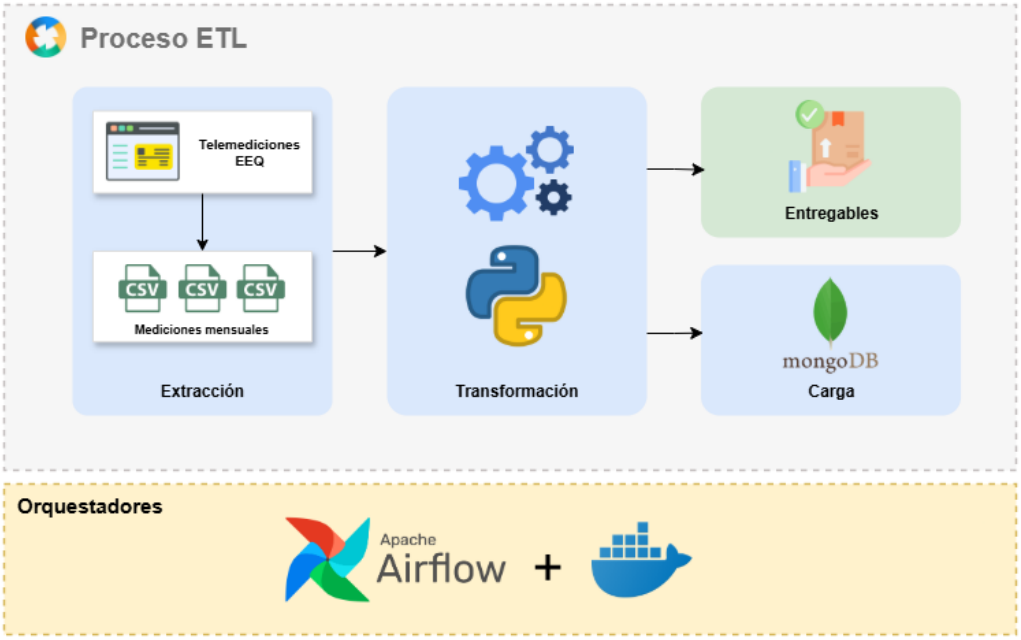
SEGMENTACIÓN DE CLIENTES NO REGULADOS DEL SECTOR ELÉCTRICO MEDIANTE APRENDIZAJE NO SUPERVISADO

Introducción

Este documento da cuenta de la integración y la organización de forma coherente de todos los entregables generados en el desarrollo del proyecto de segmentación de clientes no regulados del sector eléctrico. El objetivo general del trabajo fue caracterizar los patrones de consumo eléctrico de un conjunto de clientes no regulados de una empresa distribuidora, construyendo anualidades de curvas de carga representativas y aplicando algoritmos de aprendizaje no supervisado que agrupen los clientes en función de la forma de las curvas. La metodología que se utilizó en el trabajo fue CRISP DM; no obstante, se ha modificado la fase final en tanto que, en lugar de desplegar un modelo en producción, se decide escoger el algoritmo de agrupamiento más adecuado en función de las métricas internas de validación y la interpretabilidad de los clústeres. En línea con lo que establece la guía CRISP DM, al final del trabajo se tiene que elaborar un reporte final donde conste el proceso, el resumen de los resultados y la presentación de los productos generados.

Proceso ETL (Extracción, Transformación y Carga)

La calidad de los clústeres generados depende, en gran medida, de la preparación de datos. El proyecto se llevó a cabo partiendo de dos grupos de clientes, los cuales poseen formatos de medición diferentes; Grupo 1, que tiene archivos de medición mensuales con Demanda activa DEL y Demanda reactiva DEL y Grupo 2, donde los archivos de medición tienen energía aparente acumulada. Por cada grupo de clientes, se generó un conjunto de tareas ETL que fueron orquestadas en Apache Airflow. La imagen siguiente es un resumen del flujo general de este proceso:



Extracción de los datos.

- Grupo 1: Se recorren los archivos de medición para los diferentes clientes de forma que podamos extraer las columnas en formato: Fecha, Demanda activa DEL y Demanda reactiva DEL; determinando posteriormente los clientes únicos y generándose un diccionario que relaciona el identificador de cliente con su conjunto de medición anual.
- Grupo 2: Recorremos las carpetas que forman la información de cada cliente y concatenamos las mediciones mensuales (que incluyen la columna Fecha y la columna AS (kWh)). Las fechas son transformadas a formato cadena estrictamente para facilitar la limpieza.

Transformación de los datos

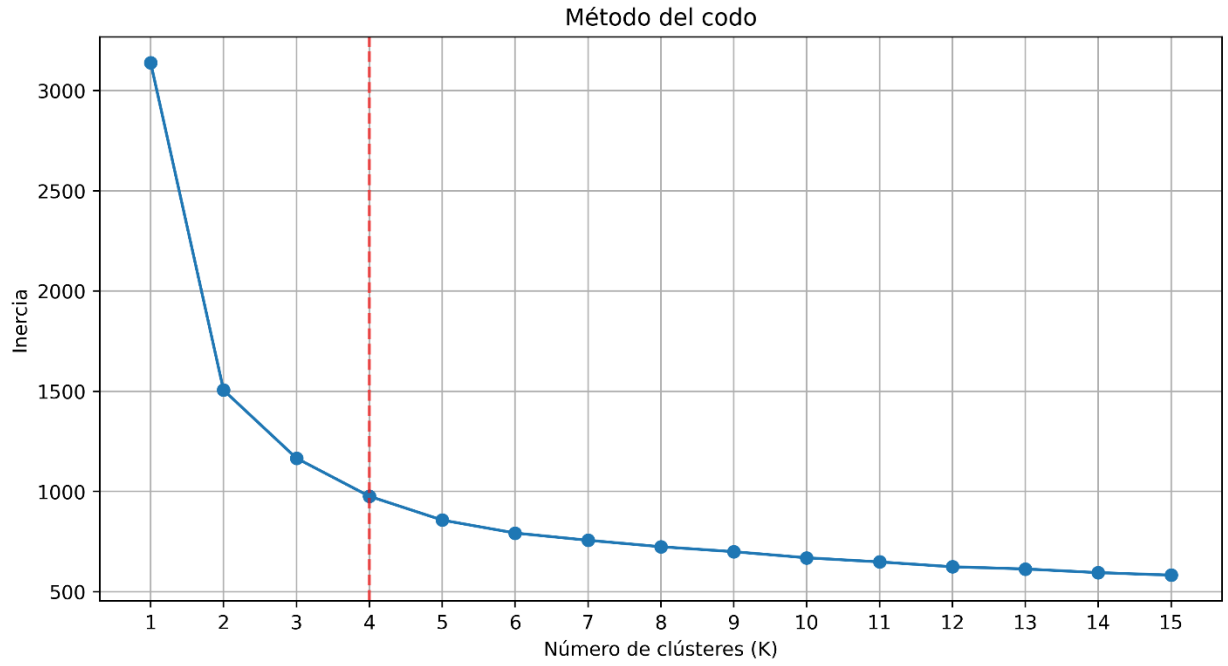
- La columna correspondiente a la Fecha se transforma en el formato único aaaa/mm/dd y se separa de la hora, generando así una columna 'Fecha' que contiene únicamente dicho dato en un formato estandarizado, y una columna 'Hora'.
- Los valores de potencia aparente de los registros vacíos son interpolados mediante interpolación cúbica con el objetivo de minimizar huecos. En el Grupo 1 se calcula la potencia aparente mediante la aplicación del teorema de Pitágoras entre la demanda activa y la demanda reactiva. En el Grupo 2 la energía aparente AS (kWh) se multiplica por 4 obteniendo un valor para cada 15 minutos.
- Los datos faltantes son rellenados haciendo uso de un interpolador de tipo spline cúbico, la decisión de usar este interpolador radica en su relevancia en la literatura.
- Se eliminan fines de semana y feriados nacionales para obtener curvas representativas de los días laborables. Cada uno de los días fue escalado entre 0 y 1 para facilitar la comparación entre clientes.
- Se procede a construir curvas de carga representativas de cada cliente agrupando las mediciones de este y obteniendo la curva promedio diaria (muestreo cada 15 min), que es a su vez la curva que tiene 96 puntos y que representa la forma habitual del perfil de cada uno de los clientes.
- Cada día fue escalado de manera individual, con el objetivo de impedir que días cuyas demandas tienen un valor más alto 'aplanen' a los demás días, el escalado individual permite conservar la forma base de la curva de cada día.

Unificación y carga.

Las curvas de los dos conjuntos de clientes se unifican en un único DataFrame, donde hay una columna del Cliente y las 96 mediciones de la curva representativa de cada cliente. Este DataFrame unificado se guarda en la base de datos MongoDB para su posterior consulta y se generan el entregable de cada cliente (curva tipo en formato CSV, curva del día de demanda máxima, gráficos y archivo de las potencias máxima y mínima), generando de este modo el entregable de cada uno de los clientes. El proceso ETL automatizado garantiza la reproducibilidad del proceso cada vez que se requiera.

Modelado y agrupación de curvas de consumo

Una vez confeccionadas las curvas que se utilizan como representativas, se seleccionaron distintos algoritmos de clustering para llevar a cabo los agrupamientos de los clientes en función de la similitud de la forma de sus curvas, es decir, se usaron K Means, Gaussian Mixture Models (GMM), Birch y Spectral Clustering. Para tal efecto se calculó el número óptimo de clústeres dependiendo del algoritmo que se estuviera utilizando, mediante el uso de la técnica del codo que mide el nivel de reducción de inercia al aumentar el número de clústeres. En el caso de K Means, la ruptura del codo se produjo alrededor de K=4, tal y como se muestra a continuación:



Optimización de los parámetros de los algoritmos

Se llevó a cabo un proceso de optimización de hiperparámetros para cada uno de los algoritmos de clustering evaluados. Dicho proceso consistió en analizar distintas configuraciones de los parámetros más relevantes de cada técnica, con el objetivo de identificar aquellas que maximizaran el valor de la correlación intra-clúster. Esta métrica fue definida como el criterio principal de evaluación, dado que cuantifica directamente el grado de semejanza entre las curvas de consumo que conforman cada grupo, constituyéndose así en un indicador adecuado para evaluar la calidad y coherencia de los clústeres generados por cada algoritmo.

Métricas de evaluación

Para poder comparar la calidad de los agrupamientos se calcularon varias métricas internas:

- Puntaje de silueta: Evalúa hasta qué punto cada centroide se ubica mejor en su propio clúster en relación con los demás. Para cada centroide se computa la distancia del centroide al resto de los centroides del clúster, que se denomina cohesión, y la distancia

del centroide al resto de los centroides del clúster más similar, que se denomina separación. La puntuación se ubica entre −1 y 1; un valor próximo a 1 indica un buen agrupamiento.

- Índice de Davies-Bouldin: Establece la similitud de cada clúster con su clúster más similar como la relación entre la dispersión interna y la distancia entre centroides. El índice total es la media de estas similitudes; cuanto menor es, mejores son la distancia entre los clústeres y la compactación de los clústeres entre sí.
- Índice de Calinski-Harabasz: También llamado criterio en relación de varianzas compara la dispersión entre clústeres con la dispersión dentro de los clústeres. Cuanto mayor es el valor, mejor desarrolla el agrupamiento, porque esto indica que el agrupamiento es denso y que los clústeres están bien separados.
- Correlación intra-clúster promedio: La correlación de Pearson fue estimada a partir de las curvas individuales de cada cliente y la curva promedio del clúster dentro del que ese cliente se encontraba; hemos tomado una mayor correlación como indicador de mayor homogeneidad.

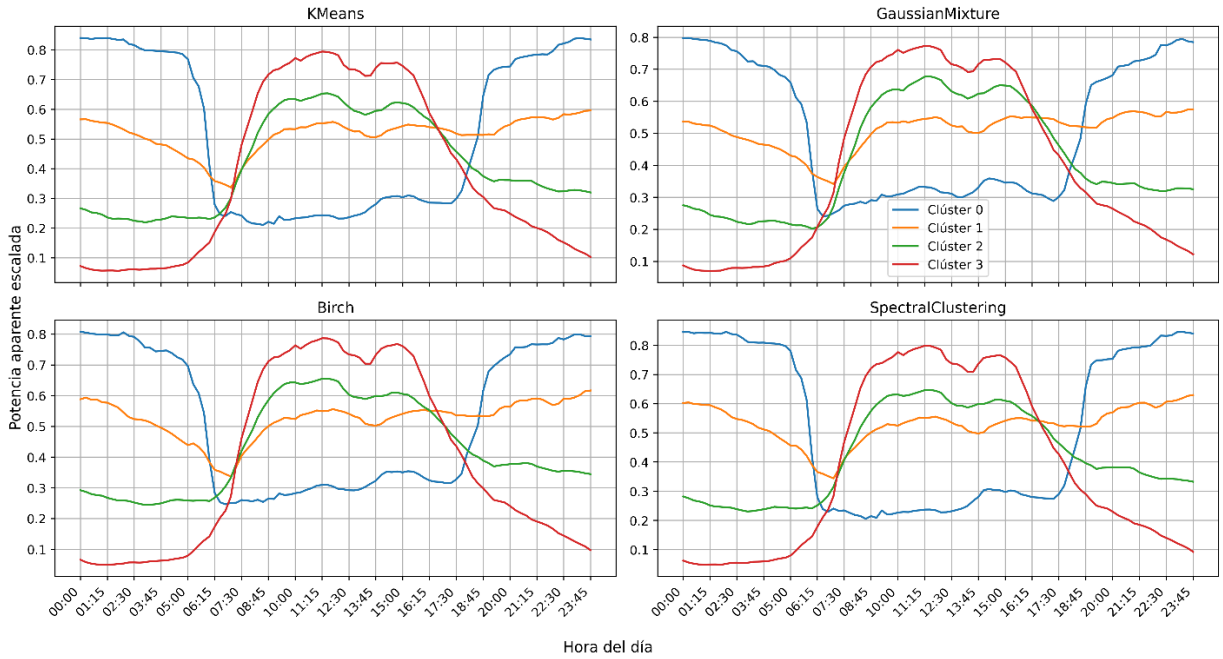
La siguiente tabla incluye la información de algunos de los resultados obtenidos. Los mejores valores de cada una de las métricas se marcan en negrita.

Métrica	K-Means	Gaussian Mixture	Birch	Spectral Clustering
Correlación intra-clúster promedio	0.7399	0.7394	0.7515	**0.7533**
Índice de Davies-Bouldin	1.1791	1.2306	1.2177	**1.1536**
Índice de Calinski-Harabasz	**283.5028**	261.7910	253.3311	270.1244
Puntaje de silueta	**0.2624**	0.2140	0.2326	0.2590

Los resultados indican que K Means obtiene la mayor puntuación en Silhouette y una máxima puntuación en el índice de Calinski Harabasz, insinuando clústeres relativamente distanciados. Los algoritmos Birch y Spectral Clustering alcanzan la máxima correlación intra-clúster, que era la prioridad de este estudio, es decir, maximizar la similitud en la forma de las curvas entre clientes pertenecientes a un mismo grupo; el índice de Davies Bouldin es un poco menor para el caso de Spectral Clustering, sugiriendo clústeres un poco más compactos.

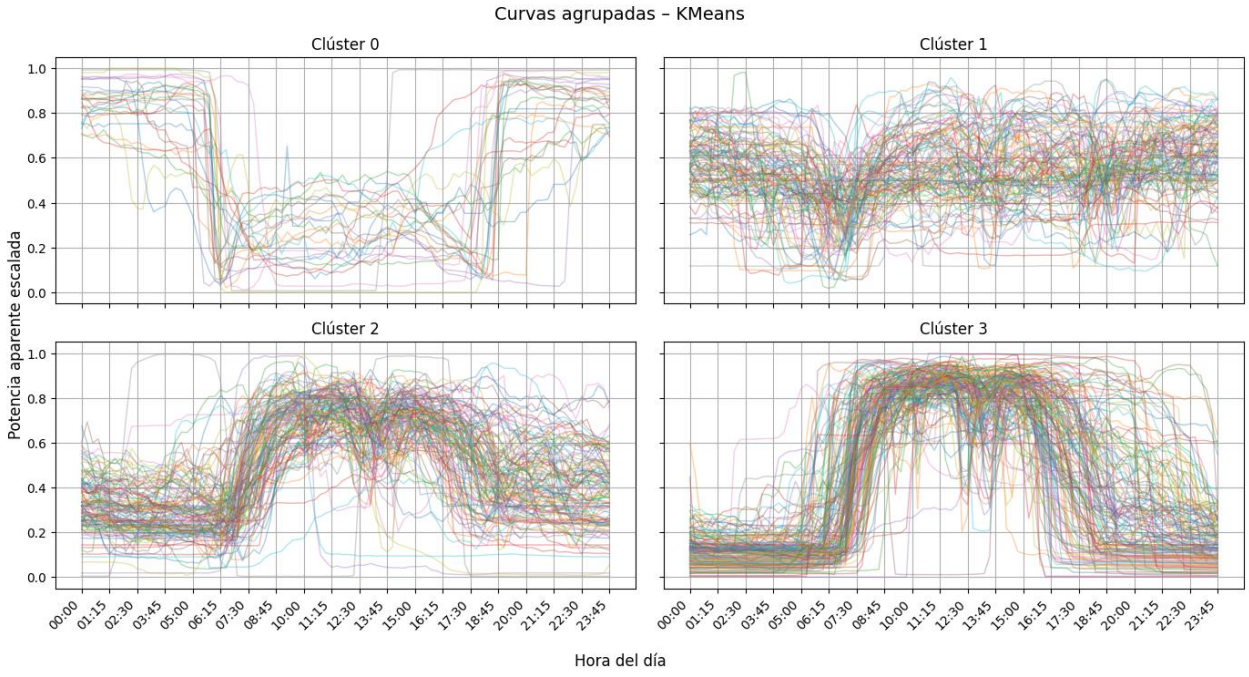
Representación gráfica de los clústeres

Para interpretar los clústeres obtenidos se adquirieron las curvas promedio de cada agrupamiento y se compararon las curvas del cliente y sus respectivas medias. En la siguiente, se muestran los patrones característicos en curva de los cuatro algoritmos en las que cada color representa un clúster diferente y la variable en el eje Y es la potencia aparente escalada (0-1) por hora del día:

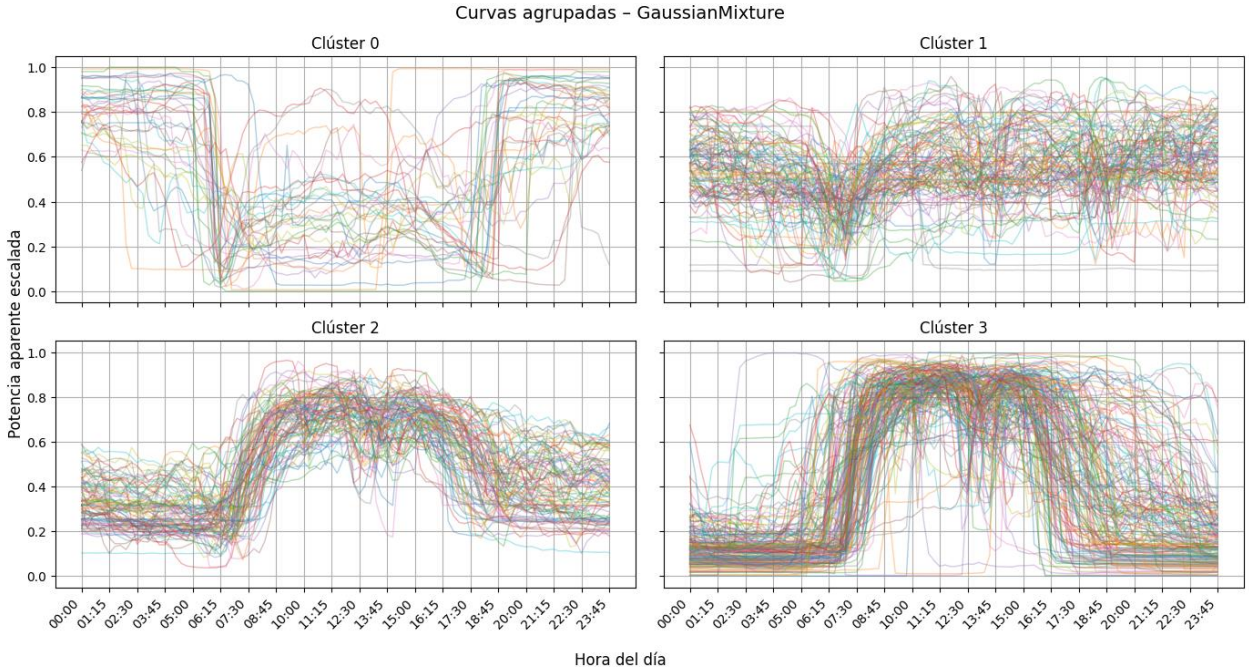


Una representación más concreta de las curvas individuales permite ver la dispersión interna. Las siguientes gráficas representan todas las curvas por clúster y algoritmo:

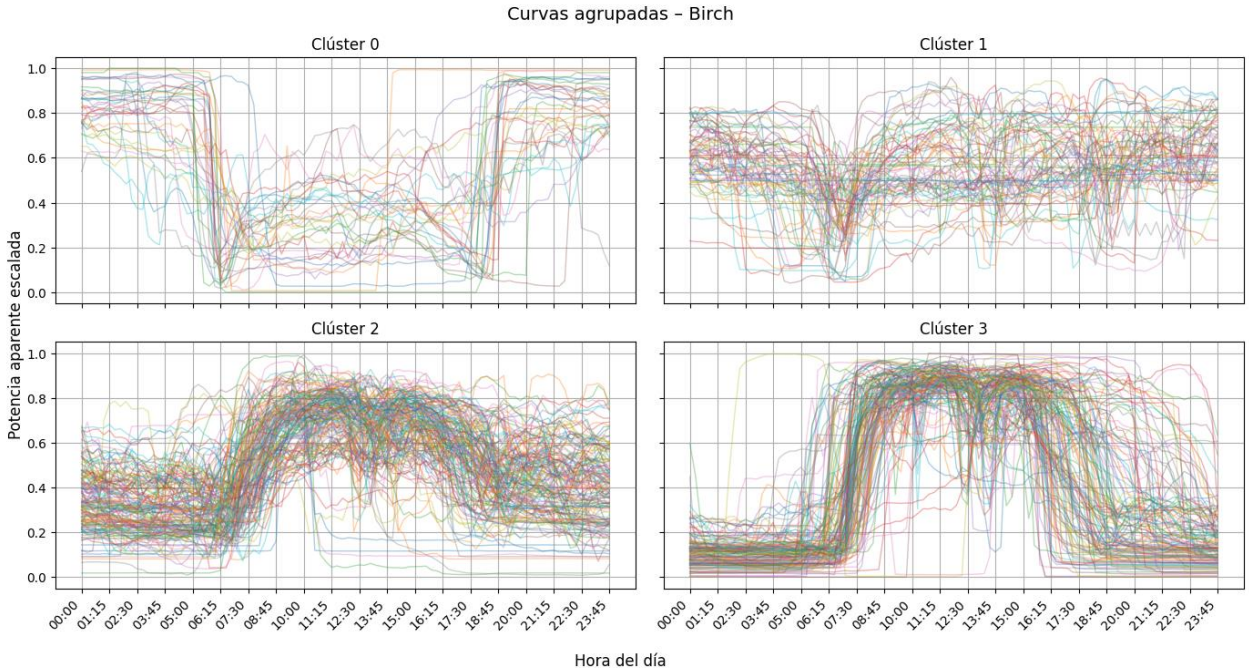
K Means



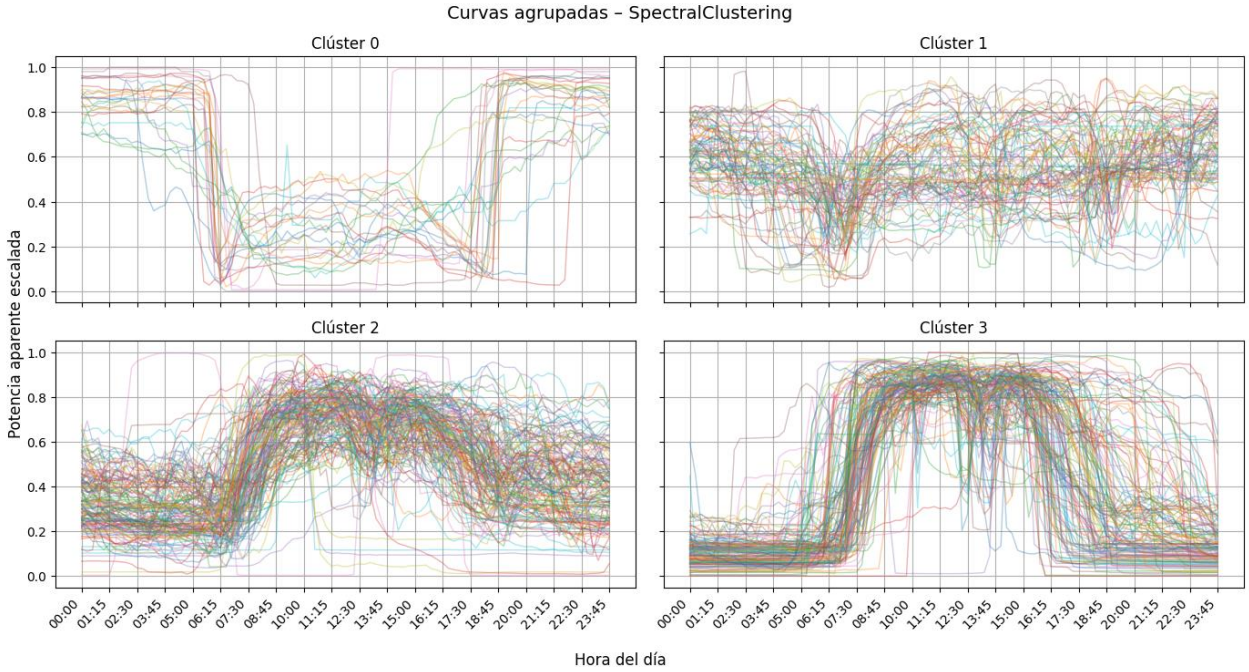
Gaussian Mixture



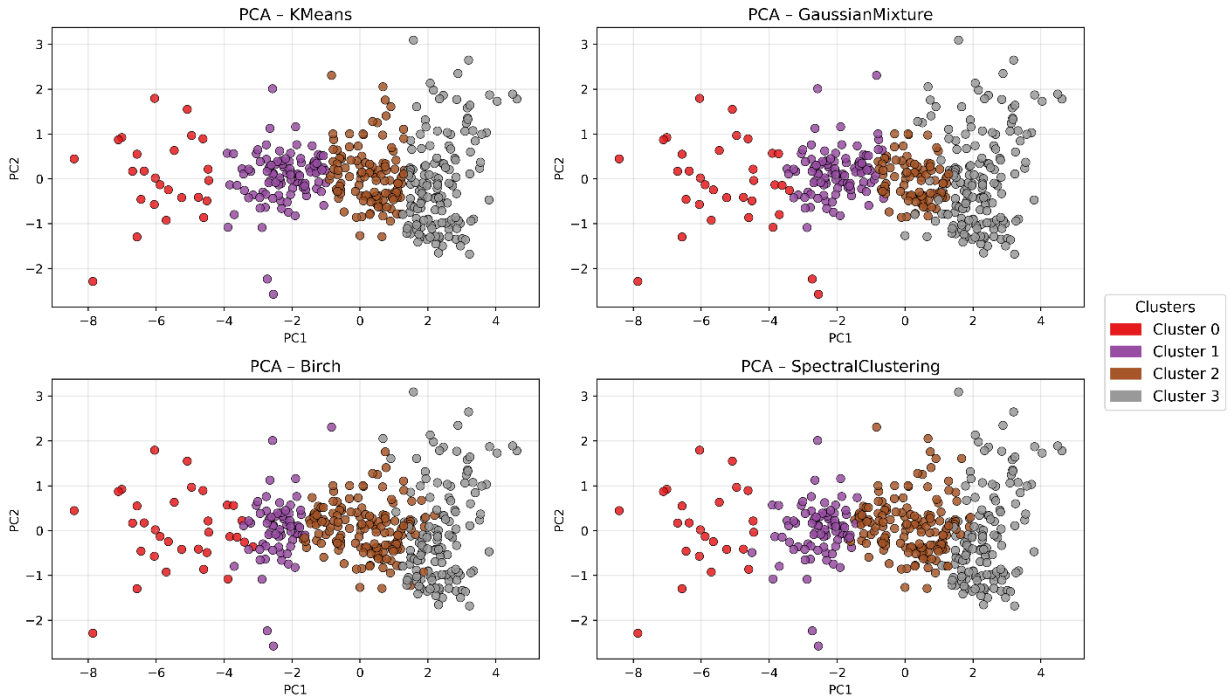
Birch



Spectral Clustering



Finalmente, y a modo de observar la separación de los grupos en el espacio de menor dimensión, se le aplica PCA a las curvas normalizadas y se representan los dos primeros componentes principales. Finalmente, como cada punto supone un cliente y los diferentes colores son claves para indicar el clúster al que se asignan:



Discusión de resultados y selección del modelo final

En base a los resultados cuantitativos (correlación intra-clúster, así como otros factores complementarios) y el análisis cualitativo (curvas tipo y proyección PCA) se escoge el algoritmo Spectral Clustering como el modelo final del componente. Esta elección se produce en base a su mayor correlación intra-clúster media (una valoración que analiza de una manera más directa la similitud estructural de las curvas).

Siguiendo esta observación, a continuación, se muestra la interpretación energética de cada clúster que ha sido identificado a través de Spectral Clustering:

- Clúster 0: consumo intensivo durante las primeras horas de la madrugada y en la noche. Se caracteriza por un comportamiento típico de clientes que trabajan fundamentalmente en la franja horaria nocturna; posiblemente estamos hablando de industrias de procesos continuos.
- Clúster 1: consumo relativamente poco intenso a la largo de la jornada. Resulta un consumo estable, aunque se produce una ligera disminución de este en las primeras horas; en este sentido se corresponde con empresas que supuestamente funcionarían de forma continua, pero menos intensamente.
- Clúster 2: consumo relativamente bajo en la madrugada seguido de un aumento notorio hacia media mañana que se sostiene hasta entrada la tarde, y representa probablemente a clientes con actividades diurnas que se agrupan en la franja horaria laboral (oficinas y comercios).
- Clúster 3: curvas con picos muy acusados en las horas centrales del día y consumo prácticamente nulo fuera de dicho rango; clientes con un tipo de operación intensiva en los horarios de carga punta (por ejemplo, fábricas o plantas de producción con turnos diurnos).

Esto permite a la distribuidora desarrollar estrategias de planificación y eficiencia diferenciadas: optimizar la gestión de la demanda, diseñar tarifas específicas, así como priorizar inversiones en infraestructuras para los clústeres de mayor repercusión en las horas de carga punta.

Conclusiones y recomendaciones

El trabajo evidenció la importancia de realizar un proceso de ETL en conjunto con técnicas de análisis no supervisado para poder obtener evidencias sobre los patrones del consumo eléctrico. La automatización mediante Apache Airflow y la construcción de curvas de carga representativas permitieron procesar grandes volúmenes de datos heterogéneos y obtener información útil para la toma de decisiones. La comparación de diferentes algoritmos de clustering pone de manifiesto que Spectral Clustering proporciona un equilibrio suficiente en cuanto a homogeneidad interna y simplicidad entre los clústeres para este caso de estudio, obteniendo el valor más alto de la métrica definida como principal, la correlación intra-clúster.

Se recomienda que para trabajos futuros se usen análisis mucho más avanzados, como estudios de proyección de la demanda, identificación de clientes con comportamientos atípicos o la aplicación de técnicas de aprendizaje profundo.

De la misma manera, la integración de variables adicionales para el análisis, como por ejemplo la actividad económica del cliente, información relacionada con el clima o algunos indicadores socioeconómicos, manteniendo como línea base el esquema de agrupación basado exclusivamente en la forma de la curva.