CAPÍTULO VII

Clasificación de conjuntos de ofertas de electricidad en el mercado diario español

Jorge Arias Martí*
Andrés M. Alonso Fernández

El mercado eléctrico en España permite a los productores de electricidad ofrecer bloques de energía a diferentes precios, generalmente relacionados con sus costes marginales, en momentos concretos del día. El operador del sistema reúne las ofertas (bloques de energía) y sus correspondientes precios de todos los participantes para formar la curva de oferta con la que se obtendrá el precio marginal de cada hora. En este trabajo se estudian los conjuntos de oferta mediante la distancia de Hausdorff y se realiza la clasificación no supervisada de estos conjuntos. Adicionalmente, se caracterizan los grupos obtenidos mediante variables de producción de energía por las distintas tecnologías y variables temporales como hora, día de la semana y mes.

Palabras clave: clasificación no supervisada, distancia de Hausdorff, conjuntos de oferta.

* Este trabajo ha sido parcialmente financiado por la Agencia Estatal de Investigación (PID2019-108311GB-I00 / AEI / 10.13039/501100011033 / PID2022-138114NB-I00). Los autores agradecen los comentarios y mejoras sugeridas por el equipo editorial del libro, Daniel Peña, Pilar Poncela y Eva Senra.

1. INTRODUCCIÓN

El precio de la electricidad es un tema de gran interés actual. Casi a diario aparecen noticias relacionadas con los precios y la influencia de las diferentes tecnologías de producción. No hay duda de que los precios de la electricidad tienen un profundo impacto tanto en la economía nacional como en la empresarial. Por ejemplo, en Khobai *et al.* (2017) se encontró que "un aumento del 4 % en los precios de la electricidad provoca que el crecimiento económico disminuya en un 0,036 % en Sudáfrica". Por supuesto, ese impacto es común a cualquier economía.

Agosti *et al.* (2007) ofrecen información sobre cómo se fijan estos precios en España. Los autores explican que el componente más importante es el mercado diario, en el que se negocia la mayor parte de la energía para cada hora del día siguiente. En este mercado los productores de energía hacen ofertas de venta mientras que los consumidores hacen ofertas de compra, especificando tanto cantidad como precio. Para cada hora del día siguiente se obtienen dos curvas:

- Curva de oferta: se obtienen mediante la agregación de las cantidades de energía ofrecidas por los productores ordenadas de forma ascendente según su precio.
- Curva de demanda: se obtienen mediante la agregación de las cantidades de energía demandadas por los consumidores ordenadas en orden descendente según su precio.

El resultado son curvas formadas por puntos [p, q], donde p hace referencia al precio y q representa la cantidad de energía que se puede comprar al precio p. Al considerar simultáneamente ambas curvas, el precio de la electricidad (precio de casación) para esa hora viene dado por el punto donde se cruzan ambas curvas (Aggarwal *et al.,* 2009). A modo de ilustración, en la figura 1 el precio para la fecha y hora indicadas rondaría los 52 euros/MWh. Los compradores con demandas ubicadas a la izquierda del punto de cruce no podrán adquirir energía en ese período (pues su oferta está por debajo del precio de mercado), y la misma situación ocurre con los productores. Si hicieron una oferta ubicada a la derecha del punto de cruce no producirán energía en ese periodo (porque la ofrecen a un precio superior al precio de mercado).

En este trabajo se estudiarán los conjuntos horarios de oferta del mercado diario entre 2017 y 2021 mediante métodos de clasificación no supervisada (*clustering*). Se intentará descubrir, por ejemplo, cuáles son los periodos en los que se ofrece menor cantidad de energía, y si corresponden a horarios nocturnos o, por ejemplo, si hay una diferencia notable entre días laborables y festivos. También se tratará de descubrir cómo cambia la estructura de generación en función de la cantidad total de energía ofrecida. El conjunto de los datos a analizar se clasifica en la categoría de datos masivos (*big data*) por dos motivos: i) su volumen, pues tenemos 43.800 conjuntos de ofertas, y ii) su variabilidad o complejidad, pues cada conjunto de ofertas tiene dimensiones diferentes.

El resto del capítulo está organizado en seis secciones. La sección segunda presenta los elementos del mercado eléctrico necesarios para entender la obtención de las curvas de oferta;

Figura 1.

Ejemplo del proceso de obtención del precio de casación

Fuente: Elaboración propia a partir de datos disponibles en https://www.omie.es/es/market-results/daily/daily-market/aggragate-suply-curves

la sección tercera presenta una medida de disimilitud entre conjuntos (distancia de Hausdorff) y su procedimiento de cálculo; en la sección cuarta se recogen los métodos de agrupamiento que se han utilizado; la sección quinta desarrolla la metodología para caracterizar los diferentes clústers; en la sección sexta se presentan los resultados obtenidos, y finalmente en la sección séptima se presentan las conclusiones y posibles extensiones.

2. OBTENCIÓN DE LAS OFERTAS Y CURVAS DE OFERTA EN EL MERCADO ELÉCTRICO ESPAÑOL

2.1. Legislación del mercado eléctrico español

En el apartado anterior se ilustró el procedimiento para calcular los precios de la electricidad que se obtienen como consecuencia de la legislación de la Comisión Nacional de Mercados y la Competencia. En BOE/CNMC (2021) se establece cómo funciona el mercado eléctrico en España. En particular, existen algunas reglas de funcionamiento que se mencionan a continuación para una mejor comprensión del mercado de suministro eléctrico. La Regla 1 establece que en el mercado diario las operaciones de compra y venta se realizan para el día siguiente para cada una de las 24 horas o períodos naturales (este número también puede ser 23 ó 25 en los días correspondientes a días de cambio de hora oficial).

Además, la Regla 30.1 dice que el precio de cada período será el resultante de la compensación realizada por el algoritmo Euphemia. Este algoritmo utiliza curvas agregadas (Regla 30.2) que se obtienen de la siguiente manera:

- Curvas de oferta: Se obtienen sumando las cantidades de energía ofertadas para la venta en orden ascendente por el precio de las mismas (Regla 30.2.1).

- Curvas de demanda: Se obtienen sumando las cantidades de energía ofrecidas para comprar en orden decreciente por el precio de las mismas (Regla 30.2.2).

Como se indicó anteriormente, se obtienen como resultado las curvas formadas por los puntos [*p, q*] y que se han ilustrado en la figura 1.

En este trabajo, se obtienen todas las curvas de oferta en el periodo de 2017 a 2021 y, para ello, se necesita la información de las ofertas para cada hora. Esta información la proporciona el operador del mercado eléctrico de la península ibérica (OMIE) y se recoge en dos tipos de ficheros, obtenidos de OMIE (b) y OMIE (a). El tipo de información contenida en cada tipo de archivo se muestra en OMIE (c) y se resume a continuación:

- Los archivos de tipo cab_aaaammdd.1, contienen el número de identificación (código) de cada oferta y la clase de la misma (compra o venta).
- Los archivos de tipo det_aaaammdd.1, contienen información detallada de las ofertas. Son de interés el número o código de identificación, que vuelve a aparecer, la cantidad de energía ofrecida y el precio de la misma.

Figura 2.

Ejemplo de un archivo cab_.

4271539	2ICUVD04IBCUR	RE	4	VNO	0.000	0.000	0.0	0.0	0.000	0.000	2500.0	0.0	0.0	120170112112326
4271540	2ICUVD05IBCUR	RE	5	VNO	0.000	0.000	0.0	0.0	0.000	0.000	1000.0	0.0	0.0	120170112112326
4271541	2ICUVD06IBCUR	RE	6	VNO	0.000	0.000	0.0	0.0	0.000	0.000	1000.0	0.0	0.0	120170112112326
4271542	2ICUVD08IBCUR	RE	8	VNO	0.000	0.000	0.0	0.0	0.000	0.000	1000.0	0.0	0.0	120170112112326

Fuente: Elaboración propia a partir de archivo disponible en https://www.omie.es/es/file-access-list

Figura 3.

Ejemplo de un archivo det_.

1717319	311	1	0.000	0.000	1.0SS
1717319	312	1	0.000	0.000	1.0SS
1717319	313	1	0.000	0.000	1.0SS
1717319	314	1	0.000	0.000	1.0SS
1717319	315	1	0.000	0.000	1.0SS
1717319	316	1	0.000	0.000	1.0SS
1717319	317	1	0.000	0.000	1.0SS
1717319	318	1	0.000	0.000	1.0SS
1717319	319	1	0.000	0.000	1.0SS
1717319	320	1	0.000	0.000	1.0SS
1717319	321	1	0.000	0.000	1.0SS
1717319	322	1	0.000	0.000	1.0SS

Fuente: Elaboración propia a partir de archivo disponible en https://www.omie.es/es/file-access-list

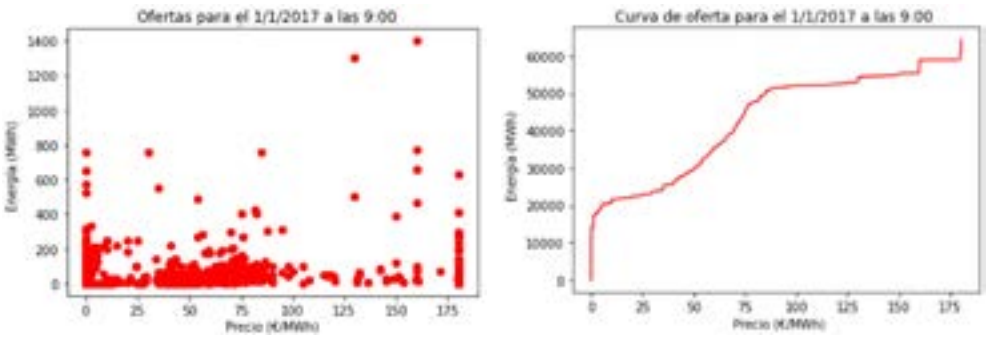
Las figuras 2 y 3 muestran un ejemplo de cada tipo de archivo.

Estos datos se ofrecen en conjuntos mensuales en archivos .zip, por lo que es necesario descargarlos y descomprimirlos.

Una vez descargados los archivos solo es necesario tomar el código y el tipo de oferta del archivo cab_aaaammdd.1 y cruzarlos con las variables del archivo det_aaaammdd.1. Iterando este procedimiento en los 1.826 archivos se obtienen las ofertas para cada periodo de cada día desde el 1 de enero de 2017 al 31 de diciembre de 2021. A partir de estas ofertas se obtienen las curvas de oferta clasificando las ofertas en orden creciente por su precio y haciendo la suma acumulada de las cantidades de energía (Regla 30.2.1). La figura 4 muestra un ejemplo de esta transformación, la cual es útil porque permite distinguir entre ofertas con el mismo precio y cantidad de energía. Sin embargo, se debe tener en cuenta que dos curvas diferentes no necesariamente contienen el mismo número de puntos. En un mismo escalón de una curva de oferta puede haber más de un punto y esto es debido a que existen varias ofertas al mismo precio.

Figura 4.

Transformación del conjunto de ofertas (a) en su curva de oferta correspondiente (b)



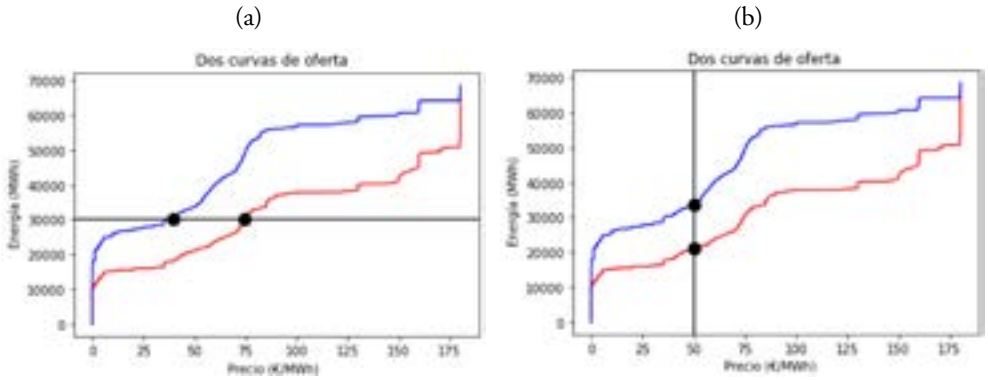
Fuente: Elaboración propia a partir de archivo disponible en https://www.omie.es/es/file-access-list

Una vez obtenidas, se pueden comparar las curvas de oferta. La figura 5 muestra dos ejemplos de curvas. Para una determinada cantidad de energía (a) la curva en rojo ofrece esa cantidad a un precio más alto, por lo que la curva en azul es más interesante desde la perspectiva del comprador. Para un precio determinado (b), la curva en azul hace referencia a una mayor cantidad de energía ofertada.

Ahora que se tienen las curvas, el siguiente paso es normalizarlas y seleccionar una medida de disimilitud entre conjuntos de puntos para poder realizar un método de agrupamiento. En este estudio se ha elegido la distancia de Hausdorff, que se presentará en la siguiente sección.

Figura 5.

Ejemplo de dos curvas de oferta



Nota: La figura (a) muestra los precios para una cantidad fija de energía, mientras que la figura (b) muestra las cantidades a un precio fijo.

Fuente: Elaboración propia a partir de archivo disponible en https://www.omie.es/es/file-access-list

3. DISTANCIA DE HAUSDORFF

Dados dos conjuntos de puntos de datos A y B, para cada punto $x \in A$ e $y \in B$ se define la distancia de Hausdorff dirigida de la siguiente manera (Taha y Hanbury, 2015):

$$\widehat{\mathrm{H}}(A,B) = \max_{x\in A}\left\{\min_{y\in B}\{\|x,y\|\}\right\}, \qquad [1]$$

donde $||\cdot, \cdot||$ puede ser cualquier norma en el espacio de los datos. En este trabajo se utiliza la norma euclidiana.

Se debe tener en cuenta que $\widehat{H}(A, B) \neq \widehat{H}(B, A)$, por lo que no se usa la distancia de Hausdorff dirigida, sino la no dirigida, que se define de la siguiente manera (Taha y Hanbury, 2015):

$$H(A, B) = \max\{\widehat{H}(A, B), \widehat{H}(B, A)\}. \qquad [2]$$

En este trabajo, A y B se corresponden con los puntos en dos curvas de oferta diferentes. La idea original era utilizar la distancia de Hausdorff aplicada a las nubes de puntos de oferta como las que se muestran en la figura 4(a), pero éstas no son propiamente un conjunto porque puede haber ofertas repetidas, es decir, ofertas con el mismo precio y cantidad. Al pasar de las ofertas a las curvas de oferta, se elimina el inconveniente de las repeticiones.

Para obtener la distancia de Hausdorff, se debe seguir el algoritmo 1:

Algoritmo 1.

Algoritmo para el cálculo de la distancia de Hausdorff

Input: Dos conjuntos de puntos, A y B.

Output: Distancia de Hausdorff.

1. Para un punto en A se calcula la distancia euclídea a todos los puntos en B. Se selecciona la menor distancia.
2. Se repite el paso anterior para todos los puntos de A.
3. De todas las distancias menores obtenidas en el paso 1, se selecciona la mayor, $\widehat{H}(A, B)$.
4. Se repiten los tres pasos anteriores intercambiando A y B para obtener $\widehat{H}(B, A)$.
5. Finalmente, se obtiene máx $\{\widehat{H}(A, B), \widehat{H}(B, A)\}$.

Fuente: Elaboración propia.

3.1. Computación de las distancias de Hausdorff

Tomando n como el número total de curvas de oferta, el objetivo es construir una matriz de distancias $n \times n$ donde cada elemento (i, j) sea la distancia de Hausdorff entre la curva i y la curva j. En este caso, $n = 43.800$ correspondientes a las curvas desde el 1 de enero de 2017 al 31 de diciembre de 2021[1]. Esto implica el cálculo de casi un billón de distancias.

Entre varios métodos de cálculo disponibles en Python, se encuentra que el más rápido para calcular la distancia de Hausdorff es usar la función *directed_hausdorff* ubicada en el módulo scipy.spatial.distance para encontrar $\widehat{H}(A, B)$ y $\widehat{H}(B, A)$ y tomar el máximo de estas dos distancias. También se consideraron las implementaciones en el módulo cuspatial. Sin embargo, a pesar de que este método era el mejor en términos de velocidad, el cálculo de toda la matriz habría llevado alrededor de tres meses, por lo que se modificó el objetivo inicial y se consideró el cálculo de las siguientes matrices de distancia:

- La primera es la matriz correspondiente a todas las curvas de 2019. Se eligió este año porque es el último antes de la pandemia del COVID-19.

- Para calcular la segunda matriz se seleccionó una hora valle (5:00) y una hora pico (12:00) para cada día. La matriz contiene todas las distancias entre todas estas horas para todos los años.

Con estas dos matrices se realizaron cuatro análisis: un análisis completo para todas las curvas de oferta de 2019; un segundo considerando solo una hora pico y una hora valle para cada día, y otros dos resultantes del estudio de estos dos tipos de curvas por separado. En la sección sexta se muestran los resultados de los dos primeros análisis, mientras que los resultados obtenidos de los dos restantes se han omitido por disponibilidad de espacio. No obstante, están disponibles mediante solicitud a los autores.

[1] En https://www.omie.es no estaban disponibles los archivos cab y det para el 1 de noviembre de 2021.

4. CLASIFICACIÓN NO SUPERVISADA

Después de calcular la matriz de distancias, es posible proceder con los métodos de agrupamiento o clasificación no supervisada. Se han considerado los procedimientos de partición alrededor de medoides (PAM) y agrupación jerárquica aglomerativa.

En la descripción del algoritmo PAM se utiliza el trabajo de Park y Jun (2009). Además del método mostrado para elegir los mediodes iniciales, los autores también proponen otras técnicas para seleccionarlos, pero en este capítulo se utilizó el enfoque heurístico. Este algoritmo se muestra a continuación:

Algoritmo 2.

Algoritmo para la obtención de la partición alrededor de medoides

Input: Un conjunto de datos o una matriz de disimilitud, y el número de grupos deseados = K.

Output: Grupo de cada observación y medoides.

1. Seleccione medoides iniciales.
   (a) Calcule la matriz de disimilitud utilizando una distancia predefinida.
   (b) Calcule v_j para la observación j:

$$v_j = \sum_{i=1}^{n} \frac{d_{ij}}{\sum_{l=1}^{n} d_{il}} \qquad [3]$$

   (c) Ordene v_j en orden ascendente y seleccione las primeras K observaciones con el valor más bajo. Estos objetos serán los medoides iniciales.
   (d) Asigne cada observación a su medoide más cercano para tener grupos iniciales.
   (e) Calcule la suma de las distancias entre cada observación y su medoide correspondiente.
2. Reemplace cada medoide con la observación que minimice la distancia total a las otras observaciones en el grupo.
3. Asigne cada observación a su medoide más cercano.
4. Calcule la suma de distancias de todos los objetos a su medioide correspondiente. Si esta suma es igual a la encontrada antes, detenga el algoritmo. De lo contrario, regrese al paso 2.

Fuente: Park y Jun (2009).

Notése que los *medoides* resultantes del algoritmo anterior son observaciones de cada grupo cuya distancia global es mínima al resto de las observaciones en el grupo, por tanto, pueden ser considerados como los representantes de los grupos.

- Por otro lado, una descripción del agrupamiento jerárquico aglomerativo puede consultarse, por ejemplo, en Nielsen (2016). El procedimiento "comienza desde los datos individuales [...] y va fusionando (iterativamente) de dos en dos los subconjuntos más cercanos" hasta que todos los datos hayan sido agrupados. Un elemento determinante en el agrupamiento jerárquico es la distancia de enlace o "distancia entre dos subconjuntos" que se denotará por $\Delta(X_i, X_j)$. Cuando los subconjuntos están compuestos sólo por un elemento, esta distancia es igual a la distancia entre las dos observaciones, pero cuando alguno o ambos subconjuntos tienen más de una observación entonces

se pueden utilizar varias definiciones de la distancia de enlace. Las siguientes son las más usadas:

- Enlace simple:

$$\Delta\left(X_i X_j\right) = \min_{x_i \in X_i, x_j \in X_j} D\left(x_i, x_j\right). \qquad [4]$$

- Enlace completo:

$$\Delta\left(X_i X_j\right) = \min_{x_i \in X_i, x_j \in X_j} D\left(x_i, x_j\right). \qquad [5]$$

- Enlace promedio

$$\Delta\left(X_i X_j\right) = \frac{1}{\left|X_i\right|\left|X_j\right|} \sum_{x_i \in X_i} \sum_{x_j \in X_j} D\left(x_i, x_j\right). \qquad [6]$$

Debido a que son los más interpretables, en la sección sexta se muestran los resultados obtenidos por el enlace promedio.

4.1. Evaluación del desempeño de la clasificación no supervisada

Para evaluar la agrupación obtenida mediante un procedimiento de clasificación no supervisada se han tenido en cuenta dos tipos de métricas:

- La primera métrica que se utiliza es el estadístico Silueta, cuya expresión es la siguiente (Kaufman y Rousseeuw, 1990):

$$s(i) = \frac{b(i) - a(i)}{\max\left\{a(i), b(i)\right\}}, \qquad [7]$$

donde $a(i)$ es la disimilitud promedio del objeto i con todos los demás objetos que pertenecen a su mismo grupo, mientras que $b(i)$ es el mínimo de todas las disimilitudes promedio entre el objeto i y el resto de los grupos, de los cuales no forma parte. El valor de $s(i)$ está entre -1 y 1. Cuando es cercana a 1 significa que la clasificación del objeto es adecuada. Si es próxima a 0, no está claro que la observación i deba asignarse a ese grupo o al grupo más cercano del resto de los grupos, y finalmente, si $s(i)$ es negativa significa que el objeto ha sido clasificado erróneamente.

Como medida global, se utiliza la media de las $s(i)$.

- La segunda métrica que se utiliza es el índice de separación, el cual está definido en Akhanli y Hennig (2020). Primero se define K como el número total de grupos. Para un objeto dado i en el grupo C_k, donde $k = \{1, 2, ..., K\}$, se calcula $d_{k:i}$ que es la distancia mínima entre el objeto i y todas las observaciones que no forman parte del grupo C_k.

Se repite el proceso para todas las observaciones en ese grupo y se ordenan todas las distancias resultantes en orden creciente. Luego, se toma una fracción igual a $[pn_k]$ de los valores más bajos de ellos, donde se sugiere que $p = 0.1$, n_k es el tamaño del grupo C_k y $[pn_k]$ representa el mayor número entero menor o igual a pn_k. El índice de separación tiene la siguiente expresión:

$$SI = \frac{1}{\sum_{k=1}^{K} [pn_k]} \sum_{k=1}^{K} \sum_{i=1}^{[pn_k]} d_{k:i}. \qquad [8]$$

A mayor índice de separación, mejor es la agrupación obtenida (Akhanli y Hennig, 2020). En la práctica, se han utilizado los procedimientos PAM y algoritmos jerárquicos para $K = 2$ hasta $K = 10$ y se ha seleccionado el número de grupos según la media del estadístico Silueta y, como segundo criterio, el índice de separación.

4.2. Evaluación de la similitud entre agrupamientos

Para evaluar la similitud de dos métodos de agrupamiento se utiliza el índice de Rand. La definición de esta métrica se puede encontrar, por ejemplo, en Warrens y van der Hoef (2020). Sean $\mathcal{A} = \{A_1, A_2, ..., A_I\}$ y $B = \{B_1, B_2, ..., B_J\}$ dos particiones diferentes de las observaciones donde I y J es el número de grupos de cada partición. El índice de Rand viene dado por la siguiente expresión:

$$R = \frac{a + d}{N}, \qquad [9]$$

donde $N = \frac{n(n-1)}{2}$ es el número total de pares de observaciones, n es el número de observaciones, a es el número de pares que pertenecen al mismo clúster en ambas particiones y d es el número de pares que no pertenecen al mismo clúster en ambas particiones. El índice de Rand está entre 0 y 1, y si $R = 1$, las particiones son idénticas.

5. DESCRIPCIÓN DE LAS CLASIFICACIONES OBTENIDAS

Una vez que se tienen las curvas divididas en grupos, el siguiente paso es caracterizar cada grupo. Para ello, se ha elaborado un conjunto de variables que contiene información relacionada con la producción eléctrica para cada hora desde el 1 de enero de 2017 al 31 de diciembre de 2021. Se distingue entre dos tipos de variables: las relacionadas con la estructura de generación eléctrica y las variables temporales (por ejemplo, si la hora es diurna o nocturna):

- Variables relacionadas con la estructura de generación: han sido recogidas del sitio web de ESIOS (Sistema de Información del Operador del Sistema) y corresponden a

los programas de generación de cada periodo en el sistema eléctrico peninsular español (BOE/SEE, 2012).

Se ha elegido la energía programada en lugar de la energía generada real porque la primera se acerca más que la segunda al concepto de curvas de oferta que se está analizando.

Las variables de este tipo que se han seleccionado son la energía programada (en MWh) para cada hora por las siguientes fuentes: biogás, biomasa, carbón, ciclo combinado (CCGT), derivados del carbón y del petróleo, hidroeléctrica, cogeneración a gas natural (NGcog), nuclear, eólica, solar fotovoltaica y solar térmica. Además, se ha incluido la variable Generación Horaria Operativa Total, que es, para cada hora, la suma de la energía programada producida por todas las fuentes.

Las fuentes que no se han tenido en cuenta son las siguientes: genéricos, geotérmicos y oceánicos, residuos domésticos y afines, subproductos mineros, energía residual y bombeo por turbinas. Se ha comprobado que, por ejemplo, en 2019 estos tipos de generación representaron en su conjunto tan sólo el 1,22 % del total de energía programada para este año.

- Variables temporales:

  - Mes: mes correspondiente a la curva.
  - Día: día de la semana correspondiente a la curva.
  - Noche: variable binaria cuyo valor es 1 si la hora es entre las 23:00 y las 6:00 horas, incluidas ambas, y 0 en caso contrario.
  - Festivo: variable binaria cuyo valor es 1 si la oferta corresponde a sábado o domingo o si corresponde a un día festivo nacional en España. Su valor es 0 en caso contrario.
  - Festivo & Verano: variable binaria cuyo valor es 1 si corresponde a un festivo o el mes correspondiente de la oferta es julio o agosto. Su valor es 0 en caso contrario.

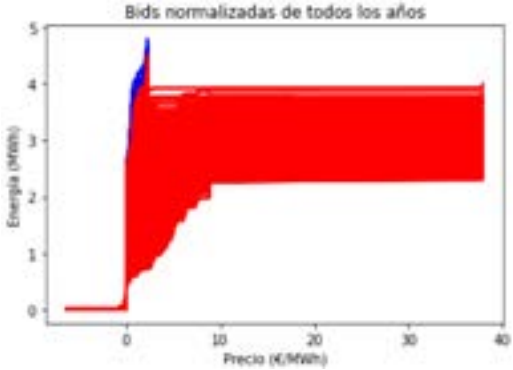
Una vez creado el conjunto de datos, se dividen todas las variables relacionadas con la estructura de generación entre la Generación Horaria Operativa Total, para obtener el porcentaje de energía generada por cada fuente para cada hora.

6. RESULTADOS Y DISCUSIÓN

Finalmente, se decidió omitir en el análisis las curvas de 2021 porque eran sustancialmente diferentes a las demás. El motivo es la gran diferencia entre los precios máximos de este año y el resto por un cambio en la legislación. La figura 6 ilustra este hecho. En la figura 6 y siguientes, se representan las curvas normalizadas, es decir, tanto los precios como las cantidades de energía se dividirán por la desviación estándar de todos los precios y todas

Figura 6.

Curvas de oferta normalizadas para el período de 2017 a 2021



Nota: Las curvas rojas corresponden a 2021 y las curvas azules corresponden al período 2017–2020.
Fuente: Elaboración propia.

las cantidades de energía, respectivamente. El objetivo de dicha normalización es que la distancia de Hausdorff no dependa de las unidades de medida.

6.1. Resultados de clasificación de las curvas de 2019 usando partición alrededor de medoides

En el cuadro 1 se muestra la media del estadístico Silueta y el índice de separación para diferentes valores de K utilizando PAM como procedimiento de clasificación. Los mejores resultados se obtienen para $K = 2$ y 3. Las clasificaciones obtenidas se describen en las siguientes subsecciones.

Cuadro 1.

Media del estadístico Silueta e Indice de separación en clasificaciones obtenidas mediante partición alrededor de medoides

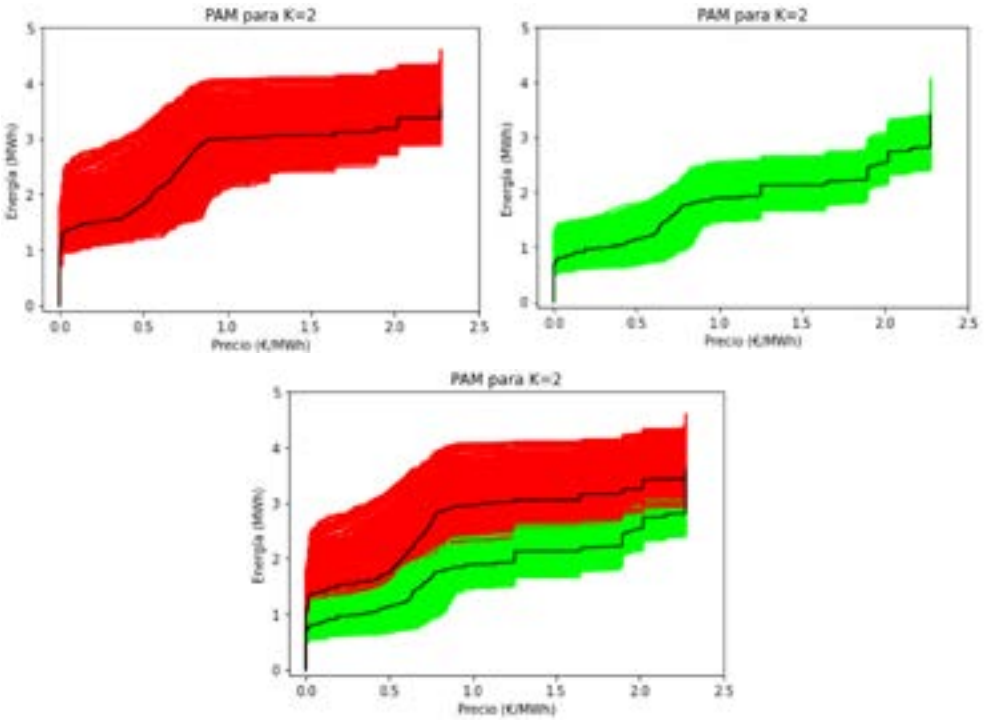
K	*Media del estadístico Silueta*	*Índice de separación*
2	0,57	0,131
3	0,43	0,069
4	0,35	0,061
5	0,33	0,047
6	0,31	0,046
7	0,28	0,044
8	0,26	0,042
9	0,23	0,041
10	0,22	0,040

Fuente: Elaboración propia.

6.1.1. Resultados utilizando PAM para K=2

Figura 7.

Curvas de oferta agrupadas mediante partición alrededor de medoides para K = 2



Nota: Las líneas en negro representan los medoides.
Fuente: Elaboración propia.

En la figura 7 se representan todas las curvas de oferta de 2019 de los grupos obtenidos con PAM para $K = 2$. Se observa que hay una cierta superposición entre las curvas verdes y rojas, lo que significa que la distancia de Hausdorff no solamente tiene en cuenta la proximidad entre dos curvas, sino también su forma. En general, las curvas verdes se refieren a ofertas más caras que las curvas rojas.

Para estos grupos, se analiza la proporción (en porcentaje) de curvas que pertenecen a cada mes. Hay que tener en cuenta que para un grupo determinado, una proporción uniforme para cada mes sería igual a 100·1/12=8,33 %, por lo que se esperan valores en torno a este número. Los resultados se muestran en el cuadro 2.

Como se ve, las curvas están distribuidas de forma más o menos uniforme en meses para cada grupo, aunque existen ligeras diferencias en los meses de enero y diciembre (mayor

Cuadro 2.

Para cada grupo definido en la figura 7, porcentaje de ofertas que corresponden a cada mes

	En.	*Feb.*	*Mar.*	*Abr.*	*May*	*Jun.*	*Jul.*	*Ag.*	*Sep.*	*Oct.*	*Nov.*	*Dic.*
■	9,05	7,89	8,41	7,95	8,11	7,85	8,21	8,21	8,07	8,10	8,70	9,39
■	7,19	7,16	8,71	8,82	9,35	9,05	9,13	9,13	8,56	9,35	7,08	6,40

Fuente: Elaboración propia.

porcentaje en las curvas rojas respecto a las verdes). Se analiza ahora si ocurre lo mismo con los días de la semana, pero ahora teniendo en cuenta que en una distribución uniforme para cada día correspondería a 100·1/7=14,28 %. Como se puede comprobar, ambos grupos parecen tener distribuciones uniformes respecto a los días de la semana.

Cuadro 3.

Para cada grupo definido en la figura 7, porcentaje de ofertas que corresponden a cada día de la semana

	Lunes	*Martes*	*Miércoles*	*Jueves*	*Viernes*	*Sábado*	*Domingo*
■	14,18	14,37	14,31	14,49	14,42	14,24	13,96
■	14,43	14,81	14,09	13,64	13,83	14,24	14,92

Fuente: Elaboración propia.

En lo que sigue se consideran las variables binarias: Noche (1 si la curva se fija para una hora entre las 23:00 y las 6:00), Festivo (1 si la curva se fija para un sábado, domingo o festivo Nacional) y Festivo & Verano (1 si la variable Festivo es 1 o el mes correspondiente de la curva es julio o agosto). En el cuadro 4 se presenta el porcentaje de curvas cuya respectiva variable es igual a 1 para cada grupo. Se observa que el 83 % de las curvas que pertenecen al clúster verde son nocturnas, mientras que sólo el 11 % de las curvas que forman el clúster rojo lo son.

Cuadro 4.

Para cada grupo definido en la figura 7, porcentaje de curvas que corresponden a días no laborables (Festivo), que corresponden a día Festivo o Verano (Festivo & Verano), o que corresponden a horas nocturnas (Noche)

	Festivo	*Festivo & Verano*	*Noche*
■	30,31	42,05	11,73
■	31,56	44,33	83,44

Fuente: Elaboración propia.

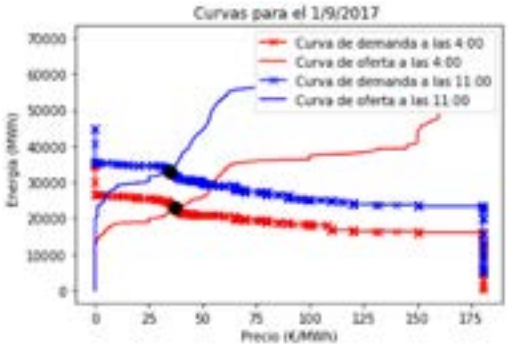
Si se mira nuevamente la figura 7, se ve que el grupo verde corresponde a curvas donde la energía es más cara, lo que puede parecer contradictorio porque se espera que la energía sea más barata en la noche. Antes de explicar este hecho, se debe analizar la figura 8. En esta figura se trazan, sin normalizar, las curvas de oferta y demanda para una hora nocturna y lo mismo

para una hora diurna. En primer lugar, se observa que la curva de demanda correspondiente a las 4 AM está por debajo (la demanda es menor) que la curva de demanda correspondiente a las 11 AM, lo que provoca que ambos precios sean similares, aunque el nocturno sigue siendo más caro (debe recordarse que el precio está dado por la intersección entre oferta y demanda). Sin embargo, se debe tener en cuenta que estas curvas corresponden a todas las ofertas realizadas para estos dos periodos, no las que corresponden únicamente a las ofertas que entran en la casación.

Se debe considerar que las ofertas que se han realizado pueden estar sujetas a una condicionalidad fijada por el productor como la que permite la Regla 40.3 de BOE/CNMC (2021), y cuando esta condicionalidad no se cumple, se hacen algunas correcciones, que al final provocan que la parte de la curva de oferta que ha casado se mueva hacia la izquierda (más barata) respecto a la original.

Figura 8.

Curvas de oferta y demanda para horas nocturnas y diurnas



Fuente: Elaboración propia.

En el cuadro 5 se recoge el porcentaje promedio de energía programada por fuente para cada grupo y el promedio total de energía:

Cuadro 5.

Para cada grupo definido en la figura 7, porcentaje promedio de energía programada por fuente para cada grupo y el promedio total de energía

	Total (MWh)	*CCGT*	*Hidro*	*NGcog*	*Nuclear*	*Eólica*	*SolarFV*	*SolarT*
■	30.202,30	21,06	10,44	10,30	21,38	20,81	4,75	2,70
■	24.044,08	19,85	8,42	12,65	26,87	22,50	0,05	0,70

	Biogás	*Biomasa*	*Carbón*	*Petróleo*
■	0,31	1,09	4,19	1,05
■	0,37	1,34	4,44	1,22

Nota: La primera columna es la energía promedio por grupo.
Fuente: Elaboración propia.

En el cuadro 5 se observa que obviamente, el porcentaje promedio de energía solar generada (fotovoltaica y térmica) es notablemente mayor en el grupo rojo (diurno) que en el verde (nocturno). El porcentaje casi nulo de energía solar parece ser sustituido en la horas nocturnas principalmente por la nuclear (diferencia de unos 6 puntos), pero también por la cogeneración. Además, la energía media total es notablemente mayor en el grupo diurno rojo que en el verde, lo cual tiene sentido.

Finalmente, se realiza una clasificación supervisada mediante el algoritmo de bosque aleatorio (*random forest*), propuesto por Breiman (2001), para predecir la etiqueta del grupo usando las variables explicativas. Se dividen los datos en conjunto de entrenamiento (80 % de los datos) y de prueba (20 % de los datos). Se obtuvo una precisión del 94,51 %. La matriz de confusión correspondiente se muestra en el cuadro 6 y las 15 variables más importantes en la clasificación se muestran en la figura 9.

Cuadro 6.

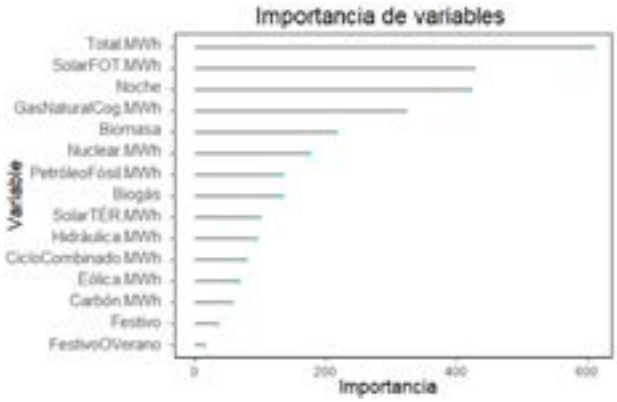
Matriz de confusión en el conjunto de prueba

	Etiqueta real	
Predicción	■	■
■	1.139	11
■	85	516

Fuente: Elaboración propia.

Figura 9.

Variables más importantes en la clasificación supervisada



Fuente: Elaboración propia.

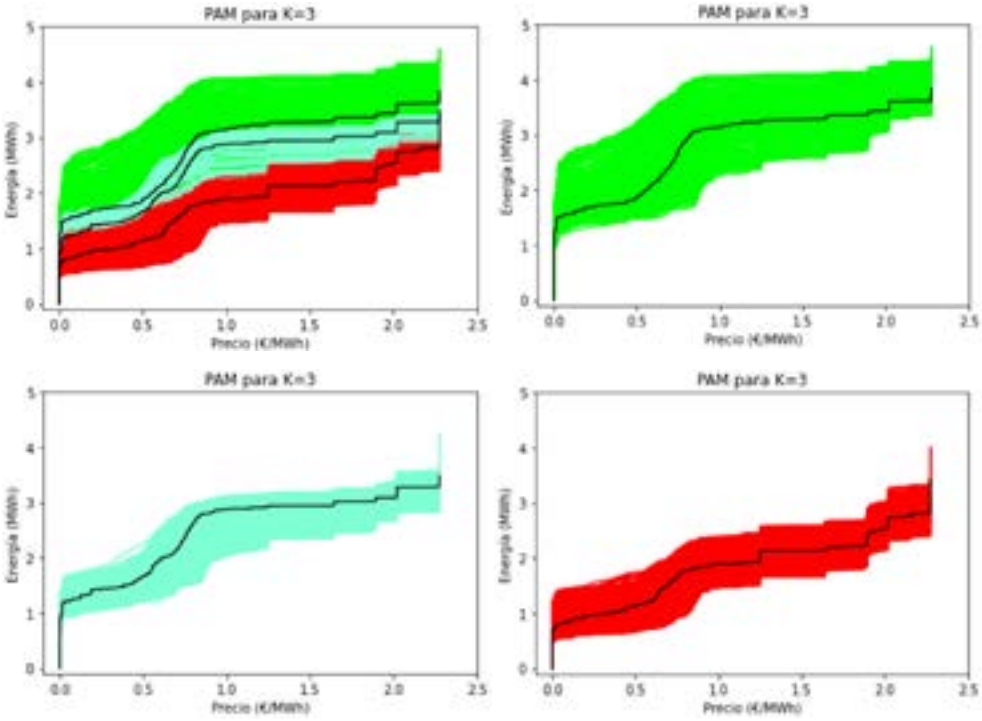
Como se observa, las variables Solar fotovoltaica y Noche son dos de las más importantes, confirmando las conclusiones anteriores.

6.1.2. Resultados utilizando PAM para K=3

Los grupos obtenidos mediante PAM para $K = 3$ se muestran en la figura 10, donde se observa nuevamente algunos solapamientos, especialmente entre las curvas verde y azul. Además, las medias de los estadísticos Silueta promedio y los índices de separación son algo más bajos en este caso, como muestra el cuadro 1.

Figura 10.

Curvas de oferta agrupadas mediante partición alrededor de medoides para $K = 3$



Nota: Las líneas en negro representan los medoides.

Fuente: Elaboración propia.

En el cuadro 7 se presenta la proporción de cada mes para cada grupo.

Un aspecto destacable que se observa en el cuadro 7 es que la proporción de curvas verdes asociadas a los meses comprendidos entre noviembre y abril es mayor que en el resto. Posteriormente se encontrará que curvas superiores, como las situadas en la parte más elevada del clúster verde, están relacionadas con una alta producción de energía eólica, pero esto se puede anticipar observando la distribución mensual de la producción con este tipo de energía. En la figura 11 se representa la energía eólica promedio programada mensualmente

Cuadro 7.

Para cada grupo definido en la figura 10, porcentaje de ofertas que corresponden a cada mes

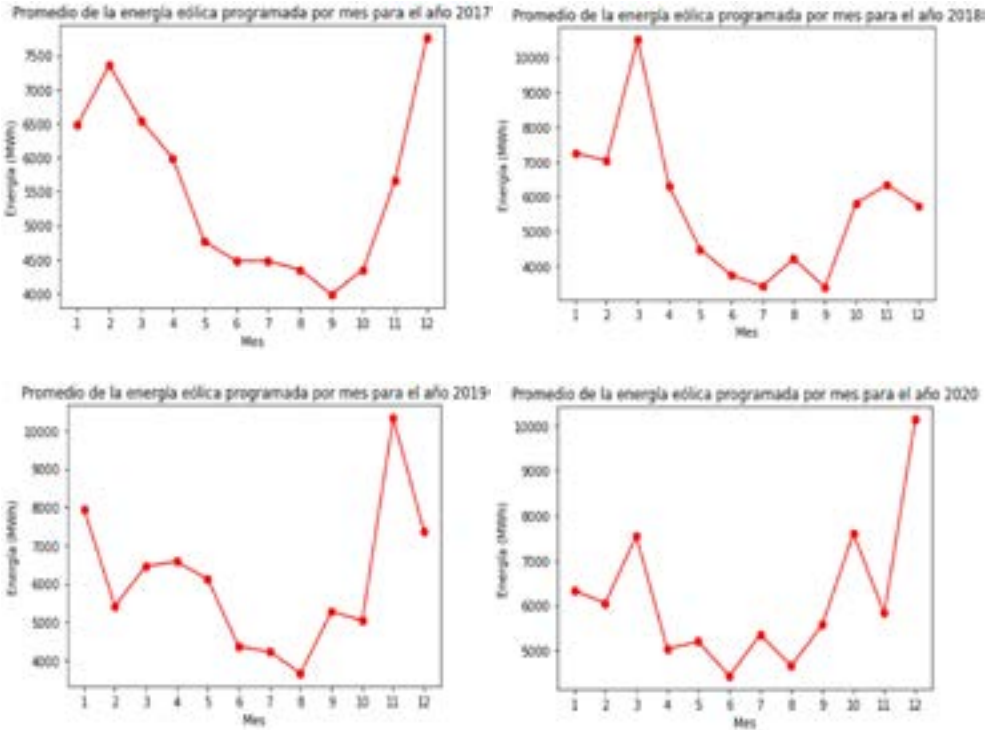
	En.	*Feb.*	*Mar.*	*Abr.*	*May*	*Jun.*	*Jul.*	*Ag.*	*Sep.*	*Oct.*	*Nov.*	*Dic.*
■	16,20	11,85	12,00	10,70	7,12	3,65	1,07	3,15	5,50	2,00	12,54	14,16
■	3,96	5,06	5,86	5,94	8,92	10,81	13,30	11,74	9,77	12,59	6,02	6,02
■	9,09	7,13	8,73	8,97	9,29	9,17	9,17	9,29	8,77	9,21	6,93	6,22

Fuente: Elaboración propia.

para cada año. Nótese que este promedio es menor en los meses de verano más septiembre y octubre, por lo que si las curvas de un grupo no se ubican en estos meses, se espera que su producción eólica asociada sea alta. Por lo tanto, se espera que las curvas verdes tengan un alto porcentaje de energía eólica.

Figura 11.

Energía eólica promedio programada por mes para los años 2017–2020



Fuente: Elaboración propia.

Además, para cada grupo se ha observado nuevamente una proporción igual para todos los días de la semana (ver cuadro 29 en el Apéndice). Sin embargo, la distinción entre curvas de día y de noche vuelve a aparecer, como muestra el cuadro 8.

Cuadro 8.

Para cada grupo definido en la figura 10, porcentaje de curvas que corresponden a días no laborables (Festivo), que corresponden a Festivo o Verano (Festivo & Verano), o que corresponden a horas nocturnas (Noche)

	Festivo	*Festivo & Verano*	*Noche*
■	29,75	32,95	6,74
■	30,54	48,44	16,15
■	31,86	44,58	85,93

Fuente: Elaboración propia.

Como se observa, el grupo azul tiene un porcentaje del 16 % de curvas que son nocturnas, mientras que en el rojo este porcentaje es del 85,3 %. Además, el porcentaje de días no laborables es similar para los tres grupos mientras que el grupo verde tiene una menor proporción de curvas Festivo & Verano, lo que se debe a la baja proporción de conjuntos de ofertas en los meses estivales.

En el cuadro 9, se presenta la estructura de generación para los tres grupos obtenidos.

Cuadro 9.

Para cada grupo definido en la figura 10, porcentaje promedio de energía programada por fuente

	Total (MWh)	*CCGT*	*Hidro*	*NGcog*	*Nuclear*	*Eólica*	*SolarFV*	*SolarT*
■	31.180,03	13,65	11,03	10,01	20,55	28,13	5,10	2,45
■	29.379,62	26,34	10,04	10,55	22,06	15,70	4,36	2,80
■	23.907,41	19,78	8,30	12,71	27,02	22,44	0,03	0,72

	Biogás	*Biomasa*	*Carbón*	*Petróleo*
■	0,30	1,04	4,82	0,99
■	0,31	1,13	3,74	1,09
■	0,37	1,35	4,47	1,22

Nota: La primera columna es la energía promedio por grupo.
Fuente: Elaboración propia.

Observando el cuadro 9 se comprueba que el grupo verde tiene un porcentaje de energía eólica notablemente superior a los demás, y cómo su energía total programada es la más alta, se puede concluir que en términos absolutos la cantidad de energía eólica que ofrecen estas curvas es también la más alta. El hecho de disponer de una cantidad tan grande de energía eólica es una posible hipótesis que explica por qué estas curvas hacen referencia a las ofertas más baratas. Además, las curvas azules tienen un mayor porcentaje de energía de ciclo

combinado, lo cual tiene sentido, porque están relacionadas con horas diurnas con un bajo porcentaje de energía eólica, por lo que se necesita algún tipo de energía de respaldo.

Finalmente, se muestran los resultados obtenidos con el algoritmo de bosque aleatorio (dividiendo nuevamente los datos en un conjunto de entrenamiento y uno de prueba). En este caso, la matriz de confusión (ver cuadro 10) ilustra que la clasificación errónea entre los grupos superior e inferior (verde y rojo, respectivamente) es casi nula, mientras que el error es mayor cuando el algoritmo intenta distinguir entre los grupos superiores y medio (azul) o entre los grupos medio e inferior. La precisión obtenida en este caso es del 91,94 %.

Cuadro 10.

Matriz de confusión en el conjunto de prueba

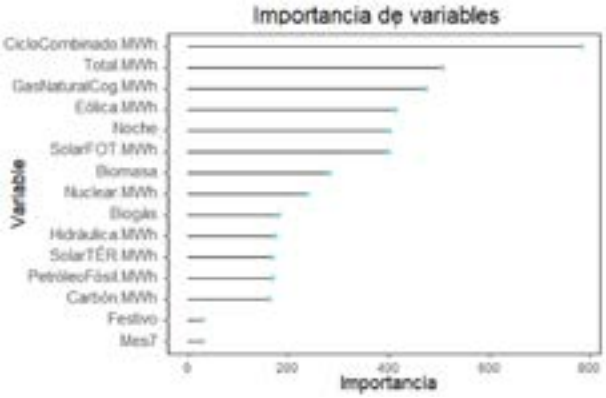
	Etiqueta real		
Predicción	■	■	■
■	469	34	1
■	48	656	16
■	2	40	484

Fuente: Elaboración propia.

El diagrama de la importancia de las variables en el procedimiento de clasificación se muestra en la figura 12. Las variables Noche y Solar fotovoltaica vuelven a ser importantes porque ayudan a distinguir el grupo rojo del resto. Además, la variable Ciclo Combinado es la más importante, lo que parece lógico porque el porcentaje medio de energía generada con ella es diferente para cada grupo. Sin embargo, ha sorprendido la importancia de la Cogeneración de Gas Natural, pero quizás también sea importante para distinguir entre el grupo rojo y el resto de los grupos. Se observa también que la energía eólica es la cuarta más importante.

Figura 12.

Variables más importantes en la clasificación supervisada



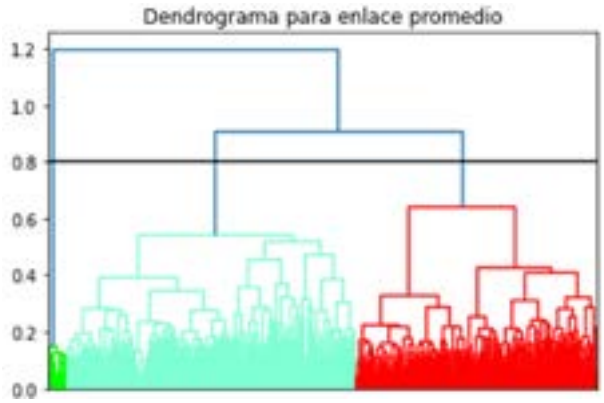
Fuente: Elaboración propia.

6.2. Resultados de clasificación de las curvas de 2019 usando un procedimiento aglomerativo

En la figura 13 se muestra el dendrograma de un agrupamiento aglomerativo usando enlace promedio (*average linkage*) aplicado a las curvas de 2019. Además, en el cuadro 11 se muestra la media del estadístico Silueta y el índice de separación para entre 2 y 10 grupos. Nuevamente son mejores los valores bajos de K. Se decide eligir K=3 puesto que K=2 conduce a una solución con un grupo muy pequeño y otro con el resto de las curvas.

Figura 13.

Dendrograma para las curvas de oferta de 2019



Nota: La línea negra corta las líneas verticales correspondientes a los tres grupos elegidos.
Fuente: Elaboración propia.

Cuadro 11.

Media del estadístico Silueta e índice de separación en clasificaciones obtenidas mediante un procedimiento aglomerativo con enlace promedio

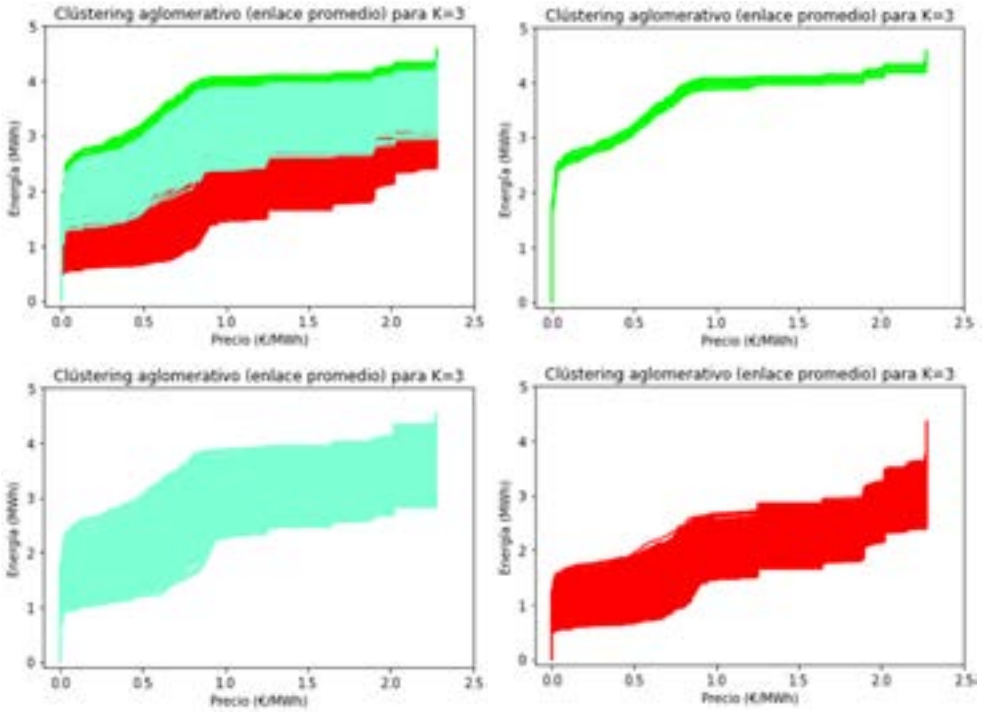
K	*Media del estadístico Silueta*	*Índice de separación*
2	0,46	0,553
3	0,51	0,145
4	0,40	0,142
5	0,40	0,084
6	0,40	0,084
7	0,39	0,084
8	0,37	0,084
9	0,27	0,084
10	0,29	0,074

Fuente: Elaboración propia.

En la figura 14 se muestran las curvas de los grupos obtenidos por agrupamiento aglomerativo para K=3. Se observa que no están tan superpuestos como los grupos obtenidos con PAM.

Figura 14.

Curvas de oferta agrupadas mediante procedimiento aglomerativo con enlace promedio para K=3



Fuente: Elaboración propia.

Se comprueba que el clúster verde está formado por 21 curvas correspondientes a las fechas y horas comprendidas entre las 10:00 y las 22:00 horas del 21 de diciembre de 2019 (Sábado) y entre las 10:00 y las 17:00 horas del 22 de diciembre de 2019 (Domingo). Para las curvas roja y azul no se ha observado ninguna distribución especial, ni en meses ni en días (ver cuadros 30 y 31 en el Apéndice).

En el cuadro 12 se presentan los porcentajes de las variables binarias en estos grupos. Nuevamente se tiene una fuerte separación entre las curvas diurnas y nocturnas.

Por otra parte, en el cuadro 13, se observa que las curvas verdes tienen un alto porcentaje de energía generada tanto por fuentes eólicas como hidráulicas, y un porcentaje bajo de

Cuadro 12.

Para cada grupo definido en la figura 14, porcentaje de curvas que corresponden a días no laborables (Festivo), que corresponden a Festivo o Verano (Festivo & Verano) o que corresponde a horas nocturnas (Noche)

	Festivo	*Festivo & Verano*	*Noche*
■	100,00	100,00	0,00
■	32,24	42,24	10,25
■	31,68	43,63	86,22

Fuente: Elaboración propia.

energía de ciclo combinado (CCGT), lo que puede explicar lo baratas que son estas ofertas. Además, los principales componentes de las curvas azules son el ciclo combinado, la nuclear y la cogeneración. En conclusión, se tienen tres grupos: uno que se refiere a las ofertas nocturnas y dos que se refieren a las diurnas. Entre ambos grupos diurnos, un grupo está relacionado con una gran cantidad de energía a bajos precios y tiene un alto porcentaje de energía generada por fuentes renovables.

Cuadro 13.

Para cada grupo definido en la figura 14, porcentaje promedio de energía programada por fuente

	Total (MWh)	*CCGT*	*Hidro*	*NGcog*	*Nuclear*	*Eólica*	*SolarFV*	*SolarT*
■	30.620,12	3,58	25,08	7,41	14,32	39,73	3,76	0,27
■	30.157,27	21,50	8,37	10,33	26,76	20,23	4,78	0,69
■	24.167,51	18,97	10,42	12,59	21,44	23,65	0,03	2,72

	Biogás	*Biomasa*	*Carbón*	*Petróleo*
■	0,29	0,91	0,00	0,33
■	0,31	1,09	4,20	1,05
■	0,37	1,34	4,47	1,21

Nota: La primera columna es la energía promedio por grupo.
Fuente: Elaboración propia.

Finalmente, se muestran los resultados de realizar una clasificación supervisada mediante un procedimiento de bosque aleatorio. Para este caso, se presentan dos situaciones diferentes: en la primera no se han dividido los datos en conjunto de entrenamiento y prueba debido al tamaño reducido del grupo más pequeño (el verde) y en la segunda se ha dividido el conjunto de datos ignorando el grupo verde. La matriz de confusión y el diagrama con la importancia de las variables para cada caso se muestran en el cuadro 14 y en el gráfico 15, respectivamente. Las precisiones son del 99,60 % para el conjunto completo de tres grupos, y del 96,56 % para el caso con dos grupos. Curiosamente, la energía eólica y la hidroeléctrica casi no son importantes, ni siquiera en el caso en que se toma en consideración el grupo verde. Ambos diagramas son muy similares y Noche y SolarFOT siguen siendo dos de las variables más relevantes.

Cuadro 14.

Matrices de confusión en el conjunto completo de datos y en el conjunto de prueba con dos grupos

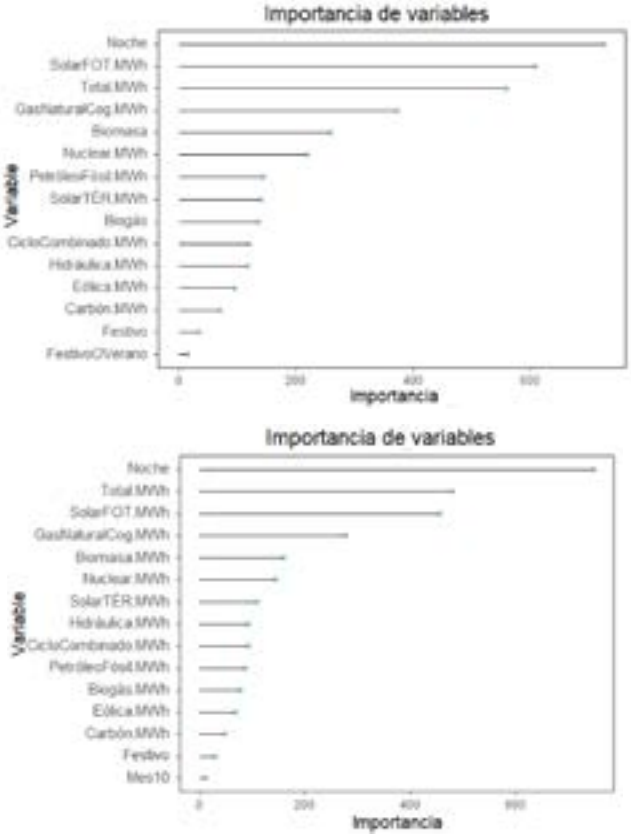
	Etiqueta real		
Predicción	■	■	■
■	21	10	0
■	0	6.055	0
■	0	25	2.649

	Etiqueta real	
Predicción	■	■
■	1.200	42
■	18	487

Fuente: Elaboración propia.

Figura 15.

Variables más importantes en la clasificación supervisada



Nota: El gráfico superior corresponde a la clasificación del conjunto completo con tres grupos y el gráfico inferior corresponde a la clasificación de los dos grupos mayoritarios.

Fuente: Elaboración propia.

6.3. Resultados de clasificación de curvas de horas pico y valle usando partición alrededor de medoides

En esta sección se toman las curvas de una hora pico (12 a.m) y una hora valle (5 a.m.) de cada día para todo el período (2017-2020). La idea es comprobar si los resultados obtenidos al analizar las curvas de 2019 son generalizables o no.

En el cuadro 15 se presentan los valores de la media del estadístico Silueta y el índice de separación.

Cuadro 15.

Media del estadístico Silueta e índice de separación en clasificaciones obtenidas mediante partición alrededor de medoides

K	*Media del estadístico Silueta*	*Índice de separación*
2	0,60	0,206
3	0,45	0,115
4	0,35	0,070
5	0,31	0,066
6	0,27	0,063
7	0,24	0,061
8	0,22	0,061
9	0,21	0,062
10	0,20	0,061

Fuente: Elaboración propia.

Nuevamente, se obtienen los valores más altos de los índices cuando K es bajo, por lo que se selecciona $K = 2$ y 3. Las clasificaciones obtenidas se describen en las siguientes subsecciones.

6.3.1. Resultados utilizando PAM para K = *2*

En la figura 16 se presentan los grupos obtenidos para $K = 2$ usando PAM. Debe tenerse en cuenta que en este caso los dos grupos no parecen estar demasiado superpuestos.

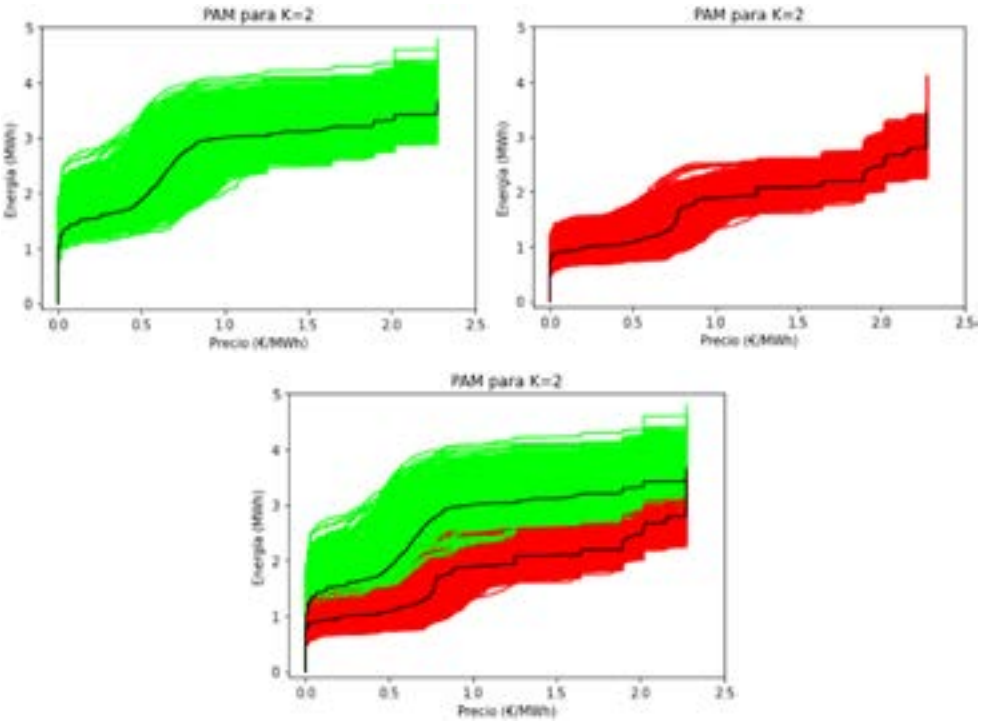
Una vez más, para cada grupo se ha observado una distribución igual en meses y en días (cuadros 32 y 33 en el Apéndice). Además, aunque no hay diferencia en la proporción de días festivos o de verano, se observa nuevamente la separación entre las curvas nocturnas y diurnas, como ilustra el cuadro 16.

Este es un buen resultado porque implica que los resultados obtenidos analizando las curvas correspondientes a 2019 pueden generalizarse a todo el período. De hecho, se puede realizar la siguiente comparación:

- Para todas estas horas punta y valle tomar únicamente las correspondientes al año 2019.
- Para todas las curvas de 2019, se toman solo las correspondientes a las 5:00 y a las 12:00.

Figura 16.

Curvas de oferta agrupadas mediante partición alrededor de medoides para K = 2



Nota: Las líneas en negro representan los medoides.
Fuente: Elaboración propia.

Cuadro 16.

Para cada grupo definido en la figura 16, porcentaje de curvas que corresponden a días no laborables (Festivo), que corresponden a un festivo o Verano (Festivo & Verano), o que corresponden a horas nocturnas (Noche)

	Festivo	*Festivo & Verano*	*Noche*
■	31,15	43,38	4,59
■	30,85	43,38	99,57

Fuente: Elaboración propia.

Tomando la partición dada por PAM para $K = 2$ para ambos conjuntos y comparándolas, se obtiene un índice de Rand de 0.99 lo que implica que hay una gran similitud entre ambas clasificaciones.

El cuadro 17 muestra la estructura de generación para cada grupo. De nuevo hay una diferencia notable en los porcentajes de energía solar, lo cual es coherente con el conjunto de datos analizado.

Cuadro 17.

Para cada grupo definido en la figura 16, porcentaje promedio de energía programada por fuente

	Total (MWh)	*CCGT*	*Hidro*	*NGcog*	*Nuclear*	*Eólica*	*SolarFV*	*SolarT*
■	31.354,80	14,52	10,98	10,45	20,30	17,87	10,17	3,77
■	23.107,13	13,53	8,74	13,79	27,43	22,27	0,06	0,50

	Biogás	*Biomasa*	*Carbón*	*Petróleo*
■	0,29	1,13	8,68	1,02
■	0,39	1,52	10,17	1,29

Nota: La primera columna es la energía promedio por grupo.
Fuente: Elaboración propia.

Finalmente, se estudia la importancia de las variables en un procedimiento de bosque aleatorio. Dividiendo entre conjuntos de entrenamiento y de prueba se obtiene que la clasificación de las curvas del conjunto de prueba ha sido correcta en el 99,31 % de los casos. El cuadro 18 muestra la matriz de confusión, y la figura 17 muestra el diagrama de importancia de las variables.

Cuadro 18.

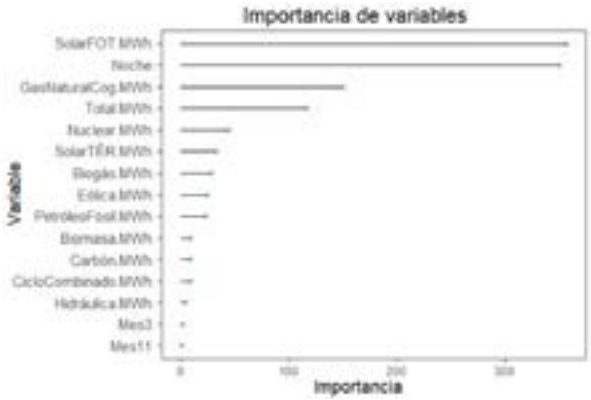
Matriz de confusión en el conjunto de prueba

	Etiqueta real	
Predicción	■	■
■	303	2
■	2	277

Fuente: Elaboración propia.

Figura 17.

Variables más importantes en la clasificación supervisada



Fuente: Elaboración propia.

La figura 17 muestra que SolarFOT y Noche son, con diferencia, las variables más relevantes para distinguir entre los dos grupos.

6.3.2. Resultados utilizando PAM para K = 3

La figura 18 representa la clasificación obtenida usando PAM para $K = 3$. En este caso, en el cuadro 19 se observa que el clúster verde tiene una mayor proporción de curvas que pertenecen a los meses comprendidos entre diciembre a abril, por lo que se espera un alto porcentaje de energía eólica. El clúster azul, por su parte, tiene una mayor proporción de curvas en julio, agosto y octubre. Por otro lado, se observa una distribución uniforme para los tres grupos al estudiar los días de la semana (ver cuadro 34 en el Apéndice).

Observando las variables binarias (cuadro 20), mientras el grupo verde muestra una menor proporción de curvas con la variable Festivo & Verano igual a 1, su proporción de días

Cuadro 19.

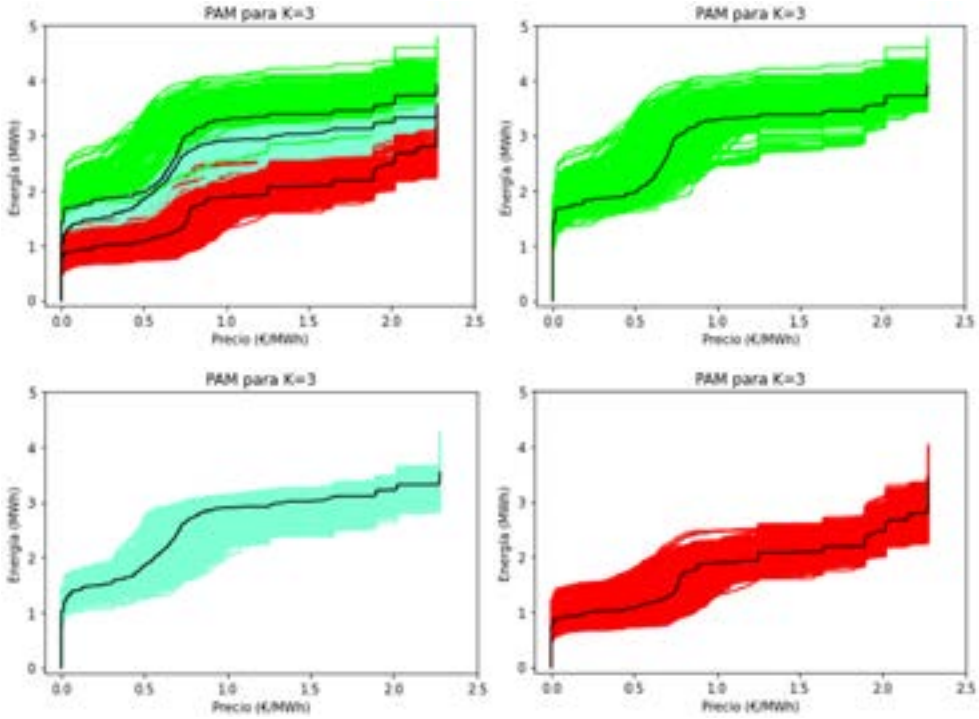
Para cada grupo definido en la figura 18, porcentaje de ofertas que corresponden a cada mes

	En.	*Feb.*	*Mar.*	*Abr.*	*May*	*Jun.*	*Jul.*	*Ag.*	*Sep.*	*Oct.*	*Nov.*	*Dic.*
■	12,05	11,73	17,10	11,40	7,98	6,03	3,26	3,09	3,42	3,75	6,84	13,36
■	7,21	5,06	4,63	5,71	8,18	8,93	11,19	11,30	1,76	11,09	8,83	7,10
■	7,76	7,76	7,25	8,48	8,92	8,70	8,99	8,99	8,63	8,85	8,41	7,25

Fuente: Elaboración propia.

Figura 18.

Curvas de oferta agrupadas mediante partición alrededor de medoides para K = 3



Nota: Las líneas en negro representan los medoides.
Fuente: Elaboración propia.

festivos es bastante similar a la de los otros dos grupos. Nuevamente, el grupo inferior (rojo, en este caso) es principalmente nocturno, mientras que los otros dos se refieren principalmente a curvas diurnas.

Cuadro 20.

Para cada grupo definido en la figura 18, porcentaje de curvas que corresponden a días no laborables (Festivo), que corresponden a un festivo o Verano (Festivo & Verano), o que corresponden a horas nocturnas (Noche)

	Festivo	*Festivo & Verano*	*Noche*
■	29,32	34,04	2,77
■	32,29	48,01	7,32
■	30,89	43,58	99,78

Fuente: Elaboración propia.

El cuadro 21 es similar al cuadro 9, donde se muestra la estructura de generación obtenida también con PAM para $K = 3$ para las curvas de 2019. Por ejemplo, en esta última, los porcentajes de energía eólica eran, del grupo superior al inferior, 28,13, 15,70 y 22,44, mientras que ahora son 25,77, 12,99 y 22,10. En general, todas las variables muestran valores similares. En conclusión, se vuelve a obtener tres grupos con las siguientes características: en primer lugar, una clara separación entre las curvas nocturnas y diurnas, y dentro de las diurnas, un grupo de ellas ofrece más energía a un precio más barato como consecuencia de la alta proporción de energía eólica.

Cuadro 21.

Para cada grupo definido en la figura 18, porcentaje promedio de energía programada por fuente

	Total (MWh)	*CCGT*	*Hidro*	*NGcog*	*Nuclear*	*Eólica*	*SolarFV*	*SolarT*
■	31.877,30	9,84	11,17	10,22	19,86	25,77	9,83	3,08
■	30.946,21	17,50	10,80	10,62	20,65	12,99	10,23	4,15
■	23.041,93	13,59	8,75	13,82	27,49	22,10	0,03	0,50

	Biogás	*Biomasa*	*Carbón*	*Petróleo*
■	0,29	1,10	6,97	0,98
■	0,30	1,15	9,78	1,04
■	0,39	1,52	10,21	1,29

Nota: La primera columna es la energía promedio por grupo.
Fuente: Elaboración propia.

Finalmente, el cuadro 22 muestra la matriz de confusión (la precisión obtenida es 91,92 %) resultante de aplicar el algoritmo de bosque aleatorio para predecir las etiquetas de los clústers usando las variables explicativas, y la figura 19 muestra el correspondiente diagrama de las 15 variables más importantes.

Cuadro 22.

Matriz de confusión en el conjunto de prueba

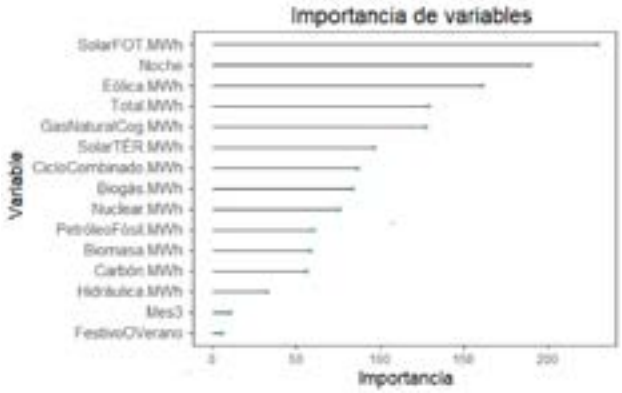
	Etiqueta real		
Predicción	■	■	■
■	98	14	0
■	23	165	3
■	1	6	272

Fuente: Elaboración propia.

La figura 19 es coherente con los hechos que se han descrito. SolarFOT y Noche vuelven a ser las variables más importantes, mientras que Eólica es la tercera.

Figura 19.

Variables más importantes en la clasificación supervisada



Fuente: Elaboración propia.

6.4. Resultados de clasificación de las curvas de horas pico y valle usando un procedimiento aglomerativo

En el cuadro 23 se muestran la media del estadístico Silueta y el índice de separación cuando se utiliza agrupamiento jerárquico aglomerativo (con enlace promedio). Nuevamente se dan valores más altos de Silueta e índice de separación cuando K es bajo. Cuando, $K = 2$ da un resultado casi igual al obtenido por PAM (índice Rand de 0.911). Para $K = 4$ se verifica la presencia de un grupo formado únicamente por cinco curvas situadas entre diciembre de 2019 y marzo de 2020, todas ellas horas valle y con un porcentaje de energía eólica entre el 30 % y el 56 %. Si $K = 5$ aparece otro clúster de cinco observaciones, siendo estas horas punta con

Cuadro 23.

Media del estadístico Silueta e índice de separación en clasificaciones obtenidas mediante un procedimiento aglomerativo con enlace promedio

K	*Media del estadístico Silueta*	*Índice de separación*
2	0,59	0,263
3	0,51	0,228
4	0,47	0,229
5	0,34	0,232
6	0,33	0,139
7	0,32	0,088
8	0,29	0,089
9	0,26	0,086
10	0,23	0,086

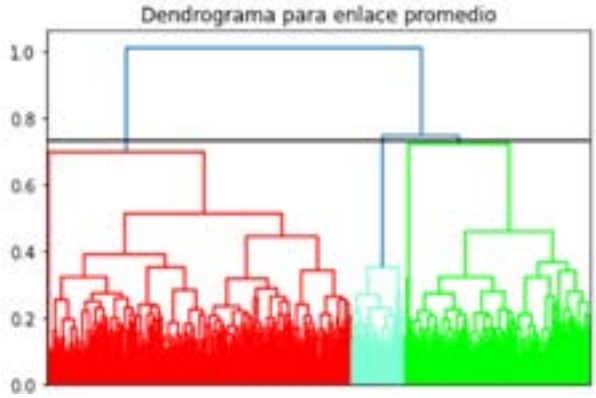
Fuente: Elaboración propia.

un bajo porcentaje de energía eólica (entre el 3 y el 12 %) cuatro de ellas en noviembre y la última en septiembre de 2020. En lo que sigue, se muestra el resultado para $K = 3$ con lo que se evitan grupos tan pequeños que podrían corresponder a días atípicos.

En la figura 20 se muestra el dendrograma y la separación correspondiente a K = 3.

Figura 20.

Dendrograma para las curvas de oferta de horas pico y valle



Nota: La línea negra corta las líneas verticales correspondientes a los tres grupos elegidos.
Fuente: Elaboración propia.

Para $K = 3$ los grupos obtenidos se muestran en la figura 21. Entre esta clasificación y la obtenida con PAM (claro, con K=3) el índice de Rand es 0.847, por lo que a pesar de algunas diferencias se esperan resultados similares en el análisis.

Figura 21.

Curvas de oferta agrupadas mediante procedimiento aglomerativo con enlace promedio para $K = 3$

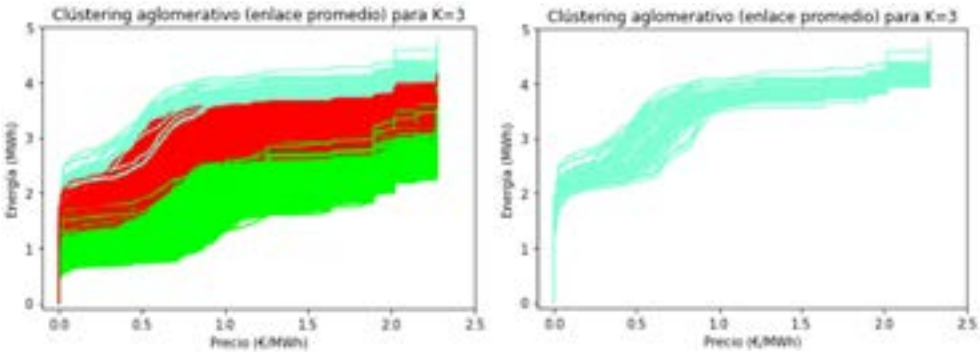
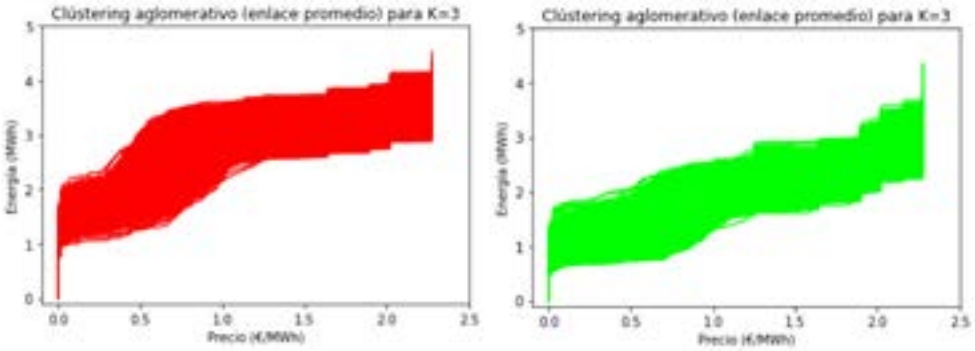


Figura 21. *(continuación)*

Curvas de oferta agrupadas mediante procedimiento aglomerativo con enlace promedio para $K = 3$



Fuente: Elaboración propia.

Primero, para el clúster azul, en su distribución en meses se observa una alta proporción de curvas correspondientes a marzo y un porcentaje nulo o muy pequeño de curvas relacionadas con los meses entre junio y octubre (ver cuadro 24). En segundo lugar, para este grupo se observa que hay una alta proporción de curvas en Lunes (cuadro 25), pero debe notarse que este grupo consta solo de 63 observaciones.

Cuadro 24.

Para cada grupo definido en la figura 21, porcentaje de ofertas que corresponden a cada mes

	En,	*Feb,*	*Mar,*	*Abr,*	*May*	*Jun,*	*Jul,*	*Ag,*	*Sep,*	*Oct,*	*Nov,*	*Dic,*
■	12,70	7,94	39,68	11,11	4,17	1,59	0,00	0,00	1,59	0,00	4,76	17,46
■	8,30	7,80	7,15	8,08	8,73	8,51	8,87	8,87	8,44	8,87	8,08	8,30
■	8,49	7,67	8,42	8,21	8,49	8,21	8,49	8,49	8,28	8,49	8,49	8,28

Fuente: Elaboración propia.

Cuadro 25.

Para cada clúster definido en la figura 21, porcentaje de ofertas que corresponden a cada día

	Lunes	*Martes*	*Miércoles*	*Jueves*	*Viernes*	*Sábado*	*Domingo*
■	23,81	14,29	12,70	11,11	14,29	9,52	14,29
■	13,88	14,23	14,31	14,38	14,23	14,59	14,38
■	14,31	14,37	14,37	14,37	14,24	14,10	14,24

Fuente: Elaboración propia.

Al observar las variables binarias en el cuadro 26, se constata que la distinción entre día y noche es casi perfecta. El grupo verde se refiere a los períodos nocturnos, mientras que los grupos rojo y azul se refieren a horas diurnas.

Cuadro 26.

Para cada grupo definido en la figura 21, porcentaje de curvas que corresponden a días no laborables (Festivo), que corresponden a un festivo o Verano (Festivo & Verano), o que corresponden a horas nocturnas (Noche)

	Festivo	*Festivo & Verano*	*Noche*
■	30,16	30,16	0,00
■	31,19	43,71	0,36
■	30,87	42,85	99,66

Fuente: Elaboración propia.

La estructura de generación se presenta en el cuadro 27. Nuevamente, el grupo superior (azul) es el que tiene la mayor proporción promedio de energía eólica. Además, para este grupo la proporción media total de energía generada por fuentes renovables (solar, eólica e hidroeléctrica) es del 61,71 %. Para el grupo rojo este porcentaje es del 37,33 % y para el verde del 23,81 %. Además, el bajo porcentaje de energía generada por ciclo combinado también diferencia al clúster azul de los otros dos.

Cuadro 27.

Para cada grupo definido en la figura 21, porcentaje promedio de energía programada por fuente

	Total (MWh)	*CCGT*	*Hidro*	*NGcog*	*Nuclear*	*Eólica*	*SolarFV*	*SolarT*
■	33.631,12	4,62	12,31	9,04	18,37	38,26	9,28	1,86
■	31.420,41	15,43	10,96	10,50	20,26	15,69	10,68	4,03
■	23.307,43	13,13	8,80	13,66	27,24	23,28	0,05	0,48

	Biogás	*Biomasa*	*Carbón*	*Petróleo*
■	0,27	1,02	3,02	0,78
■	0,29	1,13	9,16	1,03
■	0,38	1,51	9,88	1,27

Nota: La primera columna es la energía promedio programada por grupo.
Fuente: Elaboración propia.

Para la clasificación supervisada, nuevamente, se muestran los resultados para dos situaciones: la primera sin dividir los datos en conjunto de entrenamiento y prueba, y la segunda dividiendo los datos, pero sin considerar el grupo más pequeño, en este caso el azul. En el primer caso, la precisión es del 100 %, en el segundo es del 99,82 % (ver cuadro 28). Los diagramas de importancia de las variables se muestran en la figura 22.

Las variables Solar fotovoltaica y Noche vuelven a ser las más importantes en ambos diagramas, que son muy similares.

Cuadro 28.

Matrices de confusión en el conjunto completo de datos y en el conjunto de prueba con dos grupos

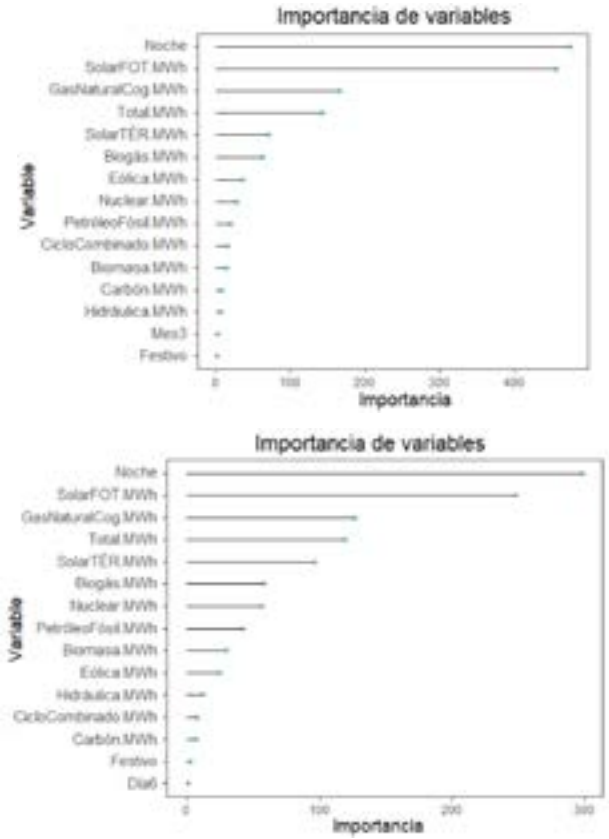
	Etiqueta real		
Predicción	■	■	■
■	63	0	0
■	0	1.398	0
■	0	0	1.461

	Etiqueta real	
Predicción	■	■
■	279	1
■	0	291

Fuente: Elaboración propia.

Figura 22.

Variables más importantes en la clasificación supervisada



Nota: El gráfico superior corresponde a la clasificación del conjunto completo con tres grupos y el gráfico inferior corresponde a la clasificación de los dos grupos mayoritarios.

Fuente: Elaboración propia.

7. CONCLUSIONES

En este proyecto, se ha encontrado una dificultad computacional notable al calcular la distancia de Hausdorff debido a la gran cantidad de cálculos iterativos que necesita, lo cual es costoso especialmente con grandes conjuntos de datos. Por ese motivo no se analizaron todas las curvas del período 2017 a 2020, y se optó por realizar un análisis para un año y para horas seleccionadas de tipo pico y valle de todo el período. Sin embargo, aun así, el estudio que se ha realizado es consistente y permite obtener conclusiones relevantes.

- En primer lugar, se han transformado los conjuntos de ofertas en curvas para poder calcular la distancia de Hausdorff entre dichos conjuntos.
- Usando esta medida de similitud para agrupar las curvas se ha observado que, en general, es preferible un valor bajo del número de grupos para obtener una buena separación entre ellos.
- No se han encontrado diferencias en las curvas de oferta respecto a los días de la semana o los días de baja actividad laboral. Si se consideran los meses del año, la distribución de las curvas cambia debido a la energía eólica.
- Este tipo de generación tiene un fuerte impacto en las curvas. Existe una clara distinción entre curvas en función de su porcentaje de energía eólica. Las curvas con un porcentaje alto hacen referencia a una gran cantidad de energía ofrecida a bajo precio.
- Además, los métodos de agrupamiento que se han utilizado también dan una separación natural entre las curvas diurnas y nocturnas.

Finalmente, una posible vía de extender esta investigación sería realizar el mismo análisis teniendo en cuenta las ofertas de demanda. Este estudio sería complementario a este proyecto y daría como resultado una comprensión más profunda de los factores que son relevantes en la fijación del precio de la energía.

Referencias

Aggarwal, S. K., Saini, L. M. y Kumar, A. (2009). Day-ahead electricity price forecasting in Victoria Electricity Market using Support Vector Machine based Model. *Power Research,* 5, pp. 37–45.

Agosti, L., Padilla, A. J. y Requejo, A. (2007). El mercado de generación eléctrica en España: Estructura, funcionamiento y resultados. *Economía Industrial,* 364, pp. 21–37.

Akhanli, S. E. y Hennig, CH. (2020). Comparing clusterings and numbers of clusters by aggregation of calibrated clustering validity indexes. *Statistics and Computing,* 30, pp. 1523– 1544.

BOE/CNMC (2021). Corrección de errores de la resolución de 6 de mayo de 2021, de la Comisión Nacional de los Mercados y la Competencia, por la que se aprueban las reglas de funcionamiento de los mercados diario e intradiario de energía eléctrica para su adaptación de los límites de oferta a los límites de casación europeos. *Boletín Oficial del Estado,* 131, pp. 67380–67543.

BOE/SEE. (2012). Resolución de 24 de julio de 2012, de la Secretaría de Estado de Energía, por la que se aprueba la modificación de los procedimientos de operación del Sistema Eléctrico Peninsular (SEP) p.o.-3.1; p.o.-3.2; p.o.-9 y p.o.-14.4 y los procedimientos de operación de los sistemas eléctricos insulares y extrapeninsulares (seie) p.o. seie-1 p.o. seie-2.2; p.o. seie-3.1; p.o. seie-7.1; p.o. seie-7.2; p.o. seie-8.2; p.o. seie-9 y p.o. seie-2.3 para su adaptación a la nueva normativa eléctrica. *Boletín Oficial del Estado,* 120, pp. 57263–57496.

Breiman, L. (2001). Random forests. *Machine Learning,* 45, pp. 5–32.

Kaufman, L. y Rousseeuw, P. (1990). *Finding Groups in Data: An Introduction To Cluster Analysis.* New York: Wiley.

Khobai, H., Mugano, G. y Le Roux, P. (2017). The impact of electricity price on economic growth in South Africa. *International Journal of Energy Economics and Policy,* 7, pp. 108–116.

Nielsen, F. (2016). Hierarchical clustering. En *Introduction to HPC with MPI for Data Science* (pp. 195–211). Springer Cham.

OMIE. Day-ahead Market bids detail. Fecha de acceso: 2022-04-10.

OMIE. Header of bids for Day-ahead Market. Fecha de acceso: 2022-04-10.

OMIE. *Modelo de Ficheros para la distribución pública de Información del mercado de electricidad. Versión 1.33.* Fecha de acceso: 2021-03-11.

Park, H.-S. y Jun, C-H. (2009). A simple and fast algorithm for k-medoids clustering. *Expert Systems and Applications,* 36, pp. 3336–3341.

Taha, A. A. y Hanbury, A. (2015). An efficient algorithm for calculating the exact Hausdorff distance. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 37, pp. 2153–2163.

Warrens, M. J. y van der Hoef, H. (2020). Understanding the Rand index. En *Advanced Studies in Classification and Data Science* (pp. 301–313). Springer.

APÉNDICE

Cuadro 29.

Para cada grupo definido en la figura 10, porcentaje de ofertas que corresponden a cada día

	Lunes	*Martes*	*Miércoles*	*Jueves*	*Viernes*	*Sábado*	*Domingo*
■	14,24	13,09	14,63	15,43	15,63	13,90	13,09
■	14,23	15,38	14,04	13,93	13,46	14,26	14,70
■	14,31	14,71	14,15	13,44	13,96	14,59	14,83

Fuente: Elaboración propia.

Cuadro 30.

Para cada grupo definido en la figura 14, porcentaje de ofertas que corresponden a cada mes

	En,	*Feb,*	*Mar,*	*Abr,*	*May*	*Jun,*	*Jul,*	*Ag,*	*Sep,*	*Oct,*	*Nov,*	*Dic,*
■	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	100,00
■	8,41	7,80	8,51	8,13	8,39	8,03	8,39	8,39	8,18	8,42	8,52	8,83
■	8,76	7,44	8,57	8,49	8,80	8,72	8,80	8,80	8,38	8,68	7,59	6,98

Fuente: Elaboración propia.

Cuadro 31.

Para cada grupo definido en la figura 14, porcentaje de ofertas que corresponden a cada día

	Lunes	*Martes*	*Miércoles*	*Jueves*	*Viernes*	*Sábado*	*Domingo*
■	0,00	0,00	0,00	0,00	0,00	61,90	38,10
■	14,33	14,58	14,35	14,33	14,53	13,97	13,89
■	14,19	14,46	14,12	14,12	13,70	14,50	14,91

Fuente: Elaboración propia.

Cuadro 32.

Para cada grupo definido en la figura 16, porcentaje de ofertas que corresponden a cada mes

	En,	*Feb,*	*Mar,*	*Abr,*	*May*	*Jun,*	*Jul,*	*Ag,*	*Sep,*	*Oct,*	*Nov,*	*Dic,*
■	8,85	7,74	9,57	8,07	8,20	7,87	8,13	8,13	7,87	8,20	7,87	9,51
■	8,09	7,73	7,30	8,38	8,80	8,59	8,88	8,88	8,59	8,80	8,59	7,37

Fuente: Elaboración propia.

Cuadro 33.

Para cada grupo definido en la figura 16, porcentaje de ofertas que corresponden a cada día

	Lunes	*Martes*	*Miércoles*	*Jueves*	*Viernes*	*Sábado*	*Domingo*
■	14,36	14,23	13,97	14,36	14,43	14,36	14,30
■	14,24	14,39	14,67	14,24	14,03	14,10	14,32

Fuente: Elaboración propia.

Cuadro 34.

Para cada grupo definido en la figura 18, porcentaje de ofertas que corresponden a cada día

	Lunes	*Martes*	*Miércoles*	*Jueves*	*Viernes*	*Sábado*	*Domingo*
■	16,29	13,84	13,52	14,50	14,66	14,66	12,54
■	13,24	14,42	14,10	14,42	14,42	14,99	15,39
■	14,14	14,43	14,79	14,14	13,92	14,21	13,36

Fuente: Elaboración propia.