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 supieron 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 Jhonny Sánchez, mi hermano de otra madre, con quien he compartido invaluables 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 Vírgen 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.

Agradezco de igual manera al ingeniero Boris Astudillo por compartir conmigo sus valiosos consejos durante mi vida universitaria y, por su constante guía y apoyo durante todo el proceso de 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

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 distribuidoras de energía, como la Empresa Eléctrica Quito (EEQ), la eficiente gestión energética es uno de los principales desafíos a enfrentar. Diversos factores como la diversificación en los hábitos de consumo y variablidad 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.

El análisis del consumo energético es un aspecto fundamental para la optimización de recursos en sectores como la distribución eléctrica y la gestión de tarifas, debido a esto, identificar patrones de consumo permite segmentar a los clientes en función de su comportamiento energético, lo cual facilita la toma de decisiones estratégicas, garantizando así una planificación energética eficiente.

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 para la segmentación de clientes 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

consolidar los datos de consumo mensual de cada cliente en una curva representativa anual. Este proceso asegura la correcta integración, transformación y normalización de los datos para que las curvas representativas sean comparables entre sí.

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 agrupamientomásrelevantes, 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 (Extracción, Transformación y Carga), 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), estos fueron seleccionados en función de su relevancia en la literatura y su aplicabilidad en el análisis de análisis de curvas de carga. Para determinar el número óptimo de agrupaciones se hara 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

1.4.1.1. 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].

1.4.1.2. 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].

1.4.1.3. Importancia de segmentar a los 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. Metodología CRISP-DM

CRISP-DM, cuyas siglas corresponden a Cross-Industry Standard Process for Data Mining, 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 [4].

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 [4]. CRISP-DM contiene un total de seis fases, tal y como se describe en [5]:

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.

1.4.3. Minería de datos

Según [6], 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 [7].

Entre sus principales funciones se destacan [7]:

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.

1.4.4. Proceso ETL

El proceso ETL (Extracción, Transformación y Carga), 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:

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].

1.4.5. 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].

1.4.5.1. 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].

1.4.5.2. Número de agrupaciones

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

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

1.4.5.3. Metodo 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].

1.4.5.4. 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:

1. **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:

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

2. **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:

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

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

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

3. **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:

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

4. **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:

$$CF = (N, LS, SS) \tag{1.5}$$

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

$$\min_{T} \; g(W_k(T), B_k(T)) \tag{1.6}$$

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

1.4.5.5. Hiperparametrización de algoritmos

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

1.4.6. 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].

1.4.7. 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:

Cuadro 1.1: Herramientas utilizadas para el desarrollo del componente

Herramienta	**Descripción de la herramienta**
Airflow	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	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	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	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	MongoDB es una base de datos no relacional basada en documentos, ofrece una gran escalabilidad y flexibilidad, además de un modelo avanzado de consultas e indexación [24].

2. Metodología

Describir el diseño o el planteamiento utilizado...

2.0.1. Flujo de trabajo propuesto

- **Extracción, Transformación y Carga de los Datos (ETL):** La primera fase consiste en la extracción, transformación y carga (ETL) de los datos de consumo energético. Los datos iniciales provienen de archivos de consumo mensual por cliente. En este paso, se construye un archivo anual para cada cliente, en el cual se agregan y consolidan los datos correspondientes a cada año. Además, los datos son escalados y normalizados para garantizar su consistencia y comparabilidad. Este proceso se lleva a cabo con la ayuda de **Apache Airflow**, el cual permite automatizar el flujo de trabajo y garantizar su ejecución eficiente. Finalmente, los datos transformados son cargados en **MongoDB**, asegurando su disponibilidad para las fases siguientes del análisis.

- **Segmentación de Clientes:** En esta fase, se procede a definir el número óptimo de grupos de clientes a través de un análisis conversacional con las partes interesadas y la aplicación de métodos como el **método del codo**. Una vez definido el número de grupos, se seleccionan y optimizan los algoritmos de agrupación más adecuados para el análisis, tales como **KMeans**, **GaussianMixture**, **Birch** y **Spectral Clustering**. Estos algoritmos se utilizan para agrupar a los clientes según la similitud de sus curvas de consumo energético anual. Los resultados obtenidos se presentan visualmente, permitiendo observar las agrupaciones y patrones emergentes en el consumo de energía de los clientes.

- **Evaluación Comparativa de los Algoritmos:** Finalmente, se realiza una evaluación comparativa de los algoritmos de agrupación aplicados, utilizando diversas métricas para medir la calidad de las agrupaciones. Entre las métricas utilizadas se encuentran el **Silhouette Score**, **SSE (Suma de Errores al Cuadrado)**, **DBI (Índice de la Diferencia de Davies-Bouldin)** y **CHI (Índice de Calinski-Harabasz)**. Estas métricas permiten analizar el rendimiento de los algoritmos y seleccionar el que mejor se adapte a los datos de consumo energético de los clientes.

1. **Proceso ETL**
   El primer proceso consiste en el desarrollo de un flujo ETL bajo el marco de trabajo de Apache Airflow, este proceso nos permitirá estructurar y preparar los datos de consumo de los clientes en curvas anuales características para

su posterior análisis de agrupación. La Figura 1 ilustra de manera detallada todas las etapas que abarca este proceso.

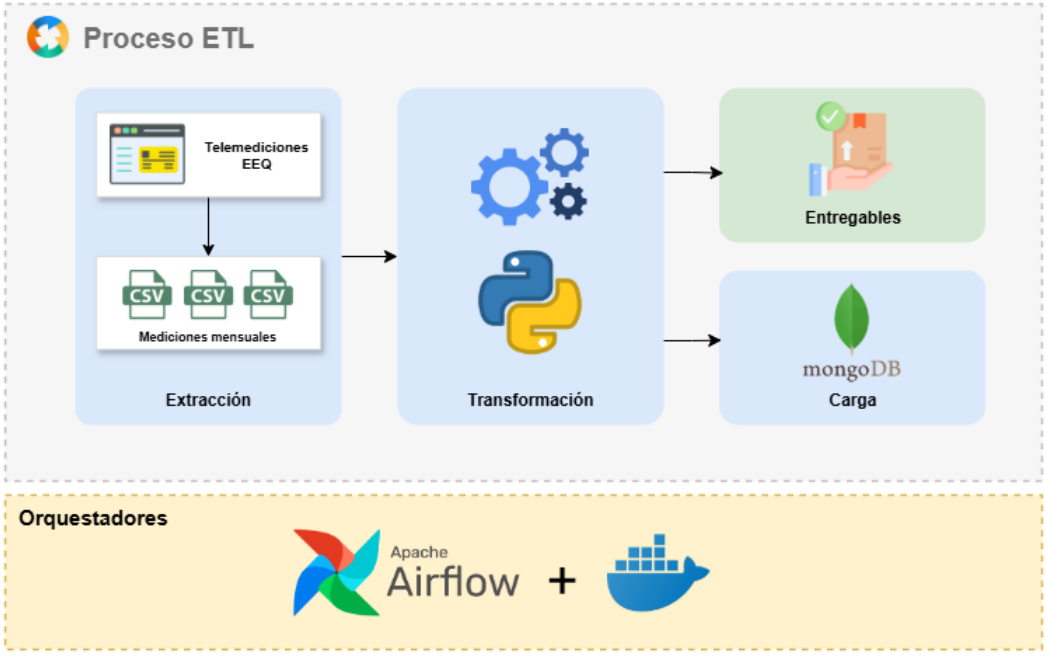

Figura 1: Proceso ETL con sus etapas

2. **Proceso de agrupación**
   El segundo proceso comprende todo el proceso de agrupamiento, en este apartado se elegirá el número de agrupaciones deseadas, se seleccionarán y aplicarán diversos algoritmos de agrupamiento para finalmente evaluar la calidad de las agrupaciones obtenidas por cada algoritmo. La Figura 2 ilustra de manera detallada todas las etapas que abarca este proceso.

Los datos utilizados para el presente componente comprenden todas las mediciones mensuales del año 2024 por cada cliente, las cuales han sido obtenidos de la página de telemediciones de la Empresa Eléctrica de Quito.

Por otro lado, la segmentación de clientes se realiza exclusivamente en función de la forma de su curva característica anual, obtenida al final del proceso ETL descrito en la Figura 1. No se consideran otros factores, como las tarifas o la geolocalización, ya que el objetivo de la parte interesada es agrupar a los clientes estrictamente según el patrón de consumo de energía reflejado en su curva característica anual.

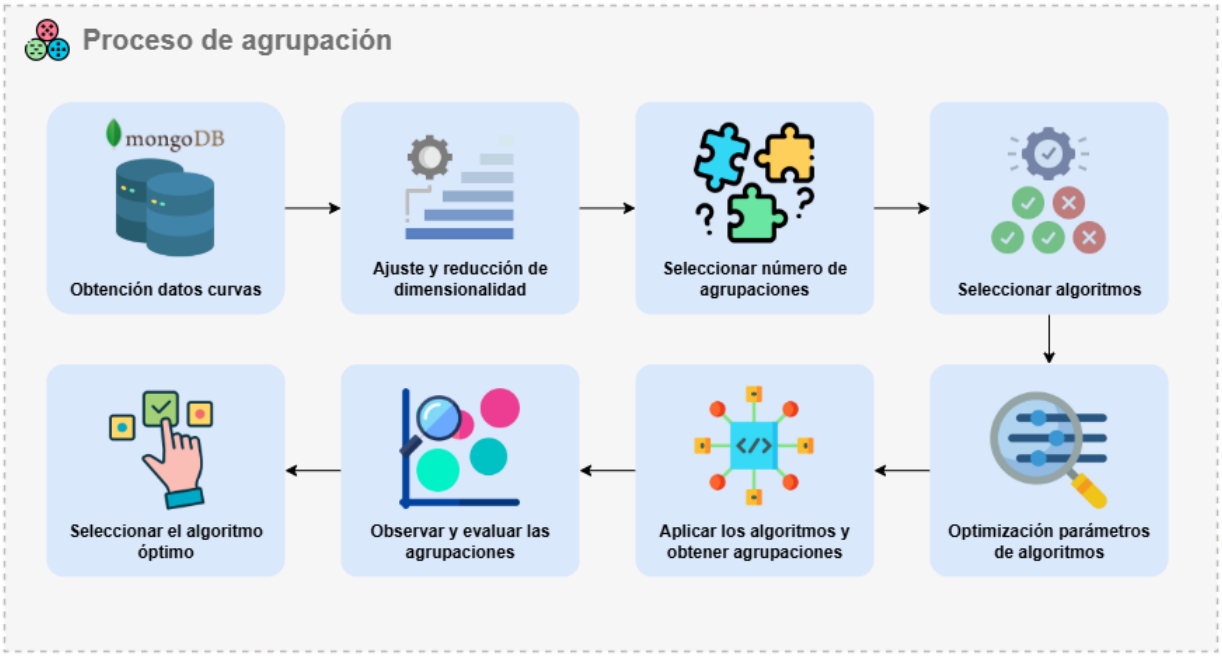

Figura 2: Proceso de agrupación con sus etapas

3. Resultados, Conclusiones y Recomendaciones

3.1. Resultados

Ejemplo de tabla:

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

Cuadro 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] 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

[5] 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

[6] 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

[7] 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

[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/0046351b96aeb4364000000Improve-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. 103 925, 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

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$_upload/fachgebiete/naumann/publications/2008/ETL_Managemen$

https://www.researchgate.net/profile/Sameer$_Shukla3/publication/369899578_Developing_Pragma$
$Pragmatic-Data-Pipelines-using-Apache-Airflow-on-Google-Cloud-Platform.pdf?origin=journalDetail_tp=eyJwYWdlIjoiam91cm5hbERldGFpbCJ9$

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