

ESCUELA POLITÉCNICA NACIONAL

FACULTAD DE INGENIERÍA EN SISTEMAS

OPTIMIZACIÓN DE SISTEMAS DE INFORMACIÓN EN CONTEXTOS 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, julio 2025

Certificaciones

Yo, **Andrés Zambrano**, 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.

NOMBRE_ESTUDIANTE

Certifico que el presente trabajo de integración curricular fue desarrollado por Andrés Zambrano, bajo mi supervisión.

NOMBRE_DIRECTOR
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 Zambrano

Josafá Aguiar

Dedicatoria

A mis padres celestiales, Dios y la santísima Virgen María, en quienes siempre he depositado toda mi fé y confianza a lo largo de toda mi trayectoria académica.

A 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 supe-
rieron alentarme y darme su apoyo incondicional para seguir adelante, sin lugar a dudas, este, y todos mis logros se los dedico a ustedes.

A Edita Vélez, mi segunda mamá, quien me cuidó durante toda mi niñez, llenándome siempre de amor, mimos y mucho cariño.

A mis padrinos, Franklin Vásquez y Silvana Barba, por acogerme con cariño en su hogar durante mis estudios universitarios, de igual manera, a mis primos, Carolina, Dennis y Pamela, quienes más que primos han sido como hermanos para mí.

A mi primo Jhonny Sánchez, quien ha sido como un hermano para mí, con quien he compartido invaluable momentos durante gran parte de mi niñez. Gracias por ser ese hermano que nunca pude tener, pero que la vida se encargó de darme.

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.

A mis amigos, compañeros de risas, retos e innumerables experiencias, que siempre han estado presentes, tanto en las buenas como en las malas.

A toda mi familia en general, quienes de manera directa o indirecta han contribuido con su granito de arena para formar la persona que soy hoy en día.

Finalmente, 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.

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 mis padres, mis dos grandes tesoros, gracias por creer en mí en todo momento, por demostrarme que con esfuerzo y dedicación todo es posible y, sobre todo, por su amor y apoyo incondicional. Gracias por tanto, gracias por ser mis padres.

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.

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.

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, agradezco a mis amigos Carlos, Alexis, Hernán, Galo, Dilan y los que faltan por nombrar, por hacer que la vida universitaria fuera mucho más llevadera. Gracias por todas las experiencias que compartimos, risas, enojos, tristezas, largas charlas, y sobre todo, la remontada del siglo en sexto semestre.

Índice general

Certificaciones	I
Declaración de autoría	II
Dedicatoria	III
Agradecimientos	IV
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
1.4.1. Sobre el sector eléctrico	4
1.4.2. Minería de datos	5
1.4.3. Proceso ETL	7
1.4.4. Aprendizaje no supervisado	7
1.4.5. Herramientas utilizadas	10
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
2.4.1. Entendimiento del negocio	15
2.4.2. Entendimiento de los datos	19
2.4.3. Preparación de los datos	24
2.4.4. Modelado	33
2.4.5. Evaluación	41
2.4.6. Despliegue	46
3. Resultados, Conclusiones y Recomendaciones	47
3.1. Resultados	47
3.2. Conclusiones	47
3.3. Recomendaciones	47
4. Referencias Bibliográficas	48

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 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 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, 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 cliente. 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 (KMeans, GaussianMixture, Birch y Spectral 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 hará uso de métodos de validación como el método del codo. Por otro lado, para la optimización de estos algoritmos se utilizará la correlación intra-cluster, esta métrica es la más adecuada pues captura de mejor manera la similitud en forma de las curvas agrupadas.
4. Los resultados incluirán la curva de carga representativa de cada cliente, la curva de carga correspondiente al día de máxima demanda, archivos .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.

1.4.1. 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 bastantes 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.

1.4.2. 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].

1.4.3. 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].

1.4.4. 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 número 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:

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

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:

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

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:

$$\text{mín}_T g(W_k(T), B_k(T)) \quad (1.6)$$

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].

1.4.5. 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 Empresa Eléctrica Quito (EEQ) es la administración eficiente de la demanda de sus clientes, especialmente la de aquel segmento que no está regulado, este grupo de clientes 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 en sus curvas de demanda. Esto 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 utilizados 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 resúmen 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. **Comprensión 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. **Comprensión 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 en esta fase la implementación del modelo en un entorno productivo, en este caso, 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 etapas 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.

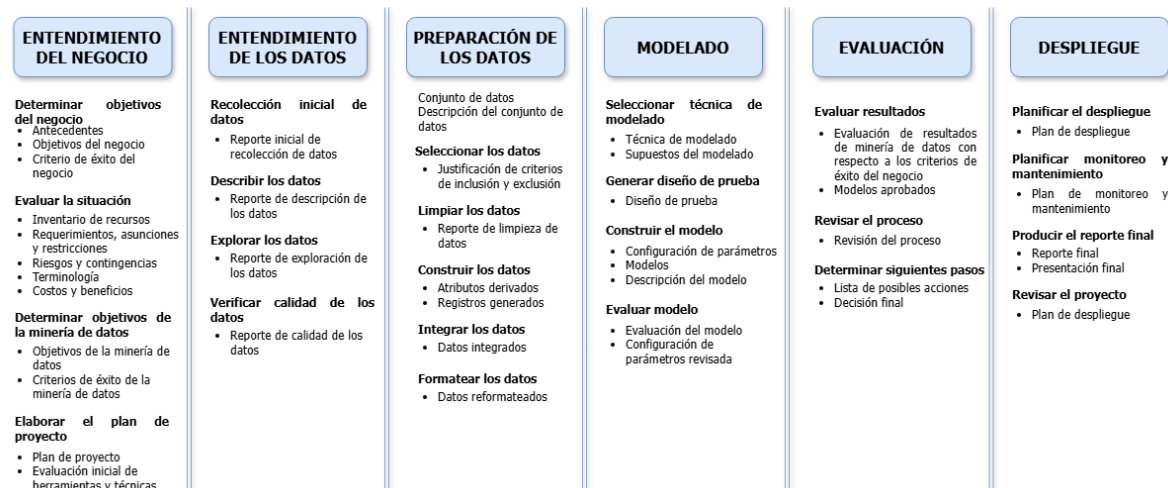


Figura 2.1: Esquematización de las fases y tareas de la metodología CRISP-DM

La Figura 2.1 presenta un esquema que sintetiza las fases y tareas que se llevarán a cabo en el desarrollo del presente componente. La figura es de elaboración propia y fue realizada a partir de los principios establecidos en [30].

2.4.1. Entendimiento del negocio

Levantamiento inicial de información

Se elaboró un mapa mental (ver Figura 2.2) como técnica inicial para recolectar y organizar la información relativa al contexto del negocio, actores involucrados, problemática, requerimientos, los objetivos y las salidas esperadas del proyecto. Esta herramienta posibilitó una visión completa del entorno y la determinación de la trazabilidad entre los elementos empresariales y los objetivos del componente.

Como se puede apreciar, el mapa mental contiene siete partes fundamentales:

- **Problemática:** Identifica las restricciones presentes en la gestión de la demanda de los clientes no regulados, las cuales se caracterizan por 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 facilitar la toma de decisiones.
- **Actores involucrados:** Se identifica a la EEQ como la entidad encargada de proveer de manera confiable y eficaz, a los clientes no regulados como los sujetos de estudio, y el 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 períodos de medición de 5 y 15 minutos, así como variables relacionadas con la energía y la potencia.
- **Salidas esperadas:** Se determinan los productos que resultarán del análisis, como las curvas de cada cliente, las curvas de demanda máxima, archivos CSV y de texto plano, además de la agrupación de los clientes en base a la similitud de sus curvas de carga características.

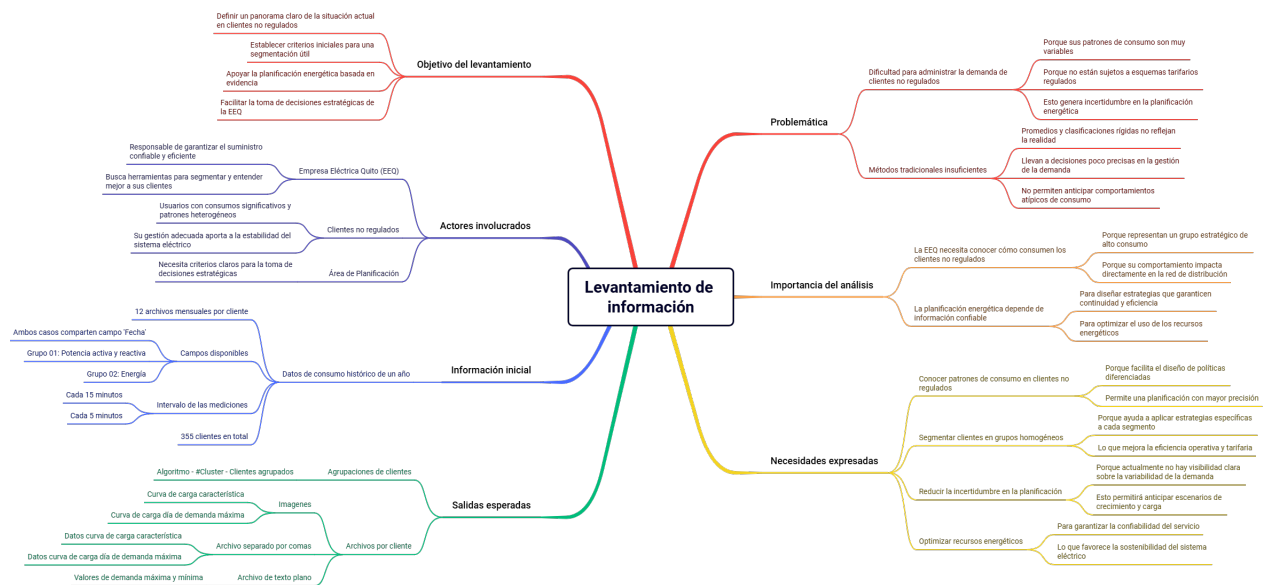


Figura 2.2: Mapa mental utilizado para levantamiento y organización de información

Determinar los objetivos del negocio

El propósito de la EEQ es determinar y agrupar a los clientes no regulados según el comportamiento energético que estos reflejen en su curva de carga características 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 objetivos específicos que surgieron son:

- Obtener una perspectiva completa de los patrones de consumo de los clientes no regulados del sector eléctrico
- Establecer un proceso de agrupación que posibilite el reconocimiento de comportamientos energéticos propios
- Promover la creación de estrategias diferenciadas para gestionar la demanda y planificar la red del sistema eléctrico.

Se considera que el proyecto tuvo éxito si las agrupaciones resultantes permiten la clara identificación de patrones energéticos característicos y si se proporciona información que pueda ser interpretable por la EEQ para la planificación estratégica.

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 den Python (con librerías como numpy, pandas, matplotlib y scikit-learn), orquestación del proceso ETL utilizando Apache Airflow sobre Docker y base de datos en nube MongoDB Atlas para guardar los datos correspondientes a las curvas de carga características
- Personal del área de Planificación de la EEQ como usuario principal de los resultados obtenidos.

■ Requerimientos y limitaciones:

- El análisis está limitado a los clientes no regulados del sector eléctrico dentro del área de concesión de la EEQ.
- Para el análisis serán considerados únicamente aquellos registros correspondientes a días laborables, todos los demás días (finde de semana y feriados) serán excluidos.
- Garantizar que los datos tengan la consistencia característica de una serie temporal
- Restringir el alcance a la creación y evaluación de los modelos de agrupamiento, sin 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 reflejan de manera precisa el comportamiento real de los clientes.

■ Riesgos detectados:

- 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...)
- Existencia de valores nulos e inconsistencias en las series temporales.
- Restricciones de capacidad computacional debido a la gran cantidad de datos.

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 un modelo de clustering que permita clasificar a los clientes no regulados en grupos homogéneos, basándose en la semejanza de sus curvas de carga características anuales, las cuales estarán previamente normalizadas, con el propósito de identificar patrones de consumo que puedan ser utilizados como insumo para la planificación estratégica.

Los objetivos específicos 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 intra-clúster promedio.)

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 y una representación visual de la curva promedio de cada clúster.

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.

El plan establecido para el proyecto se detalla a continuación:

1. Entendimiento del negocio: Recopilación de información, determinación de objetivos empresariales y técnicos.

2. Entendimiento de los datos: Análisis exploratorio de los datos históricos, validación de la calidad y consistencia de los mismos.
3. Preparación de los datos: Implementación del proceso ETL que involucra pasos como la extracción, limpieza, integración, normalización y carga de las curvas de carga características de cada cliente
4. Modelado: Aplicación y optimización de los algoritmos de clustering.
5. Evaluación: Comparación de las agrupaciones obtenidas, haciendo uso de métricas cuantificadoras de calidad de las mismas.
6. Adaptación de CRISP-DM en la fase final de despliegue: La etapa de despliegue será sustituida por un análisis comparativo entre los resultados obtenidos por cada algoritmo, utilizando visualizaciones detalladas de las curvas de carga agrupadas, con el fin de generar conocimiento de utilidad para el área de Planificación de la EEQ.

2.4.2. Entendimiento de los datos

Recolección inicial de los datos

Se llevó a cabo la recolección de datos relevantes para el análisis, los cuales corresponden a los registros históricos de consumo energético de los clientes no regulados de la EEQ a lo largo del año 2023.

Los datos fueron proporcionados directamente por un representante del área de Planificación, el cual 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, 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 estos contenían. Esta tarea permitió determinar la naturaleza de los datos y evaluar el tipo de formato que poseen. A pesar de venir en archivos separados por coma (.csv), los archivos no

necesariamente están separados por coma, algunos vienen separados por tabulaciones y otros por punto y coma, incluyen datos sobre el consumo y la temporalidad de los mismos. Se pudo notar lo siguiente:

- Aparentemente, la frecuencia de medición entre registros es de 15 minutos.
- Existen dos grupos de clientes: los que tienen únicamente mediciones de energía y los que tienen mediciones de potencia activa y reactiva.
- 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.
- Algunos archivos poseen líneas de resumen, las cuales incluyen totales ponderados agregados para intervalos específicos.
- Algunos valores dentro de los registros contienen coma como separador de miles y punto como separador decimal (ej: 1,203.239)

En la Figura 2.3 se puede apreciar lo mencionado anteriormente.

Id 786					Informe de Medidas de Puntos Frontera;				
Distribuidora EEQ S.A.					Punto Frontera: ACNOVQU03;				
Colectora Distrocuyo					Fecha;AE (kWh);AS (kWh);RE (KVarh);RS (KVarh);SE (KVah);SS (KVah);				
Marca L+G					Total D: 2023/02/28;0.000;104,224.054;0.000;34,468.932;109,799.955;86,400.000;				
Modelo RXS4					2023/03/01 00:00;0.000;1,203.239;0.000;396.692;1,267.004;900.000;				
Serie 10002684					2023/02/28 23:45;0.000;1,180.957;0.000;386.713;1,242.697;900.000;				
Medidor 90002439					2023/02/28 23:30;0.000;1,162.232;0.000;379.803;1,222.746;900.000;				
Suministro 3967					2023/02/28 23:15;0.000;1,203.980;0.000;397.491;1,267.914;900.000;				
Designación FLEXOFAMA CIA. LTDA. - (3967)					2023/02/28 23:00;0.000;1,190.779;0.000;387.437;1,252.250;900.000;				
					2023/02/28 22:45;0.000;1,217.744;0.000;415.905;1,286.987;900.000;				
DEMANDAS					2023/02/28 22:30;0.000;1,232.724;0.000;409.121;1,298.971;900.000;				
Orden	Fecha	Origen	Demanda activa DEL	Demanda reactiva DEL	2023/02/28 22:15;0.000;1,244.710;0.000;419.638;1,313.578;900.000;				
1	2023-05-07 18:15	Lectura AMR	0.03	0	2023/02/28 22:00;0.000;1,253.795;0.000;426.359;1,324.372;900.000;				
2	2023-05-07 18:30	Lectura AMR	0.0318	0	2023/02/28 21:45;0.000;1,261.706;0.000;443.740;1,337.616;900.000;				
3	2023-05-07 18:45	Lectura AMR	0.0324	0	2023/02/28 21:30;0.000;1,202.828;0.000;413.222;1,271.947;900.000;				
4	2023-05-07 19:00	Lectura AMR	0.0318	0	2023/02/28 21:15;0.000;1,186.138;0.000;386.484;1,247.552;900.000;				
5	2023-05-07 19:15	Lectura AMR	0.0336	0	2023/02/28 21:00;0.000;1,193.885;0.000;389.410;1,255.815;900.000;				

Figura 2.3: 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, 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. La Figura 2.4 presenta los resultados del análisis exploratorio de los datos.

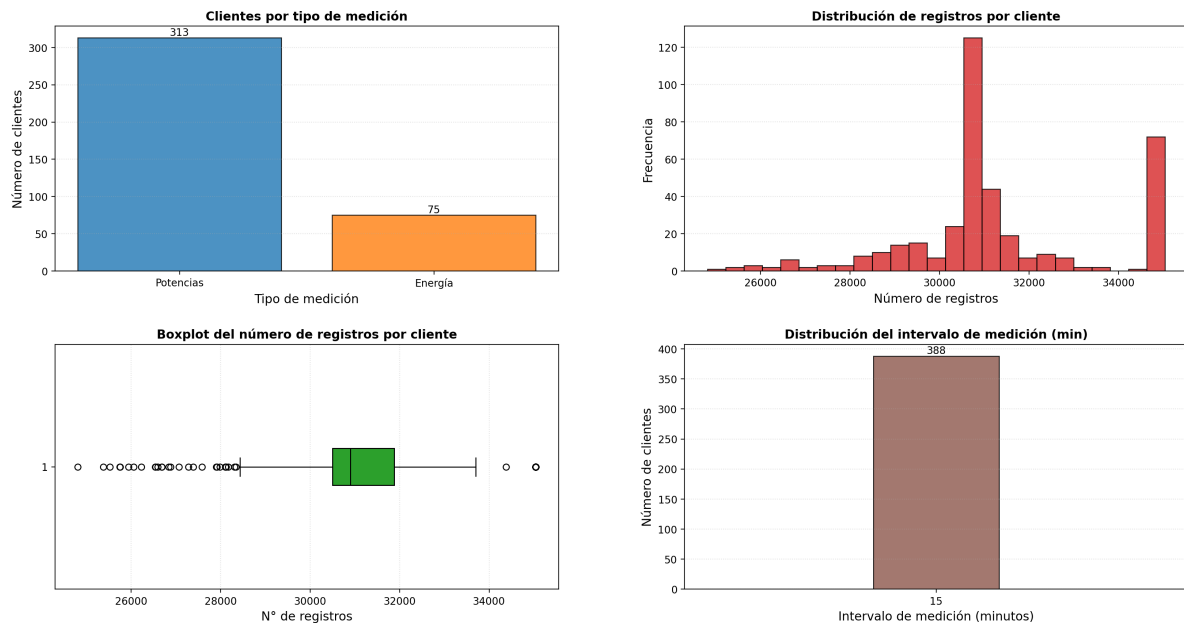


Figura 2.4: Exploración de los datos de consumo energético

La Figura 2.4 muestra cuatro gráficos:

- Gráfico de barras: Cantidad de clientes según el tipo de medición (potencias o energía).
- Histograma: Distribución del número de registros totales por cliente.
- Diagrama de caja: Variabilidad en el número de registros totales por cliente.
- Gráfico de barras: Cantidad de clientes según el intervalo de medición de sus registros.

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 35040 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 caja 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

Dado que en tareas anteriores se identificaron de manera superficial inconsistencias en los archivos de medición, tales como la variación en los formatos de fecha, uso de coma como separador de miles y la ausencia de registros, se llevó a cabo una comprobación de su calidad haciendo énfasis en los aspectos mencionados anteriormente. 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 presencia de valores nulos y ausentes.

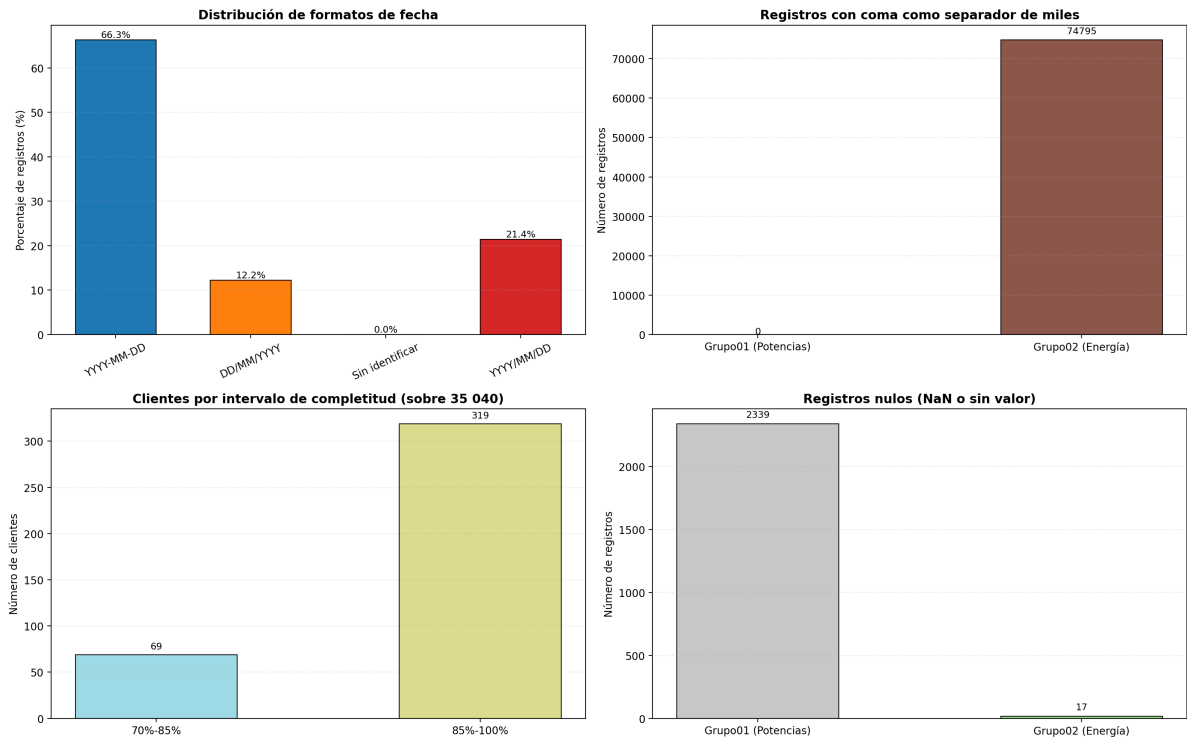


Figura 2.5: Verificación de la calidad de los datos de consumo energético

La Figura 2.5 contiene cuatro gráficos de barras que representan los resultados del análisis de la calidad de los datos:

- Distribución de registros según el formato de fecha que contienen.
- Número de registros que presentan la coma como separador de miles en cada grupo de clientes.
- Cantidad de clientes agrupados por intervalos según la completitud de sus registros [70 %, 85 %] y [85 %, 100 %] respecto al total ideal de 35040 registros.
- Número de registros nulos o sin valor identificados en cada grupo de clientes.

Como se puede apreciar en el primer gráfico (superior izquierdo), el formato más común es el YYYY-MM-DD (66,3 %), seguido por YYYY/MM/DD (21,4 %) 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 319 clientes tienen una completitud de datos mayor al 85 % del total ideal de registros anuales, mientras que 69 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 2339 registros nulos. El volumen de registros que poseen valores nulos es mínimo, 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.

2.4.3. 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 2.6, la cual es de elaboración propia. Gracias a este procedimiento se pudo orquestar de manera eficiente la extracción, transformación y carga de los datos de consumo, garantizando automatización y trazabilidad en la ejecución del flujo.

El uso de Apache Airflow permitió construir un pipeline totalmente modularizado, donde cada tarea del 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 la base de datos de MongoDB Atlas.

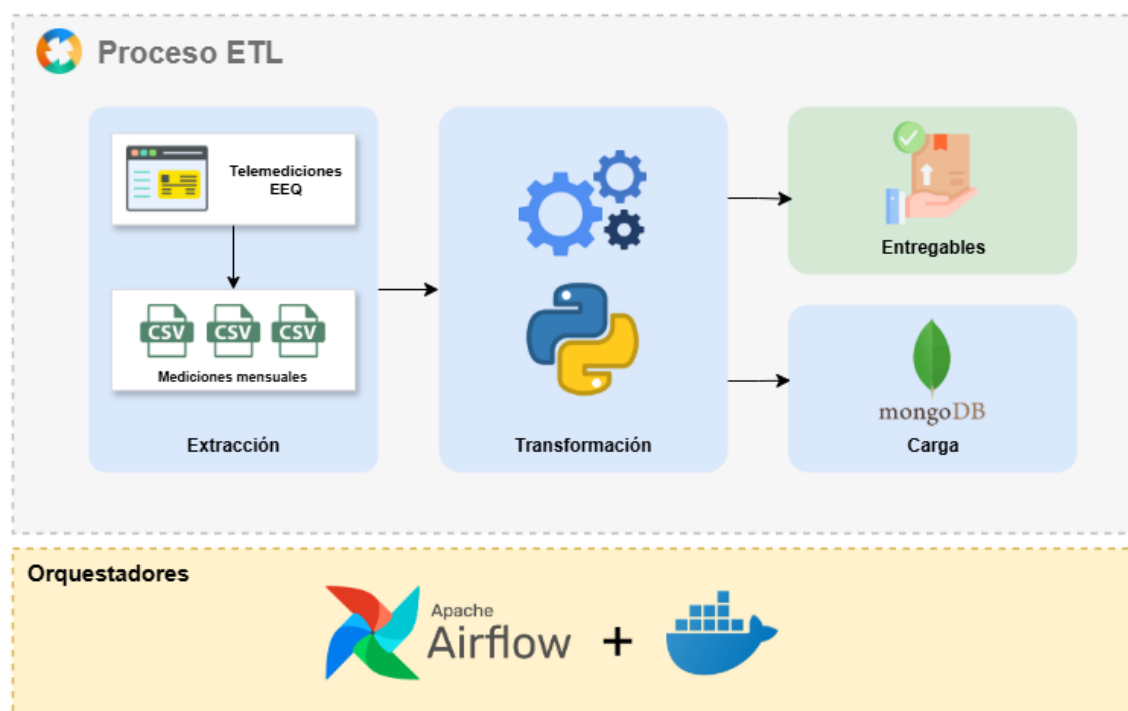


Figura 2.6: Esquematización del proceso ETL y sus etapas

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 el archivo `docker-compose.yaml`, el cual a su vez fue modificado para incluir de manera adicional tres volúmenes compartidos entre el sistema de archivos local y el entorno aislado de Airflow:

- `/data`: Contiene todos los archivos `.csv` correspondientes a las mediciones mensuales de todos los clientes no regulados.
- `/outputs`: Este volúmen fue creado con la finalidad de almacenar los entregables adicionales requeridos por la EEQ.
- `/utils`: Contiene un archivo `.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, estandarizar formato de fechas, generar las curvas de carga características, entre otros.

La Figura 2.7 ilustra la modificación realizada al archivo `docker-compose.yaml` para añadir los volúmenes compartidos mencionados anteriormente:

```
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 2.7: 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 dicho contenedor haciendo uso de la sentencia `docker exec -it IDENTIFICADOR bash`.

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 librerías como pandas, numpy, scikit-learn, pymongo, dotenv y matplotlib. La Figura 2.8 muestra todo el proceso mencionado anteriormente:

```

airflow@0c91068629f2: /opt/i  X  +  v
[ 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 2.8: Identificación de worker e instalación de librerías necesarias.

Una vez establecido el ambiente de ejecución con sus respectivas dependencias, se estructuró el proceso ETL dentro de un grafo acíclico dirigido (DAG) denominado `etl_dag_datos_consumo`. Cada nodo que se encuentra en el grafo simboliza una tarea autónoma dentro del proceso ETL, ejecutada de manera secuencial o paralela en función de las dependencias que tenga definidas.

La orquestación general de este DAG en Apache Airflow se muestra en la Figura 2.9, 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 de 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.

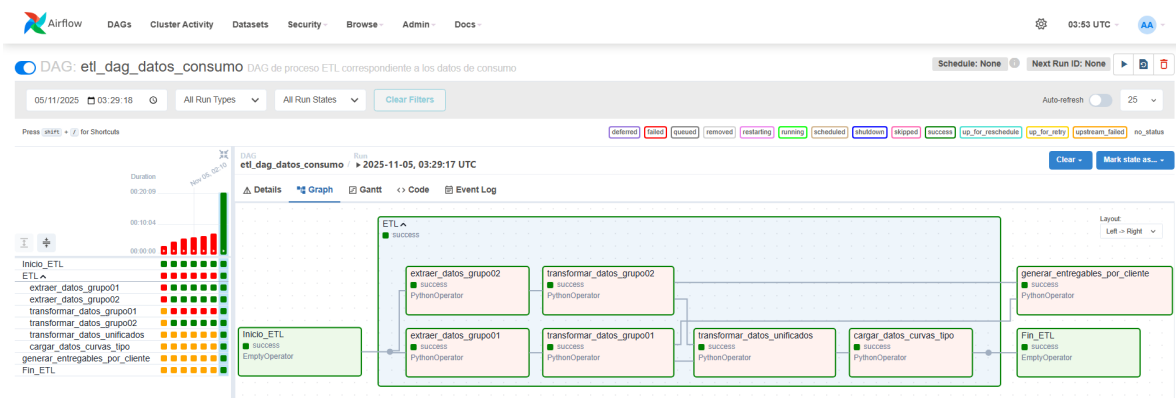


Figura 2.9: DAG de proceso ETL correspondiente a los datos de consumo

La Figura 2.10 ofrece un desglose funcional del proceso ETL implementado, con el propósito de complementar la vista general de la Figura 2.9.

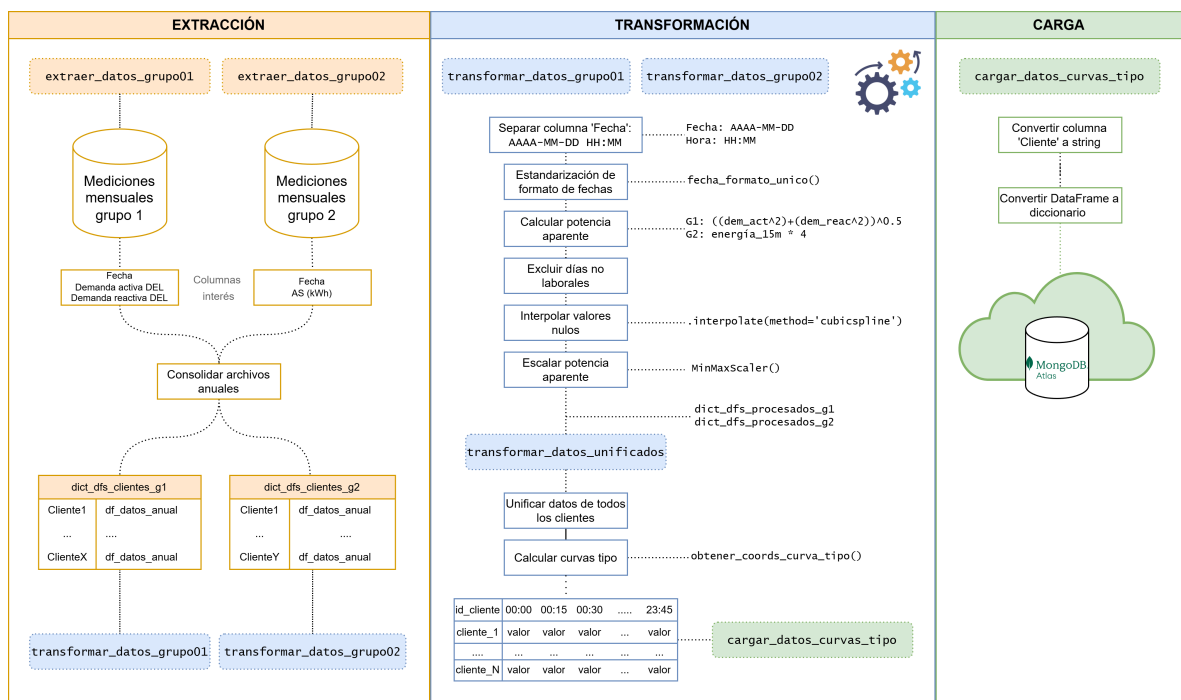


Figura 2.10: Desglose operacional del DAG de ETL: descripción de tareas.

En la Figura 2.10 esta se describen las operaciones concretas que cada tarea realiza durante las fases de extracción, transformación y carga de datos, 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 marco 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.

Seleccionar los datos

En esta tarea, se seleccionaron los conjuntos de datos que se usarían en las etapas siguientes del análisis, garantizando que coincidieran con las metas establecidas en la fase de entendimiento del negocio. Los datos se obtuvieron de los registros históricos de mediciones mensuales de los clientes no regulados de la EEQ, correspondientes a 2023.

Tomando en cuenta la estructura observada en el análisis exploratorio de los datos, se validó que el intervalo de medición entre registro es de 15 minutos, también se determinó que los archivos de medición presentan dos configuraciones distintas según el tipo de variable que registran, por lo que 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 los registros anuales consolidados de dicho cliente

Limpiar los datos

La tarea de limpieza tuvo como objetivo asegurar que los registros de medición fueran coherentes, corrigiendo errores estructurales y preparando la información durante el proceso de transformación. Esta etapa de limpieza se llevó a cabo en las tareas `transformar_datos_grupo01` y `transformar_datos_grupo02` que son parte del DAG `etl_dag_datos_consumo` y que fueron ejecutadas independientemente para cada grupo de clientes respectivamente.

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

1. Estandarización de formatos de fecha y hora: Las variaciones de formato en las fechas obstaculizaban el tratamiento uniforme de los datos como series temporales. Se implementó la función `fecha_formato_unico()` con el fin de solucionar este problema y transformar cualquier formato detectado en uno estándar (YYYY/MM/DD).
2. Tratamiento de valores nulos: Se empleó un interpolador spline cúbico para llenar los valores faltantes identificados en tareas anteriores, tomando en cuenta que estos eran pocos en comparación con el total de datos anuales por cliente. Se eligió este método debido a que posibilita la reconstrucción de los datos de manera coherente, sin perder la tendencia natural de la serie temporal.
3. Exclusión de días no laborables y feriados: Los registros correspondientes a fines de semana y feriados nacionales del año 2023 fueron eliminados. Esta exclusión se basó en que los clientes no regulados tienen patrones de consumo altamente asociados a su actividad laboral, por lo cual, los días no laborables no son útiles para el análisis.
4. Separadores de miles: En tareas anteriores se han detectado valores que poseen la coma como separador de miles y punto como separador decimal, esto se atribuye únicamente al grupo de clientes con mediciones de energía. Se eliminaron las comas y se estableció el tipo de dato como flotante para posibilitar que Python pueda procesarlos correctamente.

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 rellenar 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 estos motivos, se consideró una opción apropiada 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 manera comparable entre todos los clientes, sin distinción del tipo de medición disponible (potencias o energía).

Con este propósito, se implementaron las transformaciones correspondientes en las tareas `transformar_datos_grupo01` y `transformar_datos_grupo02` que son parte del DAG `etl_dag_datos_consumo`, 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):

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

Una vez se tiene calculada la potencia aparente para los dos grupos de clientes, 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 del 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, evi-

tando que los días con valores altos 'aplanen' al resto de registros. El procedimiento consiste en escalar cada medición de acuerdo con los valores correspondientes de su propio día, siguiendo la siguiente fórmula:

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

La normalización es esencial, debido a que posibilita que los algoritmos de clustering se enfoquen únicamente en la forma de la curva de consumo y no en su magnitud, evitando que clientes cuyas demandas son altas alteren los resultados del análisis.

La elección de escalar cada día individualmente fue el resultado de llevar a cabo un estudio comparativo entre la normalización global del año entero y la normalización diaria que se empleó. La Figura 2.11 muestra que la normalización global tiende a suavizar la curva característica debido a que hay días con valores significativamente más altos que el resto. El comportamiento mencionado anteriormente es particularmente común en clientes no regulados, cuyos patrones de consumo son muy variables debido a que están directamente relacionados a su actividad económica. Por otro lado, la normalización diaria mantiene la estructura relativa de cada día, evitando así que un día con valores mayores “aplane” la curva, dando como resultado una representación más precisa del patrón característico asociado al cliente.

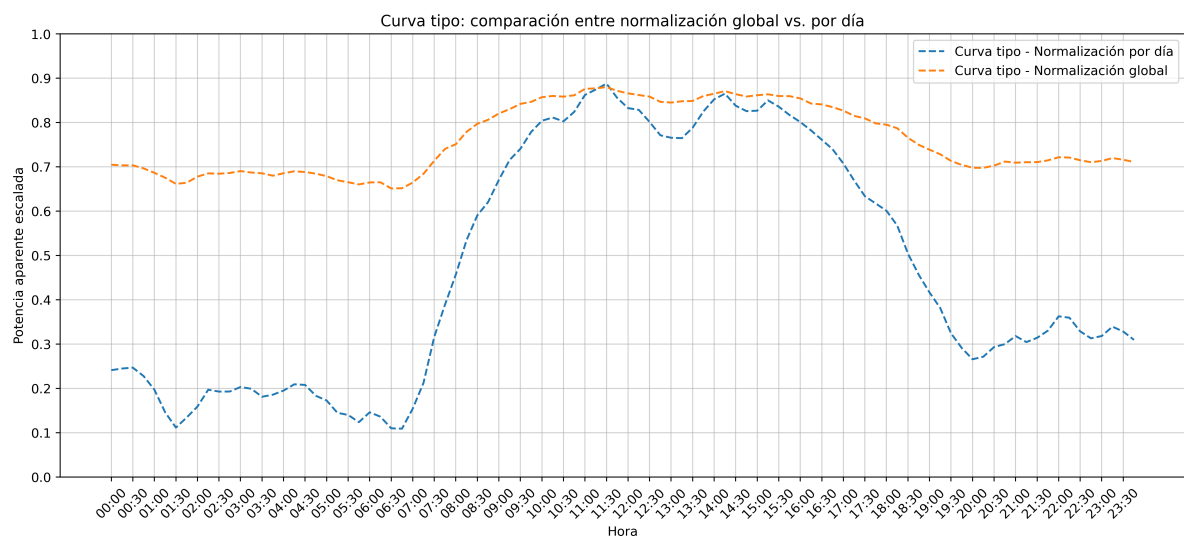


Figura 2.11: Comparación entre la curva tipo resultante de aplicar normalización diaria y normalización global.

Integrar los datos

Una vez concluido todo el proceso de transformación individual de cada grupo de clientes (selección, limpieza y construcción), se procedió a unificar los datos procesados que provenían de los dos grupos establecidos, con el fin de consolidar un único conjunto global y proceder con el cálculo de sus curvas características representativas anuales. La integración fue realizada en la tarea `transformar_datos_unificados` que es parte del DAG `etl_dag_datos_consumo`, donde se combinaron los diccionarios resultantes de los flujos de transformación individuales de cada grupo en uno solo.

Una vez consolidada toda la información, se procedió a obtener la curva tipo, con ayuda de la función `obtener_coords_curva_tipo` se realizó un agregamiento por hora, calculando la mediana de la potencia aparente escalada durante todos los días del año. Se utilizó 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, donde cada fila corresponde a un cliente y cada columna a una marca temporal (desde las 00:00 hasta las 23:45) en intervalos de cada 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 creado, con el objetivo de preaparar los datos para ser almacenados y modelados más adelante. La tarea `cargar_datos_curvas_tipo` que es parte del DAG `etl_dag_datos_consumo` fue donde se ejecutó esta operación, realizando las siguientes actividades:

1. Se comprobó que cada registro tuviera un identificador exclusivo para el cliente, convirtiendo el campo 'Cliente' a tipo cadena de caracteres, debido a que el identificador puede ser alfanumérico.
2. Se convirtió el DataFrame consolidado con los datos 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 co-

lección CurvasTipoAnuales, dando por finalizada la fase de preparación de los datos y obteniendo así un conjunto de datos totalmente limpio, estructurado, formateado y preparado para su uso en la fase de modelado.

2.4.4. Modelado

El fase de modelado es el núcleo de CRISP-DM, en la que se crean los modelos a partir del 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 que se va a utilizar y determinar como se van a valorar los resultados antes de implementar cualquier modelo. Después se desarrollan los modelos y se examinan sus resultados de acuerdo con los criterios técnicos y del negocio.

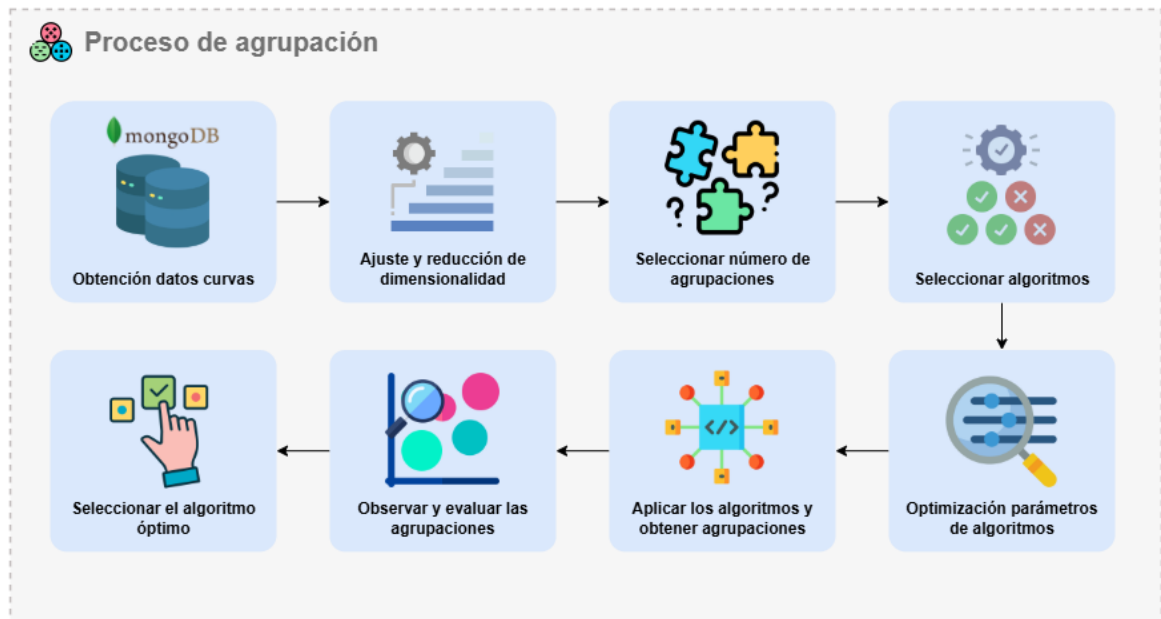


Figura 2.12: Esquematización de proceso de agrupación y sus etapas.

La Figura 2.12 representa un diagrama que sintetiza las diversas tareas que intervinieron en el proceso de agrupación utilizado en el presente trabajo. Este diagrama fue creado en base al cuaderno de trabajo utilizado para realizar clustering, se ilustra la secuencia de trabajo que se implementó: obtener las curvas desde la base de datos, ajustarlas para comparabilidad, determinar el número óptimo de agrupaciones, elegir los algoritmos y optimizar sus parámetros, aplicar los algoritmos y generar las agrupaciones para evaluarlas en base a métricas y seleccionar el algoritmo y los resultados más apropiado.

Selección 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.

El problema se enmarca en el aprendizaje no supervisado debido a que las curvas de carga son series temporales normalizadas que detallan la conducta energética de los clientes. 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].

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 Sepctral Clustering. A continuación se explica por qué hemos elegido cada uno de estos algoritmos para el presente trabajo:

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 algoritmos es especialmente eficaz en casos donde la estructura de los datos no muestran 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

Según la metodología CRISP-DM, 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 este trabajo 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 cuentan con etiquetas estándar.

El diseño de prueba tiene como objetivo analizar la capacidad de los algoritmos elegidos para agrupar a los clientes no regulados en conjuntos homogéneos, según la forma que presente su curva de carga característica anual. Con el fin de asegurar que los resultados sean comparables, se ejecutan todos los algoritmos sobre un conjunto de datos preprocesados idéntico, obtenidos como resultado de la fase de preparación de los datos.

Dado que el objetivo del análisis es encontrar similitudes en la estructura de las curvas de carga, se enfoca el diseño de la prueba en evaluar la consistencia interna y la separación entre los clústeres a través de métricas internas de validación. La Tabla 2.1 muestra la selección y justificación de las métricas utilizadas, conforme a la bibliografía especializada en la agrupación de perfiles de carga eléctrica.

Tabla 2.1: 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 tamaño. 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, si bien puede mostrar valores moderados como resultado de la superposición natural de los patrones de consumo [34].

Métrica	Descripción y propósito
Índice Davies–Bouldin (DBI)	Cuantifica la relación entre la distancia entre 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].
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 2.13 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	0.633878	0.628827	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	0.678086	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 2.13: Datos de curvas de carga obtenidas desde MongoDB Atlas.

2. Ajuste de curvas

Se llevó a cabo un ajuste adicional de las curvas de carga con la finalidad de asegurar la comparabilidad entre las curvas de carga de los clientes y evitar sesgos relacionados con desplazamientos en el eje Y. Este ajuste implicó mover cada curva de forma que su punto de partida sea ubicado aproximadamente en el origen (0,0), empleando el promedio de los primeros n registros de cada serie como referencia, lo cual permite reducir sesgos puntuales atribuido a fluctuaciones al inicio de la curva. Después, se sustrajo ese valor de cada uno de los puntos de la curva, manteniendo su forma relativa. Así, los algoritmos de agrupamiento son capaces de detectar semejanzas basándose únicamente en la forma que tiene la curva de carga.

La Figura 2.14 demuestra 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.

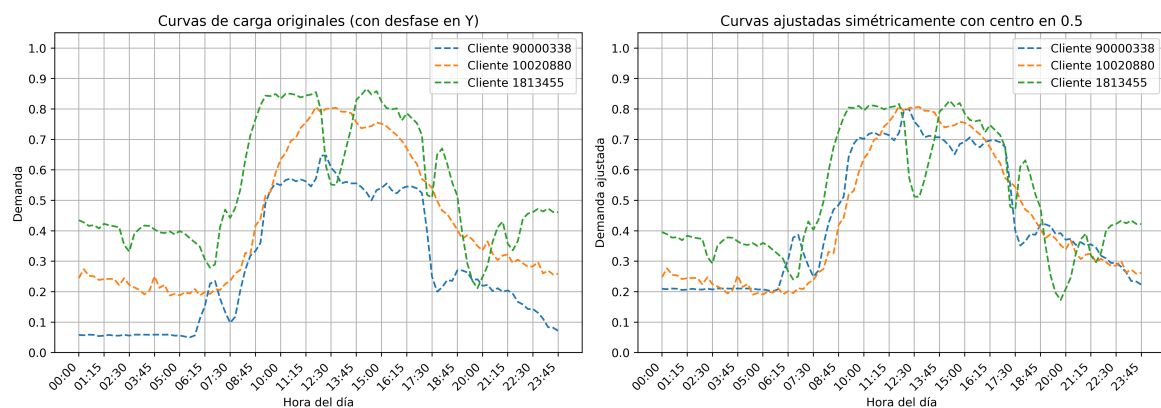


Figura 2.14: Efecto del ajuste vertical de las curvas de carga para garantizar la comparabilidad basada en la forma (tres clientes).

Además, no se usaron en este periodo 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 2.15 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.

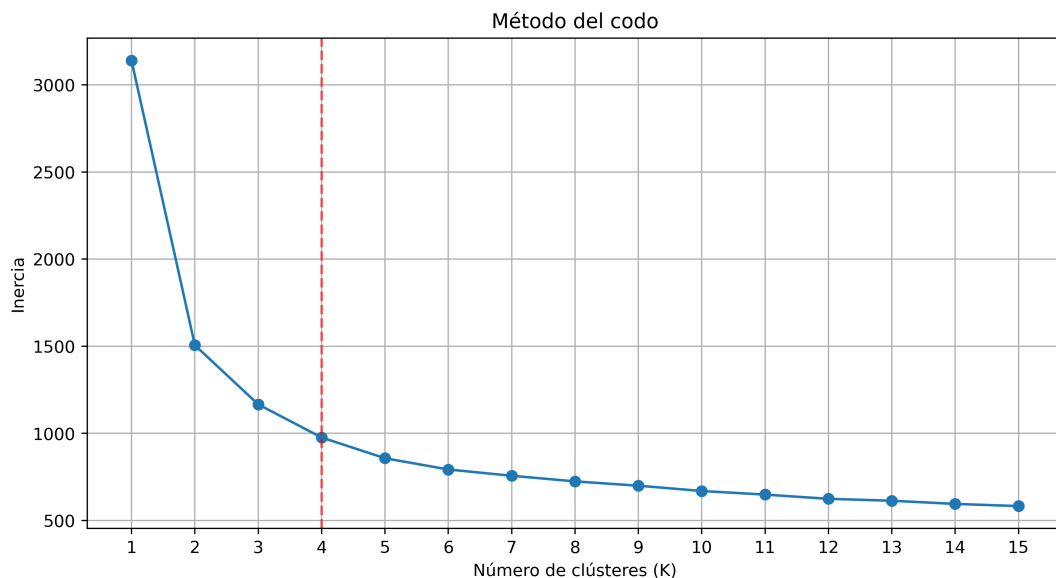


Figura 2.15: Validación del número de clústeres utilizando el método del codo sobre las curvas de carga ajustadas.

4. Hiperparametrización de los algoritmos

Se realizó un proceso de hiperparametrización para cada uno de los algoritmos de agrupamiento elegidos, después de haber establecido el número de clústeres. 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 vinculados a cada algoritmo. La Tabla 2.2 sintetiza los hiperparámetros más importantes que se tuvieron en cuenta durante el proceso de ajuste, los cuales fueron elegidos según su efecto en la calidad de las agrupaciones.

Tabla 2.2: Hiperparámetros evaluados para los algoritmos de clustering

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

La selección de los hiperparámetros óptimos se realizó utilizando métricas internas de validación, priorizando aquellas configuraciones que maximizaran 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

Se implementaron los algoritmos Birch, K-Means, Gaussian Mixture Model y Spectral Clustering en el conjunto de curvas de carga ajustadas una vez establecidos los hiperparámetros. En consecuencia, cada algoritmo le otorgó a cada cliente una etiqueta de clúster, lo que produjo cuatro grupos por técnica.

Evaluar el modelo

Con la finalidad de medir la calidad de los agrupamientos logrados y cotejar el rendimiento de los diferentes algoritmos, se llevó a cabo la evaluación de los modelos de clustering a través de métricas internas de validación.

■ Correlación intra-clúster

Se empleó la correlación intra-clúster como la métrica principal de evaluación, que se define como el promedio de la correlación entre cada curva de carga individual y la correspondiente curva media del clúster. Esta métrica posibilita la evaluación directa de la semejanza en la forma de las curvas agrupadas,

lo que resulta particularmente apropiado para series temporales normalizadas.

La correlación es una medida sólida para analizar la coherencia interna de los clústeres a partir de perfiles de carga eléctrica, porque se enfoca en las similitudes estructurales de las curvas y no en su magnitud absoluta [34], [37].

■ Métricas complementarias de validación

En un enfoque complementario, se usaron el índice de Calinski-Harabasz (CHI), el de Davies-Bouldin (DBI) y el Silhouette Score, tres métricas que se usan con frecuencia para medir la separación entre los clústeres y su compacidad interna [34], [38].

La Tabla 2.3 muestra una comparación numérica del rendimiento de cada algoritmo de agrupamiento, teniendo en cuenta las métricas complementarias y la correlación dentro del clúster.

Tabla 2.3: 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

Las características intrínsecas de las curvas de carga estudiadas en esta investigación explican los valores bajos que se han obtenido para el Silhouette Score, ya que estas presentan una variabilidad elevada y densidades heterogéneas entre clústeres. Según [40], este comportamiento es habitual en conjuntos de datos reales sobre demanda eléctrica, en los que dicho índice pierde sensibilidad frente a agrupaciones densas y patrones con una dispersión interna alta. En este escenario, los hallazgos conseguidos enfatizan la importancia de analizar el Silhouette Score de manera complementaria y dar prioridad a las métricas que se enfocan en la semejanza estructural de las curvas, como la correlación intra-clúster.

2.4.5. Evaluación

La fase de Evaluación dentro del ciclo que propone la metodología CRISP-DM tiene como propósito el de validar en integral los resultados obtenidos en la fase de

modelado, verificando que los agrupamientos alcanzados validan los objetivos previamente definidos y proporcionan información útil de cara al análisis del consumo energético de la clientela no regulada.

Esta fase tiene una especial importancia en el contexto de este trabajo, dado que aquí lo que se procura no es únicamente obtener agrupaciones matemáticamente válidas, sino ser capaz de llegar a identificar distintos, homogéneos e interpretables patrones de comportamiento energético que pudieran servir de sustrato para poder plantear procesos de planificación energética, de análisis técnico e incluso de toma de decisiones en empresas distribuidoras de energía.

Evaluar resultados

La evaluación de los resultados es la tarea clave de esta fase y del proyecto en su conjunto, ya que permite establecer en qué medida los modelos de clustering contruidos permiten conseguir el objetivo principal del análisis; el cual no es otro que segmentar los clientes no regulados en función de la forma de su curva característica anual de consumo energético.

En contraste con los problemas de aprendizaje supervisado, donde son las etiquetas de referencia las que permiten evaluar de forma directa el rendimiento, en el caso del clustering la evaluación tiene que ser abordada de forma multidimensional. En este trabajo se fundamenta en hasta tres ejes complementarios:

- Diferenciabilidad de las curvas obtenidas
- Homogeneidad interna y separación entre clústeres
- Interpretabilidad energética de los patrones identificados

De acuerdo a los tres ejes anteriormente definidos, el análisis de resultados es planteado a partir de una evaluación global de patrones que han sido identificados, abordando aspectos como diferenciabilidad y coherencia interna de los clústeres obtenidos. En este sentido, el análisis se realiza a través de los aspectos siguientes:

- Diferenciabilidad de curvas en base a su agrupación por clúster y algoritmo
Los resultados de la agrupación obtenidos a partir de la ejecución de los algoritmos K-Medoids, Gaussian Mixture Models, Birch y Spectral Clustering, dieron lugar a agrupaciones que mantenían consistencia en todos los casos, obteniendo curvas tipo promedios perfectamente diferenciables por cada clúster.

Este aspecto es relevante, ya que significa que los algoritmos han sido capaces de captar similitudes reales entre las curvas de carga de los clientes, no simplemente diferencias por consumo absoluto. Este comportamiento guarda relación directa con las decisiones tomadas en las fases de preparación y modelado de los datos, donde se normalizaron y alinearon las curvas de modo que se eliminaran sesgos por escalas y desplazamientos verticales.

Al contemplar simultáneamente las curvas tipo promedio que han generado las diferentes técnicas, se puede observar que, con independencia de la técnica empleada, los patrones de consumo identifican bien y este comportamiento se manifiesta en perfiles que pueden mostrar un comportamiento horaria estable durante el día, perfiles que muestran un comportamiento con picos muy bien definidos en un intervalo horario específico y perfiles con alta variabilidad a lo largo del día.

Es cierto que existen pequeñas variaciones en las formas exactas de las curvas obtenidas por las diferentes técnicas, pero que todas ellas acaban coincidiendo en identificar un comportamiento energético diferenciado refuerza la idea de que el abordaje es robusto y que, en consecuencia, es correcta la elección de evaluar diferentes técnicas de clustering de manera comparativa.

■ Evaluación de homogeneidad y separación mediante PCA

Con la finalidad de enriquecer el análisis de los clústeres desde una perspectiva geométrica y visual, se llevó a cabo una transformación de la dimensionalidad usando un Análisis de Componentes Principales y proyectando las curvas de las características de los clientes en un espacio bidimensional.

La representación de esta forma permite estudiar simultáneamente la separación entre los clústeres y la concentración de los clientes dentro del clúster, constituyendo una genuina evaluación de la estructura de los clústeres obtenidos.

Los resultados indican que, en los cuatro algoritmos implicados en el proceso, los clústeres muestran una distribución en la representación espacial muy bien definida, donde se observa una separación clara entre los grupos y una buena compactación de los puntos dentro de cada clúster.

Esta estructura confirma que los clientes que han quedado agrupados muestran similitudes en la forma de sus curvas de carga, y que los grupos resultantes no presentan solapamientos evidentes. Este comportamiento pone en evidencia la pertinencia de la correlación intra-clúster promedio como métrica de evaluación principal, puesto que se trata de una métrica que permite codificar de una forma más directa la semejanza entre cada curva en concreto y la curva tipo representativa de su clúster, cumpliendo exactamente con la intención del análisis.

A partir de los resultados de la tarea de evaluación de los modelos, ejecutada en la Fase 4 (Modelado) del presente trabajo, en la cual 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.

- Análisis detallado de los clústeres obtenidos mediante Spectral Clustering

A partir de la selección del algoritmo Spectral Clustering, se procedió a analizar en detalle las curvas tipo correspondientes a cada clúster generado por Spectral Clustering, considerando no únicamente las curvas promedio, sino todas las curvas individuales de los clientes que conforman cada grupo. Este análisis permite evaluar la consistencia interna de los clústeres y extraer conclusiones sobre los comportamientos energéticos característicos de cada segmento.

Clúster 1: Agrupa clientes con un perfil de consumo estable a lo largo del día, con variaciones suaves y sin picos pronunciados. Este comportamiento es característico de cargas con operación continua o procesos industriales que mantienen una demanda relativamente constante durante la jornada.

Clúster 2: Corresponde a clientes que presentan incrementos significativos de consumo en franjas horarias específicas, principalmente asociadas a horarios laborales, lo que sugiere procesos productivos concentrados en determinados intervalos del día.

Clúster 3: Incluye perfiles con alta variabilidad intradiaria, evidenciando picos y valles pronunciados en su curva de carga. Este patrón puede asociarse a procesos intermitentes, ciclos de producción no uniformes o a una combinación de diferentes tipos de carga dentro del mismo cliente.

Clúster 4: Agrupa clientes con consumos reducidos durante gran parte del día y aumentos puntuales de demanda, lo que podría relacionarse con operaciones bajo demanda, cargas específicas de corta duración o actividades con baja continuidad temporal.

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.

Revisar el proceso

En esta tarea se realizó una revisión integral del proceso seguido a lo largo del proyecto de minería de datos, 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 con los objetivos planteados, y que las decisiones adoptadas en cada etapa se encuentran debidamente justificadas. En particular, se constató que la selección de las curvas tipo como representación del consumo anual, el uso de técnicas de normalización y alineación, 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 de la presente tesis se centra en el análisis y evaluación de los modelos de clustering, las actividades asociadas a la fase de despliegue operativo no se desarrollan en este trabajo. Por tanto, la implementación del modelo seleccionado en sistemas productivos o plataformas de gestión energética no aplica dentro del alcance actual.

No obstante, se identifican como líneas futuras 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.

2.4.6. Despliegue

Planificar el despliegue

Planificar monitoreo y mantenimiento

Producir el reporte final

Revisar el proyecto

3. Resultados, Conclusiones y Recomendaciones

3.1. Resultados

Ejemplo de tabla:

No. Prueba	Resultado	Tiempo [s]
1	10	0.9
2	5	0.5

Tabla 3.1: Resultados de las pruebas realizadas

3.2. Conclusiones

3.3. Recomendaciones

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. Obtenido de: <https://arxiv.org/pdf/1702.01414>
- [4] P.-N. Tan, M. Steinbach y V. Kumar, *Introduction to Data Mining*. Pearson Education Limited, 2014. Obtenido de: https://www.ceom.ou.edu/media/docs/upload/Pang-Ning_Tan_Michael_Steinbach_Vipin_Kumar_-_Introduction_to_Data_Mining-Pe_NRDk4fi.pdf
- [5] J. Han, M. Kamber y J. Pei, *DATA MINING Concepts and Techniques*. Morgan Kaufmann, 2012. Obtenido de: <https://myweb.sabanciuniv.edu/rdehkharghani/files/2016/02/The-Morgan-Kaufmann-Series-in-Data-Management-Systems-Jiawei-Han-Micheline-Kamber-Jian-Pei-Data-Mining.-Concepts-and-Techniques-3rd-Edition-Morgan-Kaufmann-2011.pdf>
- [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 Obtenido de: <https://peerj.com/articles/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 Obtenido de: https://www.researchgate.net/publication/386874870_ETL_Process_Automation_Tools_and_Techniques

- [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 Obtenido de: <https://www.sciencedirect.com/science/article/pii/S131915781100019X>
- [10] W. H. Inmon, *Building the Data Warehouse*, 3rd. John Wiley & Sons, 2002, ISBN: 0-471-08130-2. Obtenido de: https://www.r-5.org/files/books/computers/databases/warehouses/W_H_Inmon-Building_the_Data_Warehouse-EN.pdf
- [11] V. Gour, S. S. Sarangdevot, G. S. Tanwar y A. Sharma, "Improve Performance of Extract, Transform and Load (ETL) in Data Warehouse," en *International Journal on Computer Science and Engineering (IJCSE)*, vol. 2, IJCSE, 2010, págs. 786-789. Obtenido de: https://www.researchgate.net/profile/Dr-Anand-Sharma-2/publication/49618608_Improve_Performance_of_Extract_Transform_and_Load_ETL_in_Data_Warehouse/links/0046351b96aeb43640000000Improve-Performance-of-Extract-Transform-and-Load-ETL-in-Data-Warehouse.pdf
- [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 Obtenido de: <https://www.sciencedirect.com/science/article/pii/S026427512200364X>
- [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. Obtenido de: <https://arxiv.org/abs/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 Obtenido de: <https://peerj.com/articles/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 Obtenido de: https://cdn2.hubspot.net/hubfs/55819/2018-PES-Portland-OR-USA-BIRCH_Clustering_SUBMIT%5B1%5D.pdf?t=1522697637650
- [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. Obtenido de: <https://arxiv.org/abs/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. Omid, "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, Accessed: 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 Obtenido de: <https://iopscience.iop.org/article/10.1088/1742-6596/1889/3/032009/pdf>
- [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 Obtenido de: <https://amt.copernicus.org/articles/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. Miraftebzadeh, 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, pág. 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

Ejemplo IEEE:

- **[1]** L. Carvajal, *Metodología de la Investigación Científica*. Santiago de Cali: U.S.C., 2006.

<https://ietresearch.onlinelibrary.wiley.com/doi/full/10.1049/ccs2.12080>

https://hpi.de/fileadmin/user_upload/fachgebiete/naumann/publications/2008/ETL_Management

https://www.researchgate.net/profile/Sameer_Shukla3/publication/369899578_DevelopingPragmatic-Pipelines-using-Apache-Airflow-on-Google-Cloud-Platform.pdf?origin=journalDetail&utm_source=journalDetail&utm_medium=article&utm_campaign=journalDetail

<https://www.controlrecursosyenergia.gob.ec/wp-content/uploads/downloads/2021/03/Folleto-Resumen-Estad>

5. Anexos

Anexo I. Conjunto de Datos Extensos

Anexo II. Formato de Entrevista

Anexo III. Enlaces