GRADO EN INGENIERÍA EN TECNOLOGÍAS DE TELECOMUNICACIÓN

TRABAJO FIN DE GRADO

Análisis de soluciones quasi-óptimas con Machine Learning

Autor: Alberto Martín Colino

Director: Sara Lumbreras Sancho

Co-Director: Christian Perau

Madrid

Fecha: 10/6/2025

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

Análisis de soluciones quasi-óptimas con Machine Learning

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.: Alberto Martín Colino Fecha: 10/6/ 2025

Autorizada la entrega del proyecto

EL DIRECTOR DEL PROYECTO

Fdo.: Sara Lumbreras Sancho Fecha: ..3…/ …7../ ….2025

COMILLAS
UNIVERSIDAD PONTIFICIA
ICAI

GRADO EN INGENIERÍA EN TECNOLOGÍAS DE TELECOMUNICACIÓN

TRABAJO FIN DE GRADO

Análisis de soluciones quasi-óptimas con Machine Learning

Autor: Alberto Martín Colino

Director: Sara Lumbreras Sancho

Co-Director: Christian Perau

Madrid

Fecha: 10/6/2025

Agradecimientos

Seré breve.

En primer lugar, me gustaría agradecer a todos mis compañeros de clase. Mis esfuerzos por ser el mejor delegado de clase del mundo me han llevado a ser el delegado de la mejor clase del mundo, hecho que me hace muy feliz. Ha sido un gusto y un honor compartir clase con ellos, encontrando en esta familiaridad el mejor de los motivos para madrugar cada mañana.

Por otro lado, este proyecto nunca hubiera salido adelante sin la inestimable ayuda de mi hermano, Arturo, cuya insistencia y ganas de hacer un buen trabajo han sido un ejemplo y una motivación para mí.

ANÁLISIS DE SOLUCIONES QUASI-ÓPTIMAS CON MACHINE LEARNING

Autor: Martín Colino, Alberto.
Director: Lumbreras Sancho, Sara.
Entidad Colaboradora: ICAI – Universidad Pontificia Comillas

RESUMEN DEL PROYECTO

El presente estudio aborda la optimización de costes en la expansión de la red de transmisión eléctrica mediante la evaluación de un conjunto de inversiones, distintas tanto en tipología como en horizonte temporal (2030, 2040 y 2050). Tras un análisis inicial de los datos, se aplican técnicas de aprendizaje no supervisado con el fin de identificar patrones y sinergias entre las distintas alternativas. El propósito final consiste en delimitar un conjunto de soluciones quasi-óptimas que garantice la cobertura de la demanda futura de la red buscando la máxima eficiencia económica.

Palabras clave: *Transmission Expansion Planning*, *Clustering* no supervisado, Minimización de coste, Redes de transmisión

1. Introducción

La penetración masiva de energías renovables y la electrificación de la demanda han elevado de forma inédita la variabilidad de los flujos de potencia en la red de transporte. Este nuevo escenario obliga a los operadores a ampliar y reforzar las infraestructuras con criterios de flexibilidad y resiliencia que superen el enfoque determinista tradicional. El problema se conoce como ***Transmission Expansion Planning*** (TEP) y consiste en decidir qué líneas, subestaciones y dispositivos de compensación deben construirse o modernizarse a fin de garantizar la seguridad de suministro al menor coste posible. El análisis se realiza con perspectiva a corto, medio y largo plazo. En la práctica, el TEP se aborda hoy mediante modelos de optimización que producen una solución óptima única bajo un conjunto de hipótesis concretas. No obstante, la volatilidad de precios, la evolución tecnológica y las incertidumbres climáticas evidencian la fragilidad de esta respuesta única, pues pequeñas variaciones en las condiciones pueden invalidar la planificación propuesta.

2. Definición del proyecto

Con el propósito de ofrecer a los gestores de red un abanico de alternativas igualmente competitivas y, por tanto, más robustas, este trabajo plantea un enfoque basado en aprendizaje no supervisado. El proceso comienza con un estudio detallado de una base de datos de diferentes inversiones en fuentes de energía, con un preprocesado intensivo que busca simplificar el conjunto de soluciones disponibles y extraer la máxima información posible. Esta fase inicial permite reducir la dimensionalidad del problema y mejorar la precisión de los algoritmos aplicados posteriormente. A partir del amplio conjunto de soluciones factibles generado por el optimizador, se pretende descubrir agrupaciones de configuraciones de coste similar y caracterizar los rasgos que comparten. Para ello se recurre a técnicas de reducción de dimensionalidad, que facilitan la interpretación de un espacio de variables de alta complejidad, y a seis algoritmos de *clustering*—tanto jerárquicos como no jerárquicos—capaces de revelar patrones sin necesidad de etiquetas externas ni supuestos sobre la forma de los grupos. Una vez realizados todos los análisis, se estudiarán los clusteres con mejor comportamiento en términos de coste. De este modo se pasa de una única propuesta "óptima" a un repertorio de configuraciones quasi-óptimas que conservan la eficiencia económica, pero añaden margen de maniobra frente a la incertidumbre.

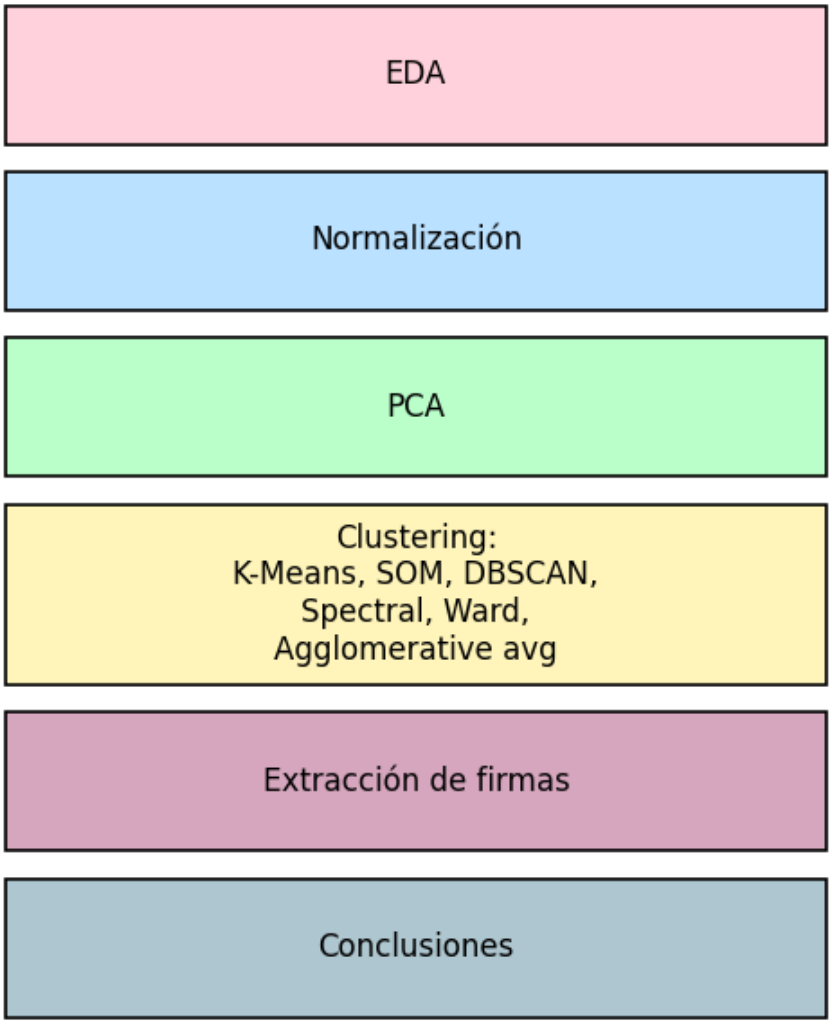
3. Descripción del modelo

El punto de partida lo constituye una exploración minuciosa del conjunto completo de soluciones facilitado por los operadores TNO (*Transmission Network Operator*) y DTI (*Dynamics Technologies Institute*), entidades que proporcionan datos relacionadas a planificación y optimización de redes eléctricas. En esta fase preliminar se depuran registros inconsistentes, se analizan las distribuciones del coste objetivo y se identifican las variables energéticas con mayor relevancia técnica. Los datos limpios se normalizan para corregir posibles sesgos de escala y, a continuación, se proyectan mediante Análisis de Componentes Principales (PCA) con el fin de reducir la dimensionalidad, mitigando así la colinealidad propia de parámetros técnico-económicos altamente correlacionados.

Como parte del propio preprocesado, se segmenta el espacio de soluciones según su eficiencia, seleccionando tanto el decil superior (10 % de menor Función objetivo) como el decil inferior, con el objetivo de comparar configuraciones altamente eficientes con aquellas de peor desempeño. Sobre el subconjunto superior se aplican,

de forma independiente, seis algoritmos de *clustering* no supervisado —*K-Means*, *Self-Organizing Map* (SOM), *DBSCAN*, *Spectral Clustering*, *Ward* y *Agglomerative-average*—. Cada técnica asigna etiquetas a las configuraciones y permite identificar el grupo cuya media de coste es más baja dentro de su propia partición.

El análisis detallado de estos clusteres —centrado en la frecuencia de activación de variables de inversión y en las combinaciones recurrentes de decisiones— permite extraer patrones robustos que se postulan como guías para una planificación energética eficiente. Además, la comparación de resultados entre algoritmos refuerza la consistencia de los hallazgos, permitiendo sintetizar un conjunto compacto de actuaciones prioritarias en los horizontes 2030, 2040 y 2050. Así, se proporciona una herramienta versátil, capaz de equilibrar eficiencia económica, robustez frente a la incertidumbre y alineación con los objetivos de transición energética.



4. Resultados

La aplicación conjunta de los algoritmos de clustering ha permitido aislar un subconjunto de configuraciones cuya media de coste es sensiblemente inferior a la del conjunto inicial. Entre los métodos evaluados, el enfoque jerárquico de enlace medio (*Agglomerative*) ofrece la agrupación más estable, lo que confirma su idoneidad para detectar familias de soluciones económicamente ventajosas. Dentro de estas familias se identifica un núcleo compacto de decisiones que se activa de forma recurrente —con independencia de la técnica empleada— y cuya presencia garantiza que el coste permanezca en la franja más baja.

El análisis de sinergias revela además que muchas de estas decisiones prioritarias se concentran en el horizonte de 2050, especialmente en proyectos vinculados a nuevas infraestructuras de gas. Esta clara inclinación hacia el largo plazo se complementa con una presencia moderada de actuaciones en 2040 y una intervención más puntual en 2030. No obstante, ciertas combinaciones de tramos iniciales en 2030 —particularmente aquellos con una función estructural temprana— junto con refuerzos potentes en 2050, muestran una sinergia destacada, optimizando el gasto sin añadir complejidad operativa.

En conjunto, los resultados apuntan a un modelo de expansión escalonada y estratégica, donde el grueso de la inversión se traslada hacia el final del horizonte, sin descuidar acciones clave a corto y medio plazo que preparan el terreno para una implantación robusta y eficiente del sistema energético futuro.

5. Conclusiones

El estudio tuvo como objetivo identificar un conjunto de soluciones quasi-óptimas en la planificación de la expansión de la red eléctrica (TEP) bajo incertidumbre. Mediante técnicas de *clustering* aplicadas a las soluciones más eficientes, se identificaron grupos con patrones de inversión consistentes.

Las inversiones relacionadas con el gas, especialmente aquellas planificadas a largo plazo para el año 2050, destacaron como decisiones recurrentes en múltiples métodos de agrupamiento. El análisis también reveló combinaciones robustas de inversiones que tienden a aparecer conjuntamente en configuraciones eficientes.

Estos resultados muestran que los métodos de *clustering* permiten descubrir patrones estratégicos de decisión dentro de espacios de soluciones complejos.

6. Referencias

[1] Ringkjøb, H.-K.; Haugan, P. M.; Sundseth, I. M. “A review of modelling tools for energy and electricity systems with large shares of variable renewables.” Renewable and Sustainable Energy Reviews, 2018.

[2] Valenzuela, D. S. A.; Latorre, L. A. “A data mining approach for classifying robust power system expansion plans under deep uncertainties.” Applied Energy, 2021.

[3] Sánchez-Úbeda, E. “Estadística II: Clustering (Lecture Notes).” Universidad Pontificia Comillas, 2020. [Online]. Available: https://www.comillas.edu/sites/default/files/clustering_estadistica_ii_sanchez_ubeda.pdf

[4] Cao, S.; Wang, B. “A Review of Evolving Challenges in Transmission Expansion Planning Problems.” IEEE Access, 2025.

[5] Perau, C.; Gläser, K.; Ruppert, M.; Fichtner, W. “Strategic Electrolyzer Placement for Green Hydrogen Production in Central Europe: A Comparative Analysis of RES-Driven and Demand-Driven Approaches Using the Synerginet Model.” 2023. [Online]. Available: https://ssrn.com/abstract=4628263

[6] Fariza, I.; Carcar, S.; Cruz Peña, J. “Iberdrola y Red Eléctrica se señalan mutuamente por el gran apagón.” Cinco Días, 29 Mayo 2025.

[7] Zhang, X.; Liu, Y.; Peng, L.; Guo, L. “Two-stage algorithm for efficient transmission expansion planning with renewable energy resources.” IET Renewable Power Generation.

[8] Han, J.; Kamber, M.; Pei, J. Data Mining: Concepts and Techniques, 3rd ed. Elsevier, 2011.

[9] García Serrano, A. Aprende Machine Learning: Programación avanzada en Python aplicada al análisis de datos, 2021.

[10] Zhang, Y. Z. A. G. D. X. “A structural k-means method for climate pattern clustering with uncertainty evaluation.” Geoscientific Model Development, 2023.

[11] Li, S. L. Y. Y. J. Z. L. J. H. “A CVaR-based integrated energy system planning considering uncertain renewable generation and energy prices.” Applied Energy, 2021.

[12] Li, C. W. Q. Y. H. L. J. “Integrated energy system planning under carbon trading mechanisms using a multi-objective CVaR-based approach.” Frontiers in Energy Research, 2024.

[13] Agrawal, R. I. T. S. A. “Mining Association Rules Between Sets of Items in Large Databases.” SIGMOD Record, 1993.

[14] Raschka, S. “MLxtend: Providing machine learning and data science utilities and extensions to Python’s scientific computing stack.” Journal of Open Source Software, 2018.

.

ANALYSIS OF QUASI-OPTIMAL SOLUTIONS WITH MACHINE LEARNING

Author: Martín Colino, Alberto.
Supervisor: Lumbreras Sancho, Sara.
Collaborating Entity: ICAI – Universidad Pontificia Comillas

ABSTRACT

This study addresses cost optimization in the expansion of the electrical transmission network by evaluating a set of investment options that vary in both type and time horizon (2030, 2040, and 2050). Following an initial data analysis, unsupervised learning techniques are applied to identify patterns and synergies among the different alternatives. The ultimate goal is to define a set of quasi-optimal solutions that ensures the future demand of the network is met while maximizing economic efficiency

Keywords: Transmission Expansion Planning, Unsupervised Clustering, Cost Minimization, Transmission Networks

1. **Introduction**

The massive penetration of renewable energy and the electrification of demand have significantly increased the variability of power flows in the transmission network. This new scenario forces system operators to expand and reinforce infrastructure using criteria of flexibility and resilience that go beyond the traditional deterministic approach. The problem is known as Transmission Expansion Planning (TEP) and involves deciding which lines, substations, and compensation devices should be built or upgraded to ensure supply security at the lowest possible cost. The analysis is conducted from a short-, medium-, and long-term perspective. In practice, TEP is currently addressed through optimization models that yield a single 'optimal' solution under a given set of assumptions. However, price volatility, technological evolution, and climate uncertainties expose the fragility of this unique response, as small changes in conditions can render the proposed plan obsolete.

2. Project definition

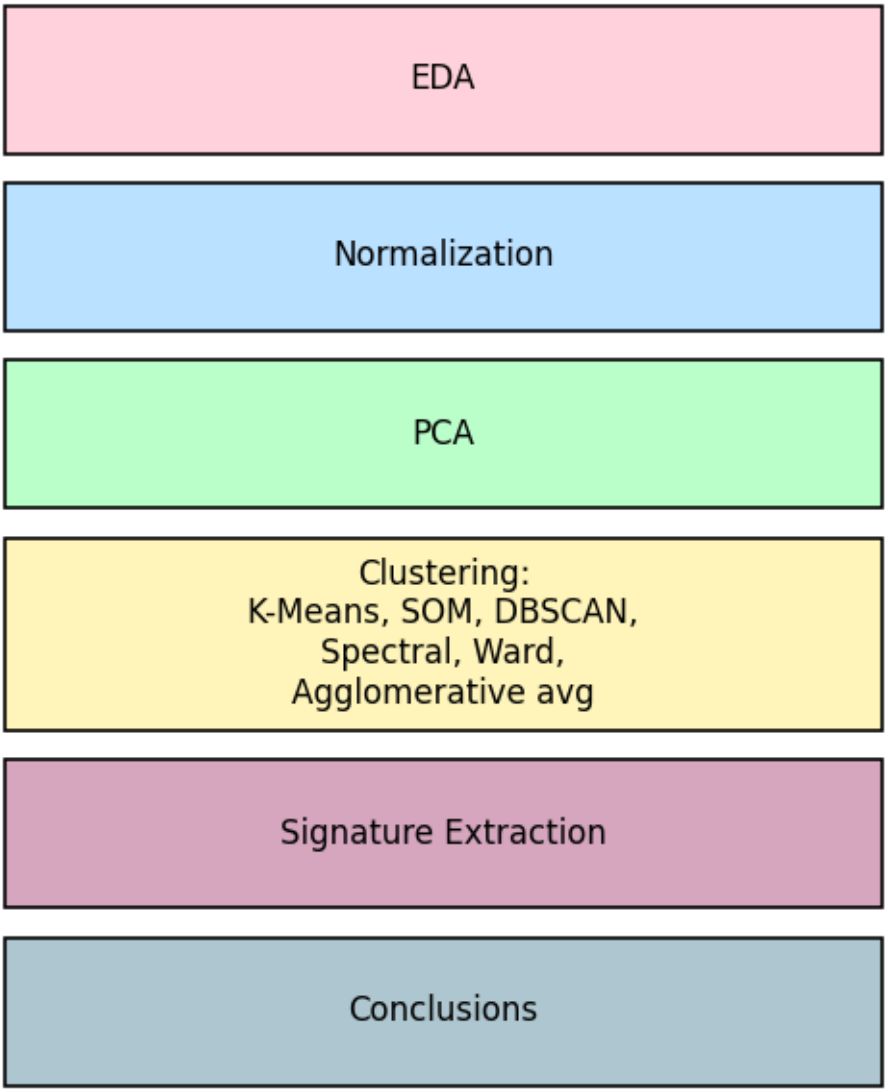
In order to network operators a range of equally competitive and, therefore, more robust alternatives, this work proposes an approach based on unsupervised learning. The process begins with a detailed study of a database of different investments in energy sources, with intensive preprocessing that seeks to simplify the set of available solutions and extract as much information as possible. This initial phase allows for reducing the dimensionality of the problem and improving the accuracy of the algorithms subsequently applied. From the large set of feasible solutions generated by the optimizer, the aim is to discover clusters of similar cost configurations and characterize their shared features. This is achieved by using dimensionality reduction techniques, which facilitate the interpretation of a highly complex variable space, and six clustering algorithms—both hierarchical and non-hierarchical—capable of revealing patterns without the need for external labels or assumptions about the shape of the groups. Once all the analyses are performed, the clusters with the best performance in terms of cost will be studied. This moves from a single "optimal" proposal to a repertoire of quasi-optimal configurations that maintain economic efficiency but add room for maneuver in the face of uncertainty.

3. Description of the model

The starting point is a thorough exploration of the full set of solutions provided by the Transmission Network Operator (TNO) and Dynamics Technologies Institute (DTI), entities that provide data related to power grid planning and optimization. In this preliminary phase, inconsistent records are refined, target cost distributions are analyzed, and the most technically relevant energy variables are identified. The cleaned data are normalized to correct for potential scaling biases and then projected using Principal Component Analysis (PCA) to reduce dimensionality, thereby mitigating the collinearity inherent in highly correlated techno-economic parameters.

As part of the preprocessing, the solution space is segmented according to efficiency, selecting both the top decile (10% with the lowest ObjFuncValue) and the bottom decile, with the goal of comparing highly efficient configurations with those of poorest performance. On the upper subset, seven unsupervised clustering algorithms are applied independently—K-Means, Self-Organizing Map (SOM), DBSCAN, Spectral Clustering, Ward, and Agglomerative-average. Each method assigns labels to the configurations and helps identify the group with the lowest average cost within its own partition.

A detailed analysis of these clusters—focused on the frequency of activated investment variables and recurrent combinations of decision make it possible to extract robust patterns that are proposed as guidelines for efficient energy planning. Moreover, comparing results across algorithms strengthens the consistency of the findings, enabling the synthesis of a compact set of priority actions for the 2030, 2040, and 2050 horizons. In this way, a versatile tool is provided, capable of balancing economic efficiency, robustness under uncertainty, and alignment with energy transition goals


EDA
Normalization
PCA
Clustering:
K-Means, SOM, DBSCAN,
Spectral, Ward,
Agglomerative avg
Signature Extraction
Conclusions

4. Results

The joint application of clustering algorithms has enabled the isolation of a subset of configurations with a significantly lower average cost compared to the initial solution set. Among the evaluated methods, the hierarchical average linkage approach (Agglomerative) provides the most stable grouping, confirming its suitability for identifying economically advantageous families of solutions. Within these families, a compact core of decisions is identified that is recurrently activated—regardless of the method used—and whose presence ensures that the cost remains in the lowest range.

The synergy analysis further reveals that many of these priority decisions are concentrated in the 2050 horizon, particularly in projects related to new gas infrastructure. This clear inclination towards the long term is complemented by a moderate presence of interventions in 2040 and a more punctual contribution in 2030. Nonetheless, certain combinations of early segments in 2030—especially those with an initial structural role—together with substantial reinforcements in 2050, exhibit notable synergies, optimizing expenditure without adding significant operational complexity.

Overall, the results point to a model of staggered and strategic expansion, in which the bulk of investment is deferred to the end of the planning horizon, while not neglecting key short- and medium-term actions that lay the groundwork for a robust and efficient deployment of the future energy system.

5. **Conclusions**

The study aimed to identify a set of quasi-optimal solutions for electricity grid expansion planning (EEP) under uncertainty. Clustering techniques applied to the most efficient solutions identified groups with consistent investment patterns.

Gas-related investments, especially those planned for the long term up to 2050, stood out as recurring decisions across multiple clustering methods. The analysis also revealed robust combinations of investments that tend to appear together in efficient configurations.

These results show that clustering methods enable the discovery of strategic decision patterns within complex solution spaces.

6. **Bibliography**

[1] Ringkjøb, H.-K.; Haugan, P. M.; Sundseth, I. M. “A review of modelling tools for energy and electricity systems with large shares of variable renewables.” Renewable and Sustainable Energy Reviews, 2018.

[2] Valenzuela, D. S. A.; Latorre, L. A. “A data mining approach for classifying robust power system expansion plans under deep uncertainties.” Applied Energy, 2021.

[3] Sánchez-Úbeda, E. “Estadística II: Clustering (Lecture Notes).” Universidad Pontificia Comillas, 2020. [Online]. Available: https://www.comillas.edu/sites/default/files/clustering_estadistica_ii_sanchez_ubeda.pdf

[4] Cao, S.; Wang, B. “A Review of Evolving Challenges in Transmission Expansion Planning Problems.” IEEE Access, 2025.

[5] Perau, C.; Gläser, K.; Ruppert, M.; Fichtner, W. “Strategic Electrolyzer Placement for Green Hydrogen Production in Central Europe: A Comparative Analysis of RES-Driven and Demand-Driven Approaches Using the Synerginet Model.” 2023. [Online]. Available: https://ssrn.com/abstract=4628263

[6] Fariza, I.; Carcar, S.; Cruz Peña, J. “Iberdrola y Red Eléctrica se señalan mutuamente por el gran apagón.” Cinco Días, 29 Mayo 2025.

[7] Zhang, X.; Liu, Y.; Peng, L.; Guo, L. “Two-stage algorithm for efficient transmission expansion planning with renewable energy resources.” IET Renewable Power Generation.

[8] Han, J.; Kamber, M.; Pei, J. Data Mining: Concepts and Techniques, 3rd ed. Elsevier, 2011.

[9] García Serrano, A. Aprende Machine Learning: Programación avanzada en Python aplicada al análisis de datos, 2021.

[10] Zhang, Y. Z. A. G. D. X. “A structural k-means method for climate pattern clustering with uncertainty evaluation.” Geoscientific Model Development, 2023.

[11] Li, S. L. Y. Y. J. Z. L. J. H. “A CVaR-based integrated energy system planning considering uncertain renewable generation and energy prices.” Applied Energy, 2021.

[12] Li, C. W. Q. Y. H. L. J. “Integrated energy system planning under carbon trading mechanisms using a multi-objective CVaR-based approach.” Frontiers in Energy Research, 2024.

[13] Agrawal, R. I. T. S. A. “Mining Association Rules Between Sets of Items in Large Databases.” SIGMOD Record, 1993.

[14] Raschka, S. “MLxtend: Providing machine learning and data science utilities and extensions to Python’s scientific computing stack.” Journal of Open Source Software, 2018..

Índice de la memoria

Índice de figuras

Índice de tablas

Capítulo 1. INTRODUCCIÓN

El objetivo principal de este Trabajo Fin de Grado es desarrollar una metodología alternativa para abordar el problema de la planificación de la expansión de redes eléctricas (***Transmission Expansion Planning*, TEP**), que permita ir más allá de la identificación de una única solución óptima. Frente al carácter determinista de los enfoques convencionales usados hasta ahora, se plantea un marco de análisis basado en técnicas y algoritmos de *Machine Learning* no supervisado, capaces de encontrar conjuntos quasi-óptimos de soluciones y hacer un análisis más profundo.

La finalidad de esta investigación es descubrir patrones robustos de decisión, ofrecer una representación estructurada del espacio de soluciones eficientes y facilitar así la toma de decisiones en contextos complicados y problemas de la vida real, caracterizados por la incertidumbre y la alta complejidad técnica. El trabajo combina un análisis exploratorio riguroso con técnicas de reducción de dimensionalidad, clustering multi-algoritmo y extracción de reglas explicativas, convirtiendo grandes volúmenes de datos en conocimiento operativo para agentes del sistema, tanto reguladores como inversores.

El uso de herramientas de *Machine Learning* permite, a partir de ese preprocesado, segmentar el conjunto de soluciones en grupos homogéneos, facilitando un análisis más preciso y enfocado [1]. Se ha puesto especial énfasis en el uso de varios modelos, de diversa índole, probando múltiples algoritmos de clustering —como *K-Means, DBSCAN*, *clustering* jerárquico o mapas autoorganizados (*SOM*)— y comparando sus resultados a través de métricas objetivas y visualizaciones interpretables. Esta aproximación permite observar si determinadas decisiones de inversión aparecen sistemáticamente en ciertos grupos de soluciones eficientes, y ayuda a construir una narrativa más rica que va más allá de la solución puntual más barata. En lugar de obtener una única respuesta, el modelo devuelve una visión estructurada del paisaje de soluciones posibles. Así pues, busca

conjuntos de soluciones que funcionen mejor cuando ambas estén activas, o incluso aquellas cuya presencia encarece el modelo y es preferible descartar.

Todo este proceso se ha implementado desde cero, transformando un conjunto denso y complejo de resultados de optimización en una herramienta analítica accesible, reproducible y escalable. [2] Se ha trabajado cuidadosamente en la trazabilidad de cada paso, documentando la lógica detrás de los filtros aplicados (media, variabilidad, correlación), la elección de los métodos de agrupación y la evaluación final de su rendimiento.

El objetivo no es solo analizar un caso concreto ni una red eléctrica en particular, sino construir una metodología replicable, útil para operadores del sistema, reguladores o agentes del sector interesados en mejorar la robustez y transparencia de sus decisiones. Este enfoque, que combina exploración estadística con aprendizaje automático, no solo permite extraer conclusiones más sólidas, sino que sienta las bases para una nueva forma de abordar los problemas de planificación en sistemas energéticos complejos y en constante evolución.

En definitiva, el presente trabajo combina la visión técnica con una vocación claramente aplicada a problemas reales. No se trata únicamente de modelizar, sino de explicar qué decisiones se repiten entre las soluciones eficientes, por qué y en qué contextos. Esta visión ampliada y segmentada del espacio de soluciones ofrece una ventaja estratégica y constituye un aporte real al campo de la planificación energética moderna.

Capítulo 2. DESCRIPCIÓN DE LAS TECNOLOGÍAS

2.1 HERRAMIENTAS Y TECNOLOGÍAS UTILIZADAS

El desarrollo completo del proyecto se ha llevado a cabo en lenguaje Python (versión 3.8.20), con un entorno de trabajo gestionado mediante Anaconda. La elección de Python frente a otros lenguajes responde a su versatilidad, su sintaxis clara y la amplia disponibilidad de bibliotecas especializadas en análisis de datos y *Machine Learning* [3], las cuales presentadas a continuación y enumeradas en la bibliografía del final del documento. A continuación, se hace una breve presentación de las herramientas empleadas, organizadas por su función principal dentro del flujo de trabajo.

2.1.1 HERRAMIENTAS PARA EXTRACCIÓN Y PROCESAMIENTO DE DATOS

Para la manipulación del conjunto de datos y el preprocesamiento intensivo, se han empleado fundamentalmente las bibliotecas **Pandas** y **NumPy**. Pandas ha permitido la carga, filtrado y tratamiento estructurado de los datos mediante estructuras tipo *DataFrame*, facilitando el manejo de tablas y el análisis estadístico de grandes volúmenes de datos, con medidas tan significativas como medias, desviaciones típicas o correlaciones. Por su parte, NumPy ha proporcionado soporte en cálculos numéricos más directos, especialmente durante operaciones de matriz y generación de máscaras lógicas para filtrado condicional.

La selección de variables más relevantes —clave para reducir la dimensionalidad del problema— se ha realizado mediante un análisis de correlaciones con la variable objetivo (*ObjFuncValue*), que representa el coste total de cada solución. Al buscarse la optimización, se aboga por aquellas soluciones con menor valor en esa variable objetivo. Tras el filtrado,

el número de columnas del conjunto de datos original se logra reducir hasta en un 95%, manteniendo únicamente aquellas variables con un alto valor informativo.

2.1.2 HERRAMIENTAS PARA LA APLICACIÓN DE ALGORITMOS DE MACHINE LEARNING

La segunda parte del proyecto, orientada fundamentalmente al descubrimiento de patrones y estrategias de decisión entre las soluciones quasi-óptimas, se ha basado en algoritmos de aprendizaje no supervisado, centrándose sobre todo en *clustering* y reducción de dimensionalidad. Estos algoritmos han sido implementados principalmente con la biblioteca **Scikit-learn (sklearn)** o **Minisom.** Se han contrastado diferentes métodos no supervisados, tanto jerárquicos como no jerárquicos:

- **KMeans** y **Spectral Clustering**, como métodos basados en la distancia y la conectividad de los datos.
- **Clustering jerárquico (*Ward*)**, útil para construir dendrogramas (Gráfico de árbol que se utiliza en este tipo de métodos y que sirve para ver paso a paso cómo se van agrupando los datos, permitiendo medir las distancias entre ellos) y analizar relaciones de pertenencia entre soluciones.
- ***DBSCAN***, empleado para detectar estructuras de densidad y soluciones atípicas.
- Además, para análisis adicionales, se ha utilizado **MiniSom**, una implementación ligera de mapas auto-organizados (SOM), útil para visualizar topologías de datos complejos y detectar similitudes en las decisiones estructurales.
- Finalmente, en la fase de interpretación y validación de resultados, se han empleado técnicas de aprendizaje supervisado mediante el modelo **XGBoost**, una implementación optimizada de gradient boosting ampliamente utilizada en problemas de regresión. Estas herramientas aportan una capa de interpretación cuantitativa adicional, permitiendo identificar qué variables estructurales están más asociadas con soluciones de bajo coste dentro del conjunto de combinaciones quasi-óptimas.

2.1.3 HERRAMIENTAS PARA VISUALIZACIÓN Y CONTRASTE DE RESULTADOS

Una parte esencial del análisis y de la extracción de conclusiones ha sido una visualización clara y comparativa de los resultados obtenidos. Para ello, se han utilizado las bibliotecas **Matplotlib** y **Seaborn**, que han facilitado la creación de histogramas, *boxplots*, gráficos de dispersión, mapas de calor de correlaciones y representaciones bidimensionales tras la aplicación de técnicas de reducción como **PCA** o **t-SNE**. Estas herramientas han sido fundamentales no solo para la comunicación de los resultados, sino también para la interpretación visual de la estructura interna de los datos y la evaluación de la calidad de los clusteres generados.

En resumen, la combinación de estas herramientas ha permitido abarcar todas las fases del proyecto: desde la selección y limpieza de los datos, pasando por la ejecución de técnicas avanzadas de *Machine Learning*, hasta la generación de conclusiones visuales que respaldan la interpretación del sistema eléctrico modelado.

Capítulo 3. ESTADO DE LA CUESTIÓN

El problema de ***Transmission Expansion Planning*** **(TEP)** ha sido ampliamente estudiado en la literatura científica como uno de los retos fundamentales en la planificación de sistemas eléctricos de potencia. Este proceso busca **determinar cómo y cuándo expandir una red de transmisión**, considerando la evolución esperada de la demanda, la capacidad instalada de generación y los costes asociados a nuevas infraestructuras. Las metodologías clásicas de optimización se han enfocado históricamente en encontrar una única solución óptima que minimice los costes, siempre y cuando cumpla las restricciones o *constraints* técnicas del sistema. Sin embargo, este enfoque presenta limitaciones evidentes en contextos con alta incertidumbre, como la integración creciente de energías renovables, la volatilidad de la demanda o la aparición de nuevas tecnologías energéticas.

En los últimos años, diversos estudios han abogado por enfoques más flexibles y resilientes, que consideren no solo una solución, sino un conjunto de alternativas viables **para dotar de mayor adaptabilidad al sistema eléctrico**. En esta línea, Cao y Bukhsh [4] realizan una revisión detallada de los desafíos actuales del TEP, destacando la importancia de incorporar incertidumbres y múltiples horizontes temporales en los modelos de expansión, así como la necesidad de herramientas analíticas más dinámicas para enfrentar la complejidad creciente de los sistemas eléctricos. Un ejemplo complementario es el trabajo de Perau et al. (2023) [5], centrado en la asignación estratégica de electrolizadores en Europa para la producción de hidrógeno verde. Esta investigación compara ubicaciones basadas en la proximidad a fuentes renovables frente a la demanda, y concluye que las decisiones de planificación basadas en generación pueden aumentar la eficiencia en la producción y reducir restricciones de red. Esta línea de trabajo ilustra cómo los modelos de optimización energética pueden enriquecerse incorporando múltiples criterios y escenarios, trascendiendo los marcos tradicionales de coste mínimo.

A partir de estas evidencias, emerge la necesidad de adoptar nuevas metodologías que permitan analizar y comparar múltiples soluciones quasi-óptimas desde diferentes dimensiones. Técnicas como el clustering, la minería de reglas de asociación o la reducción de dimensionalidad permiten no solo explorar el espacio de soluciones generadas por un modelo de optimización, sino también identificar patrones de comportamiento comunes que aporten valor en la toma de decisiones, así como extraer más informaciones de posibles sinergias entre diferentes variables. Este tipo de enfoques se posiciona como una vía prometedora para enfrentar los desafíos del TEP en un entorno energético cada vez más dinámico, híbrido y descentralizado; un futuro siempre cambiante, difícil de prever y de modelar.

Capítulo 4. DEFINICIÓN DEL TRABAJO

4.1 JUSTIFICACIÓN

La transición energética exige una planificación de redes capaz de adelantarse a escenarios de demanda volátiles, mixes de generación cada vez más renovables y restricciones medioambientales crecientes. El ***Transmission Expansion Planning* (TEP)** ha pasado de ser un problema de optimización lineal con hipótesis deterministas a convertirse en un reto multidimensional sujeto a incertidumbre tecnológica y de mercado.

El presente Trabajo Fin de Grado propone un **marco analítico basado en *Machine Learning* no supervisado** para explorar el espacio de soluciones quasi-óptimas que genera el modelo de TEP. Este enfoque aporta ventajas diferenciales frente a la búsqueda clásica de una única solución óptima:

- Permite descubrir **familias de configuraciones** cuyo coste total (*ObjFuncValue*) es competitivo.
- Identifica **patrones de decisión** que permanecen estables bajo variaciones de escenario.
- Facilita la **robustez operativa** ante *shocks* de demanda, precios o disponibilidad de recursos.

En un mercado donde cada megavatio de capacidad mal planificado puede traducirse en sobrecostes importantes o cuellos de botella regulatorios (Puntos de la red en los que, por falta de capacidad o por demoras, dificultan que la energía esté disponible a tiempo),, la posibilidad de ofrecer un repertorio compacto de alternativas igualmente eficientes resulta estratégica tanto para operadores del sistema como para inversores privados. Asimismo, los acontecimientos más recientes, como el famoso ‘apagón’ del 28 de abril de 2025, son una clara advertencia de los riesgos que puede suponer una planificación deficiente. [6]

4.1.1 BRECHA TECNOLÓGICA Y DE MERCADO

Los estudios de TEP publicados en la última década [7] siguen apoyándose, en su mayoría, en algoritmos deterministas o metaheurísticos que devuelvan **una única solución nominal**. Ejemplos de estos métodos tradicionales pueden ser *Simplex*, *Branch and Bound* o incluso *Dijkstra*, aplicados para resolver problemas de programación lineal. Sin embargo, la literatura reciente señala la necesidad de contemplar **múltiples óptimos locales** y evaluar su desempeño bajo escenarios estocásticos (variaciones de precios de CO_2, penetración de hidrógeno, etc.). Hasta la fecha:

- **No existe** una herramienta que integre clustering, minería de asociaciones y reducción de dimensionalidad para **mapear el conjunto de soluciones eficaces**.
- Los operadores TSO (Operadores de sistemas de transmisión, por sus siglas en inglés) se ven obligados a ejecutar **decenas de iteraciones manuales** para testar la robustez de su plan, con elevado coste computacional y riesgo de sesgo o de sobreajuste.
- El ecosistema académico-industrial carece de un **repositorio abierto** que combine resultados de optimización con técnicas de IA interpretables.

La propuesta de este proyecto intenta paliar esas carencias al articular un pipeline reproducible que agrupa, explica y visualiza **soluciones quasi-óptimas**, convirtiendo los datos brutos del optimizador en insumos estratégicos para la toma de decisiones.

4.1.2 RELEVANCIA CIENTÍFICA Y SOCIAL

1. **Innovación metodológica**: la aplicación conjunta de *K-Means*, *clustering* jerárquico, *Spectral Clustering* y *DBSCAN* sobre el decil más eficiente del espacio de soluciones ofrece un prisma inédito para detectar estrategias ocultas. Es un salto metodológico respecto a la aplicación de la programación lineal.

2. **Sostenibilidad**: al identificar combinaciones de inversiones que minimizan el coste del sistema manteniendo la seguridad de suministro, el proyecto contribuye a la integración acelerada de fuentes renovables y a la reducción de emisiones.
3. **Transferibilidad**: el pipeline, validado con datos de TNO y DTI, puede adaptarse a otros sistemas eléctricos o infraestructuras (gas/hidrógeno), potenciando su impacto industrial. Busca ser escalable para nuevos conjuntos de soluciones.

4.1.3 OBJETIVOS DEL PROYECTO

- **Reducción del espacio factible y foco en eficiencia**
  Aplicar un filtro inicial sobre el universo de soluciones TEP, seleccionando únicamente el 10 % de configuraciones con menor *ObjFuncValue*. Esta estrategia permite reducir el espacio factible del problema y concentrar el análisis en la franja de soluciones más competitivas desde el punto de vista del coste, facilitando el trabajo con volúmenes de datos muy grandes.
- **Segmentación multi-algoritmo y extracción de patrones eficientes**
  Ejecutar el modelo sobre este subconjunto depurado utilizando seis técnicas de *clustering* no supervisado (*K-Means*, *DBSCAN*, *Spectral Clustering*, *Ward* y *Agglomerative-average*), identificando para cada una el grupo con menor coste medio. Este paso permite observar diferencias estructurales en la toma de decisiones eficientes, según el enfoque algorítmico.
- **Anatomía de la decisión óptima y diferencias entre líneas de inversión**
  Analizar la frecuencia de activación de variables y pares de variables dentro de los clusteres eficientes. Se presta especial atención al comportamiento diferencial de cada tipo de inversión (gas, hidrógeno, electricidad…), destacando cuáles contribuyen de forma más sistemática a la reducción de costes. Esto permite construir una visión clara de qué decisiones resultan realmente imprescindibles y cuáles son prescindibles según el contexto.
- **Firma de consenso y validación estadística**
  Compilar una firma robusta a partir de las variables seleccionadas por al menos el

60 % de los métodos de clustering. Esta firma representa un núcleo de decisiones recurrentemente óptimas, y se valida estadísticamente mediante su capacidad predictiva sobre la función objetivo. Se prioriza así una interpretación fiable y generalizable del comportamiento óptimo.

- **Propuesta operativa por horizonte temporal (2030-2040-2050)**
  Traducir los resultados en recomendaciones estructuradas para operadores de red e inversores:
  • Inversiones esenciales a corto plazo (2030).
  • Decisiones condicionales y dependientes del entorno a medio plazo (2040).
  • Opciones estratégicas y de largo alcance a considerar en 2050.

Estos objetivos no solo responden a la necesidad de operar con grandes volúmenes de datos de forma eficiente, sino que garantizan una identificación robusta de decisiones clave, alineadas con los principios de sostenibilidad e innovación. El análisis, por tanto, contribuye de forma directa a los Objetivos de Desarrollo Sostenible (ODS), especialmente al **ODS 7** (energía asequible y no contaminante, más enfocado al contenido) y al **ODS 9** (industria, innovación e infraestructura, más enfocado a la metodología).

4.2 METODOLOGÍA

El trabajo explora el aprendizaje no supervisado como solución innovadora. Sin embargo, antes de aplicar cualquier técnica de análisis o segmentación, resulta clave realizar un preprocesado riguroso del conjunto de datos. Contar con una representación depurada y bien estructurada del espacio de soluciones permite reducir significativamente la complejidad del problema y aumentar la interpretabilidad de los resultados. Un “zoom” adecuado —basado en la eliminación de soluciones redundantes, la selección de variables relevantes y la normalización de escalas— no solo mejora el rendimiento de los algoritmos, sino que facilita

la extracción de conocimiento útil. Ignorar esta etapa supondría operar sobre un espacio sobredimensionado, ruidoso y poco informativo: muy complicado de manejar y de entender, comprometiendo así la utilidad del análisis posterior.

En este epígrafe se introducen brevemente los pasos que se van a realizar en el proyecto. El Capítulo 5 se dedicará a la explicación más en detalle de la base de datos, así como de su preprocesado y análisis exploratorio, mientras que el Capítulo 6 estará reservada a la explicación de los algoritmos de *Machine Learning* empleados, tanto desde un foco teórico como en relación a su aplicación en el trabajo. El Capítulo 7 unirá los resultados de ambos y extraerá conclusiones, así como una revisión más detallada de ciertos escenarios y soluciones.

4.2.1 ANÁLISIS EXPLORATORIO Y REDUCCIÓN DE VARIABLES

Se presenta el *dataset* y se hace una exploración inicial, permitiendo interpretar el significado de cada celda y sus valores. Tras ello, se hace un filtrado para reducir la dimensionalidad, basado en criterios como la variabilidad o la correlación con la función objetivo. Este trabajo se desglosa en los siguientes puntos:

- Normalización de todas las variables.
- Selección de soluciones con correlación significativa respecto a *ObjFuncValue*.
- Análisis por deciles con las soluciones ya filtradas: buscando aquellos escenarios con menor valor de la función objetivo.

4.2.2 CLUSTERING MULTI-ALGORITMO

Se explican y emplean seis métodos diferentes de aprendizaje no supervisado. Tras la ejecución de estas herramientas en Python, se trabaja con los clusters que den los mejores resultados.

- Métodos no jerárquicos (*K-Means, DBSCAN, Spectral Clustering*).
- Métodos jerárquicos (*Ward, Agglomerative average*).

- Cada algoritmo se aplica sobre el decil más barato; se selecciona el subgrupo de menor coste medio.
- Se compara el desempeño del clúster óptimo en cada uno de los algoritmos, discerniendo cuál es capaz de agrupar con mejor resultado.

4.2.3 EXTRACCIÓN DE PATRONES

A partir de los resultados obtenidos con 4.2.1 y 4.2.2, se busca profundizar y poder distinguir patrones y modelos subyacentes.

- Firma binaria (> 50 % activación) por método.
- Recuento de soluciones frecuentes en distintos clusteres
- Análisis de soluciones frecuentes por tipología y por horizonte temporal.

4.2.4 INSIGHTS CUANTITATIVOS

- Búsqueda exhaustiva de pares de variables cuya activación conjunta minimiza *ObjFuncValue*.
- Ranking de las 10 mejores combinaciones y visualización gráfica.

4.2.5 VALIDACIÓN Y CONCLUSIÓN

- Síntesis de resultados y recomendaciones.

Capítulo 5. ANÁLISIS EXPLORATORIO DE DATOS

El objetivo principal de este análisis exploratorio es **identificar qué decisiones de inversión energética están asociadas con menores costes dentro de una red eléctrica optimizada**. Debido a la alta dimensionalidad del conjunto de datos —con un número de columnas significativamente superior al de filas (iteraciones)—, el procesamiento y la interpretación directa de la información presentan una complejidad considerable tanto a nivel computacional como analítico.

Por ello, este apartado se centra en detallar el flujo de preprocesado de datos, así como en presentar los primeros resultados obtenidos a partir del análisis del conjunto completo de soluciones, que también serán relevantes de cara a las conclusiones. El enfoque inicial consiste en reducir la dimensionalidad del *dataset*, seleccionando únicamente aquellas variables que aporten valor informativo. Para ello, se aplican diferentes filtros basados en estadísticos como la media, la varianza y la correlación, lo que permite descartar atributos redundantes o irrelevantes.

Una vez reducido y depurado el conjunto de datos, se aborda el estudio de patrones y relaciones presentes en las soluciones más eficientes. En particular, se analiza el decil diez —conformado por el 10% de las soluciones con menor valor de la función objetivo—, comparando su comportamiento frente al resto de deciles y al conjunto global. Este análisis comparativo permite contextualizar el rendimiento del decil superior y extraer conclusiones más robustas sobre los factores que contribuyen a una mayor eficiencia energética.

El resultado de este análisis es una lógica clara de reducción y evaluación de datos que, además de facilitar la interpretación, constituye una herramienta robusta para realizar muestreos escalables y adaptables a distintos contextos de optimización energética.

5.1 DESCRIPCIÓN DEL CONJUNTO DE DATOS

Origen del dataset:

Los datos provienen de un modelo de optimización para la planificación de inversiones en fuentes de energía dentro de una red eléctrica diseñada para un país de Europa Central. El objetivo del modelo es minimizar el coste total de inversión, evaluando distintas configuraciones posibles.

Descripción general:

El dataset contiene múltiples columnas que representan decisiones de inversión en infraestructura energética para distintos horizontes temporales (2030, 2040, 2050), principalmente relacionadas con el gas. La columna **ObjFuncValue** representa el valor de la función objetivo a minimizar en cada iteración del modelo.

Variables incluidas:

Las variables siguen el formato:
Tipo de inversión + **Año** + **Identificador único**.
Ejemplo*: vGasInvestPotPipes_2030_430077* indica una inversión en infraestructura de gas en 2030 con identificador 430077.

Se incluyen, entre otras, las siguientes categorías de variables:

- ***vGasInvestPotPipes***: Inversiones potenciales en infraestructura de gas.
- ***vGasRepurposePipeH2***: Reconversión de gasoductos para transporte de hidrógeno.
- ***vGasRepurposePipeNG***: Reconversión de infraestructuras para gas natural licuado.
- ***vElInvestPotLinesAC***: Inversiones en líneas eléctricas de corriente alterna.

Estas variables permiten analizar distintas combinaciones temporales y tecnológicas para cubrir la demanda energética al menor coste posible. La Figura 1 analiza su distribución en el conjunto inicial de datos.

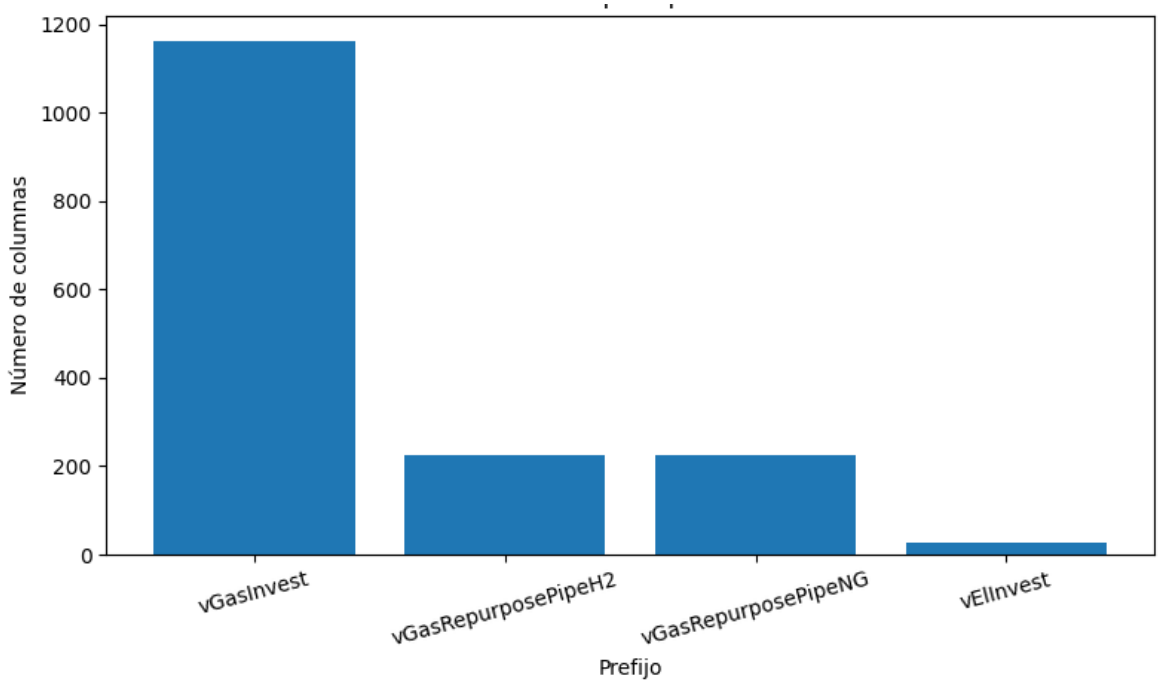


Figura 1 Distribución de soluciones de inversión en función de su tipología

Se puede ver en la Figura 1 un predominio de las soluciones enfocadas al gas, con tres posibilidades diferentes. Por otro lado, la suma de inversiones de corriente alterna apenas llega a treinta (un 2% del conjunto), siendo minoritaria en este *dataset*.

Interpretación de los valores:

Cada celda del dataset representa el grado de activación de una decisión en una configuración concreta. Aunque muchas variables toman valores cercanos a 0 o 1, no se trata de un conjunto estrictamente binario: los valores continuos provienen de una descomposición de Benders (La descomposición de Benders es un método matemático que está basado en la división en tareas de un problema complejo. Separa el TEP en un problema maestro de inversiones y subproblemas de operación. Itera proponiendo un plan, calcula su coste operativo y añade restricciones para descartar los más caros hasta converger en la mejor solución). En términos prácticos, un valor próximo a 1 indica que la inversión ha sido activada completamente en esa solución, mientras que un valor cercano a 0 representa su no activación o una contribución marginal.

A la hora de trabajar con el coste como función objetivo, es conveniente tener una imagen clara de su distribución en el *dataset,* así como de sus estadísticos principales, para poder abordar con mayor precisión fenómenos estadísticos como la presencia de outliers.

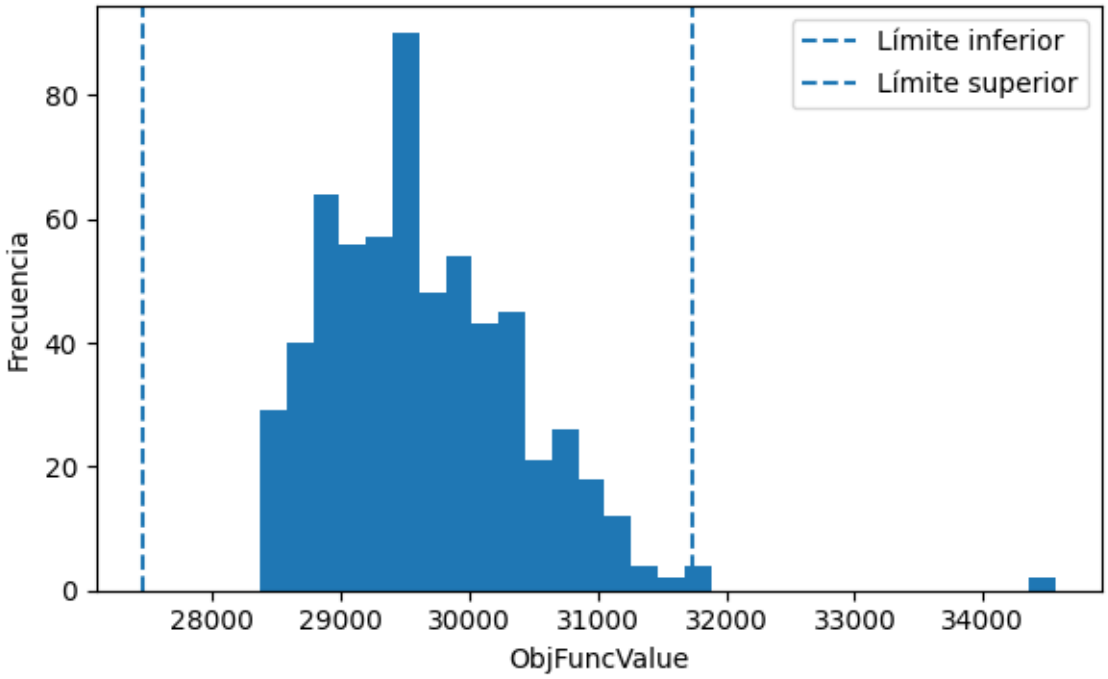


Figura 2 Distribución de función objetivo en el dataset

La Figura 2 apunta la existencia de varias soluciones con un valor de *ObjFuncValue* muy superior al límite. Más concretamente, existen dos variables con un valor de 34.568,95, más de 3000 unidades mayor que las siguientes soluciones más altas. Estas dos iteraciones no serán examinadas en este análisis.

Además, la media aritmética de *ObjFuncValue* está en torno a 29.800 unidades, con la mitad de las soluciones concentradas entre unos 29.100 y 30.500 (Esta medida se toma a partir del rango intercuartílico, que agrupa todas las soluciones entre el primer y el tercer cuartil). La media (≈29.850) y la mediana (≈29.800) coinciden estrechamente, pero la asimetría positiva (≈1,2) y una curtosis algo elevada (≈3,8) revelan una cola superior: de hecho, esto se confirma con las dos configuraciones extremas en ≈34.569 que quedan fuera del análisis para no distorsionar la interpretación. En conjunto, estos estadísticos confirman que, aunque

la mayoría de los planes de inversión mantienen costes muy similares, existen unos pocos casos atípicos que conviene descartar para un estudio más preciso

5.2 FILTRADO Y TRATAMIENTO DEL DATASET

5.2.1 ELIMINACIÓN DE VARIABLES INACTIVAS

Uno de los principales desafíos de este análisis es el **elevado número de variables** (más de 1600) frente a un número relativamente reducido de observaciones o iteraciones. Esta situación genera una alta dimensionalidad que, además de dificultar el análisis estadístico y visual, puede introducir ruido y aumentar el riesgo de overfitting o sobreajuste en eventuales modelos predictivos. Por ello, el primer paso del preprocesado se centra en una depuración estructurada del conjunto de datos, comenzando por la eliminación del modelo de aquellas variables que no aportan información significativa.

En esta fase inicial, se identifican como 'inactivas' todas aquellas variables cuya media a lo largo de todas las soluciones es exactamente cero. Este valor implica que, en ninguno de los escenarios analizados, dicha decisión fue seleccionada como parte de una solución viable. Desde el punto de vista estadístico, esta ausencia completa permite considerarlas como variables constantes, sin variabilidad alguna, lo que las convierte en candidatas naturales para ser descartadas. Al no contribuir a la discriminación entre soluciones, solo añaden complejidad sin aportar valor al análisis.

No obstante, conviene matizar que **el hecho de que una variable no haya aparecido en nuestro conjunto concreto no significa que carezca de relevancia en términos generales**. Su no utilización sistemática puede tener causas estructurales (por ejemplo, restricciones muy estrictas o ineficiencia evidente), pero también puede ser una consecuencia del tamaño limitado de nuestra muestra. Es posible que, en otras muestras con misma validez, estas variables tengan una activación media mucho mayor. Por eso, aunque estas variables se

filtren en esta etapa, es importante mantenerlas registradas. En futuros análisis con conjuntos más amplios o con diferente distribución de soluciones, podría resultar muy interesante contrastar los patrones actuales de ausencia con casos en los que dichas decisiones sí hayan sido seleccionadas, para así obtener una comprensión más robusta de su verdadero impacto.

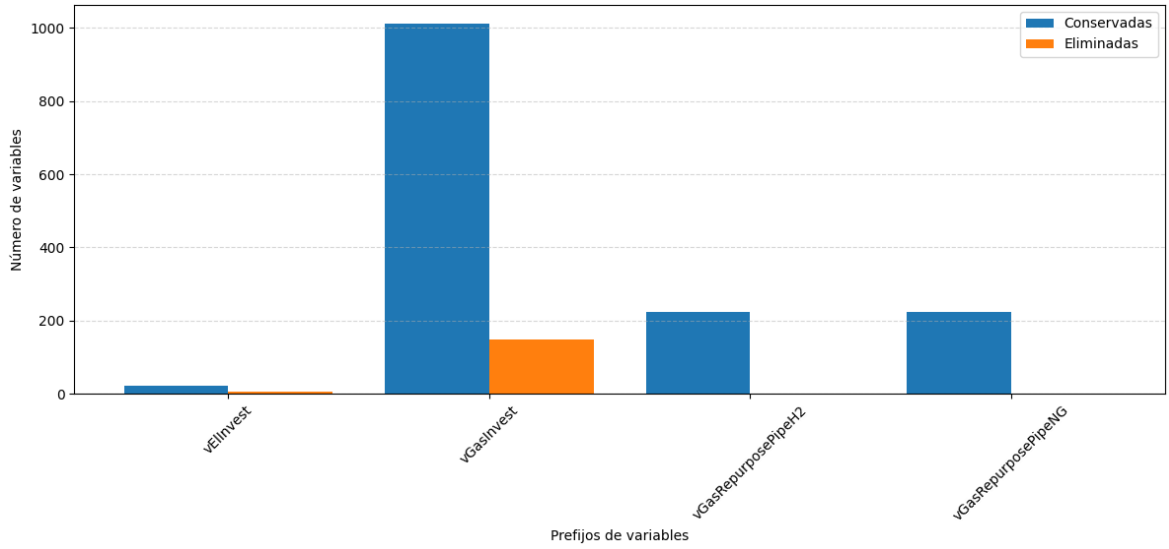


Figura 3 Evolución de número de variables tras el primer filtrado

La Figura 3 muestra el impacto del filtrado de variables inactivas por prefijo. Se aprecia que grupos como *vGasInvest* han sufrido una eliminación significativa, mientras que otros, como *vGasRepurposePipeH2* o *vGasRepurposePipeNG*, han conservado todas sus variables

---

5.2.2 FILTRADO POR BAJA VARIABILIDAD

Una vez realizado el primer paso de filtrado, se aplica un segundo criterio para descartar del análisis estadístico aquellas variables con muy poca variación entre observaciones. Se trata de columnas cuya desviación estándar es inferior a un umbral predefinido, lo cual indica que su comportamiento es prácticamente constante a lo largo de todas las iteraciones

estudiadas, tal como ya se ha indicado con aquellas variables permanentemente inactivas contempladas en el epígrafe 5.2.1.

Estas variables, aunque activas, no son relevantes desde el punto de vista analítico: no permiten distinguir entre soluciones ni aportan valor a la explicación del coste total del sistema. Su eliminación mejora la eficiencia del análisis y evita que variables con escasa variabilidad dominen o distorsionen los resultados. Este paso deja el *dataset* con 742 variables, menos de la mitad del conjunto original, pero con una riqueza informativa mucho más elevada.

Tal como se recoge en la lista adjunta mostrada en la Figura 4, muchas de las variables eliminadas en este paso presentan valores prácticamente constantes a lo largo de todas las soluciones evaluadas. Por ejemplo, variables como *vGasRepurposePipeNG_2030_30019* o *vGasRepurposePipeNG_2040_30062* muestran medias superiores a 0.96 y desviaciones estándar muy reducidas, indicando una activación sistemática en casi todos los escenarios. Otras, como *vGasInvestPotPipes_2030_430130* o *vGasInvestPotPipes_2040_431351*, apenas se activan, manteniéndose inactivas de forma constante. Desde el punto de vista estadístico, estas columnas no aportan variabilidad ni capacidad explicativa, por lo que se han descartado para mejorar la eficiencia computacional y facilitar un análisis más enfocado en las decisiones diferenciales del sistema.

No obstante, conviene subrayar que **la eliminación de estas variables no implica que carezcan de valor en términos de modelado energético**. Su activación sistemática (o su desuso absoluto) refleja decisiones robustas del modelo que se repiten en todos los casos óptimos considerados. Por tanto, aunque no contribuyan al análisis estadístico de sensibilidad o clustering, siguen formando parte esencial del conjunto de soluciones factibles y óptimas, y pueden ser relevantes en otros niveles de interpretación, planificación o implementación del sistema energético.

```
10 variables eliminadas con **menor media**:
vGasInvestPotPipes_2050_430757: media = 0.0013, std = 0.0357
vGasInvestPotPipes_2040_431224: media = 0.0022, std = 0.0298
vGasInvestPotPipes_2040_433006: media = 0.0025, std = 0.0139
vGasInvestPotPipes_2030_430098: media = 0.0025, std = 0.0504
vGasInvestPotPipes_2030_432235: media = 0.0027, std = 0.0439
vGasInvestPotPipes_2030_432580: media = 0.0027, std = 0.0130
vGasInvestPotPipes_2030_430568: media = 0.0041, std = 0.0568
vGasInvestPotPipes_2030_431213: media = 0.0045, std = 0.0648
vGasInvestPotPipes_2030_430753: media = 0.0047, std = 0.0533
vGasInvestPotPipes_2040_430088: media = 0.0064, std = 0.0796

10 variables eliminadas con **mayor media**:
vGasInvestPotPipes_2050_431714: media = 0.8906, std = 0.3124
vGasInvestPotPipes_2050_431832: media = 0.8906, std = 0.3124
vGasInvestPotPipes_2050_431186: media = 0.8915, std = 0.3086
vGasInvestPotPipes_2050_430285: media = 0.8928, std = 0.3090
vGasRepurposePipeNG_2030_30062: media = 0.9618, std = 0.1917
vGasRepurposePipeNG_2040_30062: media = 0.9618, std = 0.1917
vGasRepurposePipeNG_2050_30062: media = 0.9618, std = 0.1917
vGasRepurposePipeNG_2030_30019: media = 0.9707, std = 0.1686
vGasRepurposePipeNG_2040_30019: media = 0.9707, std = 0.1686
vGasRepurposePipeNG_2050_30019: media = 0.9707, std = 0.1686
```

Figura 4 Variables filtradas por baja variabilidad

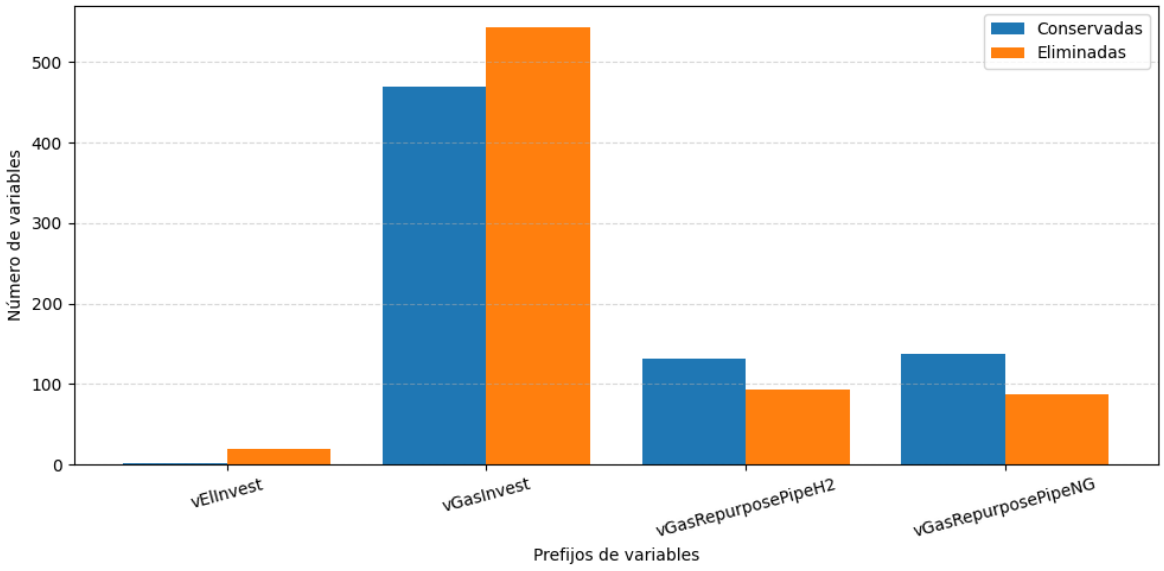


Figura 5 Evolución del número de variables tras el segundo filtrado

Se observa que, aunque *vGasInvest*, *vGasRepurposePipeH2* y *vGasRepurposePipeNG* conservan aún una parte significativa de sus variables, el volumen de columnas eliminadas es considerable. En el caso de *vElInvest*, prácticamente todas sus variables son descartadas en esta fase, lo que evidencia su escasa capacidad de diferenciación estadística

---

5.2.3 EVALUACIÓN DE LA RELEVANCIA ESTADÍSTICA DE LAS VARIABLES

Con el conjunto de datos ya depurado, se dispone ahora de una base con una dimensionalidad mucho más útil en este contexto de análisis. La última reducción se realizará en función de la correlación de cada solución con la función objetivo, buscando qué inversiones en concreto tienen mayor peso en la determinación del coste.

Para ello, se calcula la correlación lineal de cada variable con *ObjFuncValue*. Este análisis permite priorizar aquellas inversiones cuya presencia (o ausencia) está más ligada a reducciones o aumentos del coste total. Se ha elegido un criterio porcentual, buscando ese 10% de soluciones con mayor valor absoluto de correlación, no fijándose un umbral numérico concreto. Esta estrategia asegura que las variables más relevantes sean incluidas, independientemente de su valor exacto, y permite mantener una selección equilibrada entre variables positivas y negativas.

Este paso permite reducir el conjunto a solo 75 variables altamente informativas, es decir, un 4,58% del total inicial. A pesar de su reducido número, estas variables concentran una gran parte del valor explicativo del conjunto, lo que las convierte en candidatas ideales para el análisis en profundidad.

Además, se puede observar en la Figura 6 que la gran mayoría de variables relevantes tienen una correlación negativa con la función objetivo. En el contexto del trabajo, esto implica que hay muchas soluciones cuya mera presencia ya activación optimiza el coste, que es lo que se consigue al bajar la función objetivo.

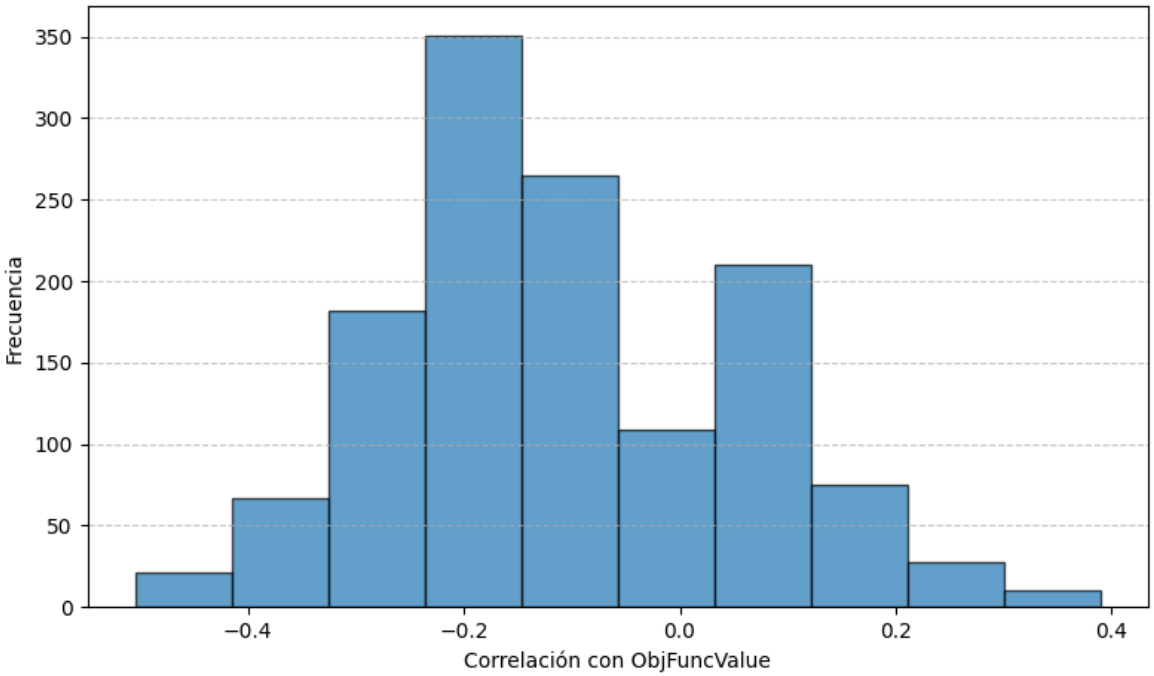


Figura 6 Correlación con la función objetivo del dataset original

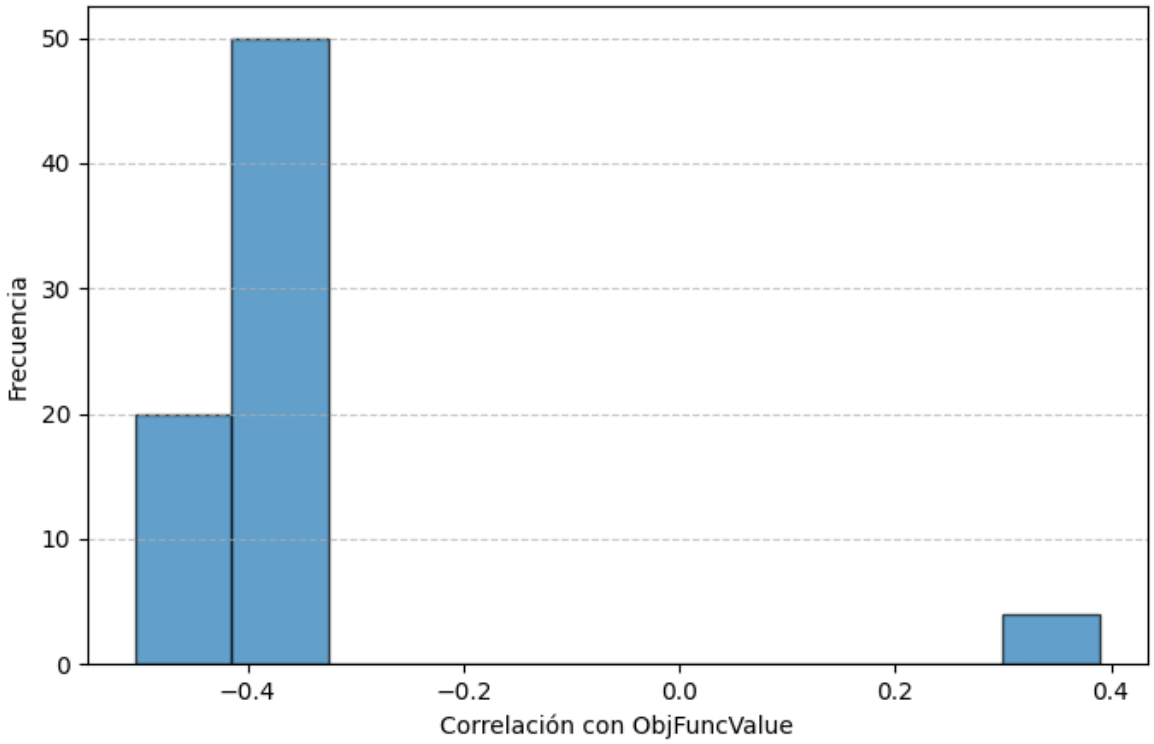


Figura 7 Correlación con la función objetivo del dataset tras el proceso de filtrado

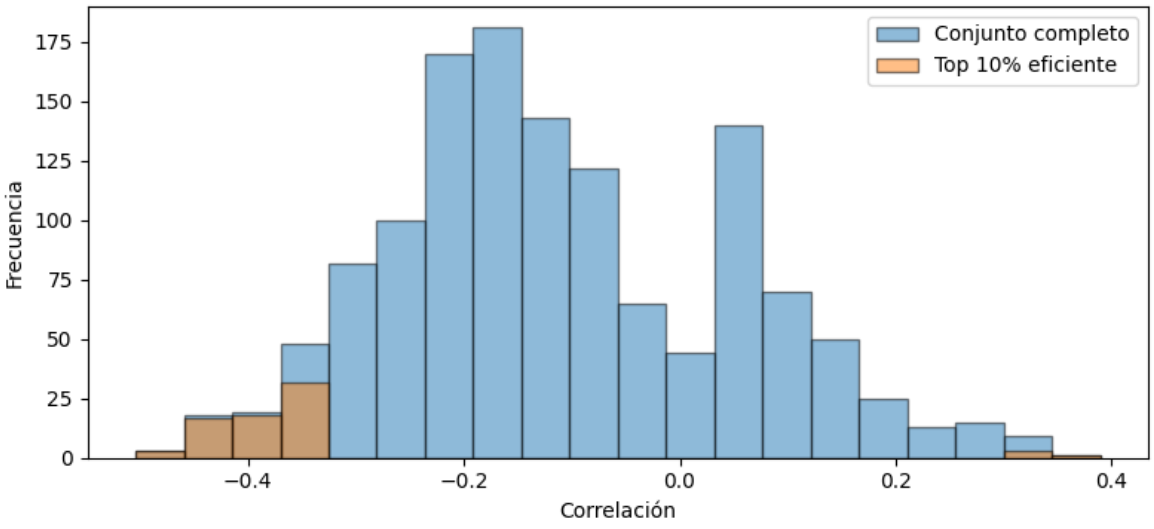


Figura 8 Comparación distribución de correlaciones antes y después del filtro

Al comparar el histograma de correlaciones del conjunto original (df_1) con el del conjunto final depurado (df_final) en la Figura 8, se aprecia una clara diferencia en la

distribución de valores. En el *dataset* original, las correlaciones están más dispersas y centradas en torno a cero, con una gran densidad de variables cuya relación con el valor de la función objetivo es prácticamente nula. Esta dispersión es esperable en un conjunto de datos masivo con muchas variables inactivas o redundantes, y refuerza la necesidad del proceso de depuración previo.

Por lo tanto, el histograma (Figura 7) generado tras aplicar los filtros de media, variabilidad y relevancia muestra una notable concentración de variables con correlaciones negativas más marcadas, indicando que las inversiones seleccionadas tienden, en su mayoría, a estar asociadas con reducciones en el coste total del sistema. Esta diferencia no solo valida el enfoque estadístico empleado, sino que también permite afinar el análisis, centrándose en aquellas decisiones de inversión que realmente contribuyen a una optimización de la función objetivo.

5.2.4 IMPACTO DE LA REDUCCIÓN DIMENSIONAL

La reducción progresiva de columnas —de 1639 a 75— ha seguido un enfoque sistemático orientado a eliminar ruido y mantener la esencia informativa del problema. Esta estrategia de filtrado y selección permite trabajar con un conjunto manejable y altamente expresivo, lo que sienta una base sólida para aplicar modelos más complejos como árboles de decisión, *clustering* o técnicas de regresión regularizada.

La Tabla 1 y la Figura 9 resumen el impacto de cada etapa en el número de variables:

Tabla 1 Evolución de número de columnas a lo largo del EDA

Paso	Descripción	Nº de columnas restantes	% respecto al total inicial
1	Número inicial de columnas del dataset	1639	100.00%
2	Tras eliminar columnas con media igual a 0	1484	90.54%
3	Tras eliminar columnas con baja variabilidad	742	45.27%
4	Tras seleccionar el 10% más correlacionado con ObjFuncValue	75	4.58%

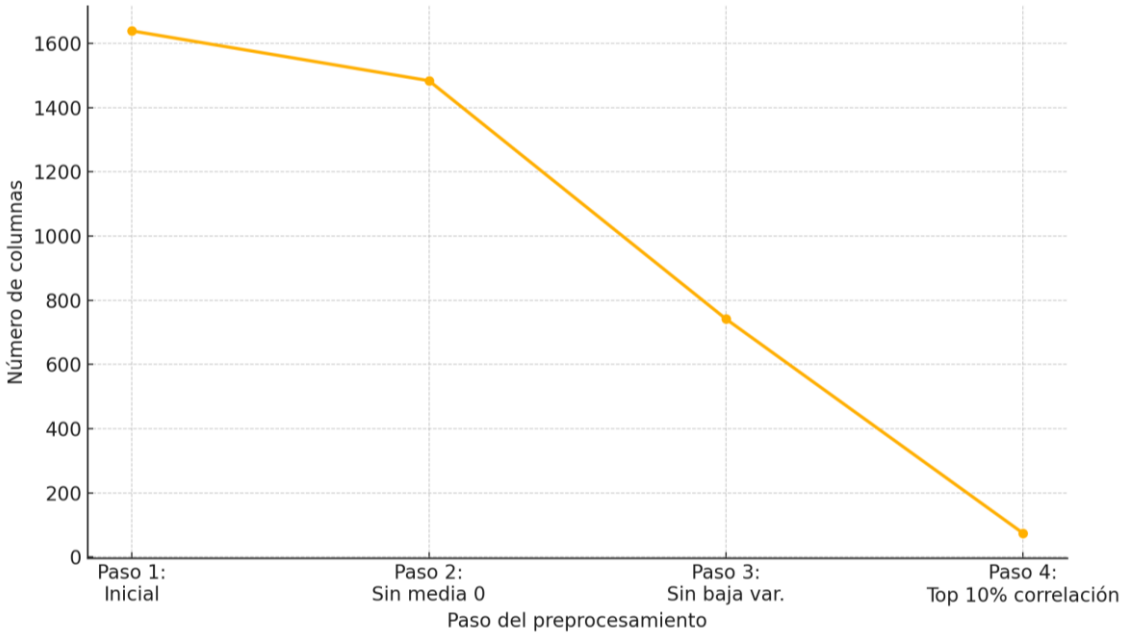


Figura 9 Evolución del número de columnas a lo largo del proceso de filtrado

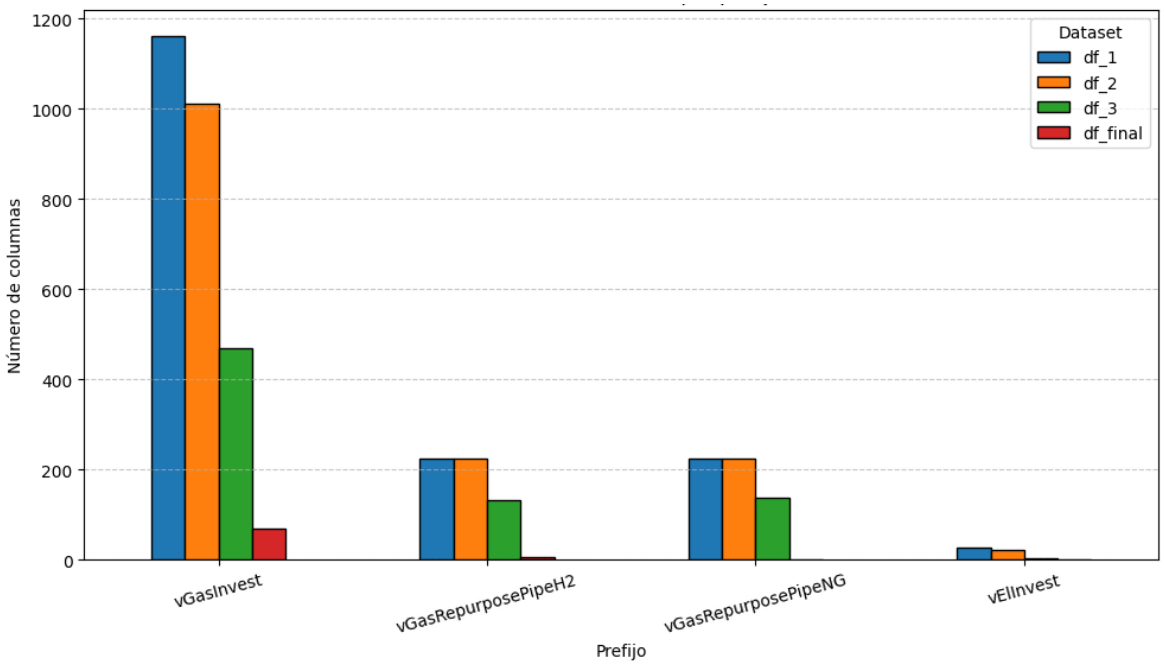


Figura 10 Evolución del número de columnas a lo largo del proceso de filtrado por tipología de inversión

El gráfico presente en la Figura 9 ilustra con claridad la evolución del número de variables correspondientes a diferentes tipos de infraestructura energética a lo largo del proceso de depuración de datos. En concreto, se observan los cambios en cuatro subconjuntos: df_1, df_2, df_3 y df_final, que representan las fases de filtrado por media, variabilidad y relevancia estadística, respectivamente.

El grupo de variables *vGasInvest* parte con un peso dominante en el conjunto original (df_1), acumulando 1161 columnas. Este número se reduce moderadamente en df_2 (1012 columnas), y cae de manera más acusada tras la eliminación de variables con baja variabilidad en df_3 (469 columnas). Finalmente, solo 68 columnas de esta categoría permanecen en el conjunto final, lo que indica una depuración selectiva que, aun conservando su relevancia, reduce drásticamente su presencia.

En el caso de *vGasRepurposePipeH2*, la reducción es también significativa: de 225 columnas iniciales se conservan solo 6 en el conjunto final. Lo mismo ocurre con *vGasRepurposePipeNG*, que pasa de 225 a 0, lo que sugiere una escasa capacidad

explicativa o redundancia estadística de estas inversiones en el contexto del modelo de coste.

Por su parte, *vElInvest*, que partía con una presencia testimonial (27 columnas), desaparece por completo en df_final. Esta categoría, aunque menor en número desde el inicio, no logra superar los filtros sucesivos. Casi todas sus variables se filtran en el segundo paso, lo que indica que poseen muy baja variabilidad.

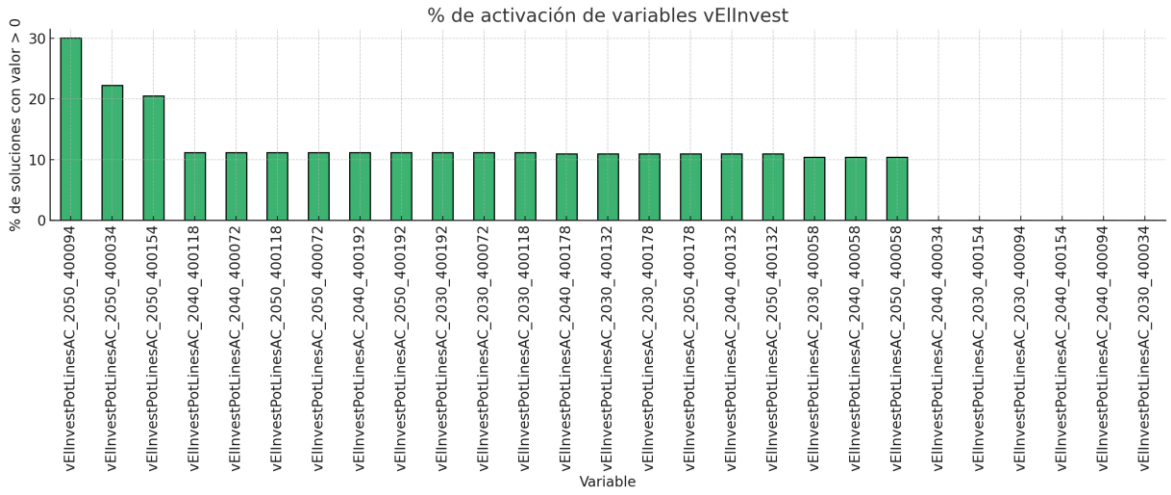


Figura 11 Porcentaje de activación de variables del tipo 'VElInvest'

Este patrón se refleja claramente en el gráfico de activación de la Figura 11: aunque algunas variables *vElInvest* superan puntualmente el 20% de uso, la mayoría apenas alcanza el 10% y varias no se activan nunca. Esta bajísima frecuencia de aparición implica una desviación casi nula, lo que impide estimar de forma fiable su correlación con la función objetivo y explica su exclusión en las etapas finales del filtrado.

En conjunto, el análisis muestra una clara concentración en aquellas variables más informativas, sacrificando volumen en favor de robustez estadística. Aunque se pierden muchos elementos, el conjunto resultante está formado por las variables que aportan una mayor capacidad explicativa del coste, sin perder de vista que las variables eliminadas no

carecen de interés técnico o estratégico en el diseño energético, sino que simplemente no resultan significativas en términos de análisis cuantitativo.

5.3 ANÁLISIS DE DECILES

5.3.1 FOCO EN LAS MEJORES SOLUCIONES

El foco de este análisis reside en buscar qué variables definen el mejor conjunto de soluciones posible. Por lo tanto, es conveniente hacer un análisis aislado de un conjunto de datos reducido que agrupe aquellas iteraciones con menor valor de la función objetivo *ObjFuncValue*.

Para ello, hemos dividido las columnas filtradas en diez subconjuntos acumulativos: el primero contiene el 10 % de las soluciones más óptimas (decil 10), el segundo el 20 %, el tercero el 30 % y así sucesivamente hasta abarcar el 100 %, que corresponde al conjunto completo. Además, hemos aislado de forma independiente el 10 % de las peores soluciones, aquellas con mayor valor de *ObjFuncValue*. De este modo, el decil con el mejor 10 % de resultados adquiere especial relevancia, ya que nos permite identificar las decisiones de inversión más frecuentes en las soluciones óptimas y compararlas con los patrones que emergen en los grupos de peor desempeño y en los conjuntos intermedios, así como contrastar con el *dataset* completo.

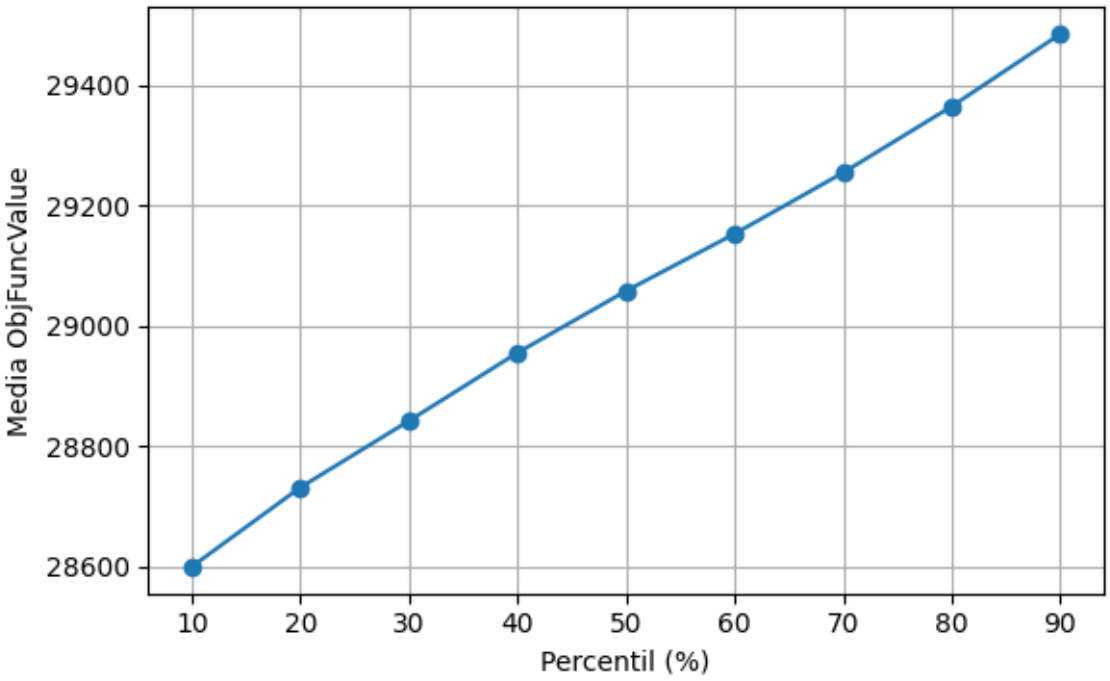


Figura 12 Media de la función objetivo en cada uno de los deciles

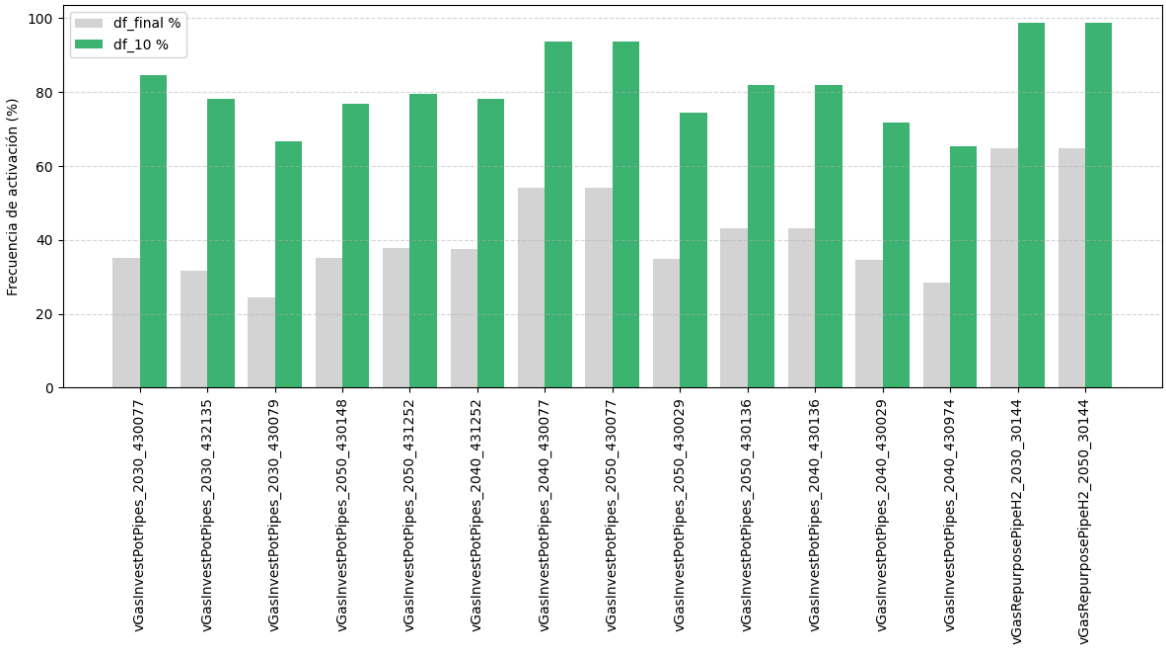


Figura 13 Comparación frecuencia de variables entre decil 10 y conjunto completo

5.3.2 COMPARACIÓN POR DECILES

Como se puede observar en la figura 13, se realiza un análisis estadístico por deciles, comparando la media de cada variable en el decil diez (mejores soluciones) con la media en el conjunto total. Esto ayudará a identificar si existen variables con una presencia mucho más frecuente en el espacio de soluciones óptimas. Para analizar el contraste, se incluye una columna que compara la frecuencia de activación en el decil diez con la frecuencia en el conjunto total, remarcando aquellas variables que son infrecuentes salvo en las buenas soluciones.

Además, se repite el análisis sobre el 10% de las peores soluciones (last_10), lo que permite contrastar las inversiones que aparecen en contextos de alto coste y comprobar si existen decisiones contraproducentes. Finalmente, se realiza una comparación directa entre el **top10** y el **last_10**, profundizando en las diferencias estructurales entre ambos grupos

Tabla 2 Comparación variables entre mejor y peor decil

Variable	All	Mejores 10	Peores 10	Diferencia
vGasInvestPotPipes_2030_432135	0.315522	0.822581	0.032258	0.507059
vGasInvestPotPipes_2030_430077	0.351145	0.854839	0.016129	0.503694
vGasInvestPotPipes_2030_430079	0.244275	0.709677	0.032258	0.465403
vGasInvestPotPipes_2050_430148	0.351145	0.806452	0.016129	0.455307
vGasInvestPotPipes_2050_431252	0.377863	0.822581	0.000000	0.444718
vGasInvestPotPipes_2040_431252	0.376590	0.806452	0.000000	0.429861
vGasInvestPotPipes_2050_430029	0.348601	0.774194	0.000000	0.425593
vGasInvestPotPipes_2050_430077	0.540712	0.951613	0.177419	0.410900
vGasInvestPotPipes_2040_430077	0.540712	0.951613	0.177419	0.410900
vGasInvestPotPipes_2040_430974	0.283715	0.693548	0.096774	0.409833

5.3.3 ANÁLISIS DE FRECUENCIA DE ACTIVACIÓN

No todas las variables se activan con valor 1. Muchas de ellas tienen valores intermedios, por lo que se analiza también su frecuencia de aparición dentro de los mejores y peores deciles. Este enfoque permite detectar variables que, aunque no estén totalmente activadas, tienen un papel recurrente en las soluciones eficientes.

Se observa, por ejemplo, que muchas variables del tipo *vGasInvestPotPipes* aparecen activas con mayor frecuencia en el top 10%, y además tienden a repetirse con distintos horizontes temporales. Esto sugiere que el año concreto (2030, 2040, 2050) puede no ser un factor determinante, sino que la decisión de invertir en una infraestructura concreta parece ser la clave, independientemente del horizonte temporal.

Los *boxplots (*Figuras 14,15 y 16) seleccionados ilustran el proceso de comparación de la activación de variables entre tres subconjuntos: conjunto completo, top 10 % y bottom 10 %. Estas gráficas evidencian de forma clara cómo ciertas variables presentan frecuencias de aparición muy superiores en las soluciones óptimas, aspecto clave para validar el filtrado por deciles y la identificación de patrones robustos de inversión.

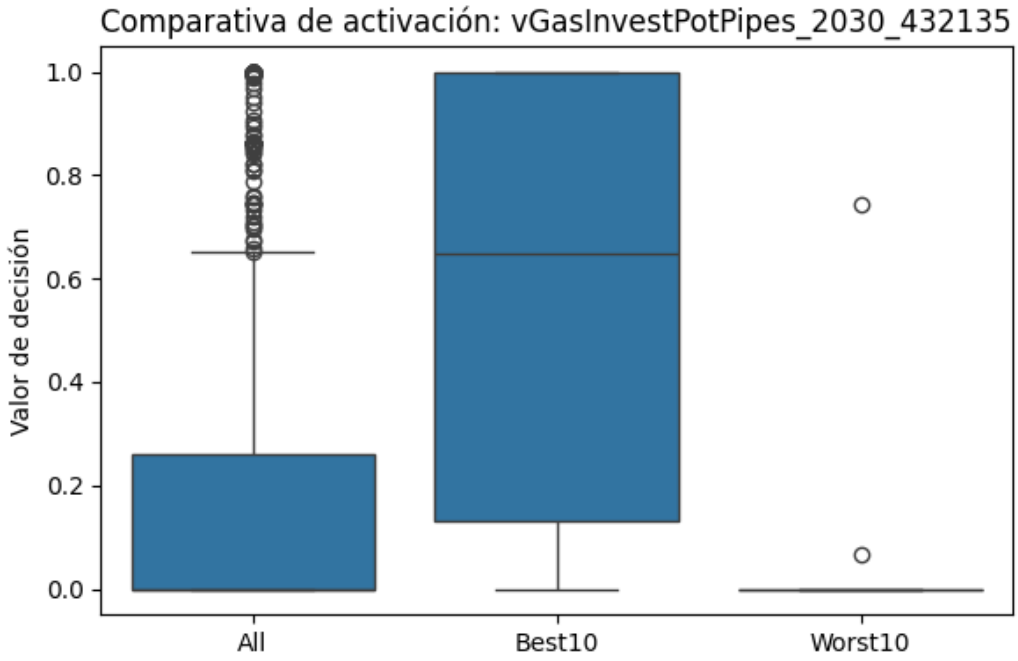


Figura 14 Primer ejemplo distribución variables en deciles

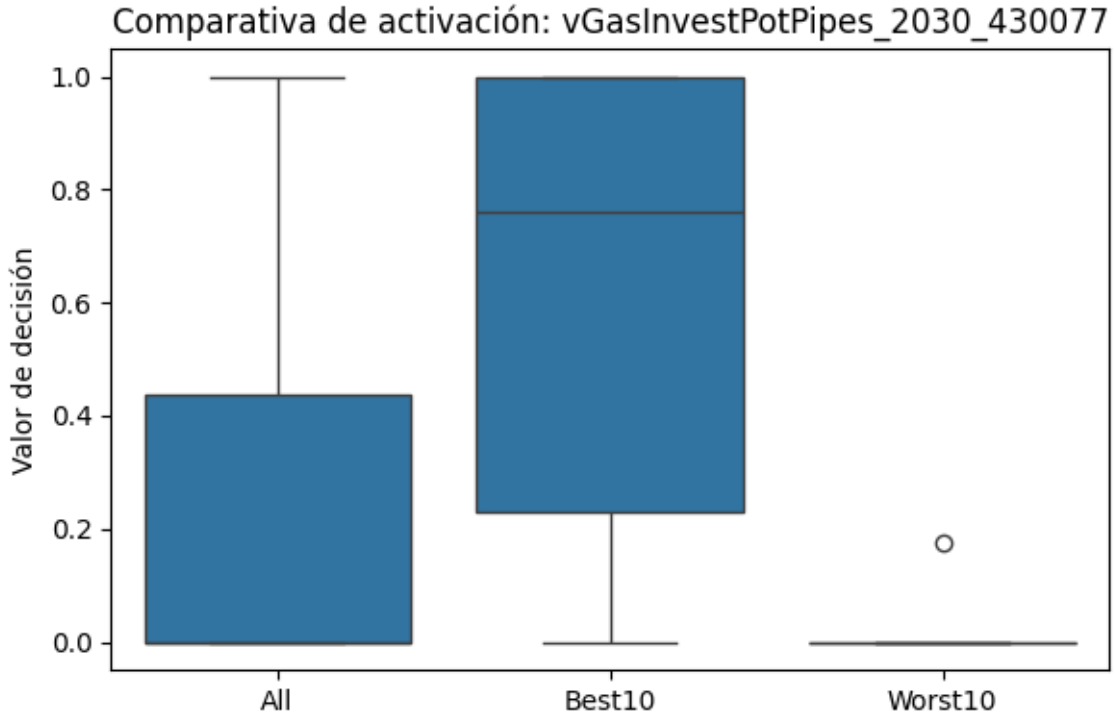


Figura 15 Segundo ejemplo distribución variables en deciles

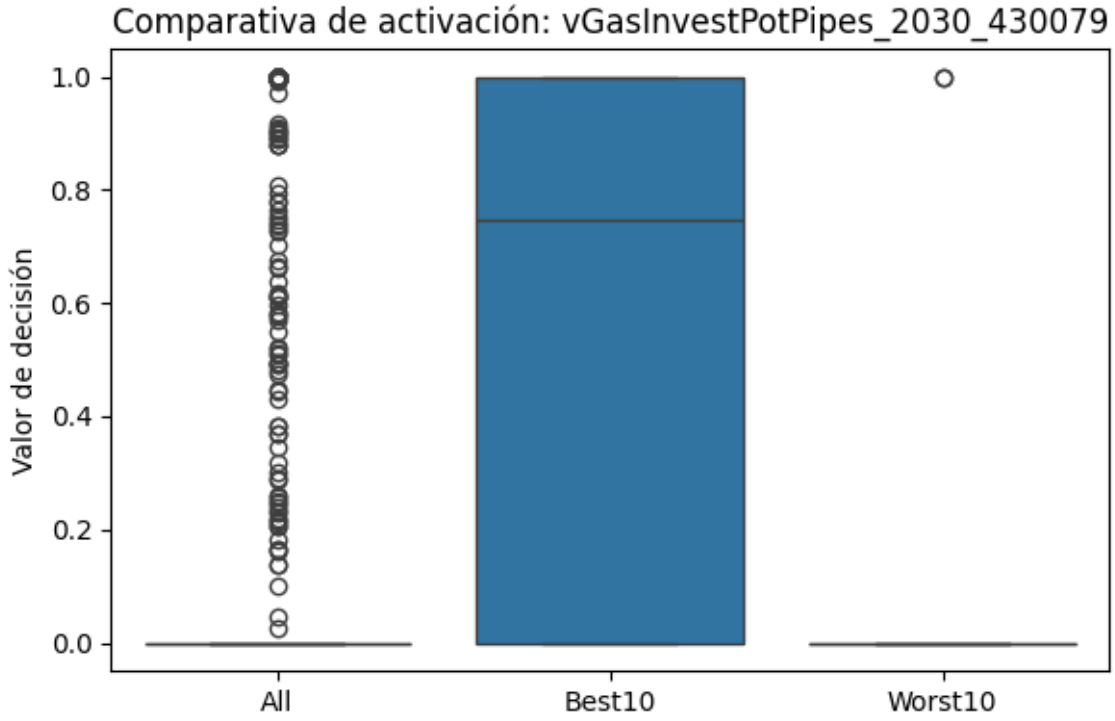


Figura 16 Tercer ejemplo distribución variables en deciles

5.4 ANÁLISIS DE CORRELACIONES

Con el subconjunto del top 10%, se calcula una matriz de correlaciones entre las 75 variables seleccionadas. El objetivo es descubrir si existen grupos de decisiones que tienden a activarse juntas, formando posibles **bloques estratégicos de inversión**.

El análisis revela correlaciones cercanas a 1 entre varias combinaciones de variables, muchas de ellas relacionadas con la misma infraestructura, pero en años distintos. Estas relaciones no se dan con la misma fuerza en el conjunto completo, lo que indica que son características exclusivas de las soluciones óptimas.

Por ejemplo, combinaciones como *vGasInvestPotPipes_2040_431502* y *vGasInvestPotPipes_2050_431513* presentan correlaciones significativamente mayores en

el top 10% que, en el conjunto general, lo que sugiere que su activación conjunta podría ser una estrategia eficaz en términos de coste total.

5.4.1 COMPARACIÓN DE CORRELACIONES: TOP 10% VS CONJUNTO COMPLETO

La siguiente tabla resume algunas de las diferencias más significativas observadas:

Estas diferencias reflejan que ciertas combinaciones de decisiones de inversión son mucho más comunes dentro de las configuraciones óptimas. Esto permite deducir que su implementación conjunta puede aportar un valor añadido real al sistema, garantizando un resultado más exitoso en esos casos.

Tabla 3 Comparación de Correlaciones

Variable 1	**Variable 2**	**Correlación Top 10 %**	**Correlación Conjunto Completo**
vGasInvestPotPipes_2050_430818	vGasInvestPotPipes_2050_431113	1.000	0.709
vGasInvestPotPipes_2040_430818	vGasInvestPotPipes_2050_431113	0.990	0.691
vGasInvestPotPipes_2050_431502	vGasInvestPotPipes_2050_431513	0.975	0.882

vGasInvestPotPipes_2040_431502	vGasInvestPotPipes_2040_431513	0.974	0.881
vGasInvestPotPipes_2050_431502	vGasInvestPotPipes_2040_431513	0.950	0.877
vGasInvestPotPipes_2050_431513	vGasInvestPotPipes_2040_431502	0.948	0.878
vGasInvestPotPipes_2040_430818	vGasInvestPotPipes_2040_431562	0.910	0.669
vGasInvestPotPipes_2040_430818	vGasInvestPotPipes_2050_431562	0.910	0.669

Capítulo 6. APLICACIÓN DE ALGORITMOS DE MACHINE LEARNING

Concluida la fase de análisis exploratorio, el trabajo se centra ahora en la identificación de patrones dentro del conjunto de soluciones quasi-óptimas del problema de *Transmission Expansion Planning* (TEP). Para tal fin se emplean algoritmos de aprendizaje **no supervisado**, pues los datos carecen de etiquetas y únicamente ofrecen el valor objetivo *ObjFuncValue*, métrica que se pretende minimizar. El propósito principal consiste en detectar agrupaciones coherentes de configuraciones de inversión, caracterizar las decisiones que las sustentan y generar conocimiento operativo para fases posteriores de planificación.

6.1 FUNDAMENTACIÓN DEL ENFOQUE NO SUPERVISADO

Los métodos no supervisados se consideran idóneos en este contexto por las siguientes razones:

- Permiten **descubrir estructuras latentes** sin imponer criterios exógenos de clasificación. Al no utilizar etiquetas, la agrupación no sigue un criterio fijo: encuentra resultados sin tener un objetivo claro marcado.
- Facilitan la **reducción de redundancias**, al agrupar soluciones con perfiles de decisión y parámetros similares.
- Ofrecen **combinaciones recurrentes** de variables que, de otro modo, permanecerían ocultas en un análisis univariado.

El proceso de agrupamiento se ejecuta sobre el **10 % de soluciones con menor *ObjFuncValue***, muestras que representan la franja de mayor interés económico. Todas las variables se normalizan y, para tareas de representación gráfica, se emplea la proyección PCA a dos componentes. [8] [9]

6.2 FAMILIAS DE ALGORITMOS Y SU APORTE ANALÍTICO

Los algoritmos seleccionados se clasifican en dos grandes categorías, cada una con un valor analítico diferenciado:

- **Métodos no jerárquicos**
  Los utilizados en este trabajo son: *K-Means, Self-Organizing Map (SOM), DBSCAN* y *Spectral Clustering*, los cuales operan con una partición única del espacio de soluciones. Esta aproximación resulta adecuada cuando se requiere determinar con rapidez el clúster que presenta el menor coste medio y, en consecuencia, extraer las variables decisorias predominantes.
- **Métodos jerárquicos**
  Ward y *Agglomerative-average* construyen una estructura de fusión progresiva que revela la secuencia de uniones entre soluciones. Dicho enfoque aporta información sobre la estabilidad de las agrupaciones y permite identificar el nivel de similitud a partir del cual la unión de ramas incrementa el coste de forma sustancial.

La utilización combinada de ambas familias se considera esencial para garantizar una visión completa del espacio de soluciones: los métodos no jerárquicos proporcionan eficiencia en la segmentación, mientras que los jerárquicos añaden una perspectiva evolutiva y de robustez.

6.3 DESCRIPCIÓN DE LOS ALGORITMOS APLICADOS

K-Means

K-Means es un algoritmo de agrupamiento (clustering) no supervisado cuya finalidad es particionar un conjunto de datos en K grupos (o clusteres) de modo que cada observación pertenezca al clúster cuyo centro (centroide) le quede más cercano.

Self-Organizing Map (SOM)

Mediante una red neuronal competitiva, el SOM proyecta el espacio de variables a una

cuadrícula 6 × 6, preservando la topología original. El análisis se concentra en los nodos con media de *ObjFuncValue* más baja, a fin de extraer firmas de activación representativas.

DBSCAN

Este método basado en densidad identifica regiones de alta concentración sin requerir la especificación previa del número de clusteres. Resulta especialmente eficaz para distinguir subconjuntos naturales de soluciones eficientes y separar puntos aislados considerados ruido.

Spectral Clustering

La técnica se apoya en la descomposición espectral de un grafo de similitudes y permite separar grupos incluso cuando las fronteras no siguen distribuciones lineales. Su aplicación busca revelar relaciones complejas entre variables decisorias.

Ward (jerárquico)

El enlace de Ward fusiona iterativamente los pares de clusteres que minimizan el incremento de varianza interna. El dendrograma resultante señala el sub-árbol cuyo coste medio es mínimo y las variables que cohesionan dicho grupo.

Agglomerative-average (jerárquico)

Con un criterio de distancia media entre clusteres, esta variante contrasta la estabilidad de las ramas detectadas por Ward y pone de relieve variables que solo adquieren relevancia bajo un umbral de fusión más laxo.

En cada caso, se determina el clúster –o nodo– con menor media de *ObjFuncValue*. A partir de dicho grupo se extraen:

1. Las variables con tasa de activación superior al 50 %.
2. Las combinaciones binarias de decisiones que registran los menores costes medios.
3. Una firma binaria exportable para simulaciones futuras.

6.4 SÍNTESIS ALGORITMOS

Tabla 4 Síntesis Metodología Algoritmos

Algoritmo	**Tipo**	**Finalidad principal**
K-Means	No jerárquico	Detectar perfiles nítidos de configuraciones de bajo coste.
SOM	No jerárquico	Representar la concentración espacial de soluciones eficientes.
DBSCAN	No jerárquico	Diferenciar grupos densos de outliers costosos.
Spectral Clustering	No jerárquico	Captar relaciones no lineales entre decisiones de inversión.
Ward	Jerárquico	Identificar el sub-árbol de fusiones de menor coste.
Agglomerative-avg	Jerárquico	Evaluar la robustez de las agrupaciones con enlace medio.

6.5 DESARROLLO Y EXPLICACIÓN DEL MÉTODO K-MEANS

Para ilustrar el procedimiento general de clustering se ha escogido un ejemplo concreto: se detalla el flujo de trabajo aplicado al método K-Means. Cabe señalar que, salvo detalles puntuales de cada método (por ejemplo, la construcción del dendrograma en métodos jerárquicos o el cálculo de la *U-Matrix* en *SOM*), la secuencia de pasos descrita es análoga al resto de algoritmos evaluados. Siempre se empieza por una normalización de los datos, se aplica el algoritmo y se examinan los distintos clusteres resultantes, poniendo el foco en aquel con el menor valor medio de *ObjFuncValue*.

6.5.1 NORMALIZACIÓN Y PROYECCIÓN PCA

En primer lugar, todas las variables decisorias se estandarizan para que posean media cero y desviación típica uno. Esta transformación elimina el sesgo de escala y garantiza que cada decisión contribuya de forma equitativa al cómputo de distancias. A continuación, se aplica un Análisis de Componentes Principales (PCA) para reducir la dimensionalidad a dos ejes, únicamente con fines de representación gráfica. El *scree plot* (gráfico de sedimentación) generado confirma que las dos primeras componentes acumulan más del 60 % de la varianza, lo cual resulta suficiente para una visualización intuitiva sin comprometer la mayor parte de la información original.

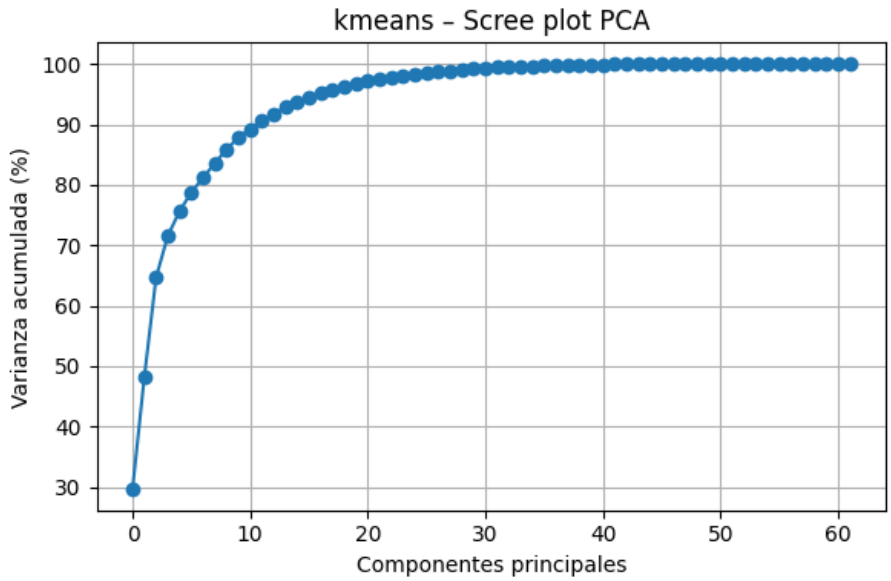


Figura 17 Varianza explicada acumulada por componentes principales

6.5.2 AJUSTE Y VALIDACIÓN DE LA PARTICIÓN

Se ejecuta *K-Means* fijando tres centroides, en línea con el objetivo de comparar resultados con los demás métodos en un escenario homogéneo. Una vez obtenidas las etiquetas, se grafica la dispersión de las observaciones sobre el plano PCA, coloreadas según su clúster. Esta vista permite comprobar a simple vista la cohesión interna de cada grupo. Para

cuantificar esta cohesión, se calcula la silueta de cada punto y su valor medio global; un histograma de silueta muestra si los grupos son compactos y están bien separados.

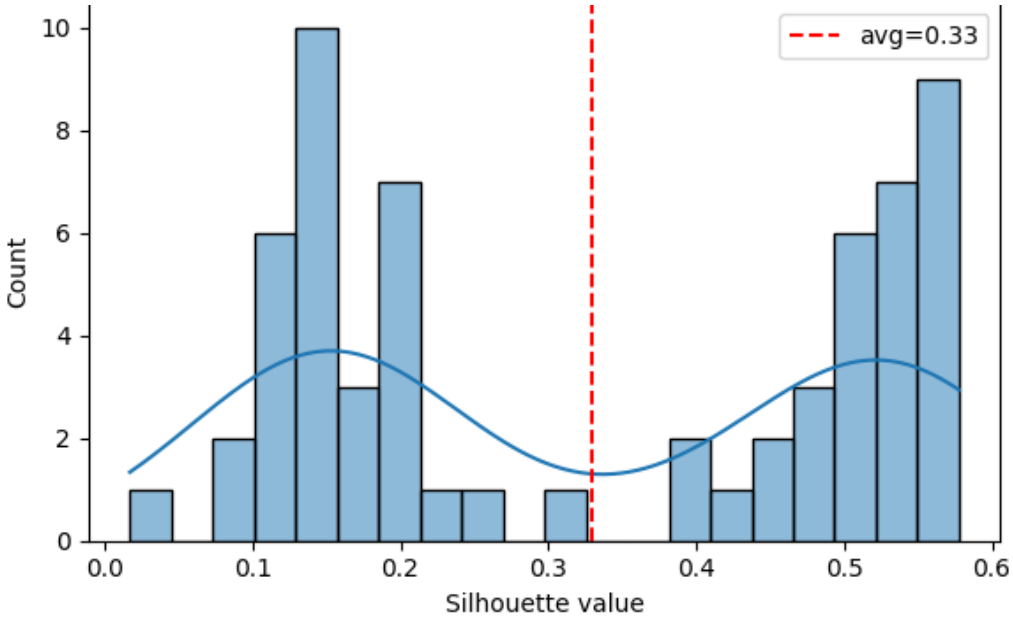


Figura 18 Número óptimo de clusteres utilizando Silhouette

6.5.3 EVALUACIÓN DE OBJFUNCVALUE POR CLÚSTER

En cada partición se analiza la distribución del coste objetivo. Mediante *boxplots* se comparan las medias y la dispersión de *ObjFuncValue* en cada uno de los tres clusteres en la Figura 20, lo que facilita identificar cuál grupo concentra las soluciones de menor coste medio. En *K-Means*, ese clúster óptimo se distingue claramente por su mediana más baja y rango más estrecho, validando su selección para el análisis posterior.

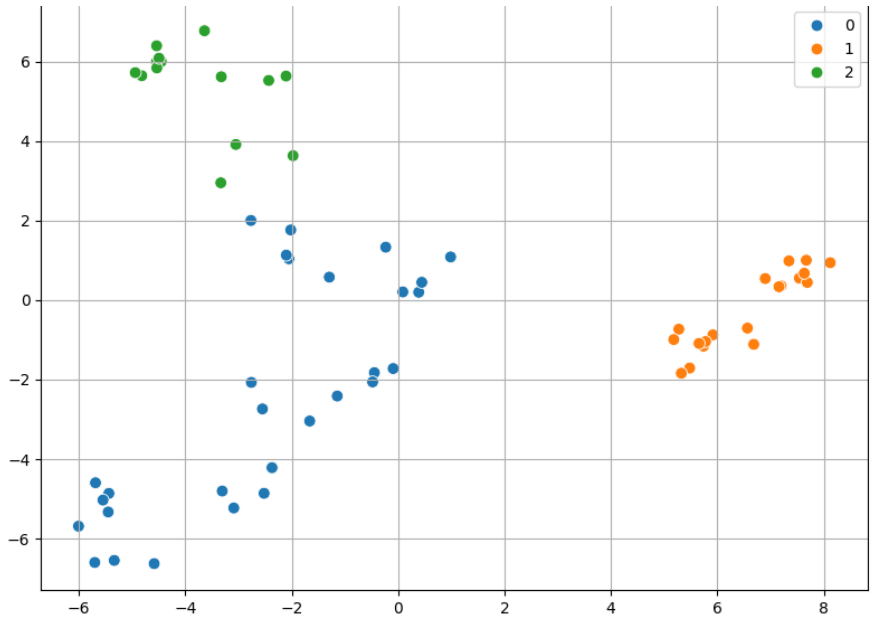


Figura 19 Distribución de los clusteres en KMeans

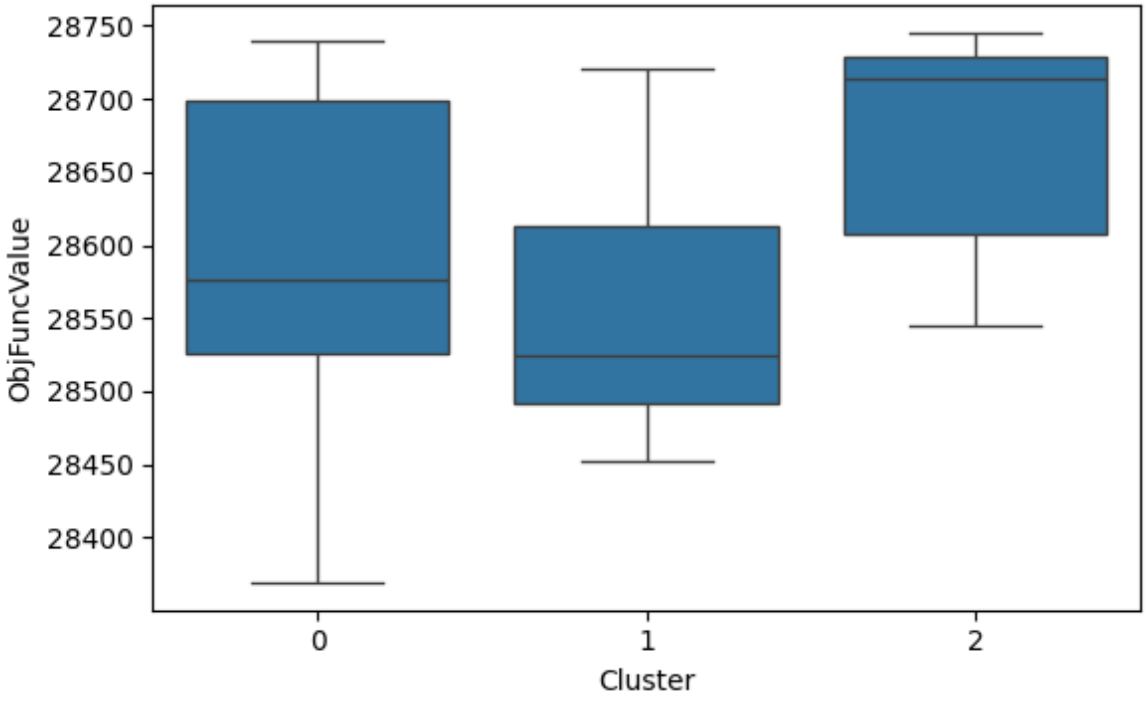


Figura 20 Media Función objetivo por clúster en KMeans

Tabla 5 Estadísticos por Clúster en K-Means

Cluster	Número de escenarios	Media ObjFuncValue	Std
0	37	28 635.74	124.50
1	22	28 587.45	116.29
2	19	28 700.94	80.88

6.5.4. PERFILADO DEL CLÚSTER ÓPTIMO Y GENERACIÓN DE FIRMAS

En el clúster óptimo se calcula la **tasa de activación** de cada variable (porcentaje de filas con valor > 0) y, a partir de ella, se crea una **firma binaria** que asigna 1 a las variables activas en más del 50 % de las soluciones y 0 en caso contrario. Con este nuevo umbral del 50 %, las inversiones en tuberías de gas (*vGasInvestPotPipes*) para los horizontes 2030, 2040 y 2050 aparecen sistemáticamente en más del 50 % de los casos, reflejando la necesidad recurrente de reforzar la red de gas en todas las etapas. En cambio, las variables de reconversión a hidrógeno (*vGasRepurposePipeH2*) quedan a 0, ya que no alcanzan ese nivel de frecuencia. Al agrupar la firma por prefijo y horizonte, se observa que la fracción promedio de variables activas de *vGasInvest* asciende de 0,55 en 2030 a 0,83 en 2040 y 0,85 en 2050, indicando un incremento progresivo de estas inversiones con el tiempo. Por el contrario, *vGasRepurposePipeH2* mantiene la variable *FirmaMedia* = 0 en todos los horizontes, lo que evidencia que dicha reconversión no es prioritaria en el conjunto de soluciones más eficientes. En conjunto, la firma binaria con umbral 50 % permite identificar de forma sencilla las decisiones de inversión esenciales y cómo evolucionan según el año y el tipo de infraestructura.

.

6.5.5. COMPARACIÓN CON OTRAS TÉCNICAS

Aunque cada método emplea un procedimiento similar —normalización, reducción por PCA, ajuste de clustering, identificación del clúster óptimo y perfilado de variables— sus resultados permiten evaluar la consistencia de signos: las variables recurrentes en *K-Means* suelen aparecer también en *SOM*, *DBSCAN* y las métricas jerárquicas (*Ward*, enlace promedio) aportan una visión evolutiva de las mismas. Esta convergencia entre diferentes familias de algoritmos confirma la solidez de los patrones descubiertos y sugiere un conjunto de decisiones prioritarias con alta confianza técnica.

En definitiva, el caso de *K-Means* ejemplifica un flujo de trabajo replicable: la combinación de PCA para visualización, métricas de silueta y boxplots de coste para validar la partición, proporciona un esquema claro para extraer conclusiones operativas, extensible al resto de métodos no supervisados.

6.6 RESULTADOS DE LOS DISTINTOS ALGORITMOS

Una vez ejecutados los seis algoritmos de *clustering* seleccionados sobre el subconjunto de soluciones más prometedoras (el 10 % de menor *ObjFuncValue*), se lleva a cabo un análisis comparativo de los clusteres óptimos detectados por cada método. Este análisis se centra en métricas clave como la media del valor objetivo (*ObjFuncValue*), el número de soluciones agrupadas, la distancia media intra-cluster y la media de variables activas por observación. Los resultados, resumidos en la Tabla 5, permiten identificar diferencias relevantes en el comportamiento de cada técnica y sus implicaciones prácticas dentro del contexto del TEP.

En términos de eficiencia económica, el método *Agglomerative Clustering* con enlace promedio destaca como el más eficaz, con una media de *ObjFuncValue* de 28.565, claramente inferior al resto de métodos con clusteres de tamaño representativo. Aunque el método *Self-Organizing Map* (*SOM*) obtiene un valor medio aún más bajo (28.438), su

clúster óptimo contiene solo 4 observaciones, lo que limita seriamente su relevancia estadística y capacidad de generalización. En este sentido, la solución de *Agglomerative* logra un equilibrio más robusto entre calidad y tamaño del clúster, agrupando 13 observaciones con comportamiento consistente.

El rendimiento de otros métodos como K-Means, Ward, y DBSCAN se sitúa en un rango similar, con medias en torno a 28.587, pero agrupando entre 19 y 22 observaciones, lo que sugiere una segmentación algo más amplia, aunque potencialmente menos precisa desde el punto de vista económico. El método Spectral Clustering, aunque logra el clúster más extenso (72 observaciones), presenta la media de coste más alta (28.630) y la mayor distancia intra-cluster con diferencia (17.47), como se aprecia claramente en la Figura 20, lo que indica una agrupación más difusa y heterogénea.

Un indicador complementario de gran interés es la media de variables activas por observación dentro del clúster óptimo (ver Figura 22). Aquí observamos cómo métodos como *K-Means*, *Ward* y *DBSCAN* tienden a identificar configuraciones de soluciones con un número elevado de decisiones activas (alrededor de 73), lo cual puede interpretarse como una señal de riqueza estructural o complejidad inversora. En cambio, *Agglomerative* se caracteriza por una activación media sensiblemente más baja (51.38), lo cual sugiere una firma más depurada y eficiente, donde un menor número de decisiones resulta suficiente para alcanzar soluciones de alta calidad. Esta diferencia refuerza el atractivo del método aglomerativo como herramienta explicativa para construir estrategias de planificación más concretas y racionalizadas.

Además, al observar la Tabla 6, que compara gráficamente los valores medios de *ObjFuncValue*, se visualiza claramente la ventaja relativa de *agglomerative* frente a los demás, situándose consistentemente en la parte baja del ranking. Por otro lado, la dispersión interna de cada clúster se ha medido mediante la distancia media intra-cluster, un indicador de cohesión interna que muestra valores estables y moderados (entre 4.0 y 5.1) en todos los métodos salvo en spectral. Esto refuerza la percepción de que el clúster de

agglomerative no solo es más económico, sino también más compacto y estructurado. Cuando se utiliza el adjetivo 'estructurado' para definir un clúster, se hace referencia a la baja distancia intra-cluster media que hay entre sus distintos componentes, la cual nos permite hablar de soluciones muy similares.

Por último, cabe señalar que la metodología seguida permite no solo evaluar el rendimiento global de cada técnica, sino también extraer las firmas binarias de activación de cada clúster óptimo y analizar la cohesión de los grupos tanto desde el punto de vista económico como estructural. Estas capas de análisis, combinadas con las métricas objetivas y visuales aportadas, permiten trazar un mapa robusto y reproducible del espacio de soluciones eficientes en TEP.

En conclusión, *Agglomerative Clustering* aparece como el método más equilibrado entre eficiencia, compacidad y claridad interpretativa, mientras que métodos como SOM o *spectral*, pese a su valor en términos exploratorios, ofrecen resultados más extremos o menos estructurados. Esta diversidad de resultados justifica la elección de un enfoque *multiclustering* y evidencia el valor añadido de aplicar aprendizaje no supervisado en problemas de planificación energética de alta dimensionalidad y complejidad.

Tabla 6 Clusteres óptimos de los distintos algoritmos

Método	**Nº observaciones**	**Media ObjFuncValue**	**Std ObjFuncValue**	**Media nº activos**	**Distancia intra**
Som	4	28438.74	50.52	54.75	4.07
agglomerative	13	28565.36	124.9	51.38	5.11
Dbscan	19	28577.06	117.95	72.79	4.81

Kmeans	22	28587.45	116.29	73.77	4.67
Ward	22	28587.45	116.29	73.77	4.67
Spectral	72	28630.18	119.65	55.07	17.47

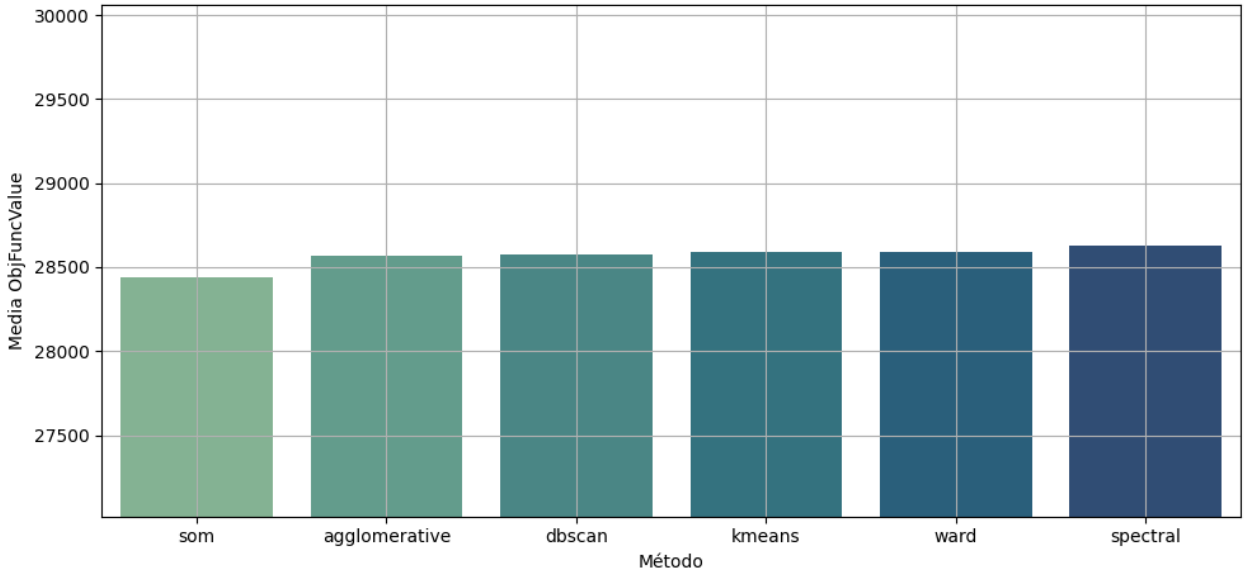


Figura 21 Media función objetivo en clúster óptimo por método

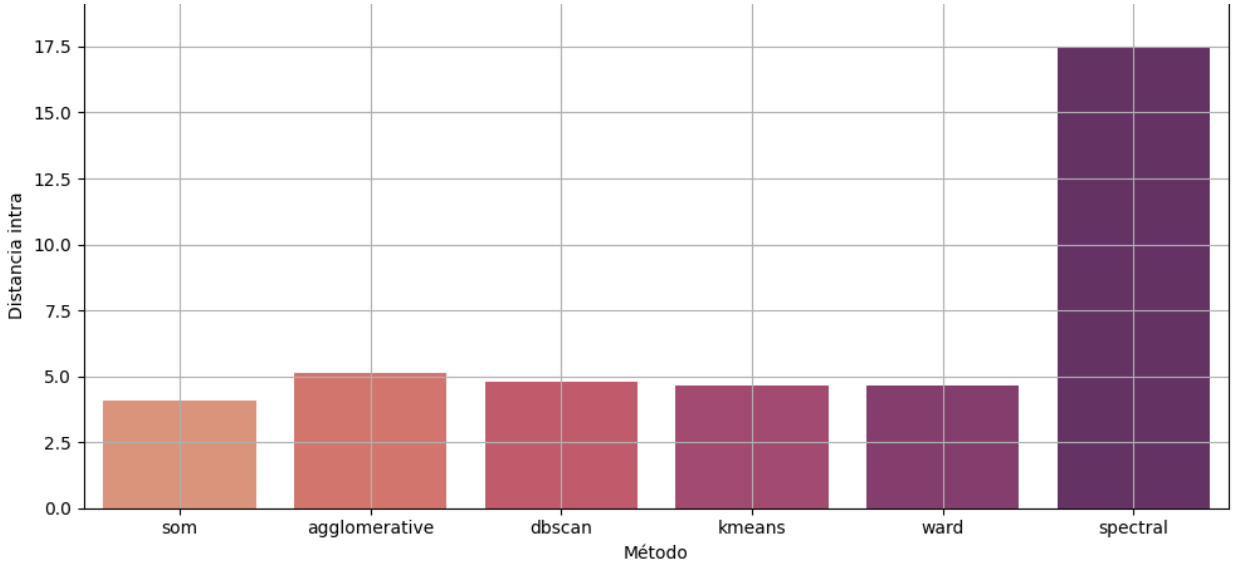


Figura 22 Distancia intra-cluster por método

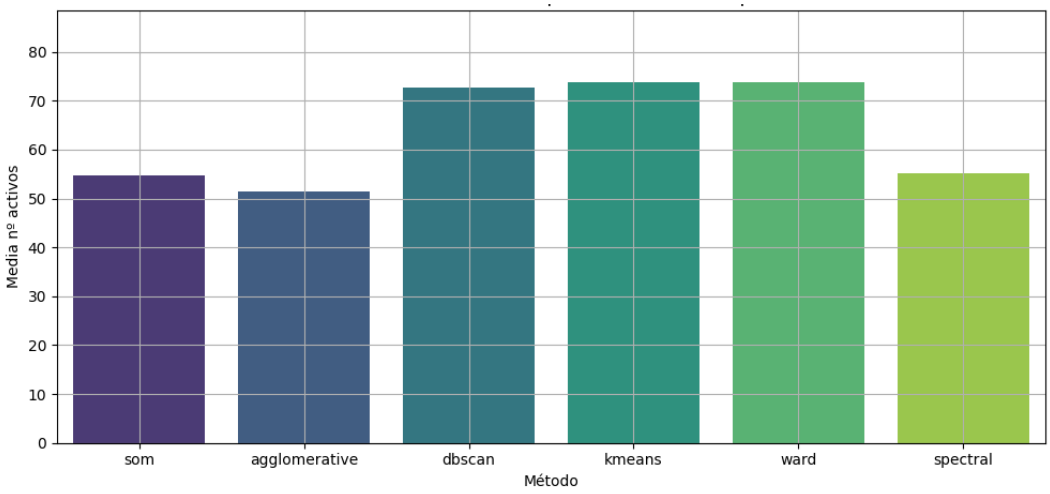


Figura 23 Media de variables activas por método

6.7 CLÚSTER ÓPTIMO: ESTUDIO EN DETALLE DEL MÉTODO AGGLOMERATIVE

El método *Agglomerative Clustering* ha demostrado ser el más eficaz de los probados, al devolver el clúster óptimo con la menor media de *ObjFuncValue*, tal como se observa en el

boxplot comparativo de los tres grupos. Esta eficiencia lo convierte en la referencia principal para extraer conclusiones.

Al analizar en detalle ese clúster óptimo, se observa una presencia sistemática de decisiones de inversión en gas, especialmente variables *vGasInvestPotPipes* y, en menor medida, *vGasRepurposePipeH2*. Como muestra la Figura 25, las decisiones más recurrentes se concentran en los horizontes de 2040 y 2050, lo que sugiere una preferencia estratégica por inversiones a largo plazo, compatibles con escenarios de transición energética. Esta firma de activación sintetiza las decisiones más robustas del modelo.

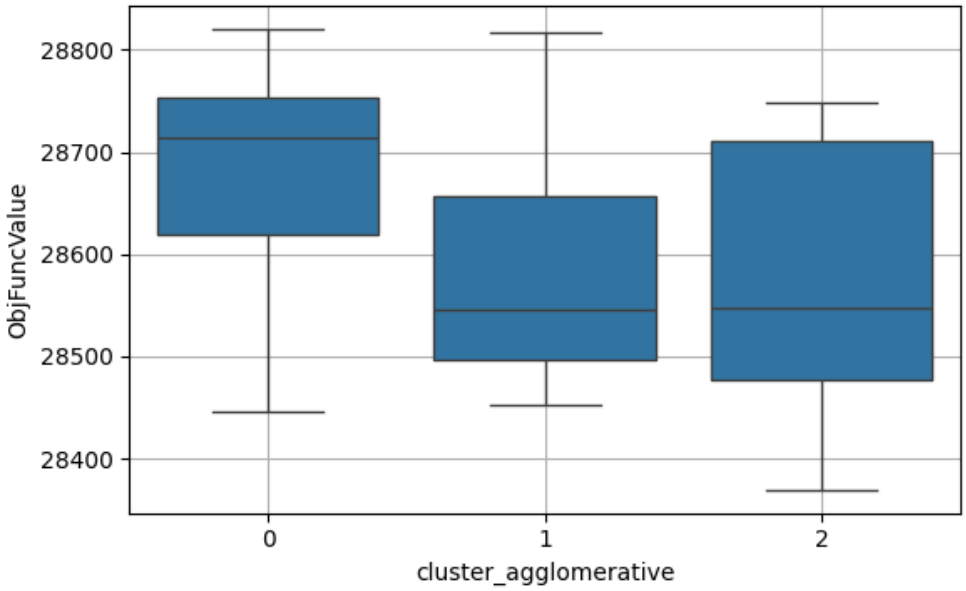


Figura 24Media de Función Objetivo en clusteres por método Agglomerative

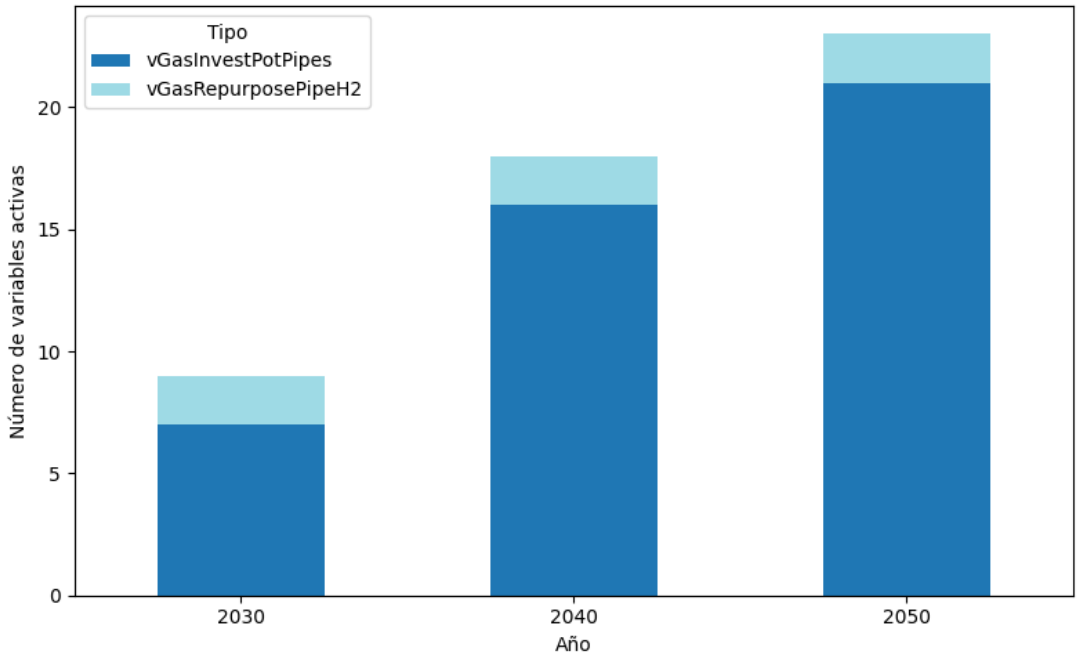


Figura 25 Distribución de variables en clúster óptimo por horizonte temporal y tipología

Capítulo 7. IDENTIFICACIÓN DE REGLAS FRECUENTES EN EL ESPACIO DE SOLUCIONES

El presente capítulo integra los resultados procedentes del análisis exploratorio (Capítulo 5) y de la aplicación de los algoritmos de clustering (Capítulo 6), con el objetivo de responder a los interrogantes planteados al inicio del trabajo y de ofrecer pautas operativas para la planificación de la red de transmisión en los horizontes 2030-2040-2050.

7.1 ALGORITMO CON MEJOR PODER DISCRIMINANTE

La evaluación comparativa entre métodos de clustering confirma la superioridad de *Agglomerative-average*, que obtiene la media de *ObjFuncValue* más baja (~28 565 u.m.) frente al resto de técnicas. Aunque la diferencia con otros enfoques como *Ward* no es de gran magnitud, la estabilidad de los resultados y la forma en que este método agrupa soluciones con decisiones coherentes y estructuradas refuerzan su valor. Al observar las variables predominantes en su clúster óptimo, se aprecia una clara tendencia hacia **inversiones en gas y en el horizonte temporal de 2050**, lo que sugiere que, incluso en fases preliminares del análisis, este método es capaz de anticipar las configuraciones de mayor relevancia técnica y económica. Esta capacidad para captar decisiones clave con poder explicativo convierte al algoritmo *Agglomerative-average* en el más eficaz del conjunto evaluado.

7.2 VARIABLES DE CONSENSO

Las conclusiones extraídas a partir del análisis multiclustering refuerzan la utilidad del enfoque basado en firmas binarias y activación de variables para la interpretación de soluciones quasi-óptimas en el problema de planificación de redes de gas. Para ello, se ha

generado una matriz de activación por método (*k-means, DBSCAN, Ward, Agglomerative y Spectral*), en la que se codifica si una variable concreta está activa (>50 % de frecuencia en el clúster óptimo) o no. Esta matriz permite visualizar la robustez de ciertas decisiones y diferenciar las más recurrentes frente a las específicas de cada técnica.

Con base en esta matriz, se ha calculado el número de métodos en los que cada variable está presente (es decir, activada en su mejor clúster). Este conteo ha permitido identificar las denominadas variables de consenso: aquellas activadas por al menos el 60 % de los métodos (es decir, en cuatro o más de los seis algoritmos analizados). Estas variables constituyen una especie de "núcleo duro" de decisiones robustas, que aparecen de forma sistemática en los subconjuntos más eficientes, independientemente del algoritmo de segmentación empleado. Muchas de ellas corresponden a inversiones en tuberías nuevas de gas en horizontes 2040 y 2050, lo que refuerza las tendencias observadas previamente.

En este sentido, conviene detenerse brevemente en las variables que componen los pares más eficientes. La mayoría hacen referencia a inversiones en nuevas tuberías de gas (*vGasInvestPotPipes*) y se distribuyen entre los horizontes 2030, 2040 y, especialmente, 2050. Este patrón temporal sugiere que los planes de expansión que retrasan parte de la inversión hacia el final del horizonte pueden beneficiarse de una mayor flexibilidad o de condiciones de red más favorables. Asimismo, ciertos identificadores como 430028, 430033, 430143 o 431900 se repiten con frecuencia, lo que indica que estas ubicaciones específicas podrían actuar como nodos críticos en la arquitectura de la red futura. En consecuencia, las variables asociadas a estos nodos y años podrían considerarse prioritarias para cualquier estrategia de desarrollo robusta.

Tabla 7 Combinaciones eficientes de soluciones

	Variable 1	Variable 2	Valor Medio	Recuento
1	vGasInvestPotPipes_2030_430079	vGasInvestPotPipes_2050_430143	28852.58	39
2	vGasInvestPotPipes_2040_430033	vGasInvestPotPipes_2050_430143	28898.51	90
3	vGasInvestPotPipes_2050_430033	vGasInvestPotPipes_2050_430143	28898.51	90
4	vGasInvestPotPipes_2040_430028	vGasInvestPotPipes_2050_430143	28938.21	107
5	vGasInvestPotPipes_2050_430146	vGasInvestPotPipes_2030_430079	28939.50	49
6	vGasInvestPotPipes_2040_430028	vGasInvestPotPipes_2050_430071	28942.87	98
7	vGasInvestPotPipes_2040_431967	vGasInvestPotPipes_2040_430028	28946.06	90
8	vGasInvestPotPipes_2050_431967	vGasInvestPotPipes_2040_430028	28946.06	90
9	vGasInvestPotPipes_2050_431900	vGasInvestPotPipes_2040_430033	28953.49	33
10	vGasInvestPotPipes_2050_431900	vGasInvestPotPipes_2050_430033	28953.49	33

La Tabla 7 recoge las combinaciones de pares de variables decisorias que, cuando se activan simultáneamente, producen las menores medias de *ObjFuncValue* dentro del conjunto de soluciones evaluado. Se trata de un análisis complementario al enfoque de clustering, centrado no en familias de soluciones, sino en relaciones binarias concretas entre inversiones.

Cada par incluye dos variables de decisión activadas (> 0), y la métrica considerada es la media del coste objetivo (*ObjFuncValue*) entre todas las soluciones que activan ambas simultáneamente. Las combinaciones con menor coste medio indican sinergias especialmente eficientes en la red de transmisión.

7.3 COMBINACIONES BINARIAS DE MENOR COSTE

Tras el filtrado progresivo aplicado al conjunto de soluciones, se ha pasado de analizar aproximadamente 1600 configuraciones válidas distribuidas en unos 750 escenarios iniciales, a concentrar el estudio en unas 60 soluciones dentro de un subconjunto reducido de 75 escenarios altamente eficientes. Esta depuración ha permitido centrar el análisis en un espacio de decisión más relevante, donde se encuentran las estrategias con mejor desempeño en términos de coste y viabilidad. Además, se han realizado investigaciones complementarias que, aunque fuera del pipeline principal, han permitido reforzar ciertas intuiciones, como la importancia estratégica de variables específicas como *vElInvest*, que pese a su baja frecuencia global, presentan una gran activación en configuraciones muy eficientes.

Después de realizar un análisis exploratorio detallado y seleccionar el decil más eficiente del conjunto de soluciones —aquel 10 % de configuraciones con menor *ObjFuncValue*—, se

procedió a aplicar técnicas de clustering no supervisado con el objetivo de segmentar las soluciones quasi-óptimas en grupos representativos. Esta fase de multiclustering permitió comparar distintos métodos (como *k-means*, *DBSCAN*, *Ward*, *Spectral* o *Agglomerative*) y, a partir de los clusteres óptimos de cada uno, identificar qué variables de decisión aparecen con mayor regularidad en los subconjuntos más eficientes.

El análisis de activación cruzada entre métodos reveló la existencia de 70 variables que aparecen activas en más del 60 % de los clusteres óptimos, lo que las convierte en variables de consenso. De estas, la inmensa mayoría (64) corresponden a inversiones potenciales en gas (*vGasInvestPotPipes*), mientras que únicamente 6 se vinculan a infraestructuras para reconversión de hidrógeno (*vGasRepurposePipeH2*). En cuanto al horizonte temporal, la tendencia es clara: predominan las decisiones planificadas para 2050 (34 variables), seguidas por 2040 (25) y 2030 (11). Esto sugiere que las inversiones más robustas y eficientes tienden a concentrarse en el largo plazo, alineándose con la lógica de planificación a futuro del sistema energético.

A partir del estudio de combinaciones binarias de estas variables de consenso, se identificaron numerosos pares de decisiones con muy buena performance en términos de coste medio. Aunque algunas de estas combinaciones presentan una frecuencia baja —por ejemplo, la mejor combinación 2030+2050 aparece en 39 de los 500 escenarios del decil eficiente—, su capacidad de ahorro es notable. En cambio, otras combinaciones como 2040+2050 o 2050+2050 muestran un equilibrio más realista entre frecuencia y eficiencia, siendo las más recurrentes dentro del conjunto eficiente. Estas sinergias entre decisiones sugieren que ciertas parejas de inversiones pueden constituir núcleos estratégicos del sistema, optimizando el coste total sin comprometer la robustez del escenario.

En conjunto, este análisis revela no solo cuáles son las inversiones con mejor comportamiento individual, sino también qué configuraciones de variables tienden a reforzarse mutuamente en distintos escenarios, y cuáles podrían ser consideradas como decisiones ancla en cualquier estrategia de planificación robusta. La metodología basada en multiclustering ha demostrado ser una herramienta eficaz para extraer conocimiento de alto valor a partir de soluciones optimizadas, permitiendo avanzar hacia una toma de decisiones más informada, precisa y resiliente.

7.4 RECOMENDACIONES OPERATIVAS

A partir de dichas variables, se ha realizado un estudio exhaustivo de combinaciones binarias (pares de inversiones coactivas) **reintegrando el conjunto completo de los 786 escenarios originales**. Aunque el análisis multiclustering y el filtrado por eficiencia habían reducido la muestra operativa a unas 60 soluciones, esta etapa final vuelve a contemplar **todos los escenarios posibles**, garantizando así una evaluación robusta de la frecuencia real y el impacto global de cada combinación en el universo completo de soluciones.

Se analizaron más de **2 400 pares distintos**, evaluando dos métricas clave: el **número de escenarios en que ambas variables se activan simultáneamente** y el **coste medio asociado** (*ObjFuncValue*). Los resultados se reflejan en el gráfico de dispersión adjunto, donde se aprecia claramente el *trade-off* entre rentabilidad económica y robustez.

Las combinaciones más baratas —como **430079 + 430143**, con ≈ 28 853 u.m. de coste medio— tienen una frecuencia moderada (≈ 39 escenarios). Por otro lado, combinaciones como **430028 + 430143** y **430033 + 430143**, con un coste ligeramente superior (≈ 28 940 u.m.), aparecen en **más de 100 escenarios distintos**, lo que las convierte en decisiones consistentes y de elevada repetibilidad.

Esta dualidad entre ahorro unitario y robustez operativa queda recogida en la tabla inferior, que presenta las **10 combinaciones con menor coste medio** y su **frecuencia de aparición** sobre los 786 escenarios originales. Las más eficientes tienden a involucrar tramos del

horizonte 2050, lo cual refuerza el patrón ya detectado en el análisis previo: **el largo plazo alberga las decisiones estratégicas más determinantes para la sostenibilidad técnica y económica del sistema energético**.

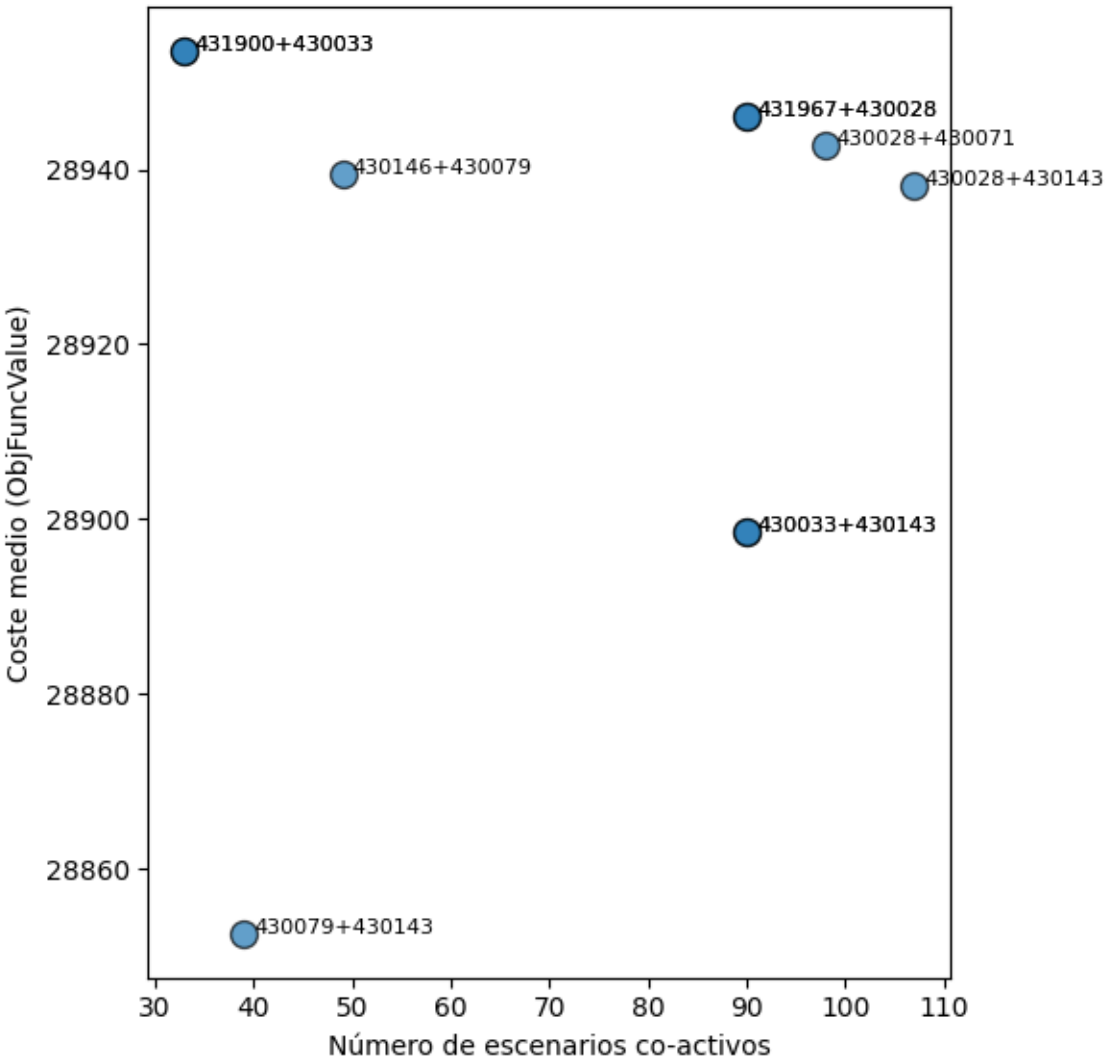


Figura 26 Mapa de combinaciones óptimas

7.5 CONCLUSIÓN

El objetivo central de este trabajo ha sido identificar conjuntos de soluciones quasi-óptimas en la planificación de la red de transmisión (TEP) bajo incertidumbre. Para ello, se ha aplicado un enfoque de análisis *multiclustering* sobre el decil de soluciones con menor coste (ObjFuncValue), tras depurar un espacio inicial de más de 1 600 configuraciones y 750 escenarios.

Entre los métodos evaluados, el algoritmo *Agglomerative Clustering* (enlace promedio) ha destacado por su capacidad para agrupar soluciones de bajo coste medio (≈ 28 565 u.m.) con estructuras coherentes y compactas. A partir de su clúster óptimo, se identificaron variables con alta frecuencia de activación, como *vGasInvestPotPipes_2050_430143*, *vGasInvestPotPipes_2040_430028* o *vGasInvestPotPipes_2050_430033*, que se repiten también en otros métodos, consolidándose como decisiones clave.

El análisis de combinaciones binarias refuerza esta conclusión, al revelar pares de inversiones con excelente desempeño tanto en coste medio como en frecuencia de aparición. Por ejemplo, la combinación 430033 + 430143 presenta un coste medio competitivo (≈ 28 898 u.m.) y aparece en más de 90 escenarios distintos, lo que la convierte en una decisión robusta y recurrente.

Capítulo 8. LÍNEAS DE TRABAJO FUTURAS

Aunque el marco propuesto ya permite identificar configuraciones de inversión robustas y detecta ciertas soluciones con impacto más positivo que otras, tres extensiones se consideran como alternativas viables en trabajos futuros por su relevancia académica y su aplicabilidad industrial inmediata.

8.1 CLUSTERING ADAPTATIVO BAJO INCERTIDUMBRE OPERATIVA

Los escenarios de demanda y de generación renovable cambian con rapidez; por lo tanto, resulta conveniente dotar al proceso de clustering de un componente adaptativo que reentrene los clusteres cuando se detecten desviaciones significativas. El concepto está basado en el funcionamiento de las redes neuronales, que se reentrenan a base de iteraciones. La literatura reciente en planificación estocástica sugiere el uso de algoritmos de agrupamiento dinámico que combinan K-Means con umbrales de resiliencia para redistribuir las observaciones cuando la varianza interna del grupo supera un valor crítico [10]. Integrar este enfoque permitiría actualizar la firma de consenso sin repetir todo el pipeline y reduciría el riesgo de obsolescencia de las recomendaciones en contextos de alta volatilidad, lo cual haría el proyecto consistente en el tiempo.

8.2 EVALUACIÓN DE RIESGO MEDIANTE MÉTRICAS DE COLA

En la fase actual el rendimiento de cada clúster se evalúa mediante la media de ObjFuncValue. Sin embargo, la toma de decisiones a largo plazo requiere conocer el riesgo de escenarios extremos. Ampliar el análisis con indicadores como *Value-at-Risk* (VaR) y *Conditional Value-at-Risk* (CVaR) sobre el coste total proporcionaría una medida explícita de la exposición a desviaciones adversas. Estudios recientes en expansión de redes con criterios CVaR demuestran que la selección de configuraciones basada en el percentil 95 de la distribución de costes puede reducir las pérdidas esperadas en situaciones de demanda pico hasta un 20 % sin incrementar el CAPEX medio [11] [12]. Incorporar este módulo de

riesgo reforzaría la robustez de las inversiones recomendadas y alinearía el modelo con las prácticas de gestión de activos de los operadores TSO.

8.3 MINERÍA DE REGLAS DE ASOCIACIÓN (ARM) APLICADA A CONFIGURACIONES EFICIENTES

Aunque el presente trabajo ha explorado técnicas avanzadas de segmentación y análisis multiclustering, también se contempló inicialmente el uso de minería de reglas de asociación (ARM) como herramienta complementaria para identificar patrones frecuentes de coactivación entre inversiones. Esta técnica se basa en la identificación de reglas del tipo *"si A, entonces B"*, muy empleadas en problemas de configuración binaria y selección de portfolios energéticos.

Sin embargo, **ARM no ha sido finalmente implementado** en el cuerpo central del análisis. El motivo principal es de naturaleza técnica: la **base de datos original se genera a partir de un procedimiento de descomposición de Benders**, lo que produce matrices de soluciones con **valores continuos o no estrictamente binarios**. Esto imposibilita el uso directo de ARM, que exige estructuras puramente discretas para calcular medidas como el *support*, *confidence* o *lift*. [13]

Aunque sería teóricamente posible aplicar una binarización previa —por ejemplo, mediante umbrales de activación sobre las variables continuas—, el coste computacional asociado y el riesgo de distorsión de la información original hacen que esta opción no resulte robusta en el contexto actual. No obstante, en futuros desarrollos, el diseño de bases de datos alternativas con decisiones ya consolidadas en formato binario permitiría retomar esta vía metodológica. En ese escenario, la implementación de ARM podría llevarse a cabo de forma eficiente mediante la librería mlxtend de Python, desarrollada por Raschka [14] , que ofrece utilidades específicas para el cálculo de reglas frecuentes y métricas asociadas a partir de estructuras transaccionales. Esta aproximación sería especialmente valiosa en etapas de

validación industrial o en entornos donde se requiera la generación de reglas interpretables y reutilizables.

Capítulo 9. BIBLIOGRAFÍA

La bibliografía recopilada en este trabajo abarca tres bloques fundamentales de conocimiento. En primer lugar, se incluyen artículos y revisiones científicas centradas en la planificación de la expansión de redes de transmisión (TEP), que permiten contextualizar el problema desde una perspectiva técnica, regulatoria y operativa, comparando los desafíos actuales frente a escenarios de alta penetración renovable e incertidumbre estructural.

En segundo lugar, se han consultado fuentes orientadas al aprendizaje automático no supervisado, con especial énfasis en técnicas de clustering, reducción de dimensionalidad y minería de reglas, tanto desde una perspectiva teórica como desde una práctica, indagando en su implementación desde Python.

Finalmente, se proporciona una descripción detallada de las bibliotecas de Python utilizadas en el desarrollo del análisis.

[1] H.-K. H. P. M. &. S. I. M. Ringkjøb, «A review of modelling tools for energy and electricity systems with large shares of variable renewables,» *Renewable and Sustainable Energy Reviews,* 2018.

[2] D. S. A. &. L. A. Valenzuela, « A data mining approach for classifying robust power system expansion plans under deep uncertainties,» *Applied Energy,* 2021.

[3] E. Sánchez-Úbeda, «Estadística II: Clustering (Lecture Notes) — Universidad Pontificia Comillas,» 2020. [En línea]. Available: https://www.comillas.edu/sites/default/files/clustering_estadistica_ii_sanchez_ubeda.pdf.

[4] S. &. B. W. Cao, «A Review of Evolving Challenges in Transmission Expansion Planning Problems,» *IEEE Access,* 2025.

[5] C. G. K. R. M. &. F. W. Perau, «Strategic Electrolyzer Placement for Green Hydrogen Production in Central Europe: A Comparative Analysis of RES-Driven and Demand-Driven Approaches Using the Synerginet Model,» 2023. [En línea]. Available: https://ssrn.com/abstract=4628263.

[6] I. Fariza, S. Carcar y J. Cruz Peña, «Iberdrola y Red Eléctrica se señalan mutuamente por el gran apagón,» *Cinco Días,* 29 Mayo 2025.

[7] X. Z. Y. L. P. &. L. G. Zhang, «Two-stage algorithm for efficient transmission expansion planning with renewable energy resources».*IET Renewable Power Generation.*

[8] J. K. M. P. J. Han, Data Mining: Concepts and Techniques, 2011.

[9] A. García Serrano, Aprende Machine Learning: Programación avanzada en Python aplicada al análisis de datos, 2021.

[10] Y. Z. A. G. D. X. Zhang, «A structural k-means method for climate pattern clustering with uncertainty evaluation,» *Geoscientific Model Development,* 2023.

[11] S. L. Y. Y. J. Z. L. J. H. Li, « A CVaR-based integrated energy system planning considering uncertain renewable generation and energy prices,» *Applied Energy,* 2021.

[12] C. W. Q. Y. H. L. J. Li, «Integrated energy system planning under carbon trading mechanisms using a multi-objective CVaR-based approach,» *Frontiers in Energy Research,* 2024.

[13] R. I. T. S. A. Agrawal, «Mining Association Rules Between Sets of Items in Large Databases,» *SIGMOD Record,* 1993.

[14] S. Raschka, «MLxtend: Providing machine learning and data science utilities and extensions to Python's scientific computing stack,» *Journal of Open Source Software,* 2018.

[15] J. K. M. P. J. Han, Data Mining: Concepts and Techniques, 2011.

Bibliotecas de Python utilizadas

- **MiniSom**: Implementación minimalista en Python del algoritmo *Self-Organizing Maps* (SOM), desarrollada por Giuseppe Vettigli. Utilizada para análisis no supervisado y visualización de topologías de datos.
- **Scikit-learn (sklearn)**: Biblioteca desarrollada por la comunidad científica (INRIA, entre otros), ampliamente utilizada para algoritmos de *Machine Learning* como clustering, reducción de dimensionalidad y minería de reglas.
- **Pandas**: Biblioteca desarrollada por Wes McKinney y la comunidad de código abierto. Fundamental para la manipulación, limpieza y análisis estructurado de datos en forma de tablas (*DataFrames*).
- **NumPy**: Biblioteca de cálculo numérico de alto rendimiento, base de muchas otras librerías científicas. Desarrollada por Travis Oliphant y colaboradores.
- **Matplotlib**: Biblioteca para la creación de gráficos 2D en Python, desarrollada originalmente por John D. Hunter. Permite una visualización detallada y personalizable de resultados.
- **Seaborn**: Extensión de Matplotlib creada por Michael Waskom para realizar visualizaciones estadísticas más atractivas y complejas con menos código.

ANEXO I: ALINEACIÓN DEL PROYECTO CON OBJETIVOS DE DESARROLLO SOSTENIBLES

El presente Trabajo de Fin de Grado se alinea con la Organización de las Naciones Unidas en la consecución de los objetivos de desarrollo sostenible (ODS). Atendiendo al carácter y al contenido de este proyecto, contribuye especialmente al ODS 7: Energía asequible y no contaminante y al ODS 9: Industria, innovación e infraestructura. Esta conexión determina tanto el fin y la aplicación de los resultados de este informe como la metodología de trabajo y las herramientas utilizadas para llevarla a cabo.

En un contexto de transición a fuentes de energías renovables, planificar adecuadamente la expansión de las redes de transmisión eléctrica se vuelve fundamental. Este trabajo, especialmente bajo escenarios de incertidumbre, no solo debe poner el foco en soluciones técnicamente viables, sino también sostenibles, tanto en el presente como en el futuro. Se entiende la sostenibilidad siguiendo la definición clásica enunciada en el congreso de la ONU de Brundtland en 1987: buscando una red que pueda satisfacer las necesidades presentes sin comprometer las futuras, sino más bien reforzándolas.

Para ello, a través del uso de algoritmos de análisis de datos para detectar combinaciones robustas de inversión, se contribuye al cumplimiento del **ODS 7** al favorecer un sistema eléctrico más eficiente, estable y accesible, que facilite el despliegue de energías limpias y reduzca la dependencia de fuentes tradicionales.

Por otro lado, la contribución al **ODS 9** es clara desde la metodología planteada y las tecnologías empleadas. Este trabajo aplica herramientas avanzadas de análisis de datos y Machine Learning para resolver un problema de gran escala en el ámbito energético, lo que representa una clara apuesta por la innovación tecnológica al servicio de infraestructuras más sostenibles. El uso de técnicas como el clustering no supervisado o diversos estadísticos para un preprocesado intensivo permite, no solo adquirir una mayor precisión respecto a la toma de decisiones, sino también sentar las bases para sistemas de planificación más inteligentes, automatizados y escalables para escenarios futuros con mayores conjuntos de datos. En este sentido, el proyecto se enmarca en la nueva ola de digitalización

y modernización de infraestructuras críticas, imprescindible para afrontar los retos energéticos que se planteen

Figura 1 ODS 7 y 9: Energía Sostenible e Infraestructuras Innovadoras

.

Figura 26 ODS 7 y 9: Energía Sostenible e Infraestructuras Innovadoras