COMILLAS
UNIVERSIDAD PONTIFICIA
ICAI

GRADO EN INGENIERÍA EN INGENIERÍA EN TECNOLOGÍAS INDUSTRIALES

TRABAJO DE FIN DE GRADO

APLICACIÓN DE TÉCNICAS DE CLÚSTERIZACIÓN Y *MACHINE LEARNING* PARA EL ESTUDIO DE LOS PERFILES DE DESVÍO DE LA RED ELÉCTRICA ESPAÑOLA

Autor: Conde Cortizas, Juan

Director: Saz-Orozco Huang, Pablo Carlos del

Entidad Colaboradora: ICAI – Universidad Pontificia Comillas

Firma Autor: Visto bueno director:

Juan Conde Cortizas **Sí**/No

Madrid

Abril de 2025

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

Aplicación de técnicas de clusterización y *machine learning* para el estudio de los perfiles de desvío de la Red Eléctrica Española

en la ETS de Ingeniería - ICAI de la Universidad Pontificia Comillas en el

curso académico 2024/25 es de mi autoría, original e inédito y

no ha sido presentado con anterioridad a otros efectos.

El Proyecto no es plagio de otro, ni total ni parcialmente y la información que ha sido

tomada de otros documentos está debidamente referenciada.

Fdo.: Juan Conde Cortizas Fecha: 24/08/2025

Juan Conde Cortizas

Autorizada la entrega del proyecto

Pablo Carlos del Saz-Orozco Huang

Fdo.: Pablo Carlos del Saz-Orozco Huang Fecha: 26/ 08/ 2025

DEL SAZ OROZCO HUANG PABLO CARLOS - 44578440R

Firmado digitalmente por DEL SAZ OROZCO HUANG PABLO CARLOS - 44578440R
Fecha: 2025.08.26 11:27:13 +02'00'

APLICACIÓN DE TÉCNICAS DE CLÚSTERIZACIÓN Y APRENDIZAJE PROFUNDO PARA EL ESTUDIO DE LOS DESVÍOS DE LA RED ELÉCTRICA ESPAÑOLA

Autor: Conde Cortizas, Juan

Director: Saz-Orozco Huang, Pablo Carlos del

Entidad Colaboradora: ICAI – Universidad Pontificia Comillas

Resumen del Proyecto

El presente proyecto se centra en la detección y caracterización de patrones de desvío que permitirían optimizar la estrategia de compra para comercializadoras españolas a través del análisis de desvíos entre la demanda prevista y la demanda real, aplicando técnicas de inteligencia artificial como K-means (aprendizaje no supervisado) y Random Forest (aprendizaje supervisado). El objetivo fundamental es identificar patrones en los datos históricos que expliquen los desvíos y, a partir de ellos, proponer estrategias económicas que permitan a una comercializadora reducir costes asociados a dichos desvíos. Este trabajo no pretende desarrollar un modelo predictivo convencional, sino un enfoque aplicado que aporte conclusiones prácticas al mercado eléctrico.

Palabras Clave: Sistema eléctrico, desvíos, aprendizaje automático, K-means, Random Forest, estrategias económicas, comercializadora.

1. Introducción

En los últimos años, el sistema eléctrico español ha experimentado un notable incremento de la complejidad debido a la integración masiva de fuentes renovables variables y a la mayor volatilidad en la demanda. Este escenario exige herramientas avanzadas que permitan comprender mejor las causas de los desvíos diarios. Dichos desvíos, definidos como la diferencia entre la demanda prevista y la real [REE25], suponen un coste económico importante, ya que deben ser compensados mediante servicios de ajuste. Actualmente, Red Eléctrica de España (REE) utiliza modelos predictivos avanzados, principalmente basados en aprendizaje supervisado y deep learning [VARA24], con errores inferiores al 5%. Sin embargo, aún existe un margen de mejora en la interpretación de los patrones de los desvíos, ámbito en el cual se enmarca este proyecto.

2. Estado de la cuestión

La literatura especializada ha abordado el problema de los desvíos eléctricos desde diferentes perspectivas. Por un lado, se encuentran los modelos estadísticos clásicos como ARIMA [KONG19], que ofrecen simplicidad e interpretabilidad, pero con limitaciones frente a patrones no lineales complejos. Por otro lado, el auge de las técnicas de machine learning y deep learning ha permitido entrenar modelos de predicción más precisos y

flexibles. Random Forest, en particular, se ha consolidado como una herramienta robusta para la clasificación y predicción en contextos energéticos, gracias a su capacidad de manejar gran cantidad de variables heterogéneas [JAME23]. En paralelo, técnicas de clústerización como K-means se han empleado para segmentar perfiles de consumo o generación, permitiendo identificar patrones comunes en los datos. Sin embargo, la aplicación de estas técnicas no supervisadas para el diseño de estrategias económicas de una comercializadora sigue siendo incipiente [BRUN24], lo que justifica la originalidad de este proyecto.

3. Definición del proyecto

El objetivo principal del proyecto es doble. En primer lugar, identificar y caracterizar perfiles de desvíos eléctricos a través de técnicas de clústerización, utilizando únicamente variables conocidas de antemano (como la temperatura, el tipo de día o la potencia programada). En segundo lugar, evaluar estrategias económicas basadas en dichos perfiles para optimizar la compra de energía de una comercializadora. La hipótesis central es que un mejor entendimiento de los patrones de desvío permitirá diseñar estrategias que reduzcan los costes de ajuste y aporten ventajas competitivas en el mercado. La originalidad radica en integrar un análisis descriptivo con un análisis aplicado económico, generando resultados de impacto directo en la rentabilidad de una comercializadora mediana.

4. Descripción del modelo y metodología

La metodología seguida se estructura en cinco etapas principales. Primero, se recopilaron los datos históricos de demanda, generación prevista y potencia instalada de Red Eléctrica de España (REE), así como información meteorológica de AEMET. En segundo lugar, se aplicó un preprocesamiento exhaustivo para homogeneizar formatos, tratar valores faltantes y normalizar variables. En la tercera fase, se empleó el algoritmo K-means para agrupar días con características similares, obteniendo diferentes perfiles de desvío. Más adelante, los clústeres resultantes fueron validados mediante Random Forest, alcanzando una precisión del 92%, lo que demuestra la solidez de los grupos identificados. Por último, se desarrollaron unas estrategias adecuadas a los datos para intentar aprovechar el perfil de desvío visto en la fase 3 para poder explotarlo y conseguir así un impacto económico positivo.

5. Resultados

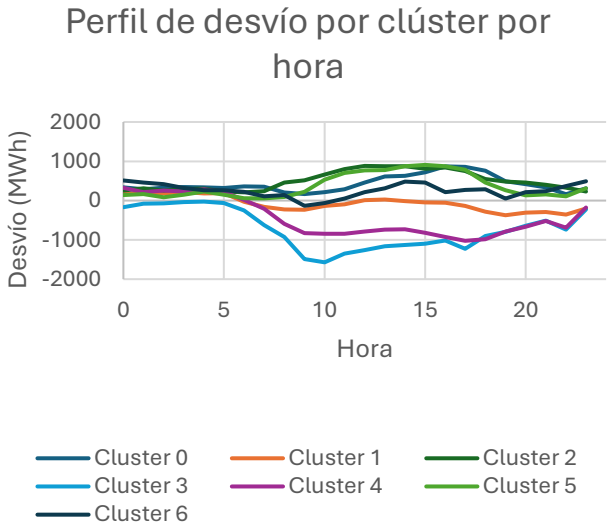
Los resultados del análisis muestran que la clústerización permitió identificar perfiles de días claramente diferenciados, estos fueron denominados como “clúster” y sus características fueron estudiadas y utilizadas para entrenar el modelo basado en Random Forest correspondiente a la segunda parte del trabajo. Un ejemplo de las características de un clúster sería:

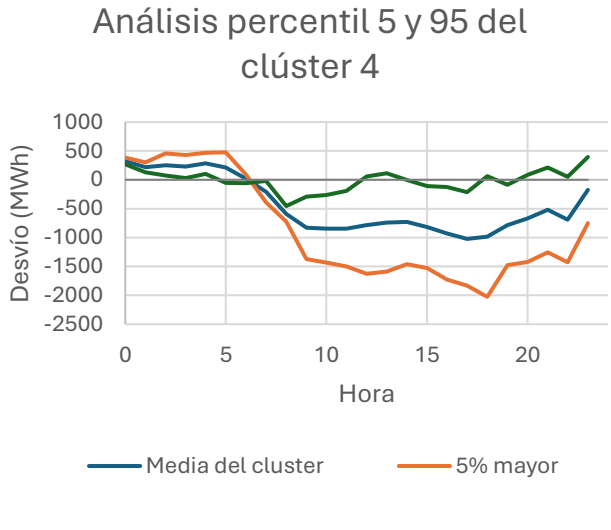
Tabla 1: Desviaciones de las variables del clúster 0

Variable	Madrugada (0-5)	Mañana (6-11)	Tarde (12-17)	Noche (18-23)
Prevision_Foto	0,0%	-6,3%	-13,8%	28,1%
Eólica_Diaria	-25,1%	-37,6%	-32,6%	-19,5%
Instalada_Foto	-35,6%	-35,6%	-35,6%	-35,6%
Instalada_Eolica	-6,5%	-6,5%	-6,5%	-6,5%
Prev_Demanda	5,8%	2,2%	8,5%	2,4%
Disp_Nuclear	8,6%	8,7%	8,6%	8,5%
Festividad	-54,1%	-54,1%	-54,1%	-54,1%
Lunes	-37,8%	-37,8%	-37,8%	-37,8%
MartesJueves	-99,0%	-99,0%	-99,0%	-99,0%
ViernesDomingoFestivo	74,9%	74,9%	74,9%	74,9%
Temp_Madrid	61,0%	58,1%	49,1%	50,2%
Temp_Tenerife	11,9%	8,3%	6,8%	8,7%
Temp_Sevilla	44,8%	38,4%	33,9%	38,0%
Temp_Santiago	35,8%	33,3%	29,3%	30,5%
Temp_Coruña	23,5%	19,8%	21,0%	23,7%
Temp_Zaragoza	57,8%	48,6%	43,2%	49,2%

Como se puede observar, existen características que guían la tendencia de un clúster hacia un tipo de día u otro, lo cuál demuestra un hallazgo en sí mismo, pero lo que verdaderamente aportó a este trabajo fueron el análisis de los desvíos de estos clústeres, siguiendo estos, en muchos casos, una tendencia común:

Ilustración 1: Perfil de desvío por clúster por hora e Ilustración 2: Análisis percentil 5 y 95 del clúster 4





Algunos clústeres presentan desvíos positivos recurrentes en horas de la mañana, mientras que otros reflejan desvíos negativos concentrados en horas de alta generación renovable. Estos hallazgos revelan la influencia de factores como la estacionalidad, los días festivos o la penetración de tecnologías renovables en la red. La validación con Random Forest confirmó que los clústeres son coherentes y reproducibles en años distintos, alcanzando un 92% de acierto en la clasificación.

Tabla 2: Matriz de confusión del modelo de Random Forest

		Predicho						
		0	**1**	**2**	**3**	**4**	**5**	**6**
Clúster real	**0**	0	0	0	0	0	0	0
	1	0	117	0	0	0	0	0
	2	0	0	0	0	0	0	0
	3	0	0	0	114	0	0	24
	4	0	0	0	0	60	0	0
	5	0	0	0	0	0	0	0
	6	0	0	0	4	0	0	47

Donde se puede observar donde la precisión es muy alta salvo para el clúster 6. Esto es debido a las características de este clúster y lo parecidas que son al número 3, diferenciándose en muy pocas y causando esta confusión. Además, el hecho de que no aparezcan clústeres como el 0,2 o 5 es una señal del sesgo temporal de los datos, puesto que variables como la potencia instalada eólica y fotovoltaica son las que caracterizaban dichos clústeres y estas aumentan inherentemente con el tiempo. Sesgo que se ha demostrado de poca significación para el desarrollo de este proyecto.

En la segunda parte del proyecto se diseñaron dos estrategias económicas fundamentadas en los perfiles de desvío previamente identificados. Con el fin de garantizar que estas estrategias dependieran únicamente del análisis realizado y no de los desvíos concretos observados en determinados días, se asumió que los desvíos de una comercializadora puntual podían modelarse como una variable aleatoria con distribución normal centrada en cero. Esta hipótesis resulta razonable, ya que una comercializadora mediana, al atender a un conjunto limitado de clientes, no debería presentar sesgos sistemáticos en sus desvíos, sino más bien fluctuaciones aleatorias en torno a cero.

Sobre esta base se plantearon dos aproximaciones. La **Estrategia 1** consiste en aplicar un ajuste moderado sobre la energía comprada en el mercado, anticipando los desvíos más probables según el clúster al que pertenezca el día. Se trata de una estrategia conservadora que busca minimizar penalizaciones sin alterar en exceso el perfil de compra original. La **Estrategia 2**, en cambio, es más agresiva y se activa únicamente cuando el nivel de confianza de la predicción es elevado. En estas circunstancias, la comercializadora aprovecha al máximo las señales de precio y desvío, comprando más o menos energía de forma deliberada en las horas críticas, con el objetivo de obtener un

beneficio económico mayor. Ambas estrategias fueron evaluadas en términos económicos, comparando el coste total de adquisición de energía con y sin su aplicación. Para comprobar la robustez de los resultados y descartar la influencia de valores extremos, se realizó además un análisis de variabilidad, eliminando progresivamente las horas con precios de desvío más elevados. Los resultados fueron los siguientes:

Tabla 3: Análisis de Variabilidad

	Coste Original	Coste Con estrategia 1	Coste con estrategia 2
Caso Base	1.514.024,41 €	1.303.835,86 €	1.191.019,92 €
Diferencia (% Caso Base)	**0,0%**	**-13,9%**	**-21,3%**
Eliminado el 1% extremos	1.443.368,72 €	1.253.294,24 €	1.175.369,74 €
Diferencia (% Caso Base)	**-4,7%**	**-13,2%**	**-18,6%**
Eliminado el 3% extremos	1.392.914,76 €	1.219.944,29 €	1.161.756,21 €
Diferencia (% Caso Base)	**-8,0%**	**-12,4%**	**-16,6%**
Eliminado el 5% extremos	1.326.641,45 €	1.176.072,72 €	1.152.182,98 €
Diferencia (% Caso Base)	**-12,4%**	**-11,3%**	**-13,2%**
Eliminado el 10% extremos	1.183.915,82 €	1.068.565,70 €	1.114.074,64 €
Diferencia (% Caso Base)	**-21,8%**	**-9,7%**	**-5,9%**

Los resultados muestran que la Estrategia 1 aporta una reducción de costes del 14% de media, mientras que la Estrategia 2 puede alcanzar ahorros superiores al 20%, aunque con mayor dependencia de valores extremos. Además, se puede observar que, incluso al suprimir los valores más desfavorables para el caso base (que podrían haber resultado ventajosos para las estrategias), el impacto económico de ambas seguía siendo positivo, aunque se reduce sustancialmente para la Estrategia 2. Estos hallazgos confirman que las estrategias propuestas no dependen únicamente de situaciones extremas del mercado, sino que presentan un efecto consistente de reducción de costes bajo distintas condiciones.

6. Conclusiones

Las conclusiones principales de este proyecto confirman que las técnicas de inteligencia artificial aplicadas al análisis de desvíos no solo permiten caracterizar patrones ocultos en los datos históricos, sino también trasladar dicho conocimiento a la práctica económica. El impacto económico estimado para una comercializadora mediana se sitúa entre un 14% y un 22% de ahorro en los costes de compra de energía, lo que supone un ahorro de más de 150.000 euros en el escenario simulado. Estos resultados demuestran la utilidad práctica del enfoque, que combina análisis descriptivo y supervisado con una aplicación económica concreta.

La Estrategia 1 se muestra más consistente y robusta frente a la variabilidad de los datos, aunque con menor rentabilidad en escenarios típicos, mientras que la Estrategia 2 ofrece retornos más altos, pero principalmente gracias a su aprovechamiento de valores extremos. En conjunto, ambas aproximaciones abren la puerta a combinaciones híbridas que podrían mejorar la estabilidad y la rentabilidad simultáneamente.

Finalmente, este trabajo confirma que el análisis de desvíos no debe limitarse al plano técnico de REE, sino que puede tener un impacto directo en la rentabilidad de las comercializadoras. Sin embargo, también se advierte que un uso intensivo de estas estrategias podría llamar la atención de los reguladores, como la CNMC, que establece derechos y obligaciones en los servicios de ajuste del sistema [CNMC22]. Esto subraya la necesidad de equilibrar la optimización económica con la estabilidad y sostenibilidad del sistema eléctrico en su conjunto.

7. Referencias

A continuación, se presentan las referencias más representativas y utilizadas a lo largo de la memoria:

[VARA24] Varas, A. (2024). *Modelos de predicción de la demanda en sistemas eléctricos mediante aprendizaje profundo*. Revista de Energía y Regulación, 15(2), pp. 45–61.

[MANR20] Manrique, L. (2020). *Introducción al aprendizaje supervisado y no supervisado en Machine Learning*. Editorial UPM, Madrid.

[FRAN18] Franco, D. (2018). *Machine Learning básico: conceptos y aplicaciones*. Barcelona: Ediciones UPC.

[GONZ18] González, J. (2018). *Deep Learning: Fundamentos y aplicaciones en ingeniería*. McGraw Hill Interamericana, Madrid.

[KONG19] Kong, W., Dong, Z. Y., Jia, Y., Hill, D. J., Xu, Y., & Zhang, Y. (2019). "Short-term residential load forecasting based on LSTM recurrent neural network". *IEEE Transactions on Smart Grid*, 10(1), pp. 841–851.

[MARI16] Mariano, J., & Ruiz, F. (2016). *Modelos avanzados de predicción energética: de ARIMA a redes neuronales*. Editorial UNED, Madrid.

[JAME23] James, G., Witten, D., Hastie, T., Tibshirani, R., & Grosse, R. (2023). *An Introduction to Statistical Learning: with Applications in Python*. Springer Texts in Statistics.

[CNMC22] Comisión Nacional de los Mercados y la Competencia (2022). *Derechos de cobro y obligaciones de pago por los servicios de ajuste del sistema eléctrico*. Informe técnico de supervisión.

[BRUN24] Bruneel, J., Van der Elst, S., & Meeus, L. (2024). "Managing imbalance price risk with intraday trading strategies in single-price electricity markets". *Energy Economics*, 126, 106857.

APPLICATION OF CLUSTERING AND DEEP LEARNING TECHNIQUES FOR THE STUDY OF SPANISH POWER GRID DEVIATIONS

Author: Conde Cortizas, Juan

Principal: Saz-Orozco Huang, Pablo Carlos del

Collaborating Institution: ICAI - Comillas Pontifical University

Project Summary

This project focuses on the optimization of the operation of the Spanish electricity system through the analysis of deviations between the forecasted demand and the real demand, applying artificial intelligence techniques such as K-means (unsupervised learning) and Random Forest (supervised learning). The main objective is to identify patterns in the historical data that explain the deviations and based on them, to propose economic strategies that allow a marketer to reduce costs associated with these deviations. This work does not intend to develop a conventional predictive model, but an explanatory and applied approach that brings practical conclusions to the electricity market.

Keywords: Electricity system, deviations, machine learning, K-means, Random Forest, economic strategies, trading company.

1. Introduction

In recent years, the Spanish electricity system has experienced a remarkable increase in complexity due to the massive integration of variable renewable sources and to the higher volatility in demand. This scenario requires advanced tools to better understand the causes of daily deviations. Such deviations, defined as the difference between forecasted and actual demand [REE25], entail a significant economic cost, as they must be compensated through adjustment services. Currently, Red Eléctrica de España (REE) uses advanced predictive models, mainly based on supervised learning and deep learning [VARA24], with errors of less than 5%. However, there is still room for improvement in the interpretation of the causes of deviations, which is the scope of this project.

2. State of the art

The specialized literature has approached the problem of electrical deviations from different perspectives. On the one hand, there are the classical statistical models such as ARIMA [KONG19], which offer simplicity and interpretability, but with limitations in the face of complex nonlinear patterns. On the other hand, the rise of machine learning and deep learning techniques has made it possible to train more accurate prediction models and flexible. Random Forest has established itself as a robust tool for classification and prediction in energy contexts, thanks to its ability to handle a large number of heterogeneous variables [JAME23]. In parallel, clustering techniques such as K-means have been used to segment consumption or generation profiles, allowing the identification of common patterns in the data. However, the application of these unsupervised techniques

for the design of economic strategies of a marketer is still incipient [BRUN24], which justifies the originality of this project.

3. Project definition

The main objective of the project is twofold. First, to identify and characterize electrical deviation profiles through clustering techniques, using only variables known in advance (such as temperature, type of day or programmed power). Secondly, to evaluate economic strategies based on these profiles to optimize the purchase of energy from a retailer. The central hypothesis is that a better understanding of deviation patterns will allow the design of strategies that reduce adjustment costs and provide competitive advantages in the market. The originality lies in integrating a descriptive analysis with an applied economic analysis, generating results that have a direct impact on the profitability of a medium-sized retailer.

4. Description of the model and methodology

The methodology followed is structured in five main stages. First, historical data on demand, forecasted generation and installed power were collected from Red Eléctrica de España (REE), as well as meteorological information from AEMET. Second, exhaustive preprocessing was applied to homogenize formats, treat missing values, and normalize variables. In the third phase, the K-means algorithm was used to group days with similar characteristics, obtaining different deviation profiles. Later, the resulting clusters were validated by Random Forest, reaching an accuracy of 92%, which demonstrates the robustness of the identified groups. Finally, strategies appropriate to the data were developed to try to take advantage of the drift profile seen in phase 3 in order to exploit it and thus achieve a positive economic impact.

5. Results

The results of the analysis show that the clustering allowed the identification of clearly differentiated profiles of days, these were denominated as "clusters", and their characteristics were studied and used to train the Random Forest based model corresponding to the second part of the work. An example of the characteristics of a cluster would be:

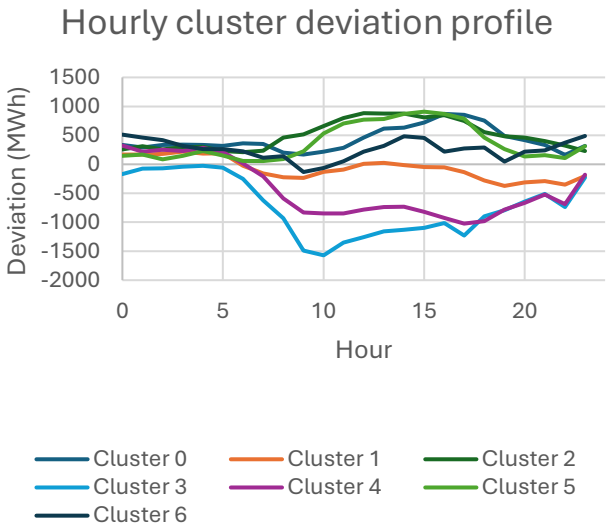
Table1 : Deviations of cluster 0 variables.

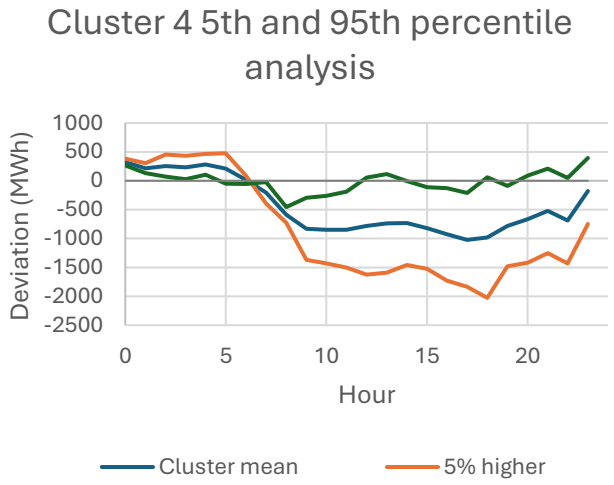
Variable	Morning (0-5)	Morning (6-11)	Afternoon (12-17)	Evening (18-23)
Forecast_Photo	0,0%	-6,3%	-13,8%	28,1%
Wind_Daily_Wind	-25,1%	-37,6%	-32,6%	-19,5%
Installed_Photo	-35,6%	-35,6%	-35,6%	-35,6%
Installed_Wind	-6,5%	-6,5%	-6,5%	-6,5%
Prev_Demand	5,8%	2,2%	8,5%	2,4%
Disp_Nuclear	8,6%	8,7%	8,6%	8,5%
Holidays	-54,1%	-54,1%	-54,1%	-54,1%
Monday	-37,8%	-37,8%	-37,8%	-37,8%

TuesdayThursday	-99,0%	-99,0%	-99,0%	-99,0%
FridaySundayHoliday	74,9%	74,9%	74,9%	74,9%
Temp_Madrid	61,0%	58,1%	49,1%	50,2%
Temp_Tenerife	11,9%	8,3%	6,8%	8,7%
Temp_Sevilla	44,8%	38,4%	33,9%	38,0%
Temp_Santiago	35,8%	33,3%	29,3%	30,5%
Temp_Coruña	23,5%	19,8%	21,0%	23,7%
Temp_Zaragoza	57,8%	48,6%	43,2%	49,2%

As can be seen, there are characteristics that guide the tendency of a cluster towards one type of day or another, which is a finding in itself, but what really contributed to this work was the analysis of the deviations of these clusters, which in many cases follow a common trend:

Illustration3 : Deviation profile per cluster per hour and Illustration4 : Cluster 4 5th and 95th percentile analysis.





Some clusters show recurring positive deviations in the morning hours, while others reflect negative deviations concentrated in hours of high renewable generation. These findings reveal the influence of factors such as seasonality, holidays, or the penetration of renewable technologies in the grid. Validation with Random Forest confirmed that the clusters are consistent and reproducible in different years, reaching 92% classification accuracy.

Table2 : Confusion matrix of the Random Forest model

		Predicted						
		0	**1**	**2**	**3**	**4**	**5**	**6**
Actual cluster	**0**	0	0	0	0	0	0	0
	1	0	117	0	0	0	0	0
	2	0	0	0	0	0	0	0
	3	0	0	0	114	0	0	24
	4	0	0	0	0	60	0	0

	5	0	0	0	0	0	0	0
	6	0	0	0	4	0	0	47

Where it can be seen where the accuracy is very high except for cluster 6. This is due to the characteristics of this cluster and how similar they are to number 3, differing in very few and causing this confusion. In addition, the fact that clusters such as 0,2 or 5 do not appear is a sign of the temporal bias of the data, since variables such as wind and PV installed power are the ones that characterized these clusters, and these inherently increase over time. This bias has proven to be of little significance for the development of this project.

In the second part of the project, two economic strategies were designed based on the previously identified diversion profiles. In order to ensure that these strategies depended only on the analysis performed and not on the specific deviations observed on certain days, it was assumed that the deviations of a point trader could be modelled as a random variable with a normal distribution cantered at zero. This assumption is reasonable, since a medium-sized marketer, serving a limited set of customers, should not have systematic biases in its deviations, but rather random fluctuations around zero.

On this basis, two approaches were considered. **Strategy 1** consists of applying a moderate adjustment on the energy purchased in the market, anticipating the most probable deviations according to the cluster to which the day belongs. This is a conservative strategy that seeks to minimize penalties without excessively altering the original purchase profile. **Strategy 2**, on the other hand, is more aggressive and is activated only when the confidence level of the prediction is high. In these circumstances, the trader takes full advantage of the price and deviation signals, deliberately buying energy at critical hours, with the aim of obtaining a higher economic benefit. Both strategies were evaluated in economic terms, comparing the total cost of energy procurement with and without their application. To check the robustness of the results and rule out the influence of extreme values, a variability analysis was also performed, progressively eliminating the hours with higher deviation prices. The results were as follows:

Table3 : Variability Analysis

	Original Cost	Cost With strategy 1	Cost with strategy 2
Base Case	1.514.024,41 €	1.303.835,86 €	1.191.019,92 €
Difference (% Base Case)	**0,0%**	**-13,9%**	**-21,3%**
Eliminating 1% extremes	1.443.368,72 €	1.253.294,24 €	1.175.369,74 €
Difference (% Base Case)	**-4,7%**	**-13,2%**	**-18,6%**
3% eliminated at extremes	1.392.914,76 €	1.219.944,29 €	1.161.756,21 €
Difference (% Base Case)	**-8,0%**	**-12,4%**	**-16,6%**
5% eliminated at the end of the year	1.326.641,45 €	1.176.072,72 €	1.152.182,98 €
Difference (% Base Case)	**-12,4%**	**-11,3%**	**-13,2%**
Eliminating 10% extremes	1.183.915,82 €	1.068.565,70 €	1.114.074,64 €
Difference (% Base Case)	**-21,8%**	**-9,7%**	**-5,9%**

The results show that Strategy 1 brings an average cost reduction of 14%, while Strategy 2 can achieve savings of more than 20%, although with greater dependence on extreme

values. Furthermore, it can be observed that even when removing the most unfavorable values for the base case (which could have been advantageous for the strategies), the economic impact of both was still positive, although it is substantially reduced for Strategy 2. These findings confirm that the proposed strategies do not depend only on extreme market situations but show a consistent cost reduction effect under different conditions.

6. Conclusions

The main conclusions of this project confirm that artificial intelligence techniques applied to drift analysis not only allow to characterize hidden patterns in historical data, but also to translate such knowledge into economic practice. The estimated economic impact for a medium-sized retailer is between 14% and 22% savings in energy purchase costs, which represents a saving of more than 150,000 euros in the simulated scenario. These results demonstrate the practical utility of the approach, which combines descriptive and supervised analysis with concrete economic application.

Strategy 1 proves to be more consistent and robust in the face of data variability, albeit with lower returns in typical scenarios, while Strategy 2 offers higher returns, but mainly thanks to its exploitation of extreme values. Taken together, both approaches open the door to hybrid combinations that could improve stability and profitability simultaneously.

Finally, this paper confirms that deviation analysis should not be limited to the technical level of REE but can have a direct impact on the profitability of marketers. However, it also warns that an intensive use of these strategies could attract the attention of regulators, such as the CNMC, which establishes rights and obligations in system adjustment services [CNMC22]. This underscores the need to balance economic optimization with the stability and sustainability of the electricity system as a whole.

7. References

The following are the most representative references used throughout the report:

[VARA24] Varas, A. (2024). *Demand forecasting models in power systems using deep learning*. Journal of Energy and Regulation, 15(2), pp. 45-61.

[MANR20] Manrique, L. (2020). *Introduction to supervised and unsupervised learning in Machine Learning*. Editorial UPM, Madrid.

[FRAN18] Franco, D. (2018). *Machine Learning basics: concepts and applications*. Barcelona: Ediciones UPC.

[GONZ18] González, J. (2018). *Deep Learning: fundamentals and applications in engineering*. McGraw Hill Interamericana, Madrid.

[KONG19] Kong, W., Dong, Z. Y., Jia, Y., Hill, D. J., Xu, Y., & Zhang, Y. (2019). "Short-term residential load forecasting based on LSTM recurrent neural network". *IEEE Transactions on Smart Grid*, 10(1), pp. 841-851.

[MARI16] Mariano, J., & Ruiz, F. (2016). *Advanced energy prediction models: from ARIMA to neural networks*. Editorial UNED, Madrid.

[JAME23] James, G., Witten, D., Hastie, T., Tibshirani, R., & Grosse, R. (2023). *An Introduction to Statistical Learning: with Applications in Python*. Springer Texts in Statistics.

[CNMC22] Comisión Nacional de los Mercados y la Competencia (2022). *Collection rights and payment obligations for electricity system adjustment services*. Supervisory technical report.

[BRUN24] Bruneel, J., Van der Elst, S., & Meeus, L. (2024). "Managing imbalance price risk with intraday trading strategies in single-price electricity markets." *Energy Economics,* 126, 106857.

Contenido

Índice de Ilustraciones

Índice de Tablas

Índice de Anexos

Capítulo 1.-Introducción y planteamiento del proyecto

1.1.- Introducción

En los últimos años se ha disparado el uso de modelos y técnicas basadas en Inteligencia Artificial (IA) y Machine Learning (ML) en aplicaciones de predicción en múltiples campos, incluyendo el sector energético [LARA20]. En particular, la predicción de la demanda y el precio de la electricidad se ha beneficiado del uso de redes neuronales artificiales, máquinas de soporte vectorial y árboles de regresión, entre otras técnicas. Más recientemente, el aprendizaje profundo (deep learning) ha emergido como un enfoque poderoso para las series temporales energéticas, permitiendo captar patrones complejos en los datos históricos de demanda eléctrica [MALL04].

En el presente trabajo no se pretende desarrollar un modelo predictivo tradicional de demanda, ya que por un lado dichos modelos requieren herramientas de alta complejidad computacional y ya han alcanzado errores medios inferiores al 5% [REDE25] en las previsiones de Red Eléctrica. En lugar de ello, el proyecto se centra en un enfoque descriptivo o explicativo, con el objetivo de entender las relaciones entre ciertas variables que influyen en los desvíos del sistema eléctrico. Los desvíos hacen referencia a las discrepancias diarias entre la generación prevista y la real en la red, las cuales suponen un coste económico notable al tener que ser compensadas mediante servicios de ajuste [CARB07]. Reducir estos desvíos implicaría una operación más eficiente de la red eléctrica. Por tanto, este estudio buscará identificar patrones en los datos históricos que puedan explicar por qué ocurren ciertos desvíos, más que predecir su valor exacto con antelación. Más que explicar, el modelo se basa en predecir el perfil de ciertos días identificando patrones comunes entre ellos.

En concreto, el proyecto propone desarrollar un modelo descriptivo capaz de relacionar variables meteorológicas y operativas con los desvíos netos diarios de la red eléctrica española. Se emplearán datos de temperatura diaria proporcionados por la Agencia Estatal de Meteorología (AEMET) junto con datos diarios de desvío neto de Red Eléctrica de España (REE), complementados con información sobre la potencia eléctrica instalada y el número de generadores en operación cada día. Estos últimos datos ofrecen indicios sobre las previsiones de generación gestionadas por REE. Mediante técnicas descriptivas de análisis de datos, como la clústerización (clustering) con K-means, se dejará que sea el propio modelo el que encuentre agrupaciones y relaciones ocultas en los datos sin imponer hipótesis a priori. Con este primer análisis exploratorio se espera descubrir correlaciones o patrones significativos entre las variables (por ejemplo, entre determinadas condiciones de temperatura y determinados niveles de desvío).

Dado que este análisis inicial constituye únicamente un primer paso, el proyecto se completa con una segunda fase centrada en el estudio económico de los desvíos. En esta parte se analizará cómo una comercializadora eléctrica podría aprovechar el conocimiento de los patrones de desvío previamente identificados para diseñar estrategias que reduzcan el impacto de las penalizaciones y, al mismo tiempo, optimicen la rentabilidad de sus operaciones. De este modo, el análisis técnico no se limita a

caracterizar los desvíos, sino que se traduce en propuestas prácticas de actuación en el mercado eléctrico, permitiendo evaluar de forma cuantitativa el efecto económico de dichas estrategias.

En resumen, el proyecto propone una metodología híbrida: primero un análisis no supervisado con clústering para encontrar grupos naturales en los datos históricos, y posteriormente un enfoque aplicado de carácter económico, en el que se diseñan y evalúan estrategias basadas en los clústeres identificados. Este planteamiento busca ofrecer conclusiones prácticas que permitan a una comercializadora mejorar sus resultados financieros y, al mismo tiempo, contribuir a un funcionamiento más eficiente del sistema eléctrico en su conjunto.

1.2.- Motivación

La motivación principal para realizar este proyecto radica en el gran potencial no aprovechado de las técnicas de análisis de datos aplicadas al mercado eléctrico, no solo desde la perspectiva predictiva, sino también desde la económica. A pesar de los avances mencionados en modelos de predicción de la demanda, su uso limitado a la mera estimación numérica deja espacio para enfoques que permitan identificar patrones recurrentes en los desvíos y, a partir de ellos, diseñar estrategias económicas adaptadas al tipo de día que se prevé. Actualmente, los operadores del sistema (REE) cuentan con estimaciones muy precisas de la demanda y la generación previstas. Esto, sumado a la elevada complejidad de dichos modelos, hace poco atractivo centrar un Trabajo de Fin de Grado en mejorar marginalmente la precisión de las previsiones. En cambio, existe una necesidad no cubierta de estudiar el sistema desde un punto de vista aplicado: ¿qué condiciones caracterizan a los días con mayores desvíos? ¿cómo puede una comercializadora anticiparse a esos escenarios y aprovechar esa información para reducir sus costes de desvío? Analizar estas cuestiones permite dotar tanto al operador como a los agentes de mercado de herramientas prácticas de anticipación y reacción, orientadas a una gestión más eficiente de los recursos energéticos y económicos.

Otro motivo que impulsa este proyecto es la creciente importancia de las técnicas de machine learning y deep learning en el sector industrial en general. Se está viviendo un momento en que la IA está ganando protagonismo en la gestión de redes eléctricas inteligentes (smart grids) [JUDG24], integrando energías renovables, respuesta de la demanda y almacenamiento. En particular, las técnicas de IA generativa han irrumpido recientemente con fuerza, por ejemplo, en ámbitos como la generación de imágenes, texto y música [APPL24], y explorar su aplicación en el contexto energético es una oportunidad para contribuir a un campo de investigación emergente. La integración de energías renovables intermitentes y la necesidad de equilibrar en tiempo real oferta y demanda hacen que la gestión de los desvíos adquiera una relevancia creciente. En este contexto, explorar la aplicación de métodos como la clústerización y los modelos supervisados (por ejemplo, Random Forest) para caracterizar y predecir el tipo de día eléctrico no solo contribuye a comprender mejor el sistema, sino que también abre la puerta a diseñar estrategias económicas que mejoren los márgenes de las comercializadoras. El poco uso previo de estas técnicas con un enfoque económico, frente a su demostrado potencial en otros sectores como las finanzas o la logística, refuerza la motivación de aplicarlas en este campo.

En resumen, este TFG se inicia, en primer lugar, con la convicción de que los modelos explicativos basados en datos pueden plantear nuevas perspectivas para optimizar la

operación del sistema eléctrico. En segundo lugar, porque existe un vacío de conocimiento aplicado en la utilización conjunta de técnicas de clústerización y modelos predictivos para el diseño de estrategias económicas en el mercado eléctrico. Y, en tercer lugar, por el interés académico y profesional de explorar cómo estas herramientas pueden contribuir no solo a reducir los costes de los agentes, sino también a apoyar el equilibrio del sistema eléctrico en su conjunto. La motivación última es, por tanto, doble: demostrar que es posible generar un impacto económico positivo para una comercializadora mediante estrategias inteligentes basadas en datos, y al mismo tiempo contribuir a un sistema eléctrico más eficiente y sostenible, en línea con los objetivos de Red Eléctrica y del mercado eléctrico español [REE25].

1.3.- Objetivos del proyecto

Los objetivos concretos que se persiguen con este proyecto TFG son los siguientes:

- **Identificar relaciones y patrones entre las variables meteorológicas, operativas y los desvíos mediante clústering**: Se realizará un análisis exploratorio de los datos históricos utilizando el algoritmo K-means para agrupar días con características similares. En particular, se buscará encontrar grupos de días en función de su perfil de temperatura, grado de festividad, potencia programada (eólica, solar, nuclear, etc.) y estacionalidad (estación en la que se encuentren), y por último de la demanda prevista sin usar los datos de los desvíos resultantes. Este estudio estadístico inicial permitirá establecer si existen tipos de días (o situaciones) que tienden a asociarse con desvíos altos o bajos. Por ejemplo, podría descubrirse que cierto patrón de temperaturas semanales junto con determinados niveles de generación prevista conduce sistemáticamente a desvíos significativos, lo cual proporcionaría información valiosa sobre el comportamiento del sistema.
- **Validar la solidez y el grado de acierto de los clústeres**: Una vez alcanzado un buen nivel de detalle en la caracterización de los clústeres, se desarrollará un modelo de validación utilizando técnicas de aprendizaje supervisado como Random Forest. El objetivo será comprobar la fiabilidad de los grupos obtenidos y su capacidad de generalizar a nuevos datos. Para ello, se utilizarán series de datos de un año distinto al de entrenamiento, en principio el año siguiente en el calendario. Se analizará hasta qué punto el modelo es capaz de predecir correctamente a qué grupo pertenecerán los días de este nuevo conjunto basándose únicamente en las variables conocidas de antemano, como la temperatura, el tipo de generación programada o la estacionalidad, sin utilizar aún los datos reales de desvíos. La finalidad es anticipar el perfil de desvíos que podría presentar cada día clasificándolo en el grupo adecuado, lo que permitiría ajustar los modelos de previsión de Red Eléctrica para estar mejor preparados según el tipo de jornada que se espera.
- **Diseñar y evaluar estrategias económicas basadas en los clústeres identificados**: Un objetivo fundamental es aprovechar la información obtenida de la clusterización y la validación predictiva para plantear estrategias de compra y gestión de energía que reduzcan los costes derivados de los desvíos. Se desarrollarán distintos enfoques estratégicos, desde los más conservadores hasta los más agresivos, con el fin de estudiar cómo una comercializadora podría minimizar el impacto económico negativo de los desvíos y, al mismo tiempo, contribuir al equilibrio del sistema.

- **Analizar el comportamiento de los clústeres en escenarios anómalos o extremos**: Una vez identificados los clústeres, se pondrá especial atención a los días considerados atípicos (por ejemplo, con temperaturas muy extremas o con generación renovable fuera de lo previsto). Se estudiará cómo estos días afectan a la clasificación y cómo podrían condicionar las estrategias económicas diseñadas, con el fin de evaluar la robustez de las propuestas.
- **Comparar y evaluar la efectividad de las técnicas y estrategias propuestas, estableciendo recomendaciones**: Como objetivo final, el proyecto evaluará los resultados obtenidos con el enfoque de clústering combinado con los modelos predictivos y las estrategias económicas, determinando qué configuraciones ofrecen un mejor entendimiento de los desvíos y una mayor reducción de costes. Se espera elaborar una comparación de las fortalezas y limitaciones de cada enfoque en este contexto (por ejemplo, K-means es sencillo de interpretar, pero puede ser sensible a variables dominantes, mientras que Random Forest ofrece robustez predictiva, aunque con menor interpretabilidad). A partir de esta comparación, se propondrán líneas futuras de investigación o mejoras, como podría ser integrar más variables, ajustar dinámicamente las estrategias dentro de un mismo día o explorar métodos híbridos para optimizar aún más la operación económica de una comercializadora.

1.4.- Metodología de Trabajo

Para alcanzar los objetivos propuestos, se seguirá una metodología estructurada en varias etapas, combinando labores de obtención de datos, análisis exploratorio, desarrollo de modelos y evaluación de resultados. A continuación, se describen los pasos previstos:

1. **Recopilación de datos**: En primer lugar, se obtendrán y consolidarán las distintas fuentes de datos necesarias. Esto incluye la descarga de datos meteorológicos diarios, con los que se elaborarán los perfiles diarios de temperatura, proporcionados por AEMET. Por otro lado, se recopilarán los datos del sistema eléctrico disponibles públicamente a través de Red Eléctrica de España (REE). Entre estos últimos estarán los desvíos netos diarios del sistema (diferencia entre la demanda real y la prevista cada día), así como datos que influyen en la previsión, como la potencia instalada por tecnologías y la potencia prevista o generación programada diariamente. REE publica gran parte de estos datos en plataformas de transparencia y bases de datos históricas, por lo que se espera reunir, por ejemplo, varios meses de histórico para tener suficiente información. Una vez obtenidos, los datos serán integrados en un conjunto unificado, probablemente a nivel de día (cada instancia representando un día con sus variables asociadas), con lo que se dispondrá para cada día de la evolución de temperatura, de potencia prevista y de los desvíos por hora, para posteriormente aplicar las distintas técnicas.
2. **Preprocesamiento y análisis exploratorio**: Antes de aplicar modelos avanzados, se realizará un cuidadoso preprocesamiento de los datos. Esto implicará limpiar valores faltantes o erróneos, homogeneizar formatos (por ejemplo, asegurarse de que las fechas de AEMET y REE coinciden exactamente) y crear variables derivadas útiles. Posiblemente se normalizarán las variables para evitar escalas muy diferentes (por ejemplo, escalar la temperatura y la potencia a intervalos comparables).

3. **Agrupación de datos con K-means**: Con el conjunto de datos preprocesado, se aplicará el algoritmo de clústerización K-means para agrupar los días con patrones similares. Aquí cada día será tratado como un punto en un espacio multidimensional cuyos componentes podrían ser, por ejemplo: temperatura media del día, desviación neta en MWh, porcentaje de error de previsión, potencia eólica instalada, etc. Una de las formas más relevantes para clústerizar será comparar los perfiles de los días con las predicciones del día anterior y de la semana anterior. Es decir, se analizarán patrones semanales (por ejemplo, los lunes con respecto a lunes anteriores), teniendo en cuenta también particularidades como festivos o periodos especiales (Navidad, Semana Santa, etc.). Después, se elegirá un número de clústeres K apropiado (probando varios valores y utilizando métricas como el *Elbow method* [GEEK24] o el coeficiente de silueta). El resultado esperado será un conjunto de clústeres de días; cada uno se interpretará examinando sus características promedio (por ejemplo, "días de invierno fríos con alta sobreestimación de demanda", "días templados con ligera infraestimación", etc.). Esta etapa permitirá etiquetar los datos en grupos homogéneos y verificar si efectivamente existen patrones diferenciados de desvíos asociados a distintas condiciones.
4. **Modelo de validación supervisada**: Una vez obtenidos e interpretados los clústeres, se procederá a comprobar su validez y utilidad a la hora de clasificar días futuros. Para ello, se desarrollará un modelo predictivo utilizando técnicas de aprendizaje supervisado como Random Forest. El objetivo será evaluar hasta qué punto, con la información disponible de antemano (como la temperatura, la programación de generación o el tipo de día), es posible asignar correctamente cada nuevo día a uno de los clústers ya definidos. Se utilizarán series de datos de un año distinto al de entrenamiento para comprobar si los patrones identificados se mantienen estables con el paso del tiempo. Si la clasificación de los nuevos días en los clústers originales es consistente, se podrá concluir que estos grupos capturan relaciones estructurales relevantes del sistema eléctrico. Esto permitirá anticipar el perfil de desvíos asociado a cada jornada y mejorar la toma de decisiones operativas.
5. **Diseño de estrategias económicas**: Una vez validados los clústers y asegurada la capacidad de predecir a qué tipo de día pertenece cada jornada, se procederá a diseñar y aplicar distintas estrategias económicas. Estas estrategias consistirán en ajustar de forma deliberada las compras de energía de una comercializadora en función de la tipología de día identificada. Se plantearán tanto enfoques conservadores como más agresivos, con el fin de estudiar cómo se pueden reducir los costes derivados de los desvíos y, al mismo tiempo, contribuir al equilibrio del sistema. El análisis se apoyará en escenarios históricos y simulados, evaluando el impacto económico que tendría la aplicación de estas estrategias frente a una política estándar de compra de energía.
6. **Análisis de resultados y validación de hipótesis**: Tras aplicar las estrategias, se realizará un estudio comparativo entre los costes simulados de una comercializadora ficticia de los desvíos con y sin estrategia. Se evaluará en qué medida se logra una reducción significativa de los costes y qué factores condicionan la eficacia de cada enfoque. También se analizarán casos extremos o atípicos (picos de demanda inesperados, sobreproducción renovable, etc.), para comprobar la robustez de las estrategias en contextos adversos. A partir de estos resultados, se elaborarán recomendaciones prácticas para comercializadoras y se extraerán conclusiones sobre la viabilidad real de estas técnicas en el mercado eléctrico.

Capítulo 2.- Descripción de las tecnologías (estado de la técnica)

Actualmente, Red Eléctrica de España (REE) basa sus modelos de previsión en técnicas avanzadas de inteligencia artificial, principalmente redes neuronales profundas, dentro del marco del aprendizaje supervisado [VARA24]. Para contextualizar estos conceptos, resulta útil repasar brevemente las diferencias entre aprendizaje supervisado y no supervisado en *machine learning* (ML).

En el aprendizaje supervisado, el modelo se entrena con datos de entrada etiquetados (es decir, con conocimiento del resultado esperado), de modo que aprende a mapear entradas a salidas deseadas. Por ejemplo, si se quiere entrenar un modelo para reconocer patos en imágenes mediante aprendizaje supervisado, se proporcionarían numerosas fotos de patos etiquetadas como "pato" para que aprenda sus características distintivas; así, cuando se le muestre una nueva imagen, podrá identificar si se trata de un pato porque ha sido entrenado explícitamente para ello [MANR20]. En cambio, en el aprendizaje no supervisado no se entregan etiquetas ni resultados esperados: el modelo recibe únicamente datos brutos y debe descubrir por sí mismo estructuras o agrupaciones ocultas. Siguiendo el ejemplo, un modelo no supervisado de clasificación de animales recibiría imágenes de distintos animales sin etiquetar, y podría agrupar las fotos de patos separándolas de las de otras especies, identificando similitudes sin haber sido programado explícitamente para ello [FRAN18].

En este contexto, técnicas como la clústerización K-means entran en la categoría de aprendizaje no supervisado, ya que el objetivo consiste en que el algoritmo agrupe días con características similares sin indicarle de antemano a qué categoría pertenece cada perfil.

Otro concepto clave es el aprendizaje profundo (*deep learning*), que hace referencia a modelos de redes neuronales con múltiples capas internas. Este enfoque multicapa se denomina "profundo" porque la red contiene varias capas ocultas entre la entrada y la salida, permitiendo aprender representaciones de los datos en niveles sucesivamente más abstractos [GONZ18]. Este avance permitió entrenar redes con muchas más capas que antes, popularizando el término *deep learning* y extendiendo su uso a ámbitos como la predicción energética. Aunque los modelos estadísticos clásicos (como ARIMA) siguen empleándose por su simplicidad e interpretabilidad, en los últimos años se ha extendido el uso de modelos de *deep learning* como las redes LSTM o convolucionales temporales, capaces de capturar relaciones no lineales y dinámicas complejas en la demanda eléctrica [KONG19] [MARI16].

Dentro de las técnicas supervisadas, una de las más relevantes en este trabajo es Random Forest. Este algoritmo fue introducido a finales de los años noventa por Leo Breiman [BREI01] como una mejora de los árboles de decisión clásicos, con el objetivo de reducir el sobreajuste y mejorar la robustez de las predicciones. Su idea principal es entrenar múltiples árboles de decisión sobre subconjuntos aleatorios de los datos y de las variables (técnica conocida como *bagging* o *bootstrap aggregating*) y combinar sus resultados mediante votación o promedio. Este enfoque, inspirado en los métodos de ensamblado, ha demostrado ser muy eficaz tanto en clasificación como en regresión, proporcionando resultados estables y de alta precisión incluso en problemas con datos ruidosos o correlacionados [JAME23].

En el ámbito energético, Random Forest ha sido aplicado a tareas como la predicción de la demanda a corto plazo, la estimación de la generación renovable, la predicción de precios marginales horarios e incluso la clasificación de perfiles de consumo. Su éxito radica en su capacidad de manejar grandes cantidades de variables heterogéneas (climáticas, operativas y de calendario) y en su habilidad para modelar interacciones no lineales sin necesidad de especificarlas a priori. Varios estudios muestran que Random Forest puede alcanzar precisiones comparables a modelos más complejos de *deep learning*, pero con menor coste computacional y mayor interpretabilidad, lo que lo hace especialmente útil en entornos industriales donde la transparencia es crítica [JAME23].

Más allá de la predicción, el presente trabajo plantea la novedad de trasladar estos resultados al ámbito de las estrategias económicas de una comercializadora. En este punto, cabe señalar que la literatura previa sobre estrategias basadas en desvíos es limitada, ya que la mayor parte de los estudios se centran en la predicción de demanda, oferta o precios de equilibrio. Sin embargo, existen antecedentes relevantes en mercados eléctricos europeos que analizan la gestión de riesgos de desvíos y los costes asociados al balance. Por ejemplo, la CNMC ha documentado cómo los desvíos implican derechos de cobro y obligaciones de pago en el mercado de ajustes, donde las señales de precio buscan incentivar que los agentes reduzcan los desequilibrios que generan [CNMC22].

En general, las estrategias de los agentes se han orientado hacia la optimización del *portfolio* de compra y la gestión de coberturas en el mercado a plazo. Lo novedoso de este proyecto es proponer estrategias directamente basadas en la tipología de días detectada mediante clústeres y en la predicción del perfil de desvíos, lo que permite anticiparse y decidir de forma estratégica cuándo conviene sobredimensionar o infradimensionar la energía adquirida.

Hasta ahora, la literatura académica que aborda los desvíos eléctricos se ha centrado principalmente en aspectos relacionados con la gestión de la demanda o la integración de renovables, con el objetivo de reducir la incertidumbre en la operación del sistema. Sin embargo, en los últimos años comienzan a surgir investigaciones que, aunque todavía incipientes, demuestran un creciente interés en explotar las señales de balance o los precios de desvío como mecanismo para diseñar estrategias de operación optimizadas.

Un ejemplo representativo sería una estrategia de *trading* intradiario adaptativa que busca gestionar de forma explícita el riesgo asociado al *imbalance price* en mercados de precio único. Los autores utilizan modelos de mezcla probabilística para ajustar las posiciones de los agentes frente a las desviaciones, demostrando que este tipo de esquemas pueden mejorar tanto la rentabilidad como la robustez frente a la volatilidad del mercado [BRUN24].

Por su parte, hay otros que proponen un marco de mercado alternativo en el que cualquier actor, incluidas comercializadoras de tamaño medio, puede participar de forma activa en la negociación de desequilibrios. En su planteamiento, el sistema no solo se beneficia de los grandes generadores o agregadores, sino también de la agregación de pequeñas decisiones de múltiples agentes. De este modo, se refuerza la idea de que los desvíos no deben considerarse únicamente como un "fallo" del sistema, sino como una oportunidad para generar valor económico al tiempo que se apoya la estabilidad de la red [Lago21].

Finalmente, también se ha estudiado la implementación de estrategias de precios escalonados de compensación para consumidores y agregadores, analizando cómo la

estructura de precios puede incentivar la corrección de desvíos en tiempo real. Su modelo demuestra que la introducción de estos mecanismos no solo reduce los costes de ajuste, sino que también promueve un comportamiento cooperativo en los agentes del mercado. Este enfoque resulta especialmente relevante en escenarios regulados como el español, donde los precios de desvío están fuertemente vinculados a los servicios de ajuste y donde cualquier incentivo adicional puede mejorar la eficiencia global del sistema [YANG24].

En conjunto, estas aportaciones reflejan que, aunque todavía existe un vacío significativo en la investigación de estrategias económicas basadas en desvíos específicamente aplicables a comercializadoras, ya se empieza a vislumbrar una tendencia en esa dirección. El presente trabajo se inserta precisamente en ese hueco, proponiendo un enfoque original que combina la **clusterización de días**, la **validación predictiva mediante Random Forest** y el desarrollo de **estrategias económicas aplicables** a una comercializadora mediana. Con ello, se avanza no solo en el análisis descriptivo de los desvíos, sino también en demostrar que estas técnicas pueden convertirse en una herramienta práctica para optimizar costes y mejorar la eficiencia de mercado.

Capítulo 3.- Descripción del modelo desarrollado

3.1.- Recopilación de datos

Para entender la metodología de este trabajo, es necesario describir detalladamente los datos empleados en el modelo. Como se ha comentado, el estudio abarca dos fases principales: una primera fase donde se agrupan los días según sus propias características (maximizando la cohesión intra-clúster y la separación inter-clúster), y una segunda fase donde se espera obtener un impacto económico positivo para una comercializadora que implante técnicas similares. Antes de entrar en la base teórica de estas técnicas, conviene presentar las distintas variables consideradas en el análisis. Las variables del modelo se pueden clasificar en varias categorías:

Datos de corte temporal: Cómo se ha comentado anteriormente, una de las principales variables el tiempo, ya que, tanto para la primera como para la segunda, es importante que los datos tengan un indicativo temporal. En primer lugar, como identificativo de los datos, para saber qué día es cual, y a que grupo pertenece cada día. Y en segundo lugar para saber dónde hacer el corte temporal para el modelo basado en machine learning supervisado. Para ello se usarán los siguientes indicadores:

- **Día:** El día que corresponde. Los datos que se usarán empezaran el 1/04/2019 y acabaran el 30/03/2025 para abarcar 6 años enteros o 2190 días, lo cual proporcionará datos suficientes para el modelo propuesto.
- **Hora:** Para muchos de los datos que se usarán es necesario una división horaria, ya sea porque los datos están proporcionados de esta forma, o porque se ha encontrado una forma de apartarlo a una división horaria. La razón principal es por los desvíos, ya que los datos proporcionados por Red Eléctrica son horarios y así se puede realizar un perfil detallado por día.

Datos relacionados con REE: Este segundo grupo de variables están directamente referenciados de REE, ya que han sido obtenidos a través de su portal ESIOs de OpenData, estos datos consisten en lo siguiente:

- **Desvíos (MWh):** Diferencia entre la potencia programa y la potencia real de REE por día por hora. Esta será nuestra variable a predecir, ya que estos desvíos suelen ser costosos tanto para generadores, como comercializadores y para la propia red.
- **Foto_Diaria (MWh):** corresponde a la **"Previsión diaria D+1 de generación fotovoltaica"**, es decir, a la predicción de la energía que se espera genere la tecnología solar fotovoltaica para el día siguiente, agrupada por días. Correspondiente al indicador 1779 de ESIOS [ESIO25].
- **Eólica_Diaria (MWh):** Corresponde a la "**Previsión diaria D+1 de generación eólica"** calculado por el Modelo de Predicción Eólica del Operador del Sistema. Cada hora del día D+1 cuenta con un valor previsto de megavatios (MW) de potencia eólica instalada. Tiene que ver con datos de demanda, tiempo atmosférico al día siguiente y también potencia instalada. Corresponde con el indicador 1485 de ESIOS [ESIO 25]
- **Instalada_Foto (MW):** Corresponde a la **"Potencia instalada de energía solar fotovoltaica"**, es decir, a la capacidad total instalada de parques solares fotovoltaicos en el sistema eléctrico español, agrupada mensualmente y desagregada por comunidades autónomas en megavatios (MW). Tiene que ver con la evolución del mix renovable y la planificación de la infraestructura de generación. Corresponde con el indicador **1486** de ESIOS [ESIOS 25].

- **Instalada_Eolica (MW):** Corresponde a la **"Potencia instalada de energía eólica (terrestre y marina)"**, es decir, a la capacidad total instalada de parques eólicos en el sistema eléctrico español, agrupada mensualmente y desagregada por comunidades autónomas en megavatios (MW). Tiene que ver con la evolución del mix renovable y la planificación de la infraestructura de generación. Corresponde con el indicador **1485** de ESIOS [ESIOS 25].
- **Prev_Demanda (MWh):** Corresponde a la **"Previsión diaria D+1 de demanda eléctrica"**, es decir, a la estimación horaria de la carga que se espera consumir en cada hora del día siguiente, calculada por el Operador del Sistema a partir de datos históricos de consumo, previsiones meteorológicas y patrones de demanda. Corresponde con el indicador **1775** de ESIOS [ESIOS 25].
- **Disp_Nuclear (MW):** Corresponde a la **"Potencia disponible de generación Nuclear horizonte horario"**, es decir, a la capacidad neta instalada de los reactores nucleares conectados al sistema eléctrico español, medida en megavatios (MW) y utilizada para la planificación de la generación y la seguridad del suministro. Corresponde con el indicador **474** de ESIOS [ESIOS 25].

Una vez acabado con los datos de REE, se continuará con los datos que caracterizan con el tipo de día con el que se corresponde:

- **Día de la semana:** para ver el día de la semana que fuese. Como estos datos son de carácter cualitativos, para poder trabajar con ellos en este trabajo se convirtieron en variables dicotómicas binarias (1 o 0). En este caso, no hacía falta una codificación dummy, es decir, poner una como base, esto es porque no se trata de un modelo de regresión el que se está tratando sino una clústerización, y todas las posibles variables de aspecto que puedan aportar información son necesarias. Por lo que se han codificado tres variables distintitas:
    - **Lunes:** Variable dicotómica, 1 si es lunes, y 0 si no lo es.
    - **Martes-Jueves:** Variable dicotómica, 1 si es martes, miércoles o jueves, y 0 si no lo es.
    - **Viernes-Domingo/Festivo:** Variable dicotómica, 1 si es viernes, sábado o domingo o festivo nacional, y 0 si no lo es.
- **Festividad:** Es el grado de festividad del 0 al 1 a nivel nacional debido a las distintas fiestas, ya sean nacionales, locales o regionales de las distintas comunidades autónomas. Se dividen en tres tipos:
    - **Festividad nacional:** Son los días de descanso establecidos por el Estado que **se aplican en todo el territorio español**, recogidos en el Estatuto de los Trabajadores y publicados anualmente en el Boletín Oficial del Estado (BOE). Ejemplos: Año Nuevo (1 de enero), Día del Trabajo (1 de mayo), Fiesta Nacional de España (12 de octubre).
    - **Festividad autonómica/local:** Son los días festivos que cada **Comunidad Autónoma** puede señalar dentro del límite anual de 14 días. Habitualmente, además de los nacionales, cada autonomía elige hasta **4 festivos propios** para conmemorar celebraciones de ámbito regional (por ejemplo, Día de Andalucía, Día de la Comunidad de Madrid).
    - **Festividad regional:** Son los festivos que cada **ayuntamiento** determina, hasta un máximo de 2 por año, para celebrar fiestas **municipales** (patronales, históricas o tradicionales). Su aplicación queda limitada al término municipal correspondiente.

  Para calcular el grado de festividad de cada día, que supondría ver cuanta proporción de la población española está en un estado de festividad cada día. Esto afectaría a la predicción de la demanda eléctrica y es por eso por lo que es

importante calcularlo en este trabajo. Para ello, primero se sacan los datos de la población de cada comunidad del INE de los últimos 5 años. Y a partir de ahí, y de un fichero con el que se contaba ya con todas las festividades de España por provincia de los últimos 5 años, con ello se puede calcular el grado de festividad de España para los distintos días del año. Para ello:

- Para calcular esta variable compuesta, se procede del siguiente modo: si el día es festivo a nivel nacional se asigna Festividad = 1 automáticamente (pues el 100% de la población está de fiesta ese día). Si el día es festivo solo en ciertas Comunidades Autónomas (festivo autonómico o regional), se asigna un valor proporcional a la fracción de la población española que tiene descanso ese día. Por ejemplo, el Jueves Santo es festivo en la mayoría de las regiones (excepto algunas como Cataluña, que trasladan la festividad al Lunes de Pascua), de modo que aproximadamente el 72% de la población española está de festivo ese día; en consecuencia, a un Jueves Santo se le asignaría Festividad ≈ 0,72 (72%). Por último, para las fiestas regionales, se aplica un coeficiente corrector a la provincia que equivale a: la media entre, por un lado, la población de la capital de la provincia sobre el total de la provincia, y por otro, la población media de un municipio de esa provincia sobre la población total de la provincia. Para verlo más claro, se ejemplifica con el siguiente caso:

Para Madrid, primero se dividirían la población de la capital (Madrid) sobre la población de la provincia:

$$\boldsymbol{Coef_1} = \frac{Población\ de\ la\ ciudad\ de\ Madrid}{Población\ de\ la\ provincia} = \frac{3.463.311}{7.001.715} = \mathbf{49,08\%}$$

Luego se calcularía el promedio poblacional por municipio en Madrid:

$$\frac{Población\ de\ la\ provincia}{nº\ municipios\ en\ la\ provincia} = \frac{7.001.715}{179} = \mathbf{39.118} \rightarrow$$

$$\boldsymbol{Coef_2} = \frac{Población\ promedio\ de\ un\ mucinipio\ de\ Madrid}{Población\ de\ la\ provincia}$$

$$= \mathbf{39.118/7.001.715 = 0,56\%}$$

$$\overline{\boldsymbol{Coef}} = \frac{\boldsymbol{Coef_1} + \boldsymbol{Coef_2}}{\mathbf{2}} = \mathbf{24,82\%}$$

Coeficiente para la provincia de Madrid =24,82%

Para otras provincias este coeficiente puede variar, algunos ejemplos serían:

Tabla 4: Coeficientes de distintas provincias de España para calcular el grado de festividad

Ciudad	Coeficiente
Madrid	24,82%
Cuenca	13,64%
Orense	17,74%
Cádiz	5,52%

El porqué de este coeficiente calculado de esta forma, es debido a que las festividades regionales no están aclarados a cuantos municipios de la provincia afectan, y como podrían afectar tanto a la capital como a cualquier número de los municipios de la provincia, esta aproximación se relaciona con el comportamiento real de las festividades regionales de una manera plausible. Pues presupone que para una festividad regional cualquiera, solo entre el 5% y el 25% de la provincia celebra esa festividad, que no parece nada demasiado destacable y que no afectará de manera sustancial a los datos.

Con estos datos y aproximaciones ya se puede calcular el grado de festividad de España para cualquier día pasado o futuro. Este indicador resulta útil porque un mayor grado de festividad suele conllevar patrones distintos de demanda eléctrica (fines de semana y festivos suelen tener menor demanda que días laborables) y potencialmente afecta a los desvíos.

Selección de estaciones meteorológicas y perfil de temperatura: Dado que la temperatura es una variable esencial que afecta tanto a la demanda eléctrica como a la producción renovable, se ha prestado especial atención a su representación. La Agencia Estatal de Meteorología (AEMET) proporciona datos diarios de temperatura de numerosas estaciones repartidas por el territorio (incluyendo temperatura media diaria, máxima y mínima, precipitaciones, etc.). Incluir todas las estaciones meteorológicas de España habría sido inviable y redundante para este estudio, por lo que se seleccionó un conjunto reducido de estaciones representativas de las principales regiones climáticas del país. España cuenta con una gran diversidad climática que, de forma simplificada, puede agruparse en cinco tipologías climáticas principales, cada una de las cuales está representada en este trabajo por una o más estaciones meteorológicas cuyos registros capturan fielmente las variaciones de ese clima. España cuenta con una gran diversidad climática que puede agruparse en cinco grandes tipologías según la clasificación Köppen [BRIT24], cada una representada por estaciones que capturan fielmente sus variaciones térmicas:

1. **Clima oceánico (Atlántico noroeste) [Cfb]: c**aracterística de Galicia, con temperaturas suaves y precipitaciones bien distribuidas a lo largo del año. Para reflejar tanto la influencia marítima como la ligera continentalización hacia el interior, se emplean dos estaciones: A Coruña (situada en la costa, expuesta al Atlántico) y Santiago de Compostela (más interior). Sus series horarias combinadas sintetizan adecuadamente el perfil térmico de esta región y ponderan su peso demográfico dentro del conjunto nacional
2. **Clima subtropical de Canarias [Csb]:** propio del archipiélago canario, con inviernos muy suaves y veranos cálidos pero moderados por la brisa marina, presentando oscilaciones diarias reducidas. Se utiliza la estación de Tenerife Norte – Los Rodeos como proxy del conjunto de las islas, asegurando la cobertura de este clima singular (en Canarias reside cerca del 10% de la población nacional)
3. **Clima mediterráneo costero (sur, Sevilla) [Csa]:** representado por la cuenca baja del Guadalquivir y Andalucía occidental. Se selecciona la estación de Sevilla – Aeropuerto (San Pablo), cuyas series reflejan veranos muy secos y calurosos (a menudo por encima de 40 °C) e inviernos templados típicos del litoral sur mediterráneo. Sus elevados picos térmicos estivales capturan la amplitud característica de esta zona y suponen un peso relevante en la media nacional
4. **Clima mediterráneo continental (interior, Madrid) [CSA continentalizado]:** clima propio de la Meseta central. Se toma la estación de Madrid – Retiro, ubicada

en el centro de la ciudad, que registra inviernos fríos y veranos muy calurosos con grandes oscilaciones diarias y estacionales. La zona metropolitana de Madrid concentra más del 15% de la población española, por lo que sus datos son esenciales para la media ponderada nacional. La estación del Retiro, además, presenta series prácticamente completas y capta bien la isla de calor urbana, lo que la hace muy representativa del clima continental urbano de la región centro

5. **Clima mediterráneo continental semiárido (interior noreste) [BSk]**: variante propia del valle del Ebro y zonas del interior norte peninsular, con menor humedad. Se incluye la estación de Zaragoza – Aeropuerto, que registra un régimen térmico seco y continental, con amplias oscilaciones, ligeramente distinto al de Madrid. Su inclusión afina la representación de los climas interiores septentrionales de España.

Además, para seleccionar la estación dentro de cada una de las ciudades se definieron los siguientes criterios: Para cada ciudad seleccionada se eligió, dentro de lo posible, la estación de AEMET más cercana al núcleo urbano, con series horarias lo más completas posibles y representativas de las condiciones locales. Siguiendo estos criterios, las estaciones meteorológicas "más representativas" para nuestras seis localidades fueron: Madrid (Parque del Retiro, en el corazón urbano, con mínima tasa de datos faltantes) ; Sevilla (Aeropuerto de San Pablo, mantenimiento continuo y refleja bien la amplitud térmica veraniega) ; Zaragoza (Aeropuerto, datos fiables hora a hora, captura las condiciones secas del valle del Ebro) ; A Coruña (Aeropuerto de Alvedro, ubicado en la ría, series largas y completas, sin grandes interferencias urbanas) ; Santiago de Compostela (Aeropuerto de Lavacolla, registros exhaustivos que reflejan el clima interior húmedo gallego) ; y Tenerife Norte (Aeropuerto Los Rodeos, estación de referencia canaria, cubre bien el régimen subtropical insular) .

Por último, no es suficiente con elegir las estaciones de manera correcta, sino que también es importante el diseñar un perfil de temperatura adecuado a cada hora del día, puesto que los datos de la AEMET no son suficientes para esto. Estos datos solo contienen cierta información meteorológica en concreto son de esta forma [AEMET25]:

Ilustración 5: Ejemplo de FICHERO DE Datos De la AEMET


```
[{
"fecha" : "2025-01-26",
"indicativo" : "3195",
"nombre" : "MADRID, RETIRO",
"provincia" : "MADRID",
"altitud" : "667",
"tmed" : "8,8",
"prec" : "5,2",
"tmin" : "5,4",
"horatmin" : "05:50",
"tmax" : "12,2",
"horatmax" : "23:59",
"dir" : "22",
"velmedia" : "2,2",
"racha" : "10,0",
"horaracha" : "18:20",
"presMax" : "942,2",
"horaPresMax" : "01",
```

"presMin" : "933,6",
"horaPresMin" : "24",
"hrMedia" : "88",
"hrMax" : "98",
"horaHrMax" : "22:20",
"hrMin" : "82",
"horaHrMin" : "17:40"
}, {
"fecha" : "2025-01-27",
"indicativo" : "3195",
"nombre" : "MADRID, RETIRO",
"provincia" : "MADRID",
"altitud" : "667",
"tmed" : "10,4",
"prec" : "1,4",
"tmin" : "6,2",
"horatmin" : "23:59",
"tmax" : "14,5",
"horatmax" : "13:50",
"dir" : "24",
"velmedia" : "6,7",
"racha" : "17,2",
"horaracha" : "09:20",
"presMax" : "936,6",
"horaPresMax" : "24",
"presMin" : "929,7",
"horaPresMin" : "15",
"hrMedia" : "85",
"hrMax" : "99",
"horaHrMax" : "06:10",
"hrMin" : "69",
"horaHrMin" : "Varias”

Como se puede observar en la Ilustración 1, los datos crudos obtenidos directamente de la AEMET no son suficientes para obtener cualquier tipo de relación con el resto de variables propuestas en el problema, además de no poder relacionarse con los desvíos, por lo que se llevará a cabo una transformación basada en la relación sinusoidal de las temperaturas con su temperatura media y los tiempos de temperatura mínima máxima y el valor de cada una de estas variables [MAUR00]. Para esto se desarrolló el siguiente modelo:

$$\mathrm{T(t)} = \mathrm{TM} + [\mathrm{a} * \cos(\frac{\pi \mathrm{t}}{12}) + \mathrm{b} * \mathrm{sen}(\frac{\pi \mathrm{t}}{12})]$$

Dónde se puede encontrar una aproximación a la variación diaria de la temperatura a través de un solo armónico (ciclo senoidal de periodo 24 h) alrededor de la media diaria T_M. Además, Al buscar los extremos de esta función, resolviendo $\frac{dT}{dt} = 0$, se demuestra que el valor máximo se alcanza cuando el vector (a,b) está alineado con $(\cos\left(\frac{\pi \mathrm{t}}{12}\right), \mathrm{sen}(\frac{\pi \mathrm{t}}{12}))$ de modo que:

$$\begin{cases} T_{max} = T_M + \sqrt{a^2 + b^2} \\ T_{min} = T_M - \sqrt{a^2 + b^2} \end{cases}$$

Esto implica que la amplitud del ciclo diario,($\sqrt{a^2 + b^2}$), coincide con la mitad de la diferencia entre la temperatura máxima y mínima del día, calibrando así las constantes a y b. Este primer armónico es suficiente para reproducir con alta fidelidad la curva diurna , con un error típico por debajo de 1 °C, y se ha validado tanto en superficie como mediante observaciones satelitales [FAL18], así como en estudios de validación de modelos de ciclo diurno que muestran que "se ajusta un modelo armónico con el periodo correspondiente a la frecuencia dominante de los datos, y los residuales resultantes se emplean para identificar la siguiente frecuencia relevante en la serie" [GOE16].

Tras definir todas las variables anteriores, es necesario integrar las distintas fuentes de datos en un único conjunto coherente para cada día. Se disponen, por un lado, de variables diarias (ej. potencia instalada, festivos) y, por otro, de variables horarias (ej. desvíos, generación prevista). En este trabajo se construyó una base de datos consolidada donde cada fila corresponde a un día, y las variables horarias se representaron mediante 24 valores (uno por hora) por día. Esto da lugar a vectores de dimensión 24×p por día, donde p es la cantidad de variables consideradas. A modo ilustrativo, a continuación, se muestra un extracto de los datos por hora correspondientes a un día concreto (el 09/10/2020):

Tabla 5: Ejemplo del perfil del día 09/10/2020 para su clústerización

Día	Hora	Desvíos (MWh)	Previsión Foto (MWh)	Eólica_Diaria (MWh)	Instalada_Foto (MW)	Instalada_Eolica (MW)	Prev_Demanda (MWh)	Disp_Nuclear (MW)
09/10/2020	0:00:00	706,3	0	2.136,00	11.149,20	27.225,09	24.555,00	6.125,50
09/10/2020	1:00:00	756,4	0	2.122,00	11.149,20	27.225,09	23.336,00	6.125,50
09/10/2020	2:00:00	962,6	0	2.023,00	11.149,20	27.225,09	22.586,00	6.125,50
09/10/2020	3:00:00	1.052,70	0	1.923,00	11.149,20	27.225,09	22.234,00	6.125,50
09/10/2020	4:00:00	808,9	0	1.862,00	11.149,20	27.225,09	22.153,00	6.125,50
09/10/2020	5:00:00	1.007,30	0	1.711,00	11.149,20	27.225,09	22.510,00	6.125,50
09/10/2020	6:00:00	593	0	1.580,00	11.149,20	27.225,09	24.427,00	6.125,50
09/10/2020	7:00:00	-491,5	0	1.470,00	11.149,20	27.225,09	27.379,00	6.125,50
09/10/2020	8:00:00	-721,4	326,6	1.264,00	11.149,20	27.225,09	28.420,00	6.125,50
09/10/2020	9:00:00	-795,6	2.068,60	1.097,00	11.149,20	27.225,09	29.532,00	6.125,50
09/10/2020	10:00:00	768,3	4.099,00	950	11.149,20	27.225,09	30.087,00	6.125,50
09/10/2020	11:00:00	860,8	5.394,20	976	11.149,20	27.225,09	30.368,00	6.125,50
09/10/2020	12:00:00	857,9	6.033,00	1.049,00	11.149,20	27.225,09	30.849,00	6.125,50
09/10/2020	13:00:00	669	6.357,60	1.169,00	11.149,20	27.225,09	30.915,00	6.125,50
09/10/2020	14:00:00	120,4	6.325,20	1.221,00	11.149,20	27.225,09	30.000,00	6.125,50
09/10/2020	15:00:00	747,9	6.044,00	1.335,00	11.149,20	27.225,09	28.853,00	6.125,50
09/10/2020	16:00:00	1.057,20	5.359,60	1.563,00	11.149,20	27.225,09	28.418,00	6.125,50
09/10/2020	17:00:00	1.035,40	4.137,20	2.164,00	11.149,20	27.225,09	28.223,00	6.125,50
09/10/2020	18:00:00	559,1	2.177,50	3.008,00	11.149,20	27.225,09	28.087,00	6.125,50
09/10/2020	19:00:00	416,1	382,8	4.036,00	11.149,20	27.225,09	28.546,00	6.125,50
09/10/2020	20:00:00	-204,6	0	4.514,00	11.149,20	27.225,09	30.657,00	6.125,50
09/10/2020	21:00:00	-164,3	0	5.065,00	11.149,20	27.225,09	29.980,00	6.125,50
09/10/2020	22:00:00	362,9	0	5.718,00	11.149,20	27.225,09	27.275,00	6.125,50
09/10/2020	23:00:00	478,4	0	6.092,00	11.149,20	27.225,09	25.171,00	6.125,50

Tabla 6: Continuación Tabla 2

Día	Hora	Festividad	Día de la semana	Lunes	Martes-Jueves	Viernes-Domingo/Festivo	Temp_Madrid (ºC)	Temp_Tenerife (ºC)
09/10/2020	0:00:00	0,11	D	0,00	0,00	1,00	14,89	12,37
09/10/2020	1:00:00	0,11	D	0,00	0,00	1,00	13,54	11,22
09/10/2020	2:00:00	0,11	D	0,00	0,00	1,00	12,57	10,36
09/10/2020	3:00:00	0,11	D	0,00	0,00	1,00	12,06	9,84
09/10/2020	4:00:00	0,11	D	0,00	0,00	1,00	12,03	9,70
09/10/2020	5:00:00	0,11	D	0,00	0,00	1,00	12,49	9,94
09/10/2020	6:00:00	0,11	D	0,00	0,00	1,00	13,41	10,56
09/10/2020	7:00:00	0,11	D	0,00	0,00	1,00	14,73	11,51
09/10/2020	8:00:00	0,11	D	0,00	0,00	1,00	16,35	12,73
09/10/2020	9:00:00	0,11	D	0,00	0,00	1,00	18,16	14,12
09/10/2020	10:00:00	0,11	D	0,00	0,00	1,00	20,04	15,60
09/10/2020	11:00:00	0,11	D	0,00	0,00	1,00	21,87	17,07
09/10/2020	12:00:00	0,11	D	0,00	0,00	1,00	23,51	18,43
09/10/2020	13:00:00	0,11	D	0,00	0,00	1,00	24,86	19,58
09/10/2020	14:00:00	0,11	D	0,00	0,00	1,00	25,83	20,44
09/10/2020	15:00:00	0,11	D	0,00	0,00	1,00	26,34	20,96
09/10/2020	16:00:00	0,11	D	0,00	0,00	1,00	26,37	21,10
09/10/2020	17:00:00	0,11	D	0,00	0,00	1,00	25,91	20,86
09/10/2020	18:00:00	0,11	D	0,00	0,00	1,00	24,99	20,24
09/10/2020	19:00:00	0,11	D	0,00	0,00	1,00	23,67	19,29
09/10/2020	20:00:00	0,11	D	0,00	0,00	1,00	22,05	18,07
09/10/2020	21:00:00	0,11	D	0,00	0,00	1,00	20,24	16,68
09/10/2020	22:00:00	0,11	D	0,00	0,00	1,00	18,36	15,20
09/10/2020	23:00:00	0,11	D	0,00	0,00	1,00	16,53	13,73

Tabla 7: Continuación Tabla 3

Día	Hora	Temp_Sevilla (ºC)	Temp_Santiago (ºC)	Temp_Coruña (ºC)	Temp_Zaragoza (ºC)
09/10/2020	0:00:00	21,84	11,26	16,27	17,25
09/10/2020	1:00:00	19,79	10,50	16,11	15,64
09/10/2020	2:00:00	17,99	10,11	16,12	14,27
09/10/2020	3:00:00	16,56	10,11	16,31	13,25
09/10/2020	4:00:00	15,59	10,50	16,66	12,64
09/10/2020	5:00:00	15,15	11,26	17,15	12,49
09/10/2020	6:00:00	15,28	12,34	17,75	12,80
09/10/2020	7:00:00	15,96	13,66	18,41	13,55
09/10/2020	8:00:00	17,15	15,14	19,09	14,69
09/10/2020	9:00:00	18,76	16,66	19,74	16,15
09/10/2020	10:00:00	20,69	18,14	20,33	17,82
09/10/2020	11:00:00	22,81	19,46	20,80	19,60
09/10/2020	12:00:00	24,96	20,54	21,13	21,35
09/10/2020	13:00:00	27,01	21,30	21,29	22,96
09/10/2020	14:00:00	28,81	21,69	21,28	24,33
09/10/2020	15:00:00	30,24	21,69	21,09	25,35
09/10/2020	16:00:00	31,21	21,30	20,74	25,96
09/10/2020	17:00:00	31,65	20,54	20,25	26,11
09/10/2020	18:00:00	31,52	19,46	19,65	25,80
09/10/2020	19:00:00	30,84	18,14	18,99	25,05
09/10/2020	20:00:00	29,65	16,66	18,31	23,91
09/10/2020	21:00:00	28,04	15,14	17,66	22,45
09/10/2020	22:00:00	26,11	13,66	17,07	20,78
09/10/2020	23:00:00	23,99	12,34	16,60	19,00

Como se observa en este ejemplo, cada día queda descrito por un conjunto extenso de valores. Las variables operativas (desvíos, previsiones de generación renovable, demanda, etc.) presentan variaciones a lo largo de las horas del día, mientras que las variables de tipo de día (festividad, indicadores de lunes, martes-jueves, etc.) mantienen el mismo valor en todas las horas de una jornada determinada. Asimismo, se dispone de la temperatura horaria estimada en cada una de las seis localidades meteorológicas seleccionadas (Madrid, Tenerife, Sevilla, Santiago, A Coruña y Zaragoza), que representan las distintas regiones climáticas de España (como se describe más abajo). Con este proceso de integración de datos, se obtiene finalmente una matriz de datos completa donde cada día está asociado a un vector de características que incluye toda la información conocida de antemano para ese día (variables temporales, operativas y meteorológicas). Esta matriz de datos consolidada será la base sobre la cual se aplicarán los algoritmos de agrupación (clústering) y los posteriores análisis.

3.2.- Clústerización de los días

El algoritmo K-means es una técnica de agrupamiento (clústering) no supervisado ampliamente utilizada para particionar un conjunto de datos en K grupos o clústeres distintos. En esencia, K-means busca dividir n observaciones en K clústeres no solapados, de modo que cada observación pertenezca a un único grupo. La calidad de un

agrupamiento se evalúa mediante la variación interna de los grupos: idealmente, los puntos dentro de un mismo clúster deben ser lo más similares posible entre sí. Matemáticamente, K-means intenta minimizar la suma de las distancias al cuadrado de cada punto a su centroide (lo que se conoce como suma de cuadrados intra-clúster o inercia). En otras palabras, el algoritmo trata de encontrar las particiones que minimicen la dispersión de los datos dentro de cada grupo [JAME23]. Matemáticamente el algoritmo a minimizar es:

Ecuación 1: Ecuación a minimizar en el algoritmo K-means [JAME23]

$$\underset{C_1,\ldots,C_K}{\text{minimize}} \left\{ \sum_{k=1}^{K} \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2 \right\}$$

El procedimiento de K-means es iterativo. Primero se fija el número de clústeres K, y se asigna aleatoriamente a cada observación un número de 1 a K, inicializando así una partición inicial. A continuación, el algoritmo alterna entre dos pasos fundamentales: (a) recalcular el centroide de cada clúster con las asignaciones actuales, y (b) reasignar cada observación al clúster con el centroide más cercano (según la distancia euclídea) [JAME 23]. El centroide de un clúster es simplemente el vector promedio de las coordenadas de las observaciones en ese grupo (de ahí el nombre K-means, “K medias”). Después de recalcular los centroides, en la siguiente iteración cada punto “mira” qué centroide le queda más próximo y cambia de grupo si corresponde. Este ciclo de recalcular promedios y reagrupar puntos se repite hasta que las asignaciones de los puntos ya no cambian, lo cual indica que se ha alcanzado la convergencia. El algoritmo está garantizado a disminuir la inercia total en cada iteración, aproximándose progresivamente a un mínimo local de la variación intra-clúster. Finalmente, produce como resultado la asignación definitiva de cada dato a un clúster y las coordenadas de los K centroides representativos. Cabe destacar que el óptimo hallado por K-means es local (no necesariamente el global) y depende de la inicialización aleatoria; por ello es común ejecutar el algoritmo varias veces con distintas semillas o bien utilizar inicializaciones inteligentes (por ejemplo, el método K-means++) para mejorar la solución. De hecho, [GERÓ19] se recomienda emplear K-means++ por defecto, ya que este método selecciona centros iniciales alejados entre sí y tiende a converger hacia agrupaciones de mejor calidad que una inicialización puramente aleatoria. Otro aspecto importante es la sensibilidad de K-means a valores atípicos: un outlier o dato extremo puede desplazar significativamente un centroide, distorsionando el resultado. Por ello es buena práctica preprocesar los datos (normalizar las escalas o eliminar outliers extremos, entre otros) y elegir cuidadosamente K para obtener resultados significativos.

En suma, K-means es sencillo y eficiente para datos numéricos, pero requiere tomar decisiones previas importantes, siendo la elección de K la más crítica. Una limitación clave de K-means es que exige fijar de antemano el número de clústeres K. Seleccionar un K inapropiado puede llevar a agrupaciones poco útiles: si K es demasiado pequeño, se mezclarán en un mismo clúster datos que en realidad son heterogéneos; si K es demasiado grande, se corre el riesgo de subdividir grupos naturales en clústeres artificiales. [TREV18] Se señala que esta necesidad de preespecificar K es un inconveniente de K-means frente a

métodos como el clústering jerárquico. Dado que no existe una fórmula exacta para el K óptimo, se recurre a criterios empíricos. Uno de los métodos más difundidos es la regla del codo (también llamada método del codo o método de las inercias). Este procedimiento gráfico consiste en ejecutar K-means variando K en un rango (por ejemplo, probando K=1, 2,...,10) y, para cada valor, calcular la inercia intra-clúster total (también conocida como SSE, *sum of squared errors*). Luego se grafica dicha inercia en el eje vertical frente al número de clústeres en el eje horizontal. A medida que aumenta K, la inercia siempre disminuye (ya que cuantos más clústeres, más cerca estarán los puntos de sus centroides); sin embargo, la ganancia marginal por añadir clústeres empieza a decrecer a partir de cierto punto. En la curva resultante se suele observar un punto de inflexión con forma de codo: inicialmente la inercia cae pronunciadamente al incrementar K (lo que indica que agregar clústeres mejora sustancialmente la compactación de los 1 2 3 4 5 6 7 8 9 10 11 grupos), pero llega un punto a partir del cual la curva se aplana y las mejoras adicionales son mínimas. Ese punto de inflexión –el "codo" de la gráfica– señala un K para el cual la reducción de inercia deja de ser significativa, y por tanto se interpreta como el número óptimo de clústeres.

Para ello se describirá el método a utilizar para localizar el codo, y este es el de derivadas discretas. Este se describe como

$$\Delta J(K) \;=\; J(K+1) - J(K) \;\leq 0,$$

y la **segunda diferencia** o curvatura discreta

$$\Delta^2 J(K) \;=\; J(K+2) - 2\,J(K+1) + J(K) \;=\; \Delta J(K+1) - \Delta J(K)$$

Un **criterio de codo** ampliamente usado es

$$K^* \;=\; \arg\min \; \Delta^2 J(K)$$

es decir, elegir el K donde la **curvatura hacia abajo** es más intensa (el cambio de pendiente es más brusco): a partir de ahí, los descensos de J por aumentar K son menores y típicamente corresponden a sobre-segmentación.

En la práctica, la determinación visual del codo puede ser algo subjetiva, pero suele dar un criterio razonable para elegir K: antes del codo se estaba sub-agrupando la estructura de los datos (alta variación interna), y después del codo se estaría sobre-ajustando ruido o diferencias muy pequeñas Aunque el método del codo no garantiza una elección perfecta de K en todos los casos, proporciona una guía intuitiva basada en la disminución de la variabilidad intragrupo [JAME23], complementándose con otras métricas como el coeficiente de silueta si se requiere un análisis más robusto. En resumen, la regla del codo ofrece un equilibrio entre simplicidad y efectividad para estimar un K adecuado, motivo por el cual es habitualmente enseñada en cursos introductorios de *Machine Learning*.

Random Forest es un algoritmo de aprendizaje supervisado que mejora la capacidad predictiva y la robustez de los modelos al combinar múltiples árboles de decisión. A diferencia de K-means, que es no supervisado, Random Forest se utiliza, entre otros, para tareas de **clasificación** y **regresión**. Utiliza una técnica de *bagging* (Bootstrap Aggregating) e introduce aleatoriedad en el proceso de creación de árboles, lo que reduce el

sobreajuste. Cada árbol se entrena con un subconjunto aleatorio de los datos y selecciona aleatoriamente un subconjunto de características en cada nodo para realizar la división. Esta aleatoriedad genera árboles de decisión variados que, al combinarse, proporcionan un modelo más robusto y menos susceptible al sobreajuste. En la predicción, Random Forest calcula el **voto mayoritario** en clasificación o el **promedio** en regresión a partir de los resultados de todos los árboles.

La principal ventaja de Random Forest es su **capacidad para manejar datos ruidosos** y su **resistencia al sobreajuste**, lo que lo convierte en una herramienta poderosa en Machine Learning, especialmente cuando se dispone de grandes conjuntos de datos con muchas variables. Sin embargo, pierde algo de **interpretabilidad**, ya que, a diferencia de un solo árbol de decisión, es difícil entender el modelo completo debido a la cantidad de árboles involucrados. A pesar de ello, se pueden obtener medidas de **importancia de las variables** para evaluar qué características son más relevantes para el modelo.

En resumen, tanto **K-means** como **Random Forest** son técnicas fundamentales en el análisis de datos: K-means es útil para descubrir agrupaciones ocultas en datos no etiquetados, mientras que Random Forest proporciona un modelo predictivo robusto y eficiente en datos etiquetados. Ambos algoritmos ofrecen capacidades complementarias en la ingeniería de datos y son esenciales en un repertorio de herramientas de Machine Learning.

3.4.- Implantación numérica

Para llevar a la práctica las técnicas descritas, se desarrollaron una serie de códigos en Python que permiten su implementación sobre los datos recopilados. En primer lugar, se prepararon los datos para aplicar K-means. Como se indicó, cada día está representado por un vector de dimensión 24×p (24 horas por p variables). Fue necesario entonces aplanar la estructura de datos original (donde cada fila del dataframe correspondía a una hora de un día) en una nueva estructura donde cada fila corresponda a un día completo con todas sus horas concatenadas. Seguidamente, se normalizaron las variables para evitar que sus distintas escalas distorsionaran el cálculo de distancias: se aplicó una normalización Min-Max variable por variable, escalando cada característica numérica al rango [0, 1]. Sin este paso de escalado, las variables medidas en unidades mayores (por ejemplo, megavatios de potencia) dominarían la métrica de distancia euclídea y eclipsarían la influencia de otras variables con valores numéricos más pequeños (como la festividad codificada de 0 a 1 o las temperaturas, medidas en decenas de grados). Al llevar todas las variables al mismo rango, se homogeneiza la contribución de cada variable-hora en el espacio de características, asegurando que el centroide de cada clúster refleje fielmente el perfil completo del día y no se vea sesgado por magnitudes heterogéneas. Además, esta normalización mitiga el problema de la concentración de distancias en espacios de alta dimensión, donde la disparidad de escalas podría reducir el contraste entre observaciones. En resumen, sin los pasos de aplanado y escalado, el algoritmo K-means no podría comparar correctamente perfiles diarios completos ni formar clústeres coherentes, comprometiendo tanto la convergencia del método como la interpretabilidad de los grupos hallados. La preparación de los datos descrita se implementó con las bibliotecas estándar de Python. A modo de ejemplo, el fragmento de código siguiente ilustra el proceso de aplanado y normalización de la matriz de días:

```
# Aplanar los datos para K-means
```

```
flattened_data = pivoted_df.map(lambda x: np.array(x) if isinstance(x,
list) else np.array([x]))
flattened_data = np.array(flattened_data.values.tolist())  # Convertir a
un array de NumPy
flattened_data = flattened_data.reshape(flattened_data.shape[0], -1)  #
Aplanar las horas

# Normalizar las variables (escalar al rango 0-1)
scaler = MinMaxScaler()
normalized_data = scaler.fit_transform(flattened_data)
```

A continuación, se procedió a aplicar el algoritmo K-means sobre los datos transformados. Se comenzó declarando un número inicial de clústeres arbitrario (por ejemplo, $K = 3$) con el fin de realizar una primera partición y posteriormente evaluar cuál sería el K más adecuado. Una característica de K-means es que el usuario debe especificar a priori cuántos clústeres quiere encontrar; sin embargo, existe un número óptimo K^* que conviene determinar objetivamente en lugar de fijarlo de antemano al azar. Para estimar el adecuado, se empleó la llamada regla del codo (elbow method). Esta técnica gráfica consiste en ejecutar K-means variando K en un rango (en nuestro caso probando K = 1, 2, ..., 10) y, para cada valor, calcular la inercia intra-clúster total (es decir, la suma de cuadrados intra-clúster). A medida que K aumenta, naturalmente disminuye (ya que más clústeres explican mejor la variabilidad), pero llega un punto en que el beneficio marginal por añadir un clúster extra decae notablemente, formando un "codo" en la gráfica de J vs K. Formalmente, se implementó una variante cuantitativa del método del codo basada en la curvatura discreta de la curva inercia–K. Sea la inercia obtenida con K clústeres; se computó primero la primera diferencia (pendiente entre puntos consecutivos, que es negativa porque decrece al aumentar K) y luego la segunda diferencia discreta:

$$\Delta^2 J(K) = J(K+2) - 2J(K+1) + J(K) = \Delta J(K+1) - \Delta J(K),$$

que mide el **cambio de pendiente** (curvatura). El "codo" se identifica en el K cuyo Δ2J es **mínimo** (el más negativo), es decir, donde aparecen rendimientos decrecientes claros: añadir un clúster extra ya no reduce sustancialmente la inercia y empieza a fragmentar clústeres existentes. Para ello, se ha construido la siguiente tabla que demuestra el número óptimo de clústers:

Tabla 8: Inercia J(K) de los clústeres, mejoras y segunda derivada

K	*J(K)*	*ΔJ(K)=J(K−1)−J(K)*	*Δ^2J(K)=J(K+1)−2J(K)+J(K-1)*
1	54640,2	-	-
2	34157,5	20482,7	-
3	22761,6	11396,0	9086,7
4	19365,7	3395,9	8000,1
5	18294,7	1071,0	2324,9
6	17482,0	812,7	258,3
7	16768,0	714,1	98,6
8	14596,8	2171,2	-1457,1
9	14326,9	269,9	1901,3
10	13758,1	568,9	-299,0

De esta tabla se pueden hacer las siguientes observaciones:

1. Se observa una meseta clara entre K=4 y K=7: las mejoras decrecen y se estabilizan en el rango 700–1100.
2. Justo después, al pasar a K=8, aparece un **repunte** hasta 2 171.2, indicio de **sobre-segmentación** (romper un clúster para capturar subestructura puntual). Este quiebre se cuantifica con la **segunda diferencia** (curvatura discreta) $\Delta^2 J(K) = J(K+1) - 2J(K) + J(K-1)$: para K=8, **−1 457.10, el mínimo** (más negativo) en el punto asociado a **K=7**, que es, por definición del método del codo, donde la pendiente "rompe" y comienzan rendimientos claramente decrecientes.
3. En términos de varianza explicada, [JAME23]

$$R^2(K) = 1 - \frac{J(K)}{J(1)}$$

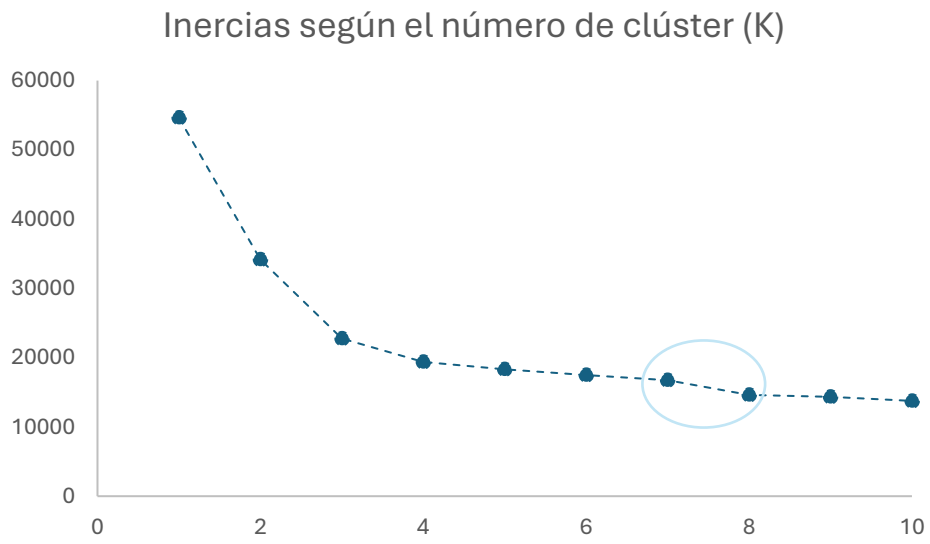
$$con\ K = 7\ se\ obtiene\ R^2(7) \approx 0.6931\ (69.31\ \%),$$

$$y\ subir\ a\ K = 8\ solo\ aporta\ hasta\ R^2(8) \approx 0.7329$$

(**+3.97 p.p.**) a costa de complejidad adicional y potencial inestabilidad de clústeres.

Por lo tanto, el criterio cuantitativo del codo (curvatura mínima) y la lectura de mejoras marginales respaldan fijar **K=7** como solución parsimoniosa y estable. También existe una opción gráfica, que es la que se mostrará a continuación que permite visualizar rápidamente este fenómeno:

Ilustración 6: Elbow Method



Una vez que ya se ha encontrado, cual es el número óptimo de clústeres, por dos métodos distintos, se puede empezar a analizar las características de estos distintos clústeres. Para ello, se empezará por describir algunos de los resultados del análisis que van a permitir no solo describir, sino caracterizar los distintos clústeres. Para ello se ha elaborado la una tabla con las medias globales por variable para cuatro franjas horarias: Madrugada, Mañana, Tarde y Noche, no porque el análisis de clústeres se haya hecho de esta forma,

sino para dar una explicación menos repetitiva del análisis, puesto que ir observando hora por hora, para esta primera caracterización, no aporta tanto en este momento del análisis. Esta es la Tabla de resúmenes global:

Tabla 9: Medias globales por día por variable (2018-2025)

Variable	Madrugada (0-5)	Mañana (6-11)	Tarde (12-17)	Noche (18-23)
Prevision_Foto	0,0	3389,7	7576,3	1044,7
Eólica_Diaria	6821,9	6008,2	6446,8	7219,4
Instalada_Foto	17856,3	17857,3	17857,3	17857,3
Instalada_Eolica	28737,6	28737,9	28737,9	28737,9
Prev_Demanda	22689,3	27395,5	29139,3	29203,1
Disp_Nuclear	6441,4	6440,7	6439,2	6441,8
Festividad	0,0	0,0	0,0	0,0
Lunes	0,2	0,2	0,2	0,2
MartesJueves	0,3	0,3	0,3	0,3
ViernesDomingoFestivo	0,5	0,5	0,5	0,5
Temp_Madrid	11,9	14,7	20,6	17,9
Temp_Tenerife	22,1	26,7	28,1	22,8
Temp_Sevilla	14,5	17,5	25,1	22,1
Temp_Santiago	11,1	14,1	18,8	15,9
Temp_Coruña	12,4	15,8	17,9	14,6
Temp_Zaragoza	12,4	15,1	21,4	18,7

Algunas primeras conclusiones de esta tabla serían, la **fotovoltaica** no aporta de madrugada, **sube con fuerza por la mañana** y **alcanza su máximo en la tarde** antes de caer al anochecer. La **eólica** es más bien **nocturna y de madrugada** (ligeramente superior en esas franjas que en la mañana), lo que cuadra con los regímenes de viento típicos. La **demanda** arranca baja de madrugada, **sube por la mañana**, **toca techo en la tarde** y se mantiene **muy alta en la noche**; es decir, existe una curva pronosticada con "valle nocturno–madrugada y doble cresta tarde/noche". La **nuclear** permanece prácticamente **plana** a lo largo del día, como base del sistema. En el calendario, el reparto medio de días (**0,2 lunes**, **0,3 martes-jueves**, **0,5 fin de semana/festivo**) recuerda que, si un clúster se aparta mucho de esas proporciones, probablemente agrupe jornadas de un mismo tipo. En **temperaturas**, Madrid y Zaragoza muestran una **amplitud diaria mayor** (mañanas más frescas y tardes más cálidas), rasgo típico **continental**; A Coruña y, sobre todo, **Tenerife** son **más templadas** y con **amplitud menor**, rasgo **litoral/subtropical**. Esta fotografía global no es para describir por describir: es **la línea cero** contra la que se va a comparar cada clúster.

¿Cómo se separarán y nombrarán los clústeres? Muy sencillo: para cada clúster se observará **en qué franjas** se aparta **por encima o por debajo** de ese patrón medio y **con qué combinación** de variables lo hace. No se busca saber si "tiene más o menos" a lo bruto, sino **cuándo** concentra la energía o la demanda, porque el **timing** es lo que más cambia el perfil de desvíos. Así, si un clúster tiene **mucha eólica de noche** respecto al global, **se llamará nocturno-eólico**; si su **fotovoltaica pesa más por la tarde** que por la mañana frente al patrón medio, será **vespertino-solar**; si a eso se suma **demanda más alta en la noche** de lo normal, se hablará de **punta tardía**. Del mismo modo, si el clúster está **sobrerrepresentado en fines de semana/festivos** frente al 0,5 global, lo se

etiquetará como **“de fin de semana”**; si, además, presenta **mañanas más frescas en Madrid/Zaragoza** y **tardes más cálidas** que la media, se hablará de que es **“continental”**; si por el contrario es más templado y con amplitud pequeña en A Coruña/Tenerife, se llamará **“litoral/subtropical”**. La nuclear seguirá siendo un buen **detector de contingencias**: si un clúster muestra una base distinta a la media, se podrá añadir el matiz **“con evento en base”**.

Para que esto no sea subjetivo, se trabajará con una lógica de **alto/medio/bajo** por franja que se define **comparando todos los clústeres entre sí**: lo que quede claramente por encima de lo habitual en un grupo se marcará como **alto**, lo claramente por debajo como **bajo** y el resto como **neutro**. Con esa codificación es muy fácil **construir el nombre** del clúster a partir de sus rasgos dominantes y, sobre todo, explicar **por qué**: “alto en eólica de madrugada y noche, alto en demanda de noche, FV más cargada a la tarde que la media → clúster nocturno-eólico con punta tardía y sesgo vespertino de FV”. Si un clúster cumple varias etiquetas, se ordenarán los rasgos por **fuerza** y se hablará de **uno principal** y **uno secundario**

Para **defender visualmente** estas diferencias se usarán unos mapas de calor que indican las principales características de estos grupos.

Clúster 0 — “Fin de semanas calurosos”

Este grupo está formado predominantemente por días comprendidos entre viernes y domingo (o festivos) que presentan temperaturas muy por encima de la media. Por el grado de festividad relativamente bajo en sus datos (inferior a 1 en la variable Festividad), se deduce que la mayoría de estos días son fines de semana normales (sábado o domingo) más que festivos nacionales específicos. Este sesgo hacia los fines de semana aparecerá de forma recurrente en varios clústeres: las variables dummy de días de la semana tienden a sesgar la agrupación para diferenciar claramente días laborables de días no laborables, ya que esta condición influye en muchas otras variables (principalmente la demanda eléctrica). Además de su componente de calendario, el clúster 0 se caracteriza por una baja contribución de la generación eólica (muy por debajo de la media) y por un pico tardío de generación fotovoltaica: en estos días se observa que la producción solar alcanza valores altos a últimas horas de la tarde/noche en comparación con el perfil promedio. Este comportamiento atípico de la fotovoltaica a última hora podría explicarse, en parte, por las altas temperaturas generalizadas en todo el país durante esos días, que retrasan el descenso de la irradiación útil o prolongan el periodo de alta generación solar al final del día. En conjunto, el clúster 0 correspondería a jornadas de fin de semana en época estival (calurosas) con *oversupply* fotovoltaico tardío y poca producción eólica.

Tabla 10: Desviaciones de las variables del clúster 0

Variable	Madrugada (0-5)	Mañana (6-11)	Tarde (12-17)	Noche (18-23)
Prevision_Foto	0,0%	-6,3%	-13,8%	28,1%
Eólica_Diaria	-25,1%	-37,6%	-32,6%	-19,5%
Instalada_Foto	-35,6%	-35,6%	-35,6%	-35,6%
Instalada_Eolica	-6,5%	-6,5%	-6,5%	-6,5%
Prev_Demanda	5,8%	2,2%	8,5%	2,4%
Disp_Nuclear	8,6%	8,7%	8,6%	8,5%
Festividad	-54,1%	-54,1%	-54,1%	-54,1%
Lunes	-37,8%	-37,8%	-37,8%	-37,8%
MartesJueves	-99,0%	-99,0%	-99,0%	-99,0%
ViernesDomingoFestivo	74,9%	74,9%	74,9%	74,9%
Temp_Madrid	61,0%	58,1%	49,1%	50,2%
Temp_Tenerife	11,9%	8,3%	6,8%	8,7%
Temp_Sevilla	44,8%	38,4%	33,9%	38,0%
Temp_Santiago	35,8%	33,3%	29,3%	30,5%
Temp_Coruña	23,5%	19,8%	21,0%	23,7%
Temp_Zaragoza	57,8%	48,6%	43,2%	49,2%

Clúster 1 — "Día estándar"

Este clúster agrupa la mayoría de los días laborables convencionales, típicamente aquellos de martes a jueves que no presentan ninguna particularidad destacada. Es el clúster más habitual en cuanto a frecuencia, lo cual es natural dado que los días laborales comunes son también los más numerosos en la realidad. Al ser el tipo de día más repetido, las tendencias internas de este clúster acaban siendo muy parecidas a las tendencias globales de todo el conjunto de datos. De hecho, las variables en este grupo tienden a situarse muy cerca de sus valores medios globales (desviaciones cercanas a 0%). Dicho de otro modo, el clúster 1 representa el comportamiento medio o base del sistema: demanda y generación previstas sin desviaciones notables, temperaturas en torno a la media histórica, etc. Constituye una especie de "patrón de referencia" frente al cual los demás clústeres mostrarán desviaciones.

Tabla 11: Desviaciones de las variables del clúster 1

Variable	Madrugada (0-5)	Mañana (6-11)	Tarde (12-17)	Noche (18-23)
Prevision_Foto	0,0%	1,4%	0,1%	0,1%
Eólica_Diaria	-1,8%	-2,3%	-2,0%	-1,8%
Instalada_Foto	-0,8%	-0,8%	-0,8%	-0,8%
Instalada_Eolica	-0,1%	-0,1%	-0,1%	-0,1%
Prev_Demanda	0,2%	0,7%	0,6%	0,3%
Disp_Nuclear	-0,1%	0,0%	0,0%	0,1%
Festividad	-81,7%	-81,7%	-81,7%	-81,7%
Lunes	-96,9%	-96,9%	-96,9%	-96,9%
MartesJueves	147,9%	147,9%	147,9%	147,9%
ViernesDomingoFestivo	-63,2%	-63,2%	-63,2%	-63,2%
Temp_Madrid	0,5%	0,4%	0,7%	0,7%
Temp_Tenerife	-1,6%	-1,7%	0,1%	0,7%
Temp_Sevilla	-0,2%	-0,2%	-0,2%	-0,2%
Temp_Santiago	0,9%	1,2%	0,6%	0,1%
Temp_Coruña	1,2%	-0,8%	-0,2%	1,3%
Temp_Zaragoza	1,8%	0,3%	0,4%	1,2%

Clúster 2 — "Lunes laborables fríos"

En este grupo aparecen principalmente los días clasificados como "lunes" laborales que se distinguen por ser significativamente más fríos de lo normal. Se trata del primer clúster donde un día de la semana específico (el lunes) cobra protagonismo. Estos lunes fríos ocurren tanto en el interior peninsular como en zonas costeras e incluso en Canarias, es decir, abarcan situaciones de bajas temperaturas generalizadas. Como consecuencia de las temperaturas más bajas, en este clúster se observa que la generación renovable prevista es inferior a la media, especialmente la fotovoltaica (los días fríos suelen tener menos radiación solar útil). Efectivamente, la previsión de generación solar de estos días se sitúa muy por debajo de la del resto de clústeres (en torno a un 40% menor en las horas centrales), acorde con la menor insolación típica de días invernales. Por lo demás, el clúster 2 refleja condiciones de un inicio de semana con demanda relativamente alta (propia de un lunes laboral) pero limitada en capacidad de generación renovable.

Tabla 12: Desviaciones de las variables del clúster 2

Variable	Madrugada (0-5)	Mañana (6-11)	Tarde (12-17)	Noche (18-23)
Prevision_Foto	0,0%	-41,1%	-38,0%	-40,4%
Eólica_Diaria	-5,0%	-5,8%	-4,8%	-5,5%
Instalada_Foto	-38,4%	-38,4%	-38,4%	-38,4%
Instalada_Eolica	-6,8%	-6,8%	-6,8%	-6,8%
Prev_Demanda	1,6%	2,6%	4,0%	1,5%
Disp_Nuclear	0,0%	0,1%	0,2%	0,2%
Festividad	-68,8%	-68,8%	-68,8%	-68,8%
Lunes	468,3%	468,3%	468,3%	468,3%
MartesJueves	-82,6%	-82,6%	-82,6%	-82,6%
ViernesDomingoFestivo	-96,0%	-96,0%	-96,0%	-96,0%
Temp_Madrid	-5,6%	-4,4%	-1,8%	-2,3%
Temp_Tenerife	-4,6%	-4,6%	-5,7%	-5,8%
Temp_Sevilla	0,6%	-0,9%	-1,4%	-0,4%
Temp_Santiago	-8,5%	-8,8%	-1,1%	0,1%
Temp_Coruña	2,7%	6,4%	3,2%	-0,3%
Temp_Zaragoza	-4,7%	-4,8%	-3,6%	-3,6%

Clúster 3 — "Fin de semana soleado"

Este clúster consiste en fines de semana (o festivos) caracterizados por una altísima producción solar instalada y prevista, combinada con temperaturas por debajo de la media. En estos días, la potencia fotovoltaica instalada y la previsión de generación solar alcanzan picos muy altos en comparación con el promedio, lo que indica que corresponden a periodos posteriores a una gran expansión solar (por ejemplo, años recientes) o días excepcionalmente despejados. Sin embargo, la temperatura ambiente resulta inferiores a lo normal, lo cual podría parecer contradictorio. Esta combinación. mucho sol, pero ambiente frío, sugiere posibles desvíos elevados: si se ha pronosticado una alta generación fotovoltaica, pero en la práctica las temperaturas bajas reducen algo la eficiencia o acompañan a menor demanda, es posible que la producción real no alcance completamente lo previsto o que no pueda absorberse toda la generación, generando desvíos. En otras palabras, el clúster 3 agrupa días de fin de semana (baja demanda) en épocas frías pero muy soleadas, donde se pueden producir excedentes de generación renovable respecto a la demanda, sobre todo durante las horas centrales del día. Esto podría ser un primer indicio de clúster asociado a desvíos netos positivos (exceso de generación programada) en ciertas franjas horarias.

Tabla 13: Desviaciones de las variables del clúster 3

Variable	Madrugada (0-5)	Mañana (6-11)	Tarde (12-17)	Noche (18-23)
Prevision_Foto	0,0%	23,6%	39,4%	23,9%
Eólica_Diaria	9,6%	8,7%	6,0%	8,0%
Instalada_Foto	47,6%	47,6%	47,6%	47,6%
Instalada_Eolica	7,9%	7,9%	7,9%	7,9%
Prev_Demanda	-0,6%	-1,2%	-2,8%	1,1%
Disp_Nuclear	7,4%	7,4%	7,2%	7,3%
Festividad	100,0%	100,0%	100,0%	100,0%
Lunes	-100,0%	-100,0%	-100,0%	-100,0%
MartesJueves	-90,3%	-90,3%	-90,3%	-90,3%
ViernesDomingoFestivo	89,0%	89,0%	89,0%	89,0%
Temp_Madrid	-6,9%	-8,4%	-5,3%	-3,7%
Temp_Tenerife	-1,2%	-2,5%	-1,2%	0,2%
Temp_Sevilla	-8,2%	-7,5%	-5,4%	-5,7%
Temp_Santiago	-2,4%	-3,5%	-4,4%	-3,7%
Temp_Coruña	-10,6%	-4,3%	-1,9%	-6,9%
Temp_Zaragoza	-8,4%	-7,4%	-6,6%	-7,0%

Clúster 4 — "Lunes laborables cálidos"

Este clúster también corresponde mayoritariamente a lunes laborales, pero a diferencia del clúster 2, aquí se trata de lunes atípicamente cálidos. Se distingue del caso de lunes fríos (clúster 2) principalmente en que tanto la generación eólica como la fotovoltaica previstas están por encima de la media, favorecidas por condiciones meteorológicas más benignas (vientos presentes y cielos despejados, respectivamente). Asimismo, las temperaturas en este grupo están muy por encima del promedio, especialmente durante la madrugada y la mañana. En estos lunes cálidos, por tanto, hay mayor disponibilidad de energías renovables que en un lunes típico: la curva de generación prevista muestra más aporte eólico nocturno y más producción solar diurna de lo usual en un comienzo de semana. Al mismo tiempo, al ser lunes (día laboral tras el fin de semana), la demanda prevista arranca relativamente baja de madrugada, pero sube con fuerza durante el día. En resumen, el clúster 4 representa lunes con temperaturas inusualmente altas y abundante recurso renovable, lo que podría implicar menores desvíos si se aprovecha bien la generación, o desvíos por exceso en caso de sobre-predicción de generación.

Tabla 14: Desviaciones de las variables del clúster 4

Variable	Madrugada (0-5)	Mañana (6-11)	Tarde (12-17)	Noche (18-23)
Prevision_Foto	0,0%	41,0%	38,5%	42,4%
Eólica_Diaria	7,1%	5,6%	3,0%	3,9%
Instalada_Foto	39,6%	39,6%	39,6%	39,6%
Instalada_Eolica	7,1%	7,1%	7,1%	7,1%
Prev_Demanda	-1,4%	-1,8%	-3,2%	-0,9%
Disp_Nuclear	0,5%	0,4%	0,5%	0,3%
Festividad	-82,2%	-82,2%	-82,2%	-82,2%
Lunes	420,9%	420,9%	420,9%	420,9%
MartesJueves	-58,8%	-58,8%	-58,8%	-58,8%
ViernesDomingoFestivo	-96,1%	-96,1%	-96,1%	-96,1%
Temp_Madrid	6,5%	5,7%	2,7%	3,0%
Temp_Tenerife	6,1%	3,8%	2,6%	4,1%
Temp_Sevilla	1,6%	3,6%	1,8%	0,1%
Temp_Santiago	7,1%	5,1%	2,8%	3,8%
Temp_Coruña	-1,7%	-1,3%	-3,1%	-3,7%
Temp_Zaragoza	2,5%	6,1%	4,0%	1,3%

Clúster 5 — "Días festivos fríos"

En este grupo se ven claramente representados muchos días festivos (nacionales o periodos vacacionales) que, además, registran temperaturas muy bajas. El grado de festividad medio de estos días es elevado, indicando que una buena parte coincide con festivos señalados (Navidad, Año Nuevo u otros puentes festivos), y la demanda eléctrica prevista suele ser considerablemente menor a la de un día laborable estándar. Por otra parte, las temperaturas gélidas de madrugada son una nota característica de este clúster, lo que repercute en la previsión solar (que es muchísimo más baja que la media, dado que los días fríos suelen ser invierno con menos horas de sol). En cambio, la generación eólica prevista no disminuye tanto, e incluso en las horas centrales del día tiende a ser superior a la de otros clústeres, indicando que en estos días fríos sopla el viento a media jornada ("levanta" a mitad del día). En suma, el clúster 5 refleja días no laborables con clima invernal muy frío, donde la producción renovable esperada se inclina más hacia la eólica que hacia la solar, y la demanda es reducida. Este cóctel puede dar lugar a desvíos importantes si la predicción de demanda o de generación no se ajusta a condiciones tan extremas.

Tabla 15: Desviaciones de las variables del clúster 5

Variable	Madrugada (0-5)	Mañana (6-11)	Tarde (12-17)	Noche (18-23)
Prevision_Foto	0,0%	-60,2%	-45,2%	-68,2%
Eólica_Diaria	8,2%	17,2%	15,3%	5,8%
Instalada_Foto	-37,5%	-37,5%	-37,5%	-37,5%
Instalada_Eolica	-6,5%	-6,5%	-6,5%	-6,5%
Prev_Demanda	-0,6%	1,4%	0,5%	0,2%
Disp_Nuclear	-1,1%	-1,1%	-1,1%	-1,1%
Festividad	155,2%	155,1%	155,1%	155,1%
Lunes	-100,0%	-100,0%	-100,0%	-100,0%
MartesJueves	-18,7%	-18,7%	-18,7%	-18,7%
ViernesDomingoFestivo	43,6%	43,6%	43,6%	43,6%
Temp_Madrid	-32,7%	-31,2%	-28,2%	-28,9%
Temp_Tenerife	-5,6%	-3,8%	-3,5%	-4,8%
Temp_Sevilla	-21,8%	-19,9%	-17,6%	-18,7%
Temp_Santiago	-23,1%	-20,7%	-17,9%	-18,7%
Temp_Coruña	-9,2%	-11,6%	-9,6%	-6,1%
Temp_Zaragoza	-33,6%	-28,2%	-24,3%	-27,8%

Clúster 6 — "Días festivos cálidos"

El último clúster identificado también agrupa días festivos o fines de semana, pero en este caso se trata de días muy cálidos. Es similar al clúster 5 en el sentido de que son días no laborables (con baja demanda prevista), pero difiere en las condiciones meteorológicas: aquí las temperaturas son muchísimo más altas que las típicas del clúster 5.Se caracteriza, por ejemplo, por festivos en verano u olas de calor en periodos vacacionales. En consecuencia, la demanda prevista en estos días es aún menor que la de clústeres laborables (por el efecto festivo) e incluso algo más baja de lo esperable por la temperatura (ya que en días muy cálidos la demanda de ciertas horas punta puede aplanarse, por menor actividad). Al mismo tiempo, la disponibilidad de generación renovable es muy elevada: la previsión de generación fotovoltaica es extremadamente alta (muy por encima de la media, llegando a más de 60% superior en ciertas horas) y la eólica también presenta aportes positivos. Este escenario, baja demanda, pero alta oferta renovable en días festivos cálidos, sugiere desvíos elevados en sentido contrario al clúster 5: aquí podrían darse excedentes de generación muy por encima de lo pronosticado para la demanda real. En otras palabras, el clúster 6 podría asociarse a desvíos netos negativos (generación real excedentaria respecto a la programada, es decir, sobrante) particularmente en las horas centrales del día, cuando la producción solar sobrepasa con creces la demanda reducida de un festivo.

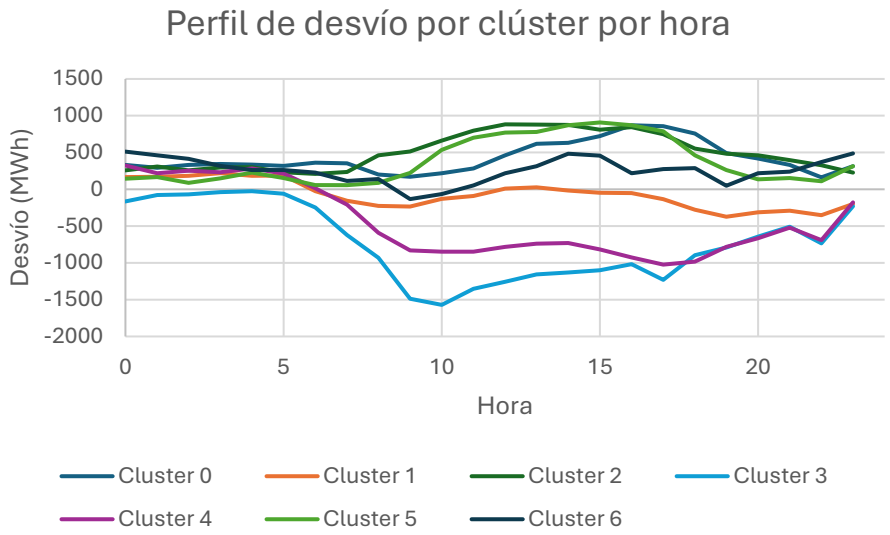
Tabla 16: Desviaciones de las variables del clúster 6

Variable	Madrugada (0-5)	Mañana (6-11)	Tarde (12-17)	Noche (18-23)
Prevision_Foto	0,0%	61,7%	23,3%	43,4%
Eólica_Diaria	0,1%	2,6%	5,5%	3,9%
Instalada_Foto	21,6%	21,6%	21,6%	21,6%
Instalada_Eolica	4,6%	4,6%	4,6%	4,6%
Prev_Demanda	-4,8%	-5,4%	-7,3%	-6,1%
Disp_Nuclear	-18,5%	-18,6%	-18,7%	-18,5%
Festividad	35,4%	35,5%	35,5%	35,5%
Lunes	-54,3%	-54,3%	-54,3%	-54,3%
MartesJueves	-100,0%	-100,0%	-100,0%	-100,0%
ViernesDomingoFestivo	80,7%	80,7%	80,7%	80,7%
Temp_Madrid	2,5%	5,2%	3,4%	1,3%
Temp_Tenerife	1,1%	3,2%	2,4%	0,9%
Temp_Sevilla	3,7%	5,3%	4,5%	3,1%
Temp_Santiago	4,7%	6,0%	4,4%	3,1%
Temp_Coruña	4,0%	4,7%	-1,4%	-3,1%
Temp_Zaragoza	7,6%	8,1%	6,2%	6,2%

Una vez caracterizados los clústeres en función de las variables independientes, el siguiente paso es examinar los desvíos asociados a cada tipo de día. Hay que recordar que el objetivo principal de este trabajo es comprobar si los clústeres (formados sin usar la información de desvíos) presentan algún patrón distintivo en sus desvíos diarios, de modo que se pueda anticipar el comportamiento de los desvíos conociendo el tipo de día. En esta sección se lleva a cabo ese análisis: se calcula el perfil de desvío correspondiente a cada clúster y se evalúa la variabilidad interna de dichos desvíos.

Para ello, se tomó el conjunto de días asignado a cada clúster y se calculó la media de los desvíos horarios de todos esos días, así como la desviación estándar correspondiente, obteniendo un perfil medio y una dispersión para cada grupo. La Ilustración 2 muestra, de manera comparativa, las curvas medias de desvío horario para cada uno de los 7 clústeres. En el eje horizontal se representa la hora del día (0 a 23 h) y en el vertical el desvío medio en MWh.

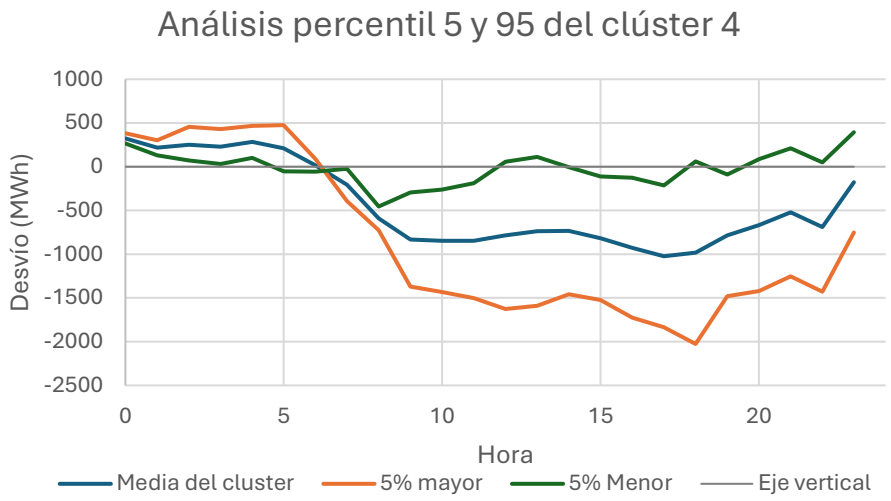
Ilustración 7: Perfil de desvío medio por clúster por hora



Del gráfico comparativo se desprenden hallazgos interesantes: las medias de desvío difieren notablemente según el clúster, lo que confirma que el tipo de día (según nuestras variables de entrada) influye en el comportamiento de los desvíos. Por ejemplo, algunos clústeres presentan desvíos medios positivos en ciertas franjas y otros negativos. En particular, se observa que en clústeres como el 3 (asociados a escenarios de alta generación renovable y demanda baja) tienden a mostrar desvíos netos negativos (exceso de generación sobre la demanda prevista, es decir, generación no aprovechada) durante las horas diurnas , se ven curvas con valores medios negativos en esas horas. Por otro lado, clústeres como el 4 (lunes cálidos) muestran desvíos medios positivos por la mañana y negativos por la tarde, indicando posibles déficits matutinos y excesos vespertinos de generación respecto a lo programado. Esto sugiere que, efectivamente, siguiendo las variables analizadas se pueden distinguir patrones de desvío típicos de cada grupo: conocer de antemano a qué clúster va a pertenecer un día da indicios sobre en qué horas es probable que tenga excedentes o déficits de generación. Ahora bien, no basta con conocer la media de los desvíos; es igualmente importante evaluar la dispersión interna de cada clúster. Si dentro de un mismo clúster los desvíos varían mucho de un día a otro (es decir, una desviación estándar muy alta), entonces el patrón medio podría no ser fiable. Una preocupación derivada del gráfico anterior es precisamente si algunos clústeres presentan desviaciones típicas tan grandes que sus medias no resulten estadísticamente significativas. Para investigar esto, se realizó un análisis de los valores extremos de los desvíos en cada clúster: se examinaron, para cada grupo, los casos más extremos de desvíos para ver si están sesgando la media o indicando falta de cohesión. Con el fin de no extender el análisis repetitivamente para todos los clústeres, se decidió ilustrar este estudio de sensibilidad con un clúster representativo y resumir los demás de forma cualitativa. En particular, se eligió el clúster 4 (lunes cálidos) para un análisis detallado de sus desvíos extremos, dado que (1) este clúster, junto con el 3, mostró una tendencia de desvíos opuesta a la de la mayoría (tiende a desvíos negativos en varias horas, frente a otros con desvíos positivos), y (2) si la mayoría de los días de este clúster presentasen desvíos negativos significativos, se podría plantear una estrategia específica para aprovechar esos desvíos. Se comparó el 5% de los días con menores desvíos del clúster 4 con el 5% de los días con mayores desvíos, para comprobar si alguno de estos subconjuntos extremos

estaba distorsionando la media global del clúster más de lo deseable. Los resultados de esta comparación se presentan en la Ilustración 3, que muestra los perfiles de desvío correspondientes al percentil 5 y al percentil 95 dentro del clúster 4:

Ilustración 8: Análisis percentil 5 y 95 del clúster 4



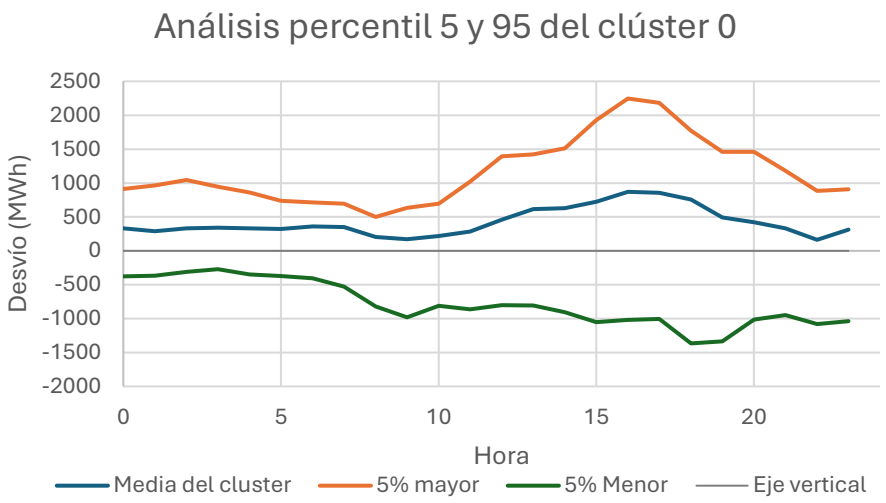
Es muy importante entender el análisis que se ha llevado a cabo para que no haya confusión en este tipo de gráficos. Lo que se ha llevado a cabo es un análisis de variabilidad, en el que para cada clúster, se ha observado el tipo de desvío, ya sea mayoritariamente negativo o mayoritariamente positivo, y se han seleccionado el 5% de los valores no más negativos a una hora concreta, sino los días cuyos perfiles globales suponían ser más negativos, y en casos como este el del clúster 4, como la mayoría del perfil es negativo, el 5% menor no son positivos, sino que simplemente son menos negativos y más cercanos al 0, causando ese cruce e las 6 a.m., puesto que los días del 5% de los valores menores, tienen un desvío positivo por la mañana, pero eso no implica que tengan uno mucho más negativo más adelante.

Los resultados del análisis de extremos permiten concluir que, pese a la variabilidad interna, sí existe una relación consistente entre el tipo de clúster y el perfil de desvíos. En el caso concreto del clúster 4 (lunes cálidos), su perfil de desvíos horario está muy marcado: como se comentó, presenta desvíos positivos por la mañana (la generación real resulta insuficiente en las primeras horas, o la demanda real supera a la prevista) y desvíos negativos muy acentuados por la tarde (exceso de generación respecto a la demanda). Esta pauta se mantiene incluso considerando la dispersión: aunque haya días con desviaciones mayores o menores, la mayoría sigue esa tendencia general. Por tanto, se podría diseñar una estrategia operativa que, al identificar que el día siguiente será de tipo “lunes cálido” (clúster 4), anticipe excedentes de generación vespertinos y déficit matutinos, permitiendo tomar medidas como modificar programas de generación o activar reservas para equilibrar.

Ahora bien, no todos los clústeres ofrecen patrones tan claramente aprovechables. Por ejemplo, el clúster 0 (fines de semana calurosos) muestra un comportamiento interno más complejo: del análisis de sus extremos se desprende que el 5% de los días con mayores desvíos en este clúster presentan desvíos muy positivos (generación real muy por debajo de la programada), mientras que el 5% con menores desvíos son fuertemente negativos (generación muy por encima de lo previsto). Es decir, dentro del propio clúster 0 hay días con comportamientos de desvío opuestos. Esto sugiere que la distribución de desvíos en

ese grupo es bastante dispersa y bimodal, lo cual dificulta extraer una única estrategia de ajuste válida para todos esos días. En este sentido, el clúster 0 podría considerarse menos consistente en términos de desvíos, representando quizás un conjunto heterogéneo de situaciones de fin de semana con comportamientos variados (algunos casos con falta de generación, otros con exceso). Este hecho plantea una posible limitación para futuros análisis: habrá clústeres que requieran una subdivisión adicional o el uso de información extra para poder predecir sus desvíos con precisión.

Ilustración 9: Análisis percentil 5 y 95 del clúster 0



Y esta correlación va variando según el clúster, hay algunos muy estables donde los desvíos siguen todos la misma tendencia (recordando como el 5% mayor es siempre el grupo del 5% que en mayor absoluto tuviesen más media, por lo que puede ser para algunos clústeres negativo y para otros positivo, y pudiendo cruzar al del 5% mayor, si el perfil de esos días así lo demuestra). Teniendo así dos grupos claros, un primero que incluye a clústeres como el 1 el 2 o el 5:

Ilustración 10: Análisis percentil 5 y 95 del clúster 1

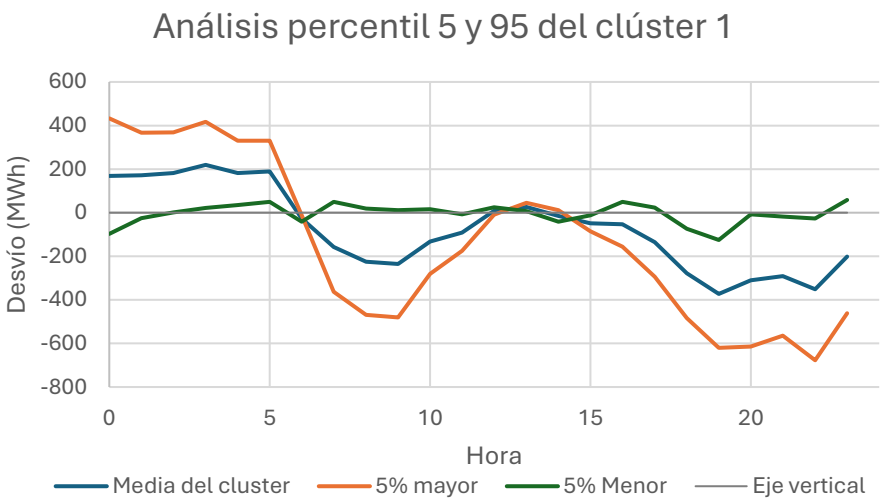


Ilustración 11: Análisis percentil 5 y 95 del clúster 2

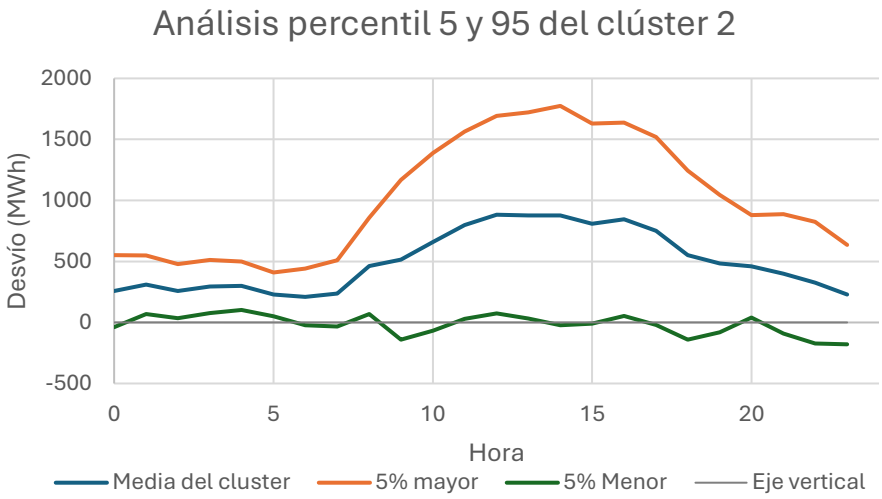
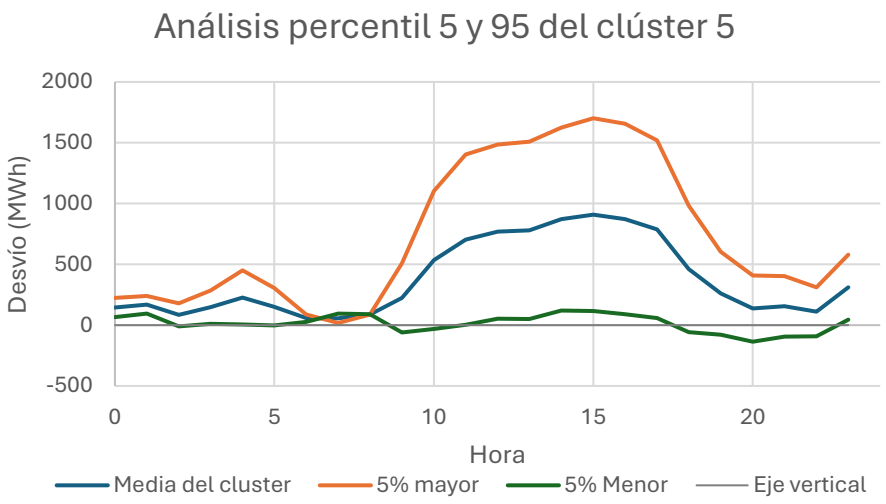


Ilustración 12: Análisis percentil 5 y 95 del clúster 5



Y otro grupo, que presenta comportamientos mucho más dispares y difíciles de predecir como el 3 y el 6, donde un por la complicación de los desvíos (como puede ser el 6) y otro por que algunos de los desvíos son negativos y otros muchos positivos, como puede ser el 3. Esto puede ser un indicativo de la similitud de estos días y que el algoritmo K-means los ha podido confundir, teniendo así un perfil de desvíos poco claro. Estos serían sus perfiles:

Ilustración 13:Análisis percentil 5 y 95 del clúster 3

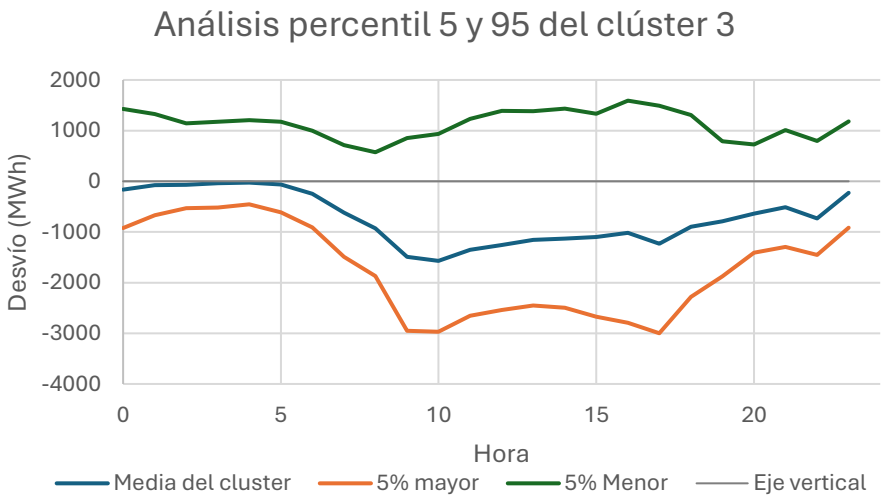
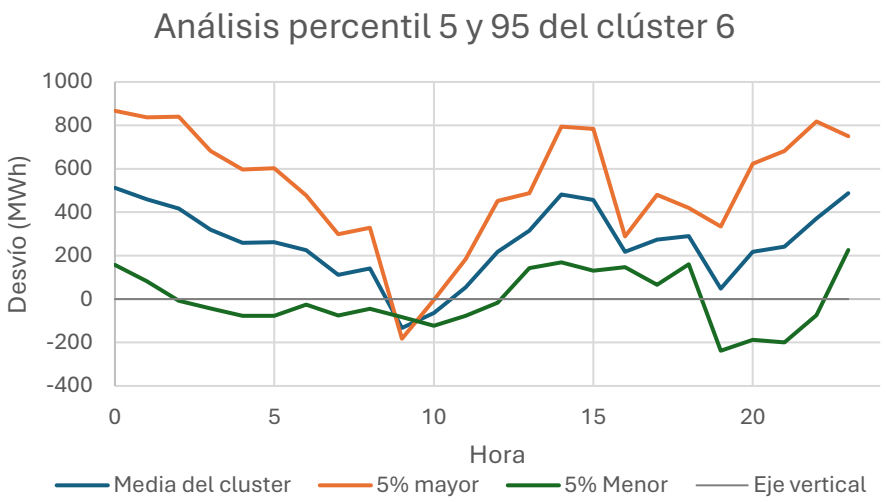


Ilustración 14: Análisis percentil 5 y 95 del clúster 6



En conjunto, estos resultados son prometedores: demuestran que existe cierta correlación entre los tipos de días (clústeres) y los patrones de desvío del sistema. Esto cumple en buena medida el objetivo planteado de relacionar variables operativas y meteorológicas con los desvíos diarios. Sin embargo, también ponen de manifiesto la necesidad de un análisis cuidadoso clúster a clúster, ya que algunos grupos presentan variabilidad interna elevada que podría limitar la capacidad predictiva simple. En los capítulos siguientes se extraen conclusiones generales y se proponen recomendaciones, así como posibles líneas de trabajo futuro para abordar las limitaciones observadas (por ejemplo, refinamiento de clústeres o incorporación de nuevas variables al análisis).

Capítulo 4.- Estudio Económico

4.1.- Introducción a los precios de los desvíos

La segunda parte y esencial de este proyecto, pasa por darle una visión económica a todo lo elaborado anteriormente, para ello, se propondrá la siguiente estrategia, basada en la siguiente pregunta ¿Si se fuese capaz de predecir el tipo de día que va a ser el día siguiente, gracias a un modelo basado en machine learning, se podría sacar ventaja económica de esa información?

Ante dicha pregunta, es imperativo entender bien cómo funciona el mercado eléctrico español, sobre todo el que tiene que ver con los desvíos a tiempo real. Estos desvíos son la diferencia entre la energía que un agente programó (ofertó/ compró en el mercado) y la energía realmente medida en tiempo real. Cuando esa diferencia es positiva o negativa, el Operador del Sistema (REE, en el caso de España) clasifica el desvío como “a subir” o “a bajar” según su sentido:

- **Desvíos a bajar**: Ocurren cuando la producción medida es menor que la programada o el consumo real es mayor que el programado [REE25]. En otras palabras, hay déficit de energía: falta generación o sobra demanda respecto al plan. El sistema, para mantener el balance, tiene que aumentar generación (o reducir consumos de bombeo) vía servicios de ajuste en tiempo real. Un ejemplo sería un generador que se quedó corto (produjo menos de lo previsto) o un consumidor que consumió de más; en ambos casos el sistema necesita energía adicional.
- **Desvíos a subir:** Se dan cuando la producción medida es mayor que la programada o el consumo real es menor que el previsto [REE25]. Aquí hay un exceso de energía: sobra generación o falta demanda frente al programa. El operador debe reducir producción (o incrementar consumos flexibles, como bombeos) a través de los mercados de ajuste para restablecer el equilibrio. Un ejemplo es un productor que generó de más o un cliente que consumió menos de lo esperado; ambos aportan energía no solicitada que el sistema debe absorber. En resumen, un desvío "a bajar" implica que el sistema tuvo que subir generación para cubrir un déficit, mientras que un desvío "a subir" implica que el sistema tuvo que bajar generación para eliminar un excedente. Esta nomenclatura se define desde la perspectiva del sistema (operador) pero aplica igual a cualquier agente que se desvía. Todos los desvíos involuntarios producen un desequilibrio generación-demanda que el operador corrige activando reservas y servicios de balance en tiempo real.

Una vez, vistos los desvíos habrá que entender cuál es el precio asociado a esos desvíos. Los precios de desvío a subir y a bajar son los precios unitarios (€/MWh) a los que se liquidan esas diferencias de energía frente al programa. Estos precios los determina el operador del sistema ex-post (a posteriori), en función de los costes reales de equilibrar el sistema en cada intervalo horario [DURÁ14]. En la práctica, España ha operado históricamente con un esquema de doble precio de desvío, ligado al precio del mercado diario pero ajustado según la necesidad de balance real:

- **El precio del desvío a bajar** (para agentes con déficit de energía) nunca será inferior al precio del mercado diario [CNMC22]. Esto significa que quien se quedó corto de energía pagará como mínimo el mismo precio que la energía en el mercado principal, y si equilibrar ese déficit resultó caro para el sistema, pagará más. En términos regulatorios, si el sistema estaba deficitario en esa hora, el precio de desvío a bajar se fija como el máximo entre el precio marginal diario (PMD) y el precio medio de las activaciones para subir (es decir, el coste medio de la energía extra movilizada para cubrir el faltante) [CNMC22]. Así se asegura que un agente que cause un desvío negativo no se beneficie de que la energía de ajuste pudiera ser más barata que el mercado; por el contrario, suele incurrir en un sobrecoste si su desvío ocurre en momentos de escasez de energía.
- El precio del desvío a subir (para agentes con exceso de energía) nunca superará al precio del mercado diario [CNMC22]. Es decir, quien entregó energía de más o consumió de menos no cobrará por ese excedente un precio mayor que el de mercado, y si remover dicho excedente tuvo un coste menor para el operador, se le pagará menos. En la práctica, si el sistema estaba excedentario (sobraba generación) en esa hora, el precio de desvío a subir se calcula como el mínimo entre el precio diario y el precio medio de las activaciones para bajar (el coste medio de reducir generación o desviar energía sobrante) [CNMC22]. De esta forma, un agente que provoque un desvío positivo no obtiene ventaja vendiendo energía no programada a un precio mayor que el base; de hecho, usualmente recibirá un precio reducido si su desvío ocurre cuando hay exceso de oferta en el sistema.

En condiciones normales, cuando el sistema no requiere acciones de balance extraordinarias en una dirección, el precio de desvío tiende a coincidir con el precio del mercado diario. De hecho, si en una hora dada no hubo desequilibrio neto significativo, los desvíos de todos se liquidan al precio marginal horario habitual (no hay penalización). Sin embargo, en horas donde sí hizo falta activar regulaciones (ya sea subir o bajar producción), los precios de desvío se apartan del precio base para reflejar el coste de esas intervenciones del operador. En resumen: el desvío a bajar suele encarecerse por encima del mercado en escenarios de déficit de energía, y el desvío a subir se abarata por debajo del mercado en escenarios de excedente, siempre respetando las reglas de no ser peor que el caso base para quien ayudó al sistema ni mejor para quien lo empeoró [CNMC22].

¿Cómo obtiene REE esos precios? Simplificadamente, los calcula ponderando los precios de todas las acciones de ajuste activadas en la hora (regulación secundaria, terciaria, gestión de desvíos, etc.) y comparándolos con el precio del mercado diario de esa hora [BOE22]. El resultado es publicado en las liquidaciones horarias de desvíos. Por ejemplo, si a una determinada hora el sistema quedó corto y fue necesario activar generación adicional (terciaria o desvíos) a 150 €/MWh mientras el precio diario era 140 €/MWh, el precio de desvío a bajar para esa hora sería 150 €/MWh (el mayor de ambos) y ese será el precio que paguen por su energía faltante los agentes deficitarios. Por el contrario, si en otra hora sobró energía y el operador tuvo que reducir producción (p.ej. recortando generación renovable a un coste implícito de 20 €/MWh cuando el precio diario era 30 €/MWh), el precio de desvío a subir resultante sería 20 €/MWh (el menor de ambos) y ese será el precio al que se liquidará la energía excedentaria de quienes se desviaron al alza. En todo caso, estos precios se determinan a posteriori una vez conocida la realidad del sistema; no son fijos de antemano, sino que dependen de las condiciones de balance de cada hora y de las ofertas en los mercados de ajuste correspondientes [BOE22].

A partir de recientes regulaciones europeas de balance (Reglamento EB GL), el sistema español está evolucionando hacia un esquema de precio único de desvío por hora en lugar del dual tradicional, de forma que todos los desvíos se liquidarían al mismo precio marginal de balance por hora, independientemente de su signo. [UE17]

Los precios de los desvíos están directamente ligados al estado de equilibrio o desequilibrio del sistema eléctrico en tiempo real. En particular, el volumen total de desvío del sistema en cada intervalo (es decir, cuánto excedente o déficit neto hubo considerado todos los agentes) influye fuertemente en el precio resultante de los desvíos para ese intervalo. Si la suma de los desvíos de todos los participantes es grande, significa que el operador ha tenido que hacer ajustes significativos (movilizar muchas reservas o alterar la producción/consumo programado) para mantener la frecuencia y el balance; consecuentemente, mayores ajustes implican mayores costes, que se traducen en precios de desvío más extremos (más altos en caso de déficit, o más bajos en caso de excedentes).

En situaciones tranquilas, sin grandes desviaciones agregadas, suele no haber una desviación fuerte del precio de desvío respecto al mercado. De hecho, como se indicó, si el sistema está prácticamente equilibrado o con ligera sobra/defecto, el precio de los desvíos acaba siendo el del mercado diario (no hay penalización neta) [CNMC22]. Esto explica por qué en promedio los precios de desvío tienden a seguir la misma tendencia que el mercado mayorista base. Sin embargo, ante eventos imprevistos o mala planificación colectiva, los precios de desvío pueden dispararse o desplomarse respecto al mercado, reflejando el esfuerzo de balance: Por ejemplo, si súbitamente una gran central (como podría ser una nuclear) sufre una avería y sale de servicio, el sistema enfrenta un déficit brusco de generación. REE deberá activar rápidamente reservas de generación (y/o importar energía si es posible) a precios probablemente muy elevados en ese momento. Ese coste extra se repartirá entre todos los agentes desviados en esa hora, resultando en un precio de desvío a bajar muy alto (muy por encima del precio de mercado de esa hora) que penalizará a cualquier agente que estuviese corto de energía. De hecho, casos extremos en los servicios de ajuste han llegado a precios de cientos o miles de euros por MWh en situaciones críticas de desbalance. Todos los consumidores y productores finalmente terminan asumiendo parte de ese coste en las liquidaciones de desvíos, ya que es un coste conjunto de mantener la estabilidad del sistema. Inversamente, si en una determinada hora hay una sobreoferta masiva de generación (por ejemplo, se preveía mucha demanda o menos viento, pero finalmente la demanda cayó o el viento sopló más de lo esperado, generando excedentes), el operador tendrá que recortar producción (p. ej., ordenando a parques eólicos que reduzcan potencia, o desviando energía a bombeos). Esa energía sobrante puede incluso no tener salida, lo que puede llevar a precios de balance muy bajos o cercanos a cero en el desvío a subir de esa hora. En tal caso, los generadores que produjeron de más cobrarán por ese excedente un precio bajo (posiblemente cercano a 0 €/ MWh si había que tirar energía) y los consumidores que consumieron de menos verán que su energía no consumida vale muy poco en esa hora. Nuevamente, el precio de desvío refleja el coste que tuvo el sistema para lidiar con la abundancia de energía: si tuvo que pagar a algunos generadores para que se desconectaran, ese pago reduce el precio resultante que reciben quienes entregaron energía no deseada. En síntesis, los precios de los desvíos actúan como señales económicas del desequilibrio del sistema: penalizan las desviaciones que agravan el desbalance en momentos críticos y, en cambio, a las desviaciones que ayudan a corregir el equilibrio se les aplica el precio base (no sufren penalización). Un agente que, por casualidad o buena gestión, se desvía a favor del sistema

(por ejemplo, genera de más cuando faltaba energía, o consume de menos en un pico de demanda) generalmente liquidará ese desvío al precio de mercado como si lo hubiera planificado así, sin coste adicional. Por el contrario, quien se desvía en contra de las necesidades del sistema (faltando a su suministro en un momento de déficit o volcando excedente en un momento de sobra) pagará un precio peor que el de mercado, asumiendo su parte del coste de las medidas de emergencia tomadas por REE.

Esta relación tiene importancia económica significativa: incurrir en desvíos en los momentos "equivocados" puede ser muy costoso. Por ejemplo, un consumidor directo que termine demandando 0,33 MWh por encima de lo previsto en una hora de sistema escaso podría pagar por esa energía un ~20% más caro que el precio OMIE de esa hora (porque se le aplica el precio de desvío a bajar, que incluye recargo). Por el contrario, si ese desvío extra de consumo ocurre en una hora sin tensión (o si hubiera reducido consumo en hora de déficit), el coste adicional sería nulo o muy pequeño. Todos los agentes, por tanto, tienen el incentivo de minimizar sus desvíos o al menos alinearlos con la condición del sistema para evitar penalizaciones.

Y ¿cómo aplicaría entonces la primera parte de este trabajo a la segunda? Sencillo, si se puede saber qué tipo de día va a ser mañana, se puede explotar de alguna forma, y todo gracias a la agrupación en clústeres. Gracias a esto se puede intentar asumir que tipo de desvíos va a tener el sistema un día determinado, e intentar apoyarlo en vez de ir en contra de él.

El objetivo fundamental de este enfoque es doble:

1. **Reducir el impacto de los desvíos:** La clusterización permite identificar en qué condiciones (tipo de día, clima, generación renovable) los desvíos son más altos, lo que puede ayudar a ajustar las previsiones de generación o consumo y así reducir la probabilidad de incurrir en penalizaciones económicas por desvíos. Esto es especialmente útil si se supiese, por ejemplo, que, en un día con ciertas características, el sistema tendrá una **probabilidad alta de tener desvíos a subir**, es decir, cuando la energía generada supera la demanda prevista. En este caso, si **ya se predice que el sistema va a estar excedentario**, se podría optar por **pedir más energía** de la que realmente se necesita. De este modo, si posteriormente los precios de desvío a bajar son elevados, en lugar de pagar por un desvío negativo, por ejemplo:

   Si normalmente a cierta hora se tuviesen un desvío de 5MWH positivos, es decir, se ha consumido de menos, ahora se tendría que liquidarlo de nuevo al sistema, y esa energía se cobraría a un precio PDS<PMD por lo que se perdería parte del dinero. Mientras que si se pudiese saber que los PDS<<PMD, mientras que los PDB≈PMD, la penalización sería de una cantidad menor.

   Esta estrategia no solo depende de minimizar los desvíos, sino de **anticipar correctamente el comportamiento del sistema** y actuar para mitigar las penalizaciones asociadas.

2. **Prever escenarios de desvíos y reaccionar adecuadamente:** Si se pudiese saber, por ejemplo, que en un escenario con **exceso de generación** o **bajo consumo** el precio de los desvíos a subir será bajo (o incluso negativo), se pueden aprovechar esos **periodos de precios bajos** para **almacenar más energía** o **ajustar la**

producción a la baja, evitando así tener que pagar por un desvío a subir. La clave aquí es **no esperar a que el desvío ocurra**, sino tomar decisiones proactivas basadas en los patrones históricos identificados por la clústerización, lo que permite adaptarse al contexto y evitar penalizaciones por desvíos imprevistos.

Por ejemplo, aplicando técnicas de **clústerización** a los datos históricos de desvíos y las condiciones operativas del sistema (como la generación renovable, la demanda de electricidad y las previsiones meteorológicas), se pueden encontrar agrupaciones de días o periodos con **características similares** que conducen a desvíos recurrentes. Estos clústeres pueden incluir:

- **Escenarios de sobreoferta o desvíos a subir:** Días con alta generación renovable (por ejemplo, mucho viento o sol) y baja demanda, donde el sistema se ve obligado a **reducir generación** para evitar la sobrecarga. Si, al analizar el clúster, se sabe que **es muy probable que haya desvíos a subir**, se pueden ajustar las compras de energía o almacenamiento de manera anticipada, **evitando la penalización económica** por excedentes.
- **Escenarios de escasez o desvíos a bajar:** Días con baja producción renovable y alta demanda (por ejemplo, días fríos sin viento), donde el sistema necesita **comprar energía adicional** a precios elevados. Si este clúster es identificado con antelación, se puede **garantizar un suministro adecuado** y ajustar las previsiones de consumo para **evitar quedar cortos**, reduciendo así la probabilidad de tener que pagar penalizaciones elevadas por desvíos negativos.

Lo importante en este enfoque es que, al identificar los **patrones históricos de desvíos y precios**, se pueden **anticipar las condiciones que provocarán los mayores desvíos** y actuar en consecuencia. La **clústerización** ayuda a agrupar estos **escenarios típicos** y proporciona una base sólida para hacer **ajustes estratégicos** en la operación, ya sea aumentando la previsión de energía en momentos de alta generación o asegurando un suministro adecuado cuando el sistema esté en riesgo de déficit.

En resumen, la clústerización no solo permite **identificar patrones y escenarios de desvíos**, sino que ofrece una herramienta para **prever y mitigar los costes asociados a esos desvíos**. Con una correcta identificación de estos patrones, los agentes del mercado pueden tomar decisiones proactivas, ajustando sus planes de consumo y generación para **reducir el impacto de las penalizaciones por desvíos** y mejorar la eficiencia del sistema.

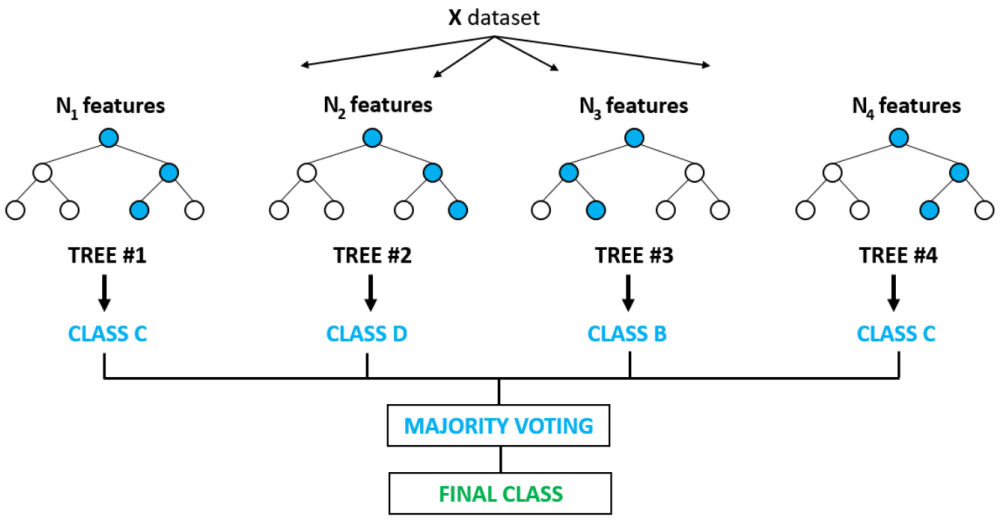
4.2.- Predicción del tipo de día

Ahora que se ha visto la importancia que podría tener el predecir el tipo de día, ahora se procederá a hacerlo. Para ello, se empezará usando el algoritmo Random Forest, que se basa en ir entrenando, en base a distintos árboles de segmentación distintos y sesgados, características comunes, para luego, predecir con unos datos de testeo cuál ha sido el tipo de día.

El algoritmo Random Forest, ya explicado en el estado de la cuestión, se basa en elaborar distintos árboles de decisión, cada uno de manera aleatoria. La ventaja de este método es que elabora muchos árboles de decisión aleatorios, eligiendo de cada uno la característica más destacable para luego poder tomar la decisión al clasificar. Al hacer esto, se consigue que, aunque los distintos árboles estén sesgados, la decisión final tenga la menor varianza posible y disminuye el sesgo general que puedan tener los datos, consiguiendo un modelo

con una capacidad de predicción bastante alta [GERÓ19]. Con los distintos árboles, luego intenta extrapolar a los nuevos datos, y según las distintas decisiones de cada árbol toma una decisión. Gráficamente lo que hace es:

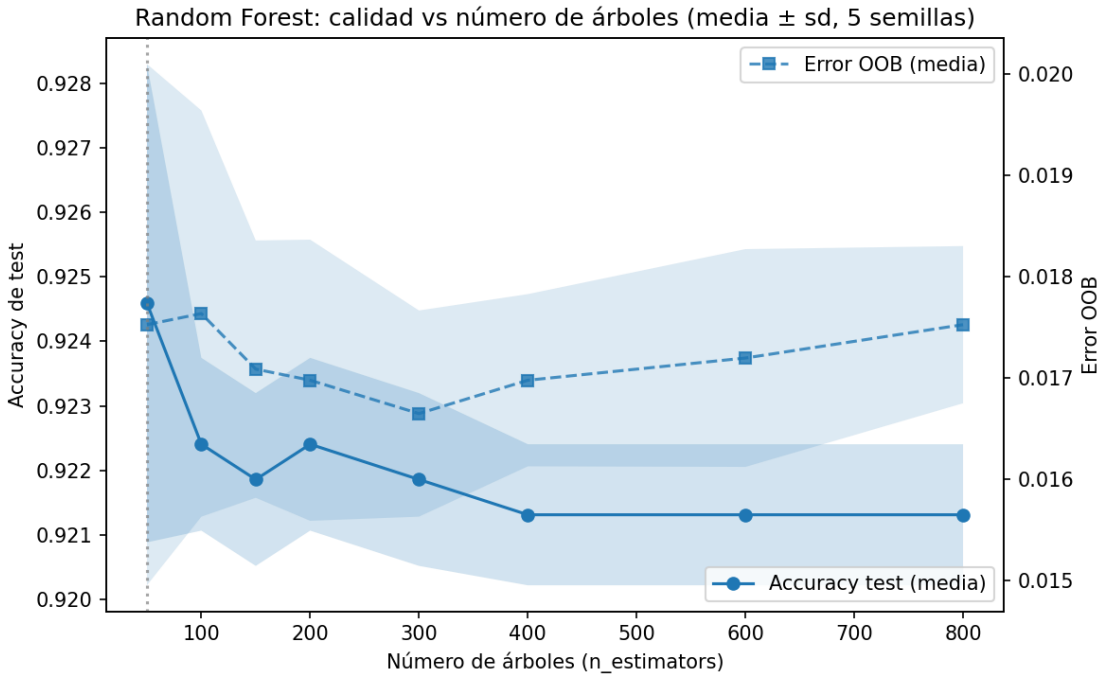
Ilustración 15: Gráfico de actuación del Random Forest [GÁME21]



El siguiente paso, será ir elaborando el algoritmo en Python y con el código ir evaluando los resultados hasta encontrar el modelo adecuado. Para ello, lo primero que se tendrá que hacer es definir cuáles van a ser los datos de entrenamiento y cuáles serán los de testeo. Una de las razones fundamentales por las que se fue atrás tanto en el tiempo es tener la capacidad suficiente para entrenar a nuestro modelo, así que se usarán los datos del 1 de abril de 2018 al 30 de marzo de 2024 como entrenamiento (es decir 5 años formarán todo el conjunto de datos de entrenamiento), y los que vayan desde el 1 de abril del 2024 hasta el 30 de marzo de 2025 como testeo.

El siguiente paso será ajustar el número de árboles adecuados para el problema en cuestión, para ello el primer paso es comprobar cuál es el número óptimo a través de ver cuando se estabiliza el error estándar por aumentar el número de árboles. Para ello, se llevará a cabo una comprobación con el *out-of-bag* (OOB) *score*. En un Random Forest, cada árbol se entrena con una muestra *bootstrap* del conjunto de entrenamiento; por eso, de cada observación, algunos árboles **no la ven** al entrenar. El **OOB** aprovecha esto para estimar el error "gratis": para cada observación de entrenamiento se hace una predicción **solo con los árboles que no la usaron** y se compara con su etiqueta real; al promediar sobre todas, se obtiene una precisión OOB. En **scikit-learn**, esa métrica se expone como `rf.oob_score_` (precisión) y su complemento es el **error OOB** (`1 - oob_score_`) [GERÓ19]. La otra medida es la precisión *(accuracy),* esto es el número de predicciones correctas hechas. Por último, habrá que medir la desviación estándar de estas medidas empezando por distintas semillas, u agrupaciones, pues no se querría que por elegir una semilla u otra se estuviera sesgando la decisión. Comparando estos dos parámetros, y sus desviaciones típicas, se puede ver cuál es el número óptimo de árboles. Estos fueron los resultados:

Ilustración 16: OOB score y accuracy test



En la Figura se aprecia que el modelo entra en **meseta** a partir de ~100–150 árboles. Con **50** árboles se obtiene la **mayor accuracy media** (0,9246), pero con **variabilidad** entre semillas más alta (desv. típica 0,003706 ≈ 0,371 %). Al pasar a **100–200** árboles la accuracy de test apenas varía (0,9224 con 100 y **0,9224 con 200**, es decir, −0,22 p.p. respecto a 50), mientras que la **variabilidad** cae a **~0,00134** (≈ 0,134 %). El **error OOB** disminuye con el tamaño del bosque hasta **300** árboles (mínimo 0,01665), siendo prácticamente igual en **200** (0,01698). Por criterios de **estabilidad** y **parquedad**, se fija **n_estimators = 200**: mantiene una precisión equivalente a la mejor alternativa en meseta, reduce la variabilidad frente a 50 árboles y alcanza un OOB casi mínimo. (Si se priorizara exclusivamente el OOB, **300** árboles sería una elección igualmente válida con un coste computacional algo mayor).

Con esto, ya se puede realizar las predicciones con nuestro modelo de Random Forest. Todo el código relacionado con el análisis a continuación se encuentra en el Anexo 9. Tras ejecutar el código aquí están los resultados principales y la matriz de confusión. Lo primero que se puede observar es que no se han predicho días del tipo 0 ni del tipo 2, esto se debe, a un sesgo de los datos que luego se tratará en las conclusiones, ya que para estos clústeres, la potencia instalada fotovoltaica es mucho menor que la media, y esto es algo producido por los datos de tipo temporal con los que se trabaja en este modelo, ya que la potencia instalada sigue aumentando año tras año, y nunca disminuye, por lo que los clústeres 0 y 2, se han quedado atrás tecnológicamente y es posible que no sean detectados nunca más debido a que la potencia instalada solo ha aumentado. Aunque esto pueda suponer una limitación, luego en las conclusiones se explicará por que se ha decidido seguir con esta distribución de los datos a pesar de limitaciones como esta.

La matriz de confusión del modelo es:

Tabla 17: Matriz de confusión del modelo de Random Forest

		Predicho						
		0	**1**	**2**	**3**	**4**	**5**	**6**
Clúster real	**0**	0	0	0	0	0	0	0
	1	0	117	0	0	0	0	0
	2	0	0	0	0	0	0	0
	3	0	0	0	114	0	0	24
	4	0	0	0	0	60	0	0
	5	0	0	0	0	0	0	0
	6	0	0	0	4	0	0	47

Y la precisión tanto global como de cada uno de los clústeres:

Tabla 18: Tasa de acierto del modelo Random Forest

	Precisión
Global	92,35%
Clúster 1	100,00%
Clúster 3	96,61%
Clúster 4	100,00%
Clúster 6	66,20%

Como se puede observar, la precisión general del modelo es muy alta 92,35%, por lo que se puede estar bastante convencidos de que es un modelo fiable. De todas formas, se pondrá a prueba uno de los fallos del modelo para llegar a entender si existe algún tipo de sesgo interno, para ello se estudiará un ejemplo de día que no fue clasificado correctamente, como podría ser el 2 de abril de 2024, que fue predicho como 6 cuando verdaderamente era del clúster número 3.

- Si se observa el porcentaje de árboles que lo clasificaron como parte del clúster 6, se puede ver que es un 60% de ellos, mientras que del clúster 3, tan solo un 38%.
- Para seguir profundizando en este ejemplo, se observarán los detalles que más guiaron a dicha decisión como pueden ser las variables que más han guiado a elegir el clúster 3 por encima del 6 en este caso concreto. Para ello, se observarán las dos variables locales que más han afectado a la decisión y se clasificarán según la aparición en las rutas de los árboles de decisión (Imagen 5 se pueden ver las rutas de decisión de manera gráfica), y segundo la importancia en la decisión. Están serían:
    - La potencia instalada Fotovoltaica a nivel nacional (que como es una constante para todo el día, afecta a la decisión con un peso 24 veces mayor al estar divida la decisión por las 24 horas del día)
    - La segunda, la temperatura en Zaragoza, presente en 190 de los 200 árboles de decisión, pues parece ser que este día las temperaturas medias en zaragoza fueron entre 4 y 10ºC menores de lo esperado, y han sido estas dos variables las que han hecho que este día haya sido clasificado como un tipo y no como otro.

- Con este análisis, se puede entrever como la similitud entre estos dos clústeres es grande, puesto que solo les diferencian cosas pequeñas como pueden ser días más fríos de lo que deberían, o que se ha instalado más capacidad fotovoltaica.

Siguiendo ese ejemplo concreto, se ha comprobado si la confusión **3→6** es "lógica", es decir, si ambos clústeres son realmente parecidos en el espacio de características. Para ello se comparan sus **centroides estandarizados** (cada componente de las 24×p variables está en z-score, por lo que todas pesan por igual) con tres índices complementarios.

1. En primer lugar, la **distancia euclídea entre centroides en z-espacio** cuantifica la **proximidad absoluta**: valores pequeños implican clústeres próximos. En el par 3–6 es **10,792**, mientras que la **media de todas las distancias inter-clúster** es **17,894** con **desviación típica 2,987**; esto sitúa a 3–6 **2,38 sigmas por debajo** de la media (percentil ≈ **4,8**), por lo que está entre los pares **más cercanos** del conjunto [TAN19].
2. La **similitud coseno** mide cuánto **apuntan en la misma dirección** los dos centroides (1 = idénticos, 0 = ortogonales, −1 = opuestos). El valor **0,454** indica **alineación moderada**: no son iguales, pero sí guardan una orientación común apreciable [TAN19].
3. La **correlación de Pearson** capta la **co-variación lineal global** (misma escala relativa de subperfiles a lo largo de horas y variables, independientemente del nivel medio). El valor **0,478** refuerza esa **similitud de forma** de rango **moderado-positivo** [TAN19].

La tabla de "Top pares confundidos" muestra, además, que la confusión es **asimétrica** (24 casos **3→6** frente a 4 **6→3**). Esa asimetría es habitual cuando un clúster tiene **más representación** o cuando su **frontera de decisión** está más cerca del otro: el clasificador se inclina más veces hacia el clúster con mayor masa o con centroides próximos. En conjunto, la combinación de: (a) **distancia euclídea muy por debajo** de la media inter-clúster (percentil 4,8), (b) **similitud coseno ≈ 0,45** y (c) **correlación ≈ 0,48**, respalda que 3 y 6 son **perfiles moderadamente parecidos**. Por tanto, el error observado (92,35 % de acierto global y confusión concentrada entre 3 y 6) es **coherente con la geometría de los datos** y no invalida la interpretación: el modelo confunde, sobre todo, **tipos de día próximos** en términos de su perfil horario.

Para terminar de dar perspectiva a las distancias euclídeas, aquí se adjunta las distancias entre cada par de clústeres:

Ilustración 17: Distancia Euclídea entre clústeres

	0	1	2	3	4	5	6
0		19,50	20,86	18,48	23,45	18,61	18,03
1			18,02	16,55	18,17	17,03	16,73
2				21,62	12,86	17,67	21,01
3					16,99	14,67	**10,79**
4						22,08	18,26
5							14,41
6							

Con esto se puede concluir que el análisis llevado a cabo en este capítulo es coherente y completo, y se puede empezar a diseñar una estrategia para la compensación de desvíos.

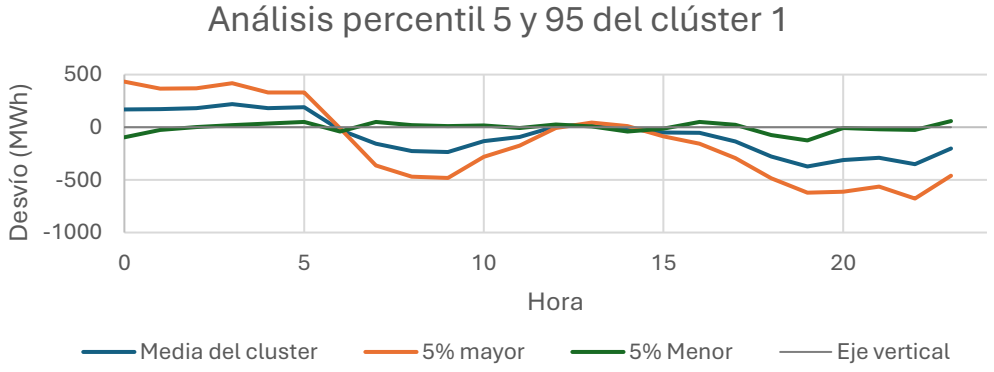
4.3.- Estrategia económica

Una vez ya se es capaz de predecir con un 92,35%, el siguiente paso es conseguir una estrategia eficiente para poder tener un impacto económico positivo. Para ello, en este subcapítulo se estudiará desde el lado de una comercializadora eléctrica, estas son "aquellas sociedades mercantiles, o sociedades cooperativas de consumidores y usuarios, que, accediendo a las redes de transporte o distribución, adquieren energía para su venta a los consumidores, a otros sujetos del sistema o para realizar operaciones de intercambio internacional en los términos establecidos en la ley [BOE24/2013]".

El objetivo de este estudio será obtener una estrategia por la cual una comercializadora pudiese tener una estrategia que le permitiese ahorrarse un dinero llevando una estrategia contraria a la esperada por el mercado. Haciendo referencia con esto a que, si se supiese, que, por el tipo de día, la tendencia del mercado va a estar en desvíos a subir, por lo cual, los precios de esos desvíos van a ser muy altos, se puede suponer que la estrategia óptima ese día sería el quedarse "cortos" con un nivel de incertidumbre medio o bajo, sabiendo que el precio del ajuste va a ser menor que el que se tendría si se tuviese una estrategia estándar acorde con el resto del mercado. Al fin y al cabo, este trabajo es simplemente es un estudio del mercado del sector eléctrico y su comportamiento ante cierto tipo de día, pues las variables a analizar , como son la temperatura, la predicción de demanda, la potencia nuclear disponible, son variables que usan ambas, comercializadores y REE para predecir demandas y generación, y si por estas variables están teniendo algún tipo de sesgo, la comercializadora ficticia, como consumidora de ese sistema, puede aprovecharlo para, y esto es lo más importante de todo, apoyar al sistema, puesto que está ayudando a equilibrarlo en horas de despunte de desvío hacia un lado o hacia otro. Entonces se puede tener un beneficio económico mientras se apoya al sistema hacia el equilibrio lo cual supone un doble beneficio, tanto para el sistema como para la comercializadora.

Una vez visto como se quiere apoyar al sistema, se va a intentar desarrollar una estrategia para llevar a cabo precisamente esto. Para ello se empezará por analizar, por ejemplo, los desvíos del Clúster 1, puesto que el 0 no está presente en nuestros datos y así se lleva a cabo el análisis con un clúster que no se ha estudiado en el capítulo anterior. Este sería su perfil de desvíos:

Ilustración 18: Media y Análisis percentil 5 y 95 del clúster 1



Del análisis realizado se infiere que, durante las primeras horas de la mañana, los desvíos nacionales presentan habitualmente valores positivos, lo que indica que, en promedio, el consumo efectivo resulta inferior al previsto. En estas circunstancias, una comercializadora que programe una cantidad de energía ligeramente superior en dicho intervalo temporal tenderá a obtener una compensación favorable. Un ejemplo representativo se observa el 05/04/2024 a las 4:00 a.m., momento en el que el precio de desvío a subir (PDS) alcanzó los 8,46 €/MWh, mientras que el precio de desvío a bajar (PDB) se situó en 1,55 €/MWh. En este escenario, una posición larga de 5,46 MWh resultaría económicamente más ventajosa que una posición corta de 1 MWh, dada la marcada diferencia entre ambos precios.

No obstante, debe señalarse que, en determinados clústeres, caracterizados por una elevada desviación típica en la distribución horaria de los desvíos, pueden coexistir intervalos con signo positivo y negativo. En tales casos se hace necesario adoptar un enfoque más conservador para evitar incrementos indeseados del riesgo. Con el objetivo de generalizar la estrategia, se considera inadecuado definir reglas distintas para cada clúster de manera aislada, pues ello limitaría el alcance del análisis. El planteamiento óptimo consiste en disponer de un procedimiento aplicable a cualquier día, cuya eficacia pueda mejorarse gracias a la información específica que aporta cada clúster, logrando así beneficios consistentes y sostenibles.

4.3.1 Estrategias propuestas

Estrategia 1

Se define el ajuste óptimo de los desvíos (Δ) con la expresión:

Ecuación 2: Ecuación que modula la estrategia 1

$$\Delta = k \cdot \mu \cdot \left(1 - e^{-\frac{|\mu|}{\sigma}}\right).$$

donde k es un factor de corrección específico de cada clúster, μ es la media de los desvíos históricos de ese clúster y σ su desviación típica. Este planteamiento sigue la lógica de muchas otras funciones usadas en otros lugares de la ingeniería, como la función de fricción tipo Stribeck, que se usa para suavizar comportamientos cercanos al 0, en la que el tamaño de la posición se ajusta por el cociente señal-ruido $\frac{|\mu|}{\sigma}$ y por un multiplicador que representa la confianza estadística.

El término $(1 - e^{-\frac{|\mu|}{\sigma}})$ actúa como un suavizador:

- Cuando | μ | es pequeño en relación a σ, el patrón débil o volátil y este factor se aproxima a cero, reduciendo Δ y limitando el riesgo.

- Cuando | μ | es grande frente a σ (patrón fuerte y estable), el factor se acerca a 1, y el ajuste es prácticamente $k \cdot \mu$ maximizando la oportunidad.

El parámetro k se calibra de forma empírica para cada clúster en función de su previsibilidad: un valor alto en clústeres muy consistentes, más bajo en clústeres inestables. Así, la estrategia busca capturar el beneficio esperado del sesgo de desvíos del clúster reduciendo exposición en situaciones inciertas, equilibrando rentabilidad y riesgo.

Estrategia 2

Se define el ajuste óptimo de los desvíos (Δ) con la expresión:

Ecuación 3: Ecuación que modula la estrategia 2

$$\Delta = \begin{cases} 0, & \text{si } \frac{|\mu|}{\sigma} \leq z_0, \\ k \cdot \mu, & \text{si } \frac{|\mu|}{\sigma} > z_0 \ , \end{cases}$$

Donde μ es la media histórica del desvío del clúster, σ su desviación típica (con lo que $\frac{|\mu|}{\sigma}$ es un **z-score** o relación señal/ruido comparable entre clústeres), z_0 es el **umbral mínimo de significancia/robustez** y k es un multiplicador de confianza por clúster. Operativamente, la regla introduce una **zona muerta** alrededor de cero: si el efecto no es "grande" respecto a su volatilidad, se **no-opera** ($\Delta = 0$) para evitar falsas alarmas; si la señal es suficientemente clara, se aplica el ajuste proporcional $k \cdot \mu$. Así se controla el riesgo (no se reacciona a ruido) y, a la vez, se captura el sesgo cuando es estadísticamente creíble, dejando a k la calibración del tamaño de la acción por clúster.

Esta lógica de **umbral duro** ("actuar/no actuar") aparece en varios ámbitos bien asentados:

- **Control estadístico de procesos (gráficos de Shewhart)**: se disparan acciones solo cuando la estadística estandarizada rebasa límites tipo "$\pm 3\sigma$". Es exactamente la idea de "si | Z |> $z0$, actúa; si no, no actúes" [NIST25].
- **Denoising por** *hard-thresholding* **en** *wavelets*: el estimador mantiene un coeficiente **solo si** su magnitud supera un umbral y, si no, lo pone a cero; es la misma estructura por tramos (fuera del umbral → 0; por encima → valor proporcional). Es un referente clásico en estadística no paramétrica [CAI99].

4.3.2 Análisis de estrategias

Una vez ya están definidas las estrategias que se van a seguir, ya se puede empezar a calcular si verdaderamente se puede obtener un beneficio económico de esta estrategia, para ello se ha desarrollado la siguiente tabla:

Tabla 19: Tabla resumen estrategias

Día	Hora	Grupo	PMD	PDS	PDB	Penalización Subir	Penalización bajar
01/04/2024	0:00:00	3	0,7	0	2,99	0,7	2,29
01/04/2024	1:00:00	3	0,02	0,01	26,21	0,01	26,19
01/04/2024	2:00:00	3	0	-5	3,25	5	3,25
01/04/2024	3:00:00	3	0	2,81	2,81	-2,81	2,81
01/04/2024	4:00:00	3	0	1,29	1,29	-1,29	1,29
01/04/2024	5:00:00	3	0	3,01	3,01	-3,01	3,01
01/04/2024	6:00:00	3	0,13	-2,31	3,12	2,44	2,99

Donde, el grupo es el clúster predicho por el Random Forest del modelo del aparatado 4.2 Predicción del tipo de día. El PMD es el precio del mercado diario del indicador 600 de la web de ESIOS [ESIO25], después el PDS y PDB son los precios de los desvíos a subir y bajar respectivamente sacados de los indicadores 763 y 764 de ESIOS [ESIO25]. Finalmente, las penalizaciones a subir y bajar son, respectivamente, cuanto te cobran de más por ese MWh que has consumido de más o de menos. En el caso de los desvíos a subir, la penalización corresponde a la pérdida asociada a los MWh no consumidos, cuantificada como (PMD−PDS). Por su parte, en los desvíos a bajar, la penalización se traduce en un coste adicional equivalente a (PDB−PMD) por haber consumido más de lo adquirido en el mercado diario. En ambos supuestos, la atención se centra en la magnitud de dichas penalizaciones, dado que constituyen el factor determinante en este estudio, cuyo objetivo fundamental es minimizar su impacto económico.

Para proceder a la aplicación de la Estrategia 1 resulta necesario determinar los valores de k que se utilizarán en el análisis. Tras un proceso de optimización preliminar, se obtuvieron los siguientes parámetros:

Tabla 20: Factores de confianza de la estrategia 1

Clúster	$k_{i\ \text{estrategia 1}}$
0	n.a.
1	0,0272
2	n.a.
3	0,0414
4	0,0309
5	n.a.
6	-0,0175

Lo cual demuestra coherencia con el resto del trabajo, puesto que el clúster 3 y 4, se han usado de ejemplos de clústeres con desvíos altos negativos más fáciles de contrarrestar, mientras que el 1 y el 6 tienen más variabilidad, además de los problemas de predicción que se tuvieron en el apartado anterior para poder predecir con precisión el clúster 6.

Siguiendo con el estudio de la estrategia 1 se obtendrían las siguientes conclusiones:

Tabla 21: Tabla resumen Estrategia 1

Compra original	Estrategia 1	Compra con estrategia	Desvío Original	Consumo real	Desvío con estrategia
100,0	0,1	100,1	15,0	115,0	14,9
100,0	5,8	105,8	7,8	107,8	2,0
100,0	6,9	106,9	7,4	107,4	0,6
100,0	7,4	107,4	8,5	108,5	1,1
100,0	7,6	107,6	7,3	107,3	-0,3
100,0	8,5	108,5	9,8	109,8	1,2
100,0	8,7	108,7	7,6	107,6	-1,1

Desde un punto de vista metodológico, resulta conveniente homogeneizar las magnitudes empleadas en el análisis con el fin de evitar sesgos de escala y garantizar que los resultados puedan extrapolarse a distintos agentes del sistema. Por este motivo, se adopta la hipótesis de que la comercializadora adquiere 100 MWh en todas las horas del día. Esta simplificación no altera el sentido del estudio, pero permite dotar a los resultados de una magnitud relativa fácilmente interpretable y comparable, como si se tratase de una de las cien comercializadoras más grandes del mercado español.

Sobre esta base, la columna correspondiente a la Estrategia 1 se calcula aplicando la Ecuación 1, definida previamente, con los coeficientes recogidos en la Tabla 13. La compra con estrategia se obtiene sumando a la compra original el valor de Δ derivado de la Estrategia 1.

En este primer análisis, el **desvío original** hace referencia a los desvíos registrados por Red Eléctrica de España (REE) en cada hora del día, es decir, a las discrepancias reales entre la demanda prevista y la demanda final observada en el sistema. Con el objetivo de adaptarlos a la escala de una comercializadora mediana, que es el caso de estudio de este trabajo, dichos valores se han normalizado dividiéndolos entre 100. De esta manera, en lugar de reflejar los desvíos globales del sistema, se obtiene una magnitud relativa que resulta más representativa del volumen de energía con el que opera una comercializadora de este tamaño. Conviene señalar que este procedimiento constituye únicamente una primera aproximación, cuyo propósito es contextualizar las estrategias en una escala manejable. Más adelante se analizará un escenario más realista en el que los desvíos no se consideran los ya conocidos del sistema, sino que se modelan como aleatorios alrededor de cero, lo que se ajusta mejor al comportamiento esperado de una comercializadora concreta. Finalmente, el desvío con estrategia refleja la desviación efectiva de la comercializadora en el caso de aplicar la Estrategia 1.

Finalmente, como resultados finales que se recopilan en esta última tabla:

Tabla 22: Tabla resumen beneficios estrategia 1

Diferencia de desvío	Coste original	Coste Con estrategia	Desvíos netos originales	Desvíos estrategia netos
0,1	34,4	34,1	15,0	14,9
5,8	204,9	53,7	7,8	2,0
6,9	24,1	1,8	7,4	0,6
7,4	24,0	3,1	8,5	1,1
6,9	9,4	0,5	7,3	0,3
8,5	29,4	3,7	9,8	1,2
6,5	22,7	2,7	7,6	1,1

Donde el coste original es el desvío original por la penalización a subir o a bajar, según la dirección del desvío, y el coste con estrategia es lo mismo, pero con el desvío ajustado por la estrategia que se ha seguido. Por último, los desvíos netos sería el valor absoluto de los desvíos para ver cómo es el comportamiento con la estrategia frente a no tener la estrategia. Si se extiende el procedimiento a todo el año pasado, estos serían los resultados:

Tabla 23: Resumen Conclusiones Estrategia 1

Estrategia 1	Suma Original	Suma estrategia	Diferencia
Coste	3.568.070,27 €	3.235.333,03 €	-9%
Desvíos	80538,731	81913,92039	2%

Los resultados obtenidos muestran que la aplicación de la Estrategia 1 resulta rentable, alcanzando un beneficio adicional de **332.737 €.** No obstante, debe señalarse que este planteamiento incrementa el volumen total de desvíos de la comercializadora, aunque lo hace en la dirección que contribuye a apoyar el equilibrio de la red en momentos de necesidad. Aun así, es importante considerar que este tipo de actuaciones no siempre es bien recibido por el operador del sistema [CNMC22], por lo que, aunque en el contexto actual la estrategia sea económicamente ventajosa, podría no serlo en escenarios regulatorios futuros.

Siguiendo con el mismo razonamiento, se procede ahora a analizar los resultados derivados de la implementación de la Estrategia 2. En este caso, el principal inconveniente radica en que se trata de un enfoque considerablemente más arriesgado: cuando el nivel de confianza supera el umbral establecido por el filtro, el modelo tiende a realizar ajustes de gran magnitud. Por esta razón, se ha introducido una limitación que impide solicitar más del 50 % adicional o menos del 50 % respecto a la cantidad prevista originalmente. De no aplicarse esta restricción, en determinadas horas caracterizadas por un nivel de confianza muy elevado, la estrategia podría aprovechar en exceso la situación, distorsionando el balance del sistema al adquirir o reducir cantidades excesivas de energía.

Para formalizar este análisis se definen los filtros z_0 y los niveles de confianza k, que servirán como base para evaluar el desempeño de la Estrategia 2:

Tabla 24: Factores de confianza de la estrategia 2 y filtro

Clúster	$k_{i\ estrategia\ 2}$	$z_{0i\ estrategia\ 2}$
0	0	0
1	0,02	0,0468
2	0	0
3	0,09	0,324
4	0,02	0,02
5	0	0
6	0,01	0,2

Una vez vistos, las constantes, se llevará a cabo el desarrollo de la estrategia:

Tabla 25: Tabla resumen Estrategia 2

Compra original	Umbral	Estrategia (1=Sí / 0=No?	Estrategia 2	Estrategia 2 ajustada	Compra con estrategia	Desvío Original
100	1,00	1	0	0,1	100,1	15,0
100	0,18	0	0	0,0	100,0	7,8
100	0,24	0	0	0,0	100,0	7,4
100	0,29	0	0	0,0	100,0	8,5
100	0,28	0	0	0,0	100,0	7,3
...						
100	0,34	1	-411	-50	50	4,3

Siendo, en este caso, el umbral el $\frac{|\mu|}{\sigma}$ calculado y que dependiendo si es mayor o no del filtro z_0, se llevará a cabo la estrategia, como está indicado en la siguiente columna. Después la siguiente columna sería aplicar la estrategia 2, teniendo esta vez una columna adicional para ajustar lo mencionando anteriormente, no superar el 50% de aumento o disminución del consumo habitual de la comercializadora a esa hora. Esto, por supuesto es ajustable, pero es un valor razonable para este primer análisis. Siguiendo con la línea de pensamiento se obtendría la siguiente tabla:

Tabla 26: Tabla resumen beneficios estrategia 2

Coste original	Coste Con estrategia	Desvíos netos	Desvíos estrategia netos
34,4	34,1	15,0	14,9
204,9	204,9	7,8	7,8
24,1	24,1	7,4	7,4
24,0	24,0	8,5	8,5
9,4	9,4	7,3	7,3
...			
12,2	154,2	4,3	54,3

Donde se ha añadido una fila adicional al final donde se ve un ejemplo del problema de esta estrategia frente a la otra, y es que en momentos donde se aplica la estrategia, se es muy agresivo teniendo tanto recompensas altas, como penalizaciones altas como se puede ver en esta última fila, donde el precio para esa hora es 12x veces mayor al precio original. También ocurre esto en el otro sentido, pero es importante remarcarlo. Con todo esto, aquí estarían los resultados:

Tabla 27: Resumen Conclusiones Estrategia 2

Estrategia 2	Suma Original	Suma estrategia	Diferencia
Coste	3.568.070,27 €	3.143.013,99 €	-12%
Desvíos	80538,731	96475,90558	20%

Como puede observarse, esta estrategia arroja un **beneficio económico superior, de 425.056 €**, pese a que los desvíos resulten mayores en este caso. Es posible que la estrategia sea **demasiado agresiva** y se beneficie de **horas pico** que arrastran el resultado hacia un lado u otro; o bien que los **desvíos originales**, por estar muy ligados al precio, eleven el coste total de la comercializadora y el análisis pierda capacidad explicativa. Para **dar mayor profundidad técnica** y aportar **robustez** a esta parte final, se añadió el siguiente análisis: en realidad, **una comercializadora no conoce ex ante sus desvíos hora a hora**; si los conociera, el problema sería trivial.

Dado que una comercializadora individual atiende únicamente a un conjunto relativamente reducido de clientes dentro del sistema eléctrico, sus desvíos no pueden considerarse perfectamente correlacionados con el desvío agregado del sistema. En otras palabras, aunque el sistema en su conjunto presente un desvío neto positivo o negativo, ello no implica necesariamente que cada comercializadora, de forma aislada, reproduzca el mismo patrón. Por el contrario, los desvíos de un agente individual son en gran medida **idiosincráticos y aleatorios**, fruto de la agregación de múltiples comportamientos de consumo y generación que, a nivel micro, no son completamente previsibles.

En este sentido, resulta más realista suponer que los desvíos de una comercializadora tienden a distribuirse en torno a cero, lo que refleja un desempeño operativo razonablemente ajustado a sus previsiones, pero con cierta variabilidad inherente a la incertidumbre del mercado minorista. Para capturar este comportamiento, se modelan los

desvíos como una variable aleatoria con distribución normal de media cero y una desviación típica acotada $N(0,\ 10)$ de manera que un desvío en valor absoluto superior a 30 sea altamente improbable. Este planteamiento es coherente con el **teorema central del límite**, ya que la suma de múltiples errores de predicción individuales de los clientes tiende a aproximarse a una distribución normal alrededor de la media.

La principal ventaja de este enfoque radica en que los desvíos simulados no arrastran el sesgo estructural asociado al desvío total del sistema, que sí puede estar correlacionado con factores de precio o de generación renovable que tenía de forma inherente el caso anterior, lo que permite aislar mejor el efecto de la estrategia propuesta. De este modo, si la estrategia continúa ofreciendo un impacto económico positivo bajo la hipótesis de desvíos aleatorios, podrá concluirse que su validez no depende de las particularidades de los desvíos reales del sistema, sino de la robustez intrínseca del método. En otras palabras, se demostraría que la mejora económica obtenida proviene de la lógica de la estrategia en sí misma, y no de un sesgo específico de los datos históricos.

Por lo que, llevando a cabo el mismo análisis, con los **mismos coeficientes** que en el caso anterior (véanse Tablas 13 y 17) y se aplica el **mismo protocolo de análisis** (Anexos 11 y 12). Los resultados comparativos son:

Tabla 28: Comparativa de la estrategia 1

	Estrategia 1	Suma Original	Suma estrategia	Diferencia
Caso 1: Desvíos Reales	Coste 1	3.568.070,27 €	3.235.333,03 €	-9%
	Desvíos 1	80538,73	81913,92	2%
Caso 2: Desvíos Aleatorios	Coste 2	1.514.024,41 €	1.303.835,86 €	-14%
	Desvíos 2	70149,84	79613,91	13%

Tabla 29: Comparativa de la estrategia 2

	Estrategia 2	Suma Original	Suma estrategia	Diferencia
Caso 1: Desvíos Reales	Coste	3.568.070,27 €	3.121.575,67 €	-13%
	Desvíos	80538,73	100517,21	25%
Caso 2: Desvíos Aleatorios	Coste	1.514.024,41 €	1.191.019,92 €	-21%
	Desvíos	70149,84	97589,29	39%

Los resultados presentados en las Tablas 21 y 22 son claramente positivos: no solo se consigue un impacto económico favorable, sino que, en términos relativos, este resulta más elevado en el escenario con desvíos aleatorios. Este hecho respalda la hipótesis de que, en el **Caso 1**, los desvíos estaban sesgados y generaban una presión adicional sobre los costes. Por el contrario, cuando los desvíos se distribuyen en torno a cero, el ahorro persiste incluso si la magnitud absoluta de los desvíos aumenta. Este incremento es coherente con la naturaleza del planteamiento: al partir de una base aleatoria, la estrategia orienta la posición en la dirección prevista (según el tipo de día), y pese a ello el resultado económico agregado sigue siendo positivo.

Una cuestión que debe considerarse es la posibilidad de que el ahorro observado provenga en gran medida de valores extremos y que, en ausencia de dichas horas pico, la eficacia de las estrategias disminuya. Con el objetivo de evaluar esta preocupación, se ha realizado un análisis de variabilidad en el que se estudia el efecto de eliminar las horas con precios más elevados, reduciendo así la influencia de observaciones extremas en los resultados. Este procedimiento se aplicó en cuatro escenarios diferentes, eliminando el 1 %, 3 %, 5 % y 10 % de las horas más extremas, exclusivamente en el **Caso 2**, por considerarse el más representativo de la realidad. Los resultados obtenidos se muestran a continuación:

Tabla 30: Análisis de Variabilidad

	Coste Original	Coste Con estrategia 1	Coste con estrategia 2
Caso Base	1.514.024,41 €	1.303.835,86 €	1.191.019,92 €
Diferencia (% Caso Base)	**0,0%**	**-13,9%**	**-21,3%**
Eliminado el 1% extremos	1.443.368,72 €	1.253.294,24 €	1.175.369,74 €
Diferencia (% Caso Base)	**-4,7%**	**-13,2%**	**-18,6%**
Eliminado el 3% extremos	1.392.914,76 €	1.219.944,29 €	1.161.756,21 €
Diferencia (% Caso Base)	**-8,0%**	**-12,4%**	**-16,6%**
Eliminado el 5% extremos	1.326.641,45 €	1.176.072,72 €	1.152.182,98 €
Diferencia (% Caso Base)	**-12,4%**	**-11,3%**	**-13,2%**
Eliminado el 10% extremos	1.183.915,82 €	1.068.565,70 €	1.114.074,64 €
Diferencia (% Caso Base)	**-21,8%**	**-9,7%**	**-5,9%**

Las conclusiones principales derivadas de la Tabla 23 muestran que, incluso tras eliminar hasta el 10 % de los valores más extremos, la **Estrategia 1** mantiene una mejora de hasta un 9,7 % respecto a la ausencia de estrategia. Este resultado confirma su solidez y capacidad para generar beneficios aun en escenarios en los que se reducen de forma significativa los efectos de las horas pico.

En el caso de la **Estrategia 2**, los resultados evidencian una notable pérdida de eficacia al eliminar valores extremos, lo cual era previsible dado su carácter intrínsecamente más agresivo. En consecuencia, su rendimiento está más condicionado por situaciones de precios extremos, que representan una parte sustancial de su rentabilidad. No obstante, esta diferencia entre ambas estrategias aporta un valor añadido al análisis, al señalar la necesidad de explorar posibles ajustes o variantes metodológicas que permitan incrementar la robustez del enfoque más agresivo. Estas cuestiones se desarrollarán en el apartado de conclusiones generales del trabajo.

Capítulo 5.- Conclusiones

5.1.- Conclusiones y limitaciones sobre la metodología

La metodología seguida en este trabajo presenta un orden lógico y se explica en sí misma, de modo que la mayoría de las decisiones metodológicas quedan justificadas a lo largo del propio desarrollo. No obstante, existe un aspecto que merece ser tratado de forma particular: la elección concreta de las variables y el modo en que se ha trabajado con ellas.

Al analizar las tablas de caracterización de los clústeres se identificaron dos factores especialmente relevantes. En primer lugar, los **días de la semana**, que constituyen tres de las 17 variables iniciales y generan un sesgo considerable, ya que cuando una de ellas toma valor 1, las otras dos necesariamente toman valor 0. Esta codificación produce diferencias significativas en las distancias euclídeas únicamente por la definición de dichas variables. En segundo lugar, la **potencia instalada** (tanto fotovoltaica como eólica), que introduce un sesgo temporal evidente, dado que aumenta de forma estructural con el paso de los años. Como consecuencia, determinados clústeres podían quedar condicionados por un efecto puramente temporal, sin aportar información adicional de carácter operativo.

Esta situación podría llevar a pensar que los datos no habían sido tratados de la forma más adecuada y que existían alternativas metodológicas más consistentes. Para contrastar esta hipótesis se llevó a cabo un estudio paralelo en el que se eliminaron las variables correspondientes a los días de la semana y se incorporó una nueva variable: la **previsión fotovoltaica relativa**, definida como la previsión de generación fotovoltaica dividida entre la potencia instalada. De esta forma, se mitigaba el sesgo temporal asociado al crecimiento progresivo de la capacidad instalada.

Tras dicho análisis, los resultados permitieron sostener la validez de mantener el planteamiento original. En primer lugar, se examinó el **RMSE de los centroides**, observándose una reducción marginal al pasar de 910,059 en el modelo original a 909,930 en la versión revisada. Este indicador mide el error cuadrático medio de las distancias entre los puntos de cada clúster y su respectivo centroide, de manera que valores menores implican una mayor compacidad interna de los grupos y, por tanto, un ajuste potencialmente mejor del modelo de agrupamiento. Sin embargo, la disminución registrada fue mínima (inferior al 0,02 %), lo que demuestra que la calidad de la compactación interna de los clústeres apenas varía entre ambas configuraciones. En términos prácticos, esta diferencia no modifica la interpretación global del modelo: ambos escenarios presentan un comportamiento equivalente, sin que pueda afirmarse que la versión revisada suponga una mejora cualitativa significativa respecto a la versión original.

En segundo lugar, se analizaron las desviaciones típicas de los desvíos eléctricos en la versión revisada del modelo (sin las variables correspondientes a los días de la semana y con la nueva variable de ajuste relativa). Se observó que las desviaciones asociadas a los distintos clústeres eran ahora mucho mayores y que los perfiles resultaban considerablemente más inestables que en la versión original.

Dado que esta nueva configuración no aportaba una mejora cualitativa del modelo, ni generaba perfiles de desvío de mayor interés para el estudio, se decidió mantener la versión inicial. Con ello, el análisis realizado condujo a conclusiones sólidas y prometedoras. Entre ellas, cabe destacar, en primer lugar, que, aunque los clústeres estén

fuertemente condicionados por las variables de los días de la semana, también lo están los propios perfiles de desvío. Este hecho ha resultado útil para la interpretación en numerosos casos. En segundo lugar, incluso con una parte de los perfiles ya era posible identificar irregularidades que apuntaban a potenciales problemas en la predicción eléctrica. Por ejemplo, en el clúster 6 se observó que la disponibilidad de generación renovable era especialmente elevada: la previsión de generación fotovoltaica se situaba muy por encima de la media, llegando a ser superior en más de un 60 % en determinadas horas, y la eólica también presentaba aportaciones positivas. Este escenario, caracterizado por baja demanda, pero alta oferta renovable en días festivos cálidos, sugiere desvíos elevados en sentido contrario al clúster 5: en este caso, podrían producirse excedentes de generación muy superiores a lo previsto respecto a la demanda real.

Más allá de estas observaciones, es importante reconocer que el trabajo está condicionado por limitaciones inherentes a los métodos empleados, en particular **K-means** y **Random Forest**.

En lo relativo a **K-means**, uno de sus principales inconvenientes es la necesidad de fijar a priori el número de clústeres K. La elección de un valor inadecuado puede conducir a segmentaciones poco representativas: un K demasiado bajo agrupa patrones heterogéneos en un mismo clúster, mientras que un K excesivo genera subdivisiones artificiales que no responden a dinámicas reales. Además, el algoritmo asume implícitamente que los clústeres son esféricos y de tamaño similar, lo que no siempre es cierto en datos de naturaleza energética, donde los perfiles pueden ser altamente asimétricos o presentar colas pesadas. Otro aspecto a considerar es su sensibilidad a los valores atípicos: la presencia de unos pocos desvíos extremos puede desplazar de manera significativa los centroides y alterar la asignación global. Finalmente, K-means converge hacia óptimos locales, de manera que los resultados dependen de la inicialización de los centroides, lo que obliga a realizar múltiples ejecuciones para garantizar cierta estabilidad [JAME23].

En el caso de **Random Forest**, sus limitaciones son de naturaleza distinta. Aunque se trata de un modelo robusto y de gran capacidad predictiva, puede perder interpretabilidad al combinar un elevado número de árboles, lo que dificulta el análisis directo de las reglas de decisión. Asimismo, si bien la técnica reduce sustancialmente la varianza frente a un árbol único, puede mostrar un sesgo comparable al de estos, lo que implica que no siempre logra capturar patrones muy complejos de interacción entre variables. Otro aspecto señalado en la literatura es su mayor exigencia computacional: la construcción de cientos o miles de árboles incrementa los tiempos de entrenamiento y la demanda de memoria, especialmente con conjuntos de datos de alta dimensionalidad. Por último, aunque Random Forest suele proporcionar estimaciones de importancia de variables, estas pueden estar sesgadas en favor de predictores con más niveles categóricos o con mayor variabilidad, lo que debe tenerse en cuenta al interpretar los resultados [JAME23].

En conjunto, tanto K-means como Random Forest constituyen herramientas valiosas para el análisis de desvíos eléctricos y la predicción de patrones de comportamiento del sistema, pero es fundamental reconocer sus limitaciones y el modo en que estas pueden condicionar la interpretación de los resultados.

En términos generales, la aplicación de las técnicas de aprendizaje automático seleccionadas ha ofrecido resultados altamente satisfactorios. Por un lado, la clusterización mediante **K-means** ha permitido identificar agrupaciones de días y perfiles

horarios con características similares, ofreciendo una caracterización clara de escenarios de operación del sistema eléctrico. Estos clústeres, definidos a partir de variables tanto de demanda como de generación renovable, han facilitado la interpretación de patrones complejos y han revelado situaciones recurrentes de desvíos eléctricos, lo que constituye una base sólida para la toma de decisiones estratégicas. Por otro lado, el uso de **Random Forest** ha proporcionado un modelo predictivo con una precisión cercana al 92 %, lo que demuestra su capacidad para anticipar correctamente el tipo de día que se observará en el sistema. Esta capacidad predictiva no solo valida la robustez del enfoque metodológico adoptado, sino que además abre la posibilidad de transformar dicho conocimiento en una ventaja económica para las comercializadoras, aspecto que se analizará en el siguiente apartado.

5.2.- Conclusiones sobre los resultados

Los resultados obtenidos en este trabajo muestran de forma clara el enorme potencial que puede tener la aplicación de técnicas avanzadas de análisis de datos y predicción al ámbito eléctrico. Recordando la tabla de análisis final:

Tabla 31: Copia Tabla 27: Análisis de Variabilidad

	Coste Original	Coste Con estrategia 1	Coste con estrategia 2
Caso Base	1.514.024,41 €	1.303.835,86 €	1.191.019,92 €
Diferencia (% Caso Base)	**0,0%**	**-13,9%**	**-21,3%**
Eliminado el 1% extremos	1.443.368,72 €	1.253.294,24 €	1.175.369,74 €
Diferencia (% Caso Base)	**-4,7%**	**-13,2%**	**-18,6%**
Eliminado el 3% extremos	1.392.914,76 €	1.219.944,29 €	1.161.756,21 €
Diferencia (% Caso Base)	**-8,0%**	**-12,4%**	**-16,6%**
Eliminado el 5% extremos	1.326.641,45 €	1.176.072,72 €	1.152.182,98 €
Diferencia (% Caso Base)	**-12,4%**	**-11,3%**	**-13,2%**
Eliminado el 10% extremos	1.183.915,82 €	1.068.565,70 €	1.114.074,64 €
Diferencia (% Caso Base)	**-21,8%**	**-9,7%**	**-5,9%**

Haber conseguido un **ahorro económico directo de entre el 14 % y el 22 % en la compra de energía de una comercializadora de tamaño medio**, sin necesidad de inversiones adicionales en infraestructura ni modificaciones en la compra de distintas fuentes en el sector energético, constituye un hito muy relevante. En términos absolutos, este planteamiento ha permitido demostrar la posibilidad de alcanzar ahorros cercanos a **150.000 €** simplemente ajustando la estrategia de compra de energía en función del tipo de día previsto. Estos resultados ponen de manifiesto que la explotación inteligente de la información disponible puede transformarse en una ventaja competitiva sustancial, y evidencian cómo la digitalización y la ciencia de datos están cambiando la forma en la que los agentes del mercado eléctrico toman sus decisiones.

En primer lugar, se ha comprobado que tanto los desvíos del sistema como las estrategias aplicadas presentan una cierta dependencia de los valores extremos. Sin embargo, incluso eliminando el 10 % de las horas con precios más altos en el escenario de desvíos aleatorios

(lo que equivale a descartar los episodios más favorables para la estrategia, donde grandes desvíos coinciden con precios elevados, y que suponen los desvíos económicos más importantes para el escenario base), los resultados económicos siguen siendo positivos. Este hecho refuerza la validez del enfoque y demuestra que no se trata de una mera explotación de episodios puntuales, sino de un planteamiento que aporta beneficios también en condiciones de operación habituales.

En segundo lugar, la comparación entre las dos metodologías estratégicas propuestas muestra matices importantes. La **Estrategia 1** se caracteriza por ser más **sólida y consistente** frente a la variabilidad, manteniendo un rendimiento positivo incluso cuando se reducen los valores extremos. A cambio, su rentabilidad en días típicos resulta más moderada. La **Estrategia 2**, en contraste, presenta un **impacto económico más elevado**, aunque más dependiente de episodios extremos. Esto se debe tanto a su diseño , basado en filtros de confianza y actúa de forma más agresiva cuando la señal es clara, como a su capacidad para aprovechar situaciones en las que la diferencia entre previsión y realidad es especialmente marcada.

Estas conclusiones abren la puerta a futuros desarrollos metodológicos, en particular a la **combinación híbrida de ambas estrategias**. Un enfoque mixto permitiría garantizar un ahorro estable en situaciones regulares gracias a la consistencia de la Estrategia 1, y al mismo tiempo capturar rentabilidades adicionales en escenarios extremos gracias a la capacidad de la Estrategia 2. Este planteamiento, además de equilibrar rentabilidad y riesgo, podría adaptarse dinámicamente según la volatilidad esperada de los desvíos y los precios.

Ahora bien, no debe olvidarse que este tipo de actuaciones también tiene implicaciones regulatorias. Tal y como señala la **Comisión Nacional de los Mercados y la Competencia [CNMC22]** el diseño de los mecanismos de desvíos y servicios de ajuste persigue, en última instancia, incentivar comportamientos que favorezcan la estabilidad y el equilibrio del sistema. En este sentido, si REE detectara que determinadas comercializadoras aplican sistemáticamente estrategias de este tipo para beneficiarse económicamente, aunque nominalmente apoyen al sistema, podría considerar introducir modificaciones normativas o ajustes en el diseño de liquidación de los desvíos con el fin de evitar un aprovechamiento excesivo del mecanismo.

En conclusión, este trabajo demuestra que es posible lograr **ahorros económicos significativos** y, a la vez, contribuir al equilibrio del sistema eléctrico mediante un uso inteligente de los datos. Sin embargo, también pone de relieve que el éxito de estas estrategias depende tanto de la robustez técnica de los modelos como de la evolución futura del marco regulatorio, lo que convierte a este ámbito en un terreno fértil para el desarrollo de nuevas líneas de investigación y de herramientas de apoyo a la toma de decisiones en el sector energético.

5.3.- Recomendaciones para futuros estudios

Por último, si se quisiera seguir avanzando con el trabajo para conseguir desarrollar mejor la técnica, estos serían los dos caminos principales de actuación.

En primer lugar, respecto a la **primera parte del trabajo**, sería necesario profundizar en el análisis de los perfiles de desvío. Durante el estudio se observaron discordancias relevantes en determinados clústeres, como el 3 o el 6, donde las previsiones

meteorológicas, de demanda y de potencia generada parecían contradictorias, reflejándose posteriormente en los perfiles de desvío. Un estudio más detallado de estas incongruencias permitiría comprender con mayor precisión los factores que originan desvíos estructurales y, en consecuencia, contribuir a mejorar las previsiones de demanda y generación elaboradas por Red Eléctrica de España. Reducir los desvíos generales del sistema tendría un impacto positivo para el mercado en su conjunto, aunque supusiera que las estrategias de la segunda parte del trabajo resultasen menos explotables para una comercializadora concreta.

En segundo lugar, en lo relativo a la **segunda parte del trabajo**, el ámbito de mejora más prometedor radica en el diseño de la estrategia a seguir. Las estrategias aplicadas en este estudio son deliberadamente sencillas, con el objetivo de demostrar que el conocimiento previo de los perfiles de desvío permite obtener un beneficio económico. No obstante, existen enfoques más sofisticados, tanto en ingeniería como en economía, por ejemplo, en los mercados financieros, en los que se emplean **estrategias dinámicas** en lugar de estáticas. Estas no solo se ajustan al tipo de día previsto, sino que evolucionan en función de la información que se va revelando a lo largo de la jornada. Así, si se clasifica un día en el clúster 1 (caracterizado por desvíos positivos en la mañana y negativos en la tarde), un modelo dinámico podría ajustar de forma adaptativa la estrategia según la magnitud real de los desvíos observados en las primeras horas, decidiendo si conviene adoptar una posición más agresiva o conservadora en el transcurso del día.

Este segundo camino resulta especialmente atractivo para futuros proyectos, pues la optimización de estrategias dinámicas, o incluso la combinación híbrida de las estrategias propuestas en este trabajo, podría incrementar los márgenes de ahorro por encima del 10 % ya alcanzado. Asimismo, la incorporación de técnicas de aprendizaje más avanzadas, como **modelos de series temporales basados en LSTM o Transformers**, permitiría anticipar mejor los cambios intradía en la demanda y la generación, aumentando la capacidad predictiva y, con ello, el impacto económico.

Finalmente, debe destacarse que cualquier estrategia de este tipo está condicionada también por el **marco regulatorio y el papel del operador del sistema**. Tal y como señala la Comisión Nacional de los Mercados y la Competencia [CNMC22], los mecanismos de desvíos buscan incentivar comportamientos que favorezcan el equilibrio del sistema. En este contexto, si Red Eléctrica de España detectase que determinadas comercializadoras emplean sistemáticamente estas técnicas para maximizar su beneficio, podría considerar la introducción de ajustes normativos que limiten un aprovechamiento excesivo del mecanismo. Este aspecto resalta la necesidad de que cualquier proyecto futuro combine la dimensión técnica y económica con una visión regulatoria, asegurando que las mejoras propuestas no solo generen beneficios privados, sino que también contribuyan a los objetivos globales de estabilidad y eficiencia del sistema eléctrico.

Capítulo 6.- Bibliografía

[LARA20] Lara-Benítez, P.; Carranza-García, M.; Luna-Romera, J.M.; Riquelme, J.C. Temporal Convolutional Networks Applied to Energy-Related Time Series Forecasting. *Appl. Sci.* , *10*, 2322. 2020

[VARA24] Varas Yuste, G., Predicción de la demanda eléctrica en España mediante redes neuronales profundas, Trabajo Fin de Grado, Universidad Pontificia Comillas, Madrid, 2024.

[REDE25] Red Eléctrica de España, Seguimiento de la demanda de energía eléctrica, www.ree.es. 2025

[CARB07] Carbajo Josa, A., Los mercados eléctricos y los servicios de ajuste del sistema, ISSN 0422-2784, Nº 364, (Ejemplar dedicado a: Ajustes regulatorios en el sector eléctrico español), págs. 55-62. 2007

[CNMC22] Comisión Nacional de los Mercados y la Competencia (2022) : *Derechos de cobro y obligaciones de pago por los servicios de ajuste del sistema*

[GERÓ19] Géron, A. (2019). Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems. O'Reilly., , págs. 586-596. 2019

[MANR20] Manrique Rojas, Esperanza. Machine Learning: analysis of programming languages and development tools Revista Ibérica de Sistemas e Tecnologias de Informação; Lousada Iss. E28, : 586-599. (Apr 2020)

[FRAN18] François-Lavet, V., Henderson, P., Islam, R., Bellemare, M. G., & Pineau, J. An Introduction to Deep Reinforcement Learning. Foundations And Trends® In Machine Learning, 11(3-4), 219-354. (2018).

[GONZ18] González Muñiz, A. Aplicaciones de técnicas de inteligencia artificial basadas en aprendizaje profundo (deep learning) al análisis y mejora de la eficiencia de procesos industriales. Universidad de Oviedo, 2018

[KONG19] W. Kong, Z. Y. Dong, Y. Jia, D. J. Hill, Y. Xu and Y. Zhang, "Short-Term Residential Load Forecasting Based on LSTM Recurrent Neural Network," in IEEE Transactions on Smart Grid, vol. 10, no. 1, pp. 841-851, Jan. 2019

[MARI16] Marino, D. L.; Amarasinghe, K. and Manic, M.; "Building energy load forecasting using Deep Neural Networks," IECON 2016 - 42nd Annual Conference of the IEEE Industrial Electronics Society, Florence, Italy, pp. 7046-7051, 2016

[JAME23] James, G., Witten, D., Hastie, T., Tibshirani, R., & Grosse, R., An Introduction to Statistical Learning: with Applications in Python, Springer Texts in Statistics, Springer, 2023.

[BREI01] Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32. doi: 10.1023/A:1010933404324

[BRUN24] Bruneel, J., Van der Elst, S., & Meeus, L. (2024). "Managing imbalance price risk with intraday trading strategies in single-price electricity markets". *Energy Economics*, 126, 106857.

[YANG24] Yang, H., Zhang, Z., Liang, R., & Zhao, W. (2024). Research on time-of-use compensation pricing strategies for load aggregators based on user demand response. *Frontiers In Energy Research*, *12*. https://doi.org/10.3389/fenrg.2024.1442194

[PERE19] Pérez López, A., Predictor de precios del mercado mayorista de la electricidad, Proyecto Fin de Máster, dirigido por J. M. Riquelme Santos, Universidad de Sevilla, 2019.

[SAIN24] Sainz-Trápaga de Garnica, M., Modelos Reg-ARIMA para la predicción de demanda eléctrica en el archipiélago balear, Trabajo Fin de Grado, Universidad Politécnica de Madrid, Madrid, 2022

[JUDG22] Judge, M. A., Franzitta, V., Curto, D., Guercio, A., Cirrincione, G., & Khattak, H. A., A comprehensive review of artificial intelligence approaches for smart grid integration and optimization, Energy Conversion and Management: X, vol. 21, p. 100724, 2024

[APPL24] Appleute, IA generativa: visión general, [en línea], disponible en: https://www.appleute.de/es/app-entwickler-bibliothek/ia-generativa-vision-general/, consultado el 12 de abril de 2025.

[REE25] Redeia, Misión y visión de Red Eléctrica, [en línea], disponible en: https://www.redeia.com/es/conocenos/ree-en-2-minutos/mision-y-vision, consultado el 17 de abril de 2025.

[NUEN25] Naciones Unidas, Objetivo de Desarrollo Sostenible 7: Energía asequible y no contaminante, [en línea], disponible en: https://www.un.org/sustainabledevelopment/es/energy/, 2025.
Naciones Unidas, Objetivo de Desarrollo Sostenible 9: Industria, innovación e infraestructura, [en línea], disponible en: https://www.un.org/sustainabledevelopment/es/infrastructure/, 2025.
Naciones Unidas, Objetivo de Desarrollo Sostenible 13: Acción por el clima, [en línea], disponible en: https://www.un.org/sustainabledevelopment/es/climate-change-2/, 2025.

[GEEK24] GeeksforGeeks, *Elbow Method for optimal value of k in KMeans*, [en línea], disponible en: https://www.geeksforgeeks.org/elbow-method-for-optimal-value-of-k-in-kmeans/, consultado el 20 de abril de 2025

[REE25] Glosario de REE de muchs de los términos eléctricos url: https://www.ree.es/es/glosario

[ESIO25] ESIOS previsión diaria fotovoltaica. url: https://www.esios.ree.es/es/analisis/1779

[ESIO25] ESIOS previsión diaria fotovoltaica. url: https://www.esios.ree.es/es/analisis/1777

[ESIO25] ESIOS previsión diaria fotovoltaica. url: https://www.esios.ree.es/es/analisis/1485

[ESIO25] ESIOS previsión diaria fotovoltaica. url: https://www.esios.ree.es/es/analisis/1486

[INE24] Instituto Nacional de Estadística (INE). (2024). *Población por provincias y sexo*. Disponible en: https://www.ine.es/jaxiT3/Datos.htm?t=2852

[WIKI25] Wikipedia. (2025). *Anexo: Provincias de España por código postal*. Disponible en: https://es.wikipedia.org/wiki/Anexo:Provincias_de_Espa%C3%B1a_por_c%C3%B3digo_postal

[INE24] Instituto Nacional de Estadística (INE). (2024). *Encuesta Continua de Población (ECP): resultados*. Disponible en: https://www.ine.es/dyngs/INEbase/operacion.htm?c=Estadistica_C&cid=1254736177095&menu=ultiDatos&idp=1254735572981

[BRIT24] Encyclopaedia Britannica. (2024). *Köppen climate classification*. Disponible en: https://www.britannica.com/science/Koppen-climate-classification

[AEME25] Agencia Estatal de Meteorología (AEMET). (2025). *Centro de descargas de productos AEMET (Open Data)*. Disponible en: https://opendata.aemet.es/centrodedescargas/productosAEMET

[MAUR00] Mauro M. A. & Ramírez P. C. (2000). Obtención de la curva diaria de temperatura: aplicación para el cálculo de los grados día de calefacción. VI Congreso Internacional Ciencias de la Tierra: Seminario Contaminación Atmosférica y Estrategias de Control. Universidad de Santiago de Chile

[FAL18] Hong F., Zhan W., Göttsche F. M., Liu Z., Zhou J., Huang F., Lai J. & Li M. (2018). Comprehensive assessment of four-parameter diurnal land surface temperature cycle models under clear-sky. *ISPRS Journal of Photogrammetry and Remote Sensing*, 142, 190–204

[GOE16] Goela P. C. (2016). Time series analysis of data for sea surface temperature: harmonic modelling of daily cycles. *Journal of Atmospheric and Oceanic Technology*, S0924796316301142

[UE17] Reglamento (UE) 2017/2195 del Parlamento Europeo y del Consejo, de 23 de noviembre de 2017, por el que se establecen las normas de balance para los sistemas eléctricos. *Diario Oficial de la Unión Europea*, L 312, 6-53.

[DURÁ14] Durán Esteban, I. M. (2014) *Ofertas Estratégicas de generación Eólica al mercado español de electricidad,* Universidad Carlos III de Madrid, Departamento de ingeniería Eléctrica

[BOE22] Boletín Oficial del Estado 2022 *Real Decreto-ley 10/2022, de 13 de mayo, por el que se establece con carácter temporal un mecanismo de ajuste de costes de producción para la reducción del precio de la electricidad en el mercado mayorista*.

[GÁME21] Gámez Badouin, J. I. (2021) *Random Forest;* Rpub consultado el 1/08/2025 en https://rpubs.com/jigbadouin/randomforest1

[TAN19] Tan, P.-N., Steinbach, M., Karpatne, A., & Kumar, V. (2019). *Introduction to Data Mining* (2nd ed.). Pearson.

[BOE213] Boletín Oficial del Estado, Ley 24/2013, de 26 de diciembre, del Sector Eléctrico, artículo 6.1.f)

[ESIO25] ESIOS Precio del Mercado diario url: https://www.esios.ree.es/es/analisis/600

[ESIO25] ESIOS Precio de los desvíos a subir url: https://www.esios.ree.es/es/analisis/763

[ESIO25] ESIOS Precio de los desvíos a bajar url: https://www.esios.ree.es/es/analisis/764

[NIST25] NIST (National Institute of Standards and Technology). (s. f.). *2.2.2.1. Shewhart control chart*. En **NIST/SEMATECH e-Handbook of Statistical Methods** (sección Measurement Process Characterization).

[CAI99] Cai, T. T. (1999). *Adaptive wavelet estimation: A block thresholding and oracle inequality approach*. **The Annals of Statistics, 27**(3), 898–924.

Capítulo 7.- Apéndices

Anexo 1: Alineación con lo ODS

El presente proyecto se alinea con varios de los Objetivos de Desarrollo Sostenible de la Agenda 2030 de la ONU, contribuyendo a ellos de la siguiente manera:

- **ODS 7: Energía asequible y no contaminante.** Este objetivo busca garantizar el acceso a una energía fiable, sostenible y moderna para todos, así como mejorar la eficiencia energética. El proyecto contribuye al ODS 7 al proponer métodos para optimizar la operación del sistema eléctrico, reduciendo ineficiencias (desvíos) en la gestión de la oferta y la demanda de electricidad. Si logramos disminuir los desvíos diarios, se minimizan las necesidades de energía de reserva y ajustes de última hora, lo que redunda en una red más eficiente y estable. A largo plazo, una mejor gestión de los desvíos puede facilitar la integración de energías renovables (al comprender mejor las causas de los desequilibrios), apoyando una transición hacia fuentes limpias sin comprometer la confiabilidad del suministro [**NUEN25**].
- **ODS 9: Industria, innovación e infraestructura.** El objetivo 9 promueve la modernización de las infraestructuras y la adopción de tecnologías innovadoras en la industria. Este TFG se inscribe claramente en esa línea al aplicar técnicas de IA avanzadas (machine learning supervisado y no supervisado) en la infraestructura eléctrica nacional. Supone una innovación metodológica en la forma de analizar datos de la red, fomentando la introducción de herramientas de la llamada Industria 4.0 en el sector eléctrico. Los resultados podrían traducirse en prácticas operativas más inteligentes dentro de REE, fortaleciendo la resiliencia y eficiencia de la infraestructura eléctrica mediante la ciencia de datos. [NUEN25]
- **ODS 13: Acción por el clima**. De forma indirecta, el proyecto también apoya el ODS 13, que insta a adoptar medidas urgentes contra el cambio climático. Una red eléctrica operada de manera más eficiente (con menos desvíos y desperdicios de energía) implica un uso más racional de los recursos energéticos, lo que contribuye a reducir las emisiones de gases de efecto invernadero asociadas a la generación de electricidad de respaldo. Además, al facilitar la incorporación de energía renovable variable (gracias a una mejor comprensión y anticipación de sus desviaciones), se está promoviendo una matriz eléctrica más limpia. En suma, optimizar la gestión eléctrica mediante IA ayuda a mitigar el impacto ambiental del sector energético, alineándose con las metas climáticas [NUEN25].

Anexo 2: Código de K-means

```
variables = ['Prevision_Foto', 'Eólica_Diaria', 'Instalada_Foto',
'Instalada_Eolica',
             'Prev_Demanda', 'Disp_Nuclear', 'Festividad', 'Lunes',
'Martes-Jueves',
             'Viernes-Domingo/Festivo', 'Temp_Madrid', 'Temp_Tenerife',
'Temp_Sevilla',
             'Temp_Santiago', 'Temp_Coruña', 'Temp_Zaragoza']

import pandas as pd
import datetime
```

```
# Cargar el archivo Excel
df = pd.read_excel("Datos_K.xlsx")

# Convertir 'Fecha' a tipo date
df['Fecha'] = pd.to_datetime(df['Fecha']).dt.date

# Corregir las horas: si están normalizadas (0-1), multiplicarlas por 24
if df['Hora'].max() <= 1:
    df['Hora'] = df['Hora'] * 24

# Convertir 'Hora' a entero (0-23)
df['Hora'] = df['Hora'].apply(lambda x: int(x) if not pd.isnull(x) else
x)

# Reemplazar valores nulos en la columna 'Festividad' con 0
df['Festividad'] = df['Festividad'].fillna(0)


# Pivotar el DataFrame para que cada fecha tenga 24 columnas (una por
hora)
pivoted_df = df_grouped.pivot(index='Fecha', columns='Hora',
values=variables)

# Asegurarse de que las columnas están ordenadas por hora
pivoted_df = pivoted_df.sort_index(axis=1)
# Rellenar valores NaN con la media entre el dato anterior y el siguiente
pivoted_df = pivoted_df.interpolate(method='linear', axis=0,
limit_direction='both')

# Mostrar las primeras filas del DataFrame reorganizado
print(pivoted_df.head())
from sklearn.clúster import KMeans
from sklearn.preprocessing import MinMaxScaler
from sklearn.clúster import KMeans
from sklearn.preprocessing import MinMaxScaler
import numpy as np
import pandas as pd

# Asegurarnos de que los datos están en la estructura correcta
pivoted_df = df_grouped.pivot(index='Fecha', columns='Hora',
values=variables)

# Aplanar los datos para K-means
flattened_data = pivoted_df.map(lambda x: np.array(x) if isinstance(x,
list) else np.array([x]))
flattened_data = np.array(flattened_data.values.tolist())  # Convertir a
un array de NumPy
flattened_data = flattened_data.reshape(flattened_data.shape[0], -1)  #
Aplanar las horas
```

```
# Normalizar las variables (escalar al rango 0-1)
scaler = MinMaxScaler()
normalized_data = scaler.fit_transform(flattened_data)

# Número de clústers
n_clústers = 7  # Puedes ajustar este valor según tus necesidades
print(f"Usando {n_clústers} clústers para K-means.")

# Aplicar K-means
kmeans = KMeans(n_clústers=n_clústers, random_state=42)
kmeans.fit(normalized_data)

# Agregar los labels al DataFrame original
pivoted_df['Clúster'] = kmeans.labels_

# Mostrar los resultados
print("Centroides de los clústers:")
print(kmeans.clúster_centers_)
print("\nAsignación de clústers:")
print(pivoted_df[['Clúster']])
# Método del codo
inertia = []
k_values = range(1, 11)  # Probar con 1 a 10 clústers
for k in k_values:
    kmeans = KMeans(n_clústers=k, random_state=42)
    kmeans.fit(normalized_data)
    inertia.append(kmeans.inertia_)

# Guardar los valores de k y sus inercias en un archivo Excel
k_inertia_df = pd.DataFrame({"k_values": k_values, "inertia": inertia})
k_inertia_df.to_csv("k_values.csv", index=False)

print("Valores de k y sus inercias guardados en 'k_values.xlsx'.")

# Guardar pivoted_df en un archivo CSV
pivoted_df.to_csv("pivoted_df.csv", index=False)

# Guardar los datos normalizados en un archivo .npy
np.save("normalized_data.npy", normalized_data)
# Asegurarse de que 'Fecha' esté como columna en el DataFrame
if 'Fecha' not in pivoted_df.columns:
    pivoted_df.reset_index(inplace=True)  # Mover el índice 'Fecha' a una
columna

# Aplanar el MultiIndex de las columnas
pivoted_df.columns = ['_'.join(map(str, col)).strip() if isinstance(col,
tuple) else col for col in pivoted_df.columns]
```

```
# Guardar los datos con los clústers en un archivo Excel
pivoted_df.to_excel("pivoted_df_with_clústers.xlsx", index=False)
print("Archivo 'pivoted_df_with_clústers.xlsx' generado correctamente.")
```

Anexo 3: Elbow Method

```
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np

# Leer el archivo CSV con los datos de k_values
k_values_df = pd.read_csv("k_values.csv")

# Extraer los valores de k y sus inercias
k_values = k_values_df["k_values"].tolist()
inertia = k_values_df["inertia"].tolist()

# Graficar el método del codo
plt.figure(figsize=(8, 5))
plt.plot(k_values, inertia, marker='o', linestyle='--')
plt.title('Método del Codo')
plt.xlabel('Número de Clústers (k)')
plt.ylabel('Inercia')
plt.xticks(k_values)
plt.grid()
plt.show()

# Determinar el número óptimo de clústers (codo)
differences = np.diff(inertia)  # Diferencias entre puntos consecutivos
second_differences = np.diff(differences)  # Segunda derivada
optimal_k_index = np.argmin(second_differences) + 1  # Índice del codo
(ajustado por np.diff)
optimal_k = k_values[optimal_k_index]

# Mostrar el número óptimo de clústers
print(f"El número óptimo de clústers según el método del codo es:
{optimal_k}")
```

Anexo 4: Código para comprobar el número óptimo de árboles

```
import os, numpy as np, pandas as pd, matplotlib.pyplot as plt
from datetime import datetime
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import accuracy_score, balanced_accuracy_score

# Rutas
output_path = r"C:\Icai\TFG\Python"
pivot_file  = os.path.join(output_path, "pivoted_df_with_clústers.xlsx")
```

```
# Lectura y preparación (mismo esquema que usas)
df = pd.read_excel(pivot_file)
df = df.rename(columns={df.columns[0]: 'Fecha', df.columns[-1]:
'Clúster_KMeans'})
df['Fecha'] = pd.to_datetime(df['Fecha'])

# Split temporal: último año como test
max_date  = df['Fecha'].max()
threshold = max_date - pd.DateOffset(years=1)
train = df[df['Fecha'] <  threshold].copy()
test  = df[df['Fecha'] >= threshold].copy()

features = [c for c in df.columns if c not in ['Fecha','Clúster_KMeans']]
X_train  = train[features].values
y_train  = train['Clúster_KMeans'].values
X_test   = test[features].values
y_test   = test['Clúster_KMeans'].values

# Barrido de n_estimators con varias semillas
n_list = [50, 100, 150, 200, 300, 400, 600, 800]
seeds  = [0, 1, 2, 3, 4]

rows = []
for n in n_list:
    accs, baccs, oobs = [], [], []
    for s in seeds:
        rf = RandomForestClassifier(
            n_estimators=n, oob_score=True, n_jobs=-1, random_state=s
        )
        rf.fit(X_train, y_train)
        y_pred = rf.predict(X_test)
        accs.append(accuracy_score(y_test, y_pred))
        baccs.append(balanced_accuracy_score(y_test, y_pred))
        oobs.append(1 - rf.oob_score_)
    rows.append({
        'n_estimators': n,
        'acc_mean': np.mean(accs), 'acc_std': np.std(accs),
        'bacc_mean': np.mean(baccs), 'bacc_std': np.std(baccs),
        'oob_err_mean': np.mean(oobs), 'oob_err_std': np.std(oobs),
    })

res = pd.DataFrame(rows)

# Regla de meseta (elige el menor n cuya acc_mean esté a <= 0.002 del
máximo
# y cuyo oob_err_mean esté a <= 0.002 del mínimo)
max_acc = res['acc_mean'].max()
min_oob = res['oob_err_mean'].min()
candidatos = res[
```

```
    (max_acc - res['acc_mean'] <= 0.002) &
    (res['oob_err_mean'] - min_oob <= 0.002)
].sort_values('n_estimators')
n_opt = int(candidatos.iloc[0]['n_estimators']) if not candidatos.empty
else int(res.loc[res['acc_mean'].idxmax(), 'n_estimators'])

print("Resumen tamaño del bosque:")
print(res.to_string(index=False))
print(f"\n>>> n_opt recomendado = {n_opt} (meseta acc & OOB)")

# Guardar CSV y figura
stamp = datetime.now().strftime("%Y%m%d_%H%M%S")
csv_out = os.path.join(output_path, f"RF_ModelSize_Summary_{stamp}.csv")
res.to_csv(csv_out, index=False)

plt.figure(figsize=(8,5))
# Accuracy test (media ± sd)
plt.plot(res['n_estimators'], res['acc_mean'], marker='o',
label='Accuracy test (media)')
plt.fill_between(res['n_estimators'],
                 res['acc_mean']-res['acc_std'],
                 res['acc_mean']+res['acc_std'],
                 alpha=0.2)
# OOB error (media ± sd) en eje secundario
ax = plt.gca()
ax2 = ax.twinx()
ax2.plot(res['n_estimators'], res['oob_err_mean'], marker='s',
linestyle='--', label='Error OOB (media)', alpha=0.8)
ax2.fill_between(res['n_estimators'],
                 res['oob_err_mean']-res['oob_err_std'],
                 res['oob_err_mean']+res['oob_err_std'],
                 alpha=0.15)

# Marca n_opt
plt.axvline(n_opt, color='gray', linestyle=':', alpha=0.7)
ax.set_xlabel('Número de árboles (n_estimators)')
ax.set_ylabel('Accuracy de test')
ax2.set_ylabel('Error OOB')
ax.legend(loc='lower right')
ax2.legend(loc='upper right')
plt.title('Random Forest: calidad vs número de árboles (media ± sd, 5
semillas)')
fig_out = os.path.join(output_path, f"RF_ModelSize_Curves_{stamp}.png")
plt.tight_layout()
plt.savefig(fig_out, dpi=150)
print(f"✓ Guardados:\n  - {csv_out}\n  - {fig_out}")
```

Anexo 5: Código ejecución Random Forest y vecinos lógicos

```python
import os
import pandas as pd
import numpy as np
from collections import Counter, defaultdict
from datetime import datetime

from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import confusion_matrix, classification_report,
accuracy_score, balanced_accuracy_score
from sklearn.preprocessing import StandardScaler

# ---------------------------------------------------------------------
# 0) Parámetros y rutas
# ---------------------------------------------------------------------
output_path = r"C:\Icai\TFG\Python"
pivot_file  = os.path.join(output_path, "pivoted_df_with_clústers.xlsx")

# Ajusta si quieres otro tamaño del bosque o semilla:
N_ESTIMATORS = 200
RANDOM_STATE = 42

# ---------------------------------------------------------------------
# 1) Lectura del fichero pivotado con Clúster_KMeans
# ---------------------------------------------------------------------
df = pd.read_excel(pivot_file)
# Renombrar columnas: primera = Fecha, última = Clúster_KMeans (según tu
flujo)
df = df.rename(columns={df.columns[0]: 'Fecha', df.columns[-1]:
'Clúster_KMeans'})
df['Fecha'] = pd.to_datetime(df['Fecha'])

# ---------------------------------------------------------------------
# 2) División Train/Test (último año como test)
# ---------------------------------------------------------------------
max_date  = df['Fecha'].max()
threshold = max_date - pd.DateOffset(years=1)
train = df[df['Fecha'] <  threshold].copy()
test  = df[df['Fecha'] >= threshold].copy()

# ---------------------------------------------------------------------
# 3) Preparar X e y
# ---------------------------------------------------------------------
features = [c for c in df.columns if c not in ['Fecha','Clúster_KMeans']]
X_train  = train[features].values
y_train  = train['Clúster_KMeans'].values
X_test   = test[features].values
y_test   = test['Clúster_KMeans'].values

# ---------------------------------------------------------------------
```

```
# 4) Entrenamiento del Random Forest
# ------------------------------------------------------------------
rf = RandomForestClassifier(n_estimators=N_ESTIMATORS,
random_state=RANDOM_STATE, n_jobs=-1, oob_score=True)
rf.fit(X_train, y_train)

# ------------------------------------------------------------------
# 5) Predicción sobre test
# ------------------------------------------------------------------
y_pred       = rf.predict(X_test)
y_pred_proba = rf.predict_proba(X_test)

# ------------------------------------------------------------------
# 6) Métricas
# ------------------------------------------------------------------
acc    = accuracy_score(y_test, y_pred)
bacc   = balanced_accuracy_score(y_test, y_pred)
conf   = confusion_matrix(y_test, y_pred)
report= classification_report(y_test, y_pred, digits=3)

print(f"\n=== RF (n_estimators={N_ESTIMATORS}, seed={RANDOM_STATE})")
print(f"Split temporal: train < {threshold.date()} — test ≥
{threshold.date()} ===")
print(f"Accuracy: {acc:.4f} | Balanced Acc: {bacc:.4f} | OOB Acc:
{rf.oob_score_:.4f} (OOB Err: {1-rf.oob_score_:.4f})")
print("Matriz de confusión:\n", conf)
print("Informe de clasificación:\n", report)

# ------------------------------------------------------------------
# 7) Guardar resultados detallados (sin sobreescribir)
# ------------------------------------------------------------------
stamp = datetime.now().strftime("%Y%m%d_%H%M%S")

results = test[['Fecha','Clúster_KMeans']].copy()
results['Clúster_RF_Predicho'] = y_pred
results['Max_Prob'] = y_pred_proba.max(axis=1)
out1 = os.path.join(output_path,
f"Resultados_RF_{threshold.date()}_{stamp}.xlsx")
results.to_excel(out1, index=False)
print(f"✓ Resultados RF guardados en: {out1}")

# Comparación train+test
comp_train = train[['Fecha','Clúster_KMeans']].copy()
comp_train['Clúster_RF_Predicho'] = np.nan
comp_train['Fase'] = 'Train'
comp_test  = test[['Fecha','Clúster_KMeans']].copy()
comp_test ['Clúster_RF_Predicho'] = y_pred
comp_test ['Fase'] = 'Test'
comparison = pd.concat([comp_train, comp_test], ignore_index=True)
```

```
comparison =
comparison[['Fecha','Fase','Clúster_KMeans','Clúster_RF_Predicho']]
out2 = os.path.join(output_path,
f"Comparacion_RF_vs_KMeans_{stamp}.xlsx")
comparison.to_excel(out2, index=False)
print(f"✓ Comparación guardada en: {out2}")

# ----------------------------------------------------------------------
# 8) EXPLICABILIDAD: elegir un ejemplo "interesante" del test
#    a) Preferentemente un mal clasificado
#    b) Si no hay errores, el más "dudoso" (gap top-1 - top-2 más
pequeño)
# ----------------------------------------------------------------------
mis_idx = np.where(y_pred != y_test)[0]
if len(mis_idx) > 0:
    chosen_idx = mis_idx[0]
    motivo = "mal_clasificado"
else:
    diffs = []
    for i in range(len(y_test)):
        probs = np.sort(y_pred_proba[i])[::-1]
        gap = probs[0] - probs[1] if len(probs) > 1 else 0.0
        diffs.append(gap)
    chosen_idx = int(np.argmin(diffs))
    motivo = "mas_dudoso"

x_row   = X_test[chosen_idx].reshape(1, -1)
true_y  = y_test[chosen_idx]
pred_y  = y_pred[chosen_idx]
date_y  = test.iloc[chosen_idx]['Fecha']

print(f"\n=== Ejemplo {motivo} ===")
print(f"Fecha: {date_y} | Verdadero: {true_y} | Predicho: {pred_y}")

# 8.1) Votación de los árboles
tree_votes = [est.predict(x_row)[0] for est in rf.estimators_]
vote_count = Counter(tree_votes)
total_trees = len(rf.estimators_)
print("\nVotación por clase (árboles que votan cada clase):")
for cls, cnt in sorted(vote_count.items()):
    print(f"  Clase {cls}: {cnt}/{total_trees} árboles
({cnt/total_trees:.1%})")

# Probabilidades de bosque
probas = rf.predict_proba(x_row)[0]
for cls, p in zip(rf.classes_, probas):
    print(f"  Probabilidad bosque para clase {cls}: {p:.3f}")

# 8.2) Variables más influyentes en ESTE ejemplo (rutas de decisión)
```

```
feature_counts   = defaultdict(int)
feature_strength = defaultdict(float)

for est in rf.estimators_:
    tree = est.tree_
    node_indicator = est.decision_path(x_row)
    node_index =
node_indicator.indices[node_indicator.indptr[0]:node_indicator.indptr[1]]
    for node_id in node_index:
        f_idx = tree.feature[node_id]
        thr   = tree.threshold[node_id]
        if f_idx == -1:
            continue  # hoja
        val = x_row[0, f_idx]
        feature_counts[f_idx]   += 1
        feature_strength[f_idx] += float(abs(val - thr))  # proxy fuerza
local

feat_df = pd.DataFrame({
    'feature': [features[i] for i in feature_counts.keys()],
    'apariciones_en_rutas': list(feature_counts.values()),
    'fuerza_acumulada': [feature_strength[i] for i in
feature_counts.keys()]
})
if not feat_df.empty:
    feat_df['rank_frecuencia'] = feat_df['apariciones_en_rutas'] /
feat_df['apariciones_en_rutas'].max()
    feat_df['rank_fuerza']     = feat_df['fuerza_acumulada'] /
feat_df['fuerza_acumulada'].max()
    feat_df['score_local']     = 0.5*feat_df['rank_frecuencia'] +
0.5*feat_df['rank_fuerza']
    feat_df =
feat_df.sort_values(['score_local','apariciones_en_rutas','fuerza_acumula
da'], ascending=False)
topk = feat_df.head(10).reset_index(drop=True)

print("\nTop variables que más influyeron en ESTE ejemplo:")
if topk.empty:
    print("  (No se pudieron extraer rutas — revisa el modelo/datos.)")
else:
    for i, r in topk.iterrows():
        print(f"  {i+1:>2}. {r['feature']} | apariciones:
{int(r['apariciones_en_rutas'])} | score_local: {r['score_local']:.3f}")

# 8.3) Contexto (valor del ejemplo vs media del test)
context_rows = []
for f in topk['feature'].tolist():
    j = features.index(f)
    example_val = x_row[0, j]
```

```
    test_mean   = float(np.nanmean(X_test[:, j]))
    context_rows.append({'feature': f, 'valor_ejemplo': example_val,
'media_test': test_mean, 'delta': example_val - test_mean})
context_df = pd.DataFrame(context_rows)

# 8.4) Guardar explicación a Excel (Resumen + Top locales + Contexto)
out_exp = os.path.join(output_path,
f"Explicacion_RF_{motivo}_{stamp}.xlsx")
with pd.ExcelWriter(out_exp, engine='xlsxwriter') as writer:
    resumen = pd.DataFrame({
        'Fecha':[date_y], 'Verdadero':[true_y], 'Predicho':[pred_y],
'Motivo':[motivo]
    })
    for cls, p in zip(rf.classes_, probas):
        resumen[f'Prob_clase_{cls}'] = p
    for cls in rf.classes_:
        resumen[f'Votos_clase_{cls}'] = vote_count.get(cls, 0)
    resumen.to_excel(writer, sheet_name='Resumen', index=False)
    topk.to_excel(writer, sheet_name='Top_variables_locales',
index=False)
    context_df.to_excel(writer, sheet_name='Contexto_variables',
index=False)
print(f"✓ Explicación detallada guardada en: {out_exp}")

# ---------------------------------------------------------------------
# 9) “Vecinos lógicos”: ¿los clústeres confundidos son realmente
parecidos?
#     - Matriz de confusión (test)
#     - Centroides en espacio estandarizado (24×p)
#     - Distancia euclídea / similitud coseno / correlación Pearson
# ---------------------------------------------------------------------

# 9.1) Matriz de confusión con etiquetas ordenadas
labels = np.unique(y_train)
cm = confusion_matrix(y_test, y_pred, labels=labels)
cm_df = pd.DataFrame(cm, index=[f"true_{l}" for l in labels],
columns=[f"pred_{l}" for l in labels])
print("\nMatriz de confusión (test):")
print(cm_df)

# 9.2) Centroides por clúster en espacio estandarizado (para comparar
perfiles completos)
X_all   = df[features].values
y_all   = df['Clúster_KMeans'].values
scaler  = StandardScaler()
X_all_z = scaler.fit_transform(X_all)

centroids = {}
for l in labels:
```

```
    centroids[l] = X_all_z[y_all == l].mean(axis=0)
cent_df = pd.DataFrame(centroids).T  # filas: clúster, cols: features (z)

# Utilidades sin scipy:
def cosine_similarity(u, v):
    nu = np.linalg.norm(u); nv = np.linalg.norm(v)
    if nu == 0 or nv == 0: return 0.0
    return float(np.dot(u, v) / (nu * nv))

def pearson_corr(u, v):
    uu = u - np.mean(u); vv = v - np.mean(v)
    num = np.dot(uu, vv)
    den = np.linalg.norm(uu) * np.linalg.norm(vv)
    if den == 0: return 0.0
    return float(num / den)

# 9.3) PARES MÁS CONFUNDIDOS (excluye diagonal)
pairs = []
for i, li in enumerate(labels):
    for j, lj in enumerate(labels):
        if i == j:
            continue
        cnt = cm[i, j]
        if cnt > 0:
            pairs.append((li, lj, int(cnt)))
pairs = sorted(pairs, key=lambda x: x[2], reverse=True)

rows = []
for (a,b,cnt) in pairs[:10]:
    v1 = centroids[a]
    v2 = centroids[b]
    dist_euc = float(np.linalg.norm(v1 - v2))
    sim_cos  = cosine_similarity(v1, v2)         # 1 = idénticos, 0 =
ortogonales
    corr     = pearson_corr(v1, v2)              # 1 = misma forma global
    rows.append({
        'par_true_pred': f'{a}→{b}',
        'confusiones_test': cnt,
        'dist_euclidea_centroides_z': dist_euc,
        'similitud_coseno': sim_cos,
        'correlacion_pearson': corr
    })
pairs_df = pd.DataFrame(rows).sort_values('confusiones_test',
ascending=False)

print("\nTop pares confundidos y similitud de perfiles (centroides
estandarizados):")
if pairs_df.empty:
    print("  (No hubo confusiones en test).")
```

```
else:
    print(pairs_df.to_string(index=False))

# 9.4) Guardar a Excel (matriz, centroides y métricas de parejas)
out_nb = os.path.join(output_path,
f"RF_Confusion_and_Neighbors_{stamp}.xlsx")
with pd.ExcelWriter(out_nb, engine='xlsxwriter') as writer:
    cm_df.to_excel(writer, sheet_name='Confusion_Matrix')
    cent_df.to_excel(writer, sheet_name='Centroids_zspace')
    if not pairs_df.empty:
        pairs_df.to_excel(writer, sheet_name='Confused_Pairs_Metrics',
index=False)
print(f"✓ Vecinos lógicos guardados en: {out_nb}")

# 9.5) Interpretación rápida en consola (si hay confusiones)
if not pairs_df.empty:
    top = pairs_df.iloc[0]
    print("\nInterpretación rápida del par más confundido:")
    print(f"- Par {top['par_true_pred']} con
{int(top['confusiones_test'])} confusiones.")
    print(f"- Distancia euclídea (centroides en z):
{top['dist_euclidea_centroides_z']:.3f} (más pequeña = más parecidos).")
    print(f"- Similitud coseno: {top['similitud_coseno']:.3f} (≈1 muy
similares).")
    print(f"- Correlación Pearson: {top['correlacion_pearson']:.3f} (≈1
misma forma global).")
    print("Si coseno y Pearson son altos y la distancia es baja, la
confusión es 'lógica' (perfiles muy próximos).")
    # --- TODAS LAS DISTANCIAS / SIMILITUDES ENTRE CENTROIDES ---
C = cent_df.values
labs = cent_df.index.to_numpy()

# Distancias euclídeas (z-espacio)
D = np.linalg.norm(C[:, None, :] - C[None, :, :], axis=2)

# Similitud coseno
norms = np.linalg.norm(C, axis=1, keepdims=True)
Cn = C / np.clip(norms, 1e-12, None)
S_cos = Cn @ Cn.T  # 1 = idénticos

# Correlación de Pearson entre centroides
S_corr = np.corrcoef(C)

# DataFrames completos (K x K)
dist_df = pd.DataFrame(D, index=labs, columns=labs)
cos_df  = pd.DataFrame(S_cos, index=labs, columns=labs)
pear_df = pd.DataFrame(S_corr, index=labs, columns=labs)

# Lista de TODOS los pares ordenados por distancia (i<j)
```

```
pairs_all = []
for i in range(len(labs)):
    for j in range(i+1, len(labs)):
        pairs_all.append({
            'par': f'{labs[i]}-{labs[j]}',
            'dist_euclidea': float(D[i, j]),
            'similitud_coseno': float(S_cos[i, j]),
            'correlacion_pearson': float(S_corr[i, j])
        })
pairs_all_df = pd.DataFrame(pairs_all).sort_values('dist_euclidea',
ascending=True)

# Vecino más cercano de cada clúster
nearest = []
for i in range(len(labs)):
    j = np.argsort(D[i])[1]  # [0] sería él mismo
    nearest.append({
        'clúster': labs[i],
        'vecino_mas_cercano': labs[j],
        'dist_euclidea': float(D[i, j]),
        'similitud_coseno': float(S_cos[i, j]),
        'correlacion_pearson': float(S_corr[i, j])
    })
nearest_df = pd.DataFrame(nearest).sort_values('dist_euclidea',
ascending=True)

# --- RESUMEN EN UNA LÍNEA (par 3-6 vs todas las distancias) ---
all_d = D[np.triu_indices(len(labs), 1)]
mean_d = float(all_d.mean())
std_d  = float(all_d.std(ddof=0))
# Busca 3 y 6 si existen
lab2idx = {l: i for i, l in enumerate(labs)}
if (3 in lab2idx) and (6 in lab2idx):
    i, j = lab2idx[3], lab2idx[6]
    d36 = float(D[i, j])
    sorted_all = np.sort(all_d)
    percentile = 100.0 * (np.searchsorted(sorted_all, d36, side='right')
/ sorted_all.size)
    print(f"Dist(centroides) 3–6 = {d36:.3f}; media inter-clúster =
{mean_d:.3f}, sd = {std_d:.3f}; percentil = {percentile:.1f}")
else:
    print("Aviso: no se han encontrado ambos clústeres 3 y 6 en las
etiquetas.")
dist_df.to_excel(writer, sheet_name='Pairwise_Euclidean')
cos_df.to_excel(writer,  sheet_name='Pairwise_CosineSim')
pear_df.to_excel(writer, sheet_name='Pairwise_Pearson')
pairs_all_df.to_excel(writer, sheet_name='All_Pairs_Sorted', index=False)
nearest_df.to_excel(writer, sheet_name='Nearest_Neighbors', index=False)
```

```
C    = cent_df.values                          # filas: clúster, cols:
features estandarizadas
labs = cent_df.index.to_numpy()                # etiquetas de clúster
(p.ej., 0..6)

pares = []
for i, li in enumerate(labs):
    for j, lj in enumerate(labs):
        if j <= i:
            continue  # solo pares i<j: 0-1, 0-2, ..., 6-5
        d = float(np.linalg.norm(C[i] - C[j]))
        pares.append({'par': f'{li}-{lj}', 'dist_euclidea_z': d})

distancias_df = pd.DataFrame(pares)

# Guardar a Excel: distancias_euclideas.xlsx
out_dist = os.path.join(output_path, "distancias_euclideas.xlsx")
try:
    import xlsxwriter  # noqa
    engine = 'xlsxwriter'
except Exception:
    engine = 'openpyxl'

with pd.ExcelWriter(out_dist, engine=engine) as writer:
    distancias_df.to_excel(writer, sheet_name='distancias', index=False)

print(f"✓ Distancias euclídeas guardadas en: {out_dist}")
```

Anexo 6: Tablas Estrategia 1 Caso Desvíos Aleatorios

Compra original	Estrategia 1	Compra con estrategia	Desvío Original	Consumo real	Desvío con estrategia
100,0	0,1	100,1	-8,7	115,0	14,9
100,0	4,2	104,2	1,0	107,8	3,6
100,0	5,0	105,0	11,0	107,4	2,5
100,0	5,4	105,4	-8,6	108,5	3,2
100,0	5,5	105,5	-7,8	107,3	1,8
100,0	6,2	106,2	7,4	109,8	3,6

Diferencia de desvío	Coste original	Coste Con estrategia	Desvíos netos	Desvíos estrategia netos
-0,1	6,1	6,1	8,7	8,8
-2,1	27,0	0,0	1,0	3,1
5,0	35,7	19,6	11,0	6,0
-5,4	24,2	39,3	8,6	14,0
-5,5	10,0	17,1	7,8	13,3
6,2	22,2	3,6	7,4	1,2

Anexo 7: Tablas Estrategia 2 Caso desvíos Aleatorios

Compra original	Umbral	Estrategia (1=Sí / 0=No?	Estrategia 2	Estrategia 2 ajustada
100	1,00	1	3	3,0
100	0,18	0	0	0,0
100	0,24	0	0	0,0
100	0,29	0	0	0,0
100	0,28	0	0	0,0
100	0,31	0	0	0,0

Desvío Original	Consumo real	Desvío con estrategia	Diferencia de desvío	Coste original	Coste Con estrategia	Desvíos netos	Desvíos estrategia netos
-8,7	91,3	-11,7	-3,0	6,1	8,2	8,7	11,7
1,0	101,0	1,0	0,0	27,0	27,0	1,0	1,0
11,0	111,0	11,0	0,0	35,7	35,7	11,0	11,0
-8,6	91,4	-8,6	0,0	24,2	24,2	8,6	8,6
-7,8	92,2	-7,8	0,0	10,0	10,0	7,8	7,8
7,4	107,4	7,4	0,0	22,2	22,2	7,4	7,4