeniería – ICAI de la Universidad Pontificia Comillas en el curso académico 2025-2026 es de mi autoría 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.

Uso de Inteligencia Artificial[1]

Declaro bajo mi responsabilidad que (indicar la opción correcta):

☐ No he utilizado Inteligencia Artificial en la elaboración del presente documento.

☑ He utilizado Inteligencia Artificial en la elaboración del presente documento y/o del Anexo B siempre en las condiciones permitidas por la Universidad Pontificia Comillas, es decir, aplicando el Nivel 2 de la Escala de Evaluación de Perkins et al. (2024): "*La IA puede utilizarse para actividades previas a la tarea, como la lluvia de ideas, la descripción y la investigación inicial. Este nivel se centra en el uso de la IA para la planificación, las síntesis y la generación de ideas, pero las evaluaciones deben hacer hincapié en la capacidad de desarrollar y refinar estas ideas de forma independiente*". En concreto, las Inteligencia Artificial ha sido empleada para:

Lluvia de ideas y planificación inicial de la estructura de la memoria (capítulos de datos, clusterización, metodología y resultados) y de la presentación para la defensa. Exploración bibliográfica preliminar sobre la predicción de precios en mercados eléctricos, como punto de partida del estado de la cuestión. Orientación metodológica en la planificación del análisis: sugerencias sobre métricas de evaluación, referencias de comparación entre modelos y análisis complementarios, como el desglose del rendimiento por hora del día o la incorporación de costes de operación. Síntesis y contraste de ideas durante el planteamiento del trabajo. Las ideas resultantes fueron seleccionadas, desarrolladas, implementadas y redactadas de forma independiente por el autor.


Firmado (alumno): Jaime Pedrós Tobaruela
Fecha: 12/07/2026

[1] Esta declaración se refiere al uso de la Inteligencia Artificial generativa para realizar los documentos del Proyecto (Anexo B y Memoria). No aplica a Proyectos donde, por su naturaleza, deban emplear inteligencia artificial como parte de los mismos (aplicación de técnicas de aprendizaje automático, redes neuronales, análisis de datos…)

Autorización para la entrega del Proyecto

El Director del Proyecto	El co-Director del Proyecto (si aplica)
DEL SAZ OROZCO HUANG PABLO CARLOS - 44578440R Firmado digitalmente por DEL SAZ OROZCO HUANG PABLO CARLOS - 44578440R Fecha: 2026.07.12 17:22:48 +02'00'	
Fdo: Pablo del Saz-Orozco	Fdo:
Fecha: 12/07/2026	Fecha:

Grado en Ingeniería en Tecnologías Industriales

Trabajo de fin de grado

Predicción del perfil del spread entre mercados usando técnicas de aprendizaje automático. Aplicación a Mercados Energéticos.

Autor
Pedrós Tobaruela, Jaime

Director
del Saz-Orozco Huang, Pablo Carlos

Madrid
Julio 2026

PREDICCIÓN DEL PERFIL DEL SPREAD ENTRE MERCADOS USANDO TÉCNICAS DE APRENDIZAJE AUTOMÁTICO. APLICACIÓN A MERCADOS ENERGÉTICOS

Autor: Pedrós Tobaruela, Jaime.

Director: del Saz-Orozco Huang, Pablo Carlos.

Entidad Colaboradora: ICAI - Universidad Pontificia Comillas.

RESUMEN DEL PROYECTO

Introducción

En el mercado eléctrico ibérico, el mercado diario fija cada mediodía un precio para cada una de las veinticuatro horas del día siguiente, y los mercados intradiarios corrigen después esos precios a medida que llega información nueva sobre demanda, generación renovable o disponibilidad del sistema. La diferencia entre el precio de la primera sesión intradiaria y el precio diario para una misma hora, denominada *spread*, condensa esa corrección y tiene un valor económico directo: anticiparla permite decidir, antes de que el mercado la revele, en qué mercado conviene comprar y en cuál conviene vender.

La creciente penetración de las energías renovables y episodios como la crisis energética de 2022 han incrementado sustancialmente la volatilidad de los precios eléctricos, lo que ha impulsado el interés por la predicción de precios en mercados eléctricos (*Electricity Price Forecasting*, EPF), un campo que ha evolucionado en las últimas dos décadas desde los modelos estadísticos clásicos hacia el aprendizaje automático y, más recientemente, el aprendizaje profundo. Sin embargo, la inmensa mayoría de esta literatura se centra en predecir el nivel de precios de un único mercado, normalmente el diario; la predicción de la diferencia entre mercados ha recibido mucha menos atención, a pesar de la creciente relevancia de los mercados intradiarios en sistemas con alta penetración renovable.

Este Trabajo Fin de Grado se sitúa en ese hueco. Su objetivo general es desarrollar una herramienta de apoyo a la decisión de compra y venta de energía a

partir de la predicción del *spread* horario, utilizando exclusivamente información disponible antes del cierre del mercado diario. De él se derivan cinco objetivos específicos: construir una base de datos homogénea que permita analizar el *spread*; desarrollar un modelo de agrupación de días a partir de información anticipada; caracterizar los perfiles de *spread* de cada grupo; desarrollar modelos predictivos directos, en formulación de regresión y de clasificación de signo; y evaluar el impacto económico de dichos modelos mediante una simulación de mercado con costes de operación y contrastes de significación estadística.

Metodología

El trabajo integra, con resolución horaria y ámbito peninsular, los precios de mercado, las previsiones de demanda y de generación renovable, la disponibilidad de potencia nuclear, la potencia instalada, el precio del gas natural TTF como variable exógena de coste, variables meteorológicas agregadas de ocho ciudades de referencia y un calendario laboral enriquecido con el porcentaje de población afectada por cada festivo. Todo el proceso respeta un criterio estrictamente anticipado: solo se emplea información disponible antes del cierre del mercado diario. Tras la depuración y homogeneización de fuentes, la muestra analítica final quedó formada por 1.763 días entre enero de 2019 y diciembre de 2025.

El análisis se desarrolla en dos fases. En la primera, técnicas de agrupación no supervisada (K-means) organizan los días en ocho regímenes de mercado a partir únicamente de las variables explicativas disponibles con antelación, sin utilizar el *spread* como entrada. Se compararon 35 configuraciones de K-means y clusterización jerárquica, evaluadas mediante los índices de Silhouette, Calinski-Harabasz y Davies-Bouldin. En la segunda fase se plantea la predicción supervisada del *spread* horario mediante *gradient boosting* (LightGBM), en dos formulaciones complementarias que comparten la misma matriz de características: una de regresión, que estima el valor continuo del *spread* con una función de pérdida robusta de tipo Huber, y otra de clasificación, que predice directamente su signo. Ambos modelos se ajustan mediante búsqueda bayesiana de hiperparámetros con Optuna, optimizando directamente el beneficio económico en el caso del regresor y el área bajo la curva ROC en el caso del clasificador, y se evalúan siempre mediante una validación estrictamente temporal de ventana creciente que respeta el orden cronológico de las observaciones.

Resultados

La solución de referencia de la clusterización, K-means con ocho grupos, recupera de forma emergente los principales regímenes históricos del mercado eléctrico

español (el periodo previo a la crisis energética, la propia crisis del gas de 2022 y el régimen reciente de expansión fotovoltaica) e identifica una tipología especialmente nítida: los días festivos, un grupo pequeño pero con el *spread* más negativo y extremo de la muestra. Los contrastes de ANOVA y Kruskal-Wallis confirman diferencias significativas en el *spread* medio diario entre clústeres para las soluciones de siete y ocho grupos ($p < 0{,}001$), mientras que una solución más agregada de solo tres grupos no alcanza significación estadística.

En cuanto a la predicción, la magnitud exacta del *spread* resulta difícil de anticipar ($R^2 = 0{,}018$), pero su dirección contiene una señal anticipable real: el regresor alcanza un área bajo la curva ROC de 0,590 y el clasificador de 0,584, frente al 0,565 de la persistencia y el 0,500 del azar. La Tabla 1 y la Figura 1 resumen la comparación entre modelos y referencias.

Modelo	Acc	Bal. acc	AUC	PnL total	Hit
Regresor (Optuna)	0,594	0,561	0,590	7.022	0,584
Clasificador (Optuna)	0,616	0,523	0,584	7.676	0,575
Persistencia	0,561	0,541	0,565	5.146	0,555
Clase mayoritaria	0,604	0,500	0,500	5.662	0,557
Aleatorio	0,509	0,508	0,500	1.230	0,510

Tabla 1: Comparación de modelos y referencias en el conjunto de prueba.

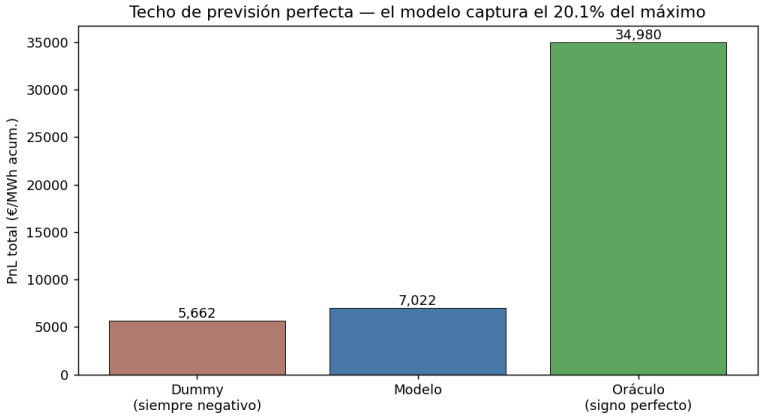

Figura 1: PnL del modelo y de la clase mayoritaria frente al máximo alcanzable mediante una previsión perfecta (oráculo). El modelo obtiene un 24 % más de beneficio que la clase mayoritaria y captura en torno al 20 % del máximo teórico.

Traducida en una estrategia de operación con posiciones de 1 MWh por hora, el regresor acumula un beneficio de 7.022 € en el conjunto de prueba, frente a los 5.662 € de la regla de apostar siempre al signo mayoritario y a los 34.980 € que

obtendría un oráculo con conocimiento perfecto del signo (el modelo captura en torno al 20 % del máximo teórico). Esta ventaja resulta estadísticamente significativa mediante un test de permutación ($p = 0{,}0005$). El análisis desagregado por hora del día muestra que la ventaja del modelo no se concentra donde el signo es más previsible, sino donde el sesgo estructural negativo del mercado se debilita: en las horas centrales del día, coincidiendo con la producción fotovoltaica, la regla de apostar siempre a la baja llega a perder dinero, mientras que el modelo mantiene un resultado positivo en las veinticuatro horas.

Conclusiones

El *spread* entre el mercado diario y el intradiario del MIBEL es, con la información disponible antes de la decisión de mercado, un fenómeno mayoritariamente ruidoso pero no completamente impredecible. Los modelos desarrollados capturan una fracción modesta, en torno a una quinta parte, del beneficio que sería alcanzable con información perfecta, y esa fracción depende del régimen del sistema y del tramo horario considerado. La etiqueta de clúster, incorporada como variable de entrada del modelo predictivo, resultó no aportar información incremental, aunque la propia clusterización sí resulta útil para explicar dónde y por qué el modelo genera valor. Entre las principales limitaciones del estudio destacan el uso de observaciones meteorológicas en lugar de previsiones históricas y la naturaleza necesariamente simplificada de la estrategia de simulación económica empleada. Como líneas de trabajo futuro se proponen la regresión por cuantiles, la extensión del análisis a otras sesiones del mercado intradiario, la incorporación de previsiones meteorológicas históricas y la comparación con arquitecturas de aprendizaje profundo.

Referencias

[1] Ciarreta, A., Pizarro-Irizar, C. y Zarraga, A. (2020). Renewable energy regulation and structural breaks: An empirical analysis of Spanish electricity price volatility. *Energy Economics*, 88, 104749.

[2] Diaz, G., Coto, J. y Gomez-Aleixandre, J. (2019). Prediction and explanation of the formation of the Spanish day-ahead electricity price through machine learning regression. *Applied Energy*, 239, 610-625.

[3] Jedrzejewski, A., Lago, J., Marcjasz, G. y Weron, R. (2022). Electricity price forecasting: The dawn of machine learning. *IEEE Power and Energy Magazine*, 20(3), 24-31.

[4] O'Connor, C., Bahloul, M., Prestwich, S. y Visentin, A. (2025). A review

of electricity price forecasting models in the day-ahead, intra-day, and balancing markets. *Energies*, 18(12), 3097.

[5] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q. y Liu, T. (2017). LightGBM: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems*, 30, 3146-3154.

[6] Akiba, T., Sano, S., Yanase, T., Ohta, T. y Koyama, M. (2019). Optuna: A next-generation hyperparameter optimization framework. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2623-2631.

PREDICTION OF THE SPREAD PROFILE BETWEEN MARKETS USING MACHINE LEARNING TECHNIQUES. APPLICATION TO ENERGY MARKETS

Introduction

In the Iberian electricity market, the day-ahead market sets a price for every hour of the following day at noon, and the intraday markets then correct those prices as new information arrives on demand, renewable generation or system availability. The difference between the price of the first intraday session and the day-ahead price for the same hour, referred to as the *spread*, condenses that correction and has direct economic value: anticipating it makes it possible to decide, before the market reveals it, where it is worth buying and where it is worth selling.

The growing penetration of renewable energy and episodes such as the 2022 energy crisis have substantially increased electricity price volatility, which has driven interest in Electricity Price Forecasting (EPF), a field that has evolved over the last two decades from classical statistical models towards machine learning and, more recently, deep learning. However, the vast majority of this literature focuses on forecasting the price level of a single market, usually the day-ahead one; forecasting the difference between markets has received much less attention, despite the growing relevance of intraday markets in systems with high renewable penetration.

This Bachelor Thesis addresses that gap. Its general objective is to develop a decision-support tool for buying and selling energy based on the prediction of the hourly *spread*, using exclusively information available before the closure of the day-ahead market. Five specific objectives follow from it: building a homogeneous database that allows the *spread* to be analysed; developing a day-clustering model based on information available in advance; characterising the *spread* profiles of each group; developing direct predictive models, both in regression and sign classification form; and evaluating the economic impact of these models through a market simulation with transaction costs and statistical significance tests.

Methodology

The work integrates, at hourly resolution and peninsular scope, market prices, demand and renewable generation forecasts, nuclear power availability, installed capacity, the Dutch TTF natural gas price as an exogenous cost variable, aggregated weather variables from eight reference cities, and a working-day calendar enri-

ched with the percentage of population affected by each public holiday. The whole process follows a strictly forward-looking criterion: only information available before the closure of the day-ahead market is used. After cleaning and homogenising the sources, the final analytical sample consisted of 1,763 days between January 2019 and December 2025.

The analysis unfolds in two stages. In the first, unsupervised clustering (K-means) groups the days into eight market regimes based solely on the explanatory variables available in advance, without using the *spread* as an input. Thirty-five configurations of K-means and hierarchical clustering were compared, evaluated using the Silhouette, Calinski-Harabasz and Davies-Bouldin indices. In the second stage, the hourly *spread* is predicted through gradient boosting (LightGBM), in two complementary formulations that share the same feature matrix: a regression that estimates the continuous value of the *spread* with a robust Huber loss, and a classification that directly predicts its sign. Both models are tuned through Bayesian hyperparameter search with Optuna, directly optimising economic profit for the regressor and the area under the ROC curve for the classifier, and are always evaluated using a strictly temporal expanding-window validation that respects the chronological order of the observations.

Results

The reference clustering solution, K-means with eight groups, recovers in an emergent way the main historical regimes of the Spanish electricity market (the period before the energy crisis, the 2022 gas crisis itself, and the recent regime of photovoltaic expansion) and identifies a particularly clear typology: public holidays, a small group with the most negative and extreme *spread* in the sample. ANOVA and Kruskal-Wallis tests confirm significant differences in the mean daily *spread* between clusters for the seven- and eight-group solutions ($p < 0{,}001$), whereas a more aggregated three-group solution does not reach statistical significance.

Regarding prediction, the exact magnitude of the *spread* proves difficult to anticipate ($R^2 = 0{,}018$), but its direction carries a real anticipable signal: the regressor reaches an area under the ROC curve of 0.590 and the classifier of 0.584, compared with 0.565 for persistence and 0.500 for chance. Table 1 and Figure 1 summarise the comparison between models and baselines.

Translated into a trading strategy with 1 MWh positions per hour, the regressor accumulates a profit of €7,022 on the test set, compared with €5,662 for the rule of always betting on the majority sign and €34,980 that would be obtained by an oracle with perfect knowledge of the sign (the model captures around 20 % of the theoretical maximum). This advantage is statistically significant according to a

Model	Acc	Bal. acc	AUC	Total PnL	Hit
Regressor (Optuna)	0.594	0.561	0.590	7,022	0.584
Classifier (Optuna)	0.616	0.523	0.584	7,676	0.575
Persistence	0.561	0.541	0.565	5,146	0.555
Majority class	0.604	0.500	0.500	5,662	0.557
Random	0.509	0.508	0.500	1,230	0.510

Tabla 2: Comparison of models and baselines on the test set.

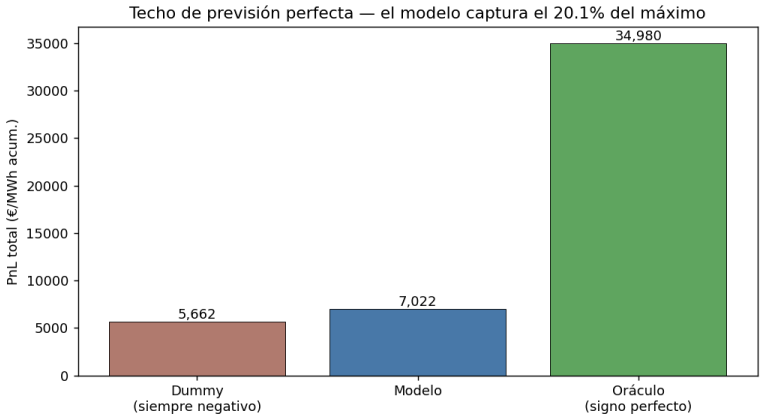

Figura 2: PnL of the model and of the majority class against the maximum achievable through a perfect forecast (oracle). The model obtains 24 % more profit than the majority class and captures around 20 % of the theoretical maximum.

permutation test ($p = 0{,}0005$). The hour-by-hour analysis shows that the model's advantage is not concentrated where the sign is most predictable, but where the market's negative structural bias weakens: during the central hours of the day, coinciding with photovoltaic production, the rule of always betting downward ends up losing money, while the model maintains a positive result across all twenty-four hours.

Conclusions

The *spread* between the MIBEL day-ahead and intraday markets is, with the information available before the market decision, a largely noisy but not entirely unpredictable phenomenon. The models developed capture a modest fraction, around one-fifth, of the profit that would be attainable with perfect information, and that fraction depends on the system's regime and on the hour of day considered. The cluster label, included as an input variable of the predictive model, was found to add no incremental information, although the clustering itself remains

useful for explaining where and why the model generates value. The main limitations of the study include the use of observed weather data rather than historical forecasts and the necessarily simplified nature of the economic simulation strategy employed. Proposed lines of future work include quantile regression, extending the analysis to other intraday market sessions, incorporating historical weather forecasts and comparing the approach with deep learning architectures.

References

[1] Ciarreta, A., Pizarro-Irizar, C. and Zarraga, A. (2020). Renewable energy regulation and structural breaks: An empirical analysis of Spanish electricity price volatility. *Energy Economics*, 88, 104749.

[2] Diaz, G., Coto, J. and Gomez-Aleixandre, J. (2019). Prediction and explanation of the formation of the Spanish day-ahead electricity price through machine learning regression. *Applied Energy*, 239, 610-625.

[3] Jedrzejewski, A., Lago, J., Marcjasz, G. and Weron, R. (2022). Electricity price forecasting: The dawn of machine learning. *IEEE Power and Energy Magazine*, 20(3), 24-31.

[4] O'Connor, C., Bahloul, M., Prestwich, S. and Visentin, A. (2025). A review of electricity price forecasting models in the day-ahead, intra-day, and balancing markets. *Energies*, 18(12), 3097.

[5] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q. and Liu, T. (2017). LightGBM: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems*, 30, 3146-3154.

[6] Akiba, T., Sano, S., Yanase, T., Ohta, T. and Koyama, M. (2019). Optuna: A next-generation hyperparameter optimization framework. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2623-2631.

Agradecimientos

A mi director, Pablo Carlos del Saz-Orozco Huang, por proponerme el tema de este trabajo, por orientarme en la búsqueda de bibliografía y por su disponibilidad para resolver las dudas que le planteé.

A mi familia, por su apoyo incondicional durante estos años de carrera y muy especialmente, durante los meses de elaboración de este trabajo. Gracias en particular a mi hermana Sofía, que ya conocía de primera mano el mundo de la inteligencia artificial y de las herramientas empleadas en este proyecto, y cuyas conversaciones, ideas y paciencia para explicarme lo que a mí me costaba entender han dejado huella en más de una página de esta memoria.

Y, por supuesto, a todos los que de un modo u otro han acompañado este camino.

Índice general

Índice de figuras

Índice de tablas

Capítulo 1

Introducción

1.1. Contexto: el mercado ibérico de electricidad

Los mercados eléctricos constituyen uno de los pilares fundamentales del sistema energético moderno, actuando como el principal mecanismo para la asignación eficiente de recursos y la formación de precios de la electricidad. Su papel resulta especialmente relevante en el corto plazo: son los mercados de entrega más próxima al tiempo real (el mercado diario y los mercados intradiarios) los que coordinan, hora a hora, la posición final de generadores y consumidores a medida que se dispone de información cada vez más precisa sobre el estado del sistema.

En el contexto de la península ibérica, estos mercados se organizan dentro del Mercado Ibérico de Electricidad (MIBEL), operativo desde mediados de la década de 2000, que integra a España y Portugal con el objetivo de fomentar la competencia, la transparencia y la eficiencia en la negociación de energía eléctrica. Dentro del MIBEL, el Operador del Mercado Ibérico de Energía – Polo Español (OMIE) es el responsable de la organización y supervisión del mercado diario y de los mercados intradiarios, conocidos conjuntamente como mercado *spot*, donde se casan las ofertas de compra y venta de electricidad con distintos horizontes temporales. Por su parte, el Operador del Mercado Ibérico de Energía – Polo Portugués (OMIP) gestiona los mercados a plazo, orientados a la cobertura de posiciones a medio y largo plazo.

El número y la configuración de las sesiones intradiarias no han sido estables en el tiempo: han evolucionado con las sucesivas reformas de acoplamiento de mercados a escala europea, orientadas a integrar la negociación de corto plazo entre

países y a facilitar el ajuste continuo de posiciones. Esta evolución es relevante para el planteamiento de este trabajo, ya que condiciona qué información está efectivamente disponible en cada momento y obliga a acotar con precisión, como se hará más adelante, a qué sesión concreta del mercado intradiario se refiere el análisis.

La Tabla 1.1 resume los horarios límite establecidos en las Reglas de Funcionamiento del Mercado para la subasta diaria y para las seis sesiones del mercado intradiario de subastas de ámbito MIBEL vigentes en el momento de realización de este trabajo. Las tres primeras sesiones se celebran el día anterior a la entrega y cubren el conjunto de las veinticuatro horas del día siguiente; las tres últimas se celebran ya durante el propio día de entrega y cubren únicamente las horas restantes de esa jornada. Junto a estas subastas, un mercado intradiario continuo de ámbito europeo (*Single Intraday Coupling*, SIDC) permite ajustar posiciones de forma continua hasta una hora antes de la entrega física [1].

Tabla 1.1: Cronograma horario de la subasta diaria y de las seis sesiones del mercado intradiario de subastas del MIBEL. D indica el día en que se celebra la sesión; D+1, el día de entrega. Fuente: OMIE, *Detalle del funcionamiento del mercado intradiario* [1].

Mercado	Día	Apertura	Cierre / casación	Horizonte de programación
Diario	D	–	12:00	24 h (1–24, D+1)
Intradiario, sesión 1	D	14:00	15:00	24 h (1–24, D+1)
Intradiario, sesión 2	D	17:00	17:50	28 h (21–24 D, 1–24 D+1)
Intradiario, sesión 3	D	21:00	21:50	24 h (1–24, D+1)
Intradiario, sesión 4	D+1	01:00	01:50	20 h (5–24)
Intradiario, sesión 5	D+1	04:00	04:50	17 h (8–24)
Intradiario, sesión 6	D+1	09:00	09:50	12 h (13–24)

En los últimos años, la creciente penetración de las energías renovables, especialmente la eólica y la solar, ha incrementado de forma notable la volatilidad de los precios eléctricos, sobre todo en sistemas con baja flexibilidad y alta participación renovable [2, 3, 4, 5, 6]. Esta volatilidad está fuertemente influida por factores exógenos como las condiciones meteorológicas, la evolución de la demanda eléctrica, la disponibilidad de generación convencional o el carácter laboral o festivo de los días, lo que complica la anticipación del comportamiento de los mercados. A ello se sumó la crisis energética de 2022, que llevó los precios del gas y de la electricidad europeos a niveles sin precedentes: episodios de este tipo muestran que el sistema no atraviesa un régimen único y estable, sino sucesivos periodos con dinámicas de precio distintas, una idea que se retomará con detalle en la clusterización

de días del capítulo 3.

En este contexto, la capacidad de anticipar el perfil del *spread* entre los distintos mercados eléctricos representa una herramienta de gran valor para los agentes que operan en ambos mercados simultáneamente: generadores que deciden en qué mercado asignar su producción, comercializadoras que gestionan carteras de clientes expuestas a ambos precios, o agregadores que ajustan posiciones de pequeños productores y consumidores flexibles. Para todos ellos, anticipar el signo y la magnitud del *spread* permite optimizar las estrategias de compra y venta de energía y mejorar la gestión del riesgo asociado a la volatilidad de precios, sin necesidad de esperar a que el mercado intradiario revele la corrección de precio a posteriori.

El presente Trabajo Fin de Grado se centra en el análisis y la predicción de dicho *spread* de precios entre el mercado diario y los mercados intradiarios del MIBEL mediante el uso de técnicas de aprendizaje automático, con el objetivo de identificar patrones recurrentes en su comportamiento y de desarrollar modelos predictivos que se apoyen exclusivamente en información disponible con antelación a la decisión de mercado.

1.2. El problema: el *spread* entre el mercado diario y el intradiario

Este trabajo se centra en esa diferencia, conocida como *spread* entre mercados: para cada hora de entrega, la resta entre el precio del mercado intradiario y el precio que esa misma hora obtuvo en el mercado diario. Su interpretación económica es directa. Si el intradiario cotiza por encima del diario, un agente que hubiese comprado en el diario y vendido en el intradiario habría obtenido un margen positivo; si cotiza por debajo, la posición rentable habría sido la contraria. Más allá de esta lectura de arbitraje, el *spread* condensa información valiosa para cualquier agente que opere en ambos mercados: para una comercializadora o un generador, anticipar su signo y su magnitud ayuda a decidir en qué mercado situar la energía, cuánto margen reservar para el ajuste fino y cómo gestionar el riesgo de precio asociado a la operativa [2, 4].

Dos rasgos hacen que este problema sea a la vez interesante y difícil. El primero es que el *spread* no se comporta igual a todas las horas: su perfil intradía cambia con el contexto del sistema, de modo que el objeto natural de predicción no es un único valor diario sino el perfil horario completo. El segundo es que se trata de una variable muy ruidosa, concentrada en torno a cero pero con colas pesadas,

en la que separar la señal anticipable del ruido irreducible constituye en sí mismo una pregunta de investigación. De las distintas sesiones del mercado intradiario, el trabajo emplea la primera, por ser la de mayor cobertura horaria y liquidez y la más próxima en el tiempo al cierre del mercado diario.

1.3. Estado de la cuestión

La predicción de precios en mercados eléctricos, conocida en la literatura como *Electricity Price Forecasting* (EPF), ha evolucionado significativamente en las últimas dos décadas, pasando de los modelos estadísticos tradicionales a técnicas avanzadas de aprendizaje automático [7]. Históricamente, la predicción de series temporales financieras y energéticas se apoyaba en modelos de regresión lineal y en métodos como ARIMA. Estos enfoques sentaron las bases para comprender la dinámica de los precios, pero presentan limitaciones importantes frente a la volatilidad, los picos abruptos y las relaciones no lineales que caracterizan a los mercados modernos [8, 7]. Estudios recientes muestran que, aunque ARIMA puede capturar tendencias generales, suele fallar al anticipar cambios repentinos o extremos en los precios [9].

Para superar estas limitaciones, la investigación se ha orientado hacia los algoritmos de aprendizaje automático. Técnicas como los árboles de decisión, los bosques aleatorios o las máquinas de vectores soporte emergieron como alternativas robustas, capaces de modelar interacciones complejas y no lineales entre variables [10]. Dentro de los métodos de conjunto, las técnicas de *boosting*, que entrenan modelos sencillos de forma secuencial, corrigiendo cada uno los errores de los anteriores, se han consolidado como una herramienta de referencia. En el contexto específico del mercado eléctrico español, Díaz et al. [10] validaron la eficacia de los árboles de regresión con gradiente, mostrando que estos modelos no solo reducen el error respecto a los métodos tradicionales, sino que capturan la influencia no lineal de variables exógenas como la generación eólica o la disponibilidad térmica en la formación del precio. Implementaciones modernas de esta familia, como XGBoost o LightGBM, han mostrado un rendimiento destacado en la predicción de precios a corto plazo [9].

En paralelo, una parte creciente de la literatura ha adoptado técnicas de aprendizaje profundo. Las redes neuronales recurrentes y, en particular, las redes LSTM permiten aprender dependencias de largo plazo en las series de precios [11, 12], y arquitecturas híbridas que combinan capas convolucionales con LSTM han superado en precisión tanto a los modelos estadísticos como a los de *boosting* en algunos

mercados europeos [9]. El aprendizaje profundo se emplea también en aplicaciones directamente ligadas a la operativa, como la puja virtual en mercados mayoristas [13], y trabajos recientes exploran cuestiones de diseño experimental como la longitud óptima de las ventanas de entrenamiento en presencia de cambios de régimen [14].

Ahora bien, la inmensa mayoría de esta literatura se centra en predecir el precio de un mercado concreto, casi siempre el diario. La predicción de la *diferencia* de precios entre el mercado diario y el intradiario ha recibido una atención mucho menor: las revisiones recientes del área señalan que los mercados intradiarios y de balance están notablemente menos estudiados que el diario, a pesar de su importancia creciente en sistemas con alta penetración renovable [15]. Este trabajo se sitúa en ese hueco: en lugar de predecir el nivel de precios, aborda directamente el *spread* horario entre ambos mercados para el caso español, con dos señas de identidad metodológicas: el uso estricto de información disponible antes del cierre del mercado diario y la evaluación de los modelos no solo con métricas estadísticas, sino también con una métrica económica que simula la operativa. En cuanto a la técnica, se opta por el *gradient boosting* sobre árboles en lugar del aprendizaje profundo, una elección adecuada para abordar un problema de series temporales mediante una representación tabular con variables heterogéneas y retardos explícitos, sobre muestras de tamaño moderado como las de este problema, por las razones que se detallan en el capítulo 4.

1.4. Motivación

La motivación principal de este proyecto surge de la necesidad de disponer de herramientas que permitan reducir la incertidumbre asociada a la operativa en mercados eléctricos. La volatilidad de precios, y en particular la asociada al *spread* entre los mercados diario e intradiario, ha sido documentada en numerosos estudios, que muestran que los incrementos de generación renovable intermitente aumentan la frecuencia de episodios de precios extremos y de oscilaciones horarias significativas [2, 4, 3, 5, 6]. Esta volatilidad genera riesgos económicos para los agentes, especialmente en periodos de márgenes ajustados, de modo que incluso mejoras modestas en la capacidad de anticipar el *spread* pueden traducirse en un beneficio tangible. Desde un punto de vista práctico, anticipar su signo y su magnitud permite diseñar estrategias de compra y venta más eficientes, ajustando la decisión de operar al grado de confianza de la predicción.

Desde el punto de vista académico, el proyecto resulta interesante por aplicar

técnicas de aprendizaje automático a un problema real y poco explorado del sector energético. Estudios recientes han demostrado que los modelos de aprendizaje automático y profundo pueden capturar patrones complejos en series de precios eléctricos, superando a los métodos estadísticos clásicos en precisión y adaptabilidad [15, 14, 13]. La combinación que aquí se propone (agrupación no supervisada de días y predicción supervisada directa) permite evaluar dos enfoques distintos sobre el mismo problema y analizar sus ventajas y limitaciones.

A estas dos motivaciones se añade una tercera, de carácter metodológico. En un problema tan ruidoso como este, es fácil producir evaluaciones engañosamente optimistas: basta con mezclar información futura en el entrenamiento o con compararse contra referencias poco exigentes. Por ello, el trabajo pone un énfasis particular en cuantificar con honestidad qué parte del *spread* es realmente anticipable: validación estrictamente temporal, comparación con reglas de referencia exigentes, contrastes de significación estadística y análisis del efecto de los costes de transacción. El resultado de interés no es solo el rendimiento de los modelos, sino la delimitación rigurosa de su alcance real.

1.5. Objetivos del trabajo

El objetivo general del proyecto es desarrollar una herramienta de apoyo a la toma de decisiones en la compra y venta de energía, orientada a determinar, a partir de la predicción del *spread* de precios entre el mercado diario y el intradiario, qué opción resulta más favorable en cada situación, utilizando únicamente información disponible antes del cierre del mercado diario. Este objetivo general se concreta en cinco objetivos específicos:

1. Caracterizar el comportamiento del spread horario y su relación con las variables explicativas del sistema, a partir de una base de datos homogénea de ámbito peninsular y resolución horaria construida específicamente para este trabajo.
2. Desarrollar un modelo de agrupación de días basado exclusivamente en variables disponibles con antelación al cierre del mercado diario.
3. Caracterizar los perfiles de *spread* asociados a cada grupo de días identificado y valorar su utilidad operativa.
4. Desarrollar modelos predictivos directos del *spread* horario mediante técnicas de aprendizaje automático, tanto en formulación de regresión como de

clasificación direccional.

5. Evaluar el impacto económico de los modelos desarrollados mediante la simulación de una estrategia de mercado, incorporando costes de operación y umbrales de confianza, y contrastar la significación estadística de los resultados.

1.6. Enfoque metodológico y alcance

La resolución del problema se aborda en fases sucesivas que se corresponden con la estructura del documento. En una primera fase se recopilan y depuran series históricas de precios del mercado diario y de la primera sesión del intradiario, previsiones de demanda y de generación renovable [1], disponibilidad de potencia nuclear, potencia instalada, precio del gas natural como variable exógena de coste, información meteorológica y variables de calendario, construidas estas últimas con una ponderación poblacional del carácter festivo de cada fecha. El resultado es una base analítica de ámbito peninsular con resolución horaria que cubre el periodo 2019-2025.

En la segunda fase se aplican técnicas de aprendizaje no supervisado (*clustering*) para agrupar los días con características similares a partir de las variables explicativas, sin utilizar el *spread* como entrada, y se analiza si los grupos resultantes presentan perfiles de *spread* diferenciados. En la tercera fase se plantea la predicción supervisada del *spread* horario mediante *gradient boosting*, en dos formulaciones complementarias: la regresión de su valor y la clasificación de su signo. Todo el proceso respeta un criterio estrictamente anticipado, solo se emplea información disponible antes del cierre del mercado diario, y una validación temporal con ventana creciente que imita las condiciones reales de despliegue. Por último, los modelos se someten a una evaluación económica que simula la estrategia de operación, incorpora costes por operación y umbrales de confianza, y contrasta si la ventaja obtenida frente a reglas de referencia es estadísticamente significativa.

Conviene delimitar desde el principio el alcance del estudio. El análisis se circunscribe al sistema peninsular español, al periodo 2019-2025 y, esencialmente, a los días laborables, dado que la cotización del gas natural (una de las variables

[1] Una alternativa no explorada en este trabajo habría sido emplear la demanda térmica prevista (la diferencia entre la demanda prevista y la generación renovable prevista) en lugar de ambas variables por separado. Esta magnitud sintética podría generalizar mejor a periodos futuros con un mix de generación distinto, cuestión que se retoma en la sección 3.11.

empleadas) solo está disponible en esos días [2]. Las variables meteorológicas corresponden a observaciones y no a previsiones históricas, lo que constituye una aproximación cuyas implicaciones se discuten en los capítulos 2 y 5. La simulación económica, por su parte, emplea una regla de operación deliberadamente sencilla, con volumen constante, concebida para comparar modelos entre sí y no como estrategia lista para su aplicación.

1.7. Alineación con los Objetivos de Desarrollo Sostenible

El proyecto se alinea con varios Objetivos de Desarrollo Sostenible (ODS) definidos por las Naciones Unidas, principalmente desde una perspectiva metodológica y de apoyo a la toma de decisiones en el ámbito de los mercados energéticos.

En relación con el ODS 7 (Energía asequible y no contaminante), el trabajo contribuye de forma indirecta al estudio del funcionamiento eficiente de los mercados eléctricos, analizando el comportamiento del *spread* entre distintos mercados y desarrollando herramientas predictivas que permiten comprender mejor la dinámica de formación de precios. Una mayor capacidad de análisis y anticipación de los precios puede apoyar decisiones económicas más informadas dentro del sector energético.

Asimismo, el trabajo se relaciona con el ODS 9 (Industria, innovación e infraestructura) mediante la aplicación de técnicas de aprendizaje automático y análisis de datos a un problema real del sector energético. El desarrollo y la evaluación de modelos predictivos contribuyen a la innovación en el uso de herramientas digitales y analíticas, alineándose con los objetivos de modernización y mejora de las infraestructuras de información.

Por último, el proyecto guarda relación con el ODS 13 (Acción por el clima) en la medida en que se enmarca en el análisis de mercados energéticos en un contexto de creciente variabilidad. Aunque el proyecto no aborda de forma directa aspectos medioambientales, el estudio de herramientas que mejoran la comprensión del comportamiento de los precios resulta relevante para optimizar la integración de las fuentes renovables y asegurar la sostenibilidad económica.

[2] Una alternativa habría sido propagar hacia adelante la última cotización disponible (p. ej. la del viernes precedente) para cubrir fines de semana y festivos, lo que habría permitido ampliar la muestra más allá de los días laborables. No se aplicó en este trabajo.

1.8. Estructura del documento

El resto de la memoria se organiza como sigue. El capítulo 2 describe la obtención y construcción de la base de datos: fuentes empleadas, definición de la variable objetivo, bloques de variables explicativas y proceso de depuración. El capítulo 3 presenta la clusterización de días: la construcción de la matriz de agrupación, los métodos considerados, la selección de la solución de referencia y la caracterización de los grupos y de sus perfiles de *spread*. El capítulo 4 plantea la metodología de predicción supervisada: formulaciones del problema, matriz de características, protocolo de evaluación, modelos de referencia y ajuste de hiperparámetros. El capítulo 5 recoge los resultados: comparación de modelos, interpretación de las variables, análisis operativo y de robustez, y discusión de limitaciones y líneas de extensión. Finalmente, el capítulo 6 resume las conclusiones del trabajo.

Capítulo 2

Obtención y construcción de la base de datos

2.1. Introducción

La calidad de cualquier modelo predictivo depende en gran medida de la calidad de la base de datos utilizada para entrenarlo y evaluarlo. En un problema como el planteado en este trabajo, centrado en la predicción del *spread* entre mercados eléctricos, la fase de obtención de datos adquiere una importancia especialmente elevada por dos motivos. En primer lugar, porque el fenómeno que se pretende explicar depende de un conjunto amplio de factores exógenos y de sistema, entre ellos la demanda prevista, la generación renovable esperada, la disponibilidad de generación convencional, las condiciones meteorológicas y el carácter laboral o festivo del día. En segundo lugar, porque no todas las fuentes publican la información con la misma frecuencia, con el mismo horizonte temporal ni con el mismo ámbito geográfico.

El objetivo de este capítulo es describir de forma sistemática el proceso seguido para construir la base de datos empleada en el trabajo, desde la selección de fuentes hasta la obtención de la muestra final utilizada en la modelización. Para ello se presenta, en primer lugar, el criterio general de diseño de la base; a continuación se explica la construcción de la variable objetivo y de los distintos bloques de variables explicativas; después se detalla el proceso de descarga, integración y depuración de la información; y, por último, se discuten las principales limitaciones encontradas.

El criterio general que guía todo el proceso es operativo. La base de datos no

se concibe como una mera recopilación retrospectiva de información, sino como una aproximación a un entorno real de decisión. En consecuencia, se priorizan, en la medida de lo posible, variables disponibles con antelación suficiente para tomar decisiones antes del cierre del mercado diario, evitando utilizar como entrada el propio *spread* que se pretende predecir.

2.2. Delimitación del estudio y criterios generales de construcción

La primera decisión metodológica relevante fue definir el marco temporal y geográfico del estudio. En cuanto al horizonte temporal, se intentó recopilar el máximo histórico homogéneo posible dentro del periodo 2014–2025. Esta elección responde a una razón práctica: varias de las series empleadas, especialmente dentro del bloque de datos del sistema eléctrico, comienzan o se estabilizan en torno a 2014, lo que convierte ese año en un punto de partida razonable para la construcción de una base coherente. Aun así, la muestra finalmente utilizable no coincide exactamente con el periodo bruto de descarga, sino con la intersección temporal de todas las variables seleccionadas. Durante la recopilación se comprobó, por ejemplo, que algunas series comenzaban más tarde, que ciertas previsiones no cubrían todo el intervalo y que algunas sesiones intradiarias presentaban una cobertura incompleta en los últimos años. Por este motivo, conviene distinguir entre *base completa de obtención* y *base analítica final*.

Desde el punto de vista geográfico, el problema económico se enmarca en el Mercado Ibérico de Electricidad (MIBEL), que integra España y Portugal y articula el mercado diario y los mercados intradiarios a través de OMIE, mientras que OMIP gestiona los mercados a plazo. Sin embargo, durante la descarga y revisión de las series se observó que los indicadores finalmente utilizables no respondían todos al mismo espacio: algunos se referían a España, otros a ámbito peninsular y otros, en principio, al entorno ibérico. Para evitar incoherencias, se adoptó una decisión de homogeneización fundamental: construir toda la base analítica en **ámbito peninsular**. Esta elección afecta al tratamiento de las series del sistema eléctrico, a la potencia instalada, a la agregación meteorológica y también a las ponderaciones de población utilizadas para medir el impacto territorial de los festivos.

Además del ámbito temporal y geográfico, se fijaron desde el inicio tres criterios generales de construcción de la base. En primer lugar, se priorizó siempre que fue posible la resolución **horaria**, por ser la adecuada para estudiar el perfil

intradía del *spread*. En segundo lugar, se impuso un criterio de **coherencia entre fuentes**, evitando combinar una variable en ámbito nacional con otra en ámbito peninsular si ambas se iban a utilizar simultáneamente en el modelo. En tercer lugar, se diferenció entre variables susceptibles de ser tratadas mediante imputación razonable y variables para las que la ausencia de dato comprometía el uso de toda la observación, de modo que la política de tratamiento de valores faltantes no fuera indiscriminada.

2.3. Variable objetivo: construcción del *spread* entre mercados

La variable objetivo del trabajo es el *spread* horario entre el mercado diario y el mercado intradiario. La lógica económica de esta variable es directa: si el precio intradiario supera al precio del mercado diario en una determinada hora, existe una señal potencialmente favorable para comprar en el diario y vender en el intradiario; si ocurre lo contrario, la señal sería inversa.

Formalmente, el *spread* horario del día d y la hora h puede definirse como

$$Spread_{d,h} = P^{ID}_{d,h} - P^{MD}_{d,h}, \tag{2.1}$$

donde $P^{ID}_{d,h}$ representa el precio del mercado intradiario y $P^{MD}_{d,h}$ el precio del mercado diario en la misma hora. Esta formulación permite centrar el análisis en el comportamiento relativo entre ambos mercados y no en la predicción aislada de uno de ellos.

La construcción de esta serie exigió, en primer lugar, descargar los precios horarios del mercado diario y del mercado intradiario a partir de las plataformas oficiales del sistema eléctrico. En la práctica, la construcción de la serie objetivo no estuvo exenta de incidencias. Durante la fase de obtención se comprobó que la descarga directa desde la interfaz web de ESIOS era poco estable para periodos largos, lo que obligó a dividir varias consultas en bloques temporales más pequeños. También se detectó que ciertas exportaciones en formato CSV devolvían valores incorrectamente interpretados o inconsistentes, por lo que se optó por descargar determinados indicadores en formato de hoja de cálculo y posteriormente integrarlos. A ello se añadió una limitación concreta de cobertura: una de las sesiones intradiarias solo estaba disponible hasta 2024, lo que obligó a revisar si esa serie podía mantenerse o debía excluirse del conjunto final.

Por tanto, aunque la definición teórica de la variable objetivo es sencilla, su construcción efectiva exigió una revisión detallada de la continuidad temporal de las series y de la coherencia del emparejamiento horario entre mercados. Solo una vez asegurada esa alineación se calculó el *spread* como diferencia de precios para cada observación.

2.4. Variables explicativas del sistema eléctrico y de mercado

El núcleo principal de las variables explicativas procede del sistema eléctrico y del propio mercado energético. Las variables seleccionadas fueron: precios del mercado diario e intradiario, previsiones de demanda, previsiones de generación renovable, disponibilidad de potencia nuclear, potencia instalada por tecnología y una variable exógena asociada al coste del gas. Estas variables se eligieron porque, desde el punto de vista económico, son las que mejor representan las tensiones previsibles entre oferta y demanda y el entorno de costes marginales al que está expuesto el mercado.

La fuente principal de este bloque fue **ESIOS**, el portal de datos de Red Eléctrica. Dentro de este bloque, las variables más directamente relacionadas con el equilibrio del sistema fueron la **previsión de demanda** y las **previsiones de generación renovable**. Su inclusión responde a que la posición esperada del sistema antes del día de entrega constituye una de las principales fuentes de variación del *spread*. Si para un determinado día se anticipa demanda elevada y baja generación renovable, es razonable esperar un entorno de precios más tensionado que si ocurre lo contrario. El mismo argumento vale para la generación eólica y solar, cuyo peso creciente en la península ibérica ha incrementado la sensibilidad del precio a las condiciones meteorológicas y a los desvíos respecto a la programación.

Junto con esas previsiones se incorporó la **disponibilidad de potencia nuclear**, que actúa como aproximación a la capacidad firme disponible en el sistema. Su incorporación es especialmente relevante en un mercado como el peninsular, donde la indisponibilidad de generación nuclear puede alterar de forma apreciable el *mix* esperado y el nivel de precios.

También se incluyó la **potencia instalada por tecnología**, no tanto como variable explicativa autónoma de corto plazo como para construir magnitudes relativas más informativas. No es equivalente prever un determinado volumen de producción renovable cuando la potencia instalada total es reducida que cuando

es muy elevada. Por ello, parte de la información renovable se interpretó de forma relativa respecto a la capacidad instalada disponible. A raíz de esta lógica, una de las correcciones metodológicas adoptadas fue ajustar todas las variables derivadas a **potencia instalada peninsular**, evitando mezclarla con referencias nacionales.

Por último, este bloque incorporó el **precio del gas TTF neerlandés** como variable exógena de coste. El papel de esta variable es capturar, aunque sea de forma indirecta, el entorno de costes de las tecnologías térmicas que pueden fijar precio en el mercado eléctrico. En la práctica, durante la recopilación se observó que la serie histórica descargada para el TTF comenzaba el 23 de octubre de 2017, lo que introdujo una restricción adicional en la muestra final si se deseaba mantener esta variable en el modelo.

Desde el punto de vista operativo, la descarga de este bloque de variables fue la fase más laboriosa de toda la construcción de la base. Se detectaron varios problemas concretos: inestabilidad de la web de ESIOS para periodos largos, necesidad de partir una misma serie en varios ficheros, discrepancias en ciertas exportaciones CSV y revisión posterior de la coherencia geográfica de algunas descargas. Todo ello obligó a un proceso repetido de descarga, comprobación, concatenación y validación. En este sentido, el trabajo no consistió únicamente en obtener información publicada, sino en reconstruir series homogéneas a partir de fuentes con estructuras distintas y con cierta complejidad operativa.

2.5. Variables meteorológicas

La inclusión de variables meteorológicas responde a una doble justificación. Por una parte, las condiciones atmosféricas afectan directamente a la demanda eléctrica, especialmente a través de la temperatura. Por otra, condicionan la disponibilidad efectiva de generación renovable, sobre todo eólica y solar, y por tanto alteran de forma indirecta la formación de precios.

La fuente seleccionada para este bloque fue **AEMET OpenData**, el sistema de reutilización de información meteorológica y climatológica de la Agencia Estatal de Meteorología. En la práctica, el bloque meteorológico fue uno de los más complejos de estructurar. La dificultad inicial de navegación del portal y de comprensión de la lógica de la API quedó resuelta al adoptar un procedimiento automatizado y reproducible de descarga por estación, en lugar de una descarga manual.

Para la construcción de la variable meteorológica agregada se adoptó el criterio de utilizar como ciudades de referencia **Madrid, Barcelona, Valencia, Sevilla,**

Bilbao, Zaragoza, Valladolid y Málaga, seleccionando una única estación por ciudad con el mayor grado posible de completitud y continuidad temporal. Esta decisión evita cambios de estación dentro de una misma ciudad y reduce el riesgo de introducir discontinuidades artificiales en las series como consecuencia de cambios de emplazamiento o de características de la estación. Sobre esa base se construyó una agregación peninsular ponderada por población, coherente con el criterio geográfico general del trabajo.

Las variables finalmente consideradas dentro del bloque meteorológico fueron `tmed`, `tmax`, `tmin`, `prec`, `velmedia`, `racha` y `hrMedia`. Estas variables permiten capturar tres dimensiones distintas del entorno atmosférico: temperatura, precipitación y viento, añadiendo además una medida de humedad media. En principio, las variables térmicas son especialmente relevantes por su relación con la demanda; las de viento, por su potencial correlación con la producción eólica esperada; y la precipitación y la humedad, como indicadores complementarios del tipo de jornada meteorológica y del comportamiento de la población.

La agregación se realizó a escala diaria y peninsular, combinando las observaciones de las ocho ciudades de referencia mediante ponderaciones poblacionales. Este planteamiento tiene dos ventajas. La primera es que evita conceder el mismo peso a ciudades con tamaños poblacionales muy distintos. La segunda es que construye una variable sintética más próxima al comportamiento agregado del sistema eléctrico peninsular que la utilización aislada de una única estación. Ahora bien, esta solución también presenta una limitación importante: no reproduce la heterogeneidad meteorológica real de toda la península, sino una aproximación ponderada basada en un conjunto limitado de localizaciones. Queda como línea de trabajo futura comprobar si un modelo que incorporase las ocho ciudades de forma independiente, en lugar de su agregado ponderado, mejora la capacidad predictiva del sistema.

Una cuestión metodológica importante es que la base finalmente construida a partir de AEMET se apoyó en datos climatológicos diarios observados, y no en previsiones meteorológicas históricas reconstruidas *ex post.*Esto no responde a una decisión metodológica, sino a la disponibilidad real de los datos: AEMET OpenData no ofrece de forma gratuita un histórico de previsiones meteorológicas pasadas, únicamente observaciones climatológicas. Los datos observados son plenamente válidos para describir el contexto físico de cada día y para estudiar la relación entre condiciones meteorológicas y *spread*, pero no representan exactamente la información *ex ante* disponible para los agentes en el momento de tomar decisiones. Por tanto, este bloque se utiliza como aproximación informativa y no como previsión histórica perfecta.

En cuanto a los valores faltantes, se adoptó una política diferenciada respecto al resto de variables. En las variables meteorológicas agregadas se optó por imputar los valores ausentes mediante la **mediana** de cada variable y ciudad, calculada sobre la totalidad de la serie histórica disponible (es decir, sin distinguir estacionalidad: una observación ausente en diciembre se imputa con el mismo valor que una ausente en julio). Esta decisión es defendible porque se trata de una imputación robusta frente a valores extremos y porque, en este bloque, la pérdida de observaciones completas por pequeñas lagunas de una estación concreta podía resultar demasiado costosa. Aun así, conviene dejar constancia de que se trata de una solución conservadora y que su objetivo fue preservar la continuidad del agregado, no reconstruir con precisión perfecta el estado meteorológico de cada ciudad en todos los días.

2.6. Variables de calendario

Las variables de calendario se incorporaron para capturar patrones sistemáticos asociados al comportamiento social y económico de los días, especialmente aquellos vinculados a la diferencia entre días laborables, fines de semana y festivos.

Este bloque se estructuró en torno a dos variables. La primera fue el tipo de día, para cuya construcción se definieron cuatro categorías: Tipo 1 (lunes), Tipo 2 (martes a jueves), Tipo 3 (sábado) y Tipo 4 (viernes y domingo). Los festivos que caen en un día laborable entre semana conservan el tipo correspondiente a su día de la semana; el efecto festivo propiamente dicho se recoge, de forma independiente y más precisa, mediante la variable continua de porcentaje de población afectada que se describe a continuación. Esta clasificación permite captar de forma compacta las regularidades semanales del sistema eléctrico, sin recurrir a una desagregación completa por días que habría incrementado la complejidad de la variable sin aportar necesariamente una mejora proporcional en capacidad explicativa.

El segundo elemento, y probablemente el más original desde el punto de vista metodológico, fue la construcción de una variable continua que mide el **porcentaje de población afectada por festivo** en cada fecha. La motivación de esta variable es clara: no todos los festivos tienen el mismo impacto sobre la actividad económica ni sobre la demanda eléctrica del sistema peninsular. Un festivo nacional afecta a todo el mercado de referencia, mientras que un festivo autonómico o local solo afecta a una parte del territorio.

Desde el punto de vista normativo, el marco general de los festivos laborales en España distingue entre fiestas nacionales, autonómicas y locales. La principal dificultad práctica aparece al incorporar los **festivos locales**, ya que estos no se concentran en una única fuente estatal y pueden requerir consultas a boletines autonómicos, provinciales o municipales. Además, durante la construcción de la variable se detectó que inicialmente no se estaban teniendo en cuenta correctamente algunos **festivos trasladados**, problema que más tarde fue corregido. Este detalle es importante, ya que la calidad de la variable de calendario no dependía solo de identificar qué días eran festivos, sino de capturar correctamente los desplazamientos cuando una festividad se trasladaba y de evitar contabilizar como no laborable un día que, desde el punto de vista efectivo, sí lo era.

La variable de porcentaje de población festiva se construyó como el cociente entre la población afectada por festivo en cada fecha y la población total peninsular. Esta decisión responde a la necesidad de mantener la coherencia con el resto de la base de datos, que se ha definido en ámbito peninsular. Utilizar un denominador distinto habría generado una inconsistencia en la medición de la intensidad del efecto festivo, al compararlo con variables referidas a un espacio geográfico diferente.

Para la población se emplearon las **cifras oficiales del padrón municipal**, referidas a 1 de enero de cada año y disponibles a nivel provincial y municipal. Se optó por utilizar el padrón anual como referencia homogénea en todo el periodo, evitando combinar fuentes con distinta frecuencia temporal. Dado que la variable festiva no requiere una elevada precisión demográfica en el corto plazo, este criterio permite simplificar el tratamiento de los datos sin pérdida relevante de información.

Conviene señalar una limitación importante: no se incorporaron **festivos extraordinarios no previstos**. Esta decisión es coherente con el enfoque del trabajo, centrado en información estructural y previsible, y evita introducir arbitrariedad en la reconstrucción del calendario. En consecuencia, la variable de calendario proporciona una aproximación razonable a la intensidad territorial del efecto festivo, aunque no recoge todas las posibles excepciones de carácter institucional.

La Tabla 2.1 resume los bloques de información descritos en las secciones anteriores, junto con su fuente, ámbito geográfico, resolución temporal, cobertura y tratamiento de valores ausentes.

Tabla 2.1: Resumen de las variables de la base de datos: fuente, ámbito, resolución, cobertura de la descarga completa y tratamiento de valores ausentes. La muestra analítica final es la intersección temporal de todas las series (véase la sección 2.7).

Variable	Fuente	Ámbito	Resolución	Cobertura	Ausencias
Precio mercado diario	ESIOS (OMIE)	España	Horaria	2014–2025	Exclusión del día
Precio intradiario (sesión 1)	ESIOS (OMIE)	España	Horaria	2014–2025	Exclusión del día
Previsión de demanda D+1	ESIOS (REE)	Peninsular	Horaria	Desde 2019	Exclusión del día
Previsión eólica	ESIOS (REE)	Peninsular	Horaria	2014–2025	Exclusión del día
Previsión fotovoltaica	ESIOS (REE)	Peninsular	Horaria	2014–2025	Exclusión del día
Previsión solar térmica	ESIOS (REE)	Peninsular	Horaria	2014–2025	Exclusión del día
Disponibilidad nuclear	ESIOS (REE)	Peninsular	Horaria	2014–2025	Exclusión del día
Potencia instalada por tecnología	ESIOS (REE)	Peninsular	Mensual (expandida a diaria)	2015–2025	Exclusión del día
Precio del gas (futuro TTF)	Investing.com	Europa (TTF)	Diaria (días hábiles)	Desde oct. 2017	Exclusión del día
Meteorología (7 variables)	AEMET OpenData	8 ciudades → agregado peninsular	Diaria	2014–2025	Imputación por mediana
Festivos y % población afectada	Boletines oficiales + padrón INE	Provincial → peninsular	Diaria	2014–2025	No aplica (construida)
Tipo de día	Elaboración propia	—	Diaria	2014–2025	No aplica (construida)

2.7. Depuración final de la base de datos

Una vez descargadas las distintas fuentes, fue necesario desarrollar un proceso de integración y depuración para convertir un conjunto heterogéneo de ficheros brutos en una base analítica única. Esta fase fue tan relevante como la propia descarga, porque muchas de las dificultades del trabajo no procedieron de la ausencia total de información, sino de la necesidad de reconciliar series obtenidas en formatos, ámbitos y frecuencias diferentes.

La primera tarea consistió en la **unificación de formatos y nomenclaturas**. En la práctica, distintas fuentes devolvían fechas con estructuras diferentes, nombres de columnas no homogéneos y, en algunos casos, variables con etiquetas cambiantes según el archivo descargado. Por ello fue necesario renombrar sistemáticamente las columnas, estandarizar el formato de fecha-hora y asegurar que cada serie quedara identificada de forma inequívoca dentro de la tabla final. En el caso de AEMET, por ejemplo, se renombraron las variables por ciudad antes de agregarlas; en ESIOS, se concatenaron varios archivos de un mismo indicador descargados por tramos temporales.

La segunda tarea fue la **alineación temporal**. Aunque el modelo final trabaja a nivel horario, no todas las variables nacen con esa resolución. Algunas series son horarias por construcción, mientras que otras, como las meteorológicas agregadas o determinadas variables de calendario, son diarias y deben expandirse o vincularse a todas las horas del día correspondiente. Este paso es especialmente delicado porque cualquier desajuste horario puede generar emparejamientos erróneos entre la variable objetivo y los predictores. De ahí que se revisara cuidadosamente la continuidad temporal y se comprobara que no hubiese duplicados ni saltos anómalos.

La tercera tarea fue la **detección de incoherencias y duplicados**. La descarga por bloques, aunque necesaria, aumenta el riesgo de solapamientos entre archivos o de pérdida de observaciones en los límites de cada tramo. En consecuencia, tras unir los ficheros fue necesario comprobar la existencia de horas repetidas, días ausentes o cambios inesperados en la estructura de la serie. Una dificultad similar surgió en la construcción del calendario festivo, ya que en una misma fecha podían coincidir festivos en distintos territorios. En estos casos, fue necesario asegurar que la población de cada territorio se contabilizase una única vez, evitando duplicidades derivadas de posibles errores en la agregación.

En relación con los **valores faltantes**, se adoptó un tratamiento diferenciado por bloques. Como ya se ha explicado, en las variables meteorológicas agregadas

se optó por imputar los valores ausentes mediante la mediana histórica de cada ciudad y variable, sin ajuste estacional. En cambio, para el resto de variables del sistema eléctrico y del mercado se siguió un criterio más estricto: cuando una observación contenía ausencias relevantes en variables clave, el día correspondiente no se utilizaba en la base analítica final. Esta estrategia reduce el riesgo de introducir sesgos artificiales en variables especialmente sensibles, como el precio o las previsiones del sistema.

El resultado de este proceso fue una **muestra final de trabajo** más pequeña que la base completa inicialmente descargada, pero mucho más coherente desde el punto de vista analítico: 1.763 jornadas completas, todas ellas laborables, comprendidas entre enero de 2019 y diciembre de 2025. El inicio efectivo lo fija la disponibilidad de la previsión de demanda, y la ausencia de fines de semana se debe a que la cotización del futuro del gas solo existe en días hábiles; ambas restricciones se discuten con más detalle en el capítulo 3. Esta reducción es una consecuencia habitual en proyectos aplicados con múltiples fuentes heterogéneas: la prioridad no debe ser maximizar el número bruto de observaciones, sino garantizar que cada fila de la base represente una situación completa y comparativamente homogénea.

Una consecuencia directa de restringir la muestra a días laborables es que la categoría Tipo 3 (sábado) queda vacía en la base analítica final y no se observa en ningún análisis posterior; se conserva en la definición general del apartado 4.3.5 por coherencia conceptual, pero su columna resultante de la codificación *one-hot* es constante y no aporta información.

2.8. Valoración final y limitaciones de la base construida

La base de datos finalmente construida permite representar, con un grado razonable de detalle, la interacción entre el *spread* de mercado y sus principales determinantes estructurales. Reúne información de precios, previsiones del sistema, disponibilidad de generación firme, costes energéticos, meteorología y calendario, todo ello armonizado en una lógica peninsular y con resolución horaria para la parte central del análisis. Este planteamiento es coherente con el objetivo del trabajo: anticipar el comportamiento del *spread* a partir de información disponible con antelación y evaluar posteriormente la utilidad económica de esa predicción.

No obstante, la base presenta varias limitaciones que deben ser reconocidas

de forma explícita. La primera es la **desigual disponibilidad temporal** de las series. Aunque se tomó como referencia general el periodo 2014-2025, no todas las variables cubren homogéneamente ese intervalo. La segunda es la **heterogeneidad original de los ámbitos geográficos**, que obligó a un esfuerzo adicional de depuración para mantener una lógica peninsular coherente.La tercera es la distinta naturaleza de las variables: mientras que algunas reflejan información potencialmente disponible *ex ante*, otras, en particular las variables meteorológicas observadas, se incorporan como indicadores de las condiciones reales del día y no como previsiones meteorológicas. La cuarta es la política de imputación empleada en el bloque meteorológico: los valores ausentes se sustituyeron por la mediana histórica de cada ciudad y variable, calculada sobre la totalidad del periodo sin distinguir estacionalidad, lo que puede suavizar en exceso una ausencia puntual ocurrida en un mes atípico.

A ello se suman limitaciones prácticas derivadas de las propias fuentes. En ESIOS, la descarga masiva a través de la web resultó poco estable y obligó a trocear series y validar formatos. En AEMET, la complejidad inicial del acceso dificultó la descarga manual y justificó el paso a una estrategia automatizada. En el bloque de calendario, la correcta identificación de festivos trasladados exigió una revisión específica, y no se incorporaron festivos extraordinarios no previstos. Todo ello demuestra que la fase de obtención de datos en un trabajo como este no puede considerarse una etapa meramente instrumental, sino una parte importante del mismo.

En conjunto, puede afirmarse que la base final no es una reproducción perfecta de toda la información potencialmente relevante para el mercado, pero sí una representación sólida, documentada y coherente de los principales factores que influyen en el *spread* horario. Precisamente por ello constituye una base adecuada para la fase posterior de modelización, en la que se analizarán primero patrones de agrupación de días y, posteriormente, modelos de predicción directa del *spread*.

Capítulo 3

Clusterización de días

3.1. Introducción

Una vez construida la base analítica, el siguiente paso consiste en estudiar si los días pueden agruparse en conjuntos relativamente homogéneos a partir de la información explicativa disponible antes de la operativa de mercado. La idea de fondo es que el *spread* entre el mercado diario y el intradiario no tiene por qué comportarse de forma completamente aislada o arbitraria, sino que puede depender de configuraciones recurrentes del sistema: niveles de demanda esperada, previsiones renovables, condiciones meteorológicas, disponibilidad de generación firme y tipo de día. Si esas configuraciones se repiten, resulta razonable explorar si también se repiten determinados perfiles de *spread*. El objetivo de esta fase no es todavía predecir directamente el *spread* horario, sino ordenar los días en grupos comparables y comprobar después si esos grupos presentan características diferenciadas.

Este planteamiento encaja además con la lógica general del trabajo, orientada a utilizar información disponible con antelación y a mantener una perspectiva operativa, no meramente retrospectiva. Desde el punto de vista metodológico, esta etapa pertenece al ámbito del aprendizaje no supervisado. A diferencia de los modelos supervisados, en los que existe una variable objetivo conocida que guía el ajuste, aquí se trabaja únicamente con un conjunto de características observadas y se intenta identificar si dentro de ellas existe una estructura en forma de grupos. Por ello, esta fase debe entenderse como una herramienta previa para agrupar tipos de día. Su utilidad está en describir mejor el sistema y en preparar el análisis posterior del *spread*, pero no constituye todavía una predicción en sentido estricto.

3.2. Criterios específicos de la clusterización

Una vez definida la base analítica, en esta fase el problema ya no era tanto construir nuevas variables como decidir de qué forma representar cada día para que la agrupación tuviera sentido. En un problema de *clustering*, esta cuestión es especialmente importante, porque el resultado depende de manera directa de la matriz de entrada y del peso efectivo que termina teniendo cada bloque de información.

El primer criterio fue mantener el día completo como unidad de análisis. En esta parte del trabajo no interesaba agrupar horas aisladas, sino jornadas con un contexto parecido. Por ello, la representación de cada observación debía conservar, en la medida de lo posible, la información relevante sobre el perfil esperado del sistema a lo largo del día, y no reducirse únicamente a valores medios. Esta decisión resultaba especialmente importante en las variables horarias, ya que dos días pueden tener medias similares y, sin embargo, responder a perfiles intradiarios bastante distintos.

El segundo criterio fue intentar que los grupos resultantes siguieran siendo interpretables desde un punto de vista económico. Una opción posible habría sido resumir mucho la información desde el principio, reduciendo de forma fuerte la dimensión de la matriz. Sin embargo, eso habría hecho más difícil explicar después qué caracteriza a cada clúster. Por este motivo, se optó por mantener una representación relativamente cercana a las variables originales, aun sabiendo que eso obligaba a tratar con más cuidado los problemas de escala y de ponderación.

El tercer criterio, que en la práctica fue uno de los más importantes, consistió en evitar que unas familias de variables dominasen la agrupación simplemente por estar representadas en un mayor número de columnas. En esta aplicación conviven bloques horarios amplios, con 24 columnas por variable, y otras variables diarias resumidas en una única columna. Si se aplicase el algoritmo directamente sobre esa matriz, los bloques horarios tenderían a pesar mucho más en las distancias finales, aunque eso no implicase necesariamente que debieran tener más importancia desde el punto de vista económico. Por tanto, la construcción de la matriz de agrupación no podía consistir únicamente en unir columnas: exigía también una estrategia de normalización y ponderación que evitase ese sesgo.

3.3. Variables empleadas en la clusterización

La información utilizada para agrupar los días procede de los mismos bloques que estructuran la base analítica del trabajo. El núcleo principal lo forman las variables del sistema eléctrico y del mercado: previsión de demanda, previsiones de generación renovable, disponibilidad de potencia nuclear, potencia instalada por tecnología y una variable asociada al coste del gas. Estas magnitudes son adecuadas para la clusterización porque describen de forma bastante directa la situación esperada de equilibrio entre oferta y demanda y el entorno de costes al que está expuesto el sistema. Una jornada con demanda elevada, baja previsión renovable y menor disponibilidad de generación firme no representa el mismo contexto de mercado que otra con exceso de producción eólica o solar y menor tensión esperada del sistema.

Dentro de este bloque, las previsiones renovables tienen un interés especial. En un sistema eléctrico con un peso creciente de la eólica y la solar, el comportamiento del precio es cada vez más sensible tanto a la meteorología como a las desviaciones entre lo previsto y lo finalmente producido. Por eso, no basta con considerar la generación renovable en términos absolutos; también resulta útil relacionarla con la potencia instalada disponible, ya que una misma previsión de producción puede tener un significado distinto según la capacidad total del sistema.

Junto a estas variables se incorporan las meteorológicas. Su inclusión responde a una doble lógica. Por un lado, las condiciones atmosféricas afectan directamente a la demanda eléctrica, sobre todo a través de la temperatura (y de la sensación térmica). Por otro, guardan una relación clara con la producción renovable esperada, especialmente en el caso del viento. En la base final, estas variables se agregaron a escala diaria y peninsular a partir de ocho ciudades de referencia, ponderando por población para aproximar mejor el comportamiento agregado del sistema. Esta decisión tiene ventajas claras en términos de coherencia territorial, aunque también simplifica una realidad meteorológica bastante más heterogénea.

El tercer bloque relevante es el calendario. Las variables de calendario permiten recoger regularidades del comportamiento social y económico que afectan a la demanda y, de forma indirecta, al entorno del *spread*. En este trabajo se construyeron dos variables especialmente útiles para la clusterización. La primera es el tipo de día, definido en cuatro categorías: lunes; martes a jueves; sábado; y viernes y domingo. Como se explicó en el capítulo 2, los festivos que caen en día laborable conservan el tipo correspondiente a su día de la semana; su efecto se recoge mediante la variable continua de porcentaje de población festiva descrita a continuación. La segunda es una variable continua que mide el porcentaje de po-

blación afectada por festivo en cada fecha. Esta segunda magnitud es útil porque no todos los festivos tienen el mismo impacto sobre la actividad económica: un festivo nacional afecta a todo el sistema, mientras que uno local o autonómico solo altera una parte del territorio.

La clusterización no se apoya únicamente en niveles absolutos. Para describir bien una jornada, también puede ser relevante saber cómo se sitúa respecto a momentos cercanos. Por ello, además de los valores observados o previstos, la representación de cada día puede enriquecerse con variaciones respecto al día anterior y respecto a la semana precedente. Esta ampliación tiene sentido, porque dos días con la misma previsión de demanda o de generación eólica pueden pertenecer a contextos distintos si en un caso esa previsión supone continuidad y en otro una ruptura más marcada. Introducir esta dimensión comparativa permite que el algoritmo no agrupe solo por estados del sistema, sino también por cambios recientes en ese estado.

3.4. Construcción de la matriz de agrupación

La matriz de agrupación se construyó con una fila por día y un conjunto de columnas que recogían la información considerada relevante para describir el contexto de cada jornada. En el caso de las variables con perfil horario, se decidió mantener ese detalle intradía. Es decir, cada día quedó representado mediante las 24 observaciones correspondientes a esas magnitudes. Esta elección permitía distinguir entre jornadas que podían parecer similares en términos medios, pero que presentaban distribuciones horarias claramente diferentes.

Junto a estos bloques horarios se incorporaron variables diarias o cuasi diarias, como las meteorológicas agregadas, el precio del gas, el porcentaje de población afectada por festivo y el tipo de día. En estos casos no tenía sentido generar una estructura de 24 columnas, porque la información que aportaban era esencialmente diaria. Como consecuencia, la matriz final combinó bloques de tamaño muy distinto: algunos amplios, asociados a perfiles horarios, y otros mucho más reducidos.

Esta forma de construir la matriz tenía una ventaja clara, que era conservar información sobre la forma del día. Sin embargo, también introducía un problema importante. Al usar distancias como base de la clusterización, un bloque de 24 columnas puede acabar teniendo mucho más peso que una variable representada por una sola columna. En otras palabras, una familia de variables puede condicionar la agrupación no porque sea necesariamente más relevante, sino simplemente porque

ocupa más espacio dentro de la matriz. Este fue uno de los problemas más claros al trabajar con el código y obligó a introducir un tratamiento específico.

Por ello, la preparación de la matriz se realizó en dos pasos. En primer lugar, se estandarizó cada columna continua de forma individual. Este paso era imprescindible, ya que en la matriz convivían magnitudes medidas en unidades muy distintas, como MW, porcentajes o temperaturas. Sin estandarización, las variables con mayor rango habrían dominado las distancias de manera puramente mecánica.

No obstante, la estandarización columna a columna no resolvía por sí sola el problema principal. Aunque todas las columnas quedasen en una escala comparable, los bloques con más variables seguían acumulando más peso total en la distancia entre dos días. Por esta razón, tras la estandarización se aplicó una ponderación por bloques. La idea era que la contribución total de cada familia de variables no dependiese únicamente del número de columnas con el que estaba representada. En la práctica, esto implicó reescalar cada bloque en función de su dimensión, de modo que un conjunto horario muy amplio no eclipsara automáticamente a variables diarias más sintéticas pero también relevantes.

Desde un punto de vista conceptual, esta ponderación equivale a repartir el peso entre familias de información, y no solo entre columnas individuales. De este modo, bloques como la previsión de demanda, la previsión renovable, la información meteorológica o el calendario podían contribuir de una forma más equilibrada a la distancia final. Este punto era importante porque, sin él, el algoritmo tendería a agrupar sobre todo a partir de la forma de unas pocas curvas horarias, dejando en un segundo plano variables como la festividad, el tipo de día o el coste del gas.

En el caso de las variables categóricas, tampoco era adecuado introducirlas directamente como números enteros, ya que eso habría impuesto un orden artificial entre categorías. Por ello, variables como el tipo de día se transformaron en indicadores binarios. Con ello se evitaba, por ejemplo, que la distancia entre lunes y sábado dependiese de una codificación arbitraria en lugar de reflejar simplemente que se trata de categorías distintas.

Por último, antes de ejecutar los algoritmos se comprobó que la matriz no contuviese valores faltantes residuales y que todas las columnas utilizadas en la clusterización estuvieran definidas para el mismo conjunto de días válidos. En esta fase no era necesario repetir toda la depuración ya descrita en el capítulo anterior, pero sí asegurar que la tabla final de entrada al algoritmo fuese completa, comparable y estable desde un punto de vista numérico.

3.5. Métodos de clusterización considerados

Una vez construida la matriz de agrupación, ya estandarizada y ponderada, se aplicaron dos métodos de clusterización: K-means y clusterización jerárquica aglomerativa. La elección de estos dos enfoques responde a que ambos son métodos clásicos y muy utilizados en aprendizaje no supervisado, pero parten de lógicas distintas y, por tanto, permiten observar la estructura de los datos desde perspectivas complementarias. En ambos casos, el objetivo no era predecir directamente el spread, sino agrupar días a partir de las variables explicativas disponibles antes del cierre del mercado diario y comprobar después si los días de un mismo grupo compartían perfiles de spread similares.

Dicho de forma sencilla, estos métodos permiten responder a la siguiente pregunta: si dos días presentan condiciones parecidas en términos de previsión de demanda, generación renovable, meteorología, tipo de día o variables similares, ¿tienden también a mostrar un comportamiento parecido del spread entre mercados? Si la respuesta es afirmativa, la agrupación deja de ser solo una herramienta descriptiva y pasa a tener utilidad operativa, porque permite asociar un día futuro a un grupo ya conocido y utilizar el perfil medio de spread de ese grupo como referencia.

3.5.1. K-means

K-means es un método de partición que divide las observaciones en un número prefijado (K) de clústeres no solapados [16].La idea básica es formar grupos de modo que los días incluidos en un mismo clúster sean lo más parecidos posible entre sí. En términos más precisos, K-means minimiza exclusivamente la variabilidad interna de los clústeres, es decir, la dispersión de las observaciones respecto al centro de su grupo; el algoritmo no incorpora un término que premie explícitamente la separación entre grupos. No obstante, dado que la varianza total de la muestra se descompone en la suma de la varianza interna y la varianza entre grupos, minimizar la primera para una varianza total fija equivale, de forma indirecta, a maximizar la segunda, por lo que los días de un mismo clúster tienden también a quedar diferenciados de los de otros grupos.

El algoritmo funciona de manera iterativa. En primer lugar, fija una asignación inicial de las observaciones a los K grupos. A continuación, calcula el centroide de cada clúster, entendido como el vector de medias de las variables de los días que pertenecen a ese grupo. Después, cada día se reasigna al centroide más próximo y

el proceso se repite hasta que la asignación deja de cambiar. Por tanto, el resultado final es una partición cerrada de los días, donde cada observación pertenece a un único clúster representado por su centroide.

En este trabajo, K-means resulta útil porque produce una clasificación clara y fácil de interpretar. Una vez obtenido un conjunto de grupos, puede calcularse para cada uno de ellos el perfil medio del spread horario y comparar esos perfiles entre sí. Esto facilita pasar de una matriz con muchas variables explicativas a un número reducido de tipologías de día, lo que hace más sencilla la interpretación posterior.

No obstante, K-means también presenta limitaciones. La más importante es que obliga a fijar de antemano el número de clústeres K, cuando precisamente una de las dificultades del problema es que no se conoce cuántos tipos de día existen de forma natural en los datos. Además, el algoritmo puede converger a óptimos locales, de modo que la solución obtenida depende en parte de la asignación inicial de los grupos. Por este motivo, el método se ejecutó con múltiples inicializaciones, para reducir la probabilidad de quedarse con una partición poco representativa por efecto del azar.

3.5.2. Clusterización jerárquica aglomerativa

La clusterización jerárquica aglomerativa sigue una lógica diferente. En lugar de partir de un número de grupos fijado de antemano, comienza considerando cada día como un clúster independiente. A partir de ahí, va fusionando de forma sucesiva los dos clústeres más parecidos hasta que todas las observaciones quedan reunidas en un único grupo. El resultado de este proceso se representa mediante un dendrograma, que muestra en qué orden se producen las fusiones y a qué nivel de distancia.

La principal ventaja de este enfoque es que no obliga a decidir desde el inicio cuántos grupos se quieren conservar. El dendrograma permite explorar la estructura de similitud entre los días y observar si existen separaciones relativamente claras o si, por el contrario, las transiciones entre grupos son graduales. En otras palabras, mientras que K-means entrega directamente una partición final, la clusterización jerárquica ofrece primero una visión global de la estructura de agrupación y deja para después la decisión de cortar el árbol en un determinado número de clústeres.

Esto tiene interés en un trabajo como este, donde la agrupación de días tiene un componente exploratorio importante. Antes de decidir cuántos tipos de día son

útiles, conviene observar si los datos sugieren pocos grupos amplios o una segmentación más fina. Además, la representación jerárquica permite detectar si ciertos días quedan claramente aislados o si algunos grupos son mucho más homogéneos que otros.

Como contrapartida, la clusterización jerárquica también tiene limitaciones. Una vez que dos clústeres se fusionan en el proceso aglomerativo, esa unión ya no se puede deshacer en pasos posteriores. Por tanto, decisiones tempranas pueden condicionar la estructura final del dendrograma. Además, la partición concreta depende no solo de la distancia entre observaciones, sino también del criterio elegido para medir la distancia entre grupos.

3.5.3. Distancia y criterio de enlace

En ambos métodos era necesario definir cómo medir la similitud entre días. Dado que la matriz de agrupación ya había sido previamente estandarizada y ponderada, la referencia natural fue la distancia euclídea. Esta elección es coherente con el funcionamiento de K-means, que minimiza la variación interna de los clústeres a partir de esa métrica, y también con la formulación más habitual de la clusterización jerárquica.

La estandarización previa era especialmente importante porque las variables originales no estaban expresadas en la misma escala. Sin ese paso, las variables con rangos numéricos más altos habrían dominado el cálculo de distancias y, por tanto, la agrupación. Además, la ponderación de bloques evitaba que conjuntos de variables con muchas columnas tuvieran un peso excesivo en la similitud total entre días. De esta forma, la distancia entre dos observaciones no refleja solo diferencias brutas en magnitud, sino una comparación más equilibrada entre los distintos tipos de información incorporados a la matriz.

En el caso de la clusterización jerárquica, además de la métrica entre observaciones es necesario elegir un criterio de enlace, es decir, una regla para definir la distancia entre dos clústeres ya formados. En este trabajo se evaluaron cuatro criterios: Ward [17], completo, medio y simple. El criterio de Ward fusiona en cada paso los dos grupos cuya unión produce el menor aumento de la varianza interna total; comparte, por tanto, la lógica de compacidad de K-means y tiende a generar clústeres relativamente equilibrados. El enlace completo define la distancia entre dos grupos a partir de la pareja de observaciones más alejada entre ambos, por lo que también favorece clústeres compactos. El enlace medio utiliza el promedio de las distancias entre observaciones de ambos grupos, ofreciendo una solución

intermedia y normalmente algo menos estricta.

El enlace simple, por último, define la distancia entre grupos a partir de la pareja de observaciones más próxima. Se incluyó en la comparación por completitud, aun sabiendo que tiende a producir el denominado efecto de encadenamiento: basta con que existan observaciones próximas de forma local para que se vayan uniendo grupos muy alargados y poco compactos. Como se verá en la sección 3.8, este comportamiento se confirmó en los resultados. En un problema como este, donde interesa interpretar después perfiles medios de spread asociados a cada grupo, resulta más útil trabajar con clústeres relativamente coherentes y distinguibles que con agrupaciones demasiado difusas. Las definiciones formales de la distancia y de cada criterio de enlace se recogen en el Anexo A.

3.5.4. Justificación conjunta de ambos enfoques

En definitiva, K-means y la clusterización jerárquica no se emplearon como métodos rivales en sentido estricto, sino como herramientas complementarias. K-means aporta particiones finales claras y fáciles de explotar en el análisis posterior de los perfiles de spread. La clusterización jerárquica, por su parte, ayuda a explorar la estructura interna de los datos y a valorar si el número de grupos elegido en K-means resulta razonable o no.

Utilizar ambos métodos permite, por tanto, contrastar resultados y reducir el riesgo de que las conclusiones dependan en exceso de una única técnica. Si distintos enfoques apuntan a segmentaciones parecidas o a perfiles de spread coherentes, la interpretación gana solidez. Si, por el contrario, producen agrupaciones muy diferentes, esa discrepancia también es informativa y sugiere que la estructura de los datos no es especialmente estable o que la separación entre tipos de día es débil.

3.6. Criterios metodológicos para la evaluación posterior

La comparación entre soluciones no podía basarse únicamente en un índice interno, porque en *clustering* no existe una variable respuesta conocida que permita validar directamente la calidad del resultado. Por ello, la evaluación posterior se planteó a partir de varios criterios complementarios.

El primero fue la cohesión interna y la separación entre grupos. Una solución

razonable debía generar clústeres suficientemente compactos por dentro y, al mismo tiempo, diferenciados entre sí. Este es el criterio más habitual en clusterización y sirve para descartar particiones claramente débiles desde un punto de vista geométrico.

El segundo criterio fue el tamaño de los clústeres. En la práctica, una solución puede ofrecer buenos valores en algunos índices y, sin embargo, generar grupos muy pequeños o residuales. Eso reduce bastante su utilidad. Por este motivo, se introdujo una restricción adicional: no se consideraron adecuadas las soluciones en las que algún clúster quedaba por debajo del 2 % de los días válidos. Esta condición no asegura por sí sola que una solución sea buena, pero sí evita particiones artificiales sostenidas por grupos demasiado marginales.

El tercer criterio fue la estabilidad. Si una solución cambia de forma importante al modificar ligeramente la inicialización, el criterio de enlace o el conjunto de variables, su interés práctico es menor. En consecuencia, no bastaba con obtener una partición con buenos indicadores en una ejecución concreta; también era necesario comprobar que la estructura general fuese relativamente robusta.

El cuarto criterio fue la interpretabilidad económica. Esta cuestión era especialmente importante en este trabajo. No interesaba únicamente obtener grupos estadísticamente compactos, sino clústeres que pudiesen describirse de forma razonable en términos de demanda, renovables, meteorología y calendario. Si una solución era difícil de explicar o no permitía caracterizar con claridad qué distingue a unos días de otros, su utilidad posterior para el análisis del *spread* se reducía de forma importante.

Por último, el criterio decisivo era la utilidad de la agrupación como paso previo al análisis del *spread*. Una partición podía ser geométricamente correcta y aun así resultar poco valiosa si después no ayudaba a diferenciar comportamientos del *spread* entre grupos. Por eso, la evaluación final no debía entenderse como un problema puramente matemático, sino como una combinación entre calidad estadística e interés aplicado.

3.7. Dificultades metodológicas y limitaciones de esta fase

La primera limitación de esta fase es que el *clustering* no ofrece una validación tan directa como un modelo supervisado. No existe una respuesta observada con la

que comparar automáticamente la calidad de la partición. En consecuencia, parte de la evaluación depende del criterio del analista y de la consistencia del resultado al cambiar ciertas decisiones de diseño.

La segunda limitación es la sensibilidad a la forma de construir la matriz de entrada. En este trabajo, el resultado podía verse afectado no solo por el algoritmo elegido, sino también por decisiones previas como mantener o no el detalle horario, estandarizar las columnas, ponderar los bloques o codificar las variables categóricas de una forma u otra. Esto no es un inconveniente menor, porque implica que los grupos obtenidos no deben interpretarse como una verdad única descubierta por el algoritmo, sino como una estructura plausible bajo una determinada representación de los días.

La tercera limitación es que, incluso después de la ponderación por bloques, siguen existiendo dependencias internas importantes entre columnas. Las horas consecutivas de una misma curva suelen estar muy correlacionadas, y algo parecido ocurre entre ciertas variables del mismo bloque. La ponderación reduce el problema de sobrerrepresentación por número de columnas, pero no elimina por completo la redundancia interna de la información.

La cuarta limitación es que tanto K-means como la clusterización jerárquica obligan a asignar todos los días a algún grupo. En la práctica, puede haber jornadas atípicas que no encajen especialmente bien en ninguna tipología clara. En esos casos, el algoritmo las incorpora igualmente al clúster más cercano, lo que puede distorsionar en cierta medida la forma final de los grupos.

Finalmente, también debe tenerse en cuenta que la agrupación obtenida depende de la muestra efectiva disponible. Por tanto, los clústeres identificados describen la estructura presente en ese conjunto de días y en esa representación concreta de las variables. Su interés es claro como herramienta exploratoria y como paso previo al análisis del *spread*, pero eso no significa que deban interpretarse como categorías cerradas o permanentes fuera del contexto de la muestra analizada.

3.8. Selección de la solución de referencia

Una vez ejecutadas las distintas variantes de K-means y de clusterización jerárquica sobre la matriz construida en la sección anterior, el siguiente paso consiste en elegir, sobre la base de los criterios metodológicos definidos en la sección 3.6, qué solución (o qué pequeño conjunto de soluciones) se va a llevar adelante para caracterizar el spread. Se han evaluado siete valores de K (de 2 a 8) para K-means

y para cuatro enlaces aglomerativos (Ward, completo, medio y simple), generando un total de 35 configuraciones.

La Figura 3.1 resume la calidad geométrica interna de cada configuración mediante tres indicadores estándar: coeficiente de Silhouette [18], índice de Calinski-Harabasz [19] e índice de Davies-Bouldin [20]. Antes de interpretarla conviene definir formalmente cada métrica, ya que cada una mide un aspecto distinto de la calidad de la agrupación.

Coeficiente de Silhouette. Para cada observación i asignada a un clúster, se define $a(i)$ como la distancia media a las demás observaciones de su propio clúster, y $b(i)$ como la mínima distancia media a las observaciones de cualquier otro clúster. El coeficiente individual es

$$s(i) \;=\; \frac{b(i) - a(i)}{\max\{a(i),\, b(i)\}} \;\in\; [-1, 1], \tag{3.1}$$

y el coeficiente global es la media de los $s(i)$. Valores próximos a 1 indican que las observaciones están claramente más cerca de su propio clúster que de cualquier otro; valores próximos a 0 indican fronteras difusas; valores negativos sugieren asignaciones incorrectas. *Se busca maximizar.*

Índice de Calinski–Harabasz. Compara la dispersión entre clústeres con la dispersión interna a cada clúster:

$$\mathrm{CH} \;=\; \frac{\mathrm{tr}(B_K)}{\mathrm{tr}(W_K)} \cdot \frac{N - K}{K - 1}, \tag{3.2}$$

donde B_K y W_K son las matrices de dispersión entre clústeres e intra-clúster, N es el número total de observaciones y K el número de grupos. Valores altos indican clústeres compactos y bien separados. El término $(N - K)/(K - 1)$ penaliza el aumento del número de clústeres, por lo que el índice no crece automáticamente con K. *Se busca maximizar.*

Índice de Davies–Bouldin. Para cada par de clústeres se calcula la suma de sus dispersiones internas dividida por la distancia entre sus centroides. El índice se define como el promedio, sobre todos los clústeres, del máximo de esa razón frente a cualquier otro grupo:

$$\mathrm{DB} \;=\; \frac{1}{K} \sum_{k=1}^{K} \max_{j \neq k} \frac{\sigma_k + \sigma_j}{d(c_k,\, c_j)}, \tag{3.3}$$

donde σ_k es la dispersión media intra-clúster del grupo k y $d(c_k, c_j)$ la distancia entre sus centroides. Cuanto menor sea, mejor es la calidad de la agrupación. *Se busca minimizar.*

Las tres métricas son consistentes en su lógica general (premian clústeres compactos y separados) pero ponderan de forma distinta el efecto del tamaño y la geometría, por lo que conviene leerlas conjuntamente y no elegir una sola. En la figura, los puntos rellenos corresponden a soluciones que respetan el umbral del 2 % para el clúster más pequeño; los huecos a configuraciones con al menos un clúster por debajo de ese umbral.

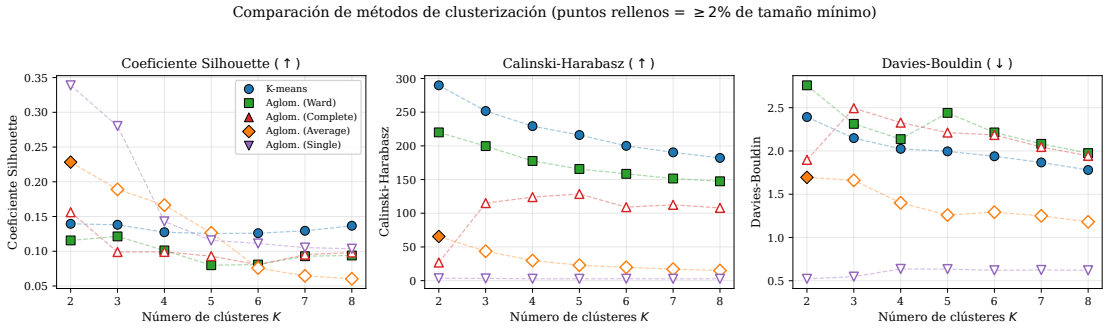

Figura 3.1: Métricas internas para cada combinación de método y número de clústeres. Los puntos rellenos cumplen el filtro de tamaño mínimo del 2 %; los huecos no.

La lectura conjunta de las tres métricas conduce a varias observaciones. En primer lugar, los valores aparentemente más favorables (Silhouette por encima de 0,2 y Davies–Bouldin por debajo de 1) aparecen exclusivamente en configuraciones aglomerativas con enlaces simple y medio. En casi todos los casos se trata de particiones extremadamente desequilibradas, que llegan a contener clústeres formados por un único día y quedan descartadas por el criterio operativo de tamaño: una partición en la que más del 99 % de los días caen en un mismo grupo no resulta útil ni para describir el sistema ni para diferenciar perfiles de spread. La única excepción es el enlace medio con $K = 2$ (Silhouette de 0,228), que supera por poco el umbral del 2 % pero se limita a aislar un grupo marginal (el 2,9 % de los días) frente al resto; una segmentación tan gruesa apenas aporta capacidad descriptiva, por lo que tampoco se ha considerado una solución útil. En segundo lugar, dentro de las configuraciones que sí respetan el umbral del 2 %, K-means se comporta de forma sistemáticamente mejor que las variantes aglomerativas en los dos índices con mayor poder discriminativo en este problema (Calinski–Harabasz y Davies–Bouldin), y de forma comparable o ligeramente superior en Silhouette. En tercer lugar, los valores absolutos de Silhouette no superan $\approx 0{,}14$ en ninguna

solución útil, lo que indica que la nube de puntos en el espacio de características no presenta grupos geométricamente muy aislados, sino más bien un continuo con regiones de mayor densidad. Esta es una observación importante: el clustering no va a recuperar fronteras nítidas, sino más bien *regímenes operativos* con transiciones suaves.

La Tabla 3.1 muestra las diez primeras configuraciones según un ranking medio agregado, calculado como el promedio de los rangos en Silhouette, Calinski–Harabasz y Davies–Bouldin, restringido a las soluciones que respetan el umbral de tamaño. Las dos primeras posiciones las comparten, empatadas, K-means con $K = 3$ y con $K = 8$, seguidas de cerca por K-means con $K = 2$, $K = 7$ y $K = 4$. Que entre las soluciones mejor clasificadas haya una partición muy agregada ($K = 3$) y una fina ($K = 8$) anticipa la decisión que se toma a continuación.

Tabla 3.1: Diez primeras configuraciones según ranking medio agregado, restringidas a las que respetan el umbral de tamaño mínimo del 2 %.

Método	K	Silh. ↑	CH ↑	DB ↓	Mín. (%)	Rk. medio
K-means	3	0,138	251,7	2,149	27,4	5,00
K-means	8	0,137	182,1	1,780	2,7	5,00
K-means	2	0,139	289,8	2,392	45,1	5,33
K-means	7	0,129	190,4	1,867	5,8	5,33
K-means	4	0,127	229,3	2,024	20,6	5,33
Aglom. (average)	2	0,228	65,5	1,695	2,9	5,67
K-means	6	0,126	200,0	1,939	12,2	5,67
K-means	5	0,125	216,1	1,996	16,9	6,33
Aglom. (ward)	3	0,121	199,6	2,313	20,4	9,33
Aglom. (ward)	2	0,116	220,0	2,759	47,0	9,67

Atendiendo a estos resultados, y aplicando los cinco criterios metodológicos de la sección 3.6, se decide trabajar con dos soluciones complementarias:

- Una solución *gruesa* con $K = 3$ clústeres, útil como descripción de primer nivel del sistema. Su tamaño mínimo es muy elevado (27 % de los días) y sus tres grupos son fácilmente narrables.
- Una solución *fina* con $K = 8$ clústeres, que será la referencia principal del trabajo. Sus tamaños están equilibrados (mínimo de 2,7 %) y sus perfiles, como se verá en las secciones siguientes, son interpretables económicamente y producen perfiles de spread diferenciados.

La elección concreta entre $K = 7$ y $K = 8$ merece un comentario aparte. Ambas soluciones son prácticamente equivalentes en calidad geométrica interna y, de hecho, el cruce de sus asignaciones muestra que se trata casi de la misma partición: siete de los ocho grupos de $K = 8$ se conservan prácticamente intactos en $K = 7$. La diferencia se concentra en los días festivos. La solución $K = 8$ los aísla como grupo propio (C3, 47 días), mientras que en $K = 7$ esos días quedan repartidos entre varios clústeres, absorbidos principalmente por el grupo de baja demanda, sin que ningún clúster supere un 8 % de población festiva media. A favor de $K = 7$ juega una estabilidad algo mayor bajo remuestreo (ARI de 0,88 frente a 0,80, Tabla 3.3), consecuencia precisamente de no contener ningún grupo pequeño. A favor de $K = 8$, una separación más significativa del spread medio diario entre clústeres (Kruskal–Wallis con $H = 41{,}7$ frente a $H = 31{,}8$) y, sobre todo, la identificación de la tipología operativamente más nítida de la muestra: como se verá, los festivos presentan el spread medio más negativo y los perfiles más extremos. Por ello se opta por $K = 8$ como solución principal. A título comparativo, las figuras del capítulo muestran también la versión $K = 7$ siempre que aporte información adicional.

3.9. Caracterización de los clústeres en el espacio de variables del sistema

Identificada la solución de referencia, conviene caracterizar cada clúster en términos de las variables originales y no únicamente en términos del algoritmo. La Figura 3.2 presenta un mapa de calor en *z-scores*de las principales variables agregadas por clúster. Los valores positivos indican que el clúster está por encima de la media de la muestra; los negativos, por debajo. Esta representación facilita identificar, de un vistazo, los rasgos económicos distintivos de cada grupo.

A partir de esta representación, y completada con los valores brutos de la Tabla 3.2, los ocho clústeres pueden describirse como sigue:

C0 ($n = 249$) Demanda media-baja, FV ligeramente por encima de la media, eólica moderada y disponibilidad nuclear claramente reducida ($z = -2{,}2$). Se trata de *días con menor capacidad firme*, distribuidos a lo largo del periodo.

C1 ($n = 390$) Demanda elevada, FV muy reducida ($z = -1{,}1$), gas bajo y potencia instalada moderada. Corresponde a *días fríos de invierno pre-crisis del gas*; es el clúster más numeroso.

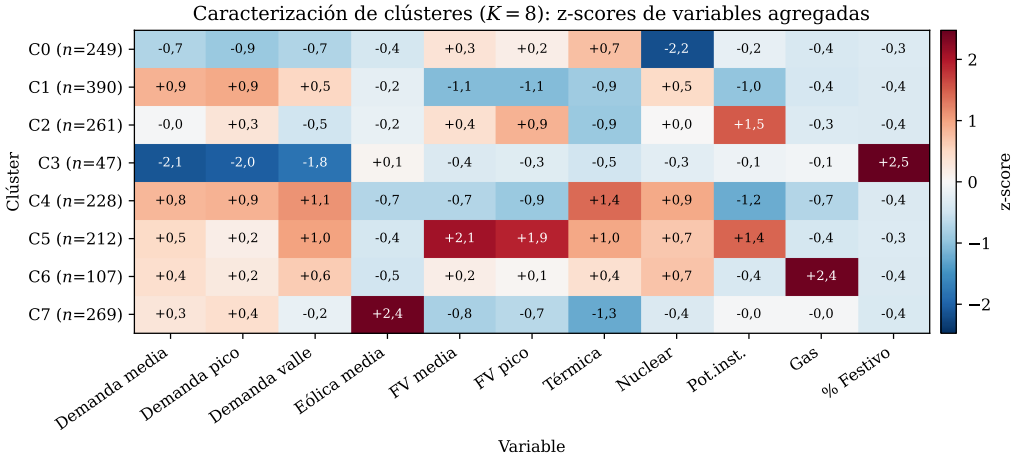

Figura 3.2: Caracterización de los ocho clústeres ($K = 8$) mediante z-scores de variables agregadas. Cada celda indica cuántas desviaciones estándar se separa la media del clúster del promedio global de la muestra.

Tabla 3.2: Caracterización compacta de los clústeres de la solución $K = 8$. Valores medios por clúster: n días; demanda y renovables en GW; gas en €/MWh; potencia instalada en GW; % festivo en porcentaje de población peninsular afectada; spread medio diario en €/MWh.

Clúster	n	Dem. (GW)	Eól. (GW)	FV med. (GW)	FV pico (GW)	Gas (€/MWh)	Pot. inst. (GW)	% Fest.	Spread (€/MWh)
C0	249	26,2	5,4	4,1	10,4	35,6	111,3	2,1 %	-0,46
C1	390	29,5	5,9	1,5	5,2	34,8	103,3	1,3 %	-0,43
C2	261	27,7	6,0	4,2	13,3	38,8	129,1	1,5 %	-0,94
C3	47	23,4	6,7	2,8	8,5	52,4	112,2	86,1 %	-1,74
C4	228	29,5	4,6	2,2	5,9	19,5	100,8	1,4 %	-0,02
C5	212	28,8	5,4	7,2	17,7	34,2	128,3	2,1 %	-0,77
C6	107	28,6	5,3	3,7	9,9	173,5	109,3	1,6 %	+0,13
C7	269	28,4	12,7	2,0	6,9	53,9	113,2	1,0 %	-0,20

C2 ($n = 261$) Gas ya normalizado, en torno a la media de la muestra, potencia instalada en su nivel más alto ($z = +1{,}5$), eólica algo por encima de la media. Se concentra en los meses fríos (febrero y octubre-diciembre) de los años 2023-2025; corresponde a los *inviernos recientes con gas normalizado y sistema expandido*. Es el clúster que aparece sólo en $K = 8$ y queda repartido en $K = 7$.

C3 ($n = 47$) El clúster más pequeño y, a la vez, el más distintivo: el % festivo alcanza un valor extremo ($z = +2{,}5$, media del 86 %) y la demanda cae casi dos desviaciones por debajo de la media. Son *días festivos con afectación territorial alta*.

C4 ($n = 228$) Demanda alta, FV con valor absoluto reducido, gas extremadamente barato ($z = -0{,}7$, media de 19,5 €/MWh, el mínimo de la muestra) y potencia instalada en el extremo inferior. Más del 80 % de los días caen entre junio y septiembre, y prácticamente todos pertenecen a 2019-2021. Son los *veranos pre-crisis del gas*; conviene subrayar que su FV es baja en términos absolutos no por estacionalidad, sino porque en ese periodo el parque fotovoltaico peninsular era todavía muy inferior al actual.

C5 ($n = 212$) FV muy elevada ($z = +2{,}1$), potencia instalada alta ($z = +1{,}4$) y demanda media-alta. Su composición es complementaria a la de C4: también concentrado en junio–septiembre, pero exclusivamente en 2023-2025. Son los *veranos posteriores a la expansión fotovoltaica*.

C6 ($n = 107$) El rasgo dominante es el gas en niveles extremos ($z = +2{,}4$, media de 173 €/MWh). Engloba los *días de la crisis energética de 2022*.

C7 ($n = 269$) Eólica extraordinariamente alta ($z = +2{,}4$, más del doble que el resto), térmica reducida. Son los *días borrascosos* con generación renovable abundante.

La Figura 3.3 confirma estas interpretaciones al mostrar la distribución de cada clúster por mes y por año. C1 se concentra en los meses fríos (octubre-marzo) de los años 2019-2021 y desaparece casi por completo a partir de 2023, al disolverse el régimen de gas barato. C4 ocupa, dentro de ese mismo periodo 2019-2021, el papel complementario en los meses de verano (junio-septiembre). C5 hereda ese rol veraniego en los años recientes (2023-2025), reflejando la expansión fotovoltaica peninsular. C6 está prácticamente confinado a 2022, el año de la crisis del gas. C2 aparece en los inviernos de 2023-2025, coincidiendo con la normalización del precio del gas y los nuevos niveles de potencia instalada.

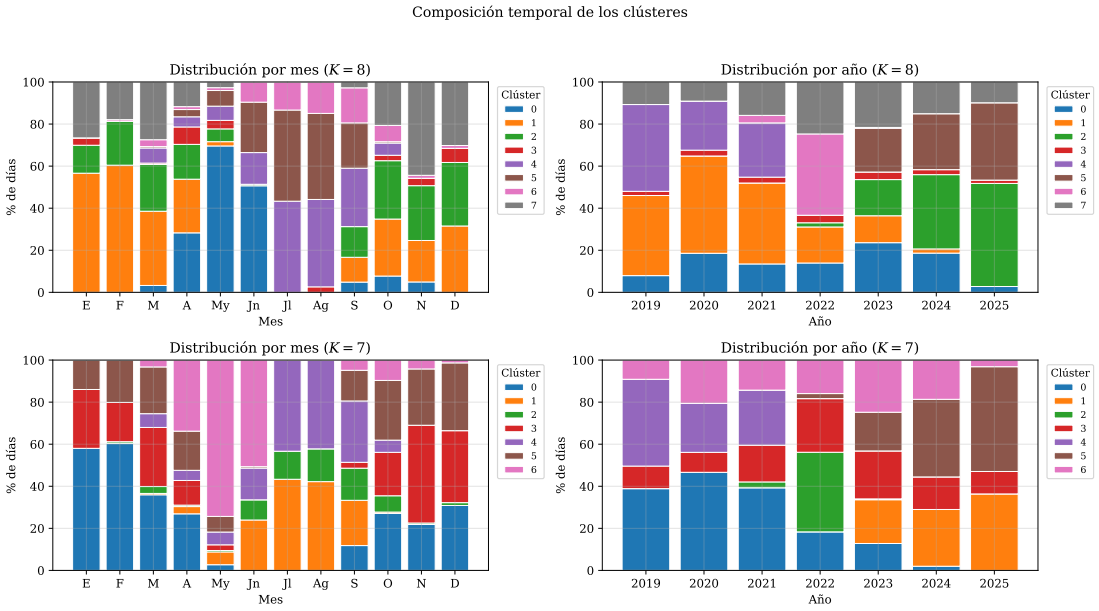

Figura 3.3: Composición temporal de los clústeres por mes y por año, para las soluciones $K = 8$ (arriba) y $K = 7$ (abajo).

Esto tiene una implicación importante. Aunque el clustering se ha construido únicamente a partir de variables explicativas del sistema (sin incorporar etiquetas temporales), los grupos resultantes recogen, de forma emergente, los principales *regímenes históricos* del mercado: el periodo pre-crisis, la propia crisis del gas de 2022 y el régimen reciente de expansión renovable con gas normalizado. Esto es consistente con la literatura, que documenta cambios estructurales significativos en el precio eléctrico español asociados a estos eventos [2, 3], y confirma que las variables utilizadas en la clusterización tienen suficiente poder descriptivo para identificar fases distintas del mercado sin necesidad de imponerlas externamente.

La Figura 3.4 completa la caracterización mostrando los perfiles horarios medios de demanda, eólica y fotovoltaica por clúster. Los perfiles confirman lo que la tabla resume escalarmente: C7 es el clúster de mayor producción eólica a todas las horas, C5 es el de mayor FV con un pico diurno claramente diferenciado del resto, y C3 (festivos) presenta una demanda significativamente reducida durante toda la jornada laboral.

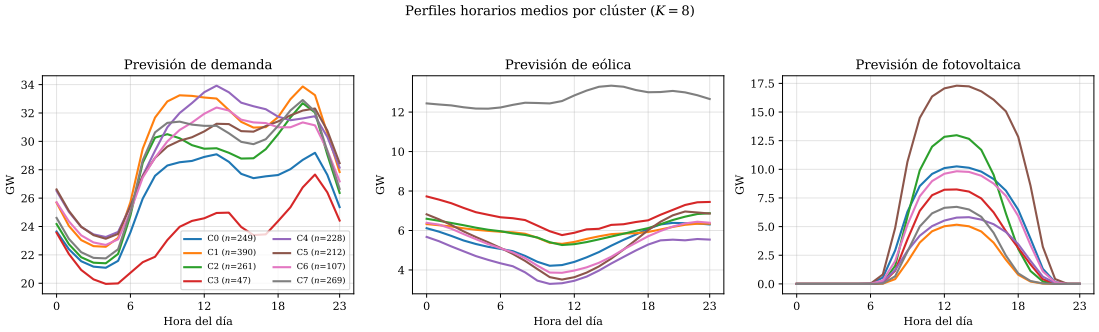

Figura 3.4: Perfiles horarios medios por clúster ($K = 8$) en las tres variables horarias de mayor peso explicativo: demanda prevista, eólica prevista y fotovoltaica prevista.

3.10. Perfiles de spread por clúster

La pregunta clave de esta fase no es si los clústeres están bien definidos en el espacio de variables explicativas, sino si, una vez agrupados los días por el contexto previo al cierre del mercado diario, los perfiles del spread $P^{ID} - P^{MD}$ difieren entre clústeres lo suficiente como para resultar útiles. Esta es la condición que separa una agrupación puramente descriptiva de una agrupación con utilidad operativa.

Conviene recordar aquí cómo se construye la variable objetivo, ya descrita en la sección 2.3. Tanto el mercado diario como el intradiario casan un precio independiente para cada una de las 24 horas de entrega del día. El mercado diario fija esos 24 precios el día anterior a la entrega; el intradiario los reajusta posteriormente, en una o varias sesiones, a medida que llega nueva información sobre demanda, generación renovable o disponibilidad de las centrales. El spread de una hora concreta es, por tanto, la diferencia entre el precio que esa misma hora de entrega obtuvo en el diario y el que obtiene en el intradiario. En este trabajo se emplea la primera sesión del mercado intradiario, por ser la de mayor cobertura horaria y liquidez. Cuando se habla del "perfil horario" del spread, el eje temporal se refiere a la hora de entrega de la energía, no al instante en que se negocia.

Para responder a la pregunta planteada se han evaluado dos tipos de evidencia. En primer lugar, se contrasta estadísticamente si las distribuciones del spread medio diario difieren entre clústeres. Antes de presentar los resultados conviene definir los tres contrastes utilizados.

El análisis de la varianza (ANOVA) contrasta la hipótesis nula de que la media del spread medio diario es idéntica en todos los clústeres. Su estadístico, denotado

F, compara la variabilidad entre grupos con la variabilidad interna de cada grupo: un valor de F grande, asociado a un p-valor pequeño, indica que las medias difieren más de lo esperable por azar. El test de Kruskal–Wallis es su análogo no paramétrico, es decir, contrasta la misma hipótesis pero sobre los rangos de las observaciones, sin asumir normalidad ni homogeneidad de varianzas. Su estadístico se denota H. Se incluyen ambos porque el spread presenta colas pesadas que se alejan de la normalidad, de modo que un resultado coincidente entre los dos refuerza la robustez de la conclusión. Finalmente, el índice de Rand ajustado (ARI) [21] no compara spreads, sino dos particiones de los mismos días entre sí. Mide el grado de acuerdo entre dos agrupaciones corrigiendo por el que se obtendría de forma puramente aleatoria, tomando el valor 1 cuando las dos particiones son idénticas y un valor próximo a 0 cuando su coincidencia es la esperable por azar. Se emplea en el análisis de estabilidad de la sección 3.11 (véase el Anexo A para su definición formal).

La Tabla 3.3 resume los resultados conjuntos de calidad interna, validación estadística y estabilidad de las tres soluciones candidatas.

Tabla 3.3: Validación cuantitativa de las tres soluciones candidatas. Las columnas Silhouette, Calinski–Harabasz y Davies–Bouldin reportan calidad geométrica interna; ANOVA y Kruskal–Wallis comprueban si el spread medio diario difiere entre clústeres; ARI mide la estabilidad de la partición por bootstrap (50 réplicas al 80 %).

K	Silh.	CH	DB	Mín. (%)	F (ANOVA)	H (KW)	ARI	σ_{ARI}
3	0,138	251,7	2,149	27,4	$2{,}45\,(p = 0{,}087)$	$5{,}3\,(p = 0{,}070)$	0,884	0,188
7	0,129	190,4	1,867	5,8	$3{,}95\,(p < 10^{-3})$	$31{,}8\,(p < 10^{-3})$	0,883	0,071
8	0,137	182,1	1,780	2,7	$4{,}24\,(p < 10^{-3})$	$41{,}7\,(p < 10^{-3})$	0,804	0,095

Para las soluciones $K = 7$ y $K = 8$, tanto el ANOVA como el test de Kruskal-Wallis rechazan con claridad la hipótesis de igualdad del spread medio diario entre clústeres ($p < 10^{-3}$ en los cuatro contrastes). Para la solución gruesa $K = 3$, en cambio, las diferencias no alcanzan la significación convencional ($p \approx 0{,}07$-$0{,}09$): tres macro-regímenes no bastan para separar comportamientos del spread. La lectura es doble: agrupar los días por el contexto disponible *ex ante* permite distinguir perfiles de spread estadísticamente diferentes, pero solo cuando la segmentación es suficientemente fina, lo que refuerza la elección de $K = 8$ como referencia principal. Conviene matizar, no obstante, que esta significación solo dice que las diferencias entre clústeres no son casualidad, no que sean grandes en términos económicos. La magnitud se examina más adelante.

Esta evidencia agregada no captura, además, el rasgo principal del comporta-

miento del spread: su perfil intradía no es plano. La Figura 3.5 muestra, para cada clúster, la trayectoria horaria del spread medio. Aunque la magnitud absoluta de las diferencias puede parecer reducida (en torno a ±2 €/MWh respecto a cero), su distribución a lo largo del día es muy distinta entre grupos.

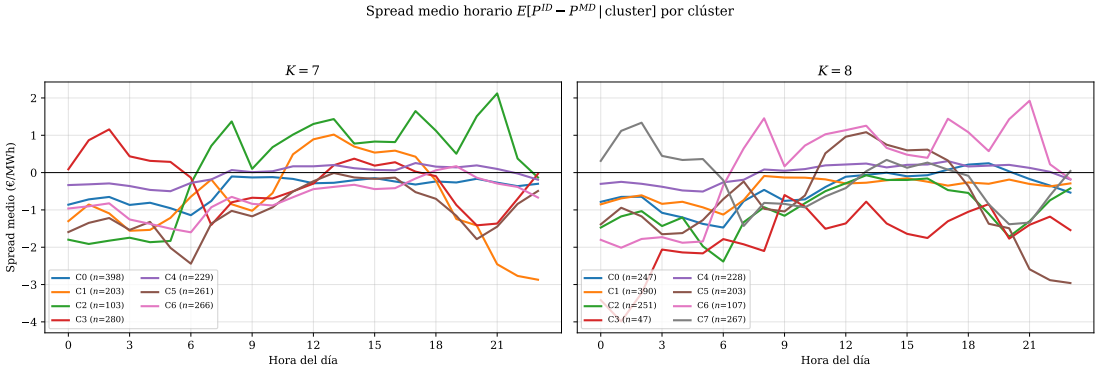

Figura 3.5: Spread medio horario $E[P^{ID} - P^{MD} \mid \text{cluster}]$ por clúster, en las soluciones $K = 7$ y $K = 8$. La línea gruesa de cada color es el promedio horario del clúster correspondiente.

Los hallazgos más relevantes son los siguientes. Algunos clústeres presentan un perfil bifásico, con cambio de signo a lo largo del día. El clúster C7 ($K = 8$), formado por días borrascosos con eólica extrema, muestra spread positivo en la madrugada (hasta +1,3 €/MWh entre las 0 h y las 5 h) y claramente negativo en el tramo final del día (hasta −1,4 €/MWh entre las 19 h y las 22 h). El clúster C6 ($K = 8$) presenta el patrón opuesto: spread muy negativo en la madrugada (hasta −2 €/MWh) y positivo al final del día (hasta +1,9 €/MWh en torno a las 21 h). Estos dos patrones opuestos encajan con lo que cabría esperar de cada régimen, aunque no se ha analizado en detalle por qué ocurren. En el caso de C7, una explicación razonable es que la gran cantidad de eólica baje mucho el precio del diario en las horas de menos demanda, y que el intradiario no llegue a corregir del todo esa bajada. En C6, durante la crisis del gas, el patrón contrario podría deberse a que las tensiones de precio aparecen sobre todo al final del día, cuando el intradiario ya refleja mejor la situación real del sistema que el diario, que se cerró el día anterior.

Otros clústeres presentan un perfil monótonamente negativo. El más claro es C3 ($K = 8$), los días festivos, cuyo spread es negativo a todas las horas y alcanza valores especialmente bajos en la madrugada (entre −3 y −4 €/MWh en las tres primeras horas del día). En estos días el intradiario tiende sistemáticamente a cotizar por debajo del diario. Una posible explicación, que no se comprueba aquí, es que el diario incorpore un margen de seguridad por la incertidumbre del festivo

y que el intradiario lo vaya corrigiendo a la baja conforme se acerca la hora de entrega.

Finalmente, varios clústeres (C1 y C4 en $K = 8$) muestran un spread medio prácticamente plano en torno a cero. Esto no significa que no exista variabilidad, ya que de hecho su desviación estándar horaria es alta, sino que en estos grupos las fuerzas que mueven el spread se compensan en promedio, sin un patrón sistemático apreciable.

La diferencia entre estos comportamientos se aprecia mejor en la Figura 3.6, que descompone el spread en tres categorías: por debajo de -1 €/MWh, dentro del rango de $[-1, +1]$ €/MWh, y por encima de $+1$ €/MWh. Esta descomposición es relevante porque, en términos prácticos, las desviaciones de pequeña magnitud no justifican una decisión de compra-venta entre mercados una vez se consideran los costes operativos asociados a la operación. El umbral de ± 1 €/MWh se ha fijado como referencia de esa magnitud mínima.

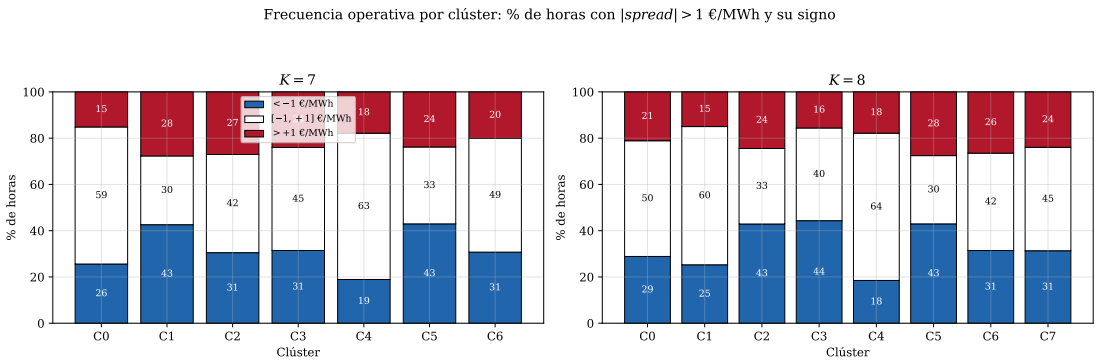

Figura 3.6: Frecuencia operativa del spread por clúster, expresada como porcentaje de horas con spread por debajo de -1 €/MWh, dentro de la banda $[-1, +1]$ €/MWh o por encima de $+1$ €/MWh.

Los contrastes entre clústeres son apreciables. El clúster C4 ($K = 8$) presenta el mayor porcentaje de horas dentro de la banda central (un 64 %), seguido de C1 (60 %). En el extremo opuesto, C5 supera el 70 % de horas situadas fuera de dicha banda y C2 se aproxima a ese nivel (67 %); en ambos predomina el extremo negativo, pero la cola positiva es también apreciable (un 28 % y un 24 % de las horas, respectivamente). El clúster festivo C3 es, en cambio, el más asimétrico: el 44 % de sus horas cae por debajo de -1 €/MWh frente a solo un 16 % por encima de $+1$ €/MWh. La distinción operativa entre clústeres no se basa, por tanto, únicamente en el signo del spread, sino también en su magnitud y en su tendencia a desviarse de cero.

La Tabla 3.4 cuantifica estos porcentajes, junto con la media y la desviación estándar del spread horario por clúster. [1]

Tabla 3.4: Frecuencia operativa del spread por clúster ($K = 8$). Cada fila es un clúster: porcentaje de horas en las que el spread $P^{ID} - P^{MD}$ cae por debajo de -1 €/MWh, dentro de $[-1, +1]$ €/MWh, o por encima de $+1$ €/MWh; media μ y desviación σ del spread horario.

Clúster	n días	$\% < -1$	$\% \in [-1, +1]$	$\% > +1$	μ (€/MWh)	σ (€/MWh)
C0	247	28,9	50,0	21,1	-0,46	4,21
C1	390	25,3	59,8	15,0	-0,43	3,34
C2	251	42,9	32,6	24,5	-0,94	5,66
C3	47	44,3	40,1	15,6	-1,74	5,66
C4	228	18,5	63,7	17,9	-0,02	1,72
C5	203	42,9	29,5	27,5	-0,77	6,21
C6	107	31,4	42,1	26,5	+0,13	7,17
C7	267	31,3	44,7	24,0	-0,20	5,92

Para completar la caracterización temporal, la Figura 3.7 presenta el desglose por clúster y por hora de los dos extremos operativos: el porcentaje de horas con spread por encima de $+1$ €/MWh (potencialmente favorable a una estrategia de venta en el intradiario) y por debajo de -1 €/MWh (favorable a la operación inversa). Esta figura resulta especialmente útil porque junta la parte horaria con la parte operativa.

Algunos resultados destacan especialmente por su valor operativo. En el clúster C3 ($K = 8$), entre las 0 h y las 6 h, el spread se sitúa por debajo de -1 €/MWh en más del 50 % de las horas, con máximos por encima del 60 %. En el clúster C5 ($K = 8$), entre las 19 h y las 23 h, ese mismo porcentaje alcanza hasta el 60 %. En el clúster C7 ($K = 8$), entre las 0 h y las 5 h, el spread supera $+1$ €/MWh en torno al 30 % de las horas. Estos son los tramos en los que la agrupación previa del día parece aportar una señal más clara. Por eso, también son las horas en las que cabría esperar que el modelo predictivo desarrollado en el capítulo siguiente tenga mayor capacidad para distinguir entre distintos comportamientos del spread.

La Figura 3.8 muestra, para los ocho clústeres, el perfil horario completo del spread mediante la mediana y el rango intercuartílico (IQR). Esta representación

[1] El número de días por clúster de esta tabla es ligeramente inferior al de la Tabla 3.2 (1.740 frente a 1.763) porque el cálculo de las frecuencias horarias exige disponer del perfil completo de 24 horas del spread, y 23 jornadas del tramo final de la muestra (una de 2023, ocho de2024 y catorce de 2025) no lo tienen por ausencia de alguno de los precios horarios necesarios.

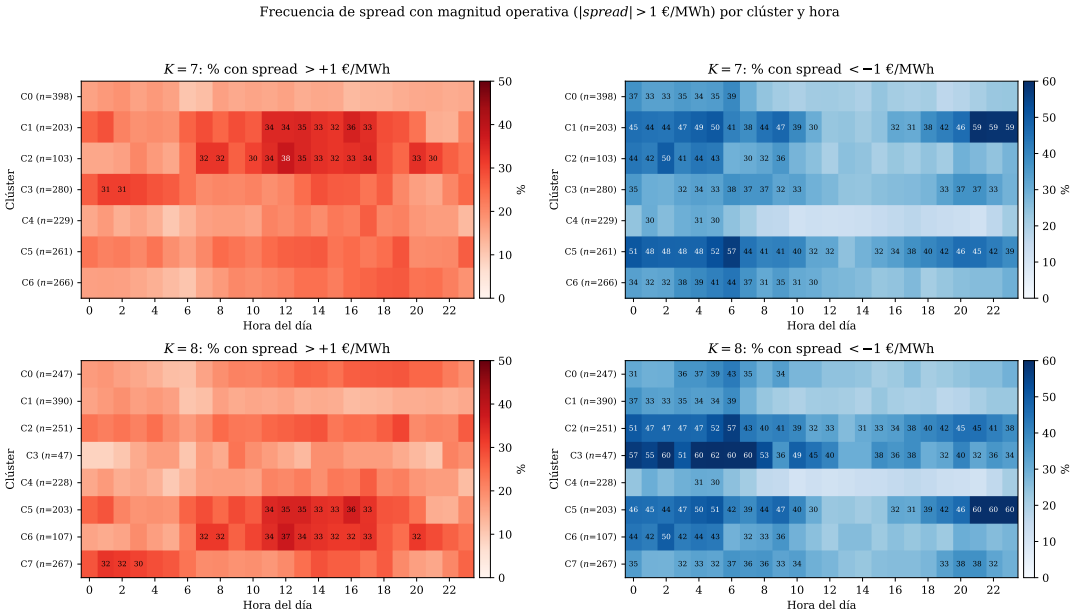

Figura 3.7: Frecuencia horaria de spread con magnitud operativa ($|\text{spread}| > 1$ €/MWh) por clúster ($K = 7$ arriba, $K = 8$ abajo). Las celdas con valores por encima del 30 % aparecen anotadas.

resulta útil porque capta mejor la asimetría del spread: la mediana es menos sensible que la media a la presencia de colas pesadas. No obstante, presenta una limitación relevante. Dado que una parte importante de los valores se concentra exactamente en cero (el 19 % de las horas), las medianas tienden a acercarse al origen, haciendo que algunos perfiles parezcan más planos de lo que son en términos esperados. Por ello, el análisis del spread medio (Figura 3.5) permite observar diferencias entre clústeres más nítidas que el análisis basado en medianas.

Como complemento, la Figura 3.9 muestra el mismo perfil horario con una banda de percentiles 5–95 en lugar del rango intercuartílico, lo que permite apreciar mejor la magnitud de las colas de la distribución. Los clústeres asociados a episodios extremos del sistema, como C6 (crisis del gas) y C7 (días borrascosos), muestran bandas notablemente más anchas que el resto, lo que confirma que su mayor variabilidad no se limita a unos pocos valores atípicos aislados.

Por completitud, la Figura 3.10 presenta la distribución del spread medio diario por clúster mediante diagramas de caja, excluyendo los valores extremos para facilitar su interpretación. Aunque los valores medianos se mantienen relativamente próximos a cero, las distribuciones muestran diferencias claras en posición y dispersión entre grupos, coherentes con los resultados de las pruebas estadísticas.

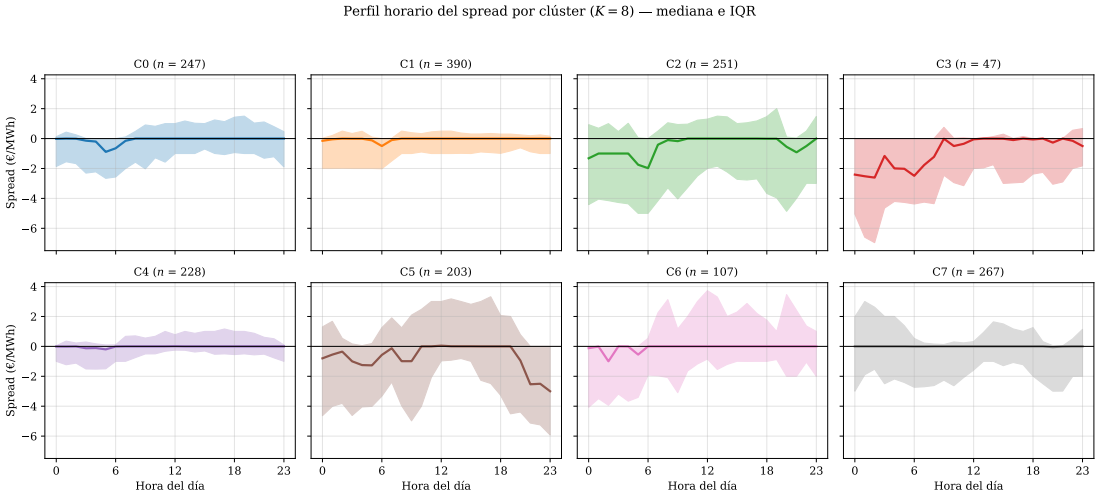

Figura 3.8: Perfil horario del spread por clúster ($K = 8$): la línea representa la mediana horaria y la banda sombreada el rango intercuartílico (Q25–Q75).

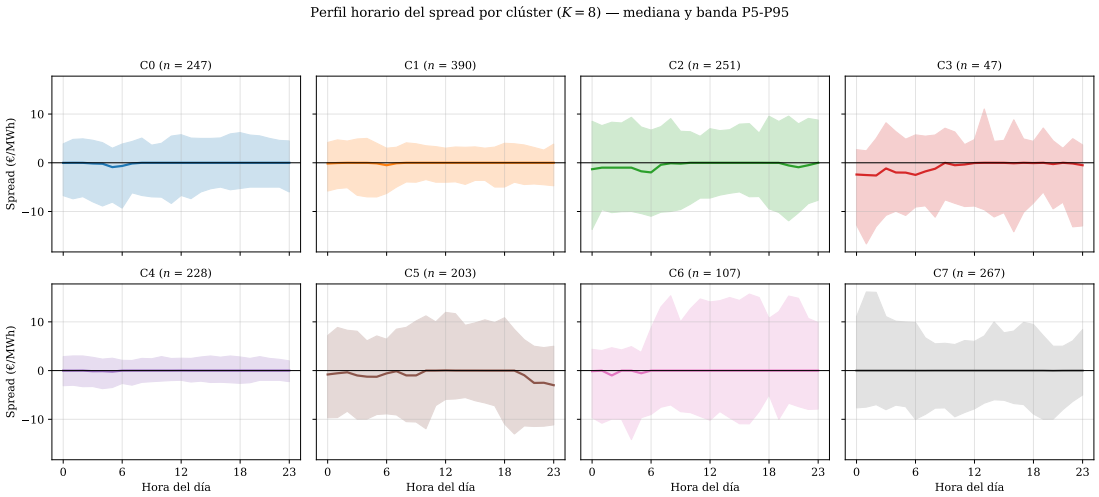

Figura 3.9: Perfil horario del spread por clúster ($K = 8$): la línea representa la mediana horaria y la banda sombreada el rango entre los percentiles 5 y 95.

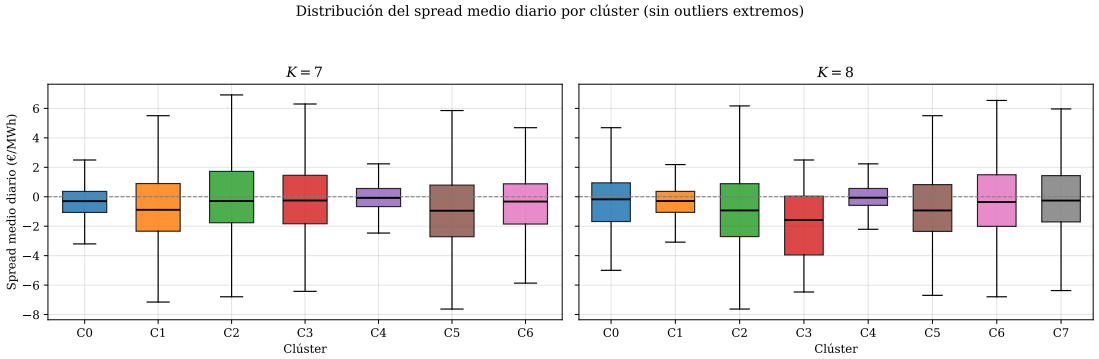

Figura 3.10: Distribución del spread medio diario por clúster (sin valores extremos, para mejorar la legibilidad).

3.11. Estabilidad y limitaciones de la agrupación obtenida

Hasta este punto se ha trabajado con una única partición ajustada sobre la muestra completa. La pregunta natural que surge ahora es si esa partición es robusta o si, por el contrario, depende demasiado del conjunto concreto de días utilizado. Para responder a esta cuestión, se ha aplicado un procedimiento de bootstrap por submuestreo: se generan 50 remuestreos sin reemplazo del 80 % de la muestra, se ajusta el algoritmo K-means en cada uno de ellos y se mide su grado de acuerdo con la partición global mediante el índice ARI ya descrito. Un valor de ARI cercano a 1 indica que la partición es estable, mientras que un valor cercano a 0 indica que cambia de forma casi arbitraria al modificar la muestra.

La Figura 3.11 muestra que las tres soluciones presentan una estabilidad razonable, con ARI medios iguales o superiores a 0,80. La lectura más informativa está en la dispersión entre réplicas. La solución gruesa $K = 3$ alcanza un ARI medio alto (0,88), pero con una desviación estándar muy elevada (0,19): en la mayoría de los remuestreos la partición se recupera casi idéntica, pero en una fracción apreciable de ellos las fronteras entre los tres macro-grupos se desplazan de forma sustancial. La solución $K = 7$ es la más consistente (0,88 con desviación 0,07). La solución $K = 8$ presenta un ARI medio algo menor (0,80), diferencia atribuible en buena parte a su clúster más pequeño, el festivo C3 (47 días), cuya recuperación exacta en submuestras del 80 % resulta más difícil, y al reparto de días frontera entre regímenes contiguos. En conjunto, lapenalización de estabilidad asociada a aislar el grupo festivo es moderada, y se considera un precio razonable a cambio de identificar la tipología más nítida y operativa de la muestra.

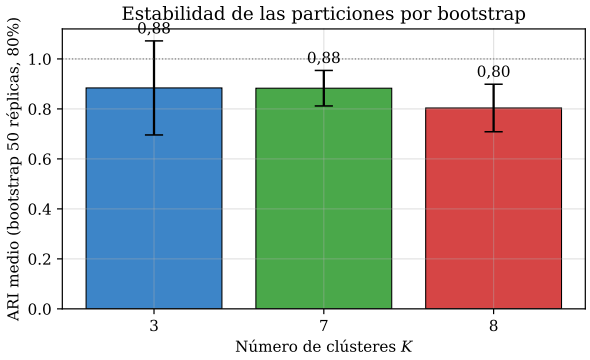

Figura 3.11: Estabilidad de las particiones K-means para $K = 3$, $K = 7$ y $K = 8$, medida por el ARI medio frente a la partición global, sobre 50 remuestreos al 80 % de la muestra. Las barras de error indican la desviación estándar entre réplicas.

Limitaciones. Las observaciones realizadas a lo largo del capítulo permiten matizar el alcance de los resultados:

- La muestra analítica final consta de 1 763 días, distribuidos entre enero de 2019 y diciembre de 2025. El periodo bruto de descarga cubría 2014–2025, pero la disponibilidad efectiva de la previsión de demanda fija el inicio en 2019. Además, la cotización de los futuros TTF solo está disponible en días laborables, lo que en la práctica excluye los sábados y domingos de la muestra. La composición resultante por tipo de día (un 18,7 % de lunes, un 61,2 % de días de martes a jueves y un 20,1 % de viernes y festivos laborables) está, por tanto, sesgada hacia los días laborables.

- Los valores absolutos del coeficiente de Silhouette se mantienen por debajo de 0,14 en todas las soluciones útiles. Esto indica que el espacio de variables explicativas no contiene grupos claramente separados desde un punto de vista geométrico, sino transiciones suaves. Por ello, los clústeres identificados deben interpretarse como regímenes operativos, más que como categorías cerradas, y cabe esperar cierto solapamiento en sus fronteras.

- Aunque las pruebas de hipótesis confirman diferencias estadísticamente significativas en el spread entre clústeres, las magnitudes esperadas son moderadas (del orden de ± 1 y ± 2 €/MWh en media horaria, con extremos puntuales de hasta -4€/MWh en la madrugada del clúster festivo). Esto sugiere que la utilidad operativa de la agrupación se concentra en los tramos horarios de mayor contraste y en las colas de la distribución, más que en la mediana del spread.

- Una parte del poder discriminativo del clustering procede del componente temporal implícito en algunas variables, como el precio del gas o la potencia instalada. Esto significa que la asignación de un día futuro a un clúster depende, en parte, del régimen general del sistema en ese momento. Por tanto, será necesario actualizar periódicamente el modelo si se quiere mantener su validez a lo largo del tiempo.

- Las variables de generación renovable (eólica y fotovoltaica) se incorporaron a la matriz de agrupación en términos absolutos (MW previstos), sin normalizar por la potencia instalada de cada tecnología. Dado que dicha potencia instalada ha crecido de forma sustancial a lo largo del periodo 2019-2025, un nivel de generación que en 2019 resultaría comparativamente alto podría ser, en términos relativos, moderado o incluso bajo en 2025. Esto introduce un riesgo de que parte de la separación entre clústeres asociados a periodos distintos (por ejemplo, entre los veranos previos y posteriores a la expansión fotovoltaica) refleje, al menos parcialmente, esta tendencia de capacidad instalada y no únicamente diferencias genuinas en las condiciones meteorológicas u operativas de cada día. Normalizar estas variables por la potencia instalada vigente en cada fecha, o emplear la demanda térmica prevista (la diferencia entre la demanda y la generación renovable prevista) como magnitud de síntesis, habría mitigado este efecto.

En conjunto, la solución $K = 8$ ofrece una representación bastante completa del sistema, permite distinguir de forma significativa distintos perfiles de spread y muestra una estabilidad suficiente bajo remuestreo como para servir de base sólida para la fase posterior de modelización supervisada. Las limitaciones detectadas no invalidan los resultados, pero sí ayudan a acotar su interpretación.

Capítulo 4

Predicción del spread mediante aprendizaje supervisado: planteamiento y metodología

4.1. Introducción

El capítulo anterior se centró en organizar el conjunto de días disponibles mediante técnicas de aprendizaje no supervisado. La agrupación de los días en clústeres permitió identificar configuraciones recurrentes del sistema a partir de variables conocidas antes del cierre del mercado diario, mostrando además que el spread entre el mercado diario y el intradiario no presenta un comportamiento uniforme, sino que adopta perfiles distintos en función del contexto previo a la operativa.

Ese análisis, sin embargo, tenía un carácter principalmente descriptivo. La clusterización permite ordenar observaciones similares y detectar patrones recurrentes, pero no proporciona por sí misma una predicción para un día futuro. El objetivo de este capítulo es, precisamente, pasar de esa descripción de regímenes históricos a un planteamiento predictivo, en el que se fija una variable objetivo y se construyen modelos capaces de anticiparla a partir de información disponible antes de la decisión.

En este trabajo, predecir no significa estimar el spread horario con precisión exacta, sino obtener una señal suficientemente informativa como para apoyar una decisión operativa. La cuestión relevante no es solo cuánto se aproxima la predicción al valor observado, sino si permite decidir razonablemente entre comprar en

el mercado diario y vender en el intradiario, adoptar la posición contraria o no operar. Esta orientación condiciona el resto del capítulo: la selección de variables, las métricas de evaluación y la comparación entre modelos se plantean siempre desde una perspectiva ligada a la decisión.

El capítulo se articula en torno a tres ideas. La primera es, como en el capítulo anterior, la construcción de una matriz de entrada estrictamente anticipada, formada únicamente por variables disponibles antes del cierre del mercado diario. La segunda es la comparación de los modelos de aprendizaje con referencias simples, que permiten valorar si el modelo aporta información adicional respecto a reglas triviales. La tercera es la adopción de una estrategia de validación temporal, en la que cada modelo se entrena con información pasada y se evalúa sobre observaciones futuras, imitando las condiciones reales en las que tendría que operar. Estas tres decisiones metodológicas preparan la lectura del capítulo de resultados.

4.2. Formulación del problema

La variable de interés sigue siendo el spread horario definido en el capítulo 2, es decir, la diferencia entre el precio del mercado intradiario y el del mercado diario para una misma hora de entrega,

$$\text{Spread}_{d,h} = P^{\text{ID}}_{d,h} - P^{\text{MD}}_{d,h}. \tag{4.1}$$

La decisión económica asociada al spread es esencialmente direccional. Si se anticipa que el precio intradiario será superior al precio diario, una posición natural consistiría en comprar en el mercado diario y vender posteriormente en el intradiario. Si se anticipa lo contrario, la posición se invertiría. Ahora bien, cuando la señal esperada es débil, la alternativa razonable no es necesariamente tomar la posición opuesta, sino abstenerse. Por tanto, aunque el primer problema predictivo se formule en términos de signo, la decisión operativa final incorporará también un criterio de confianza o umbral.

A partir de esta consideración se plantean dos formulaciones complementarias sobre la misma variable. Ambas se mantienen en paralelo porque proporcionan información distinta y porque su comparación permite analizar hasta qué punto la formulación elegida condiciona los resultados.

4.2.1. El problema natural: predecir el signo

El planteamiento más directo, y el que mejor encaja con la lógica operativa, consiste en predecir el signo del spread. Se trata de un problema de clasificación: para cada hora de cada día, el modelo debe decidir si el spread será positivo o negativo. La variable objetivo es, por tanto, una etiqueta binaria,

$$y_{d,h}^{\text{signo}} = \begin{cases} +1 & \text{si Spread}_{d,h} > 0, \\ -1 & \text{si Spread}_{d,h} \leq 0. \end{cases} \tag{4.2}$$

Las horas con spread exactamente nulo se asignan a la clase no positiva. Esta decisión responde a un criterio de codificación conservador. Como se documentó en el capítulo anterior, una proporción apreciable de las horas presenta un spread nulo o muy próximo a cero, y un valor exactamente nulo no debe interpretarse como una señal positiva. La asignación a la clase -1 permite mantener una variable objetivo binaria y resulta coherente con el ligero sesgo negativo observado en la distribución. No obstante, esta convención no implica que todas esas horas den lugar necesariamente a una operación en sentido negativo, ya que la posibilidad de abstenerse se incorpora posteriormente mediante un umbral de confianza.

La predicción del signo plantea una tarea menos exigente que la estimación del valor exacto del spread. Esta simplificación puede resultar especialmente útil en series temporales energéticas, caracterizadas por una elevada variabilidad, la presencia de valores extremos y perturbaciones difíciles de anticipar con la información disponible. En este contexto, la magnitud concreta de una variación puede ser difícil de estimar, aunque las variables explicativas sí permitan identificar parte de su dirección.

En primer lugar, la clasificación requiere menos información, puesto que basta con determinar a qué lado de cero se situará el spread, sin necesidad de anticipar con precisión su magnitud. En segundo lugar, el signo es menos sensible a las colas de la distribución. Un error de varios euros por megavatio hora penaliza de forma importante las métricas de regresión, mientras que, desde el punto de vista direccional, lo relevante es si la predicción queda a un lado u otro de cero. Por último, la concentración de observaciones en torno a cero introduce ruido en la estimación del valor, mientras que en la clasificación queda recogida por la propia definición de la clase no positiva.

Por tanto, la formulación direccional no debe interpretarse como superior a la regresión en cualquier serie temporal, sino como una alternativa potencialmente más robusta cuando la magnitud exacta de la variable presenta una predictibilidad limitada y el objetivo principal consiste en apoyar una decisión de mercado.

4.2.2. La predicción del valor como vía complementaria

El segundo planteamiento consiste en predecir el valor del spread mediante un modelo de regresión, que proporciona para cada hora una estimación continua $\hat{s}_{d,h}$. Se trata de una tarea más exigente que la clasificación del signo, pero aporta información que puede resultar útil para la decisión operativa.

La utilidad de la regresión reside en que el valor predicho ofrece una información adicional que el signo, por sí solo, no proporciona: una medida indirecta de la intensidad de la señal. La magnitud de la estimación, $|\hat{s}_{d,h}|$, indica hasta qué punto el modelo espera que el spread se aleje de cero. Así, una predicción positiva de 0,1 €/MWh no tiene la misma interpretación operativa que una predicción positiva de 5 €/MWh, aunque ambas impliquen la misma dirección. Esta diferencia permite construir reglas selectivas, en las que solo se opera cuando la señal supera un umbral de confianza, cuestión que se retoma en el capítulo de resultados. Cabe señalar que el clasificador podría ofrecer una funcionalidad similar si, en lugar de una decisión binaria, se emplease la probabilidad de pertenencia a la clase positiva que proporciona el modelo: valores próximos a 0,5 indicarían baja confianza y valores próximos a 0 o 1, alta confianza, de forma análoga a como $|\hat{s}_{d,h}|$ gradúa la confianza del regresor. Esta vía alternativa no se explora en este trabajo, que se centra en $|\hat{s}_{d,h}|$ como medida de intensidad de la señal.

Por este motivo, ambas formulaciones se mantienen. La clasificación del signo representa el problema más directamente vinculado con la decisión direccional, mientras que la regresión aporta una medida continua de la intensidad esperada del spread. Comparar ambos enfoques permite valorar si resulta preferible modelizar directamente la dirección o estimar primero el valor y traducirlo después en una decisión operativa. Las diferencias o similitudes entre sus resultados constituyen, por tanto, una cuestión relevante que se examina más adelante.

4.3. Construcción de la matriz de características

Una vez definida la variable objetivo, el siguiente paso consiste en especificar la información que recibirá el modelo. Esta etapa es de gran importancia, ya que la capacidad predictiva queda necesariamente limitada por las variables disponibles en el momento de la decisión. No se trata, por tanto, de reunir el mayor número posible de variables, sino de construir una representación informativa que respete estrictamente la dimensión temporal del problema.

La matriz empleada en este capítulo es una adaptación de la construida para el análisis de clusterización. Se mantienen los principales bloques de información ex-ante ya utilizados anteriormente, ya que el objetivo sigue siendo describir el estado anticipado del sistema antes del cierre del mercado diario. La diferencia es que ahora esas variables no se emplean para agrupar días similares, sino como entrada de un problema supervisado en el que cada observación queda asociada a una variable objetivo. Por ello, la matriz se amplía con retardos del spread, variables horarias y una estructura apilada adecuada para obtener predicciones a nivel horario.

4.3.1. El criterio anticipado

La construcción de la matriz de características respeta el criterio ex-ante introducido en el capítulo 2: se excluye toda variable que solo esté disponible una vez iniciada o finalizada la jornada que se pretende predecir, ya que su inclusión produciría una fuga de información y podría mejorar de forma artificial la capacidad predictiva del modelo.

Esta condición se cumple directamente en las variables de calendario, en las previsiones del sistema eléctrico publicadas antes de la operación y en los retardos del spread correspondientes a días anteriores. En particular, no se utilizan el precio intradiario, el spread ni otras magnitudes realizadas del día que se pretende predecir. Tampoco se incorporan datos reales de generación o demanda de esa misma jornada cuando solo pueden conocerse con posterioridad a la decisión.

Las variables meteorológicas constituyen la única excepción parcial a este criterio: como ya se explicó en el capítulo 2, corresponden a valores observados y no a las previsiones que habrían estado disponibles antes de cada jornada. Esta elección no responde a una decisión metodológica, sino a la disponibilidad real de los datos: no existe un histórico público y gratuito de previsiones meteorológicas pasadas. Esta aproximación obliga a interpretar con cautela la aportación atribuida a las variables meteorológicas en el análisis de interpretabilidad del capítulo 5, ya que su uso podría ofrecer una evaluación algo más favorable que la que se obtendría en una aplicación estrictamente ex-ante.

4.3.2. Bloques de variables

La matriz hereda los grandes bloques de información definidos en la construcción de la base de datos. El primero es el bloque del sistema eléctrico, que recoge la

previsión de demanda, las previsiones de generación renovable eólica y fotovoltaica y la disponibilidad de potencia nuclear, todas ellas publicadas antes del cierre del diario. El segundo es el bloque de mercado y coste, en el que destaca el precio del gas como variable exógena que aproxima el entorno de costes de las tecnologías térmicas. El tercero es el bloque meteorológico, con las variables agregadas de temperatura, viento, precipitación y humedad. El cuarto es el bloque de calendario, con el tipo de día y el porcentaje de población afectada por festivo.

A estos bloques se añaden dos tipos de variables construidas específicamente para la predicción. El primero está formado por los retardos del propio spread. Aunque no puede utilizarse el spread del día que se desea predecir, sí pueden incorporarse los valores de jornadas anteriores, que aportan información sobre su nivel reciente y su posible persistencia. En concreto, se incluye el perfil horario completo del día anterior y del mismo día de la semana previa, es decir, las veinticuatro horas de ambos retardos, junto con sus respectivas medias diarias. El retardo de un día recoge la persistencia a corto plazo, mientras que el retardo de siete días trata de capturar regularidades asociadas al ciclo semanal del sistema eléctrico. Este retardo semanal no se condiciona por la tipología del día de referencia: si el mismo día de la semana anterior fue, por ejemplo, un festivo, su perfil se incorpora igualmente sin ajuste adicional. El modelo dispone del tipo de día y del porcentaje de población festiva de ambas jornadas como variables independientes, por lo que puede en principio aprender a ponderar el retardo en consecuencia, pero esta corrección no está garantizada de forma explícita en la construcción de la variable.

El segundo tipo está formado por variables de estacionalidad anual construidas mediante funciones trigonométricas. El momento del año tiene una naturaleza cíclica: el 31 de diciembre y el 1 de enero son fechas consecutivas, aunque ocupen extremos opuestos en una codificación entera del día del año. Si esta variable se representase mediante un número comprendido entre 1 y 365, el modelo interpretaría ambas fechas como observaciones muy alejadas. Para evitar esta distorsión, la posición dentro del año se representa mediante un par de funciones seno y coseno,

$$\sin\left(\frac{2\pi\, d_{\text{año}}}{365{,}25}\right), \qquad \cos\left(\frac{2\pi\, d_{\text{año}}}{365{,}25}\right), \tag{4.3}$$

donde $d_{\text{año}}$ es el día del año. Esta codificación tiene la propiedad de que dos fechas próximas en el ciclo anual quedan próximas también en el espacio de las dos nuevas variables, lo que permite al modelo aprovechar la estacionalidad sin imponerle saltos artificiales. La hora del día, que también es cíclica, no se codifica de esta forma, sino que se incorpora directamente como una variable más en la representación apilada descrita en la sección 4.3.4; los modelos basados en árboles, que son los que se emplean, capturan bien su efecto sin necesidad de la transformación

trigonométrica. Junto a estas variables se añaden también el día de la semana y el mes como indicadores de calendario.

4.3.3. La información del clúster y la evaluación de su aportación

Una de las decisiones de diseño con mayor interés conceptual es la incorporación, como variable de entrada, de la pertenencia de cada día al clúster identificado en el capítulo 3. La motivación es que, si la agrupación captura regímenes del sistema asociados a perfiles de spread diferenciados, la pertenencia de una jornada a un determinado grupo puede contener información útil para anticipar su comportamiento horario. Para mantener la coherencia entre capítulos, la asignación no se recalcula en esta fase, sino que se reutiliza la partición obtenida en el análisis no supervisado y se incorpora la etiqueta correspondiente a cada día.

Conviene matizar dos puntos de coherencia temporal, de naturaleza distinta.

El primero es que la asignación de un día a un clúster se realiza a partir de variables del sistema disponibles antes del cierre del diario, de modo que conocer el clúster de un día no vulnera, en cuanto a las variables de entrada, el criterio anticipado. En una aplicación real, los centroides del modelo de clusterización quedarían fijados una vez entrenado el sistema, y la asignación de un día futuro a uno de esos centroides ya fijos no plantearía ningún problema de fuga de información, exactamente igual que ocurre con los modelos predictivos.

El segundo punto es más delicado y afecta, no a un futuro despliegue, sino a la evaluación retrospectiva presentada en este trabajo. El propio modelo de clusterización, es decir, los centroides que definen los ocho grupos, se ajustó una única vez sobre la totalidad de la muestra (2019–2025), y no se reajustó de forma incremental en cada etapa de la validación temporal, a diferencia de los modelos predictivos, que sí se reentrenan con el esquema de ventana creciente. Esto significa que, en sentido estricto, la etiqueta asignada a un día del *backtest*, por ejemplo uno de 2021, se determinó con un modelo que incorporó también información de años posteriores. Se trata, por tanto, de una fuga limitada al proceso de evaluación histórica, no al mecanismo de predicción en sí.

Dos factores acotan su impacto práctico. El primero es temporal: el efecto es mayor cuanto más alejado esté un día del final del periodo muestral, y casi inexistente para el conjunto de prueba final empleado en los resultados principales del capítulo 5, formado por los últimos 347 días de la muestra (julio de 2024 a

diciembre de 2025), muy próximos al límite superior del rango 2019–2025 usado para ajustar los centroides. Podría ser algo más relevante en el desglose anual de 2020 y 2021 de la sección 5.4.5, más alejados de ese límite. El segundo es que el análisis de interpretabilidad de la sección 5.3.4 muestra que la etiqueta de clúster es, de forma sistemática, la variable menos relevante para el modelo predictivo final, lo que limita cuánto de esa fuga puede realmente trasladarse a los resultados. Aun así, se documenta esta limitación con honestidad: una evaluación estrictamente ex-ante habría exigido reajustar los centroides de forma incremental en cada etapa, análoga a como se procede con los modelos predictivos. A ello se suma que la potencia instalada y el precio del gas, ambas variables con fuerte componente temporal, contribuyen a que la etiqueta de clúster arrastre implícitamente información sobre el régimen general del sistema en cada momento, cuestión ya señalada en el capítulo anterior.

El primero es que la asignación de un día a un clúster se realiza a partir de variables del sistema disponibles antes del cierre del diario, de modo que conocer el clúster de un día no vulnera, en cuanto a las variables de entrada, el criterio anticipado.

El segundo es más delicado y debe reconocerse explícitamente. El propio modelo de clusterización, es decir, los centroides que definen los ocho grupos, se ajustó una única vez sobre la totalidad de la muestra (2019-2025), y no se reajustó de forma incremental respecto a cada fecha, a diferencia de los modelos predictivos del presente capítulo, que sí se reentrenan con el esquema de ventana creciente. Esto significa que, en sentido estricto, la etiqueta asignada a un día concreto, por ejemplo uno de 2021, se determinó con un modelo que incorporó también información de años posteriores, y no únicamente información disponible en ese momento. Se trata de una fuga de información limitada al proceso de clusterización, que no debe confundirse con el protocolo de validación temporal de los modelos predictivos, correctamente aplicado en todo momento. Su impacto práctico sobre los resultados del capítulo 5 se considera reducido, ya que el análisis de interpretabilidad de la sección 5.3.4 muestra que la etiqueta de clúster es, de forma sistemática, la variable menos relevante para el modelo predictivo final. Aun así, se documenta esta limitación con honestidad: una aplicación estrictamente ex-ante de la clusterización exigiría reajustar los centroides de forma incremental, análoga a como se procede con los modelos predictivos. A ello se suma que la potencia instalada y el precio del gas, ambas variables con fuerte componente temporal, contribuyen a que la etiqueta de clúster arrastre implícitamente información sobre el régimen general del sistema en cada momento, cuestión ya señalada en el capítulo anterior.

Para aislar la aportación de esta variable, la pertenencia al clúster se incorpora

mediante codificación one-hot, es decir, a través de una variable binaria por grupo, por las razones que se detallan en la sección 4.3.5. Una vez incluida, se analiza su importancia dentro del modelo final en comparación con la del resto de variables. El objetivo es determinar si la etiqueta de clúster aporta información adicional una vez que el modelo ya dispone de las variables originales a partir de las cuales dicha etiqueta fue construida. El resultado de este contraste, que ayuda a delimitar el papel de la clusterización dentro del trabajo, se presenta en el capítulo de resultados.

4.3.4. Una sola hora por fila: la representación apilada

El spread es una variable horaria, por lo que cada día aporta veinticuatro observaciones, una por hora de entrega. Esta información puede organizarse de dos formas, y la elección entre ambas tiene implicaciones relevantes.

La primera opción consiste en tratar cada hora como un problema independiente y entrenar veinticuatro modelos distintos. Esta alternativa permite que cada modelo se especialice en el comportamiento de una hora concreta, pero presenta varios inconvenientes. Por una parte, multiplica el número de modelos que deben entrenarse, ajustarse y mantenerse. Por otra, reduce de forma considerable el número de observaciones disponibles para cada modelo, ya que cada uno solo recibe una observación diaria. Además, impide aprovechar de forma conjunta las regularidades compartidas por distintas horas.

La segunda opción, adoptada en este trabajo, consiste en apilar todas las horas en una única tabla. Cada fila corresponde a una hora concreta de un día determinado, y la hora se incorpora como una variable adicional mediante un índice entero comprendido entre 0 y 23. Las veinticuatro filas de una misma jornada comparten su información diaria, como el perfil completo de previsiones y los retardos del spread, y se diferencian por el valor de la hora. De este modo, se entrena un único modelo con todas las horas y todos los días.

Esta representación apilada maximiza el número de observaciones disponibles, permite que el modelo aprenda la evolución intradiaria del spread y simplifica la gestión del proceso de entrenamiento. Como contrapartida, la dependencia horaria no queda impuesta por construcción, sino que debe aprenderse a partir de la variable de hora. Este coste resulta asumible, dado que los modelos no lineales empleados son adecuados para capturar este tipo de relaciones.

4.3.5. Codificación de las variables y escalado

Antes de entrenar es necesario decidir cómo se representan numéricamente las variables categóricas y si conviene reescalar las continuas. Como ya se argumentó para el tipo de día en el capítulo 3, introducir una variable categórica como un número entero impondría un orden artificial inexistente; el problema se reproduce ahora con la etiqueta de clúster, ya que codificarla como $0, 1, 2, \ldots$ sugeriría que el clúster 1 está entre el 0 y el 2 en algún sentido, lo cual carece de significado. Por ello, tanto el tipo de día como la etiqueta de clúster se transforman mediante codificación one-hot, que crea una variable binaria por categoría y evita cualquier orden espurio entre ellas.

El tratamiento de las variables continuas depende del tipo de modelo. Los modelos lineales son sensibles a la escala, ya que sus coeficientes operan sobre variables que pueden estar expresadas en unidades muy distintas, como megavatios o euros. En esos casos, cada variable continua se estandariza restando su media y dividiendo por su desviación típica. Los modelos basados en árboles, en cambio, no requieren escalado, pues sus decisiones se construyen mediante puntos de corte que se adaptan a la escala de cada variable. Dado que el modelo principal del trabajo pertenece a esta segunda familia, la estandarización solo se aplica cuando resulta necesaria.

4.4. Protocolo de evaluación

Antes de presentar los modelos es necesario establecer cómo se evaluará su rendimiento. La elección de las métricas condiciona la interpretación de los resultados, ya que una misma predicción puede valorarse de forma distinta según se atienda al error numérico, a la capacidad direccional o a la utilidad económica. Por ello, el protocolo de evaluación se define antes de introducir los modelos, de forma que el significado de cada resultado quede fijado de antemano.

4.4.1. Partición temporal frente a aleatoria

La evaluación de un modelo requiere reservar una parte de los datos para validación y utilizar el resto para el entrenamiento. Esta separación temporal se respeta de forma estricta en el ajuste y la evaluación de los modelos predictivos; la única excepción documentada es la etiqueta de clúster, cuya construcción no siguió

este mismo esquema incremental, como se explicó en la sección 4.3.3. La forma de realizar esta separación depende de la estructura del problema. En contextos sin dependencia temporal puede recurrirse a una partición aleatoria. En este caso, sin embargo, dicha estrategia introduciría información futura de manera indirecta y produciría una evaluación poco realista.

Si la separación entre entrenamiento y validación se realizase al azar, el conjunto de entrenamiento podría contener días posteriores a algunas observaciones de validación. El modelo utilizaría así información futura para predecir el pasado, una situación incompatible con la operación real, en la que solo se dispone de datos anteriores al momento de la decisión. Los retardos del spread refuerzan este problema, ya que el spread de una jornada actúa como objetivo de ese día y como variable explicativa de jornadas posteriores. Una partición aleatoria podría generar, por tanto, dependencias entre ambos subconjuntos y producir estimaciones de rendimiento artificialmente elevadas, que se deteriorarían al aplicar el modelo a datos realmente futuros.

A esta consideración se añade la naturaleza no estacionaria del mercado eléctrico. A lo largo del periodo analizado se observan regímenes distintos, entre ellos la etapa de precios moderados anterior a la crisis energética, la crisis del gas de 2022 y el periodo posterior de expansión renovable y normalización del precio del gas. Estos regímenes fueron identificados en la clusterización del capítulo anterior. Una partición cronológica obliga al modelo a generalizar entre periodos con características diferentes, como ocurriría en una aplicación real. Una partición aleatoria mezclaría estos regímenes en entrenamiento y validación y ocultaría parte de la dificultad del problema. Por ello, la separación mantiene siempre el orden temporal.

4.4.2. Validación con ventana creciente

Reservar el último tramo del periodo como único conjunto de validación sería una opción válida, pero proporcionaría una sola medida del rendimiento y dificultaría el análisis de su estabilidad. Por ello, se adopta una estrategia de validación con ventana creciente, también denominada *expanding window* [22, 23], que reproduce, que reproduce con mayor fidelidad un escenario de despliegue con reentrenamientos periódicos.

El procedimiento se desarrolla por etapas sucesivas. En la primera, el modelo se entrena con un tramo inicial de días y se valida sobre el periodo inmediatamente posterior. A continuación, el tramo utilizado para la validación se incorpora al conjunto de entrenamiento, el modelo se vuelve a ajustar con la muestra ampliada

y se evalúa sobre el siguiente periodo. Este proceso se repite hasta agotar los datos disponibles. Así, cada observación se predice mediante un modelo entrenado únicamente con información anterior y, al mismo tiempo, se obtienen varias medidas de rendimiento que permiten analizar su estabilidad temporal. La Figura 4.1 ilustra este esquema.

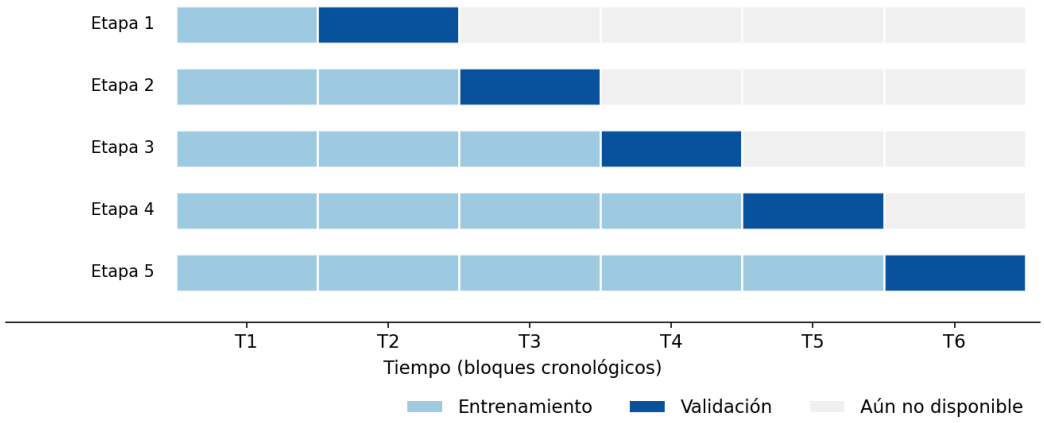

Figura 4.1: Esquema de validación con ventana creciente. En cada etapa el modelo se entrena con todos los días anteriores (tramo claro) y se valida sobre el tramo siguiente (tramo oscuro), que después se incorpora al entrenamiento de la etapa posterior. La separación respeta siempre el orden cronológico, de modo que ningún día se predice con información futura.

Este esquema cumple una doble función. Por un lado, es el procedimiento con el que se evalúan los modelos finales sobre el conjunto de prueba. Por otro, es también el que se emplea durante la búsqueda de hiperparámetros descrita más adelante, de modo que en ningún momento del proceso, ni en el ajuste ni en la evaluación, se filtra información del futuro hacia el pasado.

4.4.3. Métricas de error de la regresión

Para el modelo de regresión, que predice el valor del spread, se emplean tres métricas de error clásicas. La primera es el error absoluto medio, o MAE, que promedia el valor absoluto de la diferencia entre el spread real y el predicho,

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n} |s_i - \hat{s}_i| \,. \tag{4.4}$$

El MAE tiene la ventaja de ser fácil de interpretar, porque se expresa en las mismas unidades que el spread, esto es, en euros por megavatio hora, y de penalizar todos los errores de forma proporcional a su tamaño.

La segunda es la raíz del error cuadrático medio, o RMSE,

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(s_i - \hat{s}_i)^2}. \tag{4.5}$$

A diferencia del MAE, el RMSE eleva los errores al cuadrado antes de promediarlos, lo que hace que penalice mucho más los errores grandes. En un problema con colas pesadas como este, en el que el spread puede tomar ocasionalmente valores muy alejados de cero, el RMSE será sistemáticamente mayor que el MAE, y la diferencia entre ambos informa de la presencia de errores extremos.

La tercera es el coeficiente de determinación, R^2, que mide qué proporción de la variabilidad del spread consigue explicar el modelo respecto a la que quedaría sin explicar si simplemente se predijese siempre la media,

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(s_i - \hat{s}_i)^2}{\sum_{i=1}^{n}(s_i - \bar{s})^2}. \tag{4.6}$$

Un R^2 igual a uno indicaría una predicción perfecta, mientras que un valor igual a cero señalaría que el modelo no mejora la predicción constante basada en la media. En la interpretación de los resultados debe tenerse en cuenta que un R^2 reducido no implica necesariamente que el procedimiento de modelización sea inadecuado. En un fenómeno tan ruidoso como el spread entre mercados, una parte relevante de la variabilidad puede no ser anticipable a partir de la información ex-ante disponible. Por tanto, el nivel alcanzable por cualquier modelo está condicionado por la propia dificultad del problema.

4.4.4. Métricas de la clasificación del signo

Para el modelo de clasificación, la métrica más inmediata es la exactitud o *accuracy*, definida como la proporción de horas en las que el signo predicho coincide con el observado. Sin embargo, esta medida puede resultar engañosa cuando las clases están desequilibradas. Dado que el spread es negativo con algo más de frecuencia que positivo, una regla que predijese siempre la clase negativa alcanzaría una exactitud apreciable sin utilizar información específica de cada observación. Por tanto, la exactitud no permite distinguir por sí sola entre una capacidad discriminativa real y el aprovechamiento del desequilibrio de base.

Por ello, se emplean tres métricas menos sensibles, en distinto grado, al desequilibrio entre clases. La primera es la exactitud balanceada, que calcula el acierto dentro de cada clase por separado y promedia después ambos valores, de modo que la clase minoritaria tiene el mismo peso que la mayoritaria. La segunda es la medida F1, que combina precisión y exhaustividad en un único indicador y resulta especialmente útil cuando interesa evaluar el comportamiento sobre una clase que no domina la muestra. La tercera, utilizada como principal medida de habilidad direccional, es el área bajo la curva ROC, o AUC. Esta métrica evalúa la capacidad del modelo para ordenar las observaciones según su probabilidad de pertenecer a la clase positiva, sin depender de un umbral de decisión concreto. Un AUC de 0,5 corresponde a un modelo sin capacidad discriminativa, equivalente a una ordenación aleatoria, mientras que valores superiores indican una separación creciente entre las dos clases. Su independencia respecto al umbral y su menor sensibilidad a la proporción de cada clase justifican su uso como métrica de referencia para valorar la habilidad direccional del clasificador.

4.4.5. La métrica económica: beneficio acumulado y tasa de acierto

Las métricas anteriores describen el rendimiento estadístico, pero no reflejan de forma directa su valor económico. Dos modelos con la misma exactitud pueden presentar una utilidad operativa diferente si uno acierta principalmente en horas con spreads de gran magnitud y el otro en observaciones próximas a cero. Por ello, se incorpora una métrica económica basada en la simulación de la estrategia.

La simulación se construye del siguiente modo. Para cada hora, el modelo genera una señal direccional: comprar en el mercado diario y vender en el intradiario, o adoptar la posición contraria. Si la dirección de la señal coincide con el signo realizado del spread, la operación genera un resultado positivo; si no coincide, genera una pérdida de la misma magnitud absoluta. La suma de estos resultados sobre todas las horas operadas proporciona el beneficio acumulado para una posición normalizada de 1 MWh en cada operación. En adelante, esta magnitud se denomina PnL, por sus siglas en inglés. Junto al valor acumulado se presenta el beneficio medio por operación, calculado como el PnL dividido por el número de operaciones y expresado en euros por megavatio hora. Esta medida aproxima la rentabilidad unitaria de la estrategia y permite compararla con un posible coste por operación. Por último, la tasa de acierto, o *hit rate*, recoge la proporción de operaciones que finalizan con un resultado positivo.

Conviene precisar dos cuestiones. En primer lugar, el PnL puede aplicarse tanto

al modelo de regresión como al de clasificación, ya que ambos terminan generando una decisión de compra o venta. Por ello, constituye la medida común para comparar las dos formulaciones. En segundo lugar, la simulación básica no incorpora inicialmente los costes de transacción. Esta simplificación permite estimar el potencial bruto de la estrategia y se relaja en el capítulo de resultados, donde se introduce un coste por operación y se analiza cómo varían la rentabilidad y el número de operaciones al exigir distintos umbrales de confianza.

4.5. Modelos de referencia y modelo propuesto

Antes de presentar la técnica de aprendizaje conviene fijar las reglas sencillas frente a las que se medirá. Un modelo solo demuestra utilidad si supera referencias que no requieren aprendizaje alguno, y la elección de esas referencias determina la exigencia de la comparación.

4.5.1. Las referencias sencillas

Para el problema direccional se emplean tres reglas de referencia. La primera es la **regla aleatoria**, que asigna a cada hora un signo al azar con igual probabilidad. Su papel es únicamente delimitar el suelo del problema: cualquier modelo con información útil debe superarla con claridad.

La segunda es la **clase mayoritaria**, que predice para todas las horas el signo más frecuente del spread en el conjunto de entrenamiento. Es una referencia más exigente de lo que parece: como el spread presenta un sesgo estructural negativo, acertar por defecto la dirección dominante permite capturar una parte apreciable del beneficio sin predecir nada específico de cada hora. La comparación con esta regla es la que permite aislar la aportación del modelo más allá de ese sesgo.

La tercera es la **persistencia** del día anterior, que utiliza como predicción el comportamiento reciente del propio spread. En el problema direccional se emplea el signo del spread del día anterior en la misma hora; en el problema de regresión, el perfil horario completo de la misma jornada de la semana anterior. Esta referencia mide si el modelo aporta algo más que la inercia de la serie, ya recogida en los retardos.

En la regresión se añade una cuarta referencia con interés propio: el **perfil medio del clúster**, consistente en predecir para cada día el spread horario medio

histórico de su clúster de pertenencia. Esta regla conecta directamente con el capítulo 3 y permite cuantificar cuánta capacidad predictiva aporta, por sí sola, la agrupación de días, con la matización sobre la temporalidad de la clusterización ya señalada en la sección 4.3.3.

4.5.2. El *gradient boosting*

El modelo principal del trabajo es el *gradient boosting* el *gradient boosting* basado en árboles de decisión [24], en su implementación LightGBM [25]. Un árbol de decisión genera predicciones mediante una sucesión de divisiones sobre las variables de entrada. Por ejemplo, si la previsión de generación eólica supera un determinado umbral, la observación se asigna a una rama; en caso contrario, se dirige a la rama alternativa. El proceso se repite dentro de cada subconjunto hasta alcanzar las hojas terminales, en las que se asigna una predicción. Conviene señalar que, al igual que en la clusterización del capítulo 3, estos umbrales se aprenden sobre magnitudes absolutas de generación renovable; su generalización a periodos futuros con un mix de generación sustancialmente distinto no está garantizada, cuestión que se retoma en las limitaciones del capítulo 5. Un árbol individual es interpretable, pero suele presentar una capacidad limitada cuando las relaciones entre variables son complejas.

El *boosting* combina numerosos árboles sencillos para construir un modelo más flexible. Los árboles se añaden de forma secuencial, de manera que cada nuevo árbol trata de corregir los errores residuales del conjunto anterior. La predicción final se obtiene agregando las contribuciones de todos ellos [26]. El término *gradient* hace referencia a que cada etapa se orienta en la dirección que reduce la función de pérdida, de forma análoga a un procedimiento de descenso por gradiente. Así, la complejidad aumenta progresivamente y puede controlarse mediante regularización, lo que permite capturar relaciones no lineales limitando el riesgo de sobreajuste.

Esta técnica resulta adecuada para el problema por tres motivos. En primer lugar, permite representar relaciones no lineales sin especificar previamente su forma. La relación entre la previsión renovable y el spread, por ejemplo, no tiene por qué seguir un patrón lineal. En segundo lugar, captura interacciones entre variables, como aquellas situaciones en las que el efecto de una previsión renovable elevada depende del nivel de demanda. En tercer lugar, admite variables de naturaleza heterogénea, incluidas previsiones continuas, indicadores binarios de calendario y etiquetas de clúster, sin requerir escalado ni transformaciones complejas.

4.5.3. Un mismo modelo para los dos problemas

Una decisión deliberada del trabajo es emplear la misma técnica, el *gradient boosting*, tanto para el problema de regresión como para el de clasificación. Existe una versión de regresión, que predice el valor del spread, y una versión de clasificación, que predice su signo, pero ambas comparten la misma familia de modelos, la misma matriz de características y el mismo protocolo de evaluación. La única diferencia relevante entre ambas es la función de pérdida que cada una minimiza, adaptada a la naturaleza de su objetivo.

Esta simetría responde a un propósito metodológico. Para comparar la formulación de regresión con la de clasificación, conviene mantener constantes la familia de modelos, las variables de entrada y el procedimiento de evaluación. De este modo, las diferencias observadas pueden atribuirse principalmente a la definición del objetivo y no a cambios en otros elementos del diseño. La comparación se aproxima así a un experimento controlado sobre las dos formas de plantear el problema.

En la versión de regresión debe considerarse, además, la función de pérdida. Dado que el spread presenta colas pesadas, una pérdida puramente cuadrática otorgaría un peso excesivo a un número reducido de valores extremos, en detrimento del comportamiento en la mayor parte de las horas. Por ello, se emplea una pérdida de tipo Huber [27], que se comporta de forma cuadrática para los errores pequeños y de forma lineal para los errores grandes. Esta combinación aporta robustez frente a los valores atípicos sin perder sensibilidad en la región central de la distribución.

4.6. Ajuste de los hiperparámetros

Los modelos de aprendizaje dependen de un conjunto de parámetros que no se estiman directamente durante el entrenamiento, sino que deben definirse antes del ajuste. Estos parámetros, denominados hiperparámetros, incluyen en el caso del *gradient boosting*, entre otros, el número de árboles, el número de hojas de cada árbol, la tasa de aprendizaje, la fracción de observaciones y variables utilizada en cada iteración, los términos de regularización y el parámetro que determina la transición de la pérdida de Huber. Su elección puede influir de forma apreciable en el rendimiento, por lo que se determina mediante un procedimiento sistemático en lugar de fijarse de forma arbitraria.

4.6.1. Búsqueda con validación temporal

Para buscar los hiperparámetros se ha empleado una búsqueda bayesiana, concretamente el algoritmo de estimación basada en árboles de Parzen [28], implementado en la librería Optuna [29]. A diferencia de una búsqueda exhaustiva, que evaluaría todas las combinaciones de una rejilla, o de una búsqueda puramente aleatoria, el procedimiento bayesiano utiliza los resultados de las pruebas anteriores para estimar qué regiones del espacio de hiperparámetros son más prometedoras. Las evaluaciones posteriores se concentran en esas regiones, lo que permite encontrar configuraciones adecuadas con un número menor de intentos. Esta eficiencia resulta relevante porque cada evaluación exige entrenar y validar un modelo completo.

La evaluación de cada combinación debe respetar la estructura temporal del problema. Una partición aleatoria reintroduciría la fuga de información y seleccionaría hiperparámetros adaptados a un escenario irreal. Por ello, cada configuración se evalúa mediante el mismo esquema de ventana creciente descrito en la sección 4.4.2. La coherencia entre el procedimiento de ajuste y la evaluación final permite seleccionar los hiperparámetros por su comportamiento en condiciones comparables a las de una aplicación real.

La Figura 4.2 muestra la evolución de la búsqueda para el regresor: el valor del objetivo (el PnL medio por hora obtenido en la validación de ventana creciente) de cada una de las 600 configuraciones probadas, junto con el mejor valor acumulado. La mejora se concentra en las primeras 150 pruebas, tras las cuales el mejor valor apenas avanza: la búsqueda explora después variaciones alrededor de las regiones prometedoras sin encontrar configuraciones sustancialmente mejores, lo que indica que el número de pruebas fue suficiente. La dispersión vertical, que no desaparece a medida que avanza la búsqueda, ilustra hasta qué punto el rendimiento depende de la combinación concreta de hiperparámetros.

4.6.2. Qué se optimiza: la elección del objetivo de búsqueda

La métrica utilizada como objetivo de la búsqueda constituye una decisión metodológica relevante. Aunque podría considerarse un detalle técnico, las pruebas iniciales mostraron que su elección modifica de forma sustancial el comportamiento operativo del modelo.

La opción aparentemente natural para el modelo de regresión consiste en mi-

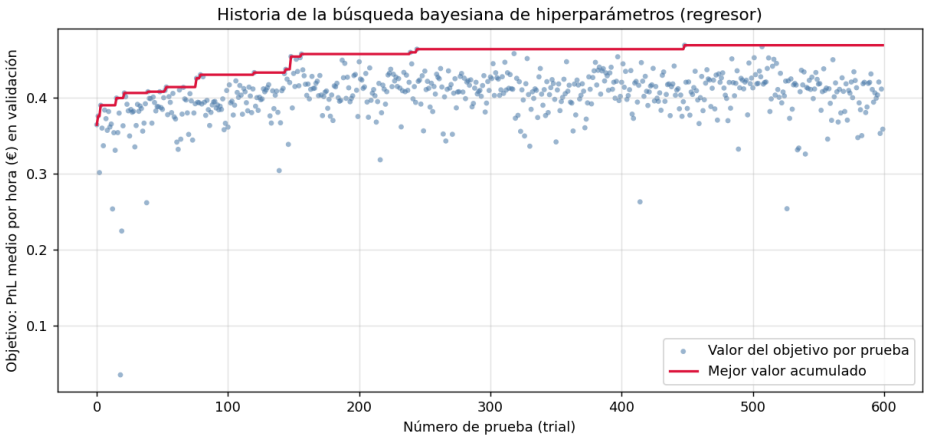

Figura 4.2: Historia de la búsqueda bayesiana de hiperparámetros del regresor. Cada punto es una configuración evaluada mediante validación temporal de ventana creciente; la línea indica el mejor valor encontrado hasta cada momento.

nimizar el error, medido mediante el MAE. Sin embargo, un primer ajuste con este criterio produjo una configuración cuyas predicciones quedaban fuertemente contraídas hacia la media del spread y apenas se alejaban de cero. Este comportamiento es coherente con la naturaleza del problema. Cuando la variable objetivo contiene una proporción elevada de ruido difícilmente anticipable, una pérdida simétrica como el MAE tiende a favorecer predicciones conservadoras, cercanas al centro de la distribución. Ante la dificultad de anticipar desviaciones concretas, el modelo reduce el error medio evitando estimaciones alejadas de cero.

El modelo optimizado por MAE obtenía una mejora marginal en esta métrica, pero perdía utilidad operativa: el beneficio acumulado descendía desde aproximadamente 6 376 € hasta 3 244 € (posiciones de 1 MWh por operación) respecto a una configuración de partida orientada a la utilidad. Expresado por operación, el beneficio medio pasaba de 1,32 €/MWh (4 826 operaciones) a 2,17 €/MWh (1 492 operaciones): la configuración ajustada por MAE llega a acertar, cuando opera, con un margen unitario mayor, pero opera con tan poca frecuencia que el resultado acumulado se reduce a menos de la mitad.[1] Como muestra la Figura 4.3, la reducción del MAE es muy pequeña, mientras que el beneficio acumulado disminuye aproximadamente a la mitad y el número de operaciones se reduce a menos de un tercio, debido a la menor intensidad de las señales generadas.

[1]Estas cifras proceden del experimento preliminar de la fase de búsqueda en el que se detectó el problema, por lo que no coinciden con las del modelo final del capítulo 5. La conclusión que motivó la decisión es, no obstante, la misma: reducir el error no implica aumentar la utilidad.

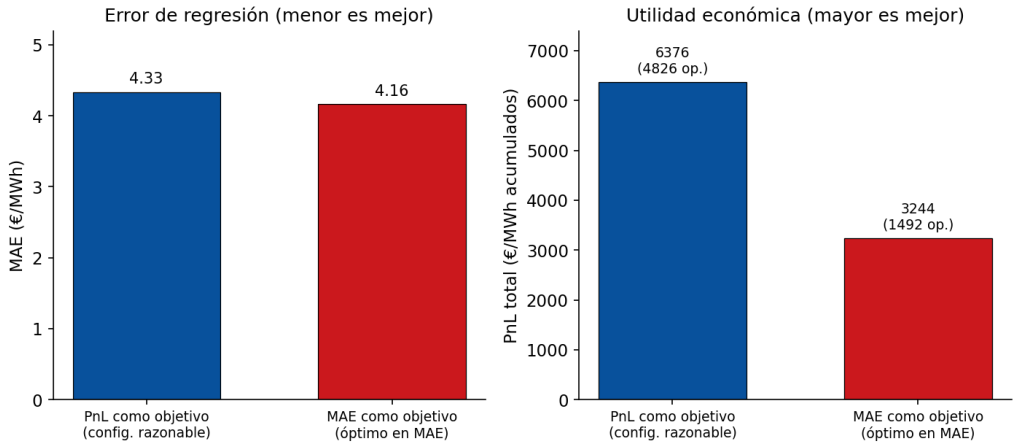

Figura 4.3: Efecto de la métrica utilizada en la búsqueda. Al optimizar el MAE (rojo) frente a una configuración orientada a la utilidad (azul), el error de regresión apenas mejora (panel izquierdo), mientras que el beneficio acumulado y el número de operaciones disminuyen de forma considerable (panel derecho). Este comportamiento es consistente con un predictor cuyas estimaciones se concentran en torno a la media.

La conclusión es que minimizar el error estadístico no equivale necesariamente a maximizar la utilidad económica. Un modelo puede obtener un error ligeramente menor y, al mismo tiempo, resultar menos adecuado para la decisión operativa si la métrica de ajuste no refleja el objetivo económico. Por este motivo, en el regresor se opta por optimizar directamente el PnL. Este criterio favorece configuraciones capaces de generar señales operativas cuando la información disponible lo justifica, en lugar de concentrar sistemáticamente las predicciones en torno a la media. En el clasificador, la búsqueda se orienta mediante el AUC, por las razones expuestas en la sección 4.4.4, ya que esta métrica evalúa directamente la capacidad discriminativa del modelo.

4.6.3. Qué hiperparámetros importan

Una vez completada la búsqueda, además de identificar la mejor configuración, resulta útil determinar qué hiperparámetros han influido en mayor medida sobre el objetivo. Para ello, se estima su importancia mediante un análisis funcional de la varianza, conocido como fANOVA [30]. Este procedimiento cuantifica qué proporción de la variación del rendimiento entre las configuraciones evaluadas puede atribuirse a cada hiperparámetro. La Figura 4.4 presenta los resultados para el re-

gresor. El clasificador se analizó de forma análoga. El Anexo B recoge el significado de cada hiperparámetro, sus implicaciones sobre el comportamiento del modelo y el valor final seleccionado en cada caso.

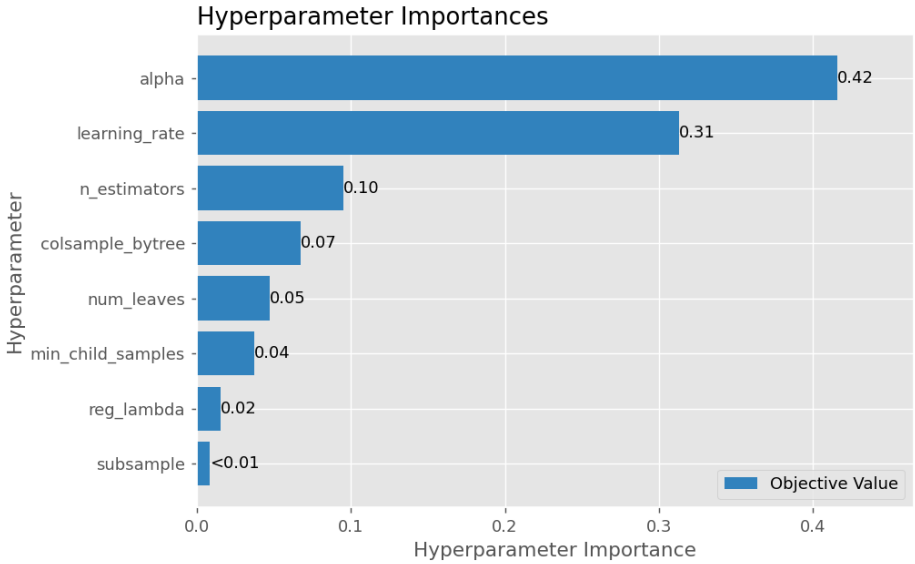

Figura 4.4: Importancia relativa de los hiperparámetros en la búsqueda con Optuna del regresor, calculada mediante fANOVA. Cada barra indica qué fracción de la variación del objetivo entre las configuraciones probadas se atribuye al hiperparámetro correspondiente.

Este análisis tiene una doble utilidad práctica. Por una parte, permite comprobar si el proceso de ajuste estaba justificado: si uno o varios hiperparámetros explican una parte relevante de la variación del rendimiento, su elección no resulta indiferente. Por otra, orienta futuros reentrenamientos, ya que permite identificar qué hiperparámetros conviene volver a ajustar con mayor atención y cuáles podrían mantenerse sin un coste apreciable de rendimiento.

4.7. Recapitulación de las configuraciones evaluadas

Antes de pasar a los resultados, conviene resumir las configuraciones que se comparan en el capítulo siguiente. Todos los modelos se evalúan respetando el mismo orden temporal. Aunque algunos predicen el valor del spread y otros su

signo, sus resultados se convierten finalmente en una decisión común de compra, venta o no operación.

En primer lugar, se consideran las referencias sencillas descritas en la sección 4.5.1: la regla aleatoria, la clase mayoritaria, la persistencia y, en el caso de la regresión, el perfil medio del clúster. En segundo lugar, se evalúa un modelo de *gradient boosting* de regresión. Sus hiperparámetros se seleccionan con Optuna, respetando la estructura temporal de los datos y utilizando el PnL como criterio de optimización. La predicción continua obtenida por este modelo se transforma después en una decisión de signo. En tercer lugar, se evalúa un modelo de *gradient boosting* de clasificación. Sus hiperparámetros también se seleccionan mediante validación temporal, pero en este caso se utiliza como criterio una métrica que evalúa la capacidad del modelo para distinguir correctamente entre spreads positivos y no positivos. De este modo, la comparación entre regresión y clasificación se realiza bajo un procedimiento de evaluación equivalente.

En ambos modelos, la información del clúster se incorpora como una variable adicional. Su aportación se analiza en el capítulo de resultados una vez considerado el efecto del resto de variables del sistema.

Esta recapitulación cierra el planteamiento metodológico. El capítulo siguiente presenta los resultados de estas configuraciones, analiza las variables en las que se apoyan los modelos, estudia la robustez de las conclusiones y discute hasta qué punto el spread entre mercados puede anticiparse a partir de información disponible antes del cierre del mercado diario.

Capítulo 5

Resultados y análisis de la predicción del spread

5.1. Introducción

Una vez definidos el planteamiento predictivo y el procedimiento de evaluación, este capítulo presenta los resultados obtenidos por las distintas configuraciones sobre un mismo conjunto de prueba temporal. En primer lugar, se comparan los modelos y las reglas de referencia mediante métricas estadísticas y económicas. Después se analiza en qué variables se apoyan las predicciones. A continuación, se estudia el efecto de los costes, se contrasta la significación estadística de la mejora y se examina la evolución de los resultados entre años. Por último, se realiza una valoración conjunta y se exponen las principales limitaciones y posibles líneas de extensión.

Todas las cifras del capítulo son estimaciones fuera de muestra: cada observación se predice con un modelo entrenado únicamente con información anterior. Conviene, no obstante, distinguir los dos marcos de evaluación empleados. La comparación entre modelos, el análisis de costes, las pruebas de significación y la comparación con el máximo teórico se calculan sobre el conjunto de prueba final, formado por el último 20 % de la muestra: 347 jornadas comprendidas entre julio de 2024 y diciembre de 2025 (8 328 horas), que no se utilizaron ni en el entrenamiento ni en la búsqueda de hiperparámetros. El análisis de estabilidad temporal de la sección 5.4.5, en cambio, necesita cubrir un periodo más largo, por lo que emplea las predicciones fuera de muestra de las sucesivas etapas de la validación con ventana creciente, que abarcan desde mediados de 2020 hasta diciembre de

2025. Por este motivo, las cifras de esa sección no son directamente comparables con las del resto del capítulo, aunque en ambos marcos se respeta siempre el orden temporal. En cualquier caso, los resultados no reproducen por completo las condiciones de una aplicación real.

5.2. Comparación de los modelos bajo un criterio común

Para comenzar, todos los modelos se evalúan mediante una decisión direccional para cada hora: comprar en el mercado diario y vender en el intradiario, o adoptar la posición contraria. En el caso del regresor, la predicción continua se transforma en una decisión a partir de su signo. De este modo, sus resultados pueden compararse directamente con los del clasificador y con las reglas de referencia. La Tabla 5.1 resume las métricas obtenidas.

Tabla 5.1: Comparación de los modelos a partir de una decisión de signo en todas las horas del conjunto de prueba. La exactitud (acc), la exactitud balanceada (bal. acc), la medida F1 y la tasa de acierto (hit) se expresan en tanto por uno. El PnL total representa el beneficio unitario acumulado para una posición de 1 MWh en cada operación.

Modelo	Acc	Bal. acc	F1	AUC	PnL total	Hit
Regresor (Optuna → signo)	0,594	**0,561**	**0,561**	**0,590**	7.022	0,584
Clasificador (Optuna)	0,616	0,523	0,445	0,584	**7.676**	0,575
Persistencia (día −1)	0,561	0,541	0,541	0,565	5.146	0,555
Clase mayoritaria	0,604	0,500	0,377	0,500	5.662	0,557
Aleatorio	0,509	0,508	0,503	0,500	1.230	0,510

Las métricas recogen aspectos distintos del rendimiento, por lo que deben interpretarse de forma conjunta. La Figura 5.1 presenta gráficamente los mismos resultados.

La exactitud muestra con claridad el efecto del desequilibrio entre clases. La regla de clase mayoritaria, que predice siempre el signo dominante del spread, alcanza un valor de 0,604, superior al del regresor (0,594) y próximo al del clasificador (0,616). Si se atendiera únicamente a esta métrica, la mejora de los modelos frente a una referencia sencilla parecería muy reducida. Este resultado confirma lo señalado en la sección 4.4.4: cuando las clases están desequilibradas, la exactitud puede favorecer a los modelos que reproducen el sesgo de la variable objetivo sin

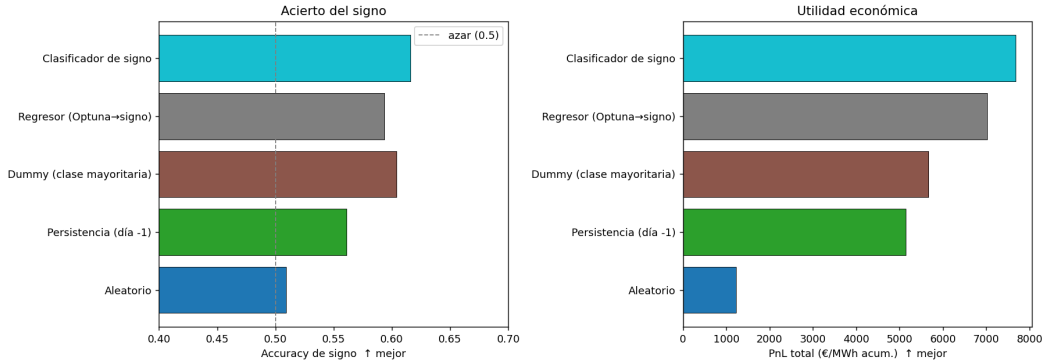

Figura 5.1: Comparación de los cinco modelos mediante las distintas métricas, una vez transformadas sus predicciones en decisiones de signo sobre el conjunto de prueba.

distinguir adecuadamente entre ambas clases.

La exactitud balanceada y la medida F1 son menos sensibles a este desequilibrio y ofrecen una lectura diferente. La regla de clase mayoritaria obtiene una exactitud balanceada de 0,500 y una F1 de 0,377, ya que concentra sus aciertos en una sola clase. El clasificador alcanza valores de 0,523 y 0,445, respectivamente, lo que indica que una parte importante de su rendimiento procede también de favorecer la clase dominante. El regresor presenta los mejores resultados en ambas métricas, con un valor de 0,561. La columna AUC confirma esta lectura: el regresor alcanza 0,590 y el clasificador 0,584, frente al 0,565 de la persistencia y el 0,5 de una ordenación sin información. Ambos modelos superan el nivel del azar, aunque los valores siguen reflejando una habilidad predictiva limitada (véase el Anexo C para la definición formal de estas métricas).

La comparación económica favorece al clasificador, que obtiene un PnL acumulado de 7,676, frente a los 7,022 del regresor. Sin embargo, la regla de clase mayoritaria alcanza 5,662, aproximadamente el 74 % del valor obtenido por el clasificador. Por tanto, una parte considerable del beneficio procede del sesgo estructural negativo del spread y puede capturarse sin realizar una predicción específica para cada hora. La mayor diferencia entre el PnL y las métricas balanceadas del clasificador sugiere que su ventaja económica está relacionada con una mayor tendencia a predecir la clase dominante. El regresor, en cambio, distribuye de forma más equilibrada sus aciertos entre las dos direcciones.

Esta aportación puede cuantificarse con exactitud. La ventaja del regresor sobre la clase mayoritaria procede, por construcción, de las horas en las que ambos

discrepan: aquellas en las que el modelo predice signo positivo. En el conjunto de prueba, el modelo adopta la posición positiva en 2.748 horas (el 33 % del total). En ellas, el spread realizado es positivo en el 48 % de los casos y nulo en el 10 %, con un valor medio de +0,25 €/MWh, de modo que estas horas aportan +680 € al modelo cuando a la regla mayoritaria le habrían costado −680 €: exactamente la diferencia de 1,360 € estimada en la sección 5.4.2. La lectura es reveladora: las señales positivas aciertan poco más de la mitad de las veces (un 54 % entre las horas con spread no nulo), pero cuando aciertan lo hacen con spreads de mayor magnitud. El valor del modelo no reside en acertar mucho, sino en identificar horas en las que apostar contra el sesgo estructural tiene esperanza positiva.

La comparación no permite identificar un modelo superior en todos los criterios. El clasificador obtiene el mayor beneficio acumulado, mientras que el regresor presenta una mejor capacidad para distinguir entre las dos direcciones. En ambos casos, la mejora respecto a las referencias es moderada y una parte relevante del resultado económico se explica por el sesgo de la variable objetivo. Estas diferencias justifican analizar por separado la capacidad predictiva y la utilidad económica.

También resulta útil comparar estos resultados con la referencia obtenida mediante la clusterización del capítulo 3. Al evaluar la predicción continua del spread, el perfil medio de cada clúster alcanza un coeficiente de determinación de 0,003 y un beneficio medio por operación de 1,29 €/MWh. El modelo de *gradient boosting* obtiene valores de 0,018 y 1,34 €/MWh, respectivamente. La persistencia semanal, que utiliza como predicción el perfil de la misma jornada de la semana anterior, presenta un coeficiente de determinación claramente negativo y queda por detrás de todas las configuraciones.. La Figura 5.2 resume esta comparación.

La proximidad entre el perfil medio de los clústeres y el modelo no lineal indica que la agrupación realizada en el capítulo anterior recoge parte de la estructura recurrente del spread. Al mismo tiempo, muestra que la mejora aportada por un modelo con un número mucho mayor de variables es reducida. Este resultado respalda la utilidad descriptiva de la clusterización, aunque no implica que la etiqueta del clúster aporte por sí sola información adicional al modelo predictivo.

El coeficiente de determinación se mantiene próximo a cero en todas las configuraciones. Esto indica que las variables y los modelos considerados explican una fracción muy pequeña de la variabilidad horaria del spread. Como se señaló en la sección 4.4.3, esta métrica debe interpretarse junto con el resto de los resultados, ya que una capacidad limitada para estimar la magnitud exacta no impide que el modelo pueda identificar parte de la dirección o del perfil medio. La Figura 5.3 muestra que el regresor reproduce la forma media del spread a lo largo del día, aunque no capture gran parte de las desviaciones de cada hora.

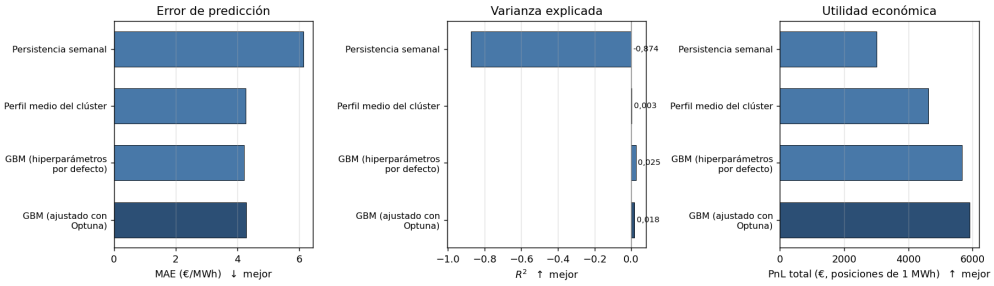

Figura 5.2: Comparación de las configuraciones de regresión: persistenci semanal (perfil del mismo día de la semana anterior), perfil medio de clúster y *gradient boosting* con hiperparámetros por defecto ajustado con Optuna. Se muestran el error absoluto medio, el coeficiente d determinación y el beneficio acumulado. El rendimiento del perfil medio de clúster es próximo al de los modelos de aprendizaje.

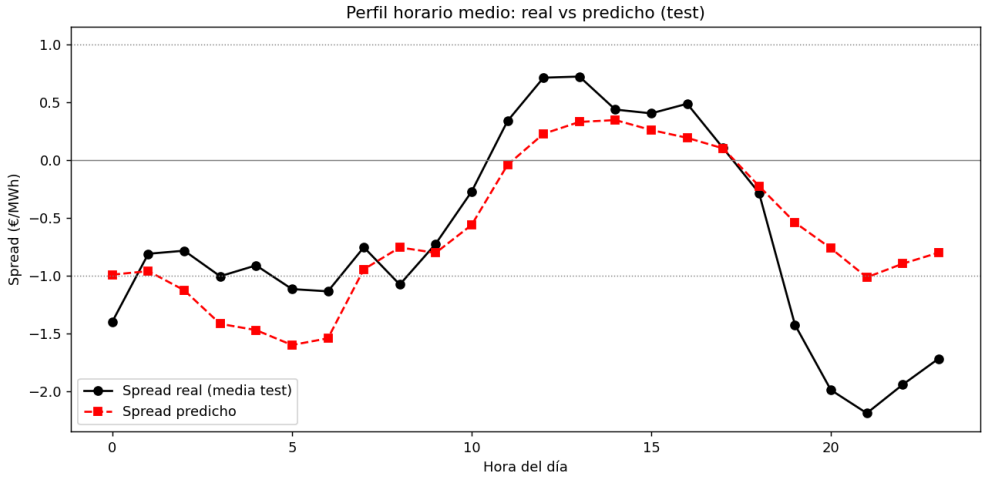

Figura 5.3: Perfil horario medio del spread real frente al predicho por el regresor sobre el conjunto de prueba. El modelo reproduce la forma media intradía, aunque la dispersión hora a hora siga siendo elevada.

5.3. Interpretación del modelo

Una vez evaluado el rendimiento, se analiza qué variables influyen en las predicciones. Este análisis permite comprobar si el modelo se apoya en información coherente con el problema y valorar la aportación específica de la etiqueta de clúster, que se incorporó como una variable adicional.

Se utilizan tres técnicas de interpretación que responden a preguntas diferentes: la importancia por permutación, la importancia por perturbación y los valores SHAP. Por esta razón, no es necesario que produzcan el mismo orden de variables. Sus resultados deben entenderse como perspectivas complementarias sobre el funcionamiento del modelo.

5.3.1. Importancia por permutación

La importancia por permutación [31] mide la reducción del rendimiento cuando los valores de una variable, o de un grupo de variables, se reordenan aleatoriamente. Este procedimiento rompe su relación con la variable objetivo y mantiene sin cambios el resto de la matriz. Cuanto mayor sea el deterioro observado, mayor es la información que el modelo obtenía de ese grupo. La definición formal de esta técnica se recoge en el Anexo D. La Figura 5.4 presenta los resultados.

La hora del día presenta la mayor importancia, seguida por la previsión eólica y los retardos del spread. El deterioro asociado al resto de grupos es reducido. En particular, la etiqueta de clúster muestra una importancia prácticamente nula y ligeramente negativa. Este valor no debe interpretarse como un efecto perjudicial estable, sino como ausencia de una mejora apreciable dentro de la variabilidad del procedimiento. En esta configuración, la pertenencia al clúster no parece aportar información adicional una vez incluidas las variables originales del sistema.

5.3.2. Importancia por perturbación

La importancia por perturbación analiza cuánto cambia la predicción al modificar una variable dentro de su rango mientras las demás permanecen fijas. A diferencia de la permutación, no mide la pérdida de rendimiento al eliminar información, sino la sensibilidad del resultado ante cambios en cada variable. Por ello, ambas técnicas pueden producir órdenes distintos. La Figura 5.5 recoge esta sensibilidad y la Figura 5.6 muestra las curvas de dependencia parcial de las variables

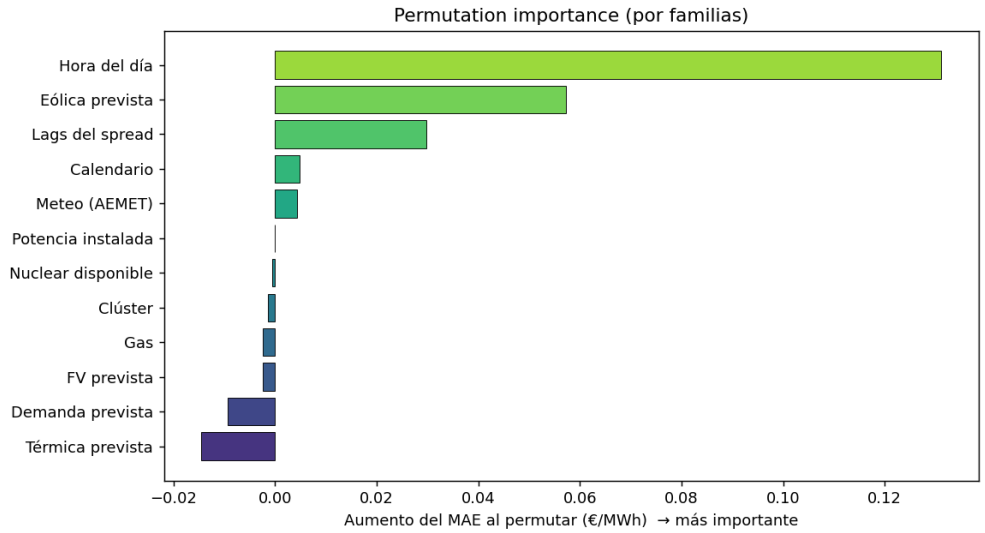

Figura 5.4: Importancia por permutación de cada grupo de variables, medida a partir del deterioro del rendimiento al reordenar sus valores. Los valores próximos a cero indican una aportación reducida.

más influyentes.

La hora del día vuelve a ser la variable más influyente, en consonancia con el perfil intradía observado en el spread. A continuación aparecen la previsión fotovoltaica y los retardos, mientras que la previsión eólica ocupa una posición algo inferior. El cambio de orden respecto a la permutación se debe a que ambas técnicas miden aspectos diferentes. La etiqueta de clúster se mantiene entre los grupos con menor influencia y su efecto vuelve a ser próximo a cero.

5.3.3. Valores SHAP

Los valores SHAP [32] descomponen cada predicción en las contribuciones asociadas a sus variables de entrada. El signo de cada contribución indica si la variable desplaza la predicción por encima o por debajo del valor de referencia, y su magnitud refleja la intensidad de ese efecto. Las contribuciones individuales pueden agregarse para obtener una medida global de importancia. Su formulación matemática, basada en el valor de Shapley de la teoría de juegos, se detalla en el Anexo D. La Figura 5.7 presenta el resultado por grupos y la Figura 5.8 muestra la distribución de las atribuciones para cada variable.

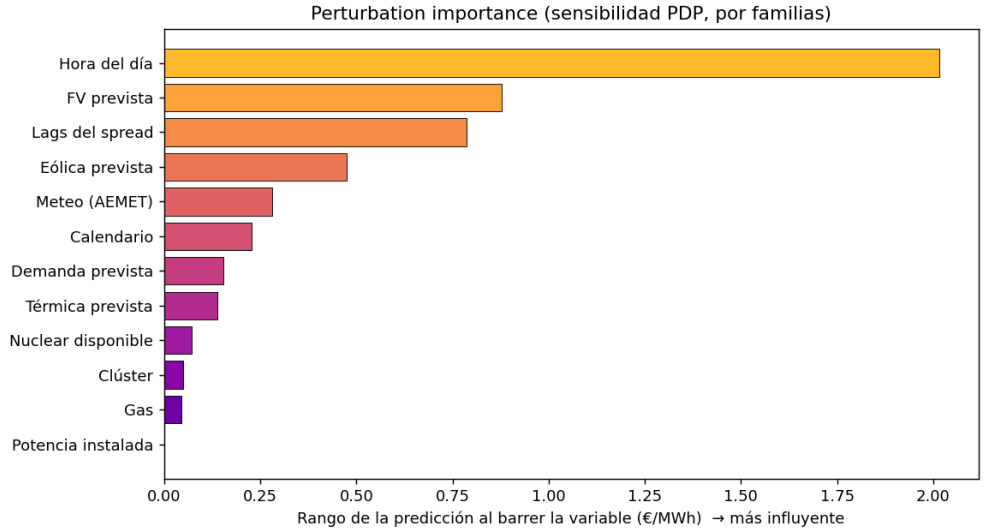

Figura 5.5: Importancia por perturbación de cada grupo de variables, medida como la sensibilidad de la predicción al recorrer el rango de la variable.

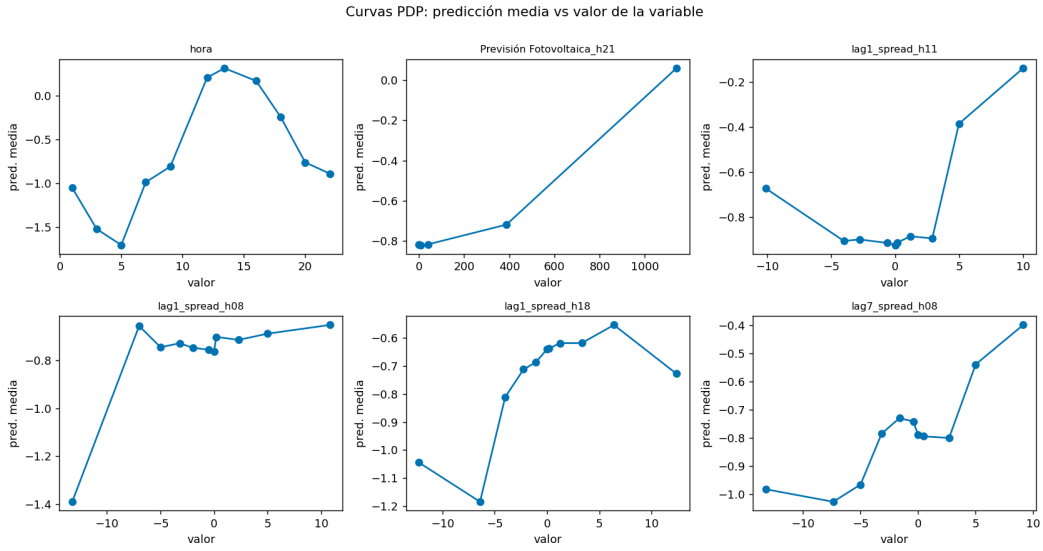

Figura 5.6: Curvas de dependencia parcial de las variables más influyentes. Cada curva muestra cómo varía la predicción media del spread al recorrer el rango de la variable correspondiente.

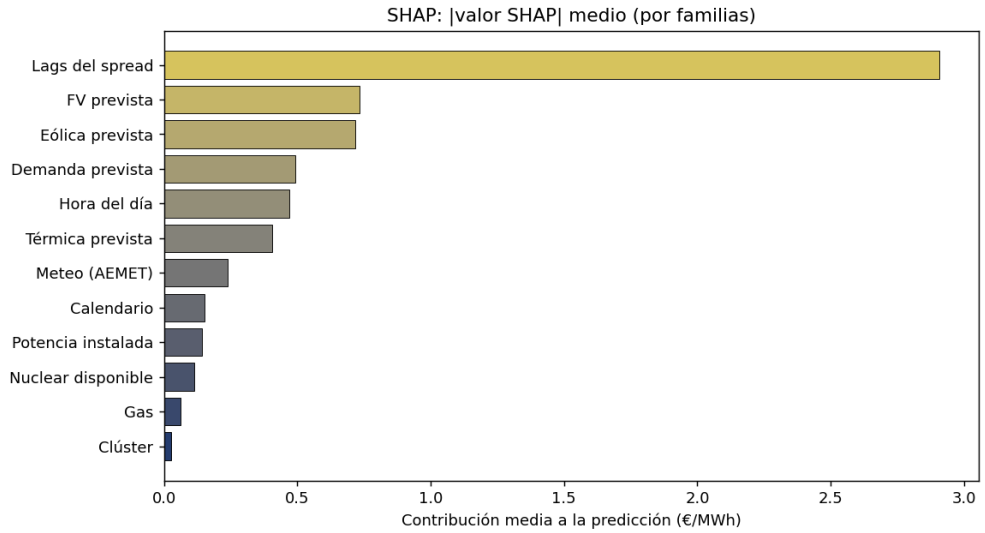

Figura 5.7: Importancia media por grupos según los valores SHAP, calculada como la magnitud media de la atribución de cada grupo sobre el conjunto de predicciones.

En este caso, los retardos del spread presentan la mayor importancia agregada, seguidos por las previsiones fotovoltaica y eólica. La hora del día ocupa una posición menos destacada. Esta diferencia está relacionada con la forma en la que se agrupan las variables: los retardos están representados por numerosas columnas, correspondientes a las horas del día anterior y del mismo día de la semana previa. Al sumar sus contribuciones, el grupo acumula un peso elevado aunque cada columna individual tenga una importancia reducida. La hora, en cambio, se representa mediante una única variable. La etiqueta de clúster vuelve a ocupar la última posición entre los grupos analizados.

5.3.4. Lectura conjunta

Las diferencias entre las tres técnicas responden a su propia definición. La permutación mide la pérdida de rendimiento al alterar la información de un grupo; la perturbación estudia la sensibilidad de la predicción ante cambios en sus valores; y SHAP reparte cada predicción entre todas las variables que intervienen en ella. Por tanto, ninguna proporciona por sí sola una ordenación definitiva.

A pesar de estas diferencias, los resultados coinciden en señalar tres fuentes principales de información: la hora del día, los retardos del spread y las previsiones de generación renovable. La hora recoge la estructura intradía; los retardos

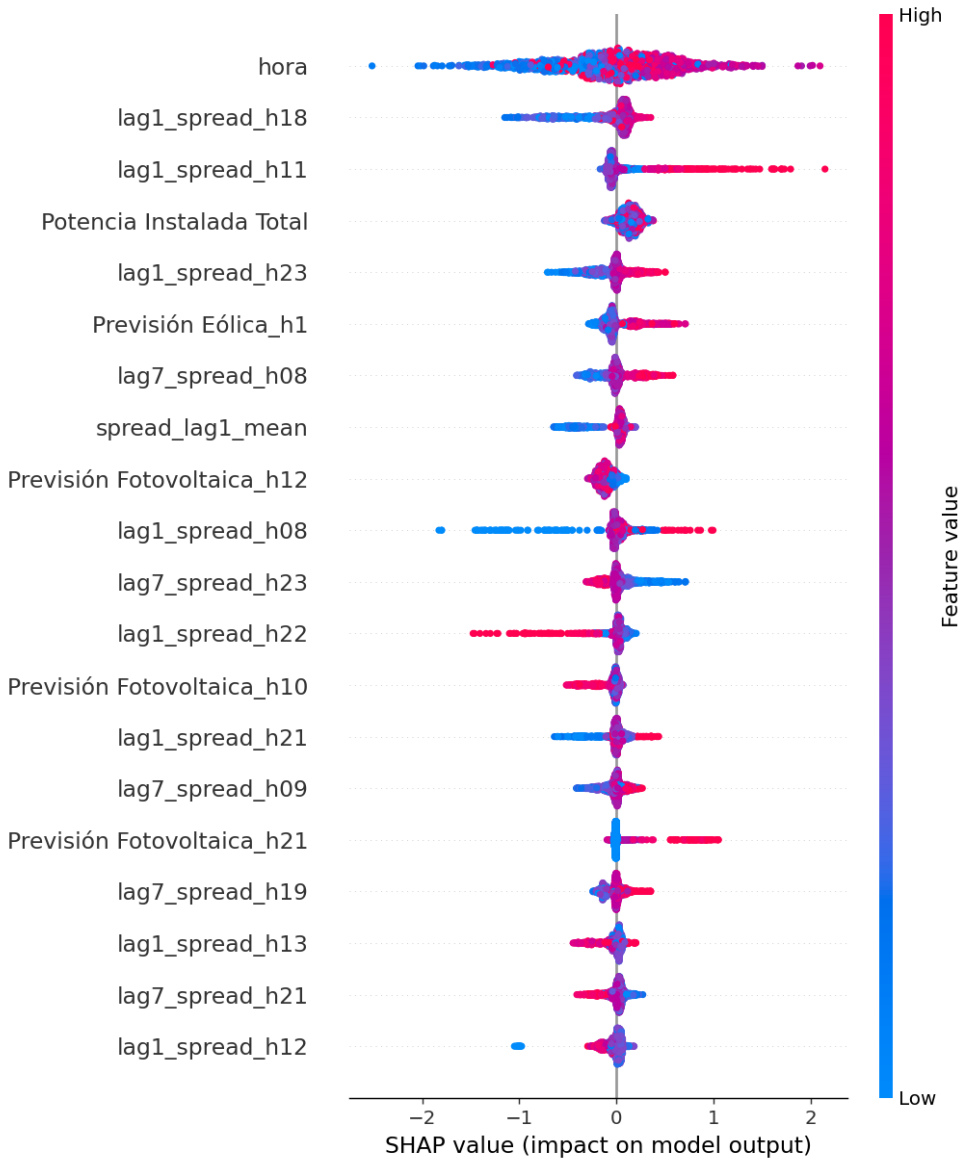

Figura 5.8: Distribución de los valores SHAP por variable. Cada punto representa una observación y su posición indica la magnitud y el sentido de la contribución de la variable a la predicción.

representan la dependencia con observaciones recientes; y las previsiones eólica y fotovoltaica describen parte de las condiciones esperadas del sistema. Por el contrario, la etiqueta de clúster aparece de forma consistente entre los grupos con menor importancia.

Este resultado no contradice el análisis del capítulo 3. La clusterización tenía como objetivo identificar y describir configuraciones recurrentes del sistema. La etiqueta asignada a cada día resume variables de demanda, generación renovable, meteorología y calendario que también se introducen directamente en el modelo supervisado. Cuando estas variables ya están disponibles con mayor nivel de detalle, la etiqueta no parece añadir información predictiva relevante. Por tanto, la utilidad descriptiva de los clústeres es compatible con su escasa aportación incremental en la predicción.

Por último, estas medidas reflejan asociaciones utilizadas por un modelo concreto y no relaciones causales. Que una variable tenga una importancia elevada no implica que una modificación de esa variable provoque por sí misma un cambio en el spread. Su contribución puede recoger también relaciones con otros factores no observados.

5.4. Análisis operativo y de robustez

Una vez caracterizado el rendimiento de los modelos, se estudian cinco aspectos adicionales: el efecto de los costes de operación, la significación estadística de la mejora observada, la distancia respecto al máximo teórico y la estabilidad de los resultados a lo largo del tiempo.

5.4.1. Costes de operación y umbral de confianza

La simulación anterior no incorpora costes de transacción. Conviene precisar qué representa este coste: no se trata de una comisión explícita cobrada por OMIE por la presentación de ofertas, ya que el número o el tamaño de las ofertas enviadas no tiene, en sí mismo, un coste diferenciado en las tarifas del operador del mercado. Se trata, en cambio, de una aproximación estilizada a los costes indirectos que sí enfrentaría un agente real al operar con volumen: el margen de ejecución (el precio efectivamente obtenido puede diferir ligeramente del precio marginal cuando se opera a mayor escala), el riesgo de desvíos o ejecución parcial, y los costes operativos y de gestión de riesgo asociados a mantener una posición activa en el

mercado. Su magnitud exacta en un caso real dependería del agente concreto, por lo que el análisis se repite para varios niveles de coste (0; 0,25; 0,5 y 1 €/MWh) con el fin de estudiar la sensibilidad del resultado a esta aproximación. Para analizar su efecto, se permite que la estrategia actúe únicamente cuando la magnitud de la predicción, $|\hat{s}|$, supera un determinado umbral. Esta magnitud se utiliza como una medida aproximada de la intensidad de la señal. La Figura 5.9 muestra el beneficio neto para distintos costes y umbrales, junto con la proporción de horas en las que se abre una posición.

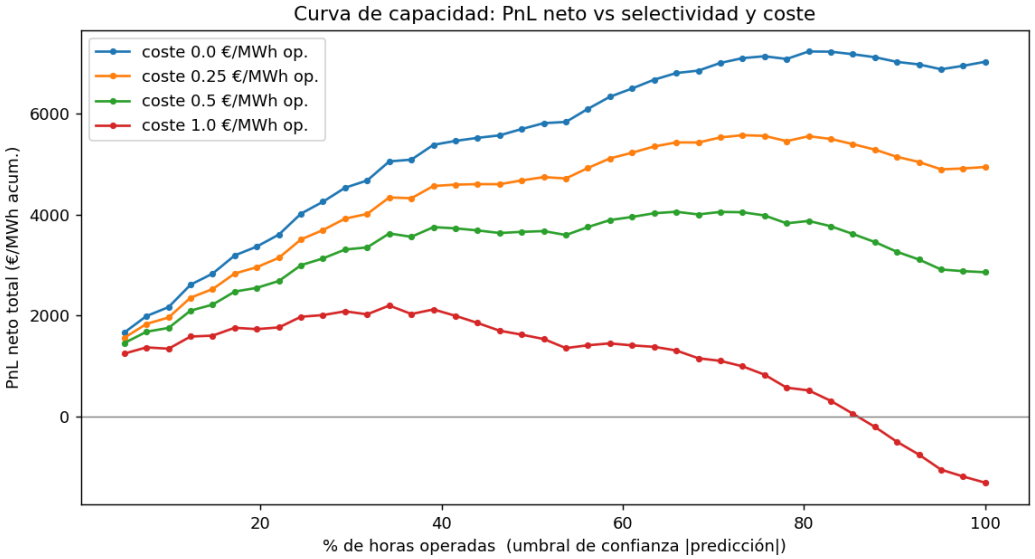

Figura 5.9: Beneficio neto en función del umbral $|\hat{s}|$ para distintos costes por operación, junto con el porcentaje de horas operadas. El aumento del coste desplaza el resultado óptimo hacia estrategias más selectivas.

En ausencia de costes, el mayor beneficio se obtiene al operar en todas las horas. Al introducir un coste de 1 €/MWh por operación, esta misma regla produce un resultado neto cercano a $-1{,}300$. En ese escenario, la estrategia vuelve a ser rentable al restringir la operación a las horas con señales de mayor magnitud. El máximo se alcanza al operar aproximadamente en el 34 % de las horas, con un beneficio neto en torno a 2.200 €/MWh en el conjunto de prueba. En general, el umbral óptimo aumenta con el coste, por lo que la estrategia reduce el número de operaciones y selecciona aquellas en las que el modelo anticipa un spread más alejado de cero. Este resultado muestra la utilidad de conservar una predicción continua, ya que $|\hat{s}|$ permite ordenar las operaciones según la intensidad estimada de la señal.

5.4.2. Significación estadística de los resultados

La mejora frente a las reglas de referencia debe contrastarse estadísticamente, ya que un beneficio positivo en un único conjunto de prueba puede aparecer por variabilidad muestral. Para ello se aplican dos procedimientos, cuyos resultados se recogen en la Figura 5.10.

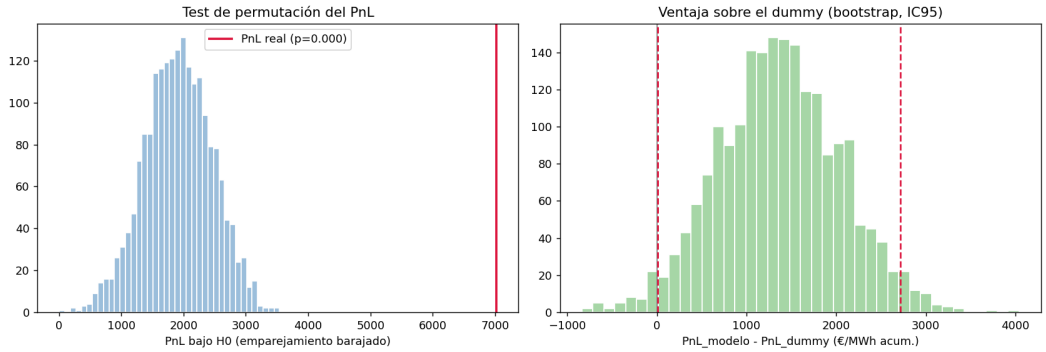

Figura 5.10: Pruebas de significación de los resultados. A la izquierda, distribución del PnL obtenida mediante permutación frente al valor observado. A la derecha, intervalo de confianza de la diferencia de PnL respecto a la clase mayoritaria.

La primera prueba es un test de permutación. Al reordenar repetidamente la relación entre las predicciones y los resultados observados se obtiene una distribución del PnL bajo la hipótesis de ausencia de asociación. El valor del modelo, 7,022, se sitúa muy por encima de la media de esta distribución, igual a 1,917, y el p-valor obtenido es 0,0005. El resultado aporta evidencia sólida de que el beneficio observado no se explica únicamente por una correspondencia aleatoria entre predicciones y resultados.

La segunda prueba utiliza remuestreo para estimar el intervalo de confianza de la diferencia de PnL entre el modelo y la clase mayoritaria. La diferencia media es de 1,360 y el intervalo del 95 % se extiende desde 7 hasta 2,718. Aunque el intervalo excluye el cero, su límite inferior se encuentra muy próximo a él. Por tanto, la mejora frente a la referencia es estadísticamente detectable, pero su magnitud presenta una incertidumbre considerable. En conjunto, las dos pruebas respaldan la existencia de una ventaja predictiva estadísticamente significativa, si bien su magnitud exacta conserva cierta incertidumbre, como refleja la amplitud del intervalo de confianza.

5.4.3. Comparación con el máximo teórico

La magnitud del rendimiento puede compararse con el máximo teórico que se obtendría al conocer de antemano el signo del spread en todas las horas. Esta referencia, denominada oráculo, no representa una estrategia aplicable, sino un límite superior para el conjunto de prueba. La Figura 5.11 compara este límite con el modelo y con la regla de clase mayoritaria.

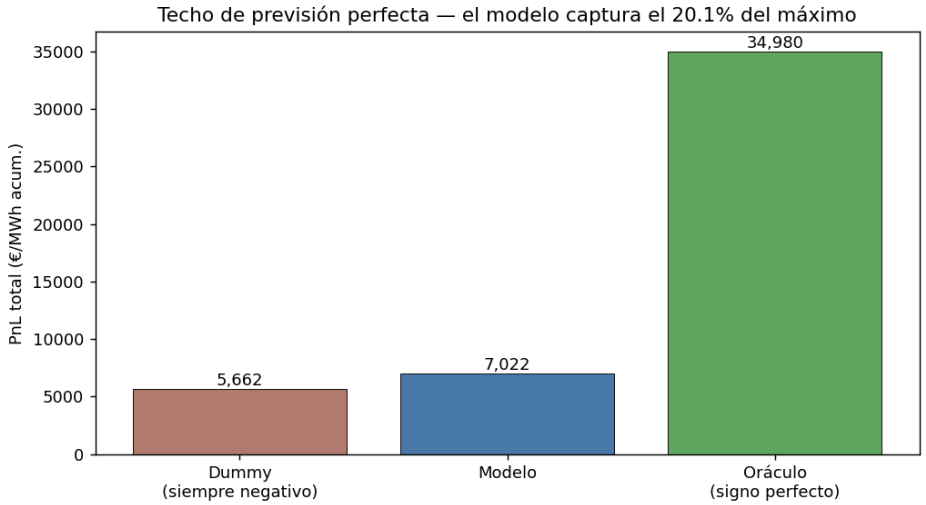

Figura 5.11: PnL del modelo y de la clase mayoritaria frente al máximo alcanzable mediante una previsión perfecta (oráculo). El modelo obtiene un 24 % más de beneficio que la clase mayoritaria y captura en torno al 20 % del máximo teórico.

El oráculo alcanzaría un PnL cercano a 35,000, mientras que el modelo obtiene 7,022, alrededor del 20 % de ese máximo. La clase mayoritaria alcanza aproximadamente el 16 %. La diferencia entre ambas estrategias equivale, por tanto, a unos cuatro puntos porcentuales del límite teórico. Esta comparación vuelve a mostrar que una parte importante del resultado procede del sesgo estructural del spread. La mejora asociada a la predicción específica de cada hora, sin embargo, no es desdeñable: el modelo obtiene un 24 % más de beneficio que la regla de clase mayoritaria, aunque el resultado conjunto se mantiene todavía lejos del límite teórico marcado por el oráculo. El resultado es compatible con una baja predictibilidad ex-ante a partir de las variables consideradas, aunque no constituye por sí solo una prueba de eficiencia del mercado.

5.4.4. Distribución horaria de la capacidad predictiva

Los análisis anteriores agregan todas las horas del día, pero el capítulo 3 mostró que el spread presenta perfiles intradía muy marcados, por lo que cabe preguntarse en qué tramos horarios se concentra la capacidad predictiva. La Figura 5.12 desglosa por hora de entrega la tasa de acierto direccional y el PnL del regresor sobre el conjunto de prueba, comparados con la clase mayoritaria.

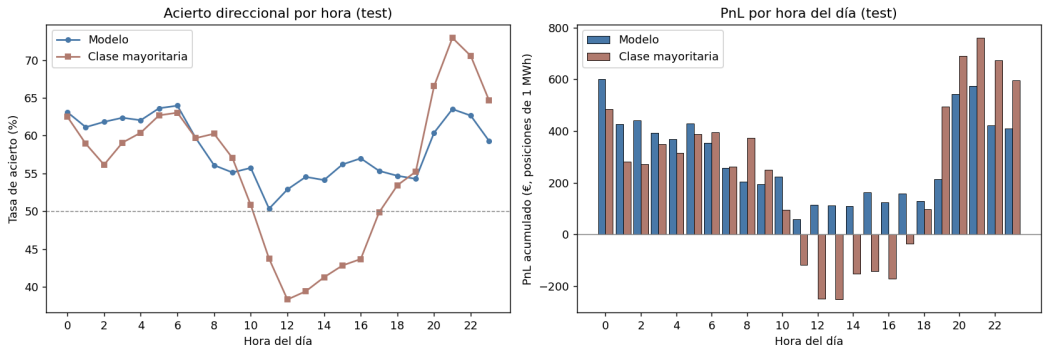

Figura 5.12: Rendimiento por hora de entrega en el conjunto de prueba: tasa de acierto direccional sobre las horas con spread no nulo (izquierda) y PnL acumulado (derecha), para el regresor y para la clase mayoritaria.

El desglose revela una estructura clara. En las horas centrales del día (aproximadamente de las 11 h a las 17 h) el sesgo negativo estructural del spread desaparece: la proporción de horas con spread positivo se aproxima al 50 %, consecuencia previsible de la expansión fotovoltaica durante el periodo de prueba, y la clase mayoritaria pasa a perder dinero, con tasas de acierto que caen hasta el 38 % a las 12 h. El modelo, en cambio, mantiene la tasa de acierto por encima del 50 % y un PnL positivo en todas las horas del día. Es en ese tramo central donde se genera la mayor parte de su ventaja frente a la referencia. En las primeras horas del día (0 h–5 h) el modelo también supera ligeramente a la mayoritaria, mientras que en la tarde-noche (19 h-23 h) ocurre lo contrario: el sesgo negativo es allí tan dominante que la regla fija resulta muy difícil de batir y el modelo, al arriesgar señales positivas, cede parte de ese beneficio.

Este resultado matiza la expectativa formulada al final del capítulo 3. Los tramos de madrugada y tarde-noche eran los de mayor contraste entre clústeres y, en la madrugada, el modelo efectivamente rinde bien; pero en la tarde-noche la mejor decisión resulta ser la sistemática, sin necesidad de modelo alguno. El valor añadido del aprendizaje automático no se encuentra, por tanto, en los tramos

donde el signo es más previsible, sino en aquellos donde el sesgo estructural se debilita y la dirección debe decidirse caso a caso.

El mismo desglose puede hacerse por regímenes, utilizando el clúster de pertenencia de cada día (Tabla 5.2). El periodo de prueba está dominado por los clústeres recientes: el invierno con sistema expandido (C2), el verano fotovoltaico (C5) y los días de eólica alta (C7). La ventaja del modelo se concentra en C5 (+1,164 € sobre la clase mayoritaria) y, muy especialmente, en C7, donde la regla fija llega a perder dinero (−482 €) mientras el modelo gana (+326 €): son precisamente los dos regímenes en los que el sesgo negativo estructural se debilita, en las horas solares del verano fotovoltaico y en la madrugada de los días borrascosos identificada en el capítulo 3. En el invierno reciente (C2), en cambio, el sesgo negativo es tan dominante que la regla mayoritaria resulta imbatible y el modelo cede parte del beneficio. En los días festivos del periodo de prueba (C3), el modelo coincide exactamente con la mayoritaria: predice signo negativo en todas sus horas, en línea con el perfil sistemáticamente negativo de ese clúster.

Tabla 5.2: Desglose del PnL del regresor en el conjunto de prueba según el clúster de pertenencia del día ($K = 8$). La diferencia frente a la clase mayoritaria mide la ventaja aportada por el modelo en cada régimen (en euros, posiciones de 1 MWh).

Clúster	Días	PnL modelo	PnL mayoritaria	Diferencia
C0 (menor capacidad firme)	6	461	596	−135
C2 (invierno reciente)	170	2.778	3.254	−477
C3 (festivos)	7	263	263	0
C5 (verano fotovoltaico)	127	3.195	2.031	+1,164
C7 (días borrascosos)	37	326	−482	+808

5.4.5. Estabilidad en el tiempo

Por último, se comprueba si el rendimiento se concentra en un periodo concreto o se mantiene a lo largo del tiempo. A diferencia de los análisis anteriores, restringidos al conjunto de prueba final, aquí se utilizan las predicciones fuera de muestra de las sucesivas etapas de la validación con ventana creciente: en cada etapa el modelo se reentrena con los mismos hiperparámetros y predice el tramo siguiente. Esto permite desglosar el rendimiento por años entre 2020 y 2025, si bien el PnL agregado de esta tabla no es comparable con el del conjunto de prueba final. La Tabla 5.3 y la Figura 5.13 recogen la evaluación anual fuera de muestra.

Las dos medidas presentan comportamientos diferentes. El AUC se mantiene entre 0,537 y 0,604, y permanece por encima de 0,5 en todos los años. Esto indica

Tabla 5.3: Rendimiento anual del modelo sobre el conjunto de prueba. El AUC mide la capacidad discriminativa; el PnL representa el beneficio unitario acumulado para una posición de 1 MWh en cada operación; y el porcentaje del oráculo indica la fracción del máximo teórico obtenida en cada año. El PnL/hora es el beneficio medio por hora operada, en €/MWh.

Año	Días	MAE	AUC	PnL	PnL/hora	% del oráculo
2020	149	0,99	0,576	497	0,14	17,4
2021	252	2,08	0,537	676	0,11	5,9
2022	251	4,20	0,580	8.027	1,33	32,7
2023	249	3,16	0,591	4.957	0,83	28,9
2024	244	3,83	0,546	2.916	0,50	14,2
2025	239	4,51	0,604	7.366	1,28	28,6

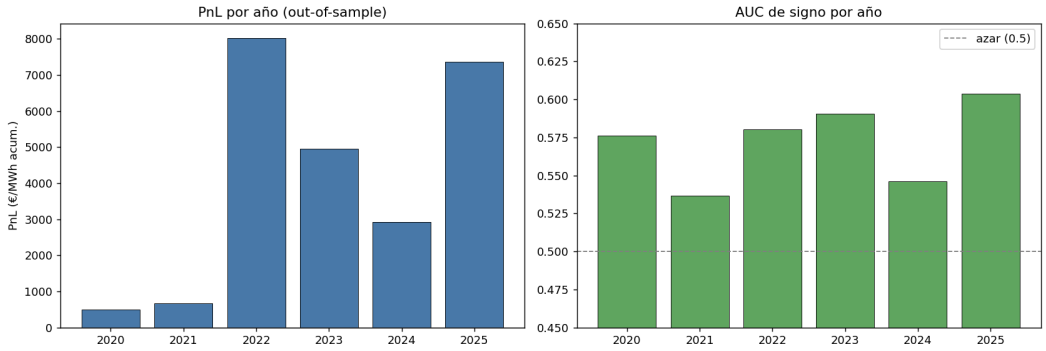

Figura 5.13: Evolución anual de la capacidad discriminativa (AUC) y del PnL. El AUC presenta una variación moderada, mientras que el resultado económico cambia de forma más acusada entre años.

una capacidad discriminativa reducida y relativamente estable, aunque no permite afirmar que la mejora sea estadísticamente significativa en cada año por separado. El PnL, en cambio, varía desde 676 en 2021 hasta 8,027 en 2022. Esta diferencia se debe a que el beneficio depende tanto de la dirección acertada como de la magnitud del spread en esas horas. Por ello, años con mayor dispersión de precios ofrecen un valor económico superior para una capacidad direccional similar. La columna de PnL por hora operada lo hace explícito: la rentabilidad unitaria oscila entre 0,11 €/MWh en 2021 y 1,33 €/MWh en 2022 para niveles comparables de AUC.Los resultados de 2022 ilustran este efecto: el modelo no alcanza allí el mayor AUC, pero sí el mayor PnL. En consecuencia, la habilidad direccional muestra una estabilidad razonable, mientras que la rentabilidad está condicionada por las características de cada periodo.

5.5. Discusión

Los resultados muestran que la magnitud horaria del spread es difícil de estimar con las variables disponibles, como refleja el coeficiente de determinación próximo a cero. En la predicción direccional sí aparece una mejora frente a las reglas de referencia, pero su magnitud es moderada. Además, una parte importante del beneficio acumulado puede obtenerse mediante la clase mayoritaria, debido al sesgo negativo de la variable objetivo. La aportación específica de los modelos se observa en la mejora sobre esa referencia, en las métricas balanceadas y en las pruebas de significación.

La validación temporal, la comparación con referencias exigentes y la incorporación de costes permiten evitar una valoración excesivamente optimista del rendimiento. Bajo este procedimiento, el regresor ofrece una mejor discriminación entre direcciones y el clasificador obtiene un PnL superior, aunque ambos se mantienen lejos del límite teórico. La evaluación anual indica, además, que la capacidad direccional no se concentra en un solo año, mientras que el beneficio varía con la magnitud de las oportunidades disponibles.

Por tanto, el principal resultado no es la obtención de una estrategia directamente aplicable, sino la cuantificación de la parte del spread que puede anticiparse con la información considerada. El análisis permite separar el efecto del sesgo estructural de la mejora aportada por los modelos y delimitar el alcance real de la predicción.

5.6. Limitaciones y líneas de extensión

La interpretación de los resultados está condicionada por varias limitaciones, algunas de las cuales ya se señalaron en capítulos anteriores.

- **Regla de operación simplificada.** La estrategia utiliza el signo predicho, un umbral fijo y un volumen constante. No adapta el tamaño de la posición a la intensidad de la señal ni representa de forma completa los costes, la liquidez o las restricciones operativas del mercado. Por ello, la simulación sirve para comparar la utilidad relativa de los modelos, pero no constituye una estrategia lista para su aplicación.
- **Meteorología observada y no prevista.** Como se explicó en el capítulo 2, se utilizan valores meteorológicos observados porque no se dispone de un histórico público, homogéneo y completo de las previsiones disponibles antes de cada jornada. Estas observaciones no reproducen exactamente la información ex-ante de un agente y pueden favorecer la evaluación del modelo. En una aplicación real, el rendimiento podría ser inferior al estimado.
- **Predominio de días laborables.** La cotización del gas solo está disponible para días laborables, lo que provoca la exclusión práctica de los fines de semana, como se explicó en el capítulo 3. En consecuencia, las conclusiones describen principalmente el comportamiento de las jornadas laborables y no deben trasladarse sin comprobación al resto de los días.
- **Cambios temporales del mercado.** La relación entre las variables explicativas y el spread puede variar como consecuencia de cambios regulatorios, tecnológicos o de condiciones de mercado. Un ejemplo concreto es el uso de variables de generación renovable en términos absolutos, tanto en la clusterización como en los umbrales de división de los árboles: dado el crecimiento sostenido de la potencia instalada, estos modelos podrían perder capacidad de generalización a medida que el mix de generación se aleje del observado durante el periodo de entrenamiento.
- **Ajuste no incremental de la clusterización.** A diferencia de los modelos predictivos, los centroides que definen los ocho clústeres del capítulo 3 se ajustaron una única vez sobre la totalidad de la muestra 2019-2025, y no se reajustaron de forma incremental en cada etapa de la validación temporal. Esto introduce una fuga de información limitada al proceso de evaluación histórica de los días más antiguos de la muestra, con un impacto reducido sobre los resultados principales, obtenidos sobre el tramo final y más reciente

de la muestra, y atenuado por la baja importancia que la etiqueta de clúster tiene en el modelo final, como se explica en la sección 4.3.3.

Estas limitaciones permiten identificar varias líneas de trabajo que podrían desarrollarse en estudios posteriores.

- **Regresión por cuantiles.** La estimación de varios cuantiles permitiría representar la incertidumbre de la predicción con más detalle que la magnitud $|\hat{s}|$. A partir de estos intervalos podría definirse una regla de operación que ajustase el volumen a la dispersión esperada del spread.
- **Otras sesiones del mercado intradiario.** El análisis se ha limitado a la primera sesión. Su extensión a las sesiones posteriores o al mercado intradiario continuo permitiría comprobar cómo cambia la capacidad predictiva a medida que se incorpora nueva información y en qué momento se concentra la utilidad económica.
- **Nuevas variables y métodos de interpretación.** La incorporación de previsiones meteorológicas históricas, información sobre intercambios transfronterizos u otras variables disponibles antes del cierre podría ampliar la señal capturada. También podrían emplearse explicaciones contrafactuales para estudiar qué cambios en las entradas habrían modificado una decisión concreta del modelo.

Estas limitaciones delimitan el alcance de los resultados y señalan los aspectos que deben mejorarse antes de una posible aplicación. Con la información y el periodo analizados, los modelos presentan una capacidad direccional reducida y relativamente estable en el desglose anual. Sin embargo, una parte mayoritaria del beneficio procede del sesgo estructural del spread y la rentabilidad varía de forma importante entre años. El análisis realizado permite cuantificar ambas componentes y establecer una referencia para futuras ampliaciones del trabajo.

Capítulo 6

Conclusiones

6.1. Síntesis del trabajo realizado

Este trabajo ha estudiado si el *spread* horario entre el mercado diario y el mercado intradiario del sistema eléctrico ibérico puede anticiparse utilizando únicamente información disponible antes del cierre del mercado diario y, en caso afirmativo, cuál es el valor económico de esa capacidad de anticipación. Para responder a esta cuestión se ha desarrollado un proceso completo de ciencia de datos aplicada, desde la construcción de la base de datos hasta la simulación económica de una estrategia de operación.

En una primera fase se construyó una base analítica de ámbito peninsular y resolución horaria que integra los precios del mercado diario y de la primera sesión del intradiario, las previsiones de demanda y de generación renovable, la disponibilidad de potencia nuclear, la potencia instalada, el precio del gas natural TTF como variable exógena de coste, variables meteorológicas agregadas y un calendario laboral enriquecido con el porcentaje de población afectada por cada festivo. Tras la depuración y homogeneización de fuentes, la muestra final quedó formada por 1.763 días laborables entre enero de 2019 y diciembre de 2025.

Sobre esa base se aplicaron, en segundo lugar, técnicas de agrupación no supervisada para comprobar si los días se organizan en configuraciones recurrentes del sistema. La solución de referencia, K-means con ocho clústeres, se seleccionó tras comparar 35 configuraciones de método y número de grupos, y se validó mediante contrastes estadísticos sobre el *spread* y análisis de estabilidad por remuestreo. En tercer lugar, se desarrollaron modelos supervisados de *gradient boosting* para

predecir el *spread* horario en dos formulaciones complementarias (regresión de su valor y clasificación de su signo), ajustados mediante búsqueda bayesiana de hiperparámetros y evaluados con validación temporal de ventana creciente. Finalmente, las predicciones se tradujeron en decisiones de mercado y se evaluaron económicamente frente a referencias exigentes, incorporando costes de operación, umbrales de confianza y contrastes de significación estadística.

6.2. Conclusiones principales

6.2.1. El sistema se organiza en regímenes reconocibles

La agrupación de días construida exclusivamente con variables conocidas antes del cierre del mercado diario, sin etiquetas temporales ni información del propio *spread*, recupera de forma emergente los principales regímenes históricos del mercado eléctrico español: el periodo previo a la crisis energética, caracterizado por precios moderados del gas; la crisis del gas de 2022, que el algoritmo identifica como un clúster propio con un precio medio del gas de 173 €/MWh; y el régimen reciente asociado a la expansión fotovoltaica y la normalización parcial de precios. Que esta estructura aparezca a partir de previsiones de demanda, renovables y variables de coste y calendario confirma que la información *ex ante* empleada contiene capacidad descriptiva relevante sobre el estado del sistema.

Dos resultados de esta fase merecen destacarse. El primero es que latipología más diferenciada de toda la muestra corresponde a los días festivos: un grupo reducido (47 días con más del 85 % de la población afectada) pero con el comportamiento más extremo del *spread*, sistemáticamente negativo a todas las horas, con una media de $-1{,}74$ €/MWh y valores próximos a -4 €/MWh durante la madrugada. La diferencia entre las soluciones de siete y ocho clústeres se reduce, en la práctica, a si este grupo aparece aislado o no, y esa capacidad de identificación fue determinante para seleccionar la solución de ocho clústeres. El segundo resultado es que la granularidad de la segmentación resulta relevante: con solo tres macrogrupos, las diferencias de *spread* entre clústeres no alcanzan significación estadística ($p \approx 0{,}07$), mientras que con siete u ocho grupos tanto ANOVA como Kruskal-Wallis rechazan claramente la igualdad ($p < 10^{-3}$). Esto sugiere que la estructura útil del *spread* no se encuentra únicamente en distinciones amplias entre estaciones o niveles de precio, sino en configuraciones más específicas del sistema.

6.2.2. La magnitud del spread es difícil de anticipar; su dirección, solo parcialmente

El valor exacto del *spread* horario resulta muy difícil de anticipar con la información considerada: el mejor modelo de regresión explica menos del 2 %de su variabilidad ($R^2 = 0{,}018$), y ni siquiera las referencias más informadas (como el perfil medio del clúster del día) superan ese orden de magnitud. Este resultado no debe interpretarse como un fallo del procedimiento, sino como una característica del propio fenómeno: el *spread* concentra un 19 % de sus horas exactamente en cero, presenta colas pesadas y depende de perturbaciones de última hora que, por definición, no están incluidas en la información disponible la víspera.

La dirección del *spread*, en cambio, sí contiene una señal anticipable. El regresor alcanza un AUC de 0,590 y el clasificador de 0,584, frente al 0,565 de la persistencia y el 0,500 del azar. Además, esta capacidad direccional no se concentra en un único periodo: en la evaluación anual entre 2020 y 2025, el AUC se mantiene siempre por encima de 0,5 (entre 0,537 y 0,604). El contraste entre ambas formulaciones apunta a una conclusión relevante para este tipo de problemas: cuando la variable objetivo está dominada por ruido y solo contiene una componente sistemática débil, la predicción direccional puede resultar más útil que la estimación precisa de la magnitud.

Igualmente relevante es identificar dónde se concentra esa señal. El desglose horario muestra que la ventaja del modelo frente a las reglas fijas no aparece en los tramos donde el signo es más previsible, sino en aquellos donde el sesgo estructural negativo del *spread* pierde intensidad. En las horas centrales del día, coincidiendo con la producción fotovoltaica, la regla de apostar siempre al signo mayoritario llega a generar pérdidas (con aciertos de solo el 38 % a las 12 h), mientras que el modelo mantiene tasas de acierto superiores al 50 % y un resultado positivo en las veinticuatro horas. El análisis por regímenes confirma esta misma interpretación: la ventaja se concentra en los veranos fotovoltaicos y en los días de elevada producción eólica (donde la referencia fija llega a ser negativa), y desaparece en el invierno reciente, donde el sesgo negativo es tan dominante que las predicciones individualizadas apenas mejoran la regla sistemática.

6.2.3. La predicción aporta valor económico positivo y significativo, aunque lejos del máximo teórico

Traducida a una estrategia de operación con posiciones de 1 MWh por hora, la predicción del regresor acumula un beneficio de 7.022 € en el conjunto de prueba (julio de 2024 a diciembre de 2025), frente a los 5.662 € de la regla de clase mayoritaria y los 34.980 € que obtendría un oráculo con previsión perfecta del signo. El modelo captura, por tanto, en torno al 20 % del máximo teórico, aproximadamente cuatro puntos porcentuales más que la mejor regla fija. Esta ventaja es estadísticamente significativa (el test de permutación arroja $p = 0{,}0005$ y el intervalo *bootstrap* de la diferencia frente a la referencia excluye el cero), aunque su magnitud exacta presenta incertidumbre (intervalo del 95 % entre 7 y 2.718 €), lo que aconseja interpretar el resultado con prudencia.

La descomposición de esa ventaja resulta especialmente informativa. El valor añadido procede del 33 % de horas en las que el modelo adopta una decisión contraria al sesgo negativo dominante. En esas horas el *spread* realizado es positivo solo ligeramente por encima de la mitad de las veces (54 % entre horas con *spread* no nulo), pero su valor medio es de +0,25 €/MWh, de modo que la decisión contraria presenta esperanza positiva. El rendimiento del modelo no proviene tanto de una elevada tasa de acierto como de su capacidad para identificar situaciones en las que resulta rentable apartarse de la regla por defecto.

Los costes de transacción no eliminan completamente este valor, aunque sí obligan a una operativa más selectiva. Con un coste de 1 €/MWh por operación, operar todas las horas produce pérdidas netas; restringiendo la operación al tercio de horas con señales más intensas, la estrategia vuelve a obtener un beneficio neto positivo (en torno a 2.200 € en el conjunto de prueba). Disponer de una predicción continua, y no únicamente del signo, permite precisamente establecer este tipo de ordenación por confianza. Por último, la rentabilidad unitaria depende de forma notable del régimen de mercado: para una capacidad direccional similar, el beneficio por hora operada osciló entre 0,11 €/MWh en 2021 y 1,33 €/MWh en 2022, lo que indica que el valor económico de acertar la dirección depende también de la magnitud de los *spreads* presentes en cada periodo.

6.2.4. Conclusiones metodológicas

El desarrollo del trabajo permite extraer también varias conclusiones metodológicas potencialmente útiles más allá de esta aplicación concreta.

La primera es que la elección de la métrica de optimización condiciona de forma decisiva el comportamiento del modelo. Cuando la búsqueda de hiperparámetros se orientó a minimizar el error absoluto medio, las predicciones tendieron a concentrarse alrededor de cero: el error estadístico apenas mejoró, pero el beneficio acumulado se redujo aproximadamente a la mitad y el número de operaciones cayó por debajo de un tercio. Alinear el criterio de optimización con el objetivo económico del problema resultó ser una de las decisiones metodológicas más relevantes del trabajo.

La segunda es que, en series temporales de mercado, el protocolo de evaluación resulta tan importante como el propio modelo. La validación estrictamente temporal, la comparación con referencias exigentes (la clase mayoritaria captura por sí sola el 74 % del beneficio del mejor modelo) y los contrastes de significación fueron esenciales para no confundir el sesgo estructural del *spread* con capacidad predictiva genuina. Una partición aleatoria y una comparación ingenua habrían conducido a conclusiones considerablemente más optimistas y, probablemente, poco robustas.

La tercera conclusión conecta las dos partes principales del trabajo. La etiqueta de clúster, incorporada como variable de entrada, resultó ser la menos importante para el modelo supervisado: las variables originales del sistema ya contienen, con mayor nivel de detalle, la información que dicha etiqueta resume. Sin embargo, esto no implica que la clusterización haya sido irrelevante. La partición en regímenes permite explicar *dónde* y *por qué* el modelo aporta valor: en qué horas, en qué tipos de día y bajo qué condiciones del sistema, y esa capacidad interpretativa resulta especialmente valiosa desde el punto de vista aplicado. La utilidad de la agrupación no reside tanto en mejorar la predicción como en facilitar la comprensión de sus resultados.

Finalmente, la fase de datos confirmó ser una parte sustancial y no meramente instrumental del trabajo. Decisiones como la homogeneización al ámbito peninsular, el tratamiento de los festivos trasladados o la limitación de la muestra a días laborables derivada de la disponibilidad de la cotización del gas condicionaron directamente el alcance y la validez del análisis posterior.

6.3. Cumplimiento de los objetivos

Los cinco objetivos específicos formulados en la sección 1.5 se han cumplido en su totalidad. Se ha construido una base de datos peninsular homogénea de resolución horaria y se ha analizado el comportamiento del *spread* en relación con las

variables del sistema (objetivo 1, capítulo 2). Se ha desarrollado un modelo de agrupación de días basado exclusivamente en información anticipada (objetivo 2) y se han caracterizado los perfiles de *spread* de cada grupo, verificando estadísticamente sus diferencias(objetivo 3, capítulo 3). Se han construido modelos predictivos directos del *spread* en formulación de regresión y de clasificación (objetivo 4, capítulos 4 y 5) y, por último, se ha evaluado su impacto económico mediante una simulación con costes, umbrales de confianza y contrastes de significación (objetivo 5, capítulo 5).

El objetivo general, desarrollar una herramienta de apoyo a la decisión basada en la predicción del *spread*, puede considerarse cumplido en los términos que permiten los resultados obtenidos: se ha identificado una señal direccional real, económicamente positiva y estadísticamente significativa (un 24, % más de beneficio que la mejor referencia simple), aunque concentrada en determinados tramos horarios y regímenes del sistema y todavía alejada del máximo teórico.

6.4. Implicaciones prácticas

Desde un punto de vista operativo, los resultados obtenidos permiten extraer varias implicaciones relevantes para un agente que opere entre ambos mercados.

En primer lugar, la estrategia de referencia en este mercado no es la ausencia de operación, sino el propio sesgo estructural del *spread*: vender en el intradiario lo comprado en el diario captura por sí solo una parte importante del beneficio bruto disponible. El modelo debe interpretarse, por tanto, como un refinamiento selectivo de esa regla y no como un sustituto completo de la misma.

En segundo lugar, el valor añadido del modelo se concentra en condiciones identificables *a priori*: las horas centrales del día, los regímenes de elevada penetración fotovoltaica y los días de abundante generacióneólica. En las horas de tarde y noche, así como en los inviernos recientes con fuerte sesgo negativo, la regla sistemática resulta difícil de mejorar.

En tercer lugar, la presencia de costes de operación hace necesaria una estrategia selectiva. El umbral de confianza óptimo aumenta con el coste de transacción, y la disponibilidad de una predicción continua proporciona el criterio de ordenación necesario para decidir cuándo operar.

Por último, los resultados muestran que la relación entre las variables del sistema y el *spread* evoluciona con el tiempo. Cualquier aplicación operativa requeriría,

por tanto, procesos periódicos de reentrenamiento y revalidación, tal como simula el esquema de ventana creciente utilizado en este trabajo.

6.5. Limitaciones

Las conclusiones anteriores deben interpretarse dentro de los límites del estudio. Las variables meteorológicas empleadas son observaciones y no las previsiones que un agente habría tenido disponibles en tiempo real, lo que puede introducir un sesgo favorable en la evaluación del modelo; esta es la limitación más relevante desde el punto de vista de la aplicabilidad *ex ante*.

Además, la clusterización del capítulo 3 se ajustó una única vez sobre la totalidad del periodo 2019–2025, sin reajustarse de forma incremental en cada etapa de la validación temporal como sí ocurre con los modelos predictivos. Esto introduce una fuga de información limitada al proceso de evaluación histórica, cuyo impacto se concentra en los años más alejados del final de la muestra y se ve atenuado por la escasa importancia que la etiqueta de clúster tiene en el modelo final.

Asimismo, el estudio se limita a la primera sesión del mercado intradiario, cuya estructura ha cambiado durante el propio periodo analizado. La simulación económica utiliza además una regla deliberadamente simple, con volumen fijo de 1 MWh por operación y sin modelar aspectos como la liquidez, el impacto en precio o las restricciones operativas, por lo que sus resultados deben interpretarse principalmente como una herramienta de comparación entre modelos y no como una estimación directa de rentabilidad alcanzable en condiciones reales.

Por último, aunque la ventaja económica frente a la mejor regla fija resulta estadísticamente significativa, su intervalo de confianza es amplio. La evidencia obtenida permite sostener que dicha ventaja existe, aunque no determinar con precisión su magnitud exacta.

6.6. Líneas futuras

El trabajo deja abiertas varias líneas de extensión naturales.

La primera es la regresión por cuantiles. La predicción puntual empleada en este trabajo solo proporciona una medida indirecta de confianza a través de la magnitud $|\hat{s}|$. La estimación de distintos cuantiles del *spread* permitiría construir intervalos de

predicción y desarrollar estrategias de operación que adapten el volumen negociado al nivel de incertidumbre.

La segunda es la extensión a otras sesiones del mercado intradiario y al mercado continuo. Cada sesión incorpora información más reciente, por lo que cabe esperar que tanto la predictibilidad del *spread* como su valor económico dependan del horizonte temporal considerado.

La tercera es la incorporación de previsiones meteorológicas históricas (*reforecasts*) en lugar de observaciones reales, eliminando así la principal amenaza a la validez *ex ante* de la evaluación. También sería de interés incorporar nuevas variables disponibles antes del cierre del mercado, como la capacidad de interconexión, los flujos transfronterizos programados, la disponibilidad hidráulica o las indisponibilidades comunicadas de generación.

La cuarta línea futura es la operacionalización del sistema predictivo: calibración de probabilidades, políticas óptimas de reentrenamiento frente a derivas de régimen y monitorización sistemática del rendimiento por tramo horario y tipo de día.

Por último, quedaría abierta la comparación con arquitecturas de aprendizaje profundo, como redes recurrentes o convolucionales aplicadas al perfil horario completo. La infraestructura de evaluación desarrollada en este trabajo (base de datos, referencias, validación temporal y métrica económica) permitiría incorporar este tipo de modelos de forma directamente comparable.

6.7. Reflexión final

El *spread* entre el mercado diario y el mercado intradiario ibérico es, con la información disponible antes del cierre del mercado diario, un fenómeno mayoritariamente ruidoso, aunque no completamente impredecible. Los modelos desarrollados capturan aproximadamente una quinta parte de lo que sería alcanzable con información perfecta, y esa capacidad predictiva no se distribuye de forma homogénea: se concentra en determinados tramos horarios, regímenes y condiciones del sistema que este trabajo ha identificado y cuantificado.

En un contexto donde la utilidad práctica de los modelos predictivos depende tanto de su robustez como de sus limitaciones, determinar cuánta señal existe, dónde se concentra y bajo qué condiciones resulta aprovechable constituye un resultado relevante en sí mismo. Más allá de la capacidad predictiva obtenida, el

trabajo aporta un marco metodológico para analizar la relación entre estructura del sistema eléctrico, predictibilidad y valor económico en mercados energéticos de corto plazo.

Apéndice A

Métodos de clusterización

Este anexo recoge las definiciones formales de los conceptos empleados en el capítulo 3: la distancia utilizada, el algoritmo K-means, los criterios de enlace jerárquico y las dos medidas empleadas para validar la solución (el índice de Rand ajustado y el procedimiento de estabilidad por remuestreo).

Distancia euclídea

Dados dos días representados como vectores $x, y \in \mathbb{R}^p$ en la matriz de agrupación ya estandarizada y ponderada, la distancia empleada en todo el capítulo es la distancia euclídea:

$$d(x, y) = \sqrt{\sum_{k=1}^{p} (x_k - y_k)^2}. \tag{A.1}$$

Algoritmo K-means

Fijado el número de clústeres K, el algoritmo procede de forma iterativa:

1. Se seleccionan K centroides iniciales.
2. Cada observación se asigna al centroide más próximo según la distancia euclídea.

3. Cada centroide se recalcula como el vector de medias de las observaciones asignadas a su clúster.

4. Los pasos 2 y 3 se repiten hasta que ninguna observación cambia de clúster (convergencia) o se alcanza un número máximo de iteraciones.

El resultado final depende de la asignación inicial de los centroides, ya que el algoritmo puede converger a un óptimo local. Para mitigar este riesgo, en este trabajo cada ajuste de K-means se repitió con 50 inicializaciones aleatorias distintas, conservando la solución con menor inercia, entendida como la suma de las distancias al cuadrado entre cada observación y el centroide de su clúster.

Criterios de enlace jerárquico

Sean A y B dos clústeres ya formados en el proceso aglomerativo. Los cuatro criterios de enlace considerados definen la distancia entre A y B de la siguiente forma:

Enlace simple Define la distancia entre dos grupos A y B como la distancia entre las dos observaciones más próximas de ambos grupos:

$$d(A, B) = \min_{x \in A,\, y \in B} d(x, y). \tag{A.2}$$

Enlace completo Define la distancia entre A y B como la distancia entre las dos observaciones más alejadas de ambos grupos:

$$d(A, B) = \max_{x \in A,\, y \in B} d(x, y). \tag{A.3}$$

Enlace medio Define la distancia entre A y B como el promedio de todas las distancias entre observaciones de ambos grupos:

$$d(A, B) = \frac{1}{|A|\,|B|} \sum_{x \in A} \sum_{y \in B} d(x, y). \tag{A.4}$$

Enlace de Ward Fusiona en cada paso los dos grupos cuya unión produce el menor incremento de la varianza interna total. Este incremento es proporcional a

$$\frac{|A|\,|B|}{|A| + |B|}\, d(c_A, c_B)^2, \tag{A.5}$$

donde c_A y c_B son los centroides de A y B.

Índice de Rand ajustado (ARI)

Dadas dos particiones de los mismos n días, el índice de Rand cuenta, sobre todos los pares posibles de observaciones, la proporción de pares para los que ambas particiones coinciden en agruparlos juntos o en separarlos. El índice de Rand ajustado corrige este valor por el acuerdo esperable por puro azar entre dos particiones aleatorias del mismo tamaño:

$$\mathrm{ARI} = \frac{\mathrm{RI} - \mathbb{E}[\mathrm{RI}]}{\mathrm{máx}(\mathrm{RI}) - \mathbb{E}[\mathrm{RI}]}, \tag{A.6}$$

donde RI es el índice de Rand sin ajustar [21]. Un valor de 1 indica coincidencia total entre las dos particiones; un valor próximo a 0, un acuerdo no mejor que el esperado por azar.

Procedimiento de estabilidad por remuestreo

Para evaluar la estabilidad de una solución de K clústeres se generaron 50 remuestreos, cada uno formado por el 80 % de los días elegidos sin reemplazo. En cada remuestreo se reajustó K-means de manera independiente y se calculó el ARI entre la partición obtenida en ese remuestreo y la partición correspondiente extraída de la solución global de referencia. La media y la desviación estándar de esos 50 valores de ARI son las que se reportan en la Figura 3.11.

Apéndice B

Hiperparámetros de los modelos

Este anexo describe el significado y el efecto de cada uno de los hiperparámetros ajustados mediante la búsqueda bayesiana con Optuna descrita en la sección 4.6, junto con el rango explorado durante la búsqueda y el valor final seleccionado para el regresor y para el clasificador.

`n_estimators` Número de árboles (iteraciones de *boosting*) que componen el modelo final. Un número mayor de árboles permite capturar relaciones más complejas, pero incrementa el riesgo de sobreajuste y el tiempo de cómputo; su efecto se compensa habitualmente con una tasa de aprendizaje más baja.

`learning_rate` Tasa de aprendizaje: factor por el que se multiplica la contribución de cada árbol nuevo antes de sumarla al modelo acumulado. Valores bajos requieren más árboles pero suelen generalizar mejor; valores altos aceleran el ajuste a costa de un mayor riesgo de sobreajuste.

`num_leaves` Número máximo de hojas (nodos terminales) por árbol. Controla la complejidad de cada árbol individual: más hojas permiten capturar interacciones más finas entre variables, pero aumentan el riesgo de sobreajuste. Este efecto es especialmente sensible en LightGBM, que crece los árboles hoja a hoja en lugar de nivel a nivel.

`min_child_samples` Número mínimo de observaciones requeridas en una hoja para que una división se acepte. Actúa como parámetro de regularización: valores altos evitan que el árbol cree hojas basadas en muy pocas observaciones (posible ruido), a costa de perder capacidad de ajuste fino.

`subsample` Fracción de las observaciones de entrenamiento, muestreada aleatoriamente, utilizada en cada iteración de *boosting*. Introduce aleatoriedad que reduce el sobreajuste y mejora la generalización, de forma análoga al *bagging*.

`colsample_bytree` Fracción de las variables explicativas, muestreada aleatoriamente, considerada al construir cada árbol. Al igual que `subsample`, introduce aleatoriedad adicional que actúa como regularización y reduce la correlación entre árboles sucesivos.

`reg_lambda` Coeficiente de regularización L_2 aplicado a los valores de predicción de las hojas. Penaliza las hojas con valores extremos, suavizando las predicciones del modelo y reduciendo el sobreajuste.

`alpha` **(solo regresor)** Parámetro δ de la función de pérdida de Huber, que fija el umbral a partir del cual el error deja de penalizarse de forma cuadrática y pasa a penalizarse de forma lineal. Valores bajos hacen que el modelo sea más robusto frente a errores grandes, acercándolo al comportamiento de un error absoluto medio; valores altos lo acercan al de un error cuadrático medio estándar.

`class_weight` **(solo clasificador)** Determina si las dos clases del problema reciben el mismo peso en la función de pérdida (`None`) o si se pondera inversamente a su frecuencia (`"balanced"`), dando más peso a la clase minoritaria. La búsqueda permitió elegir entre ambas opciones.

La Tabla B.1 resume el rango explorado durante la búsqueda y el valor final seleccionado para cada modelo. Ambas búsquedas compartieron el mismo espacio de exploración; solo difieren en la métrica objetivo (el PnL para el regresor, el AUC para el clasificador, como se explicó en la sección 4.6.2).

Dos observaciones sobre estos valores merecen comentario. En primer lugar, el número de árboles del regresor (1 200) coincide con el límite superior del rango explorado, lo que sugiere que una búsqueda con un rango más amplio podría, en principio, encontrar configuraciones ligeramente mejores; no se amplió el rango en este trabajo por motivos de coste computacional. En segundo lugar, el clasificador seleccionó `class_weight = None`, es decir, sin reponderar la clase minoritaria, un resultado coherente con el análisis de la sección 4.4.4 sobre el desequilibrio moderado de las clases del problema.

Tabla B.1: Rango de búsqueda y valor final de los hiperparámetros del regresor y del clasificador.

Hiperparámetro	Rango explorado	Valor final (regresor)	Valor final (clasificador)
`n_estimators`	$[200, 1200]$, paso 100	1 200	900
`learning_rate`	$[0{,}005, 0{,}15]$, log	0,057	0,084
`num_leaves`	$[15, 255]$	26	206
`min_child_samples`	$[10, 150]$	31	52
`subsample`	$[0{,}6, 1{,}0]$	0,730	0,740
`colsample_bytree`	$[0{,}6, 1{,}0]$	0,916	0,607
`reg_lambda`	$[0{,}0, 10{,}0]$	6,970	2,160
`alpha` (Huber)	$[0{,}5, 5{,}0]$	2,423	—
`class_weight`	{`balanced`, `None`}	—	`None`

Función de pérdida de Huber

El parámetro `alpha` del regresor corresponde al umbral δ de la función de pérdida de Huber, que combina el comportamiento cuadrático delerror cuadrático medio para errores pequeños con el comportamiento lineal del error absoluto medio para errores grandes:

$$L_\delta(r) = \begin{cases} \dfrac{1}{2}r^2 & \text{si } |r| \le \delta, \\ \delta\left(|r| - \dfrac{1}{2}\delta\right) & \text{si } |r| > \delta, \end{cases} \tag{B.1}$$

donde $r = s - \hat{s}$ es el residuo. Para $|r| \le \delta$ la pérdida coincide con la de un error cuadrático medio; para $|r| > \delta$ crece de forma lineal en lugar de cuadrática, lo que evita que un pequeño número de observaciones con residuos muy grandes (colas pesadas del *spread*) dominen el ajuste del modelo.

Mecanismo de la búsqueda bayesiana (TPE)

El algoritmo de estimación basada en árboles de Parzen (*Tree-structured Parzen Estimator*, TPE), implementado en Optuna, no evalúa el espacio de hiperparámetros de forma exhaustiva ni aleatoria. Tras un número inicial de pruebas aleatorias, TPE divide las configuraciones ya evaluadas en dos grupos según su rendimiento: las mejores (por encima de un percentil fijado) y el resto. Sobre cada grupo estima una densidad de probabilidad (mediante un estimador de Parzen, es

decir, una mezcla de distribuciones), y propone la siguiente configuración a evaluar maximizando el cociente entre la densidad del grupo de buenas configuraciones y la del grupo de malas. De esta forma, las pruebas se concentran progresivamente en las regiones del espacio de hiperparámetros que históricamente han producido mejores resultados, sin dejar de explorar puntualmente otras zonas.

Importancia de hiperparámetros (fANOVA)

El análisis funcional de la varianza (fANOVA) descompone la variación total del objetivo de búsqueda observada entre configuraciones en contribuciones atribuibles a cada hiperparámetro por separado y a sus interacciones. De forma simplificada, ajusta un modelo sustituto (un bosque aleatorio) sobre los resultados de todas las pruebas realizadas y calcula, para cada hiperparámetro, la varianza esperada del objetivo cuando dicho hiperparámetro se fija en un valor y el resto varían dentro de su rango de búsqueda. La proporción de la varianza total explicada por cada hiperparámetro es la importancia relativa que se representa en la Figura 4.4.

Apéndice C

Métricas de evaluación

Este anexo recoge, a modo de referencia, las definiciones formales de las métricas empleadas a lo largo del trabajo. Las métricas de regresión y de clasificación ya se introdujeron en las secciones 4.4.3 y 4.4.4; aquí se presentan de forma conjunta y se detalla, en particular, la diferencia entre exactitud (*accuracy*) y precisión (*precision*), dos términos que en el lenguaje coloquial se usan a menudo como sinónimos pero que en clasificación tienen significados distintos.

C.1. Métricas de clasificación: la matriz de confusión

Todas las métricas de clasificación de este trabajo se derivan de la matriz de confusión, que clasifica cada predicción según su relación con el valor observado:

- **Verdaderos positivos (VP):** horas con signo positivo observado que el modelo predijo correctamente como positivas.
- **Verdaderos negativos (VN):** horas con signo no positivo observado que el modelo predijo correctamente como no positivas.
- **Falsos positivos (FP):** horas con signo no positivo observado que el modelo predijo incorrectamente como positivas.
- **Falsos negativos (FN):** horas con signo positivo observado que el modelo predijo incorrectamente como no positivas.

A partir de estos cuatro recuentos se definen las métricas siguientes.

Exactitud (*accuracy*). Proporción de predicciones correctas sobre el total:

$$\text{Accuracy} = \frac{VP + VN}{VP + VN + FP + FN}. \tag{C.1}$$

Trata ambas clases por igual en términos absolutos, por lo que, como se explicó en la sección 4.4.4, puede resultar engañosa cuando una clase domina sobre la otra.

Precisión (*precision*). De entre todas las horas en las que el modelo predijo signo positivo, la proporción en las que acertó:

$$\text{Precision} = \frac{VP}{VP + FP}. \tag{C.2}$$

A diferencia de la exactitud, la precisión *no* tiene en cuenta los verdaderos negativos: no pregunta cuántas predicciones en total fueron correctas, sino, específicamente, cuán fiables son las predicciones positivas del modelo. Un modelo puede tener una exactitud alta y, al mismo tiempo, una precisión baja si acierta sobre todo prediciendo la clase negativa y falla con frecuencia cuando predice la clase positiva.

Exhaustividad o sensibilidad (*recall*). De entre todas las horas en las que el signo observado fue realmente positivo, la proporción que el modelo consiguió detectar:

$$\text{Recall} = \frac{VP}{VP + FN}. \tag{C.3}$$

Mientras que la precisión evalúa la fiabilidad de las predicciones positivas, la exhaustividad evalúa la capacidad del modelo para no dejar pasar por alto los casos positivos reales.

Exactitud balanceada. Media de la exhaustividad calculada por separado para cada clase:

$$\text{Bal. Acc.} = \frac{1}{2}\left(\frac{VP}{VP + FN} + \frac{VN}{VN + FP}\right). \tag{C.4}$$

Da el mismo peso a ambas clases con independencia de su frecuencia relativa, por lo que corrige la principal limitación de la exactitud simple ante clases desequilibradas.

Medida F1. Media armónica de la precisión y la exhaustividad:

$$F_1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{C.5}$$

Resulta especialmente útil cuando interesa un equilibrio entre ambas magnitudes y ninguna de las dos debe dominar la evaluación.

Área bajo la curva ROC (AUC). La curva ROC representa, para todos los umbrales de decisión posibles, la tasa de verdaderos positivos (Recall) frente a la tasa de falsos positivos ($FP/(FP + VN)$). El área bajo esa curva admite una interpretación probabilística directa: es la probabilidad de que el modelo asigne una puntuación más alta a una observación positiva elegida al azar que a una observación negativa elegida al azar. Un valor de 0,5 equivale a una ordenación aleatoria; un valor de 1, a una separación perfecta entre clases. A diferencia de las métricas anteriores, no depende de fijar un umbral de decisión concreto, lo que la hace especialmente adecuada para comparar la capacidad discriminativa de distintos modelos.

C.2. Métricas de regresión

Error absoluto medio (MAE).

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^{n} |s_i - \hat{s}_i| \,. \tag{C.6}$$

Expresado en las mismas unidades que la variable predicha (€/MWh), penaliza todos los errores de forma proporcional a su magnitud.

Raíz del error cuadrático medio (RMSE).

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (s_i - \hat{s}_i)^2}. \tag{C.7}$$

Al elevar los errores al cuadrado antes de promediarlos, penaliza en mayor medida los errores grandes que el MAE.

Coeficiente de determinación (R^2).

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (s_i - \hat{s}_i)^2}{\sum_{i=1}^{n} (s_i - \bar{s})^2}. \tag{C.8}$$

Mide qué proporción de la variabilidad de la variable objetivo consigue explicar el modelo, en comparación con la que quedaría sin explicar si se predijese siempre su valor medio. Un valor de 1 indica una predicción perfecta; un valor de 0, que el modelo no mejora la predicción constante de la media; un valor negativo, que el modelo predice peor que dicha media.

C.3. Métricas económicas

Beneficio acumulado (PnL). Suma, sobre todas las horas operadas, del producto entre el signo de la decisión adoptada y el spread realizado, para una posición normalizada de 1 MWh en cada operación (véase la sección 4.4.5).

Beneficio medio por operación. El PnL total dividido entre el número de operaciones realizadas, expresado en €/MWh; aproxima la rentabilidad unitaria de la estrategia.

Tasa de acierto (*hit rate*). Proporción de operaciones que finalizan con un resultado positivo, es decir, en las que el signo de la decisión coincide con el signo del spread realizado.

Apéndice D

Interpretabilidad y significación estadística

Este anexo formaliza las técnicas empleadas en el capítulo 5 para interpretar el modelo y para contrastar la significación estadística de sus resultados.

Importancia por permutación

Dado un modelo ya entrenado y una métrica de rendimiento M (por ejemplo, el PnL), la importancia por permutación de una variable, o de un grupo de variables, X_j se calcula en dos pasos. Primero se evalúa la métrica sobre el conjunto de prueba tal cual, obteniendo M_0. A continuación se reordena aleatoriamente (se permuta) los valores de X_j entre las observaciones, rompiendo su relación con el resto de variables y con la variable objetivo, y se recalcula la métrica sobre ese conjunto alterado, obteniendo M_{perm}. La importancia se define como

$$\text{Importancia}_j = M_0 - M_{\text{perm}}, \tag{D.1}$$

de modo que cuanto mayor sea el deterioro del rendimiento al perder la información de esa variable, mayor es su importancia [31].

Importancia por perturbación y curvas de dependencia parcial

La curva de dependencia parcial de una variable X_j se define como la predicción media del modelo cuando X_j se fija en un valor x y el resto de variables se dejan variar según su distribución empírica:

$$\mathrm{PD}_j(x) = \frac{1}{n}\sum_{i=1}^{n} f(x,\, X_{i,-j})\,, \tag{D.2}$$

donde $X_{i,-j}$ representa el vector de variables de la observación i excluyendo X_j. La importancia por perturbación mide la sensibilidad de esta curva: cuánto cambia la predicción media del modelo al recorrer X_j a lo largo de su rango, manteniendo el resto de variables en sus valores observados. A diferencia de la importancia por permutación, no mide una pérdida de rendimiento, sino la magnitud del efecto de la variable sobre la predicción.

Valores SHAP

Los valores SHAP (*SHapley Additive exPlanations*) trasladan al aprendizaje automático el concepto de valor de Shapley, procedente de la teoría de juegos cooperativos, que reparte de forma equitativa la contribución de cada jugador (aquí, cada variable) al resultado de un juego (aquí, la predicción) [32]. Formalmente, para una observación x con variables $\{1,\ldots,p\}$, el valor de Shapley de la variable j es

$$\phi_j = \sum_{S\subseteq\{1,\ldots,p\}\setminus\{j\}} \frac{|S|!\,(p-|S|-1)!}{p!}\Big[f\big(x_{S\cup\{j\}}\big) - f\big(x_S\big)\Big], \tag{D.3}$$

donde S recorre todos los subconjuntos posibles del resto de variables (todas las «coaliciones» posibles sin la variable j), y el término entre corchetes mide la contribución marginal de añadir la variable j a esa coalición. El valor ϕ_j es, por tanto, el promedio de esa contribución marginal sobre todos los posibles órdenes de incorporación de las variables. La propiedad clave de los valores SHAP es la aditividad: la suma de las contribuciones de todas las variables más un valor base (la predicción media del modelo) reproduce exactamente la predicción de la observación,

$$f(x) = \phi_0 + \sum_{j=1}^{p} \phi_j. \tag{D.4}$$

Dado que el cálculo exacto de esta suma es combinatoriamente costoso, en este trabajo se emplea TreeSHAP, un algoritmo específico para modelos basados en árboles (como LightGBM) que aprovecha su estructura para calcular los valores de Shapley exactos de forma eficiente.

Test de permutación del PnL

Bajo la hipótesis nula de que no existe una relación real entre las predicciones y los resultados observados, barajar aleatoriamente el emparejamiento entre ambos no debería alterar sistemáticamente el PnL obtenido. El procedimiento genera 2 000 barajados de esa relación, calcula el PnL resultante en cada uno para construir la distribución nula, y obtiene el p-valor como la proporción de barajados cuyo PnL iguala o supera al PnL real observado. Un p-valor pequeño indica que el resultado observado sería muy improbable si no existiese ninguna relación genuina entre predicción y resultado.

Bootstrap del intervalo de confianza

Para estimar el intervalo de confianza de la diferencia de PnL entre el modelo y la clase mayoritaria, se generan 2 000 remuestreos con reemplazo de las observaciones horarias emparejadas (predicción del modelo, predicción de la referencia, resultado real). En cada remuestreo se recalcula la diferencia de PnL entre ambas estrategias, y los percentiles 2,5 y 97,5 de la distribución resultante constituyen el intervalo de confianza del 95 %.

Apéndice E

Alineación con los Objetivos de Desarrollo Sostenible

Este anexo recoge una reflexión sobre la relación entre el presente Trabajo Fin de Grado y los Objetivos de Desarrollo Sostenible (ODS) de las Naciones Unidas. El proyecto se enmarca en el análisis de los mercados eléctricos y en el desarrollo de herramientas de aprendizaje automático para la toma de decisiones, por lo que su vínculo con los ODS es principalmente metodológico e indirecto: no aborda intervenciones físicas ni medioambientales directas, sino que aporta conocimiento y herramientas de análisis relevantes para el sector energético.

ODS 7: Energía asequible y no contaminante

El trabajo contribuye de forma indirecta al ODS 7 al estudiar el funcionamiento del mercado eléctrico ibérico en un periodo de fuerte transformación del mix de generación. La clusterización del capítulo 3 muestra, sin haber recibido ninguna etiqueta temporal, que los regímenes de mercado con alta penetración renovable (los veranos posteriores a la expansión fotovoltaica, los días de eólica abundante) presentan una dinámica de precios claramente distinta de los regímenes dominados por generación convencional, como la crisis del gas de 2022. Esta evidencia empírica ilustra de manera concreta cómo la incorporación de energías renovables está alterando la formación de precios en el sistema eléctrico, información relevante para comprender la transición hacia un sistema energético más limpio y asequible. Asimismo, la herramienta predictiva desarrollada, orientada a anticipar el comportamiento del *spread* entre mercados, puede apoyar decisiones económicas más

informadas por parte de los agentes que participan en dicha transición.

ODS 9: Industria, innovación e infraestructura

El proyecto aplica técnicas de aprendizaje automático (clusterización no supervisada y *gradient boosting*) a un problema real y poco explorado del sector energético, contribuyendo a la innovación en el uso de herramientas digitales y analíticas dentro de la industria eléctrica. Más allá de la aplicación concreta, el trabajo pone un énfasis particular en una metodología de evaluación rigurosa y honesta: validación estrictamente temporal, comparación con referencias exigentes, contrastes de significación estadística y reconocimiento explícito de las limitaciones y de las fuentes de incertidumbre de los resultados. Esta forma de proceder, orientada a evitar conclusiones artificialmente optimistas, resulta especialmente relevante en un contexto de creciente adopción de la inteligencia artificial en infraestructuras críticas como los mercados energéticos, y se alinea con los objetivos de modernización responsable de las infraestructuras de información que persigue el ODS 9.

ODS 13: Acción por el clima

El proyecto guarda relación con el ODS 13 en la medida en que se enmarca en el análisis de un sistema eléctrico en proceso de descarbonización. Uno de los resultados más relevantes del trabajo es que la capacidad predictiva del modelo se concentra precisamente en los regímenes de mayor penetración renovable (las horas centrales del día, coincidiendo con la producción fotovoltaica, y los días de generación eólica elevada), donde el comportamiento tradicional del mercado se altera de forma significativa. Esta evidencia aporta información útil sobre cómo la integración creciente de renovables está modificando la dinámica de los mercados de corto plazo, un conocimiento relevante para operadores del sistema y reguladores en la gestión de la variabilidad asociada a estas tecnologías. Aunque el proyecto no aborda de forma directa aspectos medioambientales, el estudio de herramientas que mejoran la comprensión el comportamiento de los precios resulta relevante para optimizar la integración de las fuentes renovables y apoyar la sostenibilidad económica de dicha transición.

Síntesis

En conjunto, la contribución de este trabajo a los ODS es de naturaleza metodológica y de conocimiento: no propone una intervención directa sobre el sistema energético, sino herramientas y evidencia empírica que ayudan a comprender mejor su funcionamiento en un contexto de transición hacia un modelo más renovable, asequible y sostenible.

Bibliografía

[1] OMIE. *Detalle del funcionamiento del mercado intradiario*. `https://www.omie.es/sites/default/files/inline-files/mercados_intradiario_y_continuo.pdf`. Dirección de Operación del Mercado. Consultado en 2026. Operador del Mercado Ibérico de Energía – Polo Español, S.A.

[2] Aitor Ciarreta, Cristina Pizarro-Irizar y Ainhoa Zarraga. «Renewable energy regulation and structural breaks: An empirical analysis of Spanish electricity price volatility». En: *Energy Economics* 88 (2020), pág. 104749.

[3] Marek Pavlík, František Kurimskỳ y Kamil Ševc. «Renewable Energy and Price Stability: An Analysis of Volatility and Market Shifts in the European Electricity Sector (2015–2025).» En: *Applied Sciences (2076-3417)* 15.12 (2025).

[4] Alessandro Sapio. «Greener, more integrated, and less volatile? A quantile regression analysis of Italian wholesale electricity prices». En: *Energy Policy* 126 (2019), págs. 452-469.

[5] Evangelos Kyritsis, Jonas Andersson y Apostolos Serletis. «Electricity prices, large-scale renewable integration, and policy implications». En: *Energy Policy* 101 (2017), págs. 550-560.

[6] Amelie Schischke y Andreas Rathgeber. «Unraveling European Electricity Price Volatility: The Impact of Renewables». En: *Available at SSRN 5549982* (2025).

[7] Arkadiusz Jędrzejewski et al. «Electricity price forecasting: The dawn of machine learning». En: *IEEE Power and Energy Magazine* 20.3 (2022), págs. 24-31.

[8] Alexiei Dingli y Karl Sant Fournier. «Financial time series forecasting-a machine learning approach». En: *Machine Learning and Applications: An International Journal* 4.1/2 (2017), pág. 3.

[9] Bhupesh Kumar Mishra et al. «Machine learning and deep learning prediction models for time-series: a comparative analytical study for the use case of the UK short-term electricity price prediction». En: *Discover Internet of Things* 4.1 (2024), pág. 24.

[10] Guzmán Díaz, José Coto y Javier Gómez-Aleixandre. «Prediction and explanation of the formation of the Spanish day-ahead electricity price through machine learning regression». En: *Applied Energy* 239 (2019), págs. 610-625.

[11] Hongju Yan y Hongbing Ouyang. «Financial time series prediction based on deep learning». En: *Wireless Personal Communications* 102.2 (2018), págs. 683-700.

[12] Sidra Mehtab y Jaydip Sen. «A time series analysis-based stock price prediction using machine learning and deep learning models». En: *International journal of business forecasting and marketing intelligence* 6.4 (2020), págs. 272-335.

[13] Xuesong Wang et al. «Deep learning-based electricity price forecast for virtual bidding in wholesale electricity market». En: *arXiv preprint arXiv:2412.00062* (2024).

[14] Vasilis Michalakopoulos et al. «Data-driven Day Ahead Market Prices Forecasting: A Focus on Short Training Set Windows». En: *arXiv preprint arXiv:2506.10536* (2025).

[15] Ciaran O'Connor et al. «A Review of Electricity Price Forecasting Models in the Day-Ahead, Intra-Day, and Balancing Markets». En: *Energies* 18.12 (2025), pág. 3097.

[16] Gareth James et al. *An Introduction to Statistical Learning: with Applications in Python*. Springer, 2023.

[17] Joe H. Ward. «Hierarchical grouping to optimize an objective function». En: *Journal of the American Statistical Association* 58.301 (1963), págs. 236-244.

[18] Peter J. Rousseeuw. «Silhouettes: A graphical aid to the interpretation and validation of cluster analysis». En: *Journal of Computational and Applied Mathematics* 20 (1987), págs. 53-65.

[19] Tadeusz Caliński y Jerzy Harabasz. «A dendrite method for cluster analysis». En: *Communications in Statistics* 3.1 (1974), págs. 1-27.

[20] David L. Davies y Donald W. Bouldin. «A cluster separation measure». En: *IEEE Transactions on Pattern Analysis and Machine Intelligence* PAMI-1.2 (1979), págs. 224-227.

[21] Lawrence Hubert y Phipps Arabie. «Comparing partitions». En: *Journal of Classification* 2.1 (1985), págs. 193-218.

[22] Manu Joseph y Jeff Tackes. *Modern Time Series Forecasting with Python: Explore Industry-Ready Time Series Forecasting Using Modern Machine Learning and Deep Learning*. 2.ª ed. Packt Publishing, 2024.

[23] Ben Auffarth. *Machine Learning for Time-Series with Python: Forecast, Predict, and Detect Anomalies with State-of-the-Art Machine Learning Methods*. Packt Publishing, 2021.

[24] Jerome H. Friedman. «Greedy function approximation: A gradient boosting machine». En: *The Annals of Statistics* 29.5 (2001), págs. 1189-1232.

[25] Guolin Ke et al. «LightGBM: A highly efficient gradient boosting decision tree». En: *Advances in Neural Information Processing Systems*. Vol. 30. 2017, págs. 3146-3154.

[26] Aurélien Géron. *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. 3.ª ed. O'Reilly Media, 2022.

[27] Peter J. Huber. «Robust estimation of a location parameter». En: *The Annals of Mathematical Statistics* 35.1 (1964), págs. 73-101.

[28] James Bergstra et al. «Algorithms for hyper-parameter optimization». En: *Advances in Neural Information Processing Systems*. Vol. 24. 2011.

[29] Takuya Akiba et al. «Optuna: A next-generation hyperparameter optimization framework». En: *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2019, págs. 2623-2631.

[30] Frank Hutter, Holger Hoos y Kevin Leyton-Brown. «An efficient approach for assessing hyperparameter importance». En: *Proceedings of the 31st International Conference on Machine Learning*. 2014, págs. 754-762.

[31] Aaron Fisher, Cynthia Rudin y Francesca Dominici. «All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously». En: *Journal of Machine Learning Research* 20.177 (2019), págs. 1-81.

[32] Scott M. Lundberg y Su-In Lee. «A unified approach to interpreting model predictions». En: *Advances in Neural Information Processing Systems*. Vol. 30. 2017, págs. 4765-4774.