

ESCUELA POLITÉCNICA NACIONAL

FACULTAD DE INGENIERÍA EN SISTEMAS

**OPTIMIZACIÓN DE SISTEMAS DE INFORMACIÓN EN
CONTEXTO EMPRESARIALES**

**ANÁLISIS Y SEGMENTACIÓN DE CLIENTES NO REGULADOS
DEL SECTOR ELÉCTRICO MEDIANTE ALGORITMOS DE
APRENDIZAJE NO SUPERVISADO**

**TRABAJO DE INTEGRACIÓN CURRICULAR PRESENTADO COMO
REQUISITO PARA LA OBTENCIÓN DEL TÍTULO DE INGENIERO EN
CIENCIAS DE LA COMPUTACIÓN**

ANDRÉS ANTONIO ZAMBRANO ALQUINGA
andres.zambrano03@epn.edu.ec

DIRECTOR: JOSAFÁ DE JESÚS AGUIAR PONTES
josafa.aguiar@epn.edu.ec

DMQ, enero 2026

CERTIFICACIONES

Yo, ANDRÉS ANTONIO ZAMBRANO ALQUINGA, declaro que el trabajo de integración curricular aquí descrito es de mi autoría; que no ha sido previamente presentado para ningún grado o calificación profesional; y, que he consultado las referencias bibliográficas que se incluyen en este documento.

ANDRÉS ANTONIO ZAMBRANO ALQUINGA

Certifico que el presente trabajo de integración curricular fue desarrollado por ANDRÉS ANTONIO ZAMBRANO ALQUINGA, bajo mi supervisión.

JOSAFÁ DE JESÚS AGUIAR PONTES
DIRECTOR

DECLARACIÓN DE AUTORÍA

A través de la presente declaración, afirmamos que el trabajo de integración curricular aquí descrito, así como el (los) producto(s) resultante(s) del mismo, son públicos y estarán a disposición de la comunidad a través del repositorio institucional de la Escuela Politécnica Nacional; sin embargo, la titularidad de los derechos patrimoniales nos corresponde a los autores que hemos contribuido en el desarrollo del presente trabajo; observando para el efecto las disposiciones establecidas por el órgano competente en propiedad intelectual, la normativa interna y demás normas.

**ANDRÉS ANTONIO ZAMBRANO ALQUINGA
JOSAFÁ DE JESÚS AGUIAR PONTES
BORIS ALFONSO ASTUDILLO ESPINOZA**

DEDICATORIA

Quiero dedicar este logro a los dos pilares fundamentales de mi vida, mis padres, Verito y Marco, quienes a pesar de todas las dificultades que se presentaron a lo largo del camino, nunca dudaron de mí, y en su lugar, siempre supieron alentarme y darme su apoyo incondicional para seguir adelante, este y todos mis logros se los dedico a ustedes.

A Edita Vélez, quien me cuidó desde mis primeros pasos con la entrega de una madre. Dedico este logro a la mujer que transformó mi infancia en un recuerdo feliz y que me amó como a un hijo.

A mi primo Jhonny Sánchez, el hermano que nunca pude tener, pero que la vida se encargó de darme. Dedico este logro a los años que hemos compartido juntos, a nuestra complicidad desde la niñez y a cada momento invaluable que construyó el lazo que hoy nos une.

A mis padrinos, Franklin Vásquez y Silvana Barba, les dedico este logro por haberme acogido con tanto cariño en su hogar durante mis estudios universitarios. De igual manera, a mis primos Carolina, Dennis y Pamela, quienes más que primos han sido hermanos para mí.

A la memoria de mis abuelitos, Teresa y Manuel, quienes a pesar de ya no estar físicamente conmigo, sigo sintiendo su amor y protección en cada paso que doy. Sin duda alguna, este logro no habría sido posible sin su entrega y sacrificio.

A toda mi familia en general, a quienes dedico este logro por haber contribuido de manera directa o indirecta con su granito de arena para forjar la persona que soy hoy en día.

A mis dos peluditos, Rockie y Merlín, especialmente a mi gordo, Merlín, mi más linda compañía durante mi transición por propedéutico, pasó largas noches de vela a mi lado brindándome de su cálida compañía mientras yo estudiaba.

Finalmente, este logro va dedicado a mi versión más pequeña, ese niño pequeñito que nunca dejó de soñar, ese niño que a pesar de todas las adversidades que se presentaron a lo largo del camino nunca se rindió, ese niño que tuvo que enfrentarse a muchas cosas cuando no estaba preparado. Hoy por fin deposito este triunfo en tus pequeñas manos, un triunfo que lleva tu nombre, mi nombre... nuestro nombre.

AGRADECIMIENTOS

Agradezco en primer lugar, a Dios y a la Virgen María por no desampararme nunca en ninguna etapa de mi vida, por haberme guiado en cada momento, y por empaparme de sabiduría durante toda mi transición por la universidad.

A mi madre, Verito, la reina de mi vida, eres la bendición más grande que Dios me ha dado, gracias por tu inmenso amor, por todo lo que has hecho y sigues haciendo por mí, por enseñarme que todo se puede lograr con esfuerzo y dedicación. Espero que nunca pienses que fracasaste como madre, porque yo podría escribirte un libro entero contando lo bien que lo has hecho.

A mi padre, Marco, por haber estado siempre pendiente de mí, por cumplir mis caprichos, por tu cariño incondicional y tus valiosos consejos. Gracias por inculcarme el amor a esta profesión tan bonita que es la informática y por ser mi primer maestro. Si algún día la vida me lo permite, espero poder darle a mis hijos un padre como el mío.

Quiero expresar mi más profundo agradecimiento al Ing. Boris Astudillo por su invaluable orientación, sus sabios consejos y por su constante guía y apoyo a lo largo de mi formación universitaria, en particular durante el desarrollo de mi proyecto de titulación.

Quiero agradecer de manera muy especial a mi prima Carolina Vásquez por todo lo que ha hecho por mí. Gracias Carito por ser una guía indispensable y un apoyo incondicional en mi vida, eres como una hermana para mí.

A mi alma máter, la Escuela Politécnica Nacional y a los docentes que contribuyeron a mi formación académica, por brindarme todos los conocimientos y las herramientas necesarias para desarrollarme como profesional.

Quiero agradecer a todo el equipo de la Empresa Eléctrica Quito, por su apoyo y guía durante el desarrollo de mis prácticas preprofesionales, en especial a los ingenieros e ingenieras Carolina, William, Oscar, Claudia, Isabel y Grace. Agradezco de igual manera al ingeniero Ricardo Dávila por brindarme la confianza y la oportunidad de vivir esta experiencia invaluable para mi desarrollo profesional.

Finalmente, quiero agradecer a mis amigos Alexis, Carlos, Hernán, Galo, Dilan, Salito y los que faltan por nombrar, por hacer que la vida universitaria fuera mucho más llevadera, por todas las experiencias compartidas; desde las risas y charlas interminables, hasta los enojos y tristezas.

ÍNDICE DE CONTENIDO

CERTIFICACIONES	I
DECLARACIÓN DE AUTORÍA	II
DEDICATORIA	III
AGRADECIMIENTOS	IV
ÍNDICE DE CONTENIDO	V
RESUMEN	VI
ABSTRACT	VII
1. DESCRIPCIÓN DEL COMPONENTE DESARROLLADO	1
1.1. Objetivo general	2
1.2. Objetivos específicos	2
1.3. Alcance	3
1.4. Marco Teórico	4
2. METODOLOGÍA	11
2.1. Caso de estudio	11
2.2. Brainstorming	11
2.3. CRISP-DM	12
2.4. Implementación de CRISP-DM	14
3. RESULTADOS, CONCLUSIONES Y RECOMENDACIONES	44
3.1. Resultados	44
3.2. Conclusiones	50
3.3. Recomendaciones	51
4. REFERENCIAS BIBLIOGRÁFICAS	53
5. ANEXOS	58
I. Archivo docker-compose.yaml	58
II. Código Python del DAG del proceso ETL	58
III. Cuaderno de trabajo de clustering	58
IV. Reporte final del proyecto de minería de datos	58

RESUMEN

Este Trabajo de Integración Curricular aborda un proyecto de minería de datos enfocado en la implementación de un algoritmo de aprendizaje no supervisado para segmentar clientes en grupos homogéneos a partir de sus curvas características de consumo anual. El objetivo es identificar patrones de consumo energético que permitan una planificación más eficiente y una optimización del uso de la energía en el sector eléctrico.

La metodología aplicada es CRISP-DM, con una modificación en su fase final. Dentro de la misma, se han planteado dos procesos claves a seguir: en primer lugar, se desarrolla un proceso ETL (Extraer, Transformar, Cargar) orquestado por Apache Airflow, para la consolidación y transformación de los datos mensuales en una curva característica representativa anual por cada cliente, posteriormente, en el proceso de agrupación, se seleccionan y optimizan varios algoritmos para agrupar a los clientes en base a la similitud de sus curvas de consumo.

Los resultados de cada algoritmo son evaluados mediante diversas métricas, que cuantifican la calidad de las agrupaciones, con el fin de determinar el algoritmo que ofrece las agrupaciones de mejor calidad. Los resultados de agrupación serán presentados de manera visual y cuantitativa.

Palabras clave: minería de datos, segmentación de clientes, curvas de consumo, aprendizaje no supervisado, algoritmos de clustering, planificación energética, proceso ETL, Apache Airflow, CRISP-DM.

ABSTRACT

This Curriculum Integration Project focuses on a data mining project aimed at implementing an unsupervised learning algorithm to segment clients into homogeneous groups based on their annual characteristic consumption curves. The goal is to identify energy consumption patterns that allow for more efficient planning and optimization of energy use in the electric sector.

The methodology applied is CRISP-DM, with a modification in its final phase. Within this framework, two key processes are followed: first, an ETL (Extract, Transform, Load) process orchestrated by Apache Airflow is developed to consolidate and transform monthly data into an annual representative characteristic curve for each client. Then, in the grouping process, several algorithms are selected and optimized to group clients based on the similarity of their consumption curves.

The results of each algorithm are evaluated using various metrics that quantify the quality of the groupings, in order to determine which algorithm provides the highest-quality groupings. The grouping results will be presented both visually and quantitatively.

Keywords: data mining, customer segmentation, consumption curves, unsupervised learning, clustering algorithms, energy planning, ETL process, Apache Airflow, CRISP-DM.

1. DESCRIPCIÓN DEL COMPONENTE DESARROLLADO

En el contexto actual de las empresas proveedoras de energía, como la Empresa Eléctrica Quito (EEQ), la eficiente gestión energética es uno de los principales desafíos a enfrentar. Múltiples factores como la diversificación en los hábitos de consumo y variabilidad de la demanda dificultan la planificación y diseño de estrategias eficientes que permitan responder de manera adecuada. Los métodos tradicionales de análisis, que se basan en promedios o clasificaciones rígidas resultan insuficientes para capturar dicha complejidad en los patrones de consumo de los clientes, dificultando el diseño de una planificación energética eficiente.

Con el fin de optimizar la distribución de recursos en áreas como la gestión tarifaria y la distribución eléctrica, es fundamental analizar los patrones de consumo. La identificación de estos patrones en el comportamiento energético de los clientes brinda la posibilidad de definir segmentos con características similares, permitiendo a las compañías proveedoras de energía establecer una base más firme para la toma de decisiones.

Ante esta problemática, se ha desarrollado un componente orientado a la segmentación inteligente de clientes, implementando un proyecto de minería de datos que propone un enfoque basado en técnicas de aprendizaje no supervisado con el fin de generar grupos homogéneos en función de la forma de su curva característica anual de consumo energético. El objetivo principal es identificar patrones de consumo que permitan una planificación más eficiente y optimización del uso de la energía en el sector eléctrico.

Bajo este contexto, el desarrollo del componente es realizado bajo la metodología CRISP-DM, con una ligera modificación en su fase final. Mientras que en la metodología original la fase final se centra en la implementación y despliegue del modelo, en este caso, el objetivo final es, entre todas las agrupaciones dadas por los diferentes algoritmos, escoger aquella que tenga la mejor calidad y homogeneidad, basándose en métricas de evaluación. Esta modificación de la fase final es posible debido a que CRISP-DM es sumamente flexible, y permite personalizar sus fases en función de los objetivos del proyecto.

Dentro del flujo de trabajo estructurado que propone la metodología CRISP-DM, se han definido dos procesos claves: en primer lugar, se lleva a cabo un proceso de Extracción, Transformación y Carga (ETL), orquestado por Apache Airflow, para

normalizar y consolidar los datos de consumo mensual de cada cliente en una curva representativa anual.

Posteriormente, se desarrolla el proceso de agrupación, donde se determina el número óptimo de grupos de clientes a través de un análisis conjunto con las partes interesadas y el uso de métodos de validación como el método del codo. Se implementan y optimizan diferentes algoritmos de clustering, como KMeans, GaussianMixture, Birch y Spectral Clustering, para segmentar a los clientes en base a la similitud de sus curvas de consumo. Finalmente, se evalúan los resultados de cada algoritmo utilizando diversas métricas, como Silhouette Score, SSE, Davies-Bouldin Index y Calinski-Harabasz Index, para seleccionar el algoritmo que ofrezca las mejores agrupaciones. Los resultados obtenidos serán presentados tanto de manera visual como cuantitativa, permitiendo una interpretación clara y precisa de las agrupaciones logradas.

1.1. Objetivo general

Evaluar e implementar modelos de aprendizaje no supervisado para la segmentación de clientes no regulados del sector eléctrico utilizando curvas de carga para la obtención de agrupaciones homogéneas.

1.2. Objetivos específicos

1. Levantar requerimientos para la obtención y procesamiento de los datos de consumo energético de los clientes no regulados, transformándolos en curvas de carga representativas para su almacenamiento en una base de datos.
2. Realizar una revisión literaria de los algoritmos de agrupamiento más relevantes, identificando su funcionamiento, principios y parámetros claves para su correcta optimización e implementación en la segmentación de clientes del sector eléctrico.
3. Implementar una metodología de análisis de datos para la ejecución del proceso sistemático encargado de guiar las diferentes fases.
4. Aplicar los algoritmos de clustering, utilizando métodos de validación para definir el número óptimo de agrupaciones.
5. Evaluar y presentar los resultados generados por cada algoritmo, utilizando visualizaciones detalladas de las curvas de carga agrupadas.

1.3. Alcance

Como se mencionó en la descripción del componente, el presente trabajo está enmarcado en el análisis y segmentación de clientes no regulados del sector eléctrico, a partir de la construcción de sus curvas de carga características y la posterior aplicación de algoritmos de aprendizaje no supervisado con el fin de identificar patrones de consumo energético. El alcance de este trabajo está definido bajo las siguientes consideraciones:

1. Se ha adoptado la metodología CRISP-DM como marco de referencia, con una adaptación en su fase final. Dicha fase implica originalmente el despliegue del modelo en un entorno productivo, pero en este trabajo va a enfocarse en la evaluación comparativa de los resultados obtenidos con diferentes algoritmos de clustering (K-Means, Gaussian Mixture, Birch y Spectral Clustering), donde se presentarán métricas cuantitativas así como visualizaciones interpretativas de las agrupaciones.
2. Se llevará a cabo un proceso ETL, el cual obtiene, integra, limpia y normaliza los registros históricos de consumo energético que se tienen de cada cliente, con la finalidad de generar curvas de carga que representen el comportamiento energético de cada uno. Este proceso contempla la interpolación de valores nulos, la exclusión de días no laborales, corrección de formatos inconsistentes y la normalización mediante técnicas de escalamiento.
3. Se realizará la optimización e implementación de varios algoritmos de clustering, los cuales serán seleccionados en función de su relevancia en la literatura y su aplicabilidad en el análisis de curvas de carga. Para determinar el número óptimo de agrupaciones se realizará una validación utilizando el método del codo. Por otro lado, para la optimización de estos algoritmos se llevará a cabo un proceso de ajuste de hiperparámetros, priorizando configuraciones que maximicen la similitud en la forma de las curvas de carga agrupadas.
4. Los resultados incluirán la curva de carga representativa de cada cliente, la curva de carga correspondiente al día de máxima demanda y archivos separados por coma (.csv) con las coordenadas de dichas curvas. Asimismo, se presentarán resultados visuales de los clústeres y una tabla comparativa con métricas que cuantifican la calidad de las agrupaciones generadas por cada algoritmo.
5. Para el desarrollo del presente componente se ha contemplado Python como lenguaje de programación de alto nivel, Visual Studio Code como entorno de

desarrollo integrado, bibliotecas especializadas en análisis de datos y machine learning (pandas, scikit-learn, numpy, matplotlib, entre otras), así como herramientas de orquestación, en este caso Apache Airflow sobre Docker, para la automatización del proceso ETL.

Por lo anterior expuesto el alcance del componente se limita a la construcción, aplicación y evaluación de modelos de clustering basados en la similitud de curvas de carga, sin abordar fases posteriores como despliegues productivos en entornos de la empresa distribuidora de energía.

1.4. Marco Teórico

Para comprender este trabajo y su contexto, es de gran importancia tener bases sólidas sobre los principios subyacentes que sustentan el análisis y agrupación de los clientes en función de su curva de carga. Los apartados siguientes explicarán conceptos claves dentro del desarrollo del presente componente.

Sobre el sector eléctrico

1. Clientes no regulados

Los clientes no regulados en el sector eléctrico son aquellos cuya facturación por el suministro de energía se rige estrictamente por un contrato a término, el cual es realizado entre la empresa que suministra la energía y la empresa que recibe dicha energía. Los contratos mencionados anteriormente son bilaterales[1].

Debido a la naturaleza de los contratos que se suscriben con este tipo de clientes, los patrones de consumo de energía que poseen son bastante variados respecto a los clientes regulados [1].

2. Curvas típicas (curva de carga)

Una curva de carga o también llamada curva típica es un registro gráfico que indica la demanda eléctrica que ha tenido un cliente en cada instante durante un intervalo de tiempo determinado[2].

Estas curvas de carga reflejan el patrón de consumo cotidiano que poseen los clientes, dicho patrón está directamente relacionado con las máquinas o aparatos que utilizan, así como la energía que consumen durante sus actividades[3].

3. Segmentación de clientes

Debido a la naturaleza de los clientes no regulados y, agregando el hecho de que en su mayoría son grandes clientes, segmentarlos en grupos homogéneos permite optimizar la gestión de la demanda y mejorar la planificación del suministro eléctrico. Al agrupar clientes con patrones de consumo similares, es posible diseñar estrategias más eficientes para la contratación de energía, desarrollar y optimizar modelos tarifarios y, mejorar la predicción de la demanda a futuro [2]. Además, esta segmentación ayuda a evitar el sobredimensionamiento o subdimensionamiento de la capacidad de generación y distribución, garantizando un uso más eficiente de los recursos y optimizando los costos operativos.

Minería de datos

Según [4], la minería de datos corresponde a un proceso que consiste en la extracción de información relevante a partir de un gran conjunto de datos, con el fin de encontrar patrones interesantes que sean de utilidad, los cuales de otro modo habrían pasado desapercibidos. De la misma manera, métodos tradicionales de análisis de datos son combinados con algoritmos capaces de manejar grandes volúmenes de datos [5]. Entre sus principales funciones se destacan [5]:

- 1. Caracterización/Discriminación:** Sintetizar y explicar clases o conceptos.
- 2. Patrones frecuentes y asociaciones:** Reconocer relaciones que se repiten en el conjunto de datos.
- 3. Clasificación y regresión:** Elaborar modelos para predecir clases o valores numéricos.
- 4. Agrupación:** Generar etiquetas a partir de datos sin clasificar, optimizando la similitud interior.
- 5. Detección de valores atípicos:** Reconocer datos que no se ajustan a un patrón general.

En el contexto de la minería de datos, diversas metodologías han sido propuestas con el fin de dotar de estructura y sistematicidad a este proceso. Estas metodologías proveen fases bien definidas con el fin de asegurar la coherencia entre los objetivos del proyecto y los resultados. A continuación se detallan las tres metodologías más reconocidas en la literatura:

1. KDD (Knowledge Discovery in Databases)

KDD fue el primer modelo en recibir aprobación por parte de la comunidad científica para dirigir proyectos cuyo propósito es la obtención de conocimiento a partir de grandes cantidades de datos [6]. Esta metodología plantea un proceso iterativo que incluye la selección de datos, preprocesamiento, transformación, la implementación de algoritmos y el análisis de patrones. Entre sus contribuciones relevantes destaca la distinción de la minería de datos como una fase que forma parte de un proceso más amplio [7]. A diferencia de otras metodologías posteriores, a KDD se le atribuye un enfoque fundamentalmente conceptual, debido a que establece de manera generalizada cada fase del descubrimiento de conocimiento, sin profundizarlas [6]. KDD es visto como un punto de inicio para la sistematización de la minería de datos debido a esta característica, ya que proporcionó una base para el desarrollo de modelos más integrales que surgieron en años siguientes [6].

2. CRISP-DM (Cross-Industry Standard Process for Data Mining)

La metodología CRISP-DM, establecida en el año 2000, ha logrado consolidarse como la más utilizada para proyectos vinculados con la minería de datos [6]. Comprende seis etapas: entendimiento del negocio, entendimiento de los datos, preparación de los datos, modelado, evaluación y despliegue. En este modelo se detalla claramente las tareas y actividades que se deben llevar a cabo en cada fase, lo que permite establecer una conexión entre los objetivos estratégicos y el análisis técnico, es gracias a este equilibrio que CRISP-DM se posiciona como un marco idóneo para proyectos académicos e industriales [7]. Esta metodología admite retrocesos dentro de su flujo de trabajo, también permite realizar cambios en sus fases en función de los objetivos del proyecto, lo que refuerza su naturaleza iterativa y la hace muy flexible [7].

3. SEMMA (Sample, Explore, Modify, Model, Assess)

SEMMA es un modelo desarrollado por el instituto SAS, establece una guía metodológica que estructura un proceso en cinco fases: muestreo, exploración, modificación, modelado y evaluación [6]. Cada fase está enfocada en los aspectos técnicos del tratamiento de datos y en la aplicación de algoritmos. A diferencia de CRISP-DM, SEMMA no contempla las fases de comprensión del negocio o despliegue, situándola como una metodología muy útil en la ejecución de tareas relativas al análisis de datos, pero ineficaz para tareas donde los objetivos de la organización son clave [7]. Su utilidad se encuentra en proyectos en los que la experimentación y el modelado son más importantes que integrar el conocimiento adquirido dentro de los procesos comerciales [7].

Proceso ETL

El proceso ETL, es una técnica crucial que sirve para obtener, organizar y usar los datos apropiadamente según el fin requerido, se enfoca principalmente en la unión de datos provenientes de diversas fuentes, así como de su evaluación y limpieza [8]. Tal y como sus siglas indican, este proceso involucra tres fases descritas a continuación:

1. **Extracción:** Este paso es el responsable de extraer el conjunto requerido de datos de una o más fuentes, donde cada fuente tiene sus propias características, por lo cual, se debe tener conocimiento sobre como acceder a dichas fuentes, comprender la estructura de las mismas y saber como manejar cada fuente de acuerdo a su naturaleza [9]. Este proceso termina cuando todo el conjunto de datos es consolidado en un solo repositorio [9].
2. **Transformación:** Esta segunda fase consiste en procesar los datos extraídos para que sean consistentes, limpios e integrables dentro del repositorio. Se realizan diversas tareas como reestructurar la información, convertir formatos, limpiar los datos, integrar múltiples fuentes, tratamiento de valores nulos, entre otros [10]. El objetivo es asegurar que la información esté depurada y en condiciones para su carga en el repositorio final [10].
3. **Carga:** Es la última fase, aquí los datos son almacenados en un repositorio final o en una base de datos para su posterior análisis [11].

Aprendizaje no supervisado

El aprendizaje no supervisado es un tipo de algoritmo de aprendizaje automático, utiliza únicamente datos sin etiquetar, y es usado sobre estos con el objetivo de descubrir patrones o agrupar datos que posiblemente comparten características similares entre sí [12]. En este contexto, es pertinente destacar algunos elementos clave que permitirán una mejor comprensión, tales como:

1. Clustering

Es una de las categorías del aprendizaje no supervisado, la más consolidada en la actualidad, su objetivo es la identificación de subgrupos dentro de un conjunto extenso de datos no procesados, estos subgrupos son encontrados mediante la diferenciación de características [12].

2. Número de agrupaciones

Un problema muy común al utilizar algoritmos de aprendizaje no supervisado es elegir el número de agrupaciones deseadas [13], esta elección es muy

importante debido a que puede alterar la calidad de las agrupaciones finales dadas por los algoritmos. Como se menciona en [14], esta elección puede ser totalmente subjetiva, y en la mayoría de los casos el números de agrupaciones es seleccionado en función de criterios preestablecidos, sin embargo, existen técnicas como el método del codo que ayudan a validar el número de agrupaciones y que pueden ayudar en la selección de este criterio.

3. Método del codo

Es la forma más habitual de elegir o validar el número de clústeres, este método consiste en ajustar varios modelos K-means para un rango específico de agrupaciones, normalmente desde 1 hasta un número arbitrario máximo, posteriormente se traza un gráfico que contiene el valor total de la suma de los cuadrados por cada número de clústeres frente a ese respectivo número de clústeres [15]. El objetivo es encontrar aquel valor de número de clústeres donde la gráfica muestra un 'codo' y elegir dicho valor que probablemente nos ofrezca grupos bien separados [15].

4. Algoritmos de clustering

Los algoritmos de clustering son una parte fundamental del aprendizaje no supervisado, pues facilitan el descubrimiento de estructuras y patrones ocultos dentro de un conjunto de datos sin etiquetar [16].

A continuación se describirán los algoritmos de clustering que van a ser utilizados para el desarrollo del presente componente:

a) K-Means

Algoritmo de clustering basado en centroides que organiza n puntos de datos en k clústeres según la proximidad a centroides representativos [16]. Cada centroide corresponde a la media de su clúster y el objetivo es minimizar la suma de las distancias al cuadrado entre cada punto y su centroide [16], se puede formular matemáticamente como:

$$J = \sum_{i=1}^k \sum_{\mathbf{x} \in S_i} \|\mathbf{x} - \mu_i\|^2 \quad \text{con} \quad \mu_i = \frac{1}{|S_i|} \sum_{\mathbf{x} \in S_i} \mathbf{x} \quad y \quad i = \arg \min_j \|\mathbf{x} - \mu_j\|^2 \quad (1.1)$$

b) Gaussian Mixture Models (GMM)

Modelo que asume que los datos provienen de una mezcla de Gaussianas, cada una definida por su media y covarianza [16]. Este enfoque permite representar estructuras multimodales donde K-means falla. Los parámetros se estiman con el algoritmo EM, que ajusta iterativamente

medias, covarianzas y pesos para representar mejor los datos [16]. Matemáticamente, el modelo es expresado como:

$$p(x) = \sum_{j=1}^k \pi_j N(x|\mu_j, \Sigma_j), \quad w_{ij} = \frac{\pi_j N(x_i|\mu_j, \Sigma_j)}{\sum_{l=1}^k \pi_l N(x_i|\mu_l, \Sigma_l)} \quad (1.2)$$

mientras que las actualizaciones de los parámetros en cada iteración están dadas por:

$$\pi_j = \frac{1}{n} \sum_{i=1}^n w_{ij}, \quad \mu_j = \frac{\sum_{i=1}^n w_{ij} x_i}{\sum_{i=1}^n w_{ij}}, \quad \Sigma_j = \frac{\sum_{i=1}^n w_{ij} (x_i - \mu_j)(x_i - \mu_j)^T}{\sum_{i=1}^n w_{ij}} \quad (1.3)$$

c) Spectral Clustering

Algoritmo de clustering basado en grafos, transforma los datos en una red donde los nodos representan puntos de datos y las aristas sus similitudes, a partir de esto construye la matriz Laplaciana, cuyos autovectores permiten identificar estructuras dentro del grafo y formar clústeres con alta cohesión interna [16]. El objetivo es minimizar la siguiente función:

$$\min \text{Tr}(H^T L H) \quad \text{sujeto a} \quad H^T H = I \quad (1.4)$$

d) BIRCH

Es un algoritmo de clustering de tipo jerárquico, está diseñado para trabajar con grandes volúmenes de datos, resumiendo toda la información de los mismos en una sola estructura jerárquica que tiene el nombre de CF-Tree, donde cada clúster es representado como una Clustering Feature (CF) [17], la cual está definida por:

$$CF = (N, LS, SS) \quad (1.5)$$

donde N es el número de puntos, LS la suma lineal y SS la suma de los cuadrados de los datos. El umbral de radio T se determina mediante un problema de optimización, definido como:

$$\min_T g(W_k(T), B_k(T)) \quad (1.6)$$

donde W_k mide compacidad intra-clúster y B_k separación inter-clúster [17].

5. Hiperparametrización de algoritmos

Técnica que consiste en ajustar los parámetros que controlan el comportamiento de los algoritmos de clustering, estos parámetros influyen en la calidad de las agrupaciones finales, el objetivo es encontrar aquella combinación de parámetros que ofrezca los mejores resultados en cada algoritmo [18].

6. Métricas de evaluación de agrupaciones

Son medidas de calidad que sirven para dar validación a los clústeres obtenidos por los algoritmos, estas métricas se basan en la premisa de 'Maximizar la similitud dentro de cada clúster y minimizar la similitud entre los diferentes clústeres', el objetivo es lograr clústeres compactos y lo más separados posibles entre sí [19].

Herramientas utilizadas

Para el desarrollo del componente se han considerado varias herramientas que facilitan las etapas de procesamiento, almacenamiento, análisis de los datos, e implementación de los modelos de clustering, la Tabla 1.1 los detalla:

Tabla 1.1: Herramientas utilizadas para el desarrollo del componente

Herramienta	Descripción de la herramienta
Airflow 2.10.5	Apache Airflow es una plataforma de código abierto que permite el desarrollo, programación y supervisión flujos de trabajo, utiliza Python, lo que le permite conectarse con diversas tecnologías [20].
Docker 4.43.1	Docker es una plataforma abierta utilizada para el desarrollo, envío y ejecución de aplicaciones, permite empaquetar y ejecutar aplicaciones en un entorno aislado denominado contenedor [21].
Python 3.13	Python es un lenguaje de programación de alto nivel con naturaleza interpretada, maneja estructuras de datos con un alto nivel de eficiencia y ofrece una sintaxis simple, razones por las cuales es ampliamente utilizado en campos como desarrollo web, ciencia de datos, automatización, entre otros [22].
Visual Studio Code 1.101.2	Visual Studio Code es un editor de código fuente que contiene herramientas de depuración, control de versiones y extensiones para varios lenguajes. Ofrece varias características que permiten desarrollar código eficientemente [23].
MongoDB Atlas 8.0.13	Es una base de datos no relacional administrada en nube, basada en documentos, y que brinda una gran escalabilidad y flexibilidad, además de un modelo avanzado de consultas e indexación [24].

2. METODOLOGÍA

2.1. Caso de estudio

Unos de los grandes desafíos que enfrenta la EEQ es la administración eficiente de la demanda de sus clientes, especialmente del segmento que no está regulado, este grupo es estratégico debido a su representativo nivel de consumo y a la variabilidad de sus patrones de carga. Estos usuarios, no están incluidos en un esquema tarifario regulado, por lo cual exhiben una gran diversidad de forma en sus curvas de demanda, lo cual obstaculiza enormemente la planificación energética y el diseño de estrategias destinadas a asegurar eficiencia y fiabilidad en el sistema eléctrico.

Los métodos tradicionales que han sido utilizado en la EEQ con el fin de examinar el comportamiento de la demanda (basados principalmente en clasificaciones rígidas o en el cálculo de promedios), han demostrado importantes limitaciones al no conseguir captar la complejidad de los patrones de consumo. En investigaciones anteriores realizadas por Gerencia de Planificación, se ha confirmado esta circunstancia. Los estudios mostraron que el comportamiento energético está directamente relacionado con la actividad económica del cliente no regulado y a su curva de carga característica, lo cual hace inviable que un único criterio generalizado represente apropiadamente a todo este conjunto.

2.2. Brainstorming

La lluvia de ideas, también conocida como brainstorming, es un método que se emplea en el campo de la ingeniería de requisitos y la investigación para recopilar información de manera colaborativa. Esto facilita determinar necesidades, cuestiones problemáticas y posibles perspectivas de solución durante las etapas iniciales de un proyecto. Su valor metodológico radica en su capacidad de permitir reunir un conjunto extenso de percepciones, las cuales pueden ser organizadas y analizadas con más rigor posteriormente, convirtiéndose así en un insumo fundamental para determinación del enfoque metodológico [25].

En relación con el desarrollo del presente componente, esta técnica se utilizó como método para recopilar información en reuniones con el equipo encargado del departamento de planificación de la demanda. Los métodos de análisis tradicionales, la disparidad de los perfiles de carga y la necesidad de contar con un mecanismo de segmentación que facilite la agrupar a los clientes según su comportamiento energético fueron determinados mediante este proceso.

La relevancia del uso de brainstorming en este caso de estudio se explica por el hecho de que, siendo un problema técnico y organizacional complejo, fue imprescindible obtener directamente la experiencia y la sabiduría del personal de la compañía. De este modo, esta técnica hizo posible determinar las necesidades y los problemas más importantes vinculados al estudio del comportamiento energético de los clientes no regulados, lo cual permitió establecer un punto de partida claro para el desarrollo del proyecto.

2.3. CRISP-DM

El desarrollo del presente componente se sustenta en CRISP-DM, cuyas siglas corresponden a Cross-Industry Standard Process for Data Mining, metodología que es ampliamente reconocida por su aplicabilidad en proyectos de minería de datos y por brindar un enfoque sistemático y estructurado.

La elección de esta metodología se basa en la necesidad de guiar de manera ordenada el análisis del consumo energético de los clientes no regulados del sector eléctrico, y esta opción es la que más se acopla debido a que nos permite avanzar desde la comprensión del problema hasta la obtención de resultados comparables. Además, la metodología CRISP-DM brinda la flexibilidad necesaria para realizar ajustes en sus fases en función de los objetivos que se quieran cumplir, esta característica fue clave para su elección, pues en el presente componente la fase final no contemplará un despliegue productivo como tal, sino una evaluación comparativa de la calidad de agrupaciones obtenidas por cada algoritmo.

La validez del uso de CRISP-DM para el desarrollo del presente componente es respaldada por su probado éxito en estudios anteriores similares a este. Sarnovsky y Bednár aplicaron en [26] esta metodología para realizar clustering de clientes de una empresa distribuidora de energía, donde estos fueron agrupados en función de sus curvas de carga anuales. De manera similar, Otieno adoptó CRISP-DM en [27] como base para desarrollar diversos análisis de patrones de consumo en una empresa distribuidora de energía. Incluso en otros sectores, como el de software, investigaciones como la presentada en [28] emplean CRISP-DM para estructurar procesos de clustering de clientes basados en técnicas de minería de datos. Estos antecedentes proporcionan una evidencia sólida y específica que valida la elección de CRISP-DM como metodología para el desarrollo del presente componente.

CRISP-DM es un método probado utilizado para orientar proyectos de minería de datos. Ofrece una serie de fases que resumen el ciclo vital de minería de datos, a la vez que incluye descripciones y tareas necesarias en cada fase, ayudando a estructurar un flujo de trabajo ordenado cuya secuencia no es estricta, donde se puede avanzar y retroceder entre fases de ser necesario [29].

El modelo CRISP-DM es sumamente flexible, y sus fases pueden ser personalizadas en función de los objetivos del proyecto, pudiendo crear un modelo de minería de datos que se adapte a necesidades concretas [29]. CRISP-DM contiene un total de seis fases, tal y como se describe en [30]:

1. **Comprendión del negocio:** Esta fase inicial se enfoca en analizar y comprender tanto los objetivos como los requerimientos del proyecto desde la perspectiva del negocio. Posteriormente todo este conocimiento es plasmado en un proyecto de minería de datos enfocado en alcanzar los objetivos.
2. **Comprendión de los datos:** La fase de comprensión de datos tiene como principal objetivo la 'familiarización' con los datos. Para lograr esto se realiza una recolección inicial de los datos y se procede a realizar un pequeño análisis exploratorio de los datos con el fin de comprender los datos que se tienen e identificar problemas con la calidad de los mismos.
3. **Preparación de los datos:** Esta fase es crucial en CRISP-DM, debido a que abarca todas las actividades requeridas hasta la construcción final del conjunto de datos, los cuales servirán posteriormente para la fase de modelado. Esta fase incluye tareas como la limpieza, transformación y normalización de los datos, con el fin de asegurar la calidad de estos.
4. **Modelado:** Varias herramientas de modelamiento son seleccionadas con el fin de ser aplicadas sobre nuestro conjunto de datos preparados. Los parámetros de dichas herramientas deben ser calibrados hasta obtener los valores óptimos que ofrezcan los mejores resultados.
5. **Evaluación:** En esta penúltima fase del proyecto, ya se tiene construido uno o varios modelos que aparentemente ofrecen resultados de calidad. Antes de proceder a la fase del despliegue, se realiza una evaluación del modelo, revisando cada paso ejecutado hasta la construcción final del mismo con el fin de determinar si existe algún objetivo que no haya sido abordado lo suficiente.
6. **Despliegue:** La construcción del modelo no es el final del proyecto. En función de los requerimientos, la fase de despliegue puede ser tan simple como la

generación de un reporte o tan complejo como su respectiva implementación en otros proyectos de minería de datos.

Es importante recalcar que, en el desarrollo del presente componente, la fase número seis de CRISP-DM correspondiente al despliegue fue modificada en función de los objetivos específicos del proyecto. En este caso, a diferencia de la metodología original, que incluye la implementación del modelo en un entorno productivo, esta fase tendrá un enfoque comparativo de los resultados obtenidos con los distintos algoritmos de clustering. Para ello, las agrupaciones fueron analizadas mediante el uso de métricas de evaluación que permiten cuantificar la calidad de los clústeres, con la finalidad de elegir aquellos resultados que brinden una segmentación más coherente y homogénea. Esta adaptación fue posible gracias a la flexibilidad que caracteriza a CRISP-DM, como se mencionó antes, lo que hace posible la modificación de sus estapas en base a los requerimientos específicos del proyecto, sin perder la consistencia metodológica y asegurando la validez del proceso ejecutado.

2.4. Implementación de CRISP-DM

Para la implementación de CRISP-DM en el presente componente se han tenido en cuenta las fases y tareas definidas de forma precisa y teórica en [30], las cuales han sido adaptadas (de ser necesario) y desarrolladas en función de las necesidades específicas del proyecto, sin alterar la estructura metodológica y secuencial original. La Figura 2.1 presenta un esquema que sintetiza las fases y tareas mencionadas anteriormente, la cual fue realizada a partir de los principios establecidos en [30].



Figura 2.1: Esquematización de las fases y tareas de la metodología CRISP-DM

Entendimiento del negocio

Levantamiento inicial de información

En esta tarea se realizó un levantamiento de información con la finalidad de comprender el contexto del negocio y establecer los insumos requeridos para la elaboración del componente. Este proceso se llevó a cabo con la participación de los actores que intervienen en la gestión y el análisis de la información del consumo de energía, los cuales permitieron identificar la problemática, requerimientos, restricciones técnicas, los objetivos del análisis y las posibles salidas esperadas. La Tabla 2.1 presenta los actores involucrados en el proceso, así como la descripción del rol de cada uno de ellos.

Tabla 2.1: Actores involucrados en el levantamiento de información

Actor	Rol que desempeña
Jefe de sección de planificación	Dueño del proyecto
Ingeniera eléctrica	Líder del proyecto
Tesista	Desarrollador
Tutores	Colaboradores / Supervisores

Se elaboró un mapa mental como técnica inicial para ordenar, sintetizar y estructurar la información recopilada anteriormente. Esta herramienta posibilita tener una visión completa del entorno y establecer trazabilidad entre los elementos empresariales y los objetivos del componente. La Figura 2.2 presenta el mapa mental realizado, agrupado en siete apartados:

- Problemática: Identifica las restricciones presentes en la gestión de la demanda de los clientes no regulados, asociadas a la alta variabilidad en sus patrones de consumo y métodos tradicionales de análisis insuficientes.
- Importancia del análisis: Se destaca la necesidad de que la EEQ tenga conocimiento sobre los patrones de consumo que poseen los clientes no regulados debido a que estos representan un grupo de alto consumo energético.
- Necesidades expresadas: Se detallan los requerimientos estratégicos por parte de la EEQ, tales como conocer patrones de consumo, disminuir la incertidumbre en la planificación y mejorar el uso de los recursos energéticos.
- Objetivo del levantamiento: Establecer criterios preliminares para una segmentación útil, que apoye la planificación energética fundamentada en evidencias y que facilite la toma de decisiones.

- Actores involucrados: Se identifica a la EEQ como la entidad responsable de proveer de manera confiable y eficaz la información relativa a los clientes no regulados como los sujetos de estudio, y al área de Planificación como principal usuario de los resultados.
 - Información inicial: Se describe las fuentes de datos existentes, que son los registros de consumo mensual de 388 clientes a lo largo del año 2023, con intervalos de medición establecidos entre 5 y 15 minutos, así como variables relacionadas con la energía y la potencia.
 - Salidas esperadas: Se determinan los productos que resultarán del análisis: la curva característica de cada cliente, la curva del día con máxima demanda, archivos separados por coma y de texto plano, además de las agrupaciones con los clientes segmentados en base a la similitud de sus curvas de carga características.

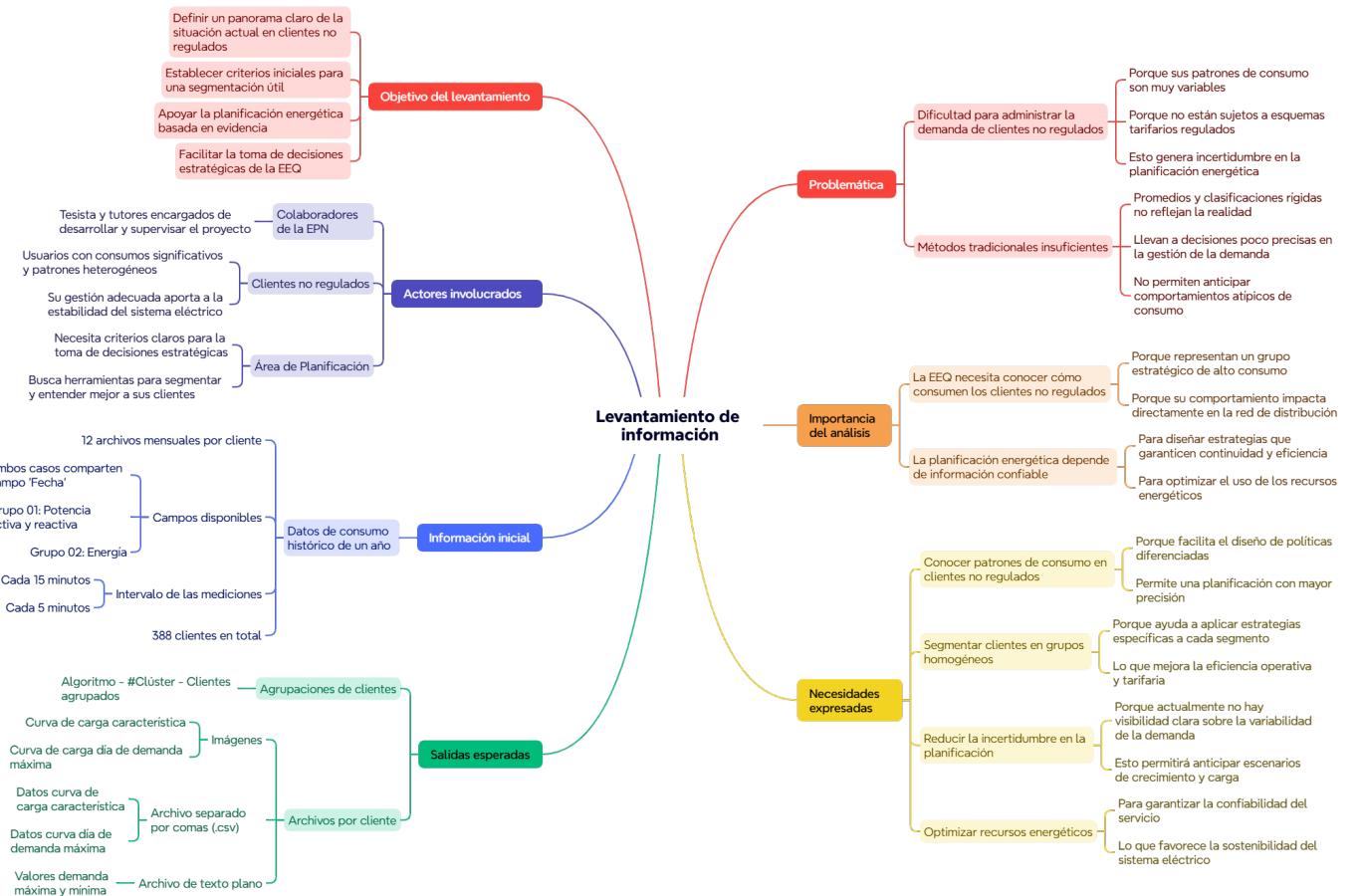


Figura 2.2: Mapa mental utilizado para levantamiento y organización de información

Determinar los objetivos del negocio

El propósito de la EEQ es generar agrupaciones de clientes no regulados según el comportamiento energético que estos reflejen en su curva de carga característica anual. Con esto, se busca simplificar la planificación energética y mejorar la gestión de la demanda a través de la identificación de patrones de consumo.

Los requerimientos específicos que surgieron son:

- Obtener una visión completa de los patrones de consumo asociados a los clientes no regulados del sector eléctrico.
- Generar agrupaciones que posibiliten el reconocimiento de comportamientos energéticos generalizados dentro del grupo de clientes.
- A partir de la información obtenida, crear estrategias diferenciadas para gestionar la demanda y planificar la red del sistema eléctrico.

Se considera que el proyecto es exitoso si las agrupaciones resultantes permiten identificar patrones energéticos característicos y generen información interpretable para la EEQ, orientada al cumplimiento de los objetivos establecidos.

Evaluar la situación

Se identificaron los recursos, las limitaciones y los supuestos requeridos para llevar a cabo el proyecto, basándose en la información resumida en el mapa mental.

- Recursos disponibles:
 - Registros históricos de consumo mensual (potencia activa, potencia reactiva o energía) del año 2023
 - Entorno de desarrollo de Python, orquestación del proceso ETL utilizando Apache Airflow sobre Docker y MongoDB Atlas como base de datos en nube para guardar los registros correspondientes a las curvas de carga características
 - Personal del área de Planificación de la EEQ como apoyo técnico y proveedor de la información requerida.
- Requerimientos:
 - Para el análisis serán considerados únicamente aquellos registros correspondientes a días laborables, los demás días serán excluidos.

- Se debe garantizar que los datos tengan la consistencia característica de una serie temporal (independientemente del intervalo)
- Se tiene que restringir el alcance a la creación y evaluación de los modelos de agrupamiento, sin contemplar un despliegue en producción.

■ Supuestos:

- Los valores atípicos reflejan comportamientos totalmente válidos propios de la naturaleza de la demanda eléctrica del sector no regulado.
- Las mediciones recopiladas reflejan de manera precisa el comportamiento energético real de los clientes.

■ Limitaciones:

- Diversidad en los formatos de los archivos de entrada (Algunos contienen solo datos de potencia, otros solo datos de energía, la fecha viene en diferentes formatos, etc...)
- El análisis está limitado a los clientes no regulados del sector eléctrico dentro del área de concesión de la EEQ.
- Existencia de valores nulos e inconsistencias en las series temporales.
- Restricciones de capacidad computacional debido al gran volumen de datos que se requiere manipular
- El análisis se realizó exclusivamente con los datos del año 2023, debido a que este período es el que mayor nivel de completitud y consistencia presenta.

Determinar los objetivos de la minería de datos

Después de haber establecido los objetivos del negocio, estos fueron convertidos a metas técnicas concretas del proceso de minería de datos. El objetivo de la minería de datos es implementar modelos de clustering que permitan clasificar a los clientes no regulados en grupos homogéneos, basándose en la semejanza de sus curvas de carga características anuales, con el propósito de identificar patrones de consumo que puedan ser utilizados como insumo para la planificación estratégica.

Los objetivos de la minería de datos son:

- Crear una base de datos no estructurada con las curvas de carga características de cada cliente a partir del proceso ETL.

- Determinar y validar número óptimo de agrupaciones utilizando técnicas como el método del codo.
- Determinar la configuración óptima de parámetros para cada algoritmo, mediante hiperparametrización.
- Implementar algoritmos de aprendizaje no supervisado (KMeans, Gaussian-Mixture, Birch y Spectral Clustering) en los datos de las curvas de carga.
- Evaluar la calidad de los clústeres utilizando métricas que cuantifican la calidad de los mismos (Silhouette Score, Davies-Bouldin Index, Calinski-Harabasz Index y correlación.)

Se considera que la minería de datos es exitosa si los modelos consiguen generar agrupaciones con una alta cohesión interna y una clara separación entre clústeres, lo cual se comprobará a través de métricas que cuantifican la calidad de los clústeres, así como mediante gráficos y técnicas de representación que faciliten el análisis visual de las agrupaciones generadas.

Elaborar el plan del proyecto

Se desarrolló un plan de ejecución, el cual fue estructurado en base a la metodología CRISP-DM y ajustado a las necesidades del presente componente, basándose en la información levantada y los objetivos establecidos previamente. La Figura 2.3 muestra el flujo definido para guiar el desarrollo del proyecto.

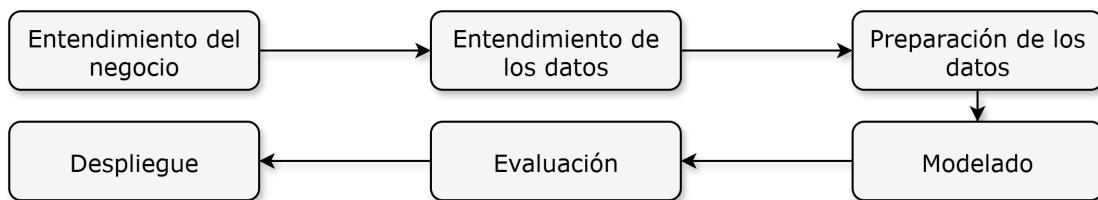


Figura 2.3: Flujo establecido para el plan de ejecución del proyecto

Entendimiento de los datos

Recolección inicial de los datos

Los datos fueron proporcionados directamente por el líder del proyecto, quien a su vez, obtuvo los datos a partir del sistema de medición comercial de la EEQ (Telemediciones). Cada cliente tiene doce archivos que corresponden a mediciones mensuales (aproximadamente 30000 registros), estos archivos pueden contener las mediciones en energía (kWh), o en potencias activa (kW) y reactiva (KVAr).

Los archivos fueron proporcionados en formatos .csv, mostrando diferencias significativas en su estructura y formato, sin embargo, mantienen información específica del valor medido, la fecha y la hora.

Descripción de los datos

Tras la recolección de datos, se llevó a cabo una revisión inicial de los archivos para determinar su estructura general y las variables que contienen. Esta tarea permitió determinar la naturaleza de los datos y evaluar el tipo de formato que poseen, donde se pudo notar lo siguiente:

- Aparentemente, la frecuencia de medición entre registros es de 15 minutos.
- Todos los archivos poseen una cabecera con información del cliente u otros aspectos relevantes.
- Existen dos grupos de clientes: los que tienen únicamente mediciones de energía (b) y los que tienen mediciones de potencia activa y reactiva (a).
- El formato de la fecha varía (ej: AAAA/MM/DD, AAAA-MM-DD, DD/MM/AAAA)
- Algunos archivos presentan ausencia de registros debido a inconsistencias o pérdidas derivadas del sistema de medición (a).
- Algunos archivos poseen líneas de resumen, las cuales incluyen totales ponderados agregados para intervalos específicos (b).
- Algunos valores dentro de los registros contienen coma como separador de miles y punto como separador decimal, por ejemplo: 1,203.239 (b).

En la Figura 2.4 se puede apreciar lo mencionado anteriormente.




Informe de Medidas de Puntos Frontera;				
Punto Frontera: ACNOVQU03;				
Fecha;AE (kWh);AS (kWh);RE (KVarh);RS (KVarh);SE (KVarh);SS (KVarh);				
Total Días: 2023/02/28;0.000;104,224.054;0.000;34,468.932;109,799.955;86,400.000;				
2023/03/01 00:00;0.000;1,203.239;0.000;396.692;1,267.004;900.000;				
2023/02/28 23:45;0.000;1,188.957;0.000;386.713;1,242.697;900.000;				
2023/02/28 23:30;0.000;1,162.232;0.000;379.803;1,222.746;900.000;				
2023/02/28 23:15;0.000;1,203.980;0.000;397.491;1,267.914;900.000;				
2023/02/28 23:00;0.000;1,198.779;0.000;387.437;1,252.258;900.000;				
2023/02/28 22:45;0.000;1,217.744;0.000;415.905;1,286.987;900.000;				
2023/02/28 22:30;0.000;1,232.724;0.000;409.121;1,298.971;900.000;				
DEMANDAS				
Orden	Fecha	Origen	Demanda activa DEL	Demanda reactiva DEL
1	2023-05-07 18:15	Lectura AMR	0.03	0
2	2023-05-07 18:30	Lectura AMR	0.0318	0
3	2023-05-07 18:45	Lectura AMR	0.0324	0
4	2023-05-07 19:00	Lectura AMR	0.0318	0
5	2023-05-07 19:15	Lectura AMR	0.0336	0

Figura 2.4: Estructura general de los archivos de datos de consumo energético

Exploración de los datos

Se llevó a cabo un análisis exploratorio de los datos con el fin de comprender de manera general su estructura, calidad y características principales. Esta tarea posibilitó el descubrimiento de distribuciones, patrones globales y eventuales anomalías que podrían tener un impacto en las etapas subsiguientes del proceso de análisis.

Se incorporaron visualizaciones que reflejan las principales cualidades del conjunto de datos (Figura 2.5), se elaboró una imagen resumen que presenta la cantidad de clientes según el tipo de medición, como se distribuyen los registros totales por cliente, la variabilidad en el número de registros mediante un diagrama de caja y el número de registros que existen en función del intervalo de medición.

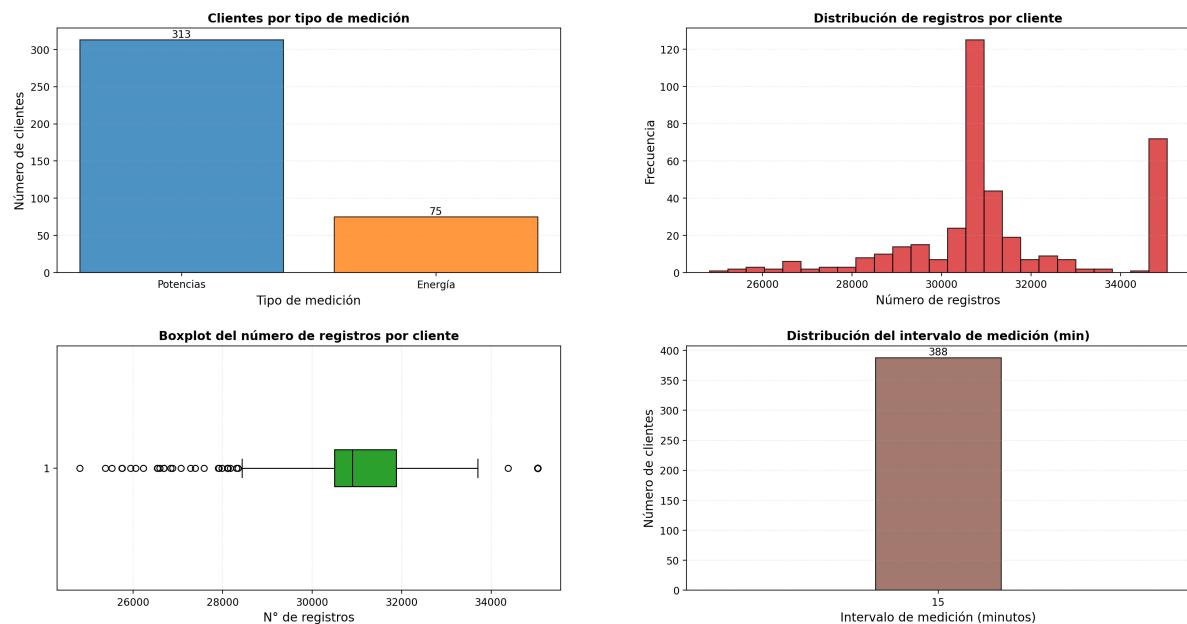


Figura 2.5: Exploración de los datos de consumo energético

El primer gráfico (superior izquierdo) indica que la mayoría de los clientes tienen registros de potencia activa y reactiva, mientras que un número menor tiene registros de energía. El segundo gráfico (superior derecho) revela que la mayoría de los clientes tienen entre 30.000 y 35.040 registros, con algunos clientes teniendo significativamente un menor, un comportamiento que puede ser causado debido a pérdida de registros desde el propio sistema de medición.

El tercer gráfico (inferior izquierdo) es un diagrama de cajas que muestra la variabilidad en el número de registros por cliente, indicando la presencia de algunos valores atípicos. Finalmente, el cuarto gráfico (inferior derecho) muestra que todos los clientes tienen un intervalo de medición de 15 minutos, lo cual es consistente con la frecuencia de medición esperada.

Verificar calidad de los datos

Dadas las inconsistencias presentes en los archivos de medición, identificadas en tareas anteriores, se llevó a cabo una comprobación de su calidad. Esta tarea posibilitó determinar el grado de completitud, coherencia y consistencia de los datos antes de proceder con su preparación.

Se llevaron a cabo varias comprobaciones con el objetivo de cuantificar y observar el impacto generado de estas anomalías sobre el conjunto de datos. Para ello, se elaboró una imagen que resume los resultados obtenidos en términos de uniformidad de formato, consistencia numérica y completitud de los datos.

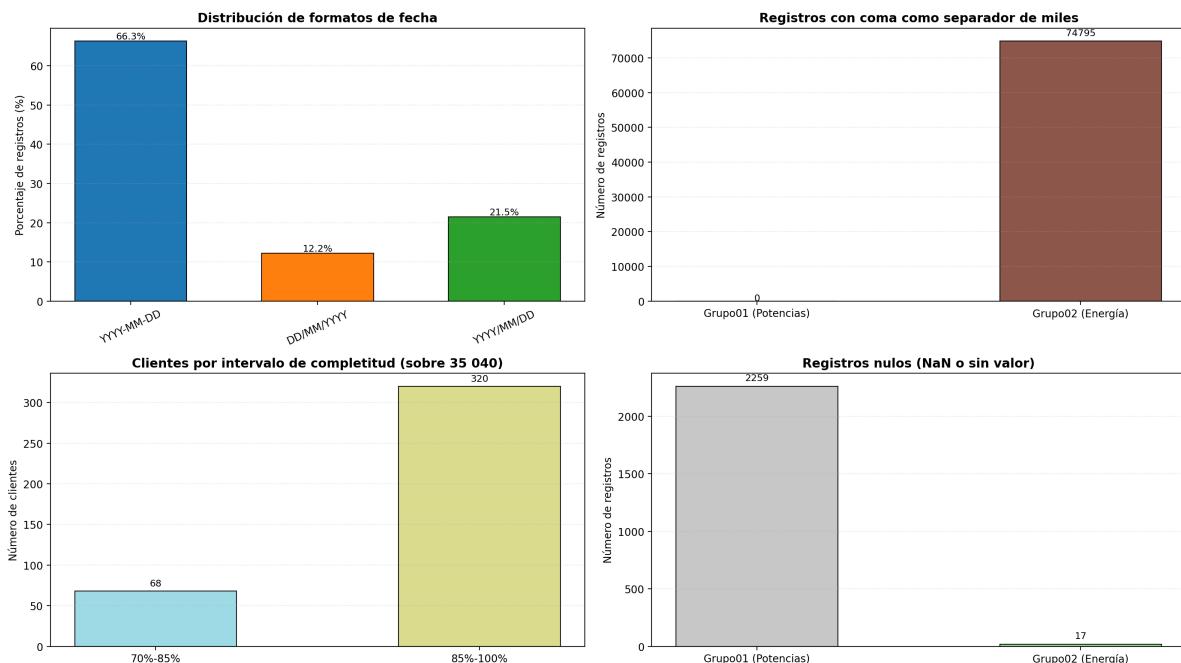


Figura 2.6: Verificación de la calidad de los datos de consumo energético

Como se puede apreciar en el primer gráfico (superior izquierdo), el formato más común es el YYYY-MM-DD (66,3 %), seguido por YY/MM/DD (21,5 %) y DD/MM/YYYY (12,2 %). Esta falta de uniformidad en el formato de fecha es crítico debido a que para construir series temporales se debe tener un formato único, se requerirá un proceso de estandarización de formato de fecha para solucionarlo.

El segundo gráfico (superior derecho) revela que dentro del grupo de clientes que poseen únicamente mediciones de energía, existen 74795 registros que tienen la coma como separador de miles. Si no se trata, esto provocará problemas relacionados con la interpretación de datos numéricos debido a que Python no trabaja con separadores de miles, y su separador decimal es el punto.

El tercer gráfico (inferior izquierdo) muestra que 320 clientes tienen una completitud de datos mayor al 85 % del total ideal de registros anuales, mientras que 68 clientes se encuentran entre el 70 % y 85 %, debido a pérdidas de información derivadas del sistema de donde fueron obtenidas las mediciones.

Por último, el cuarto gráfico (inferior derecho) indica que dentro del grupo de clientes que solo tienen mediciones de potencias, existen en total 17 registros nulos, mientras que en el grupo de clientes que posee únicamente mediciones de energía, existe un total de 2259 registros nulos. El volumen de registros que poseen valores nulos es mínimo respecto al total de registros global y por cliente, por lo que se considera apropiado utilizar un método de interpolación para llenar dichos valores. Este método posibilitará que sean estimados en función de la tendencia particular de cada cliente, conservando la coherencia temporal y el comportamiento original de los datos.

Preparación de los datos

En esta etapa, se implementó un proceso ETL a través de Apache Airflow en un entorno Docker, tal y como se ilustra en la Figura 2.7, gracias al cual se pudo orquestar de manera eficiente el flujo de procesamiento de los datos de consumo, garantizando automatización y trazabilidad.

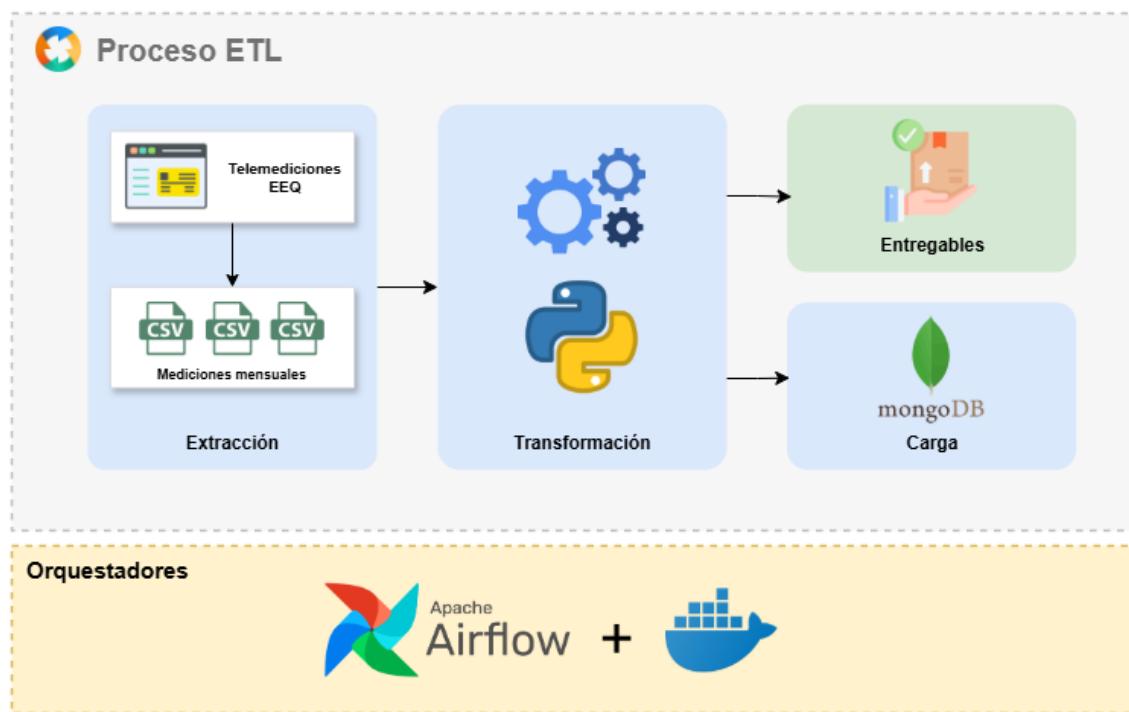


Figura 2.7: Esquematización del proceso ETL y sus etapas

El uso de Apache Airflow permitió construir un pipeline totalmente modularizado, donde cada tarea del Grafo Acíclico Dirigido (DAG) constituyó una acción concreta en el proceso ETL. Este pipeline abarca desde la lectura y consolidación de los archivos mensuales de cada cliente hasta la elaboración de las curvas de carga características y su respectiva carga en MongoDB Atlas.

Para la implementación, Apache Airflow fue desplegado en contenedores Docker utilizando la imagen oficial, la cual configura varios servicios, entre ellos, los fundamentales para el funcionamiento de Airflow [20]:

- Scheduler: Organiza y programa la ejecución de tareas establecidas en los DAGs.
- Worker: Ejecuta las tareas asignadas por el scheduler.
- Webserver: Ofrece una interfaz de usuario que permite la gestión de los DAGs y sus tareas.
- Postgres: Almacena metadatos como estado de las tareas, DAGs y logs.

Toda esta arquitectura se encuentra definida en `docker-compose.yaml`, archivo que fue modificado para incluir de manera adicional tres volúmenes compartidos entre el sistema de archivos local y el entorno aislado de Apache Airflow:

- /data: Contiene todos los archivos con formato `.csv` correspondientes a las mediciones mensuales de todos los clientes no regulados.
- /outputs: Este volumen fue creado con la finalidad de almacenar los entregables adicionales requeridos por la EEQ.
- /utils: Contiene un archivo con variables de entorno (`.env`), cuyo contenido son credenciales para comunicarse con MongoDB Atlas. Adicionalmente contiene el archivo `utilities.py`, el cual contiene funciones auxiliares para comunicación con la base de datos, estandarización de formato de fechas, generación de curvas, entre otros.

La Figura 2.8 ilustra la modificación realizada al archivo `docker-compose.yaml` para añadir los volúmenes compartidos mencionados anteriormente. El contenido completo de dicho archivo se encuentra disponible en el Anexo I.

```

YAML docker-compose.yaml
75      volumes:
76        - ${AIRFLOW_PROJ_DIR:-.}/dags:/opt/airflow/dags
77        - ${AIRFLOW_PROJ_DIR:-.}/logs:/opt/airflow/logs
78        - ${AIRFLOW_PROJ_DIR:-.}/config:/opt/airflow/config
79        - ${AIRFLOW_PROJ_DIR:-.}/plugins:/opt/airflow/plugins
80        - ${AIRFLOW_PROJ_DIR:-.}/data:/opt/airflow/data
81        - ${AIRFLOW_PROJ_DIR:-.}/utils:/opt/airflow/utils
82        - ${AIRFLOW_PROJ_DIR:-.}/outputs:/opt/airflow/outputs

```

Figura 2.8: Inclusión de volúmenes compartidos en el archivo docker-compose.yaml

Tras haber desplegado de manera exitosa el entorno de Airflow, se realizó la instalación de las librerías necesarias para la ejecución del DAG correspondiente al proceso ETL. Dado que los servicios encargados de orquestar la ejecución de las tareas son el worker y el scheduler, cuyos identificadores son 0c91068629f2 y 0219ddd747d0, se accedió a la línea de comandos propia de cada contenedor haciendo uso de la sentencia `docker exec -it IDENTIFICADOR bash`.

Una vez dentro de la CLI de cada servicio, se hizo uso del comando `python -m pip install pandas numpy scikit-learn pymongo dotenv matplotlib` para instalar las librerías requeridas por Python para el correcto funcionamiento del DAG. La Figura 2.9 muestra todo el proceso mencionado anteriormente:

```

airflow@0c91068629f2: /opt/i ~ + | 23:04:30
Andres@Andres-PCE ~ | 23:04:30
$ docker ps --format "table {{.ID}}\t{{.Names}}"
CONTAINER ID NAMES
0c91068629f2 tic-eeq-airflow-worker-1
0219ddd747d0 tic-eeq-airflow-scheduler-1
edd3cde0fc88 tic-eeq-airflow-triggerer-1
91b45806eea9 tic-eeq-airflow-webserver-1
7487c6923b8c tic-eeq-redis-1
b058113e5db0 tic-eeq-postgres-1

[ Andres@Andres-PCE ~ | 23:04:34
$ docker exec -it 0c91068629f2 bash
airflow@0c91068629f2:/opt/airflow$ python -m pip install pandas numpy scikit-learn pymongo dotenv matplotlib
[ Andres@Andres-PCE ~ | 23:07:15
$ docker exec -it 0219ddd747d0 bash
airflow@0219ddd747d0:/opt/airflow$ python -m pip install pandas numpy scikit-learn pymongo dotenv matplotlib

```

Figura 2.9: Identificación de worker e instalación de librerías necesarias.

Una vez que se ha establecido ambiente de ejecución con sus respectivas dependencias, se procedió a estructurar el proceso ETL dentro de un DAG denominado `etl_dag_datos_consumo`. Cada nodo que compone el grafo simboliza una tarea autónoma dentro del proceso ETL, ejecutada de manera secuencial o paralela en función de las dependencias que se hayan definido. El código Python correspondiente al DAG se encuentra disponible en el Anexo II.

La orquestación general de este DAG en Apache Airflow se muestra en la Figura 2.10, donde se evidencia el flujo de ejecución y la relación de dependencia entre las diferentes tareas que lo componen. A continuación se describe de manera general el propósito de cada tarea que integra dicho proceso:

- `extraer_datos_grupo01`: lee y consolida los archivos mensuales de los clientes que tienen mediciones correspondientes a demanda activa y reactiva medidas en un intervalo de 15 minutos.
- `extraer_datos_grupo02`: lo mismo que la tarea anterior con la diferencia que lo realiza para los clientes que tienen mediciones de energía medidas igualmente en un intervalo de 15 minutos.
- `transformar_datos_grupo01` y `transformar_datos_grupo02`: para cada grupo, se encargan de la limpieza, estandarización, normalización y cálculo de la potencia aparente de los datos.
- `transformar_datos_unificados`: unifica todos los datos a partir de los conjuntos procesados de los dos grupos, dando como resultado un DataFrame consolidado que contiene las curvas tipo de todos los clientes.
- `cargar_datos_curvas_tipo`: lleva a cabo el formateo final y carga el conjunto de datos consolidado en la base de datos MongoDB Atlas.
- `generar_entregables_por_cliente`: no es parte del proceso ETL como tal, genera archivos requeridos por la EEQ, tales como curva tipo, curva de demanda máxima y archivos .csv con sus respectivas coordenadas.

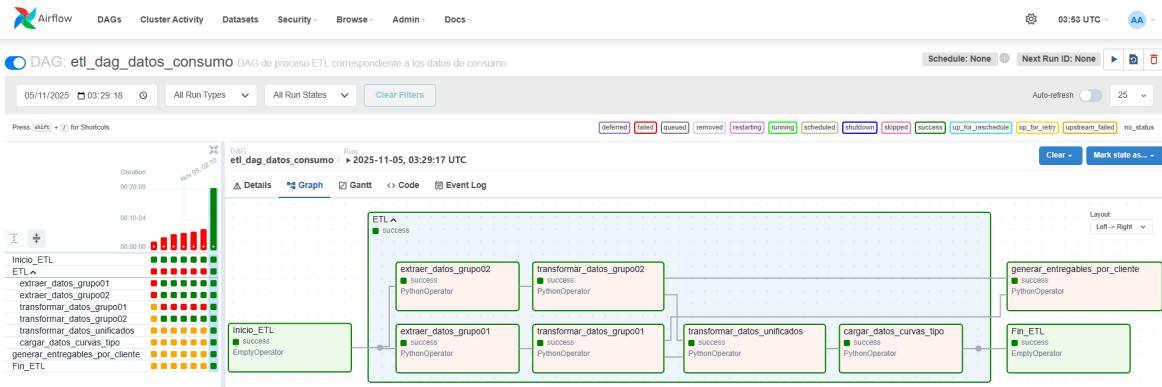


Figura 2.10: DAG de proceso ETL correspondiente a los datos de consumo

Con el propósito de complementar la vista general de la Figura 2.10, en la Figura 2.11 se describen las operaciones concretas que cada tarea realiza durante las fases

del proceso ETL, así como las tareas que integran el DAG. Este desglose permite especificar el alcance de cada tarea sobre el conjunto de datos y así brindar un flujo organizado para la elaboración detallada de los procesos de selección, limpieza, construcción, integración y formateo de los datos que se describen en las siguientes secciones.

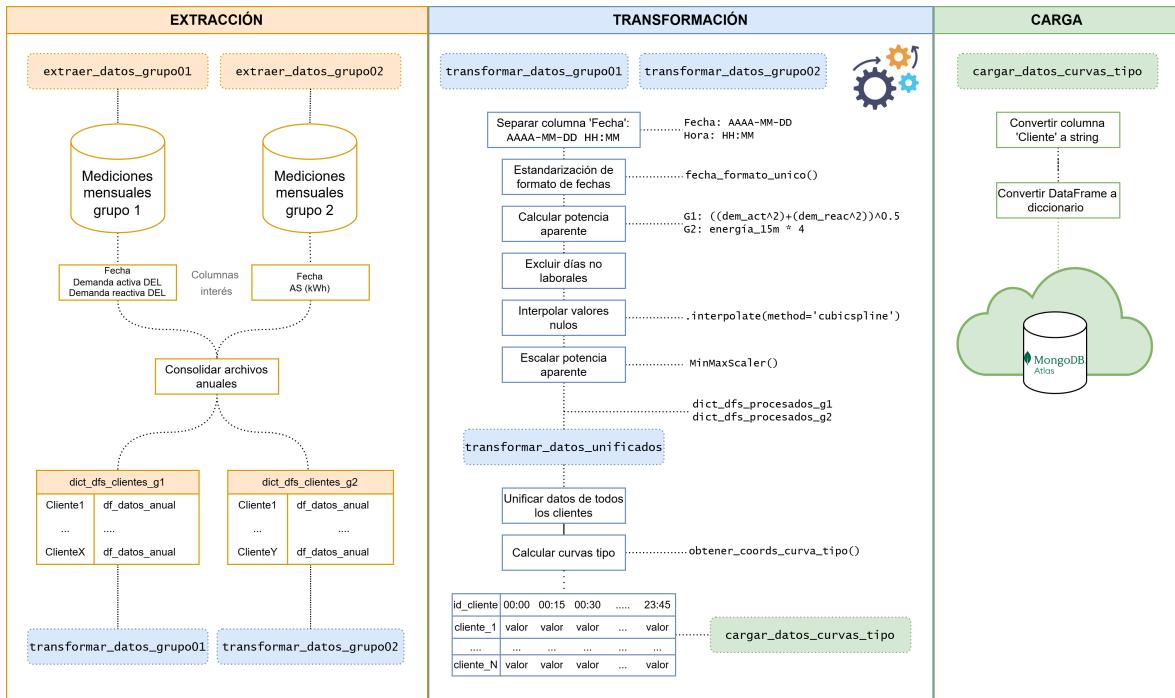


Figura 2.11: Desglose operacional del DAG de ETL: descripción de tareas.

Seleccionar los datos

En esta tarea, se hizo la selección del conjunto de datos que serán utilizados en las etapas siguientes del análisis, garantizando que coincidieran con las metas establecidas en la fase de entendimiento del negocio.

Tomando en cuenta la estructura observada en el análisis exploratorio de los datos, se establecieron dos flujos paralelos de procesamiento dentro del pipeline ETL implementado en Apache Airflow:

- Grupo 01: Constituido por clientes cuyos archivos incluían únicamente valores de demanda activa y reactiva, las columnas de interés son “Demanda activa DEL” y “Demanda Reactiva DEL”.
- Grupo 02: Constituido por clientes cuyos archivos solo incluían mediciones de energía aparente acumulada, la columna de interés es “AS (kWh)”.

Las tareas `extraer_datos_grupo01` y `extraer_datos_grupo02` se utilizaron para gestionar cada grupo independientemente en el DAG `etl_dag_datos_consumo`. Los

archivos de medición mensual de cada cliente fueron extraídos desde el directorio /data para luego consolidar dicha información en un solo conjunto anual por cada cliente. En este proceso, se conservaron solo las columnas relevantes para el análisis (Demanda activa DEL, Demanda reactiva DEL, AS (kWh) y Fecha).

Como resultado de las tareas de extracción de los datos se obtuvieron los diccionarios (clave: valor) dict_dfs_clientes_g1 y dict_dfs_clientes_g2, dentro de los cuales:

- La clave es el identificador del cliente, puede ser su nombre designado o un código numérico.
- El valor asociado a cada clave es un DataFrame que contiene todos sus registros del año consolidados.

Limpiar los datos

La tarea de limpieza tuvo como objetivo asegurar la coherencia de los registros de medición, corrigiendo errores estructurales y preparando la información durante el proceso de transformación. Esta etapa de limpieza se llevó a cabo mediante las tareas de transformación del DAG, las cuales fueron ejecutadas independientemente para cada grupo de clientes respectivamente.

Las actividades específicas que se llevaron a cabo durante el proceso de limpieza de los datos fueron las siguientes:

1. Estandarización de formatos de fecha: Todos los formatos de fecha detectados fueron estandarizados en uno solo (YYYY/MM/DD).
2. Tratamiento de valores nulos: Se empleó un interpolador spline cúbico para llenar los valores faltantes identificados en tareas anteriores.
3. Exclusión de días no laborables y feriados
4. Eliminación de separadores de miles

La elección del método de interpolación mencionado anteriormente se basó en el hecho de que ha sido empleado en trabajos similares, como en [31], que resalta los buenos resultados que ofrecen los splines cúbicos al llenar datos ausentes en series temporales; y en [32], que muestra cómo este método de interpolación posibilita conservar la tendencia general de los datos sin provocar alteraciones bruscas o valores incoherentes. Por los motivos expuestos anteriormente, se consideró una opción viable para el presente análisis.

Construir los datos

En esta tarea se generaron las variables derivadas requeridas para representar de manera uniforme el comportamiento energético de los clientes no regulados. La meta principal fue la creación de atributos que posibilitaran el análisis del consumo de una forma comparable entre todos los clientes, independientemente del tipo de medición disponible (potencias o energía).

Con este propósito, se implementaron las transformaciones correspondientes en las tareas del DAG, donde se crearon las variables potencia aparente y potencia aparente escalada.

Para el primer grupo de clientes, dado que poseen mediciones de potencia activa y reactiva (P y Q), la potencia aparente puede ser calculada aplicando el teorema de pitágoras:

$$S = \sqrt{P^2 + Q^2} \quad (2.1)$$

Por otro lado, el segundo grupo de clientes posee mediciones de energía aparente acumulada (AS (kWh)), en este caso, la potencia aparente puede ser calculada multiplicando dicho valor por cuatro (debido a que es intervalos de 15 minutos):

$$S = E_{15\text{min}} \times 4 \quad (2.2)$$

Una vez calculada la potencia aparente para los dos grupos de clientes haciendo uso de las ecuaciones (2.1) y (2.2), se procede a normalizar con el fin de garantizar la comparabilidad entre curvas de carga con diferentes magnitudes de consumo, para esto se utilizó la técnica de escalado mínimo-máximo, la cual transforma los valores de potencia aparente dentro de un rango definido, que en este caso es [0, 1].

Se decidió implementar un escalado individual para cada día de cada cliente, ya que existen días en los que la demanda es mucho más alta que en otros. Con este método se garantiza una normalización balanceada entre los diferentes días, evitando que los días con valores altos 'aplanen' al resto de registros. El procedimiento consiste en escalar cada medición con respecto a los valores mínimo y máximo correspondientes a su propio día, siguiendo la ecuación (2.3):

$$S_{\text{escalada}} = \frac{S - S_{\text{min}}}{S_{\text{max}} - S_{\text{min}}} \quad (2.3)$$

La normalización resulta esencial en este caso, debido a que posibilita que los algoritmos de clustering se enfoquen exclusivamente en la forma de la curva de consumo y no en su magnitud, evitando que clientes con demandas altas influyan de manera desproporcionada en los resultados del análisis.

La elección de escalar cada día de manera individual fue el resultado de llevar a cabo un estudio comparativo entre la normalización global aplicada al año completo y la normalización diaria que se empleó. La Figura 2.12 evidencia que la normalización global tiende a suavizar la curva característica debido a que hay días con valores significativamente más altos respecto al resto. El comportamiento mencionado anteriormente es particularmente común en clientes no regulados, cuyos patrones de consumo presentan una alta variabilidad al estar vinculados con su actividad económica. En contraste, la normalización diaria mantiene la estructura relativa, dando como resultado una representación más precisa del patrón de consumo característico asociado al cliente.

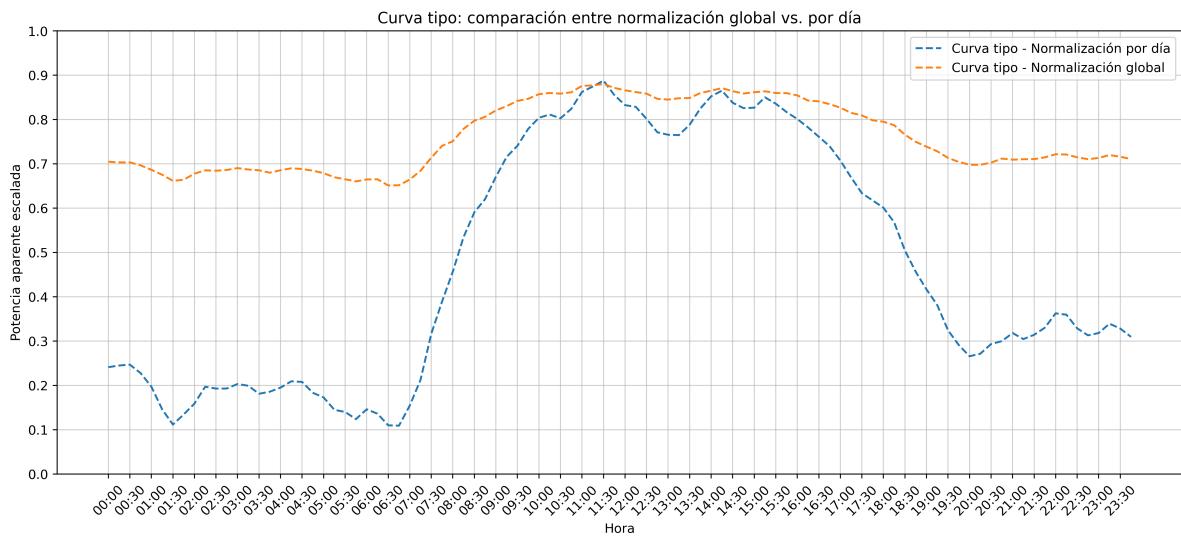


Figura 2.12: Comparación entre la curva tipo resultante de aplicar normalización diaria y normalización global.

Integrar los datos

Una vez concluido el de transformación individual de cada grupo de clientes (selección, limpieza y construcción), se procedió a unificar los datos procesados provenientes de ambos grupos, con el objetivo de consolidar un único conjunto de datos global y proceder con el cálculo de las curvas características representativas anuales. La tarea `transformar_datos_unificados` fue la encargada de realizar dicha integración, la cual forma parte del DAG `etl_dag_datos_consumo`. En dicha tarea se

combinaron los diccionarios generados en los flujos de transformación individuales de cada grupo en un único diccionario consolidado.

Una vez consolidada toda la información, se procedió a obtener la curva tipo, con ayuda de la función `obtener_coords_curva_tipo` se realizó una agregación por hora, calculando la mediana de la potencia aparente escalada a lo largo de todos los días del año. Se hizo uso de la mediana debido a que es una medida robusta frente a valores atípicos, los cuales son propios e inherentes al comportamiento energético de los clientes no regulados.

Para concluir, se integraron las curvas tipo de todos los clientes en un solo DataFrame consolidado, en el cual cada fila corresponde a un cliente y cada columna a una marca temporal única, desde las 00:00 hasta las 23:45 en intervalos de 15 minutos.

Formatear los datos

Para concluir la fase de preparación de los datos, se realizaron las modificaciones finales al formato del conjunto de curvas tipo generado en la tarea anterior, con el objetivo de preparar la información para su posterior almacenamiento y modelado. Estas operaciones se ejecutaron en la tarea `cargar_datos_curvas_tipo`, la cual es parte del DAG `etl_dag_datos_consumo`. Las actividades realizadas fueron:

1. Se comprobó que cada registro del DataFrame contara un identificador exclusivo para el cliente, convirtiendo el campo 'Cliente' a tipo cadena de caracteres, debido a que dicho identificador puede ser alfanumérico.
2. Se transformó el DataFrame consolidado de las curvas tipo en un diccionario orientado a registros, lo que permitió insertar los documentos directamente en la base de datos de MongoDB Atlas.

Finalmente, se creó un índice sobre el campo "Cliente" con el fin de optimizar consultas dentro de la base de datos y se procedió a cargar los registros en la colección `CurvasTipoAnuales`, dando por finalizada la fase de preparación de los datos y obteniendo así un conjunto de datos completamente limpio, estructurado, formateado y preparado para su uso en la fase de modelado.

Modelado

La fase de modelado es el núcleo de CRISP-DM, en la que se crean los modelos que serán aplicados sobre el conjunto de datos ya preparado. En esta fase, la

comprensión de los datos y del negocio se traduce en un grupo de algoritmos que pueden detectar patrones de consumo a partir de las curvas de carga características generadas en la fase anterior. Siguiendo la guía de CRISP-DM, es necesario escoger la técnica de modelado a utilizar y definir previamente los criterios mediante los cuales se evaluarán los resultados, antes de implementar cualquier modelo. Posteriormente se desarrollan los modelos y se analizan sus resultados de acuerdo con los criterios técnicos y del negocio.

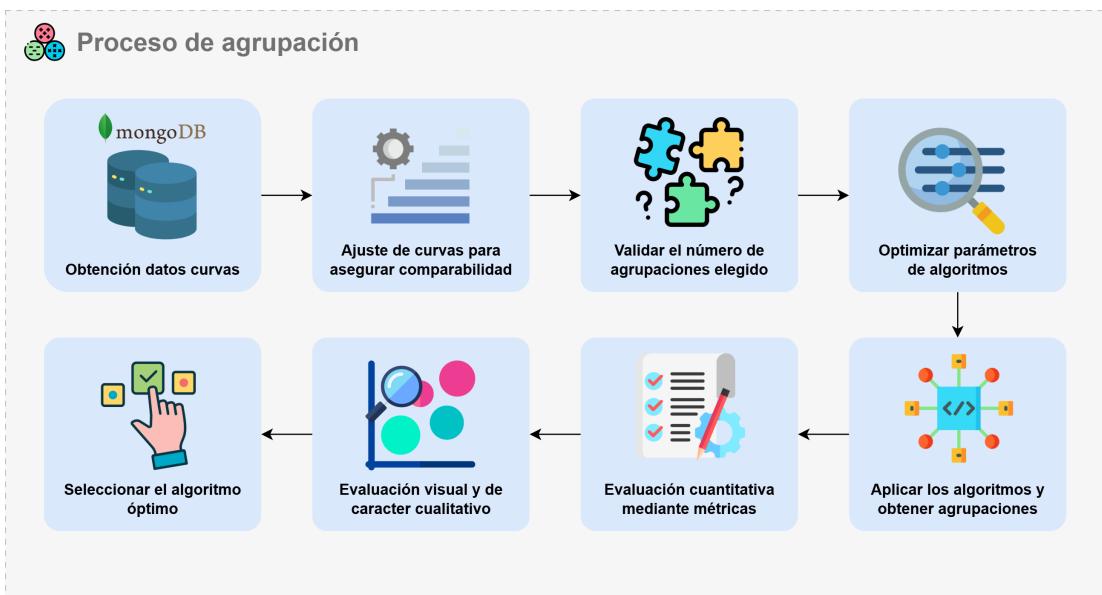


Figura 2.13: Esquematización de proceso de agrupación y sus etapas.

La Figura 2.13 presenta las diversas tareas que intervienen en el proceso de agrupación que se va a desarrollar en el presente componente, el gráfico fue creado en base al cuaderno de trabajo utilizado para realizar clustering, el cual se encuentra disponible en el Anexo III. El diagrama ilustra la secuencia de trabajo que se implementó: obtener las curvas desde la base de datos, ajustar las curvas para comparabilidad, determinar el número óptimo de agrupaciones, optimizar parámetros de los algoritmos, aplicar los algoritmos, generar agrupaciones, evaluar las agrupaciones (tanto cuantitativa como cualitativamente) y seleccionar el algoritmo con el mejor desempeño.

Seleccionar técnica de modelado

De acuerdo con la metodología CRISP-DM, la selección de la técnica de modelado busca identificar los algoritmos más adecuados para abordar el problema. Esto se hace considerando los objetivos del análisis y la naturaleza de los datos. En este estudio, el objetivo del modelado es segmentar a los clientes no regulados del sector eléctrico según la forma de su curva de carga característica anual.

Los algoritmos de agrupamiento son una opción ampliamente validada para detectar patrones de consumo energético. La agrupación de usuarios con comportamientos semejantes y el análisis de la demanda desde un punto de vista tanto operativo como de planificación son posibles gracias al clustering aplicado a perfiles de carga, según han demostrado varias investigaciones [33], [34].

Los algoritmos de agrupamiento son un opción bastante confiable para identificar patrones de consumo energético. Mediante el uso de técnicas de clustering sobre perfiles de carga es posible la agrupación de usuarios con comportamientos similares, a la vez que se puede analizar la demanda desde los puntos de vista operacional y de planificación. Diversos estudios han demostrado la efectividad de este enfoque en el análisis de perfiles de consumo eléctrico [33], [34].

Con el fin de comparar el rendimiento de varios enfoques de agrupamiento sobre un mismo conjunto de datos, se eligieron cuatro algoritmos que pertenecen a distintas familias metodológicas: K-Means, Gaussian Mixture, Birch y Spectral Clustering. A continuación se explica la razón por cual se ha escogido cada algoritmo para el desarrollo del componente:

1. K-Means: Se eligió el algoritmo K-Means por su uso extenso en aplicaciones eléctricas, sobre todo para segmentar perfiles de consumo y obtener curvas de carga que sean representativas. Según [35], K-Means es un referente básico en las investigaciones de agrupamiento de la demanda eléctrica, debido a su sencillez, su eficiencia computacional y lo fácil que es analizar los resultados. K-Means es un modelo apropiado para iniciar la evaluación de agrupaciones de curvas de carga normalizadas en el marco de este trabajo, pues posibilita reconocer grupos de clientes con conductas semejantes de manera directa. Además, su inclusión posibilita establecer una línea de referencia para comparar el rendimiento de algoritmos más complejos, como sugieren los estudios comparativos sobre técnicas de segmentación de patrones de carga [33].
2. Gaussian Mixture: Se eligió el modelo de mezclas gaussianas como un enfoque probabilístico alternativo al método K-Means, que se basa en centroides. La literatura indica que los patrones de consumo eléctrico pueden tener superposiciones entre grupos, sobre todo en clientes cuyos comportamientos son cambiantes, lo cual puede restringir la eficacia de los métodos estrictamente deterministas [36]. Gaussian Mixture Model posibilita determinar si una representación probabilística de los clústeres proporciona un perfil más preciso de

las curvas de carga de los clientes no regulados. Considerar su inclusión es relevante para examinar situaciones en las que los patrones de consumo no están bien definidos, lo cual es frecuente en este tipo de clientes, según se ha reportado en investigaciones anteriores [33].

3. Birch: Se eligió el algoritmo Birch por su capacidad de gestionar conjuntos extensos de datos con eficiencia y por su método jerárquico. Según la literatura, Birch es particularmente eficaz en aplicaciones que necesitan estabilidad y escalabilidad ante variaciones moderadas en los datos. Esto sucede, por ejemplo, en las investigaciones sobre consumo de electricidad con diversos clientes y mediciones temporales [35]. Birch posibilita la evaluación de un método alternativo de agrupamiento que no se basa solamente en la minimización de distancias globales; más bien, va construyendo una estructura resumida de los datos, lo cual permite examinar si un enfoque jerárquico posibilita la identificación de patrones de consumo que podrían no ser evidentes a través de algoritmos que se basan en centroides o en probabilidad.
4. Spectral Clustering: Con el fin de analizar un método que se basa en relaciones de similitud más sofisticadas entre las curvas de carga, se eligió finalmente el algoritmo Spectral Clustering. Según la literatura, este tipo de algoritmo es especialmente eficaz en casos donde la estructura de los datos no muestra separaciones lineales, lo cual puede ocurrir en perfiles de consumo eléctrico con conductas parecidas a lo largo de distintos intervalos de tiempo [33]. Spectral Clustering posibilita examinar si la representación de las curvas de carga como un grafo de similitudes favorece una segmentación más lógica en lo que respecta a la forma de la curva. Su incorporación completa los métodos anteriores y posibilita una comparación más exhaustiva entre diferentes métodos de agrupamiento, como lo indican las últimas revisiones en aplicaciones de agrupamiento para sistemas eléctricos [35].

Generar diseño de prueba

El propósito de crear un diseño de prueba es establecer un esquema de evaluación que posibilite examinar la calidad de los modelos de agrupamiento. Dado que el componente aborda un problema de aprendizaje no supervisado aplicado a curvas de carga eléctricas, la validación de los modelos se lleva a cabo utilizando métricas internas de calidad de agrupamiento, debido a que no se cuenta con etiquetas estándar.

Dado que el objetivo del análisis es encontrar similitudes en la estructura de las curvas de carga, el diseño de prueba se enfoca en evaluar la consistencia interna y la separación entre los clústeres mediante métricas internas de validación. En la Tabla 2.2 se presenta la selección y justificación de las métricas que se van a emplear, conforme a la bibliografía especializada en la agrupación de perfiles de carga eléctrica.

Tabla 2.2: Métricas utilizadas para el diseño de prueba y evaluación de los algoritmos de agrupamiento

Métrica	Descripción y propósito
Correlación intra-clúster	Calcula el promedio de similitud entre cada curva de carga individual y la curva media del clúster correspondiente. Al fundamentarse en la correlación, posibilita una evaluación directa de la semejanza entre las curvas, sin importar su magnitud. Esta métrica es particularmente apropiada para el estudio de perfiles de carga normalizados, cuyo propósito principal es detectar patrones de consumo que sean semejantes a lo largo del tiempo y analizar la uniformidad interna de cada conjunto [34], [37].
Silhouette Score	Analiza la calidad de la asignación de cada curva de carga a su clúster respectivo, teniendo en cuenta al mismo tiempo el agrupamiento interno del clúster y la distancia con respecto a los clústeres adyacentes. Los valores oscilan entre -1 y 1, siendo los que están más cerca de 1 un indicativo de una mejor asignación. En el contexto de las curvas de carga eléctrica, esta métrica posibilita examinar la separación global entre agrupaciones, no obstante, puede mostrar valores moderados como resultado de la superposición natural de los patrones de consumo [34].
Índice Davies–Bouldin (DBI)	Cuantifica la relación entre la distancia de los clústeres y su dispersión interna, penalizando a los agrupamientos con clústeres poco compactos o que están muy próximos. Valores más bajos del índice señalan una mejor calidad de agrupamiento. Este índice se ha empleado en investigaciones de segmentación de datos eléctricos y series temporales, incluso en aplicaciones concretas sobre perfiles de carga, gracias a su aptitud para analizar simultáneamente la separación y compacidad [38], [39].

Métrica	Descripción y propósito
Índice Calinski–Harabasz (CHI)	Evalúa la correlación entre la variabilidad intra-clúster y la variabilidad inter-clúster, priorizando agrupaciones donde los clústeres tienen cohesión interna alta y una separación global adecuada. Valores más altos del índice señalan agrupaciones mejor definidas. Se utiliza esta medida frecuentemente como criterio adicional en la validación interna de algoritmos de agrupamiento y se ha usado en investigaciones sobre segmentación de perfiles de carga eléctrica y análisis de datos energéticos [34], [38].

Construir el modelo

El modelo se construye mediante la implementación práctica de las técnicas de agrupamiento elegidas, siguiendo un flujo de trabajo lógico y organizado:

1. Obtención de los datos

Se emplearon las curvas de carga anuales características que se obtuvieron durante la etapa de preparación de los datos como insumos. Cada cliente no regulado está representado por una curva normalizada, que se crea a partir de datos históricos y se organiza en una matriz. En esta matriz, cada fila corresponde a un cliente y cada columna a un momento específico en el tiempo de la curva de carga. La Figura 2.14 muestra la estructura de esta matriz:

	Cliente	00:00	00:15	00:30	00:45	01:00	...	22:00	22:15	22:30	22:45	23:00	23:15	23:30	23:45
0	90000428	0.522064	0.529247	0.536648	0.537579	0.527764	...	0.462476	0.524822	0.560069	0.564581	0.574944	0.573480	0.582989	0.570380
1	90000537	0.362179	0.319579	0.329867	0.331490	0.315713	...	0.328011	0.333475	0.330486	0.339786	0.349709	0.354084	0.378412	0.351051
2	90000767	0.017565	0.016817	0.016490	0.017572	0.017403	...	0.016312	0.015502	0.016125	0.016614	0.016490	0.015587	0.016907	0.015465
3	90002235	0.003206	0.003219	0.003218	0.003119	0.003222	...	0.002583	0.002625	0.003249	0.003391	0.003438	0.003287	0.003216	0.003370
4	1582648	0.258588	0.268851	0.268407	0.264260	0.280052	...	0.546675	0.544861	0.539987	0.519334	0.499928	0.490922	0.453521	0.455013
...
383	SIGMAPLAST	0.629019	0.630376	0.633978	0.628027	0.628891	...	0.722260	0.691284	0.708349	0.708756	0.718092	0.726780	0.729823	0.736136
384	SINTOFIL	0.570382	0.585197	0.573659	0.597055	0.576146	...	0.618568	0.668834	0.652609	0.609928	0.658959	0.648654	0.672968	0.686380
385	SOCIEDAD INDUSTRIAL RELI CYRANO	0.507828	0.548428	0.602141	0.641924	0.676887	...	0.309762	0.366178	0.461476	0.496907	0.532044	0.538810	0.510710	0.507754
386	TEXTILES TEXSA	0.652327	0.659354	0.673829	0.678084	0.648810	...	0.681969	0.655349	0.669919	0.692258	0.722694	0.710875	0.688904	0.682519
387	VICUNHA ECUADOR	0.018601	0.016489	0.028610	0.032117	0.038422	...	0.092120	0.123175	0.128275	0.147159	0.176647	0.147915	0.163142	0.131872

Figura 2.14: Datos de curvas de carga obtenidas desde MongoDB Atlas.

2. Ajuste de curvas

Se llevó a cabo un ajuste adicional sobre las curvas de carga con el objetivo de garantizar su comparabilidad y evitar la existencia de sesgos a causa de los desplazamientos del eje Y. Este ajuste consistió en desplazar cada curva de tal manera que su posición inicial se encontrara aproximadamente en el origen, utilizando como referencia el promedio de los primeros n registros de cada serie. Este valor referencial fue posteriormente restado de todos los puntos de la curva, manteniendo invariable su forma relativa. De esta manera, los

algoritmos de agrupamiento son capaces de detectar semejanzas basándose únicamente en la forma que tiene la curva de carga.

La Figura 2.15 evidencia que, gracias al ajuste mencionado anteriormente, se eliminan los desfases en el eje Y, conservando solo la forma relativa de cada curva. Este procedimiento asegura que las curvas de carga sean comparables y posibilita que los algoritmos de clustering se basen únicamente en la semejanza de las curvas al realizar la agrupación.

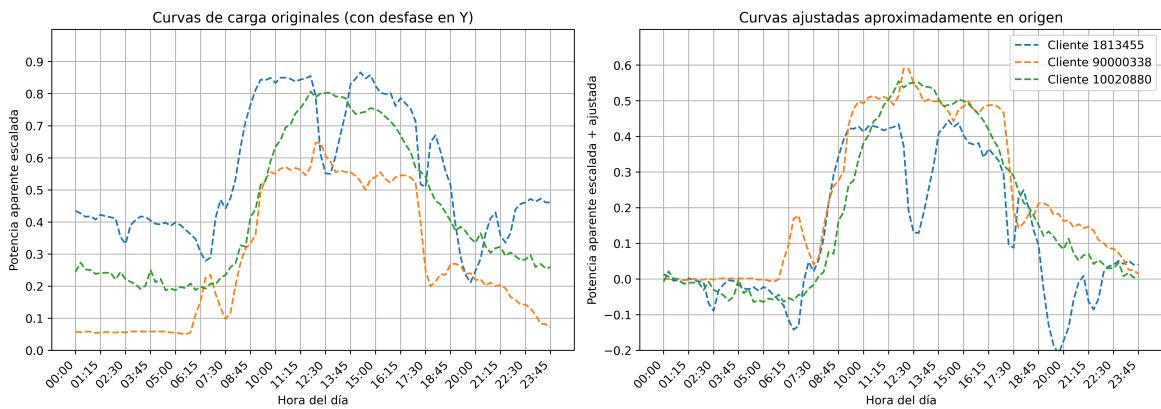


Figura 2.15: Efecto del ajuste vertical de las curvas de carga para garantizar la comparabilidad basada en la forma (tres clientes).

Además, se omitieron métodos de disminución de la dimensionalidad, como el Análisis de Componentes Principales (PCA), para mantener las propiedades temporales originales de las curvas de carga. Esta decisión se basa en lo demostrado en [40]: al analizar esquemas de agrupamiento para perfiles de demanda eléctrica, tanto con reducción de dimensionalidad como sin ella, utilizando datos reales de medidores inteligentes, concluyeron que no utilizar estos métodos permite mantener información importante relacionada a la manera y los patrones horarios del consumo, lo cual ayuda a conseguir agrupaciones más estables y representativas.

3. Definición y validación del número de clústeres

Desde el punto de vista del negocio, la parte interesada manifestó la necesidad de obtener cuatro agrupaciones, esto debido a estudios estadísticos realizados anteriormente. El método del codo se utilizó para validar esta decisión desde una perspectiva técnica, analizando cómo la suma de errores cuadráticos intra-clúster (SSE) cambiaba en función de la cantidad de clústeres. La Figura 2.16 revela que la disminución de la inercia empieza a estabilizarse desde K=4, lo cual demuestra un punto de inflexión evidente en la curva. Este

hallazgo valida el número de clústeres que la parte interesada ha solicitado, al corroborar que cuatro agrupaciones representan un balance apropiado entre la simplicidad del modelo y la compacidad interna.

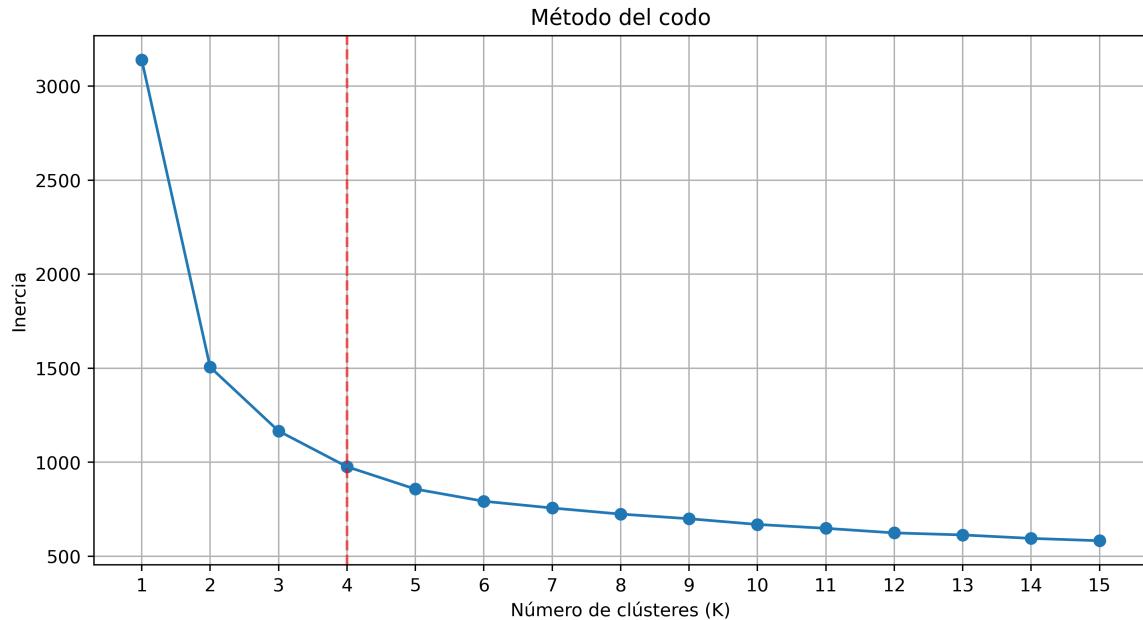


Figura 2.16: Validación del número de clústeres utilizando el método del codo sobre las curvas de carga ajustadas.

4. Hiperparametrización de los algoritmos

Se realizó un proceso de hiperparametrización para cada uno de los algoritmos de agrupamiento elegidos. Este procedimiento consistió en evaluar diferentes combinaciones de parámetros significativos para cada algoritmo, como los criterios de inicialización, el número máximo de iteraciones y parámetros particulares de cada algoritmo. La Tabla 2.3 sintetiza los hiperparámetros más importantes que se tuvieron en cuenta durante el proceso de optimización, los cuales fueron elegidos según su efecto en la calidad de las agrupaciones.

Tabla 2.3: Hiperparámetros evaluados para los algoritmos de clustering

Algoritmo	Parámetros	Valores evaluados
K-Means	n_clusters	4
	init	k-means++, random
	n_init	1, 2, 3, 5, 10, 20, 30
Gaussian Mixture	n_components	4
	covariance_type	full, diag, spherical, tied
	n_init	1, 2, 3, 5, 10, 20

Algoritmo	Parámetros	Valores evaluados
BIRCH	n_clusters threshold branching_factor	4 0.05, 0.075, 0.10, 0.125, 0.15, 0.20 10, 20, 25, 30, 40, 50
Spectral Clustering	n_clusters affinity n_neighbors eigen_solver	4 nearest_neighbors 4, 6, 8, 10, 15, 25, 35 arpack, amg

La selección de los parámetros óptimos para cada algoritmo se realizó utilizando métricas internas de validación, priorizando aquellas configuraciones que maximicen la homogeneidad interna de los clústeres. De esta manera, se garantizó que cada algoritmo fuera evaluado bajo condiciones ajustadas a las características de las curvas de carga analizadas.

5. Aplicación de los algoritmos

Una vez definidos los parámetros óptimos, se implementaron los algoritmos K-Means, Gaussian Mixture, Birch y Spectral Clustering sobre el conjunto de curvas de carga ajustadas. En consecuencia, cada algoritmo le otorgó a cada cliente una etiqueta de clúster, lo que produjo cuatro grupos por algoritmo.

Evaluar el modelo

En esta tarea se realizó la evaluación de los modelos de clustering implementados, con el fin de analizar la calidad de los agrupamientos obtenidos a partir de las curvas de carga características, se utilizaron métricas internas de validación definidas en tareas anteriores.

Los resultados cuantitativos y el análisis comparativo de los algoritmos evaluados se presentan en la sección 3.1 “Resultados” del presente documento.

Evaluación

La fase de Evaluación dentro del ciclo que propone la metodología CRISP-DM tiene como propósito validar los resultados obtenidos en la fase de modelado, verificando que los agrupamientos obtenidos cumplan con los objetivos previamente definidos y proporcionen información útil de cara al análisis del consumo energético de los clientes no regulados.

Esta fase adquiere especial relevancia en el contexto del presente trabajo, dado que aquí lo que se procura no es únicamente obtener agrupaciones matemáticamente válidas, sino identificar patrones de comportamiento energético distintos, homogéneos e interpretables, que puedan servir como guía técnica en procesos de planificación energética, de análisis técnico e incluso de toma de decisiones en la EEQ.

Evaluar resultados

La tarea de evaluación de resultados tiene como finalidad determinar en qué medida los modelos de clustering desarrollados cumplen con el objetivo del componente, orientado a la identificación de patrones de consumo energético en los clientes no regulados.

La evaluación de los resultados es realizada mediante un análisis cuantitativo y cualitativo de las agrupaciones obtenidas. Dicho análisis se presenta en la sección 3.1 “Resultados”, donde se analizan las distintas métricas de validación, las visualizaciones de los clústeres y la interpretación energética de los patrones encontrados.

Revisar el proceso

En esta tarea se realizó una revisión integral del proceso seguido a lo largo del desarrollo del presente componente bajo la metodología CRISP-DM, abarcando desde la comprensión del negocio hasta la evaluación de los modelos de clustering.

Se verificó que las fases de preparación de los datos, modelado y evaluación se ejecutaron de manera coherente en función de los requerimientos planteados, y que las decisiones adoptadas en cada etapa se encuentran debidamente justificadas. En particular, se constató que el uso de las curvas tipo como representación del consumo anual, la normalización por día, el ajuste realizado a las curvas, la hiperparametrización de los algoritmos y la elección de métricas de evaluación orientadas a la forma de las curvas fueron adecuadas para el problema abordado.

Asimismo, se confirmó que el proceso ETL implementado garantiza la reproducibilidad del análisis y que los resultados obtenidos pueden ser replicados o actualizados conforme se disponga de nueva información de consumo. No se identificaron tareas críticas omitidas ni inconsistencias en el flujo metodológico seguido.

Determinar siguientes pasos

Con base en la evaluación de los resultados y en la revisión del proceso, se determina que el proyecto ha cumplido satisfactoriamente los objetivos definidos para la segmentación de clientes no regulados del sector eléctrico.

Dado que el alcance del presente componente se centra en el análisis y evaluación de los modelos de clustering, las tareas asociadas a la fase de despliegue se desarrollan de manera adaptada al contexto de este trabajo. Por lo tanto, la implementación del modelo seleccionado en sistemas productivos o plataformas de gestión energética no forma parte del alcance actual.

No obstante, se identifican como futuras líneas de trabajo la integración del modelo de clustering en herramientas de planificación energética, la actualización periódica de las agrupaciones con nuevos datos de consumo y la incorporación de variables adicionales que permitan enriquecer la caracterización de los clientes.

Despliegue

La fase de despliegue en la metodología CRISP-DM tiene como objetivo garantizar que los resultados obtenidos durante el proceso de minería de datos sean documentados, entendidos y puedan ser utilizados de forma práctica. En proyectos orientados a entornos productivos, esta fase abarca la implementación efectiva de los modelos desarrollados. No obstante, en este trabajo el despliegue se plantea de una forma analítica y metodológica, ya que se ha determinado que el alcance del presente componente no abarca la implementación operativa.

Planificar el despliegue

El plan de despliegue no se identifica como una tarea aplicable en el presente componente, ya que el alcance excluye la implementación operativa modelo de agrupamiento en los sistemas productivos de la EEQ, limitándose únicamente en el análisis, evaluación y validación de las técnicas de segmentación de clientes en un entorno controlado.

Planificar monitoreo y mantenimiento

Esta tarea no corresponde a este trabajo, ya que los modelos generados no son implementados en un ambiente de producción, ni se soporta un proceso de trabajo que requiera el monitoreo en tiempo real o un mantenimiento transcurrido un periodo de tiempo.

Producir el reporte final

La elaboración del informe final es el principal mecanismo de despliegue de los resultados obtenidos en este trabajo, ya que el informe debe integrar de forma estructurada y coherente todo el proceso realizado bajo la guía metodológica de CRISP-DM, permitiendo que el trabajo realizado pueda quedar documentado de manera clara, tanto en lo que respecta a los aspectos metodológicos como a los resultados técnicos alcanzados.

El informe debe integrar la descripción del proceso ETL que se diseña para la construcción de curvas de carga representativas anuales, así como las decisiones adoptadas para limpiar, normalizar e interpolar los datos de consumo energético. De la misma manera, se documenta con detalle el proceso de selección e implementación de los algoritmos de aprendizaje no supervisado, detallando los criterios utilizados para validar el número óptimo de agrupaciones y la optimización de los parámetros de cada modelo.

El informe presentará de forma sistemática los resultados obtenidos por cada uno de los algoritmos de clustering, incorporando métricas cuantitativas de evaluación y representaciones gráficas de las agrupaciones generadas, así como de las curvas promedio y de las curvas individuales asociadas a cada clúster. Estas visualizaciones permitirán la interpretación de los patrones de consumo energético y la comparación entre las distintas técnicas evaluadas. El informe se encuentra disponible en el Anexo IV.

Así, el informe final tiene no solo la función de documentar lo que se ha hecho sino también un insumo técnico que puede ser utilizado por la empresa distribuidora de energía para futuros análisis relacionados con la planificación o estudios sobre la demanda eléctrica.

Revisar el proyecto

La revisión del proyecto permite reflexionar sobre el proceso de desarrollo del componente de análisis y segmentación, así como realizar una identificación de algunos elementos técnicos y metodológicos que pueden servir como referencia para trabajos posteriores.

Uno de los principales aprendizajes del proyecto es la importancia de disponer de un proceso ETL sólido y adecuadamente diseñado, lo cual cobra especial relevancia

al trabajar con datos reales de consumo que se caracterizan por tener irregularidades de formato, valores ausentes y comportamientos atípicos propios de los clientes no regulados. En este contexto, la preparación de las curvas de carga fue un punto clave para poder garantizar la calidad y robustez de los agrupamientos obtenidos.

Asimismo, se evidenció la importancia de seleccionar métricas de evaluación que en función de la naturaleza del problema. En particular, el uso de métricas de evaluación basadas en la similaridad de la forma de las curvas permitió evaluar de manera más adecuada la homogeneidad de los clústeres, en comparación con métricas clásicas que no siempre capturan correctamente el comportamiento de series temporales normalizadas.

Finalmente, el proyecto permitió demostrar que la combinación de análisis cuantitativo con herramientas de visualización facilita en gran medida la interpretación de los resultados y la comunicación de los patrones identificados a actores no expertos en minería de datos, aspecto fundamental dentro del sector eléctrico.

3. RESULTADOS, CONCLUSIONES Y RECOMENDACIONES

3.1. Resultados

En este apartado se presentan los resultados obtenidos a partir de la aplicación y evaluación de los algoritmos de clustering sobre las curvas de carga características anuales de los clientes no regulados. Los resultados son producto de las fases de Modelado y Evaluación de la metodología CRISP-DM y permiten comparar el desempeño de los algoritmos implementados, evaluar la calidad de los agrupamientos generados y seleccionar el método más adecuado para la identificación de patrones de consumo energético.

Evaluación cuantitativa de los algoritmos de clustering

Para cuantificar la calidad de los agrupamientos alcanzados y llevar a cabo la comparación del rendimiento de los diferentes algoritmos de clustering, se procedió a realizar una evaluación a partir de la cual se obtuvieron métricas internas de validación. Estas métricas permiten evaluar tanto la coherencia interna de los clústeres como su separación externa, sin necesidad de emplear información etiquetada, lo que las hace idóneas para abordar problemas de aprendizaje no supervisado.

Correlación intra-clúster promedio

La correlación intra-clúster promedio se debe considerar como la principal métrica utilizada a lo largo de este estudio, al definirse como la media de la correlación entre cada curva de carga y la curva media de su clúster, puesto que facilita la evaluación directa de la semejanza en la forma de los perfiles que se agrupan. Esto resulta coherente con los objetivos del análisis, los cuales se hallan orientados a segmentar a los clientes en función de su comportamiento energético, observable a través de la forma de su curva característica anual.

Respecto a esta métrica, se hace hincapié en la similitud estructural de las curvas obtenidas y no en su demanda absoluta, lo cual resulta especialmente adecuado para el análisis de curvas normalizadas, en las que el interés principal está relacionado con la forma del perfil de consumo [34], [37]. De este modo, los valores obtenidos evidencian diferencias en la homogeneidad de los grupos generados, constituyendo así un criterio para valorar la calidad de los agrupamientos.

Métricas internas complementarias de validación

De forma complementaria, se utilizaron las métricas internas de validación ampliamente utilizadas en la literatura como el índice de Calinski-Harabasz (CHI), el índice de Davies-Bouldin (DBI) y el Silhouette Score, con la finalidad de estudiar la compacidad y la separación de los clústeres obtenidos [34], [38]. La Tabla 3.1 muestra, de manera cuantitativa, el rendimiento de los algoritmos analizados.

Tabla 3.1: Resultados de las métricas de evaluación para los algoritmos de clustering

Métrica	K-Means	GMM	Birch	Spectral
Correlación intra-clúster promedio	0.7399	0.7431	0.7515	0.7533
Davies-Bouldin Index	1.1791	1.3374	1.2177	1.1536
Calinski-Harabasz Index	283.50	255.81	253.33	270.12
Silhouette Score	0.2624	0.2214	0.2326	0.2590

Los valores del Silhouette Score son bajos para todos los algoritmos, fenómeno habitual en los datos reales de demanda eléctrica, donde existe gran variabilidad temporal así como densidades heterogéneas [40]. En este contexto, el puntaje de silueta debe ser considerado como un criterio adicional a la correlación intra-clúster promedio, pues es empleada como el principal indicador de calidad debido a su alineación directa con la similitud estructural de las curvas de carga. En este sentido, el algoritmo Spectral Clustering es el que logra la máxima correlación intra-clúster promedio, exhibiendo un desempeño similar en las demás métricas.

Según los resultados obtenidos, se puede observar que el algoritmo Spectral Clustering es el que alcanza el mayor valor de correlación intra-clúster promedio entre los métodos de clustering analizados. Al tratarse de una métrica directamente vinculada con el criterio de similitud de la forma de la curva de carga, se la considera como el principal indicador de la calidad del agrupamiento. Asimismo, su comportamiento en las métricas complementarias es comparable al del resto de los algoritmos evaluados, sin evidenciar diferencias que desaconsejen su utilización.

Evaluación cualitativa y análisis visual de los clústeres

La evaluación de los resultados posibilita establecer el grado en que los modelos de agrupamiento creados logran el objetivo fundamental del análisis, que consiste en agrupar a los clientes no regulados según la forma de su curva característica anual de consumo. La evaluación se lleva a cabo desde un punto de vista multidimensional, pues se basa en la diferenciabilidad de las curvas logradas, la homogeneidad

interna y separación entre clústeres, además de la interpretabilidad energética de los patrones detectados. Esto posibilita una apreciación global de la coherencia y consistencia de los agrupamientos conseguidos.

Diferenciabilidad: Curvas obtenidas por clúster

Los resultados de agrupación obtenidos utilizando los algoritmos K-Means, Gaussian Mixture, Birch y Spectral Clustering evidencian agrupamientos consistentes para todos los casos, obteniendo curvas tipo claramente distinguibles para cada uno de los clústeres, lo que indica que los algoritmos han sido capaces de captar similitudes reales en cuanto a la forma de las curvas de carga de los clientes, sin verse influenciados por diferencias asociadas al consumo absoluto, aspecto directamente vinculado con las decisiones que se tomaron en las fases de preparación y de modelado de los datos. Al analizar conjuntamente las curvas tipo promedio obtenidas a partir de cada una de las técnicas de agrupamiento, se observan patrones de consumo diferenciables, independientemente del algoritmo empleado, tal y como se puede observar en la Figura 3.1.

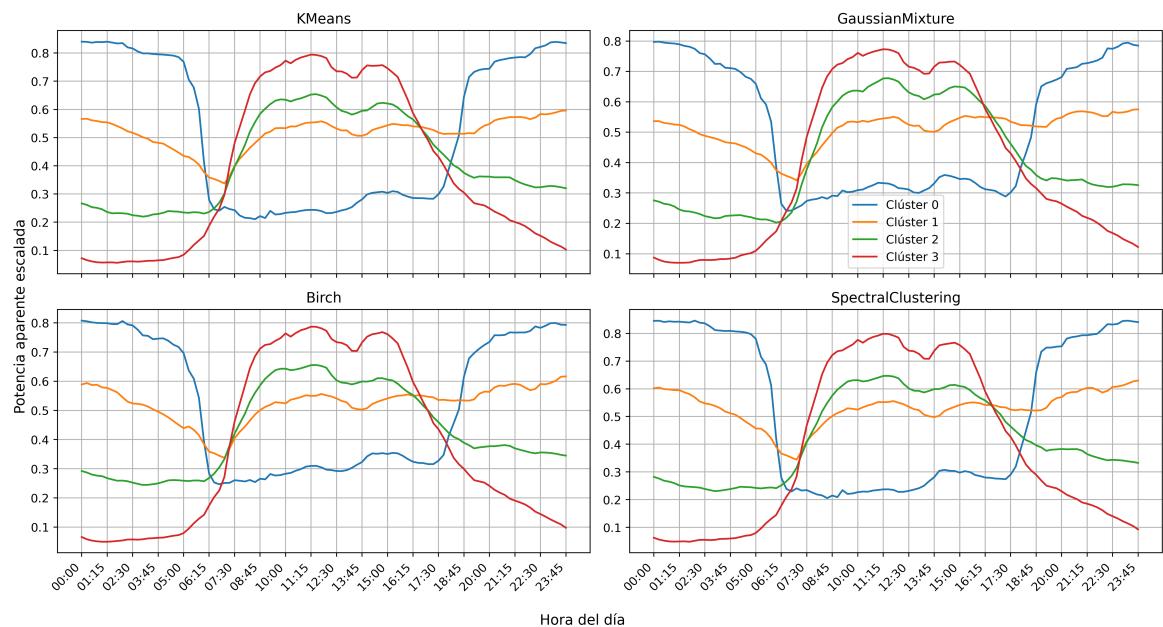


Figura 3.1: Curvas tipo representativas obtenidas para cada clúster de cada algoritmo de clustering.

Es cierto que existen pequeñas variaciones en las formas de las curvas obtenidas por las distintas técnicas; sin embargo, el hecho de que todas ellas coincidan en identificar comportamientos energéticos claramente diferenciables refuerza la idea de que el enfoque utilizado es sólido y que, por tanto, es adecuada la decisión de evaluar de manera comparativa diferentes técnicas de clustering.

Evaluación de homogeneidad y separación mediante PCA

Con el objetivo de complementar el análisis de clústeres, se aplicó Análisis de Componentes Principales (PCA) para proyectar las curvas características de los clientes en un espacio bidimensional. El uso de PCA responde únicamente a fines de exploración y visualización, un procedimiento que se utiliza con frecuencia en investigaciones sobre perfiles de consumo energético con el fin de hacer más fácil la representación gráfica de conjuntos obtenidos a través de series temporales [36], [41]. La representación obtenida posibilita el análisis simultáneo de la separación inter-clúster y la concentración de clientes dentro de cada uno. En los cuatro algoritmos estudiados, se puede notar una estructura bien definida, con una compactación interna apropiada y un distanciamiento sutil entre los clústeres, como lo ilustra la Figura 3.2.

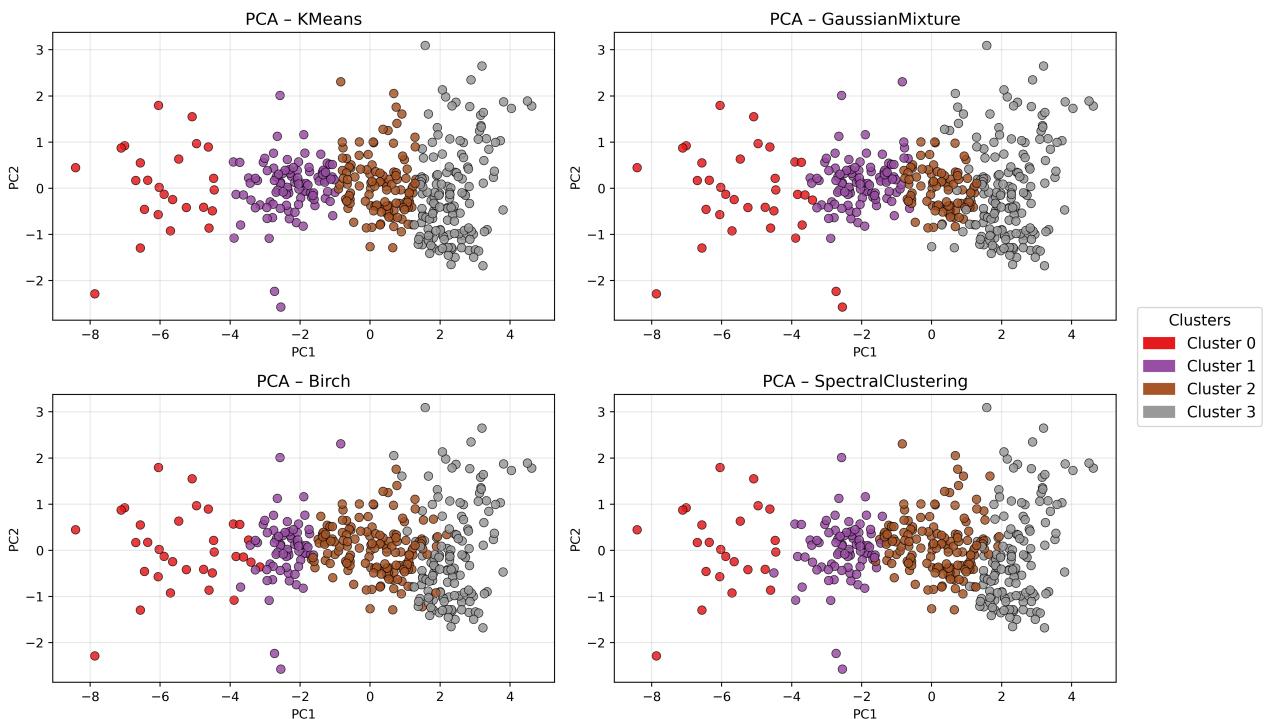


Figura 3.2: Proyección PCA bidimensional de las agrupaciones generadas por algoritmo.

Esta estructura confirma que los clientes agrupados presentan similitudes en la forma de sus curvas de carga y que los grupos obtenidos no muestran solapamientos evidentes. Este resultado respalda la pertinencia de la correlación intra-clúster promedio como métrica principal de evaluación, ya que permite medir de forma directa la semejanza entre cada curva individual y la curva tipo representativa de su clúster, en concordancia con el objetivo del análisis.

Curvas individuales por clúster

A partir de los resultados obtenidos en 3.1 “Evaluación cuantitativa de los algoritmos de clustering” del presente trabajo, donde se evaluó de manera cuantitativa la calidad de las agrupaciones generadas por los distintos algoritmos de clustering utilizando métricas internas de validación, a continuación se procede a la interpretación de los resultados de agrupamiento. Más concretamente, el análisis se enfoca en los clústeres obtenidos a partir de la aplicación del algoritmo Spectral Clustering, previamente seleccionado como el modelo que presenta el mejor desempeño.

En la Figura 3.3 se presentan, para cada clúster obtenido mediante la aplicación del algoritmo Spectral Clustering, las curvas tipo de todos los clientes que conforman cada clúster respectivamente.

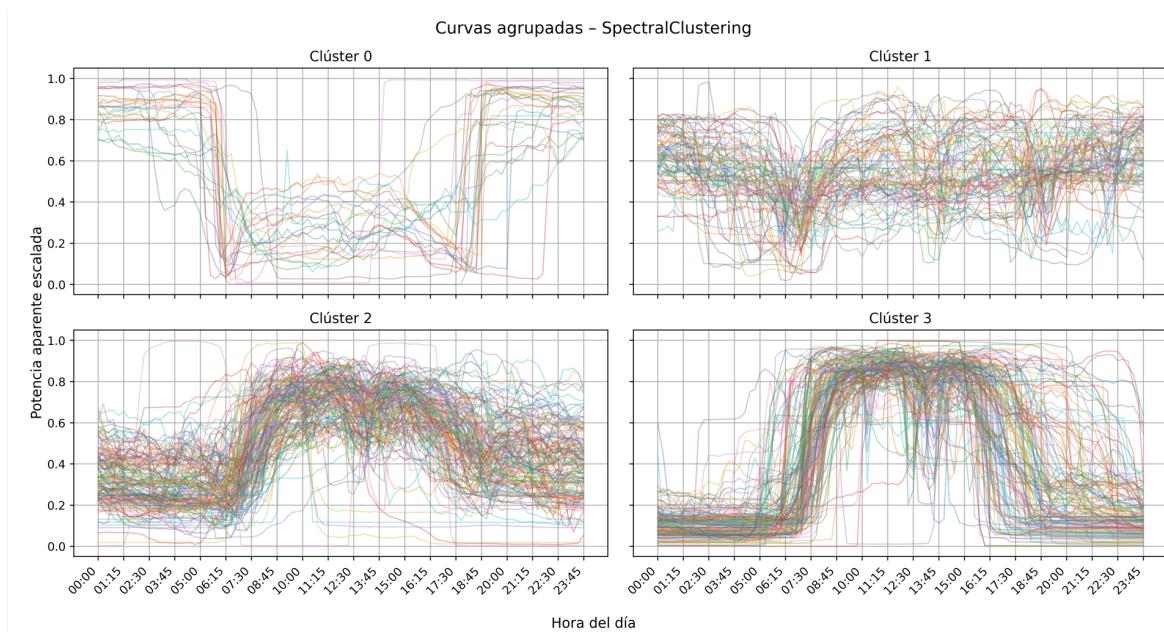


Figura 3.3: Spectral Clustering: curvas de carga de los clientes agrupados por clúster.

Interpretación energética de los clústeres obtenidos

Al observar las curvas de carga por clúster de manera conjunta, es posible percibir la coherencia interna de los grupos obtenidos y distinguir entre diferentes patrones de consumo. Siguiendo esta observación, a continuación se muestra la interpretación energética de cada clúster que ha sido identificado a través de Spectral Clustering.

- Clúster 0: Este clúster agrupa aquellos clientes cuyo perfil de consumo muestra niveles altos durante la madrugada, la forma que presentan estas curvas

nos indican patrones de funcionamiento que no corresponden a horarios laborables, lo cual indica que estos clientes realizan sus actividades en horarios nocturnos.

- Clúster 1: Los clientes agrupados en el presente clúster presentan un comportamiento estable a lo largo del día, sin variaciones claramente definidas. La forma de la curva refleja una dispersión ordinaria, lo cual indica que estos clientes mantienen su actividad sin una estructura horaria dominante.
- Clúster 2: Este clúster contiene aquellos clientes cuya forma de curva denota actividad diurna, caracterizada por un incremento sostenido del consumo en horas de la mañana, alcanzando valores elevados durante el día y disminuyendo de manera gradual y progresiva hacia la tarde y noche. Se espera que estos clientes mantengan patrones de consumo concentrados en el período diurno, sin transiciones abruptas entre estados de baja y alta demanda.
- Clúster 3: Los clientes pertenecientes a este clúster presentan una forma de curva escalonada, con consumos considerablemente bajos durante la madrugada y un ascenso abrupto en horas de la mañana, una meseta estable durante el día y un descenso igualmente abrupto al finalizar la jornada. Se espera que estos clientes presenten patrones de consumo estructurados y delimitados en intervalos de tiempo claramente definidos.

La clara diferenciación entre estos comportamientos confirma que los clústeres obtenidos no solo son estadísticamente consistentes, sino que además presentan una interpretación energética clara y útil, lo cual es fundamental para su aplicación en escenarios reales del sector eléctrico.

Selección final del mejor algoritmo

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 de carga de cada clúster, en la línea de la finalidad básica del análisis).

En consecuencia, el modelo escogido constituye la base para la interpretación de los patrones de consumo eléctrico encontrados en los consumidores no regulados y su posible utilización en los procedimientos de análisis y de planificación energética.

3.2. Conclusiones

El presente Trabajo de Integración Curricular ha permitido la evaluación e implementación de modelos de aprendizaje no supervisado para la segmentación de clientes no regulados del sector eléctrico, evidenciando que el uso de curvas características anuales de consumo se trata de un procedimiento totalmente viable desde una perspectiva técnica y siendo superior a aquellos métodos más simples como promedios o clasificaciones rígidas. La alternativa propuesta permitió poner de manifiesto la variabilidad y complejidad real que se esconden detrás de los patrones de consumo, logrando agrupaciones homogéneas y coherentes que muestran conductas energéticas diversas y valiosas para el estudio y planificación energética.

El procedimiento llevado a cabo para el levantamiento de requerimientos y el procesamiento de los datos de consumo energético posibilitaron la conversión de registros heterogéneos en curvas anuales de carga comparables y representativas, contribuyó a transformar el registro en un conjunto de datos representativos y comparables. El proceso ETL, ejecutado a través de Apache Airflow, implementó tareas de limpieza, interpolación, exclusión de días no laborables y normalización diaria, asegurando calidad y consistencia en los datos. Estos resultados demuestran que preparar los datos de manera apropiada es un paso sumamente importante para garantizar la fiabilidad de análisis posteriores.

La revisión literaria de los algoritmos de clustering permitió la identificación de los principios de funcionamiento, suposiciones y parámetros relevantes para su aplicabilidad en la segmentación de curvas de carga del sector eléctrico. Este análisis hizo posible la elección de los algoritmos y la configuración precisa de sus parámetros, brindando así una base teórica robusta para su implementación y posterior comparación en el caso de estudio del presente componente

La adopción de la metodología CRISP-DM proporcionó un marco ordenado y sistemático para orientar las diferentes etapas del análisis de datos. La adaptación realizada en la fase final, enfocada en la elección de la mejor agrupación en función de métricas de calidad y homogeneidad, fue coherente y se alineó a la naturaleza exploratoria del aprendizaje no supervisado. Esto posibilitó garantizar un procedimiento ordenado, replicable y que fuese coherente con los objetivos técnicos del proyecto.

El uso del método del codo fue fundamental para establecer el número óptimo de agrupaciones a partir del análisis de la varianza intra-clúster. Considerando que la parte interesada había manifestado la necesidad de generar cuatro agrupaciones, la aplicación de este método permitió validar dicha elección, dado que a partir de dicho número la incorporación de más clústeres no produce mejoras relevantes. De este modo, se dispuso de un criterio técnico que respalda la elección del número de clústeres a emplear en la aplicación de los diferentes algoritmos de clustering.

La evaluación de los resultados obtenidos se realizó mediante un enfoque mixto que integró análisis cualitativos y cuantitativos. La combinación de métricas como el puntaje de Silhouette, el índice de Davies-Bouldin, el índice de Calinski-Harabasz y la correlación intra-clúster promedio permitieron valorar la separación entre los clústeres y la coherencia interna desde una perspectiva cuantitativa. En contraste, para la evaluación cualitativa se utilizaron visualizaciones como proyecciones bidimensionales mediante PCA, curvas individuales por clúster y curvas promedio, lo cual hizo posible que los patrones de consumo detectados fueran interpretados claramente y que las agrupaciones conseguidas fueran validadas visualmente.

3.3. Recomendaciones

Se recomienda que, en un uso operativo posterior del TIC, se cuente con un mecanismo que permita actualizar de forma periódica las curvas características de los consumos anuales, con el fin de que los procesos de segmentación reflejen de mejor manera los cambios en los hábitos de consumo que puedan presentarse por factores económicos, productivos o tecnológicos en los clientes no regulados.

Se sugiere continuar utilizando métricas enfocadas en la similitud estructural de las curvas de carga, como la correlación intra-clúster, ya que durante el desarrollo del TIC se observó que este tipo de métricas ofrece mejores resultados que aquellas sensibles a la variabilidad propia de los perfiles de consumo eléctrico.

Se recomienda reutilizar y ampliar el flujo ETL y la infraestructura de datos implementados para el TIC, considerándolos como una base para futuros análisis energéticos, dado que su diseño modular y automatizado permite incorporar nuevos intervalos de medición sin generar cambios significativos en la estructura general del componente.

Con el objetivo de mejorar la caracterización de los segmentos obtenidos, se recomienda que en futuros proyectos se incluyan variables adicionales, como la actividad económica del cliente, información relacionada con el clima o algunos indicadores socioeconómicos, manteniendo como referencia principal el esquema de segmentación basado en la forma de la curva característica anual.

Se recomienda que el componente desarrollado sea considerado como una base metodológica para trabajos posteriores orientados a análisis más avanzados, tales como estudios de proyección de la demanda, la identificación de clientes con comportamientos atípicos o la aplicación de técnicas de aprendizaje profundo, aprovechando los procesos de procesamiento y organización de datos ya definidos.

4. REFERENCIAS BIBLIOGRÁFICAS

- [1] CONELEC, *Estadística del sector eléctrico Ecuatoriano*, 2012. Obtenido de: <https://www.controlrecursosyenergia.gob.ec/wp-content/uploads/downloads/2021/03/Folleto-Resumen-Estad%C3%ADsticas-2011.pdf>
- [2] B. Moses y O. Akanni, "The Load Curve and Load Duration Curves in Generation Planning," *Proceedings of the Second Australian International Conference on Industrial Engineering and Operations Management, Melbourne, Australia*, 2023. Obtenido de: <https://ieomsociety.org/proceedings/2023australia/245.pdf>
- [3] T. Teeraratkul, D. O'Neill y S. Lall, "Shape-Based Approach to Household Load Curve Clustering and Prediction," *Stanford University*, 2017. arXiv: 1702.01414.
- [4] P.-N. Tan, M. Steinbach y V. Kumar, *Introduction to Data Mining*. Pearson Education Limited, 2014, ISBN: 978-1-292-02615-2.
- [5] J. Han, M. Kamber y J. Pei, *DATA MINING Concepts and Techniques*. Morgan Kaufmann, 2012, ISBN: 978-0-12-381479-1.
- [6] J. M. Moine, A. S. Haedo y S. Gordillo, "Estudio comparativo de metodologías para minería de datos," en *Workshop de Investigadores en Ciencias de la Computación (WICC 2011)*, Universidad Tecnológica Nacional, Facultad Regional Rosario, 2011, págs. 1-9. Obtenido de: http://sedici.unlp.edu.ar/bitstream/handle/10915/20034/Documento_completo.pdf?sequence=1&isAllowed=y
- [7] V. Plotnikova, M. Dumas y F. Milani, "Adaptations of data mining methodologies: a systematic literature review," *PeerJ Computer Science*, vol. 6, e267, 2020. DOI: 10.7717/peerj-cs.267
- [8] S. K. Singu, "ETL Process Automation: Tools and Techniques," *ESP Journal of Engineering & Technology Advancements*, vol. 2, n.º 1, págs. 74-85, 2022, ISSN: 2583-2646. DOI: 10.56472/25832646/JETA-V2I1P110
- [9] S. H. A. El-Sappagh, A. M. A. Hendawi y A. H. E. Bastawissy, "A proposed model for data warehouse ETL processes," *Journal of King Saud University - Computer and Information Sciences*, vol. 23, n.º 2, págs. 91-104, 2011. DOI: 10.1016/S1319-1578(11)00019-X
- [10] W. H. Inmon, *Building the Data Warehouse*, 3rd. John Wiley & Sons, 2002, ISBN: 0-471-08130-2.

- [11] V. Goar, P. S. Sarangdevot, G. Tanwar y D. A. Sharma, "Improve Performance of Extract, Transform and Load (ETL) in Data Warehouse," *International Journal on Computer Science and Engineering*, vol. 2, mayo de 2010.
- [12] J. Wang y F. Biljecki, "Unsupervised machine learning in urban studies: A systematic review of applications," *Cities*, vol. 129, pág. 103925, 2022. DOI: 10.1016/j.cities.2022.103925
- [13] L. Coraggio y P. Coretto, "Selecting the number of clusters, clustering models, and algorithms. A unifying approach based on the quadratic discriminant score," *arXiv preprint*, vol. arXiv:2111.02302, 2021. arXiv: 2111.02302.
- [14] O. Lezhnina et al., "Latent Class Cluster Analysis: Selecting the Number of Clusters," *International Journal of Social Research Methodology (or similar)*, 2022. Obtenido de: <https://PMC.ncbi.nlm.nih.gov/articles/PMC9192797/>
- [15] V. J. Friedman, "A Survey of Popular R Packages for Cluster Analysis," University of Glasgow, inf. téc., 2017, E-print via University of Glasgow repository. Obtenido de: <https://eprints.gla.ac.uk/153580/7/153580.pdf>
- [16] A. A. Wani, "Comprehensive analysis of clustering algorithms: exploring limitations and innovative solutions," *PeerJ Computer Science*, vol. 10, e2286, 2024. DOI: 10.7717/peerj-cs.2286
- [17] A. D. Fontanini y J. Abreu, "A Data-Driven BIRCH Clustering Method for Extracting Typical Load Profiles for Big Data," *2018 IEEE Power & Energy Society General Meeting (PESGM)*, págs. 1-5, 2018. DOI: 10.1109/PESGM.2018.8586431
- [18] A. K. Pathak, M. Chaubey y M. Gupta, "Randomized-Grid Search for Hyperparameter Tuning in Decision Tree Model to Improve Performance of Cardiovascular Disease Classification," *arXiv preprint arXiv:2411.18234*, 2024. arXiv: 2411.18234.
- [19] D. Heras Calvo, *Título del Trabajo Fin de Grado*, Trabajo Fin de Grado, 2023. Obtenido de: https://oa.upm.es/75555/1/TFG_DANIEL_HERAS_CALVO.pdf
- [20] Apache Software Foundation. "Apache Airflow Documentation." Accedido: 8 de septiembre de 2025. Obtenido de: <https://airflow.apache.org/docs/apache-airflow/stable/index.html>
- [21] Docker Inc. "Docker Overview." Accedido: 8 de septiembre de 2025. Obtenido de: <https://docs.docker.com/get-started/docker-overview/>

- [22] Python Software Foundation. “Tutorial de Python 3.13.” Accedido: 8 de septiembre de 2025. Obtenido de: <https://docs.python.org/es/3.13/tutorial/index.html>
- [23] Microsoft Corporation. “Why Visual Studio Code.” Accedido: 8 de septiembre de 2025. Obtenido de: <https://code.visualstudio.com/docs/editor/whyvscode>
- [24] MongoDB Inc. “What is MongoDB?” Accedido: 8 de septiembre de 2025. Obtenido de: <https://www.mongodb.com/es/company/what-is-mongodb>
- [25] P. Weichbroth, “Facing the Brainstorming Theory. A Case of Requirements Elicitation,” Gdańsk University of Technology, Faculty of Management y Economics, Gdańsk, GUT FME Working Paper Series A 12/2016 (42), 2016. Obtenido de: <https://hdl.handle.net/10419/173338>
- [26] M. Sarnovský y P. Bednár, *Application of CRISP-DM Methodology for Customer Segmentation in Electricity Distribution Companies*, Óbuda University Digital Archive, Accessed: 2025-09-09, 2025. Obtenido de: https://oda.uni-obuda.hu/bitstream/handle/20.500.14044/31961/Sarnovsky_Bednar_159.pdf?sequence=1&isAllowed=y
- [27] G. O. Otieno, “A Study of Classification of Electricity Consumers by Electricity Companies in Comparison to Dynamic Data-driven Clustering Based on Consumption Patterns,” Accessed: 2025-09-09, Master’s Thesis, University of Nairobi, 2021. Obtenido de: https://erepository.uonbi.ac.ke/bitstream/handle/11295/157325/Otieno%20G_A%20Study%20of%20Classification%20of%20Electricity%20Consumers%20by%20Electricity%20Companies%20in%20Comparison%20to%20Dynamic%20Data-driven%20Clustering%20Based%20on%20Consumption%20Patterns.pdf?sequence=1&isAllowed=y
- [28] H. Javanshir, M. M. Rashidi y M. Omidi, “Clustering Customers Based on LRFM Model Using Data Mining Approach,” *International Journal of Industrial Engineering & Production Research*, vol. 32, n.º 1, págs. 19-32, 2021, Acces-sed: 2025-09-09. Obtenido de: <https://ijiepr.iust.ac.ir/article-1-1124-en.pdf>
- [29] IBM, “Guía de CRISP-DM de IBM SPSS Modeler,” *International Business Machines Corporation*, 2018. Obtenido de: https://www.ibm.com/docs/es/SS3RA7_18.4.0/pdf/ModelerCRISPDM.pdf

- [30] P. Chapman, R. Kerber, J. Clinton, T. Khabaza, T. Reinartz y R. Wirth, "The CRISP-DM Process Model," *CRISP-DM consortium*, 1999. Obtenido de: <https://mineracaodedados.wordpress.com/wp-content/uploads/2012/12/crisp-dm-no-brand.pdf>
- [31] D. A. Petrushevich, "Review of missing values procession methods in time series data," *Journal of Physics: Conference Series*, vol. 1889, n.º 032009, 2021. DOI: 10.1088/1742-6596/1889/3/032009
- [32] S. Wüst, V. Wendt, R. Linz y M. Bittner, "Smoothing data series by means of cubic splines: quality of approximation and introduction of a repeating spline approach," *Atmospheric Measurement Techniques*, vol. 10, págs. 3453-3462, 2017. DOI: 10.5194/amt-10-3453-2017
- [33] A. Rajabi, M. Eskandari, M. Jabbari Ghadi, L. Li, J. Zhang y P. Siano, "A Comparative Study of Clustering Techniques for Electrical Load Pattern Segmentation," *Renewable and Sustainable Energy Reviews*, 2019. doi: 10.1016/j.rser.2019.109628
- [34] G. Chicco, "Overview and Performance Assessment of the Clustering Methods for Electrical Load Pattern Grouping," *Energy*, vol. 42, n.º 1, págs. 68-80, 2012. DOI: 10.1016/j.energy.2011.12.031
- [35] S. M. Mirafabzadeh, C. G. Colombo, M. Longo y F. Foiadelli, "K-Means and Alternative Clustering Methods in Modern Power Systems," *IEEE Access*, vol. 11, págs. 119 596-119 633, 2023. doi: 10.1109/ACCESS.2023.3327640
- [36] S. Kallel, M. Amayri y N. Bouguila, "Clustering and Interpretability of Residential Electricity Demand Profiles," *Sensors*, vol. 25, n.º 7, págs. 2026, 2025. DOI: 10.3390/s25072026
- [37] W. Labeeuw y G. Deconinck, "Residential electrical load model based on mixture model clustering and Markov models," *IEEE Transactions on Smart Grid*, 2013. doi: 10.1109/TSG.2013.2245489
- [38] N. Li, X. Wu, J. Dong y D. Zhang, "A shape-based clustering algorithm and its application to load data," *Cognitive Computation and Systems*, vol. 5, n.º 2, págs. 109-117, 2023. doi: 10.1049/ccs2.12080
- [39] M. Halkidi, Y. Batistakis y M. Vazirgiannis, "On clustering validation techniques," *Journal of Intelligent Information Systems*, vol. 17, n.º 2, págs. 107-145, 2001.
- [40] M. Jain, T. AlSkaif y S. Dev, "Validating clustering frameworks for electric load demand profiles," *IEEE Transactions on Industrial Informatics*, 2021. doi: 10.1109/TII.2021.3065131

- [41] A. Salmina y P. Stoll, “Dimensionality Reduction and Clustering of Energy Consumption Time Series in Supermarket Buildings,” École Polytechnique Fédérale de Lausanne (EPFL), Course Project Report, 2021, Machine Learning Course Project.

5. ANEXOS

Los anexos del presente trabajo se encuentran disponibles de forma digital a través de un repositorio público en GitHub. En esta sección se presentan los hipervínculos permanentes a los archivos y documentos complementarios utilizados y generados durante el desarrollo del proyecto.

I. Archivo docker-compose.yaml

Disponible en línea

II. Código Python del DAG del proceso ETL

Disponible en línea

III. Cuaderno de trabajo de clustering

Disponible en línea

IV. Reporte final del proyecto de minería de datos

Disponible en línea