UNIVERSIDAD NACIONAL DE SAN ANTONIO ABAD DEL CUSCO

FACULTAD DE INGENIERÍA ELÉCTRICA, ELECTRÓNICA, INFORMÁTICA Y MECÁNICA

ESCUELA PROFESIONAL DE INGENIERÍA INFORMÁTICA Y DE SISTEMAS

TESIS

APLICACIÓN DE TÉCNICAS DE BIG DATA E INTELIGENCIA ARTIFICIAL PARA MEJORAR LA CAPACIDAD ANALÍTICA DE EGEMSA

PRESENTADO POR:
Br. GROVER MOREANO BRICEÑO
Br. SAUL WALDEMAR TICONA BEJAR

PARA OPTAR AL TÍTULO PROFESIONAL DE INGENIERO INFORMÁTICO Y DE SISTEMAS

ASESOR:
Mgt. JOSÉ MAURO PILLCO QUISPE

CUSCO – PERÚ
2026

Universidad Nacional de San Antonio Abad del Cusco

INFORME DE SIMILITUD

(Aprobado por Resolución Nro.CU-321-2025-UNSAAC)

El que suscribe, el **Asesor** ..JOSÉ MAURO PILLCO QUISPE..... quien aplica el software de detección de similitud al trabajo de investigación/tesistitulada: .....

APLICACIÓN DE TÉCNICAS DE BIG DATA E INTELIGENCIA ARTIFICIAL PARA MEJORAR LA CAPACIDAD ANALÍTICA DE EGEMSA

Presentado por: ..GROVER MORENO BRICEÑO..... DNI N° ..43612546.....;
presentado por: ..SAUL WALDEMAR TICONA BEJAR..... DNI N°: ..47405354.....
Para optar el título Profesional/Grado Académico de .....
INGENIERO INFORMÁTICO Y DE SISTEMAS

Informo que el trabajo de investigación ha sido sometido a revisión por ..2..... veces, mediante el Software de Similitud, conforme al Art. 6° del ***Reglamento para Uso del Sistema Detección de Similitud en la UNSAAC*** y de la evaluación de originalidad se tiene un porcentaje de .....5.....%.

Evaluación y acciones del reporte de coincidencia para trabajos de investigación conducentes a grado académico o título profesional, tesis

Porcentaje	Evaluación y Acciones	Marque con una (X)
Del 1 al 10%	No sobrepasa el porcentaje aceptado de similitud.	X
Del 11 al 30 %	Devolver al usuario para las subsanaciones.	
Mayor a 31%	El responsable de la revisión del documento emite un informe al inmediato jerárquico, conforme al reglamento, quien a su vez eleva el informe al Vicerrectorado de Investigación para que tome las acciones correspondientes; Sin perjuicio de las sanciones administrativas que correspondan de acuerdo a Ley.	

Por tanto, en mi condición de Asesor, firmo el presente informe en señal de conformidad y **adjunto** las primeras páginas del reporte del Sistema de Detección de Similitud.

Cusco, 15 de enero de 20.26

.....
Firma
Post firma: José Mauro Pillco Quispe
Nro. de DNI: 23861067
ORCID del Asesor: 0000-0002-0527-089X

Se adjunta:

1. Reporte generado por el Sistema Antiplagio.
2. Enlace del Reporte Generado por el Sistema de Detección de Similitud: **oid:** 27259:54641010 2

Grover Saul Moreano Ticona

tesis_Egemsa.pdf

Universidad Nacional San Antonio Abad del Cusco

Detalles del documento

Identificador de la entrega
trn:oid:::27259:546410102

Fecha de entrega
15 ene 2026, 9:39 a.m. GMT-5

Fecha de descarga
15 ene 2026, 9:54 a.m. GMT-5

Nombre del archivo
tesis_Egemsa.pdf

Tamaño del archivo
7.1 MB

131 páginas

21.383 palabras

120.026 caracteres

5% Similitud general

El total combinado de todas las coincidencias, incluidas las fuentes superpuestas, para ca...

Filtrado desde el informe

- Bibliografía
- Texto citado
- Texto mencionado
- Coincidencias menores (menos de 18 palabras)

Exclusiones

- N.º de coincidencias excluidas

Fuentes principales

3%	Fuentes de Internet
1%	Publicaciones
3%	Trabajos entregados (trabajos del estudiante)

Marcas de integridad

N.º de alertas de integridad para revisión

No se han detectado manipulaciones de texto sospechosas.

Los algoritmos de nuestro sistema analizan un documento en profundidad para buscar inconsistencias que permitirían distinguirlo de una entrega normal. Si advertimos algo extraño, lo marcamos como una alerta para que pueda revisarlo.

Una marca de alerta no es necesariamente un indicador de problemas. Sin embargo, recomendamos que preste atención y la revise.

Dedicatoria

A mi familia, por su amor incondicional y por enseñarme que los sueños se construyen con esfuerzo y perseverancia. A quienes creyeron en mí incluso en los momentos en que yo dudé. Este logro es tanto suyo como mío.

Agradecimiento

Agradezco profundamente el apoyo de mi familia, profesores y amigos, quienes han hecho posible la realización de este trabajo.

Resumen

En el presente trabajo de investigación se aplica un conjunto de técnicas de Big Data: Data Lake, Web Scraping, ETL (Extracción, Transformación y Carga de datos) y de Inteligencia Artificial (particularmente Machine Learning) con el propósito de mejorar la capacidad analítica de la Gerencia Comercial de EGEMSA. El problema principal radica en la ineficiencia de la recolección de datos del mercado eléctrico peruano, la deficiente consolidación y organización de estos datos y la limitada aplicación de herramientas avanzadas que permitan identificar patrones. Para abordar la problemática, se propone una arquitectura basada en Data Lake con capas Bronce, Plata y Oro, a fin de optimizar el flujo de la información su organización y posterior análisis. Además, se emplean algoritmos de clústeres (segmentación) para categorizar a los clientes de EGEMSA y facilitar la toma de decisiones basadas en datos confiables y oportunos. Los resultados evidencian una reducción significativa en los tiempos de procesamiento y mayor precisión en la identificación de oportunidades comerciales. Finalmente, se plantean recomendaciones para la automatización y el despliegue en producción de los modelos de clustering, asegurando la sostenibilidad de la solución en el largo plazo.

Palabras clave: Big data, Inteligencia artificial, Web scraping, ETL.

Abstract

This research project applies a set of Big Data techniques: Data Lake, Web Scraping, ETL (Extract, Transform, Load), and Artificial Intelligence (particularly Machine Learning) with the aim of improving the analytical capacity of EGEMSA's Commercial Management. The main problem lies in the inefficiency of data collection in the Peruvian electricity market, the poor consolidation and organization of this data, and the limited application of advanced tools that allow patterns to be identified. To address this issue, a Data Lake-based architecture with Bronze, Silver, and Gold layers is proposed in order to optimize the flow of information, its organization, and subsequent analysis. In addition, clustering (segmentation) algorithms are used to categorize EGEMSA's customers and facilitate decision-making based on reliable and timely data. The results show a significant reduction in processing times and greater accuracy in identifying business opportunities. Finally, recommendations are made for the automation and deployment of clustering models in production, ensuring the long-term sustainability of the solution.

Keywords: Big data, Artificial intelligence, Web scraping, ETL.

Lista de abreviaturas

- **EGEMSA:** Empresa de Generación Eléctrica Machupicchu S.A.
- **COES:** Comité de Operación Económica del Sistema.
- **SEIN:** Sistema Eléctrico Interconectado Nacional.
- **OSINERGMIN:** Organismo Supervisor de la Inversión en Energía y Minería.
- **MME:** Mercado Mayorista de Electricidad.
- **VTEA:** Valorizaciones de Transferencias de Energía Activa.
- **CIIU:** Clasificación Industrial Internacional Uniforme.
- **TI:** Tecnologías de Información.
- **IA:** Inteligencia Artificial.
- **ETL:** Extracción, Transformación y Carga de Datos.
- **ML:** Machine Learning (aprendizaje automático).
- **K-Means:** Algoritmo de agrupamiento basado en la definición de k centroides.
- **Data Lake:** Repositorio masivo de datos en bruto, capaz de almacenar datos estructurados, semiestructurados y no estructurados.

- **SQL:** Lenguaje de consulta estructurada (Structured Query Language).
- **OLTP :** Procesamiento de transacciones en línea (OnLine Transaction Processing).
- **PCA:** Análisis de componentes principales (Principal component analysis).
- **CV:** Costo Variable.
- **API:** Interfaz de programación de aplicaciones (Application Programming Interface).
- **RUC:** Registro Único de Contribuyente.
- **SUNAT:** Superintendencia Nacional de Aduanas y de Administración Tributaria.
- **INEI:** Instituto Nacional de Estadística e Informática.

Índice general

Índice de tablas

Índice de figuras

Capítulo 1

Introducción

1.1. Generalidades

La Empresa de Generación Eléctrica Machupicchu S.A. (EGEMSA), inmersa en el dinámico mercado eléctrico peruano, afronta crecientes retos en la gestión y análisis de grandes volúmenes de datos proporcionados por el Comité de Operación Económica del Sistema (COES). Estos datos, generados a partir de transacciones de energía (inyecciones, retiros y valorizaciones eléctricas), se caracterizan por su gran volumen, variedad y velocidad de generación, lo cual dificulta su recolección, organización y análisis.

La adopción de técnicas de Big Data e Inteligencia Artificial (IA) ofrece la oportunidad de extraer información valiosa y tomar decisiones más acertadas en un entorno de alta competitividad. Concretamente, se propone optimizar el manejo y procesamiento de datos mediante la construcción de un Data Lake, la aplicación de Web Scraping para la recolección automatizada de información del COES y la implementación de un pipeline ETL (Extracción, Transformación y Carga de Datos) para transformar y limpiar datos. Seguidamente, se emplean algoritmos de Machine Learning (K-Means) para identificar patrones y segmentar clientes.

La presente tesis demuestra cómo estas innovaciones tecnológicas, aplicadas de forma sistemática, es decir seguir un plan o secuencia de pasos para resolver el problema en EGEMSA, pueden mejorar su capacidad analítica y, en consecuencia, fortalecer su posición comercial dentro del Sistema Eléctrico Interconectado Nacional (SEIN).

1.2. Justificación

1.2.1. Conveniencia

La implementación de técnicas de Big Data y de IA responde a la creciente necesidad en la Gerencia Comercial de EGEMSA de gestionar grandes volúmenes de datos y de aprovechar el análisis de dichos datos para respaldar su posicionamiento en el mercado eléctrico.

1.2.2. Relevancia

El trabajo es relevante porque propone un enfoque replicable para empresas del sector eléctrico y organizaciones con grandes volúmenes de datos. La propuesta integra una metodología completa que abarca captura automatizada, depuración por capas (BroncePlataOro) y segmentación de clientes mediante *clustering*, validada por un experto del área comercial. De este modo, contribuye al fortalecimiento de la gestión comercial basada en analítica de datos en el sector energético.

1.2.3. Implicancias Prácticas

- Simplificación de la recolección de datos provenientes del COES.

- Organización centralizada de la información en un Data Lake.
- Segmentación y análisis de clientes con técnicas de *clustering*, lo que agiliza la toma de decisiones.

1.2.4. Valor teórico

La tesis profundiza en la aplicación práctica del paradigma Data Lake (arquitectura de medallones: Bronce, Plata, Oro) (Microsoft, 2024) y del algoritmo K-means para la solución de problemas reales de negocio, enriqueciendo la literatura y ofreciendo un referente metodológico.

1.2.5. Utilidad Metodológica

Se presenta un proceso paso a paso (recolección, limpieza, modelado) que puede ser replicado o adaptado (Ver anexos A, B y C) en otras empresas o instituciones que manejen grandes volúmenes de información y requieran extraer conocimiento.

1.2.6. Justificación del enfoque Big Data

La investigación se enmarca en Big Data porque los datos del COES superan las capacidades de las herramientas tradicionales debido a su volumen, variedad y velocidad. El reto principal no es solo el almacenamiento, sino el procesamiento y análisis oportuno de la información. Por ello, Big Data se adopta como una necesidad técnica alineada con el enfoque de las *3V*.

1.2.7. Justificación de la arquitectura Data Lake

Se justifica el uso de un Data Lake porque el problema en EGEMSA no era solo la recolección de datos, sino su consolidación y disponibilidad confiable. El Data Lake permite almacenar datos en bruto, integrar múltiples fuentes y mantener el histórico de información. Asimismo, su organización por capas (Bronce, Plata y Oro) mejora la calidad, trazabilidad y preparación de los datos para analítica avanzada y modelos de *Machine Learning*.

1.3. Planteamiento y formulación del Problema

1.3.1. Descripción del problema

EGEMSA participa en el mercado eléctrico peruano, el cual se caracteriza por una creciente competitividad y una gran producción de datos provenientes del COES. Actualmente, la Gerencia Comercial de EGEMSA enfrenta ineficiencias en la recolección y organización de dichos datos, que se agravan al no contar con herramientas de análisis que permitan identificar patrones y oportunidades de negocio de manera ágil.

1.3.2. Identificación del problema

Recolección manual y fragmentada (dividida en múltiples archivos) de datos publicados por el COES, lo que conlleva demoras y posibles errores. Limitada consolidación de información: los datos se almacenan en múltiples archivos Excel o bases locales sin un repositorio centralizado (Data Lake). Falta de herramientas de IA que faciliten la búsqueda de patrones de consumo de energía de los clientes.

1.3.3. Formulación del Problema

1.3.3.1. Problema General

¿Cómo las técnicas de Big Data (Data Lake) y de Inteligencia Artificial (particularmente Machine Learning) mejoran la capacidad analítica en la Gerencia Comercial de EGEMSA?

1.3.3.2. Problema Específicos

- Ineficiencias en la recolección de datos que limitan la disponibilidad de información actualizada sobre el consumo de energía en el mercado eléctrico.
- Dificultades en la organización y consolidación de datos, afectando la calidad y velocidad del análisis en la Gerencia Comercial de EGEMSA.
- Falta uso de herramientas de Machine Learning para identificar patrones, que limitan la capacidad analítica en la Gerencia Comercial de EGEMSA.

1.4. Alcances y limitaciones

1.4.1. Alcances

La presente investigación se circunscribe a EGEMSA y al uso de la información pública proporcionada por el COES, en el marco del SEIN. En este contexto, el estudio abarca los siguientes alcances principales:

- **Ámbito organizacional:** El trabajo se enfoca en EGEMSA, particularmente en la

problemática vinculada a la gestión, procesamiento y aprovechamiento de grandes volúmenes de datos para apoyar la toma de decisiones en el ámbito comercial.

- **Ámbito de los datos:** Se consideran los datos asociados a las transacciones de energía (inyecciones, retiros y valorizaciones eléctricas) provenientes del COES. Estos datos son utilizados como insumo principal para la construcción de la solución propuesta.

- **Ámbito tecnológico:** La investigación abarca el diseño e implementación de una arquitectura basada en:

  - Un *Data Lake* para el almacenamiento centralizado y escalable de la información.
  - Técnicas de *Web Scraping* para la recolección automatizada de datos del COES.
  - Un pipeline ETL para la depuración, estandarización y organización de los datos.
  - Algoritmos de *Machine Learning*, específicamente *K-Means*, para la identificación de patrones y la segmentación de clientes.

- **Ámbito analítico:** El estudio se orienta a demostrar que la aplicación sistemática de técnicas de Big Data e Inteligencia Artificial permite mejorar la capacidad analítica de EGEMSA y brindar insumos para fortalecer su posición comercial en el SEIN, a través de una segmentación más informada de sus clientes y el descubrimiento de patrones de comportamiento.

- **Ámbito metodológico:** La investigación se centra en el diseño, desarrollo y validación de la solución propuesta, destacando su aplicabilidad y pertinencia para EGEMSA. No se busca desarrollar nuevos algoritmos de IA, sino adaptar y aplicar metodologías existentes al contexto específico de la empresa y del mercado eléctrico peruano.

1.4.2. Limitaciones

La investigación presenta las siguientes limitaciones, que acotan el alcance de los resultados obtenidos:

- **Dependencia de la calidad y disponibilidad de datos:** La exactitud y utilidad de los análisis dependen de la calidad, completitud y consistencia de los datos proporcionados por el COES. Cualquier error, retraso o cambio en el formato de publicación puede afectar el proceso de recolección y el desempeño del pipeline ETL y de los modelos de Machine Learning.

- **Restricción al contexto de EGEMSA y del SEIN:** Los resultados y conclusiones están estrechamente vinculados a la realidad operativa y comercial de EGEMSA y al entorno regulatorio y de mercado del SEIN. Por ello, la generalización de la propuesta a otras empresas o sistemas eléctricos debe realizarse con cautela y, de ser necesario, con adaptaciones adicionales.

- **Enfoque en un algoritmo de clustering específico:** La investigación se centra en el uso del algoritmo K-Means para la segmentación de clientes. Si bien se trata de una técnica ampliamente utilizada, existen otros algoritmos de clustering y enfoques de IA que podrían generar resultados diferentes o complementarios, pero que no son abordados en el presente estudio.

- **Limitaciones en el alcance del análisis predictivo:** El componente analítico se focaliza en la identificación de patrones y la segmentación de clientes, más que en la elaboración de modelos predictivos avanzados (por ejemplo, pronósticos de demanda o ingresos). Dichos modelos podrían constituir líneas de investigación futuras, pero no forman parte del alcance de esta tesis.

- **Restricciones de recursos tecnológicos y de tiempo:** La implementación del Data Lake, el Web Scraping, el pipeline ETL y los modelos de Machine Learning se desarrolla considerando las restricciones de infraestructura tecnológica y tiempo disponibles durante la investigación. Por tal motivo, ciertas optimizaciones de rendimiento, escalabilidad o automatización total de procesos quedan fuera del alcance del presente trabajo.

1.5. Objetivos

1.5.1. Objetivo General

Mejorar la capacidad analítica en la Gerencia Comercial de EGEMSA utilizando técnicas de Big Data e Inteligencia Artificial.

1.5.2. Objetivos Específicos

- Optimizar los procesos de recolección de datos mediante el uso de técnicas de Big Data como Web Scraping, ETL y Data Lake.
- Mejorar la consolidación y organización de datos para incrementar la calidad y rapidez del análisis de datos de consumo de energía en la Gerencia Comercial de EGEMSA.
- Integrar herramientas de Machine Learning, específicamente el algoritmo K-Means, para identificar patrones de consumo de energía que fortalezcan la capacidad analítica en la Gerencia Comercial de EGEMSA.

1.6. Antecedentes

1.6.1. Antecedentes Internacionales

(Guiraldes Deck, 2020) El estudio analiza la Respuesta en Demanda (DR) en el Sistema Eléctrico Nacional (SEN) con el fin de segmentar a los clientes libres y evaluar su flexibilidad de consumo ante variaciones de precios. Utilizando datos horarios de inyecciones y retiros de energía (septiembre 2018–agosto 2019), se clasificó a los clientes según sus perfiles de consumo agrupados por estación y tipo de día y su sector económico, identificado mediante registros del Servicio de Impuestos Internos. La segmentación se realizó con técnicas de *machine learning* (*t-SNE* y *DBSCAN*) y la flexibilidad de cada grupo se caracterizó a partir de entrevistas, aportando una base para modelar mercados eléctricos con DR.

(Ramírez, 2022) La investigación analiza el consumo eléctrico mensual de los clientes regulados en Chile entre 2015 y 2021 para identificar patrones y predecir su categoría, empleando *K-Means* para la agrupación, *K-NN* para la clasificación y *PCA* para determinar las variables más relevantes. Se encontró que el tipo de cliente, el año y el mes son los factores más influyentes; más del 96 % corresponde a clientes residenciales, responsables del 50 % del consumo y de la estacionalidad mensual. Los resultados sirven como base para ajustar políticas sobre tarifas, límites de consumo invernal y eficiencia energética, recomendándose ampliar el estudio hacia la predicción del consumo eléctrico.

(Figueroa Gallardo, 2021) El trabajo desarrolla dos aplicaciones en Python que integran y visualizan datos públicos del Sistema Eléctrico Chileno, actualmente dispersos entre la Comisión Nacional de Energía (CNE) y el Coordinador Eléctrico Nacional (CEN), para ofrecerlos en una única herramienta de análisis. La primera, basada en datos de la CNE, representa la red de transmisión como nodos e incluye costos marginales y demanda proyec-

tada; la segunda, que emplea *web scraping* sobre datos del CEN, genera gráficos a partir de archivos RIO, costos marginales, mediciones y generación real. El proyecto prioriza el uso de librerías eficientes y la optimización de algoritmos, asegurando que todo el contenido visual se genere previamente y que el código esté disponible para su libre uso y modificación.

1.6.2. Antecedentes Nacionales

(Chino Espinoza, 2019) La investigación se centra en aplicar Minería de Datos Temporal (TDM) para analizar las variables de proceso de generación y distribución eléctrica registradas por el sistema SCADA de la empresa EGEMSA, el cual genera grandes volúmenes de datos que normalmente se almacenan por solo tres meses. El trabajo consistió en extraer los datos de forma segura, crear una base de datos histórica, y aplicar técnicas de TDM, expresiones regulares y *clustering* para identificar patrones y características en series temporales provenientes de sensores y medidores de campo. Siguiendo la metodología *CRISP-DM*, los resultados obtenidos se entregaron a la empresa para apoyar la toma de decisiones, la gestión de eventualidades y la planificación de mantenimientos preventivos.

(Curo Martínez, 2022) La investigación desarrolló un modelo predictivo basado en redes neuronales artificiales para pronosticar la demanda eléctrica diaria del SEIN utilizando datos históricos de 2015 a 2021 del COES, con el objetivo de reducir la demanda coincidente de un usuario libre industrial. El modelo alcanzó un MAPE (Error Porcentual Absoluto Medio) promedio de 0,993 %, permitiendo aplicar estrategias que disminuyen los costos de energía. Con un enfoque cuantitativo, aplicado y explicativo, y siguiendo el método hipotético-deductivo, se concluye que un pronóstico confiable de la demanda diaria es una herramienta efectiva para optimizar el consumo y reducir gastos en el mercado eléctrico.

(García Fernández, 2021) La tesis presenta un modelo computacional para el pronóstico a corto plazo de la demanda eléctrica peruana en el SEIN, comparando el desempeño de dos

metodologías de *machine learning*: Teoría de Resonancia Adaptativa (ARTMAP Fuzzy) y el modelo Neuro-Fuzzy (ANFIS). Incluye una propuesta de preprocesamiento de datos históricos para mejorar la precisión, evaluando escenarios con datos de 2019 y 2020 y midiendo el desempeño mediante el MAPE.

Capítulo 2

Marco teórico

2.1. Bases teóricas

2.1.1. El mercado eléctrico peruano

El mercado eléctrico peruano opera bajo la ley de Concesiones Eléctricas Decreto ley Nº 25844 (gob.pe, 2023) y su reglamento Decreto Supremo Nº 009-93-EM (gob.pe, 2023), que rigen la generación, transmisión, distribución y comercialización de electricidad en el Perú. EGEMSA, una empresa ubicada en Cusco, sigue estas leyes nacionales además de cumplir con regulaciones locales pertinentes. Este marco legal establece las normas para la operación eficiente y el suministro de energía eléctrica en el país (Sociedad Nacional de Mineria Petroleo y Energía, 2024).

De otro lado, se considera un crecimiento previsto de la máxima demanda en el sector eléctrico de 2,2 % promedio anual, lo que equivale a un incremento acumulado de 680 MW, el cual incluye la demanda de proyectos, así como un crecimiento promedio anual del PBI de 2,1 % (Banco Central de Reserva del Perú, 2023).

Dada la importancia del mercado eléctrico para el desarrollo económico y, dado que presenta particularidades que podrían tornarlo complejo, es importante repasar y comprender la organización de dicho mercado: ¿cómo se estructura?, ¿cómo se clasifican los clientes a los que se atiende?, ¿cuáles son los esquemas de comercialización que operan?. En el siguiente gráfico se muestra el esquema de la organización del mercado eléctrico (Palma, 2022).

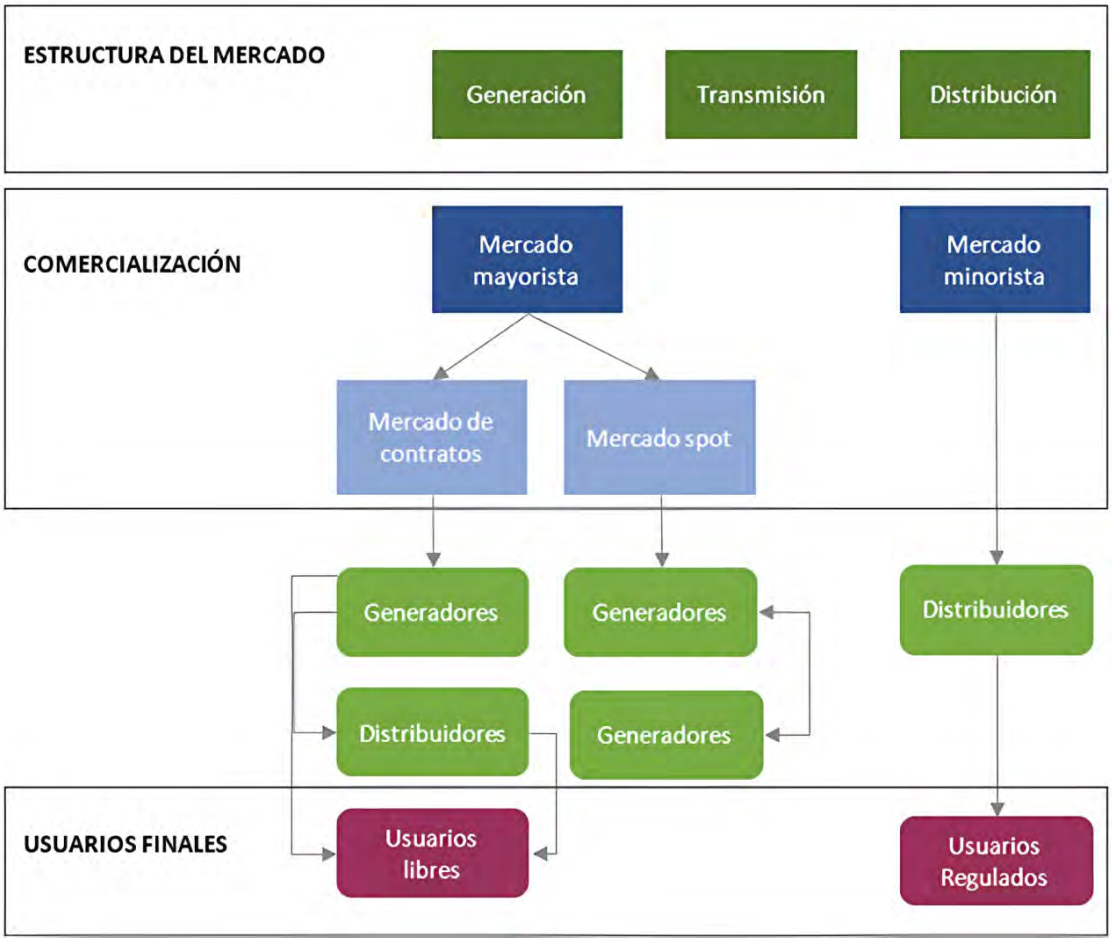

Figura 2.1: Organización del mercado eléctrico peruano

Fuente: Elaborado por (Palma, 2022)

La estructura del mercado eléctrico peruano comprende tres actividades principales: generación, transmisión y distribución. La generación eléctrica, que es la primera de las actividades del mercado eléctrico, consiste en transformar algún tipo de energía (térmica, mecánica, lumínica, entre otros) en energía eléctrica a través de la operación de centrales de genera-

ción eléctrica; la transmisión eléctrica es la actividad que transporta la electricidad desde las centrales eléctricas hacia los centros de consumo o puntos de distribución; y, por último, la distribución eléctrica tiene como función llevar el suministro de energía desde el sistema de transmisión hacia cada uno de los usuarios finales del servicio eléctrico. Estas tres actividades generación, transmisión y distribución son esenciales para el funcionamiento del sistema eléctrico (Dammert Lira, 2011).

Luego, de acuerdo a la Ley para asegurar el desarrollo eficiente de la Generación Eléctrica, Ley Nº 28832, la demanda de electricidad en el Perú tiene dos tipos de consumidores finales: los Usuarios Libres y los Usuarios Regulados. Por un lado, los Usuarios Libres obtienen esta categoría por su mayor consumo, y no están sujetos a la regulación de precios por la energía o potencia que consumen, además pueden pactar libremente los términos de sus contratos de compra de electricidad. Por otro lado, los Usuarios Regulados son los de menor consumo y se encuentran sujetos a las tarifas fijadas por el regulador OSINERGMIN, sin posibilidad de negociar los términos contractuales con quien les suministre la energía (Palma, 2022).

Tabla 2.1: *Tipos de consumidor final de electricidad según demanda máxima*

Tipo de usuario	**Demanda máxima**
Usuarios Regulados	Menor a 200 kW
Usuarios Libres	Mayor a 2 500 kW
Grandes Usuarios	Mayor a 10 MW (usuarios libres)

Fuente: Reglamento de ley de concesiones eléctricas. Decreto supremo Nº009-93-EM

Los Usuarios Libres y los Usuarios Regulados, junto con otros actores, forman parte de dos esquemas de comercialización de la electricidad: el mercado minorista y el mercado mayorista. Por un lado, la comercialización en el mercado minorista o regulado ocurre entre los Usuarios Regulados del servicio y cada operador que realiza la actividad de distribución eléctrica. Por

otro lado, la comercialización en el mercado mayorista se refiere a la que se realiza a través de transacciones bilaterales (contratos) entre los participantes de este mercado, que incluye a generadores, distribuidores y usuarios libres (denominado mercado de largo plazo o mercado de contratos); además, se refiere a la que se realiza a través de transacciones bilaterales de compra de energía y potencia sobre la base de los Costos Marginales (CMg) (denominado mercado de corto plazo o mercado spot).

Dado que la comercialización del mercado minorista se realiza dentro de la actividad de distribución, vamos a profundizar en el funcionamiento del mercado mayorista que, como se ha mencionado, está compuesto por el mercado de contratos y el mercado spot.

En el mercado de contratos, se suscriben acuerdos bilaterales de suministro entre los generadores y distribuidores, o entre generadores; asimismo, participan de estos contratos los Usuarios Libres con capacidad de negociación (por el alto nivel de consumo) que contratan directamente con el generador o distribuidor que les brinde las mejores condiciones, dichas transacciones se reúnen en el mercado libre donde compiten los generadores entre sí y, también, con los distribuidores por brindar el servicio a un usuario libre.

En el mercado spot, se realizan las transferencias de potencia y energía, determinadas por el COES, en las que pueden participar los generadores y distribuidores para atender a sus Usuarios Libres y los Grandes Usuarios Libres, con las condiciones establecidas en el reglamento del mercado mayorista de electricidad (Palma, 2022).

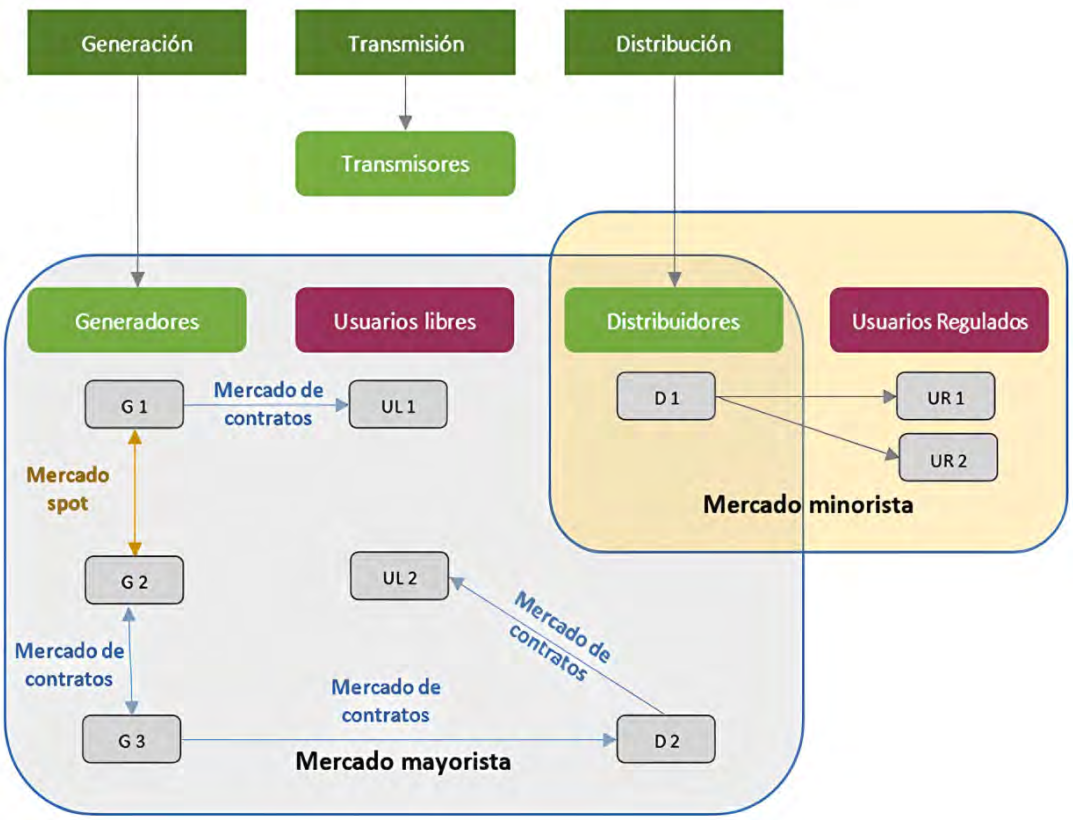

Figura 2.2: Ejemplo de la organización del mercado eléctrico

Fuente: Elaborado por (Palma, 2022)

Cabe mencionar que, tanto el mercado minorista como el mayorista, tienen una organización que corresponde al de un sistema *pool*, es decir, toda la producción de energía de los generadores es inyectada a una misma piscina (bolsa) de la cual se atienden todos los requerimientos de suministro eléctrico, tanto de los distribuidores como usuarios libres y grandes usuarios. Por lo anterior, los generadores no conocen el destino de la energía producida, ni los clientes conocen el origen de la misma. En este sentido, el despacho de energía es independiente e indistinto a los contratos financieros que puedan existir entre las partes; no obstante, esto no implica que estos no se cumplan.

Ahora bien, este sistema *pool* surge debido a que la electricidad no se puede almacenar, sino que se debe producir cuando existe demanda, razón por la que es necesario contar con

un ente encargado de la organización del mercado para la comercialización de la energía. Para el caso del Perú este papel lo desempeña el COES, quien se encarga de coordinar el despacho de energía producida por las centrales en orden de mérito, de acuerdo con sus costos variables (CV) de producción, hasta que se logre cubrir la demanda.

Para entender de mejor manera lo anterior, debemos saber que anualmente cada generador declara cuál es el CV de cada una de sus plantas de generación, los cuales son auditables; luego, el COES estima la demanda a ser abastecida, la cual fluctúa constantemente, para finalmente ordenar el despacho físico (llamar a producción) a las centrales de generación en base al CV declarado, empezando por aquellas con menor CV hasta lograr cubrir la demanda. Es decir, la decisión de producción de un generador depende de la instrucción del COES, quien luego de haberla recibido, inyecta la energía producida a un *pool* de energía para ser suministrado a las empresas distribuidoras o los clientes libres que la demanden (Palma, 2022).

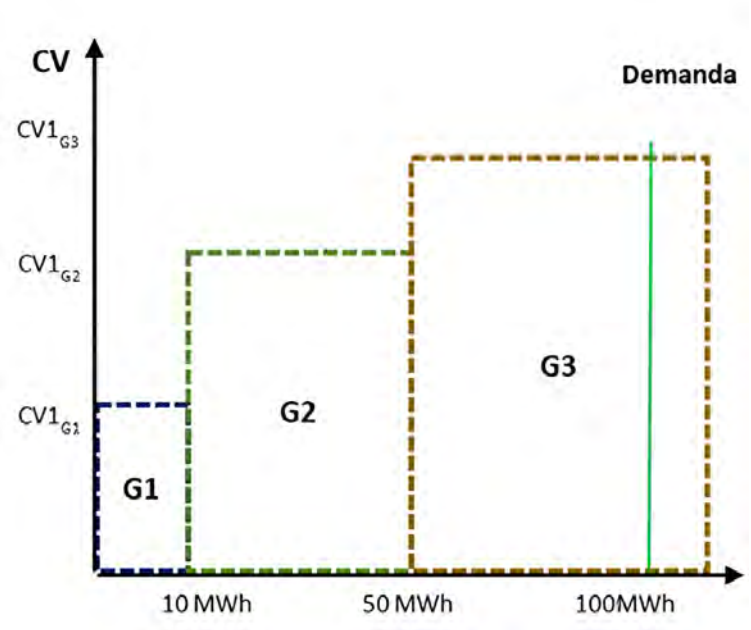

Figura 2.3: Ejemplo del despacho físico del mercado eléctrico

Fuente: Elaborado por (Palma, 2022)

Por ejemplo, asumiendo que la demanda de energía es de 100 MWh, el COES llamará a producir a los generadores disponibles ordenando el despacho primero del generador 1 (G1) ya que tiene el menor CV de todos. Sin embargo, dado que su capacidad de producción (10 MWh) no es suficiente para cubrir la demanda, llamará a producir al generador 2 (G2) que tiene una capacidad de producción de 40 MWh y finalmente al generador 3 (G3) que tiene una capacidad de producción superior a 50 MWh. De este último solo requerirá la diferencia de capacidad para completar la demanda.

El precio del mercado *spot* en el periodo de tiempo para el que el COES determinó la demanda será el que corresponde al CV del último generador que despachó CV_{G3}.

Ahora, como se mencionó, los retiros de energía de un cliente no distinguen de quién se está tomando la energía, sino más bien son atribuidos al generador que suministra la energía a dicho cliente. Por tanto, la transacción física no necesariamente se relaciona a las transacciones financieras que derivan del mercado de contratos.

Recapitulando, desde el punto de vista técnico el mercado eléctrico se estructura a través de tres actividades: generación, transmisión y distribución. La comercialización de la electricidad generada tiene como usuarios finales a los Usuarios Regulados y los Usuarios Libres, además de otros actores como los generadores y distribuidores cuyas transacciones se desarrollan en dos tipos de mercado: el minorista y el mayorista. En este último, a su vez, se tiene el mercado de contratos y el mercado *spot*. Si bien la organización del mercado eléctrico puede ser compleja, el entendimiento de los aspectos de su organización cobra mayor importancia dada la contribución de este sector en el desarrollo de la economía (Palma, 2022).

2.1.2. Situación actual de EGEMSA

EGEMSA son las siglas de la Empresa de Generación Eléctrica Machupicchu S.A., que desarrolla actividades de generación de energía eléctrica por medio de sus instalaciones ubicadas en el sur este del Perú, las cuales se encuentran conectadas al Sistema Eléctrico Interconectado Nacional (SEIN), teniendo su sede institucional en la ciudad del Cusco, Capital Arqueológica de América (EGEMSA, 2025).

EGEMSA es una empresa estatal de derecho privado que inicia sus operaciones en 1994, siendo su principal fuente de generación la Central Hidroeléctrica Machupicchu. Desde entonces se ha consolidado como una empresa abierta al avance tecnológico y respaldada por la experiencia de sus trabajadores, lo cual la ha convertido en una de las principales empresas generadoras de energía eléctrica en el territorio peruano, con grandes perspectivas de una mayor expansión en sus operaciones (EGEMSA, 2025).

Gestión de Producción

EGEMSA cuenta con una potencia instalada de 208.07 MW, distribuida en dos centrales de generación eléctrica: el 92.5 % corresponde a la Central Hidroeléctrica Machupicchu, y el 7.5 %, a la Central Térmica Dolorespata (en la actualidad, retirada de la operación comercial del COES) (EGEMSA, 2022).

En el 2022, las empresas de generación eléctrica adscritas al SEIN produjeron 56 084.20 GWh, de los cuales el 50.79 % proviene de centrales hidráulicas; el 44.30 %, de centrales térmicas, y el 4.91 % restante, de recursos energéticos renovables. Por su parte, las empresas de generación de energía de Fonafe (Electroperú, EGEMSA, Egasa, San Gabán y Egesur) registraron una producción de 9 829.61 GWh en el 2022 y representaron el 17.53 % de la energía generada por el SEIN, en el que EGEMSA contribuyó con el 2.09 % de la producción nacional. La producción de energía de la Central Hidroeléctrica Machupicchu en su conjunto

alcanzó los 1 172 683 MWh, de los cuales 504 721 MWh corresponden a los grupos Pelton, mientras que 667 962 MWh al grupo Francis (EGEMSA, 2022).

Gestión Comercial

La política comercial de EGEMSA, establecida en el Plan Estratégico, se orientó a fortalecer la gestión comercial, relacionada con la optimización de ingresos por venta de energía eléctrica y la suscripción de diferentes contratos de suministro de energía que incrementó y fidelizó su cartera de clientes.

La energía vendida a todos los clientes de EGEMSA durante el 2022 fue 1 169 523 MWh, de los cuales 334 459 MWh (28.60 %) corresponde al mercado regulado y 407 846 MWh (34.87 %) pertenece al mercado libre. La venta en el mercado eléctrico mayorista fue 427 218 MWh (36.53 %) (EGEMSA, 2022).

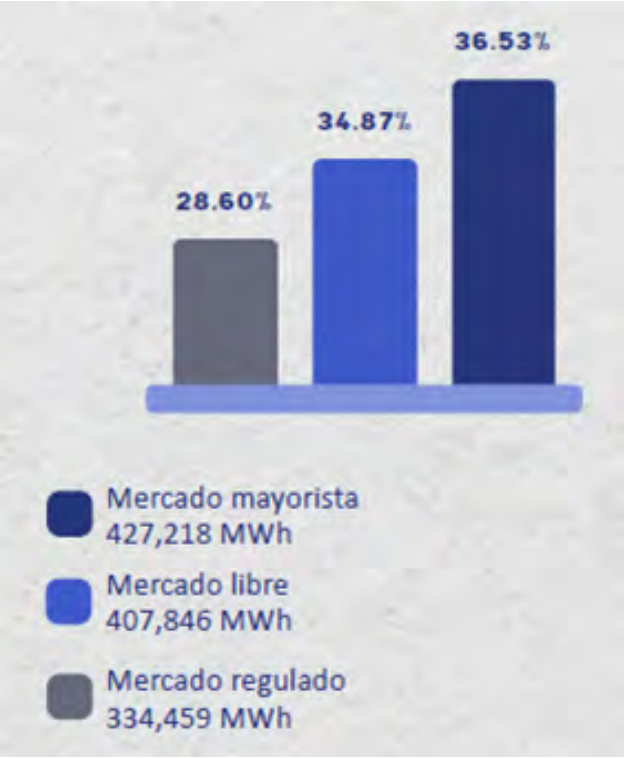

Figura 2.4: Venta de energía eléctrica por tipo de mercado

Fuente: Elaborado por (EGEMSA, 2022)

El Departamento de Comercialización es la unidad orgánica encargada de formular, ejecutar y supervisar las actividades referentes a las políticas comerciales de la empresa, así como, administrar los contratos en materia de precios y condiciones.

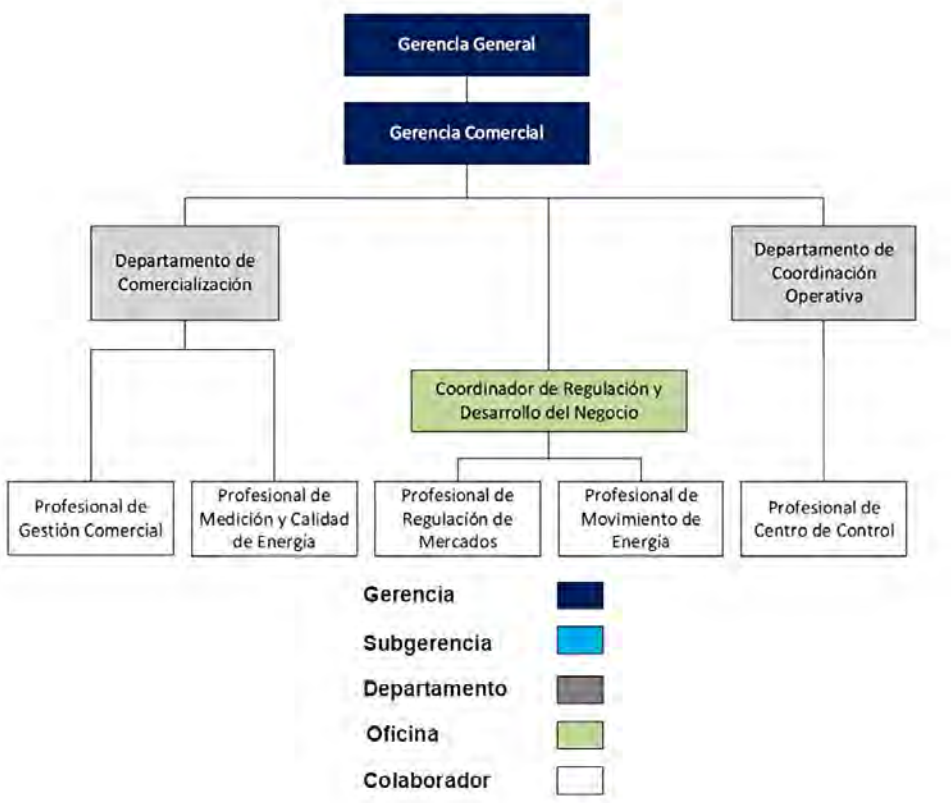

Figura 2.5: Estructura Orgánica de la Gerencia Comercial de EGEMSA

Fuente: Elaborado por (EGEMSA, 2024)

2.1.3. Big Data

Se entiende por *Big Data* el conjunto de estrategias, tecnologías y sistemas para el almacenamiento, procesamiento, análisis y visualización de conjuntos de datos complejos, que frecuentemente, pero no siempre, viene definida por volumen, velocidad y variedad (Curto, 2016; Joyanes Aguilar, 2019).

Es el acceso a grandes volúmenes de datos, pero el valor real no se encuentra en ellos,

sino en lo que podemos hacer con ellos. No es la cantidad de información lo que marca la diferencia, sino que se trata de nuestra capacidad para analizar series extensas y complejas de datos que van más allá de todo lo que hubiéramos podido hacer anteriormente. Esto significa que todas las empresas, organismos gubernamentales o cualquier persona realmente pueden utilizar el *Big Data* para mejorar la toma de decisiones (Marr, 2016; Joyanes Aguilar, 2019).

Big Data presenta desafíos y oportunidades para las organizaciones, ya que pueden extraer información valiosa, identificar patrones, tendencias y correlaciones significativas, y tomar decisiones basadas en datos a partir de estos conjuntos de datos masivos. Para gestionar y aprovechar el potencial del *Big Data*, se utilizan tecnologías y herramientas especializadas como *data lakes*, sistemas de procesamiento distribuido, análisis predictivo y *machine learning* (Fang, 2015).

Definición de Big Data

El término *Big Data* fue acuñado por Doug Laney, analista de la consultora Gartner, en 2001 (Joyanes Aguilar, 2019), para referirse a todo el conjunto de datos cuya cantidad o volumen normalmente terabytes o petabytes, velocidad y variedad exceden la capacidad de manipular y procesar la información que tienen las herramientas tradicionales. Laney se refería no sólo al volumen de datos, sino a su velocidad de generación y a la gran variedad de formatos. Este modelo se conoce como el modelo de las 3V de *Big Data*: (Joyanes Aguilar, 2019)

- **Volumen**: Tamaño global del conjunto de datos, terabytes y petabytes, aunque ya muchas empresas generan exabytes de información.
- **Velocidad**: Tiempo utilizado en la generación de los datos, así como la rapidez con que necesitan ser procesados: en tiempo real o casi en tiempo real.
- **Variedad**: Amplia gama de datos que pueden contener los conjuntos de datos que pro-

ceden de fuentes muy diversas: páginas web, texto, audio, video, fotografías, sensores, datos de máquinas, datos de dispositivos móviles, etc. Los datos se clasifican en tres tipos:

- *Estructurados*: Datos de bases de datos relacionales y heredadas, en formato tabla.
- *No estructurados*: Audio, texto, fotografías.
- *Semiestructurados*: Archivos de texto, archivos XML, etc.

(Joyanes Aguilar, 2019)

Tipos de datos en Big Data

- **Datos estructurados**: Datos tradicionales almacenados en filas y columnas (tablas) y que son los más empleados en archivos y bases de datos ordinarios de las organizaciones (Joyanes Aguilar, 2019).
- **Datos semiestructurados**: No se ajustan a un esquema fijo y explícito; no se limitan a campos determinados, mantienen marcadores para separar elementos. Tienen información poco regular, de forma que no puede ser gestionada de un modo estándar; utilizan lenguajes de marcación de hipertexto o de marcas extensibles. Ejemplos: documentos XML, HTML, datos de sensores, etc (Joyanes Aguilar, 2019).
- **Datos no estructurados**: Son los datos más complejos; se presentan en formatos que no pueden ser fácilmente manipulados por las bases de datos relacionales: archivos Word, PDF, PPT, hojas de cálculo, documentos multimedia, audio, voz, vídeo, fotografías, correos electrónicos (Joyanes Aguilar, 2019).

2.1.3.1. Data Lake

Un *Data Lake* (Lago de Datos) es un repositorio de almacenamiento que contiene una gran cantidad de datos en bruto en su formato original, incluyendo datos estructurados, semiestructurados y no estructurados, que se guardan sin ningún procesamiento (*raw data*).

Un Lago de Datos es un depósito de datos masivo y de fácil acceso para almacenar *Big Data*. Hadoop es la tecnología más utilizada para crearlos. En esencia, un *Data Lake* es un tipo de almacenamiento en el que la información almacenada tiene una estructura variable (diferentes tipos de datos, texto, imágenes, mensajes, audios, ubicaciones físicas, etc.), es masiva, de fácil y rápido acceso y resiliencia, sin atender a una lógica de negocio específica.

Los Lagos de Datos almacenan los datos en su formato más básico (en bruto), se actualizan añadiendo más información, pero nunca se modifica la información ya existente. De este modo, los *Data Lake* permiten almacenar los datos en bruto y tenerlos disponibles en todo momento y casi en tiempo real en su formato original. El uso de fuentes muy variadas permite realizar análisis complejos y modelos predictivos (Joyanes Aguilar, 2019).

Los Lagos de datos pueden almacenar datos en diversos formatos como JSON, CSV o Parquet, brindando flexibilidad para diferentes tipos de procesamiento y análisis de datos. En resumen, una arquitectura bien diseñada de lago de datos respalda el almacenamiento, procesamiento y análisis de datos para permitir que las organizaciones obtengan valiosos conocimientos de sus activos de datos (Nargesian, 2019).

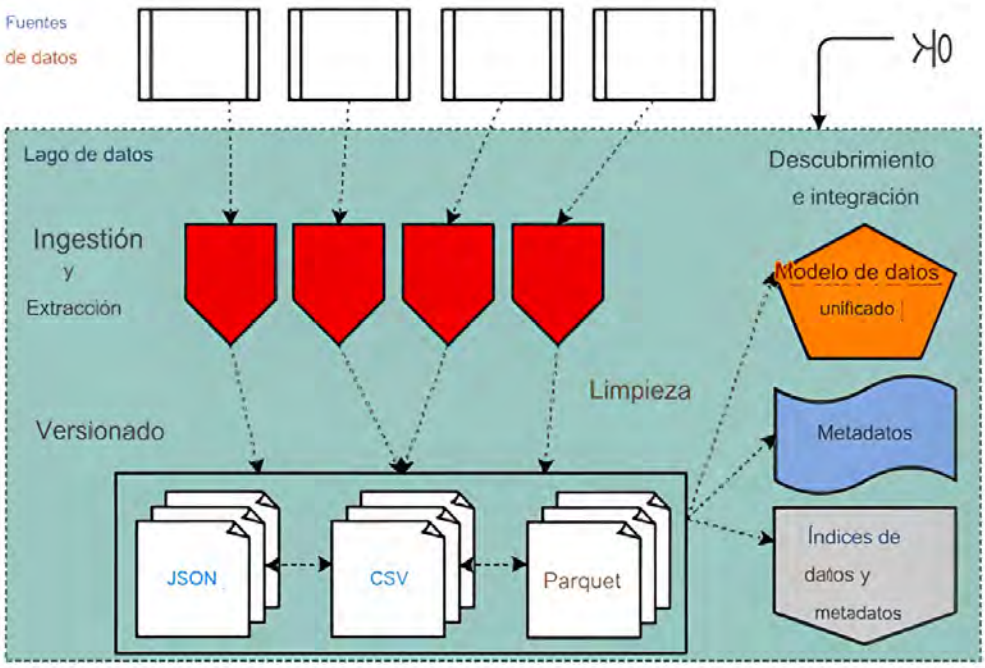

Figura 2.6: Ejemplo de sistema de gestión de lago de datos

Fuente: Elaborado por (Nargesian, 2019)

2.1.3.2. Arquitectura de Medallones (Bronce, Plata, Oro)

Una arquitectura medallón es un patrón de diseño de datos que se usa para organizar los datos lógicamente. Su objetivo es mejorar de forma incremental y progresiva la estructura y la calidad de los datos a medida que fluye a través de cada capa de la arquitectura (desde las tablas de capa *Bronce* $\Rightarrow$ *Plata* $\Rightarrow$ *Oro*).

Las arquitecturas de medallón, a veces también conocidas como arquitecturas de varios saltos, permiten que, al avanzar los datos a través de estas capas, las organizaciones mejoren incrementalmente la calidad y la confiabilidad de los datos, haciéndolos más adecuados para las aplicaciones de inteligencia empresarial y aprendizaje automático (Microsoft, 2024).

En este ejemplo de arquitectura de medallón se muestran las capas de bronce, plata y oro para su uso por parte de un equipo de operaciones empresariales.

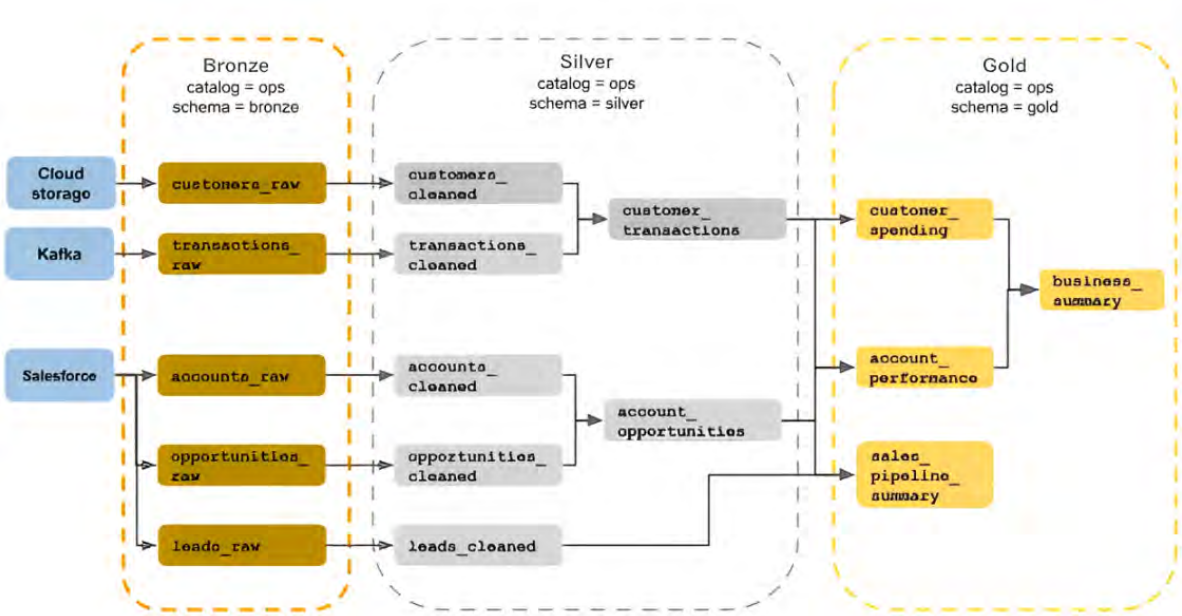

Figura 2.7: Ejemplo de una arquitectura de medallón

Fuente: Elaborado por (Microsoft, 2024)

Capa de bronce (ops.bronze)

Ingiere datos sin procesar del almacenamiento en la nube, Kafka y Salesforce. No se realiza ninguna limpieza o validación de datos en esta capa (Microsoft, 2024).

Capa de plata (ops.silver)

En esta capa se lleva a cabo la limpieza y validación de datos.

- Los datos sobre clientes y transacciones se limpian quitando valores `NULL` y eliminando registros no válidos.
- Estos conjuntos se unen para formar un nuevo conjunto denominado `customer_transactions`, que puede ser utilizado por científicos de datos para análisis predictivo.

- Las cuentas y los conjuntos de oportunidades de Salesforce se unen para crear `account_opportunities`, enriquecido con información adicional de las cuentas.
- Los datos de `leads_raw` se limpian en un conjunto denominado `leads_cleaned`.

(Microsoft, 2024)

Capa dorada (ops.gold)

Diseñada para usuarios empresariales y contiene menos conjuntos de datos que la capa plata. Incluye:

- `customer_spending`: promedio y gasto total por cliente.
- `account_performance`: rendimiento diario de cada cuenta.
- `sales_pipeline_summary`: información de la canalización de ventas de extremo a extremo.
- `business_summary`: información altamente agregada para el personal ejecutivo.

(Microsoft, 2024)

2.1.3.3. Web Scraping

La técnica de *Web Scraping* consiste en la extracción de datos, información, imágenes, documentos (entre otros) directamente desde un navegador web. Su realización puede ejecutarla un usuario de manera manual o un bot, aunque lo más frecuente es que sea de manera automatizada. Tiene variadas aplicaciones y es posible implementarlo en distintas plataformas y con distintos procesos, es decir, no hay un método único para realizar *Web Scraping* (Figueroa Gallardo, 2021).

Las extracciones de información más utilizadas son las siguientes:

- **Web Scraping estático de una sola página**: todos los datos requeridos están disponibles en una única página.
- **Web Scraping estático de varias páginas del mismo dominio**:
  - *Crawling Vertical*: extracción de información, datos o imágenes que, desde la página inicial, redirigen hacia otra.
  - *Crawling Horizontal*: extracción de información en la cual es necesario cambiar de página para recabar toda la información necesaria.
- **Web Scraping en páginas dinámicas**: extracción de información, datos, imágenes y/o documentos en páginas que requieren automatizar acciones del navegador, por ejemplo, hacer clic en un botón para cargar toda la información disponible. A diferencia de cambiar de página (donde normalmente cambia solo el número en la URL), en este caso la URL no se altera.
- **Web Scraping de API**: requiere conocer la API que define la aplicación o página web; si no se conoce, se debe analizar su estructura. Este tipo de extracción permite obtener información y recibir retroalimentación de lo ocurrido con el requerimiento del bot o usuario, y suele utilizar el formato JSON.

Lenguaje y escritura de una página web

Actualmente, los lenguajes más utilizados para escribir una página web son `JSON` (*JavaScript Object Notation*) y `XML` (*Extensible Markup Language*). Dentro de este último se encuentra `HTML` (*HyperText Markup Language*), que corresponde a un tipo de XML.

Otra opción de escritura es `PHP` (*Hypertext Preprocessor*), un lenguaje de programación utilizado para la generación de páginas web de forma dinámica, que se adapta al código

HTML. Puede considerarse como una opción para simplificar el lenguaje HTML, ya que permite introducir pequeños fragmentos de código PHP en HTML para realizar diversas funciones.

Estas incrustaciones pueden realizarse en cualquier parte del código y en más de una ocasión, alternando entre HTML y PHP en la escritura (Figueroa Gallardo, 2021).

Lenguaje HTML

Este lenguaje, usado recurrentemente para desarrollar páginas web, se basa en una estructura de árbol, donde los elementos están ordenados de manera estructurada y jerárquica.

Para entender un poco el árbol, el tronco principal es designado como `body` y cada rama corresponde a un hijo de este `body` llamado etiqueta HTML o *tag* HTML. Cada etiqueta puede o no contener más de un *tag*. La manera en que se puede diferenciar qué etiqueta es padre o hijo de otra suele representarse mediante la indentación (espacios entre el margen izquierdo y el comienzo de las letras). Para poder escribir en este lenguaje se utilizan los símbolos `< >` para encerrar el `body` o un *tag*; cada vez que se define un elemento se realiza una apertura y un cierre.

Cada *tag* puede contener en su interior una serie de elementos que lo caracterizan aún más, los cuales son llamados *atributos*; los más comunes corresponden a `class` e `id`, siendo este último un valor que no se repetirá para ninguna otra etiqueta. Adicionalmente, cada *tag* puede contener texto en su interior. Con todo lo anterior es posible armar un árbol más robusto (Figueroa Gallardo, 2021).

Librerías para extracción automatizada de datos web

- **Selenium**: herramienta con múltiples funciones para la automatización de acciones en una página web. Permite abrir una página desde el compilador, acceder a datos,

imágenes o archivos específicos, llenar formularios (como *login* y *password*), hacer clic, cambiar de pestaña, etc. Es compatible con XML y lenguaje XPATH, siendo muy útil para *Web Scraping* dinámico (Figueroa Gallardo, 2021).

- **Requests**: librería que permite acceder de manera sencilla a cualquier URL, independientemente de si requiere o no autenticación. Es muy utilizada para extraer contenido desde la web y principalmente para la descarga de archivos que, en el código fuente de la página web, contengan un enlace específico de descarga (Figueroa Gallardo, 2021).

- **Scrapy**: una de las librerías más comunes para realizar *Web Scraping*. Permite realizar peticiones al servidor y analizar la información para la recolección de datos. Sus principales usos corresponden al *crawling* vertical y horizontal, incorporando funcionalidades para recorrer múltiples páginas relacionadas por paginación (Figueroa Gallardo, 2021).

- **Beautiful Soup (bs4)**: librería utilizada para recorrer y analizar el código HTML completo de una página, pudiendo extraer información, valores o imágenes deseadas (Figueroa Gallardo, 2021).

- **Mechanize**: librería utilizada principalmente para rellenar y enviar formularios de manera automática, activar el uso de cookies y manejar acciones que podrían interrumpir la automatización del proceso de extracción de datos. Permite navegar por la web gracias a funciones que listan todos los enlaces y formularios presentes en una página (Figueroa Gallardo, 2021).

2.1.3.4. Proceso ETL

La capa *ETL* (*Extract, Transform, Load*) se centra en tres procesos principales: extracción, transformación y carga de los datos (Joyanes Aguilar, 2019).

Extracción: proceso de identificación y recolección de datos relevantes o significativos de

diferentes fuentes. Normalmente, los datos extraídos de fuentes internas y externas no están integrados y pueden estar incompletos o duplicados. Este proceso selecciona únicamente los datos significativos para la toma de decisiones en las organizaciones. Los datos extraídos se envían a un área de almacenamiento temporal denominada *Data Staging*, previa al proceso de transformación y limpieza.

Transformación: conversión de los datos utilizando un conjunto de reglas de negocio (por ejemplo, funciones de agregación) para obtener formatos consistentes que faciliten la generación de informes y análisis. Una vez que los datos se han limpiado y transformado, se almacenan nuevamente en el área temporal (*Staging Area*).

Carga: fase final en la que los datos del área de *staging* se transfieren al repositorio de destino (por ejemplo, *Data Warehouse* o *Data Marts*), normalmente a través de un almacén de datos operacional.

2.1.4. Inteligencia Artificial y Machine Learning

La inteligencia artificial (IA) se define como la capacidad de las computadoras para realizar actividades que normalmente requieren inteligencia humana. Esto abarca el uso de algoritmos, el aprendizaje a partir de datos y la toma de decisiones de manera similar a como lo haría un humano. Las máquinas basadas en IA pueden analizar grandes cantidades de información rápidamente y con significativamente menos errores en comparación con los humanos (Lasse Petteri, 2018).

de Mántaras (2017) considera que la Inteligencia Artificial tiene como objetivo diseñar algoritmos que, una vez programados, doten de comportamiento inteligente a las máquinas. Por su parte, McCarthy (2007), uno de los padres de la Inteligencia Artificial, explicó que entendía por Inteligencia Artificial (IA) "la ciencia y la ingeniería de crear máquinas inteligentes, especialmente programas de computación inteligentes. Está relacionada con la tarea

similar de utilizar ordenadores para comprender la inteligencia humana, pero la IA no se limita a métodos que sean observables biológicamente".

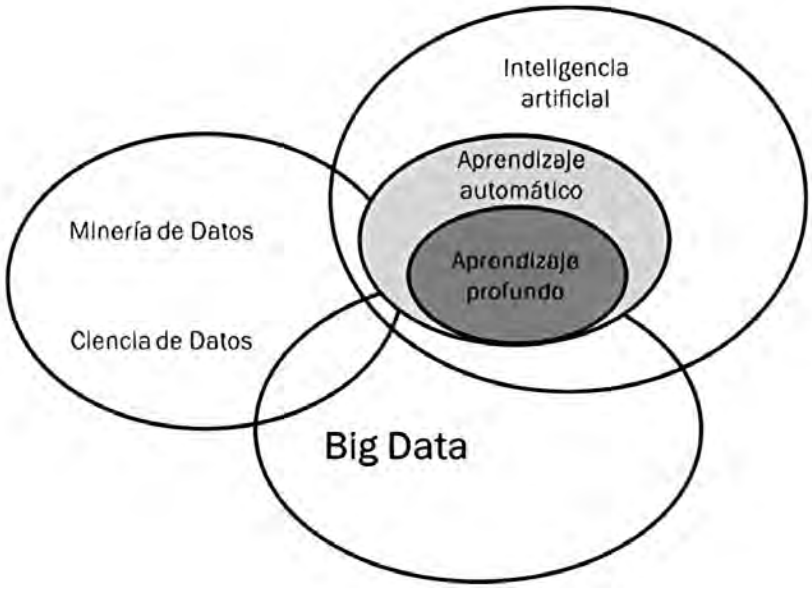

Figura 2.8: Diagrama de venn de Data Science

Fuente: (Joyanes Aguilar, 2019, Mattew Mayo, KDnuggets, 2016; citado en)

2.1.4.1. Aprendizaje Automático

Machine Learning (ML) se traduce al español como "aprendizaje automático", aunque también se traduce como "aprendizaje máquina". Se suele considerar una rama de la Inteligencia Artificial que busca construir algoritmos que permitan a las computadoras "aprender" a partir de conjuntos de datos y obtener como resultado un modelo que permita realizar predicciones, basándose en dichos datos y no en instrucciones estáticas. El aprendizaje automático es una disciplina que toma experiencias de otras disciplinas, tales como la estadística, la complejidad computacional, ciencias de la computación e ingeniería. La expansión del aprendizaje automático como disciplina complementaria o autónoma de la Inteligencia Artificial se debe, esencialmente, al diluvio de los datos (*Big Data*) que se han producido estos últimos años. Existen diferentes tipos de algoritmos que dan diferentes categorías de aprendizaje

(Joyanes Aguilar, 2019).

Aprendizaje supervisado (*Supervised Learning*). Este tipo de aprendizaje utiliza un conjunto de datos con entradas y salidas conocidas. El modelo se entrena para hacer predicciones sobre nuevos datos, ajustando sus parámetros internos para minimizar el error (Alvarez Rubio, 2024). Requiere intervención humana para indicar qué está bien y qué está mal. En muchas otras aplicaciones de la computación cognitiva, los humanos también proporcionan parte de la semántica necesaria para que los algoritmos aprendan (Joyanes Aguilar, 2019).

Aprendizaje no supervisado (*Unsupervised Learning*). En este caso, los algoritmos analizan datos que no tienen etiquetas, buscando patrones o grupos dentro de la información. Es especialmente útil en la exploración de grandes conjuntos de datos (Alvarez Rubio, 2024). La red aprende a reconocer características y a agrupar ejemplos similares (Joyanes Aguilar, 2019).

Aprendizaje reforzado (*Reinforced Learning*). Este enfoque se basa en que un agente puede aprender a tomar decisiones mediante la interacción con su entorno. Se premian las acciones correctas y se penalizan las incorrectas (Alvarez Rubio, 2024). Es un híbrido entre el aprendizaje supervisado y no supervisado, inspirado en la psicología conductista (Joyanes Aguilar, 2019).

2.1.4.2. Agrupamiento (*Clustering*)

El *agrupamiento* o *clustering* es una técnica de aprendizaje no supervisado que consiste en dividir un conjunto de datos en grupos o clústeres, de manera que los elementos dentro de cada grupo sean más similares entre sí que con los de otros grupos. `scikit-learn`, una biblioteca de Python para aprendizaje automático, ofrece diversas implementaciones de algoritmos de *clustering* que pueden adaptarse a diferentes necesidades y tipos de datos

(scikit-learn, 2025).

K-Means

El algoritmo de K-Medias es uno de los más conocidos y utilizados en la técnica de agrupamiento. Su principal objetivo es dividir un conjunto de datos en K clústeres, minimizando la variación dentro de cada clúster y maximizando la variación entre clústeres distintos (Alvarez Rubio, 2024). El método de agrupamiento no supervisado busca encontrar la distancia mínima entre un conjunto de datos y el centro de cada grupo, generando así una partición en k grupos a partir de n observaciones. Cada grupo está representado por el promedio de los puntos que lo conforman, y el valor más representativo de cada grupo se llama *centroide*. El parámetro k, que indica la cantidad de grupos a descubrir, debe establecerse previamente.

Una manera de determinar el número de grupos antes de aplicar el algoritmo K-Means es mediante el *método del codo*. Este método calcula la suma de las distancias al cuadrado desde cada punto hasta su centroide asignado en cada iteración de K-Means. Durante cada iteración, se ejecuta el algoritmo con un número distinto de grupos, lo que resulta en un gráfico que muestra la suma de las distancias al cuadrado en función del número de grupos.

Uno de los desafíos principales del algoritmo K-Means es que su resultado puede variar para un mismo conjunto de datos, debido a que los centroides iniciales se seleccionan de forma aleatoria. Esta característica tiene un impacto directo en todo el proceso del algoritmo y puede generar resultados diferentes en cada ejecución (A. Soto, 2023).

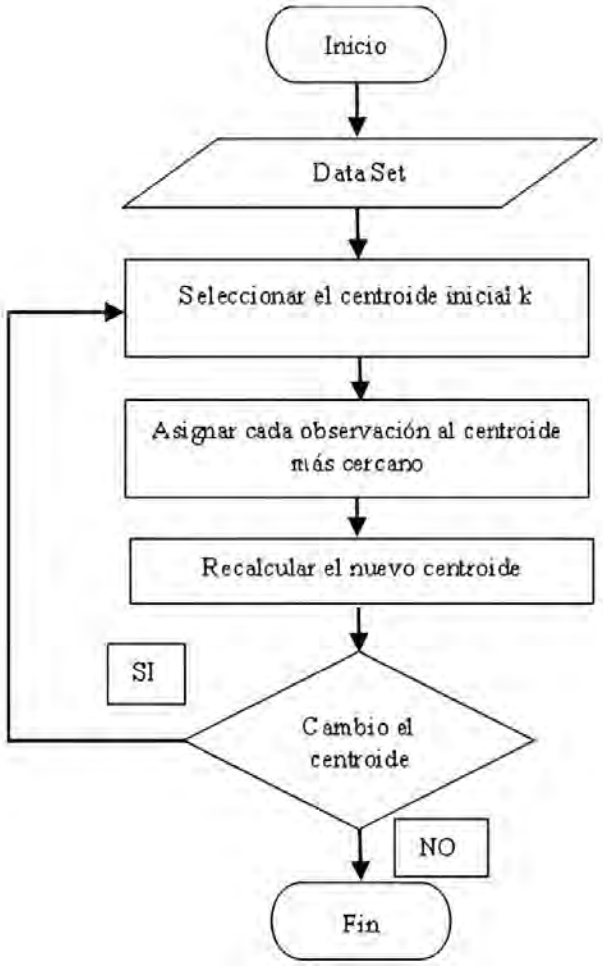

Figura 2.9: Diagrama de Flujo Algoritmo K-Means

Fuente: Elaborado por (A. Soto, 2023)

Pasos para implementar el algoritmo K-Means (A. Soto, 2023):

1. Determinar el número k de grupos que se pretenden encontrar.

2. Elegir aleatoriamente k análisis del conglomerado de datos como centroides primarios:

$$Z_1^{(0)}, Z_2^{(0)}, \ldots, Z_k^{(0)}$$

3. Asignar cada observación x al centroide más cercano utilizando la distancia euclidiana.

La asignación de la muestra x al clúster $C_i^{(k)}$ se define como:

$$x \in C_i^{(k)}, \quad \text{si} \quad d\left(x, Z_i^{(k)}\right) < d\left(x, Z_j^{(k)}\right), \quad i = 1, 2, \ldots, k; \quad i \neq j$$

4. Recalcular los centroides de cada uno de los k clústeres:

$$Z_i^{(k+1)} = \frac{1}{n_i} \sum_{x \in C_i} (k)x, \quad i = 1, 2, \ldots, k$$

   donde n_i es el número de elementos en $C_i^{(k)}$.

5. Repetir los pasos 3 y 4 hasta que las asignaciones no cambien o se alcance el número máximo de iteraciones.

DBSCAN (*Density-Based Spatial Clustering of Applications with Noise*)

El algoritmo DBSCAN es una herramienta diseñada para identificar grupos y ruido en bases de datos espaciales. Define los grupos como el conjunto más extenso de puntos conectados con una densidad específica. Entre sus ventajas se destacan su simplicidad y su capacidad para descubrir agrupaciones con características diversas, revelando valores especiales.

Para la implementación de DBSCAN se requiere de forma previa conocer dos parámetros principales:

- **Epsilon (ε)**: la distancia máxima entre dos puntos cercanos.
- **MinPts**: el número mínimo de puntos cercanos alrededor de un punto especificado para ser determinado como punto central.

Con los parámetros indicados, cada observación puede clasificarse como un punto central, un punto de borde o un punto considerado como ruido (A. Soto, 2023).

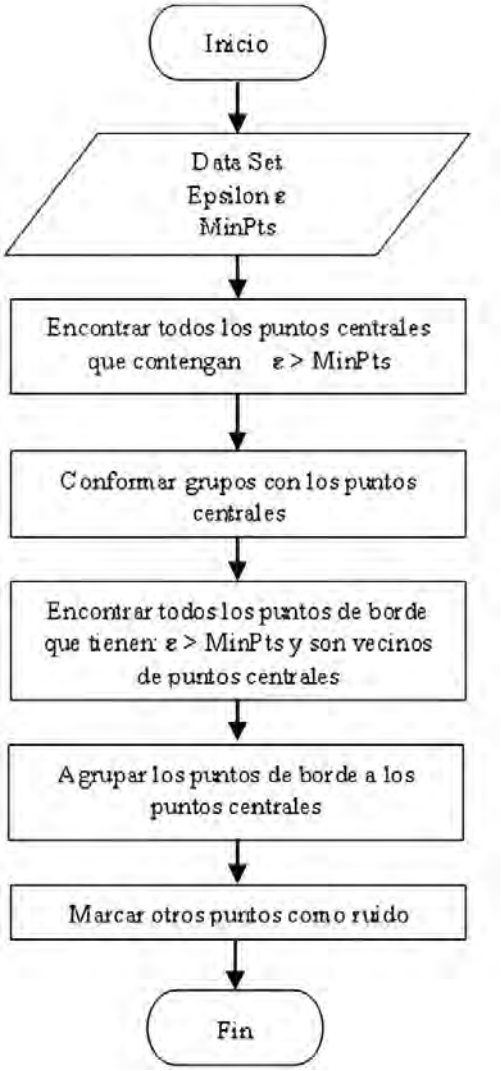

Figura 2.10: Diagrama de Flujo Algoritmo DBSCAN

Fuente: Elaborado por (A. Soto, 2023)

El proceso de DBSCAN continúa hasta que todos los objetos han sido procesados. Los puntos que no se asignan a ningún grupo se consideran puntos de ruido, mientras que aquellos que no son ni ruido ni puntos centrales se denominan puntos de borde. De esta manera, DBSCAN construye grupos donde los puntos son clasificados como puntos centrales o puntos de borde, y es posible que un grupo tenga más de un punto central.

El algoritmo comienza seleccionando un punto p arbitrario. Si p es un punto central, se

forma un grupo y se incluyen todos los objetos alcanzables desde p. Si p no es un punto central, se visita otro objeto del conjunto de datos (A. Soto, 2023).

Clustering Jerárquico

Los métodos de clústeres jerárquicos son una técnica de agrupamiento que permite organizar un conjunto de datos en una jerarquía de clústeres o grupos, desde los más generales hasta los más específicos.

Estos métodos son ampliamente utilizados en análisis de datos, exploración de datos y minería de datos debido a su capacidad para visualizar la estructura de los datos y su flexibilidad en la aplicación a diferentes tipos de conjuntos de datos (Chester and Maecker, 2015; Alvarez Rubio, 2024).

Aglomerativos

El enfoque aglomerativo es uno de los más comunes dentro de los métodos jerárquicos. En este método, el proceso de agrupamiento comienza con cada observación (o punto de datos) considerada como un clúster individual (Rodríguez, 2023; Alvarez Rubio, 2024). Por lo tanto, si se tienen N observaciones, se inicia con N clústeres. El proceso sigue avanzando y se fusionan los clústeres más cercanos en etapas sucesivas, formando clústeres más grandes.

Los pasos básicos del algoritmo aglomerativo son:

1. **Cálculo de la matriz de distancias**: Se mide la distancia entre cada par de puntos de datos, utilizando métricas como la distancia euclidiana, Manhattan o la distancia de Minkowski.
2. **Fusión de clústeres**: Se identifica el par de clústeres más cercanos y se combinan.
3. **Actualización de la matriz de distancias**: Tras la fusión, se recalcula la distancia entre los nuevos clústeres y los restantes.

4. **Repetición**: Se repiten los pasos anteriores hasta que todos los puntos se unan en un solo clúster o se alcance un número predefinido de clústeres.

El resultado de este proceso se puede representar mediante un *dendrograma*, que es una representación gráfica que muestra la relación entre los clústeres y las distancias a las que se fusionan (Alvarez Rubio, 2024).

Divisivos

El método divisivo, en contraste con el enfoque aglomerativo, comienza con todos los datos formando un único clúster y posteriormente lo divide en clústeres más pequeños (Rodríguez, 2023; Alvarez Rubio, 2024). Este método puede ser menos intuitivo que el aglomerativo, pero también ofrece ventajas dependiendo de la naturaleza de los datos.

Los pasos básicos del método divisivo son:

1. **Inicio con un único clúster**: Todos los datos se inicializan en un solo clúster.
2. **División de clústeres**: En cada iteración, se selecciona un clúster para dividir, basándose en criterios como la varianza dentro del clúster.
3. **División recursiva**: El proceso de división se repite para los clústeres obtenidos hasta que se alcanza un número deseado o se cumplen otras condiciones de parada.

Ambos métodos, aglomerativo y divisivo, ofrecen versatilidad y adaptabilidad al análisis de clústeres, siendo herramientas valiosas en la ciencia de datos y el aprendizaje automático. Sin embargo, la elección entre ellos depende de la estructura de los datos y los objetivos específicos del análisis (Alvarez Rubio, 2024).

2.1.4.3. Reducción de dimensionalidad

Principal component analysis (PCA)

Este es un algoritmo de reducción de dimensionalidad mediante cálculo de covarianza y una transformación lineal. Es decir, se reduce la dimencionalidad de datos, conservando aquellos que representan de mejor forma la variación dentro de la muestra.

A continuación, se resumen los pasos que realiza un algoritmo de pca. Para esto se contempla un conjunto de datos compuesto por N muestras (por ejemplo, clientes del sistema) y M dimensiones (por ejemplo, su consumo horario en un año).

1. **Estandarización**: Se realiza una estandarización de cada dimensión para que las variables puedan ser comparables (es distinto comparar una variable que va de 0 a 1, que una de 0 a 100).

2. **Matriz de covarianza**: Se calcula la matriz de covarianza del conjunto de datos. Esto resulta en una matriz de MxM.

3. **Vectores y valores propios**: Se calculan los vectores y valores propios de la matriz de covarianza. Los valores propios indican cuánta información de varianza contiene su vector propio asociado respecto a las demás componentes. De esta forma se ordenan los valores propios desde los que contienen mayor información a menor. Luego se selecciona un número p de componentes a preservar, que indicará la nueva dimensionalidad del conjunto. Es decir, si antes tenía M dimensiones, ahora se tienen p.

4. **Transformación**: Los vectores propios asociados a las p componentes escogidas se genera una matriz de transformación. Luego se realiza la transformación lineal al dataset inicial estandarizado (Guiraldes Deck, 2020).

2.1.4.4. Métricas de validación

Coeficiente de Silueta

El **Coeficiente de Silueta** se calcula utilizando la distancia media dentro del conglomerado (a) y la distancia media entre conglomerados más cercanos (b) para cada muestra.

El Coeficiente de Silueta de una muestra se define como:

$$s = \frac{b - a}{\text{máx}(a, b)}$$

Para aclarar, b es la distancia entre una muestra y el conglomerado más cercano que no forma parte de la muestra. Ten en cuenta que el Coeficiente de Silueta sólo se define si el número de etiquetas cumple:

$$2 \leq n_{\text{labels}} \leq n_{\text{samples}} - 1$$

Esta función devuelve el Coeficiente de Silueta medio sobre todas las muestras. Para obtener los valores de cada muestra, se utiliza `silhouette_samples`.

El mejor valor es 1 y el peor es -1. Los valores cercanos a 0 indican conglomerados superpuestos. Los valores negativos suelen indicar que una muestra ha sido asignada al conglomerado equivocado, ya que otro conglomerado es más similar (scikit-learn, 2025).

Índice de Davies-Bouldin

La puntuación se define como la medida promedio de similitud de cada clúster con su clúster más similar, donde la similitud es la razón entre las distancias dentro del clúster y las distancias entre clústeres. Por lo tanto, los clústeres que están más separados entre sí y son menos dispersos resultarán en una mejor puntuación.

La puntuación mínima es cero, y valores más bajos indican una mejor agrupación (scikit-learn, 2025).

Inercia o Suma de Cuadrados Dentro de los Clústeres (WSS)

Para obtener el K óptimo (Yajure Ramírez, 2022). empleó el método del codo, utilizando como métrica la inercia, la cual es la función de costo de los clústeres, y consiste en minimizar la sumatoria de los cuadrados de la diferencia de cada punto al centro del clúster al que pertenecen.

2.1.5. Herramientas tecnológicas adicionales

2.1.5.1. Python

Python es un lenguaje de programación potente y fácil de aprender. Tiene estructuras de datos de alto nivel eficientes y un simple pero efectivo sistema de programación orientado a objetos. La elegante sintaxis de Python y su tipado dinámico, junto a su naturaleza interpretada lo convierten en un lenguaje ideal para scripting y desarrollo rápido de aplicaciones en muchas áreas, para la mayoría de las plataformas.

El intérprete de Python y la extensa librería estándar se encuentran disponibles libremente en código fuente y de forma binaria para la mayoría de las plataformas desde la Web de Python, https://www.python.org/, y se pueden distribuir libremente. El mismo sitio también contiene distribuciones y referencias a muchos módulos libres de Python de terceros, programas, herramientas y documentación adicional.

El intérprete de Python es fácilmente extensible con funciones y tipos de datos implementados en C o C++ (u otros lenguajes que permitan ser llamados desde C). Python también es apropiado como un lenguaje para extender aplicaciones modificables (Python Software

Foundation, 2025).

2.1.5.2. Scikit-learn

Scikit-learn es un módulo de Python que integra una amplia gama de algoritmos de aprendizaje automático de última generación para problemas supervisados y no supervisados a mediana escala. Este paquete se centra en llevar el aprendizaje automático a los no especialistas utilizando un lenguaje de alto nivel de propósito general.

Se hace hincapié en la facilidad de uso, el rendimiento, la documentación y la consistencia de la API. Tiene dependencias mínimas y se distribuye bajo la licencia de software de código abierto simplificada, fomentando su uso tanto en entornos académicos como comerciales. El código fuente, los binarios y la documentación se pueden descargar desde http://scikit-learn.sourceforge.net (Pedregosa, 2011).

Scikit-learn es una biblioteca de aprendizaje automático de código abierto que admite aprendizaje supervisado y no supervisado. También proporciona varias herramientas para ajuste del modelo, preprocesamiento de datos, selección del modelo, evaluación del modelo, etc y muchas otras utilidades (scikit-learn, 2025).

2.1.5.3. Spyder IDE

Spyder es un entorno científico e IDE de código abierto, desarrollado de forma comunitaria y escrito en Python para Python. Desarrollado para combinar la potencia de una herramienta de desarrollo integral con la velocidad de un paquete de exploración de datos interactivo, todo en una interfaz fácil de usar.

Los analistas de datos, científicos e ingenieros requieren mucha experimentación, retroalimentación rápida y ciclos de iteración cortos al programar. Spyder fue diseñado desde el

inicio en torno a ese flujo de trabajo.

Se integra muy bien con las librerías científicas más populares Spyder incluye integración lista para usar con Matplotlib, Pandas y muchas otras librerías para ayudarte a trabajar más eficientemente con ellas.

Spyder ayuda a pasar de archivos individuales a módulos y paquetes estructurados y reutilizables sin perder interactividad. También incluye potentes herramientas de desarrollo de software (Spyder Project Contributors, 2025).

2.1.5.4. Visual Studio Code

Visual Studio Code cuenta con un editor de código fuente increíblemente rápido, perfecto para el uso diario. Con soporte para cientos de idiomas, VS Code lo ayuda a ser productivo al instante con resaltado de sintaxis, coincidencia de corchetes, autoindentación, selección de cajas, fragmentos y más. Los atajos de teclado intuitivos, la fácil personalización y las asignaciones de atajos de teclado aportadas por la comunidad le permiten navegar por su código con facilidad.

Para una codificación seria, a menudo se beneficiará de herramientas con más comprensión del código que solo bloques de texto. Visual Studio Code incluye soporte incorporado para la finalización del código IntelliSense, la comprensión y navegación del código semántico enriquecido y la refactorización del código.

Y cuando la codificación se pone difícil, los difíciles consiguen depuración. La depuración es a menudo la única característica que los desarrolladores pierden más en una experiencia de codificación más ágil, por lo que lo hicimos posible. Visual Studio Code incluye un depurador interactivo, por lo que puede pasar por el código fuente, inspeccionar variables, ver pilas de llamadas y ejecutar comandos en la consola.

VS Code también se integra con las herramientas de compilación y scripting para realizar tareas comunes que hacen que los flujos de trabajo cotidianos sean más rápidos. VS Code tiene soporte para Git para que pueda trabajar con el control de origen sin salir del editor, incluida la visualización de cambios pendientes (Microsoft Corporation, 2025).

2.1.5.5. Anaconda

Anaconda ofrece herramientas para construir modelos de ciencia de datos y aprendizaje automático, desplegar tu trabajo en producción y gestionar equipos de ingenieros de forma segura.

Entre las herramientas destacadas se encuentra Anaconda Navigator, una interfaz gráfica de usuario que permite lanzar aplicaciones y gestionar paquetes, entornos y canales de conda sin necesidad de utilizar la línea de comandos. Navigator puede buscar paquetes en Anaconda.org o en un repositorio local de Anaconda y está disponible para Windows, macOS y Linux.

Además, Anaconda proporciona una distribución de ciencia de datos en Python que incluye más de 300 paquetes populares de ciencia de datos y aprendizaje automático. Para una instalación más ligera, Miniconda ofrece solo conda y sus dependencias, permitiendo a los usuarios instalar paquetes según sea necesario (Anaconda, Inc., 2025).

2.1.5.6. Microsoft SQL Server

Microsoft SQL Server es un sistema de gestión de bases de datos relacionales desarrollado por Microsoft. Permite a las aplicaciones y herramientas conectarse a una instancia o base de datos y comunicarse mediante Transact-SQL (T-SQL).

Entre sus principales características se destacan:

- Inteligencia en todos tus datos: Permite consultar diversas plataformas de datos, como Azure SQL Database, Azure Cosmos DB, MySQL, PostgreSQL, MongoDB, Oracle y Teradata, sin necesidad de mover o replicar los datos.
- Elección de idioma y plataforma: Ofrece la flexibilidad de ejecutar SQL Server en entornos locales, en la nube y en el perímetro, utilizando contenedores de Windows y Linux, y gestionando implementaciones con Kubernetes.
- Rendimiento líder en el sector: SQL Server ha demostrado un rendimiento excepcional en pruebas comparativas de cargas de trabajo OLTP y almacenamiento de datos, destacando en aplicaciones del mundo real.
- Plataforma de datos segura: Proporciona una infraestructura escalable que permite a las organizaciones implementar soluciones de inteligencia empresarial en toda la empresa, ofreciendo seguridad y confiabilidad en la gestión de datos (Microsoft, 2025).

2.1.5.7. Pandas

Cuando se trabaja con datos tabulares, como los datos almacenados en hojas de cálculo o bases de datos, pandas es la herramienta adecuada para usted. pandas le ayudará para explorar, limpiar y procesar sus datos. En Pandas, una tabla de datos se llama DataFrame.

Pandas es una biblioteca de Python que permite integrar y manejar datos desde múltiples fuentes y formatos (CSV, Excel, SQL, JSON, Parquet, entre otros) mediante funciones `read_*` y `to_*`. Facilita la selección, filtrado y extracción de filas o columnas, así como la visualización de datos gracias a su integración con `Matplotlib`. Permite realizar manipulaciones eficientes sin necesidad de bucles, como crear nuevas columnas a partir de otras y calcular estadísticas básicas (media, mediana, mínimo, máximo, conteos). Ofrece herramientas avanzadas como agrupaciones, tablas pivote, transformaciones de estructura (`melt()`, `pivot()`), concatenación y fusiones de tablas. Además, tiene soporte robusto para series

temporales y manejo de fechas, y provee funciones para limpiar y procesar datos textuales, lo que lo convierte en una herramienta versátil y esencial para el análisis de datos. (Pandas, 2025).

Capítulo 3

Metodología de la investigación

La metodología se sustenta en la revisión bibliográfica exhaustiva, la recolección de datos a través de *Web Scraping* y la creación de un *Data Lake* para almacenar y gestionar la información. Posteriormente, se realizan procesos *ETL* para la limpieza y normalización de los datos en la capa Plata. En la capa Oro, se implementan algoritmos de IA (*K-Means*) para identificar patrones relevantes del consumo de energía eléctrica, siguiendo un flujo de trabajo que incorpora validaciones mediante métricas de calidad de los clústeres (por ejemplo, Método del Codo, Coeficiente de Silhouette).

3.1. Tipo, enfoque y diseño de la investigación

La investigación adopta un enfoque cuantitativo y es de tipo aplicada, orientada a resolver el problema de gestión y aprovechamiento de grandes volúmenes de datos en EGEMSA. El nivel es descriptivo - exploratorio, pues se caracteriza el comportamiento de las transacciones de energía y se exploran patrones de clientes mediante técnicas de clustering. El diseño es no experimental, transversal y retrospectivo, dado que se trabaja con datos históricos del COES sin manipular variables y considerando un periodo de análisis determinado.

Metodológicamente, el estudio se estructura en tres etapas:

1. Recolección y almacenamiento de datos mediante Web Scraping y construcción de un Data Lake.
2. Preparación y transformación a través de un pipeline ETL para limpieza, estandarización y generación de variables analíticas.
3. Modelado y análisis con el algoritmo de Machine Learning K-Means para segmentar clientes.

Se emplean métodos analítico, sintético e inductivo, y técnicas específicas como el método del codo y el coeficiente de silueta para determinar y evaluar la calidad de los clusters. Las principales métricas consideradas incluyen la inercia, el coeficiente de silueta, indicadores descriptivos por cluster (tamaños, promedios, participación en energía y valorizaciones) y criterios de relevancia comercial definidos por la Gestión Comercial de EGEMSA.

Como instrumentos, se utilizan scripts de Web Scraping, flujos ETL y un diccionario de datos, además de herramientas de software para almacenamiento, procesamiento, análisis estadístico y visualización. La validación de resultados se complementa con el juicio del experto de EGEMSA, quien evalúa la coherencia y utilidad de los clusters para apoyar la toma de decisiones en el contexto del mercado eléctrico peruano.

3.2. Proceso de ciencia de datos

Según explican (Cielen et al., 2016, p. 23), el proceso de ciencia de datos se organiza típicamente en seis pasos principales: definir el objetivo de la investigación, recopilar los datos, prepararlos, explorarlos, construir y evaluar el modelo, y finalmente presentar o automatizar los hallazgos.

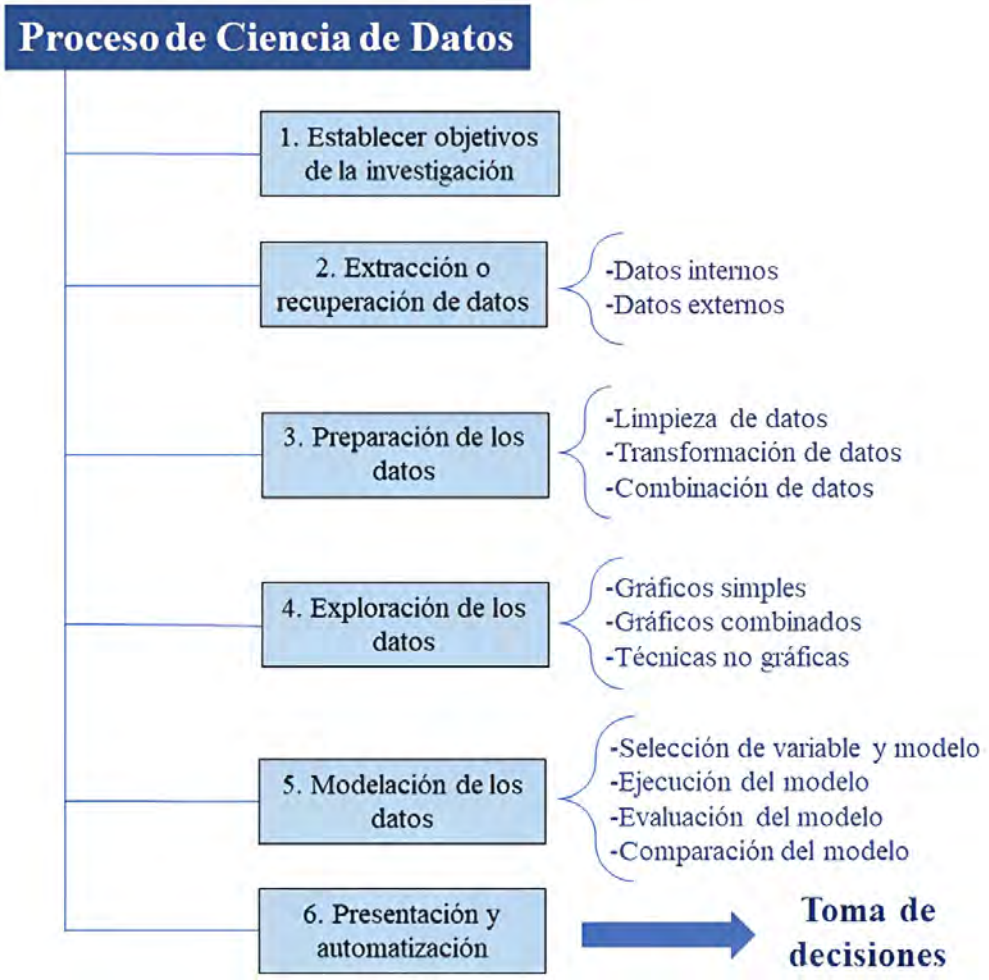

Figura 3.1: Descripción general del proceso de Machine Learning

Nota. Adaptado de "Introducing Data Science" de (Cielen et al., 2016, p. 23). Por Ramírez (2022)

3.2.1. Establecer objetivos de la investigación

En esta primera etapa, se estableció con precisión el propósito de aplicar técnicas de Big Data e Inteligencia Artificial para reforzar la capacidad analítica de la Gerencia Comercial de EGEMSA. Véase Capítulo 1, Sección 1.5 Objetivos, asegurando la alineación con la estrategia de la Gerencia Comercial y la relevancia para el contexto competitivo del mercado eléctrico peruano.

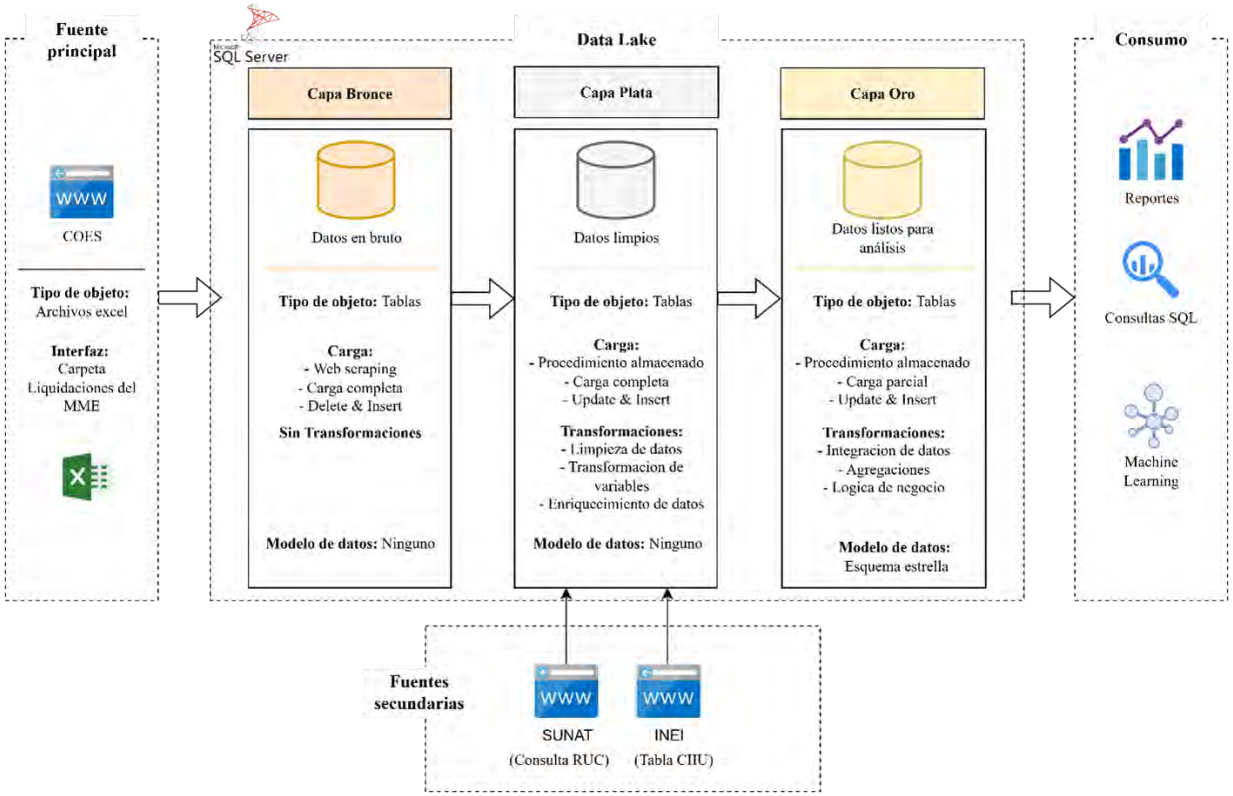

Figura 3.2: Data Lake implementado en EGEMSA

Fuente: Elaboración propia

La Figura 3.2 representa la arquitectura del Data Lake utilizada para la gestión y análisis del consumo de energía eléctrica, estructurada en tres capas: Bronce, Plata y Oro. En la capa Bronce se almacenan los datos en bruto obtenidos mediante web scraping y archivos Excel provenientes de fuentes principales como COES, sin aplicar transformaciones. En la capa Plata, los datos se limpian, normalizan y enriquecen con información de fuentes secundarias (como SUNAT e INEI) mediante procesos ETL, garantizando su calidad y coherencia. Finalmente, en la capa Oro los datos se integran y transforman siguiendo reglas de negocio, para facilitar su análisis. Los datos resultantes son utilizados en la etapa de consumo, donde se generan reportes, se ejecutan consultas SQL y se aplican algoritmos de Machine Learning para identificar patrones y obtener información valiosa para la toma de decisiones.

3.2.2. Extracción o recuperación de datos

En este paso, con el apoyo del personal especializado de la Gerencia Comercial de EGEMSA se identificó y se tuvo acceso a las fuentes de información publicados por el COES. También se contemplan mecanismos de recopilación automatizada como Web Scraping y se efectúan validaciones iniciales para asegurar la pertinencia y exactitud de los conjuntos de datos.

3.2.2.1. Datos obtenidos del COES

Se obtuvo datos del Portal Web del COES (2025) específicamente de la página de Liquidaciones del MME.

En esta página web se encuentran archivos en formato Excel (.xlsx) los cuales contienen información mensual organizada por cada año, correspondiente a las valorizaciones de transferencias de Energía realizada entre los participantes del MME del Perú.

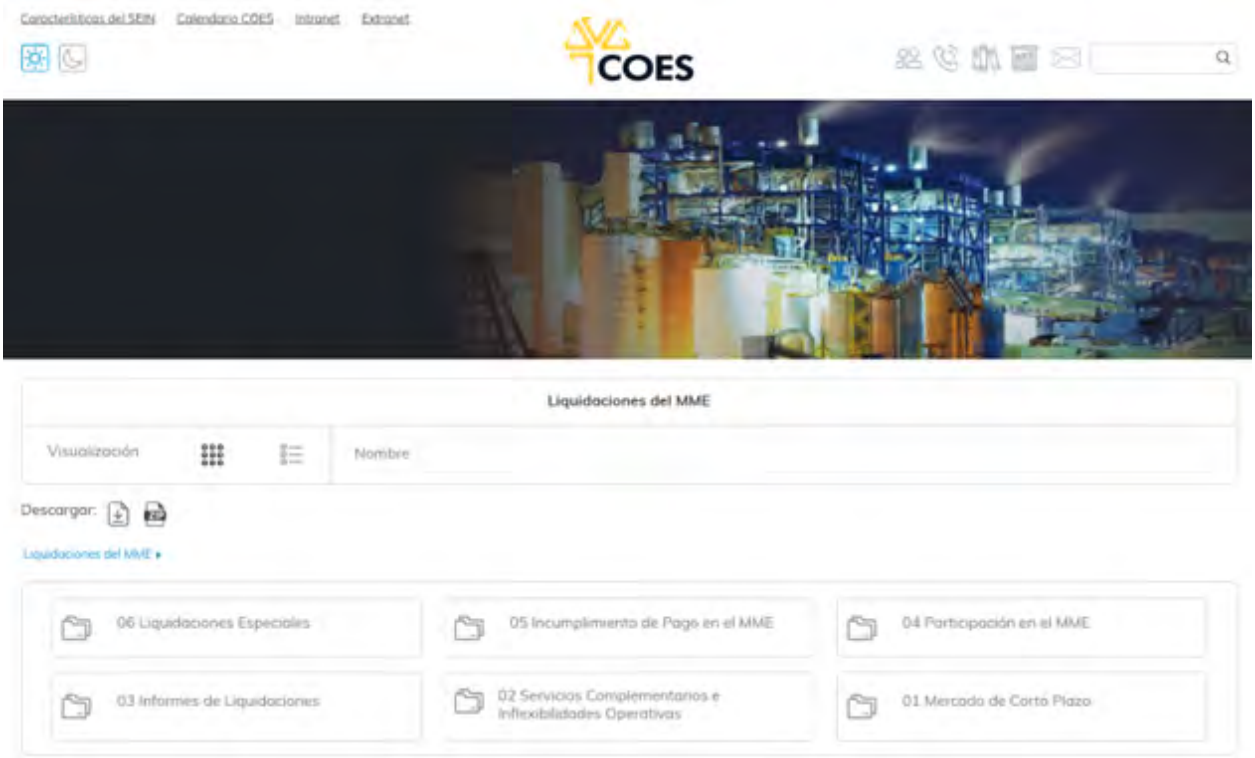

Figura 3.3: Portal web del COES

Fuente: Obtenido del portal web del COES (2025).

Para acceder a los datos, se debe seguir los siguientes pasos:

1. Acceder a la carpeta llamada “01 Mercado de Corto Plazo”
2. Acceder a la carpeta llamada “Liquidaciones VTEA”
3. Acceder a la carpeta llamada según el año (2018, 2019, 2020, etc.)
4. Acceder a la carpeta llamada según el mes (01_Enero, 02_Febrero, 03_Marzo, etc.)
5. Dependiendo si hay revisiones a la data, acceder a la carpeta llamada “Revisión [número de revisión]” (para nuestro análisis se usó la última revisión disponible a abril 2025). En caso no haya revisiones, acceder a la carpeta llamada “Mensual”
6. Identificar el archivo tipo Excel que contiene la palabra “Resumen_cuadros”

7. Click en el nombre del archivo y se iniciará la descarga correspondiente.

Navegación de ejemplo:

01 Mercado de Corto Plazo → Liquidaciones VTEA → 2024 → 01_Enero → Revisión 03 → Resumen_cuadros-0124-R3.xlsx

Web Scraping

Para obtener datos del COES de forma más rápida, se desarrolló una aplicación en Python con el objetivo de realizar *Web Scraping* vertical ya que se centra en un solo sitio web o dominio. Ver anexo A, automatizando la descarga de los archivos, es decir, automatizar los pasos previamente descritos, logrando una mejora en la recolección de los mismos.

Los archivos descargados contienen la información de interés para la carga a la base de datos. Sin embargo, dichos archivos presentan problemas como: Datos nulos, errores de tipeo e inconsistencia de datos en nombres de empresas.

Proceso ETL completo para la valorización de energía

Este proceso consta de las siguientes etapas:

1. Descarga del archivo Excel del portal.
2. Extracción de datos de la hoja de Excel descargada.
3. Transformación de los datos extraídos (limpieza, renombrado de columnas y adición de campos).
4. Carga de los datos transformados en la base de datos.

Etapa 1

En esta etapa, es donde se aplicó *Web Scraping*, que viene a ser el método principal para la descarga de archivos Excel en total 84. A continuación, se detallan las tareas involucradas:

1. Navega a la sección “Mercado de Corto Plazo”.
2. Ingresa a “Liquidaciones VTEA”.
3. Selecciona el año y mes requeridos.
4. Realiza la navegación dinámica para identificar y descargar el archivo.
5. Espera hasta que se detecte el nuevo archivo `.xlsx` en el directorio de descargas.
6. Cierra el navegador.

Etapa 2

Extrae y procesa los datos de la hoja “CUADRO 4” de un archivo Excel, utilizando la librería pandas de python como herramienta ETL.

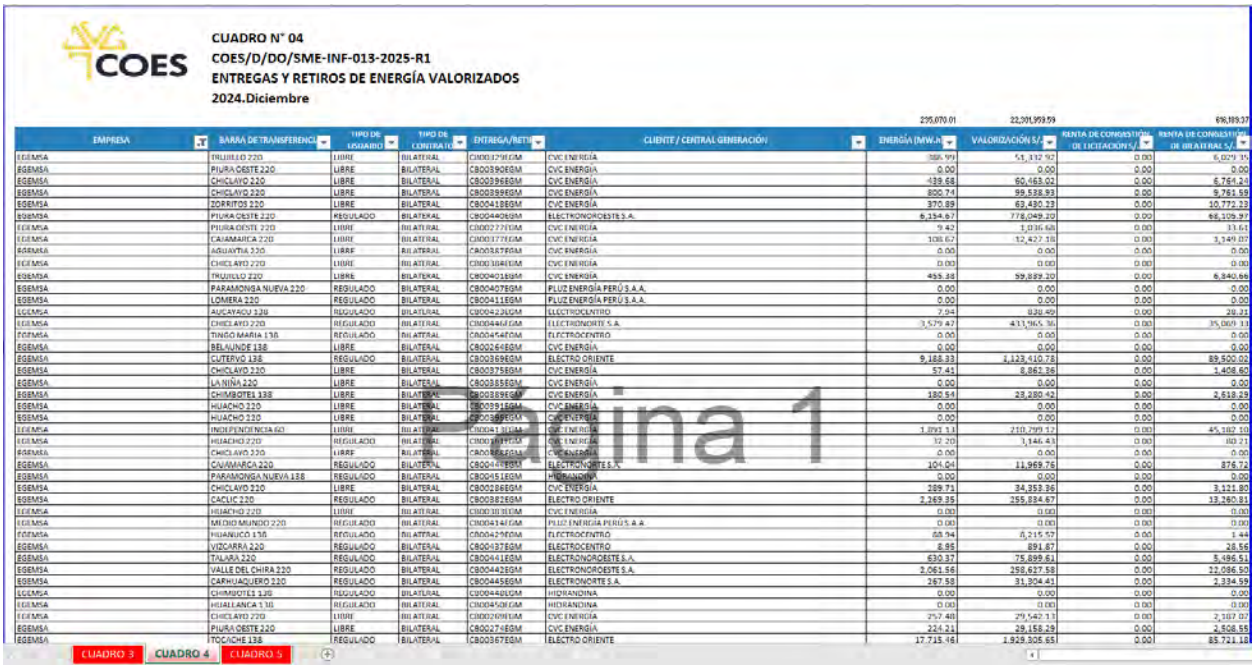

Figura 3.4: Pantallazo del archivo Excel cuadro 4

Fuente: Obtenido del portal web del COES (2025).

1. Lee el archivo Excel sin cabecera definida.

2. Delimita el área de datos de interés. Cuyas columnas se muestran en la figura 3.4.

3. Filtra las filas que tengan menos de 5 valores no nulos. Tomando éste criterio para ubicar los datos correctos para el análisis.

```python
def extract_data_from_sheet(self,path):
    data = pd.read_excel(path,sheet_name="CUADRO 4", header=None,dtype = str)
    data = self._delimit_data(data)
    # Filtrar aquellas filas que tengan por lo menos 5 valores no nulos
    data = data.dropna(thresh=5,ignore_index = True)
    return data
```

Figura 3.5: Código que muestra los 3 pasos anteriores.

Fuente: Elaboración propia (2025).

Etapa 3

Transforma los datos aplicando las siguientes operaciones:

1. Elimina espacios en blanco de cada valor de tipo cadena.

2. Renombra las columnas del cuadro 4 con nombres predefinidos: Empresa, BarraTransferencia, TipoUsuario, TipoContrato, EntregaRetiro, ClienteCentralGeneracion, EnergiaMWh, ValorizacionSoles, RentaCongestionLicitacion, RentaCongestionBilateral.

3. Agrega una columna “Periodo” con el formato “año-mes” (`yyyy-mm`).

4. Agrega una columna “FechaCreacion” con la fecha y hora actual.

Etapa 4

Carga los datos en la tabla (`bronce.ValorizacionEnergia`) de la base de datos relacional.

1. Los datos se insertan en el esquema `bronce`. Si la tabla ya existe, se añaden nuevos registros.

Librerías de Python utilizadas

- **Selenium**: nos permite navegar en el sitio web.
- **Pathlib**: se usa para identificar la ruta de los archivos.
- **Glob**: esta librería es para identificar el nombre del archivo.
- **SQLAlchemy**: nos permite la conexión a la base de datos.
- **Pandas**: se utiliza para la lectura, análisis y manipulación de datos.

Variantes de Empresas

Uno de los principales problemas al analizar la data es que los nombres de las empresas son colocados a criterio de cada **agente generador**, lo que genera nombres distintos (variantes) e incluso errores de tipeo.

Identificar variantes

Para identificar variantes, lo primero fue ordenar la data y empezar a asociar variantes por similitud de nombres:

- Podrían variar por error de tipeo de algunas letras.

Seguidamente, se usó la sintaxis `LIKE` de SQL Server para buscar por similitud de texto:

- Podrían variar cuando se agrega o no el tipo de empresa (S.A., S.A.C., S.R.L., etc.).

Los casos más difíciles se dejan para el final y se identifican analizando los archivos Excel originales, revisando mes a mes si hubo cambios de razón social y verificando quién es el generador y el cliente.

Buscar RUC asociado a la empresa

Se tuvo que navegar manualmente en la web para identificar el RUC asociado a la empresa, para los clientes de EGEMSA se realizó aproximadamente 300 busquedas para así contar con un identificador único. Esta tarea consume bastante tiempo, ya que se realiza de forma manual.

3.2.2.2. Datos obtenidos de SUNAT

Se obtuvo datos de la SUNAT (2025), a través de su servicio digital llamado "Consulta Múltiple de RUC".

Este servicio permite realizar consultas múltiples de números de RUC. Se puede consultar hasta 100 números RUC; para ello se debe grabar la lista de números RUC en un archivo de texto con extensión `.txt`, luego comprimirlo en formato `.zip` y subirlo a la plataforma. También se puede ingresar manualmente hasta 10 números de RUC en la misma página.

Luego, para ambos casos, se debe seleccionar `Enviar`; con esto se procede con la descarga de un archivo comprimido en formato `.zip`, el cual contiene un archivo de texto con extensión `.txt` que contiene los datos disponibles de la SUNAT separados por el carácter "|".

Se podrá revisar la razón social, condición de domicilio, domicilio fiscal, actividad económica, régimen tributario, entre otros gob.pe (2021).

A continuación, se describen los datos obtenidos en la consulta a SUNAT.

Tabla 3.1: *Campos y longitudes de datos obtenidos de la consulta a la SUNAT (parte 1)*

Campo	**Longitud del dato (char)**
NumeroRuc	11
Nombre ó RazonSocial	100
Tipo de Contribuyente	25
Profesión u Oficio	10
Nombre Comercial	10
Condición del Contribuyente	10
Estado del Contribuyente	10
Fecha de Inscripción	10 (dd/mm/yyyy)
Fecha de Inicio de Actividades	10 (dd/mm/yyyy)
Departamento	25
Provincia	25
Distrito	25
Dirección	120
Teléfono	25
Fax	10
Actividad de Comercio Exterior	25
Principal – CIIU	25
Secundario 1 – CIIU	25
Secundario 2 – CIIU	25

Fuente: Elaboración propia

Tabla 3.2: *Campos y longitudes de datos obtenidos de la consulta a la SUNAT (parte 2)*

Campo	**Longitud del dato (char)**
Afecto Nuevo RUS	2
Buen Contribuyente	110
Agente de Retención	110
Agente de Percepción VtaInt	110
Agente de Percepción ComLiq	110

Fuente: Elaboración propia

Posteriormente esta data es cargada a la tabla (oro.Cliente) para usarse en las siguientes etapas.

3.2.2.3. Datos de CIIU obtenidos de INEI

Se obtuvo datos de CIIU, los cuales servirán para analizar el sector económico en el que se ubican las empresas.

Los datos de CIIU se encontraron originalmente en formato `.pdf` en el libro "Clasificación Industrial Internacional Uniforme Revisión 4", publicado por el INEI (2025).

Se diseñaron cuatro tablas (*capa plata*) en la base de datos para centralizar esta información:

- `CIIU_T1_Seccion`
- `CIIU_T2_Division`
- `CIIU_T3_Grupo`

- `CIIU_T4_Clase`

Figura 3.6: Tablas de la capa plata

Fuente: Elaboración propia.

Asignación de sectores económicos

Un criterio de clasificación que se busca incorporar a los clientes es el sector económico al que pertenecen. Para ello, se utilizan los diferentes sectores establecidos por la CIIU.

Las Tablas 3.3 y 3.4 presentan los sectores acompañados de su respectivo mnemotécnico.

Tabla 3.3: Sectores económicos y sus mnemotécnicos según la CIIU (parte 1)

Nro.	**Sector económico**	**Mnemotécnico**
1	Agricultura, ganadería, silvicultura y pesca	AGRO_PESCA
2	Explotación de minas y canteras	MINERIA
3	Industrias manufactureras	MANUFACTURA
4	Suministro de electricidad, gas, vapor y aire acondicionado	ELECTRICIDAD_GAS
5	Suministro de agua; evacuación de aguas residuales, gestión de desechos y descontaminación	AGUA_DESAGUE
6	Construcción	CONSTRUCCION
7	Comercio al por mayor y al por menor; reparación de vehículos automotores y motocicletas	COMERCIO
8	Transporte y almacenamiento	TRANSPORTE
9	Actividades de alojamiento y de servicio de comidas	ALOJAMIENTO
10	Información y comunicaciones	INFORMATICA
11	Actividades financieras y de seguros	FINANCIERA
12	Actividades inmobiliarias	INMOBILIARIA
13	Actividades profesionales, científicas y técnicas	ACTI_PROFESIONALES
14	Actividades de servicios administrativos y de apoyo	SERVI_ADMINISTRATIVO

Nota. Adaptado de la Clasificación Industrial Internacional Uniforme (CIIU) Revisión 4 del INEI.

Tabla 3.4: Sectores económicos y sus mnemotécnicos según la CIIU (parte 2)

Nro.	Sector económico	Mnemotécnico
15	Administración pública y defensa; planes de seguridad social de afiliación obligatoria	ADMI_PUBLICA
16	Enseñanza	ENSEÑANZA
17	Actividades de atención de la salud humana y de asistencia social	SALUD
18	Actividades artísticas, de entretenimiento y recreativas	ARTE_ENTRETENIMIENTO
19	Otras actividades de servicios	OTROS
20	Actividades de los hogares como empleadores; actividades no diferenciadas de los hogares como productores de bienes y servicios para uso propio	ACTI_HOGARES
21	Actividades de organizaciones y órganos extraterritoriales	ACTI_ORGANIZACIONES

Nota. Adaptado de la Clasificación Industrial Internacional Uniforme (CIIU) Revisión 4 del INEI.

Configuración del Data Lake en SQL Server

Un Data Warehouse requiere definir un esquema rígido desde el inicio y cargar datos previamente estandarizados. En contraste, un Data Lake permite almacenar los datos en bruto y aplicar transformaciones progresivas por capas. Esto resulta más adecuado cuando la estructura de los datos evoluciona y se necesita conservar el dato original para asegurar la trazabilidad.

Por lo anterior, se elige un Data Lake como repositorio central, ya que el caso de estu-

dio requiere una ingesta flexible, conservación de datos en bruto, trazabilidad por capas y preparación para analítica avanzada.

Se configuró un equipo dentro de la red de EGEMSA:

- **Sistema Operativo:** Windows Server 2019 Standard
- **Memoria RAM:** 32 GB
- **Disco:** SSD 1TB
- **IP:** 10.1.21.213

Se realizó la instalación del motor de base de datos:

- **Nombre del servidor:** Microsoft SQL Server
- **Versión:** 2019
- **Nombre de la Base de Datos:** `DB_COES`
- **Modo de autenticación:** SQL Server Authentication
- **Usuario y contraseña:** Almacenados en archivo `.env`

En este servidor se recolectará toda la data necesaria para el análisis.

3.2.3. Preparación de los datos

Se organizó y transformó la información aplicando procesos de limpieza, corrección de errores y transformación de variables. Posteriormente, los datos fueron estructurados en un Data Lake utilizando la arquitectura medallón, que contempla tres capas: Bronce (datos

en bruto), Plata (datos limpios) y Oro (datos listos para análisis). Mediante un proceso ETL, los datos fueron depurados en la capa Plata y trasladados a la capa Oro para su análisis final. Para la implementación de esta arquitectura se utilizó SQL Server, definiendo esquemas, tablas y automatizando los procesos ETL con procedimientos almacenados, vistas y SQL Server Agent. Esta estructura permitió mejorar la calidad de los datos, asegurar su trazabilidad y optimizar el desempeño analítico.

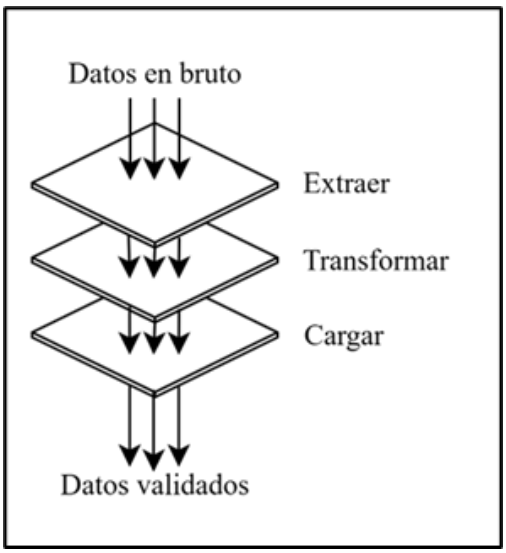

Figura 3.7: Proceso ETL

Fuente: Elaboración propia.

3.2.3.1. Diseño de la base de datos

Se implementaron los esquemas y tablas necesarios para almacenar los datos.

1. Esquemas de base de datos:

   Para la arquitectura del medallón

   - **Bronce:** Es el lugar donde se guarda temporalmente la data tal cual se cargó luego del proceso de web scraping.

En la tabla bronce.ValorizacionEnergia, los campos provenientes del webscraping fueron: Empresa, BarraTransferencia, TipoUsuario, TipoContrato, EntregaRetiro, ClienteCentralGeneracion, EnergiaMWh, ValorizacionSoles, RentaCongestionLicitacion y RentaCongestionBilateral, todos ellos son del tipo VARCHAR(250), a ellos se agregan 2 campos: Periodo de tipo VARCHAR(7) con formato yyyy-mm y FechaCreacion de tipo DATETIME.

- **Plata:** Es el lugar donde se guarda la data para su transformación y limpieza.

  Se identificó un registro erróneo en la tabla bronce.ValorizacionEnergia, en el campo Empresa cuyo valor era igual al texto EMPRESA (donde debería ingresar el nombre de una empresa generadora), por lo cual, el registro fue eliminado.

  Se aplicaron las siguientes transformaciones en la tabla bronce.ValorizacionEnergia, los campos Empresa, BarraTransferencia, TipoUsuario, TipoContrato, EntregaRetiro, ClienteCentralGeneracion fueron transformados a VARCHAR(100) mientras que los campos EnergiaMWh, ValorizacionSoles, RentaCongestionLicitacion, RentaCongestionBilateral fueron transformados a FLOAT, seguidamente en la tabla plata. ValorizacionEnergia los datos RentaCongestionLicitacion y RentaCongestionBilateral fueron transformados de NULL a ceros.

  En esta capa es donde se agregó información proveniente de la SUNAT y CIIU (obtenido de INEI) con la finalidad de enriquecer con más información el Data Lake.

- **Oro:** Es el lugar donde se almacena toda la data procesada, creada específicamente para el análisis de datos.

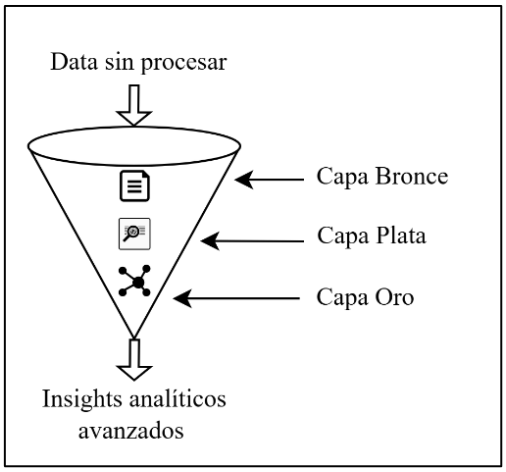

Figura 3.8: Capas de la arquitectura medallón

Fuente: Elaboración propia.

En esta capa se eliminaron los registros cuyos valores de SumaEnergiaMWh y SumaValorizacionSoles sean igual a cero, dejando la data lista para el análisis.

Para llevar control del flujo de funcionamiento

- **err:** Es el lugar donde se almacenan los errores ocurridos en procedimientos almacenados.
- **log:** Es el lugar donde se almacena el registro de los procesos ETL.

- **Modo de autenticación:** SQL Server Authentication
- **Usuario y contraseña:** Almacenados en archivo `.env`

2. Tablas:

En total se crearon 20 tablas

Tabla 3.5: Estructura de base de datos (parte 1)

Nro.	**Esquema**	**Nombre**
1	oro	BarraTransferencia
2	oro	Clientes
3	oro	Departamento
4	oro	Empresa
5	oro	Periodo
6	oro	TipoContrato
7	oro	TipoUsuario
8	oro	ValorizacionEnergia
9	plata	BarraTransferenciaVariacion
10	plata	CIIU_T1_Seccion

Nota. Elaboración propia a partir de la estructura de la base de datos.

Tabla 3.6: Estructura de base de datos (parte 2)

Nro.	Esquema	Nombre
11	plata	CIIU_T2_Division
12	plata	CIIU_T3_Grupo
13	plata	CIIU_T4_Clase
14	plata	ClientesVariacion
15	plata	EmpresaVariacion
16	plata	PrincipalCIIU
17	plata	ValorizacionEnergia
18	bronce	ValorizacionEnergia
19	err	ProcedimientoAlmacenado
20	log	CargaETL

Nota. Elaboración propia a partir de la estructura de la base de datos.

3. Procedimientos almacenados:

    a. pal_err_ProcedimientoAlmacenado_Insertar: procedimiento para insertar en la tabla err.ProcedimientoAlmacenado cualquier error ocurrido al momento de ejecutarse un procedimiento almacenado.

    b. pal_log_CargaETL_Insertar: procedimiento para insertar en la tabla log.CargaETL un registro por cada operación ejecutada en un periodo determinado.

    c. pal_ValorizacionEnergia_carga_bronce_a_plata: procedimiento para cargar la data de la tabla bronce.ValorizacionEnergia a la tabla plata.ValorizacionEnergia en un periodo determinado.

    d. pal_ValorizacionEnergia_carga_plata_a_oro: procedimiento para cargar la data de la tabla plata.ValorizacionEnergia a la tabla oro.ValorizacionEnergia en un periodo determinado.

e. pal_ValorizacionEnergia_carga_todo: procedimiento que ejecuta los dos procedimientos anteriores en un periodo determinado.

3.2.3.2. Transformación y Carga

Compuesto principalmente por *Jobs* (tareas automáticas) de SQL Server, los cuales internamente llaman a los procedimientos almacenados descritos, los cuales transforman y cargan la data del esquema `bronce` al esquema `plata` y finalmente al esquema `oro`, todo esto se realiza de manera automática.

La ejecución de los *Jobs* es totalmente configurable (fecha y horas).

Pruebas funcionales

Se realizaron pruebas del funcionamiento del ETL; para ello, se borró la data de la base de datos (BD) dejando las tablas vacías para validar nuevos registros.

- Se dejaron habilitados los *Jobs* de SQL Server.
- Se comprobó que los *Jobs* programados se ejecutaron correctamente; para eso se realizaron consultas a las tablas de BD correspondientes.
- Se validó que la data esté correctamente cargada en las tablas correspondientes comparando uno a uno con la data proveniente de los archivos Excel del COES.
- Se comprobó que la tabla `log.CargaETL` mantiene correctamente el flujo de data.
- Se provocaron errores forzadamente para validar que también funcione el registro de errores.

Diagrama de base de datos

En SQL Server se creó el diagrama “Valorización de energía” compuesto por 6 Tablas (`oro`).

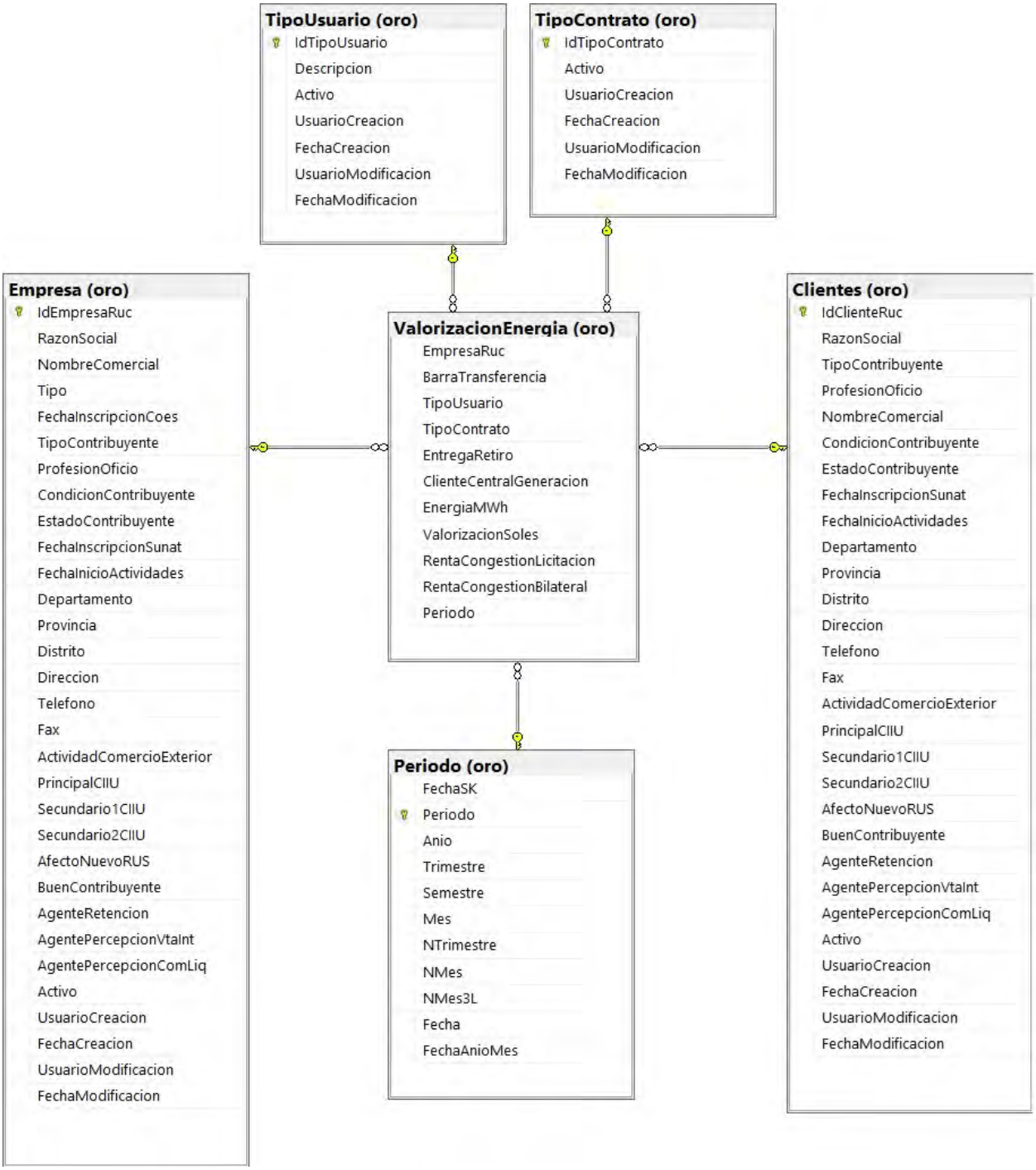

Figura 3.9: Diagrama de consumo y valorización de energía

Fuente: Elaboración propia.

Dataset de análisis

Se tiene información recolectada de **7 años (desde enero del 2018 hasta diciembre del 2024)** correspondiente al consumo mensual de energía del mercado eléctrico peruano.

Tabla Bronce: 276,997 registros, Tabla Plata: 276,997 registros, Tabla Oro: 224,929 registros

El dataset de análisis fue creado a partir de la tabla oro.ValorizacionEnergia considerando valores mayores a cero en los campos SumaEnergiaMWh y SumaValorizacionSoles. Para analizar los clientes de EGEMSA, se realizó el filtro respectivo, por lo cual, el dataset contiene **3,679 registros** y 18 campos, que se describen en las Tablas 3.7 y 3.8:

Tabla 3.7: dataset de análisis creado a partir de la tabla oro (parte 1)

Nro.	Campo	Descripción	Tipo	Valores
1	Periodo	Año y mes concatenados	Texto	Desde el 2018-01 hasta el 2024-12
2	Mes	Nombre del mes	Texto	ENERO, FEBRERO, MARZO, etc.
3	RUCEmpresa	RUC de la empresa	Texto	
4	NombreComercialEmpresa	Nombre comercial de la empresa	Texto	
5	RUCCliente	RUC del cliente	Texto	
6	RazonSocialCliente	Razón social del cliente	Texto	
7	DepartamentoCliente	Departamento del cliente según SUNAT	Texto	
8	SectorEconomicoCliente	Sector económico del cliente	Texto	Según mnemotécnico de sección del CIIU
9	TipoUsuario	Tipo de usuario	Texto	LIBRE o REGULADO
10	TipoContrato	Tipo de contrato	Texto	LICITACION o BILATERAL
11	ConsumoMensualMWh	Suma de la energía consumida	Float	En MWh

Nota. Elaboración propia a partir de la estructura de la base de datos.

Tabla 3.8: dataset de análisis creado a partir de la tabla oro (parte 2)

Nro.	Campo	Descripción	Tipo	Valores
12	SumaValorizacionSoles	Suma de la valorización de energía	Float	En soles
13	PrecioUnitario	SumaValorizacionSoles/ConsumoMensualMWh	Float	En soles/MWh
14	ConteoRetiros	Conteo de retiros por cada cliente de cada empresa según Periodo	Numérico	
15	SumaRentaCongestionLicitacion	Suma de la renta de congestión licitación	Float	En soles
16	SumaRentaCongestionBilateral	Suma de la renta de congestión bilateral	Float	En soles
17	RentaCongestionTotal	SumaRentaCongestionLicitacion + SumaRentaCongestionBilateral	Float	En soles
18	RentaCongestionUnit	(SumaRentaCongestionLicitacion + SumaRentaCongestionBilateral)/ConsumoMensualMWh	Float	En soles/MWh

Nota. Elaboración propia a partir de la estructura de la base de datos.

3.2.4. Exploración de los datos

Mediante técnicas de visualización y descriptivas, se busca comprender la distribución y comportamientos de los datos relevantes (ConsumoMensualMWh, SumaValorizacionSoles y PrecioUnitario). Se detectó relaciones iniciales y tendencias generales, así como posibles anomalías.

3.2.4.1. Estructura general del dataset

El análisis del conjunto de datos fue realizado mediante Microsoft Excel, lo que permitió una visión detallada de su estructura y características principales. El archivo contenía un total de 3 679 registros (filas) y 18 variables (columnas). El rango temporal de los datos abarcó desde enero de 2018 (2018-01) hasta diciembre de 2024 (2024-12).

Durante la exploración de valores únicos, se identificaron:

- 89 clientes distintos en la variable `RazonSocialCliente`.
- 14 departamentos correspondientes a la ubicación de los clientes.
- 11 sectores económicos representados.
- 2 tipos de usuario: LIBRE y REGULADO.
- 2 tipos de contrato: BILATERAL y LICITACIÓN.

3.2.4.2. Estadísticas Descriptivas Clave

Mediante el uso de funciones estadísticas en Excel (como PROMEDIO, MEDIANA, MIN, MAX), se calcularon los valores descriptivos de las variables cuantitativas más relevantes:

Tabla 3.9: Estadísticas descriptivas clave del dataset

Variable	**Cantidad**	**Media**	**Mediana**	**Mínimo**	**Máximo**
ConsumoMensualMWh	3,679	1,402.23	170.20	0.03	47,641.88
SumaValorizacionSoles	3,679	136,444.48	8,336.34	0.91	10,743,155.60
PrecioUnitario	3,679	67.25	35.72	7.48	825.01
ConteoRetiros	3,679	2.24	1.00	1.00	65.00
SumaRentaCongestionLicitacion	3,679	50.73	0.00	0.00	48,735.47
SumaRentaCongestionBilateral	3,679	783.83	0.79	0.00	317,139.80
RentaCongestionTotal	3,679	834.56	2.40	0.00	317,139.80
RentaCongestionUnit	3,679	0.48	0.01	0.00	30.15

Nota. Elaboración propia (2025).

Observación: en prácticamente todas las variables monetarias y de energía la media supera ampliamente la mediana, lo que indica distribuciones sesgadas con algunos valores muy elevados.

Correlaciones destacadas

Utilizando la herramienta de análisis de datos en Excel (función `PEARSON`), se evaluaron las relaciones lineales entre variables clave sugeridas por los expertos de la Genencia Comercial de EGEMSA:

- **ConsumoMensualMWh** ↔ **SumaValorizacionSoles**: 0.77 (fuerte).
- **ConsumoMensualMWh** ↔ **PrecioUnitario**: 0.12 (débil).
- **PrecioUnitario** ↔ **SumaValorizacionSoles**: 0.40 (moderada).

3.2.4.3. Tendencias temporales

A través del uso de tablas dinámicas y gráficos temporales, se identificó información importante sobre el consumo energético:

- El promedio mensual de energía total fue de aproximadamente 61,414 MWh.
- El mes con mayor consumo fue **noviembre de 2024**, con un valor cercano a 114,317 MWh.
- El mes con menor consumo fue **agosto de 2022**, con alrededor de 34,278 MWh.

Tabla 3.10: *Evolución anual de la demanda promedio mensual (MWh)*

Año	**Media mensual total (MWh)**
2018	68,072
2019	64,956
2020	52,587
2021	50,597
2022	43,814
2023	55,470
2024	94,404

Fuente: Elaboración propia

Observación: Se aprecia un descenso sostenido de 2018 hasta 2022, seguido de un repunte claro en 2023–2024.

3.2.4.4. Distribución geográfica y sectorial top 5

A través de filtros y segmentación por categorías utilizando Microsoft Excel, se determinaron las regiones y sectores económicos con mayor representación en el dataset:

Departamentos top 5

- **LIMA**: 1 957 registros (53 %)
- **ICA**: 517 registros (14 %)
- **LA LIBERTAD**: 180 registros (5 %)
- **CUSCO**: 163 registros (4 %)
- **PIURA**: 163 registros (4 %)

Sectores económicos top 5

- **MANUFACTURA**: 1 661 registros (45 %)
- **ELECTRICIDAD_GAS**: 793 registros (22 %)
- **COMERCIO**: 415 registros (11 %)
- **AGRO_PESCA**: 344 registros (9 %)
- **TRANSPORTE**: 111 registros (3 %)

3.2.4.5. Clasificación por Tipo de Usuario y Contrato

Se observaron las siguientes proporciones al analizar las categorías:

Tipo de usuario

- **LIBRE**: 3 104 registros (84 %)
- **REGULADO**: 575 registros (16 %)

Tipo de contrato

- **BILATERAL**: 3 342 registros (91 %)
- **LICITACION**: 337 registros (9 %)

3.2.4.6. Distribución de Datos por Variable

Seguidamente, se realizó el análisis de la distribución de los datos utilzando python, para ver la naturaleza de los mismos.

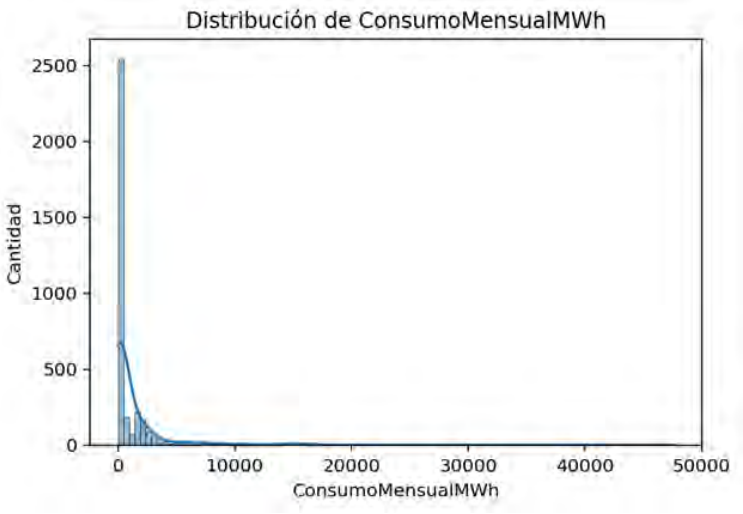

Figura 3.10: Distribución de datos del consumo mensual

Fuente: Elaboración propia.

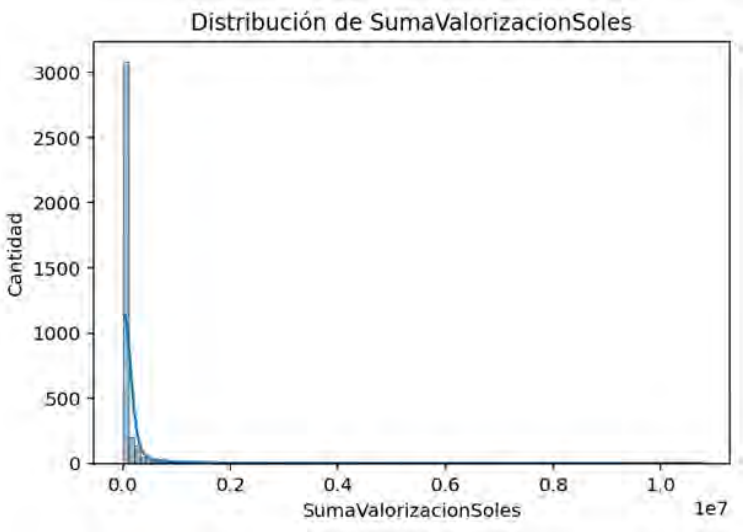

Figura 3.11: Distribución de datos de la suma de valorización

Fuente: Elaboración propia.

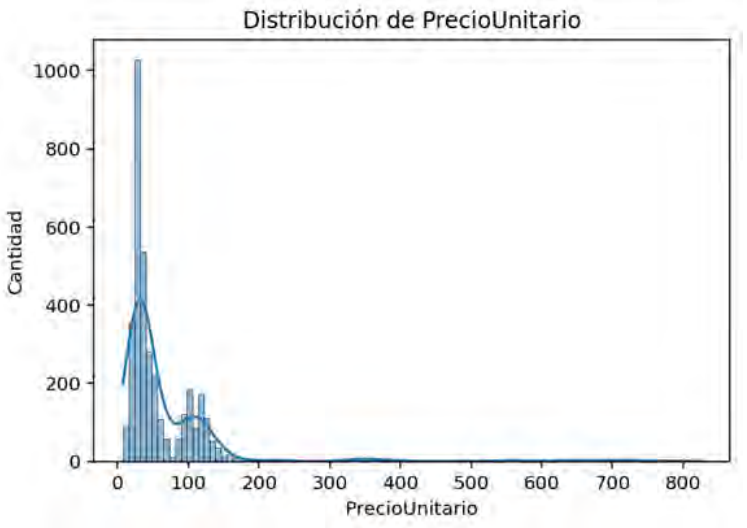

Figura 3.12: Distribución de datos del precio unitario

Fuente: Elaboración propia.

Con el fin de comprender mejor el comportamiento de las principales variables cuantitativas del dataset, se calculó un conjunto de medidas estadísticas de dispersión y forma utilizando Python, enfocándose en la desviación estándar, asimetría y curtosis. Los resultados se resumen en la siguiente tabla:

Tabla 3.11: *Medidas estadísticas de la distribución de datos*

Variable	**Desviación estándar**	**Asimetría**	**Curtosis**
ConsumoMensualMWh	3,953.95	6.60	55.63
SumaValorizacionSoles	593,505.35	9.38	111.63
PrecioUnitario	89.94	4.97	29.61

Fuente: Elaboración propia

1. **ConsumoMensualMWh**

   - La **desviación estándar** (3 953) es casi 3 veces la media (1 402), lo que indica una dispersión altísima: hay meses con consumos muy por encima y por debajo del promedio.
   - La **asimetría** de 6,60 revela una cola derecha extremadamente alargada: hay pocos meses con consumos muy altos (*outliers*) que "jalan" la distribución.
   - La **curtosis** de 55,63 ($\gg 3$) confirma colas pesadas y un pico muy pronunciado en torno a valores bajos/medios, con algunos valores extremos muy altos.

2. **SumaValorizaciónSoles**

   - La **desviación** ($\approx$ 593 505 S/) supera 4 veces la media (136 444 S/): gran volatilidad en las facturaciones mensuales.
   - La **asimetría** de 9,38 muestra aún mayor sesgo a la derecha: unos pocos meses concentran facturaciones muy elevadas.
   - La **curtosis** de 111,63 indica altísimo riesgo de *outliers*; la distribución está muy "picuda" y con colas muy gruesas.

3. **PrecioUnitario**

- La **desviación** (89,94) es mayor que la media (67,25), lo que señala que el precio unitario varía ampliamente de un registro a otro.
- Con **asimetría** de 4,97 vemos de nuevo gran sesgo a la derecha: existen precios unitarios muy elevados que distorsionan el promedio.
- La **curtosis** de 29,61 revela distribución leptocúrtica: muchas observaciones juntas en torno al centro y, a la vez, colas pesadas con valores extremos.

3.2.5. Modelación de los datos

Se utilizó Anaconda la cual proporcionó un entorno plug-and-play para la ciencia de datos al incluir Conda, un gestor de paquetes y entornos que facilitó la instalación y el aislamiento de librerías como pandas, NumPy o scikit-learn, así como Anaconda Navigator, que permitió administrar gráficamente entornos y lanzar herramientas; Spyder, distribuido con Anaconda, actuó como un IDE científico con editor de código, consola IPython interactiva, explorador de variables, visor de gráficos y depurador integrado, lo que posibilitó prototipar y depurar análisis de datos de forma ágil y reproducible.

Para comenzar con la segmentación, se cargaron los datos utilizando la librería panda, que permitió manipular y explorar los datos de forma estructurada mediante dataframes. Posteriormente, los datos se transformaron mediante un proceso de pivotado, reorganizando la estructura original en una matriz donde cada fila representaba una serie (cliente) y cada columna una característica específica (por ejemplo, el comportamiento de compra en diferentes períodos o categorías).

3.2.5.1. Estandarización de datos.

Con el fin de asegurar que cada serie (fila) tuviera una contribución equitativa al análisis, se realizó una estandarización fila a fila (row-wise). Para ello, se emplearon operaciones con NumPy, restando la media de cada fila y dividiendo por su desviación estándar correspondiente. Esta normalización evitó que diferencias de magnitud entre clientes sesgaran el proceso de agrupamiento.

Antes se tenía datos sin estandarizar como se puede apreciar en la Figura 3.13. En el cual la distribución de MWh está altamente sesgada a la derecha, con la mayoría de los valores bajos y unos pocos puntos extremos en decenas de miles.

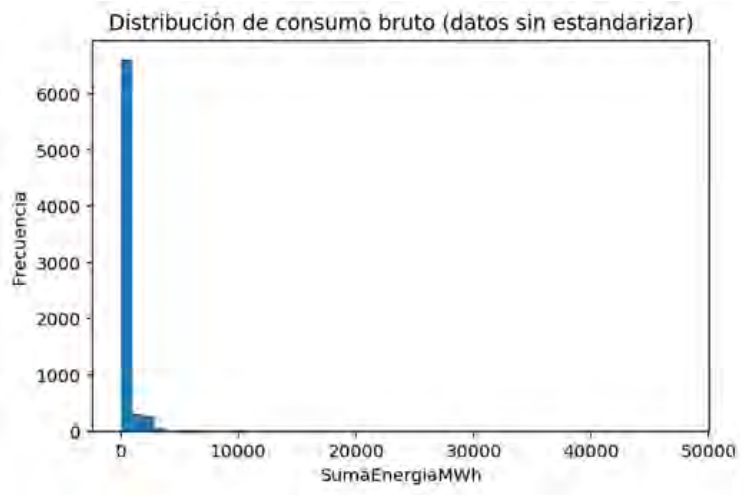

Figura 3.13: Distribución de consumo bruto

Fuente: Elaboración propia.

Después se aplica estandarización row-wise también conocido como estandarización fila a fila con el cual cada serie queda centrada en 0 y acotada entre aproximadamente –3 y +3 desviaciones estándar (con la mayoría de las observaciones en [–1, 1]), lo que uniformiza la escala de todos los clientes, facilita el clustering y evita que los consumidores de alto consumo dominen las distancias entre puntos. Ver Figura 3.14.

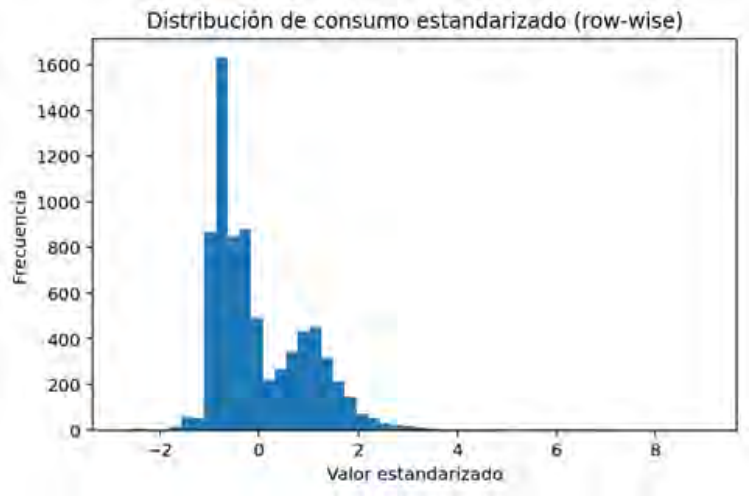

Figura 3.14: Distribución de consumo estandarizado.

Fuente: Elaboración propia.

3.2.5.2. Selección del número óptimo de clústeres

Para determinar la cantidad óptima de clústeres a utilizar, se aplicaron dos métodos complementarios:

Método del Codo (Elbow Method): Se utilizó la clase KMeans de sklearn.cluster para ajustar múltiples modelos con diferente cantidad de clústeres. Luego, se graficó la inercia (suma de errores cuadráticos intra-clúster) contra el número de clústeres mediante matplotlib.pyplot, observando el punto en el que la curva dejaba de decrecer significativamente ("codo").Ver Figura 3.15.

La inercia (qué tan juntos quedan los datos en cada grupo) baja mucho al pasar de 2 a 4 grupos, pero a partir de 4 la mejora es cada vez menor. Ahí aparece un codo que indica un buen punto de equilibrio.

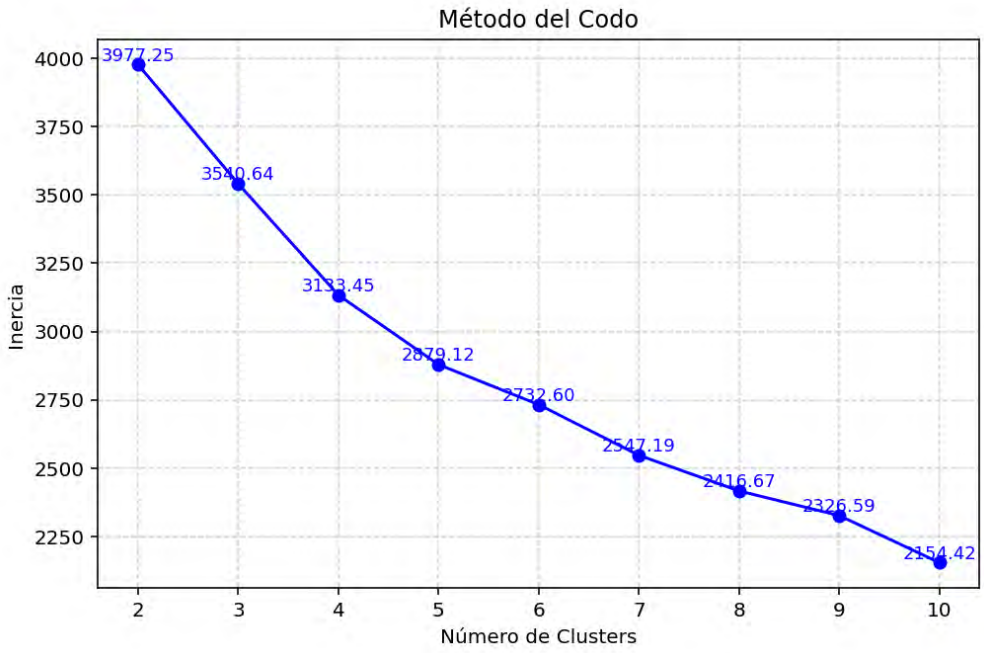

Figura 3.15: Método del codo.

Fuente: Elaboración propia (2025).

Coeficiente de Silueta: Se complementó el análisis con el cálculo del silhouette score utilizando la función silhouette_score de sklearn.metrics. Este indicador permitió evaluar la coherencia de la segmentación para cada número de clústeres, seleccionando el valor que maximizó la separación entre grupos.

Mide qué tan bien separados y definidos quedan los grupos. El valor es más alto con 2 o 3 grupos, pero sigue siendo bueno con 4–6 grupos (alrededor de 0,23). A partir de 7 baja mucho.

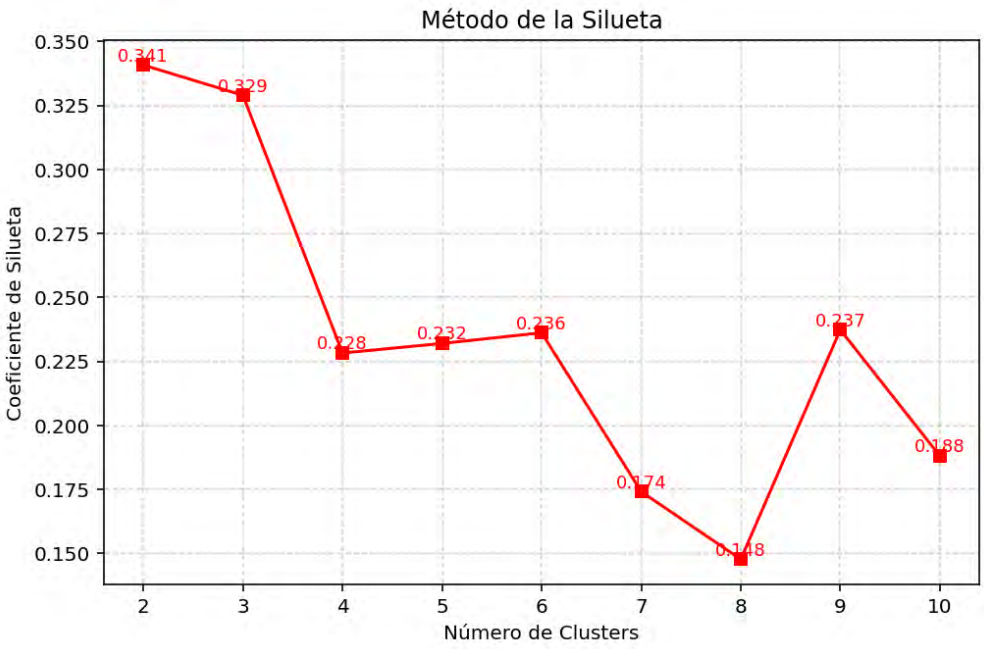

Figura 3.16: Método de la Silueta.

Fuente: Elaboración propia.

Finalmente se eligieron 4 clusters porque es el punto donde el modelo logra una buena agrupación sin volver el análisis innecesariamente complicado. Con el método del codo se observa que al pasar de 2 a 4 grupos la mejora es clara (los datos quedan mucho mejor ordenados), pero después de 4 la mejora ya es pequeña y no justifica agregar más grupos. Además, el índice de silueta muestra que entre 4 y 6 la separación entre grupos se mantiene aceptable, pero cuando se usan 7 o más grupos la calidad baja bastante. Por eso se tomó k=4, ya que permite diferenciar comportamientos de consumo de manera práctica y fácil de interpretar para tomar decisiones comerciales.

Tabla 3.12: Resultados de los métodos del Codo y Silueta

K (Número de cluster)	Inercia	Coeficiente de Silueta
2	3977.25	0.341
3	3540.64	0.329
4	3133.45	0.228
5	2879.12	0.232
6	2732.60	0.236
7	2547.19	0.174
8	2416.67	0.148
9	2326.59	0.237
10	2154.42	0.188

Fuente: Elaboración propia

3.2.5.3. Segmentación de clientes por patrones de consumo con k-means

Una vez determinado el número óptimo de clústeres, se procedió a aplicar el algoritmo K-Means para realizar la segmentación definitiva. Este algoritmo fue implementado con la clase KMeans de scikit-learn, configurando el número de clústeres previamente seleccionado. El modelo se ajustó a los datos estandarizados, y se asignó una etiqueta de clúster a cada cliente.

Se realizó una segmentación de clientes en 4 clusters basados en el perfil de consumo estandarizado (cada serie centrada en 0 y escalada). Los resultados clave son los siguientes. Ver Figura 3.17.

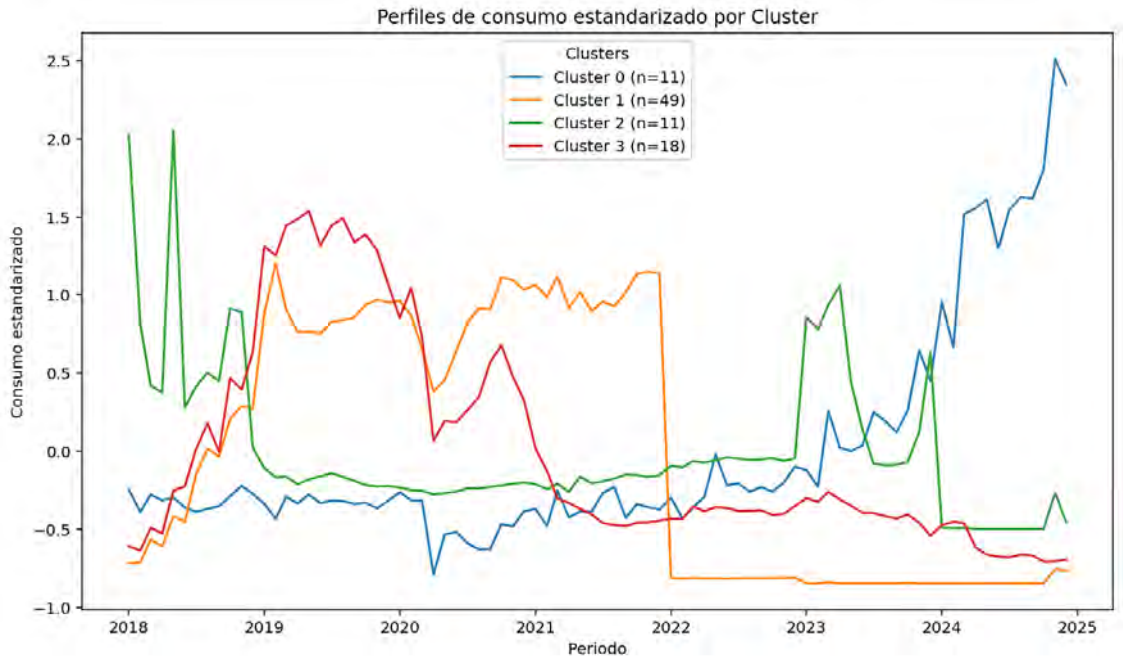

Figura 3.17: Perfiles de consumo estandarizado por cluster.

Fuente: Elaboración propia.

Cluster 0: clientes cuyo consumo aumenta de forma notable en 2023–2024, con un repunte muy marcado al final de la serie. Ver Figuras 3.18 y 3.19.

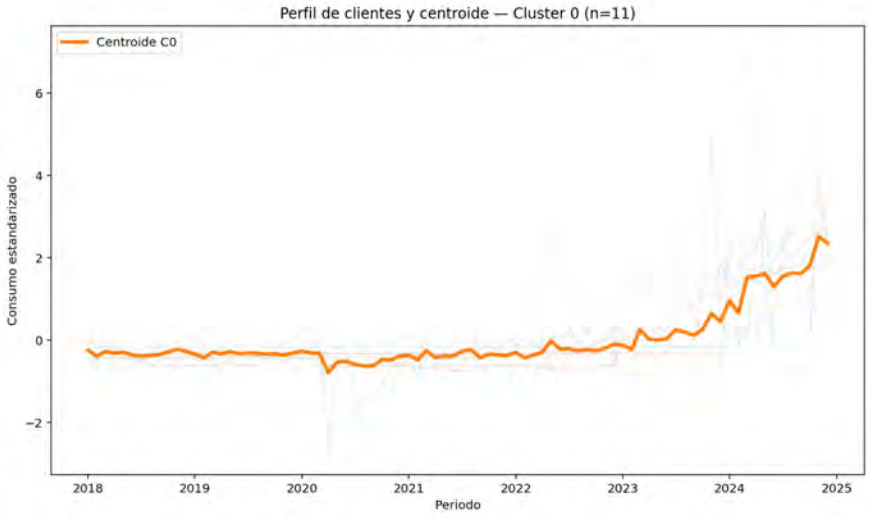

Figura 3.18: Perfil de clientes y centroide del Clúster 0.

Fuente: Elaboración propia.

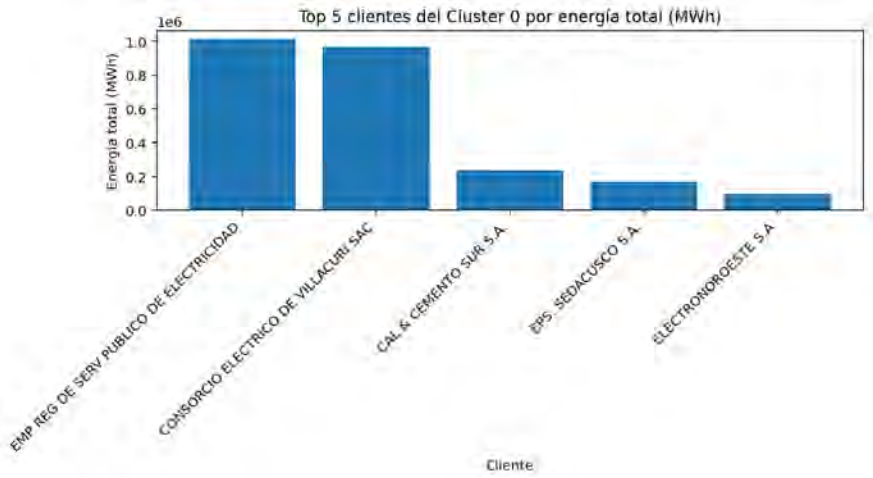

Figura 3.19: Top 5 clientes por energía total del Clúster 0.

Fuente: Elaboración propia.

Cluster 1: perfiles con crecimiento gradual hasta 2021, que experimentan una fuerte caída a comienzos de 2022 y se mantienen en niveles muy bajos y estables. Ver Figuras 3.20 y 3.21.

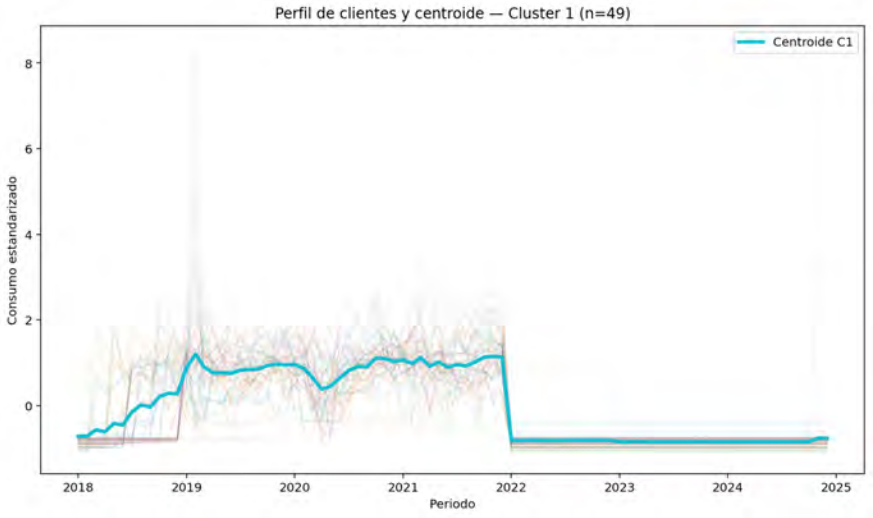

Figura 3.20: Perfil de clientes y centroide del Clúster 1.

Fuente: Elaboración propia.

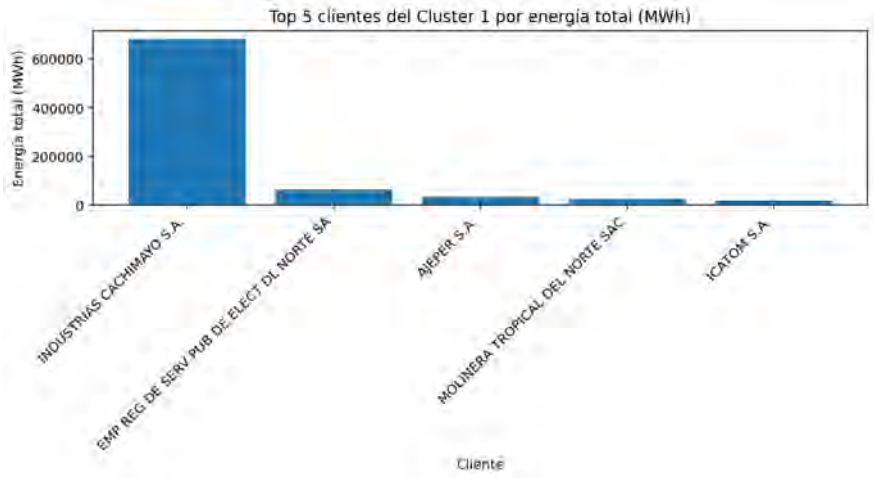

Figura 3.21: Top 5 clientes por energía total del Clúster 1.

Fuente: Elaboración propia.

Cluster 2: usuarios con patrones erráticos, caracterizados por picos aislados (por ejemplo, en 2018 y 2023) y periodos de meseta intercalados. Ver Figuras 3.22 y 3.23.

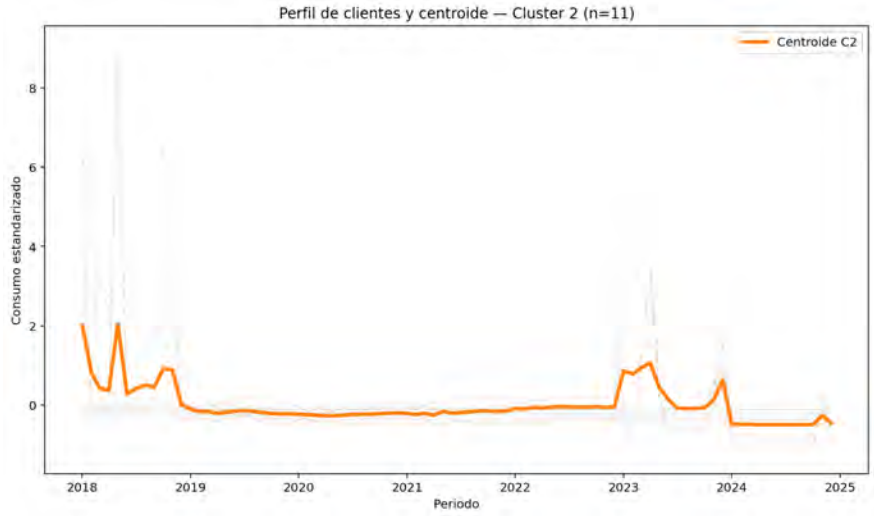

Figura 3.22: Perfil de clientes y centroide del Clúster 2.

Fuente: Elaboración propia.

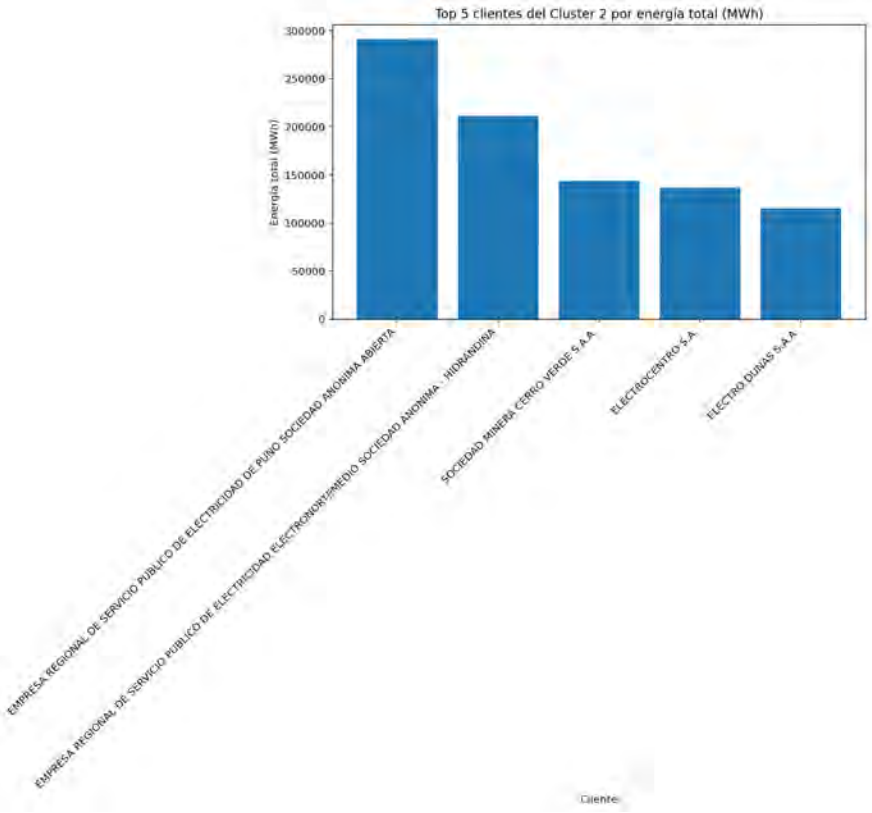

Figura 3.23: Top 5 clientes por energía total del Clúster 2.

Fuente: Elaboración propia.

Cluster 3: consumidores con alta demanda durante 2019–2020, seguida de un descenso sostenido a partir de 2021. Ver Figuras 3.24 y 3.25.

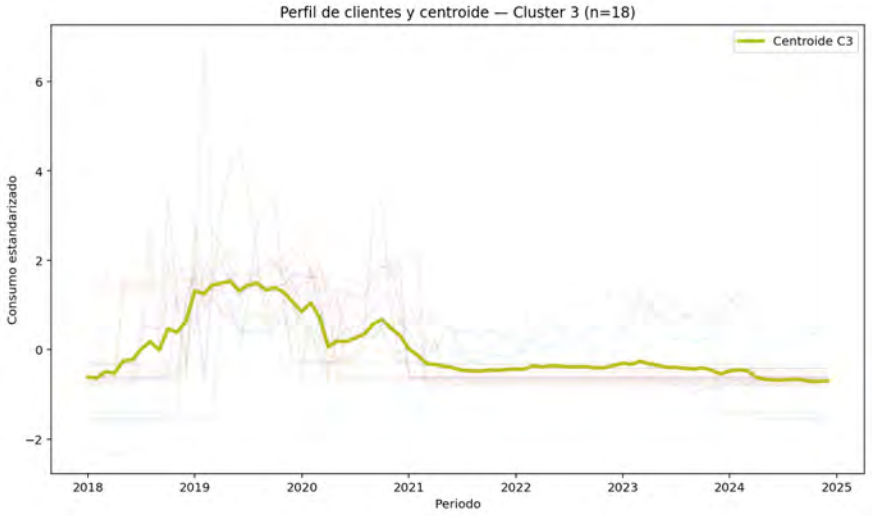

Figura 3.24: Perfil de clientes y centroide del Clúster 3.

Fuente: Elaboración propia.

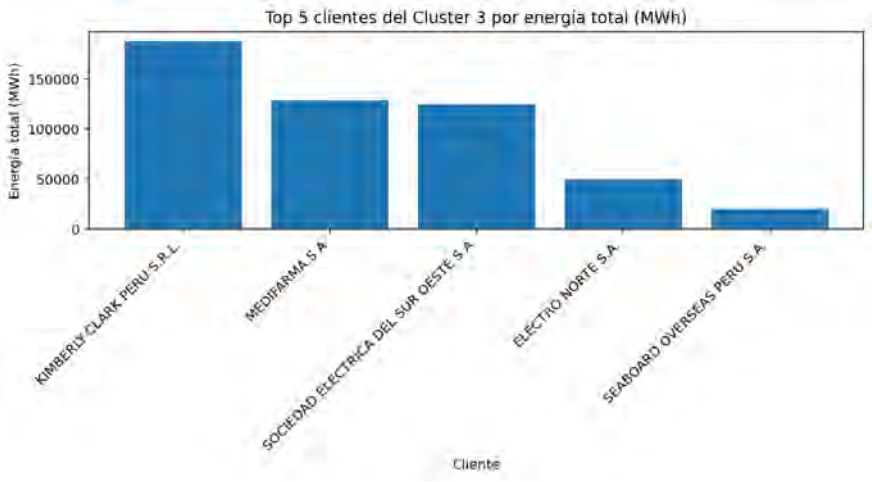

Figura 3.25: Top 5 clientes por energía total del Clúster 3.

Fuente: Elaboración propia.

Con el fin de visualizar los resultados del clustering, se redujo la dimensionalidad de los datos a dos componentes principales mediante Análisis de Componentes Principales (PCA), utilizando la clase PCA de sklearn.decomposition. Esta transformación permitió representar gráficamente los grupos en un plano bidimensional, preservando la mayor cantidad de varianza posible.

El mapa de calor que se aprecia en la Figura 3.26 muestra las cargas de cada periodo (variable) sobre los dos primeros componentes principales:

PC1:

- Cargas crecientes a lo largo del tiempo, pasando de valores bajos en 2018–2020 a valores altos en 2023–2024.
- Interpreta la tendencia general del consumo: períodos recientes, con mayor demanda agregada, contribuyen positivamente a PC1.
- Clientes con valores altos en PC1 tienden a tener un perfil de consumo más elevado en 2023–2024 comparado con su media histórica.

PC2:

- Cargas positivas moderadas en los períodos medios (2018–2020) y negativas hacia los extremos (final de 2024).
- Refleja un contraste entre consumo en periodos intermedios versus periodos finales.
- Clientes con PC2 positivo tienen relativamente más consumo en 2019–2020; con PC2 negativo, destacan en consumo muy reciente (finales de 2024).

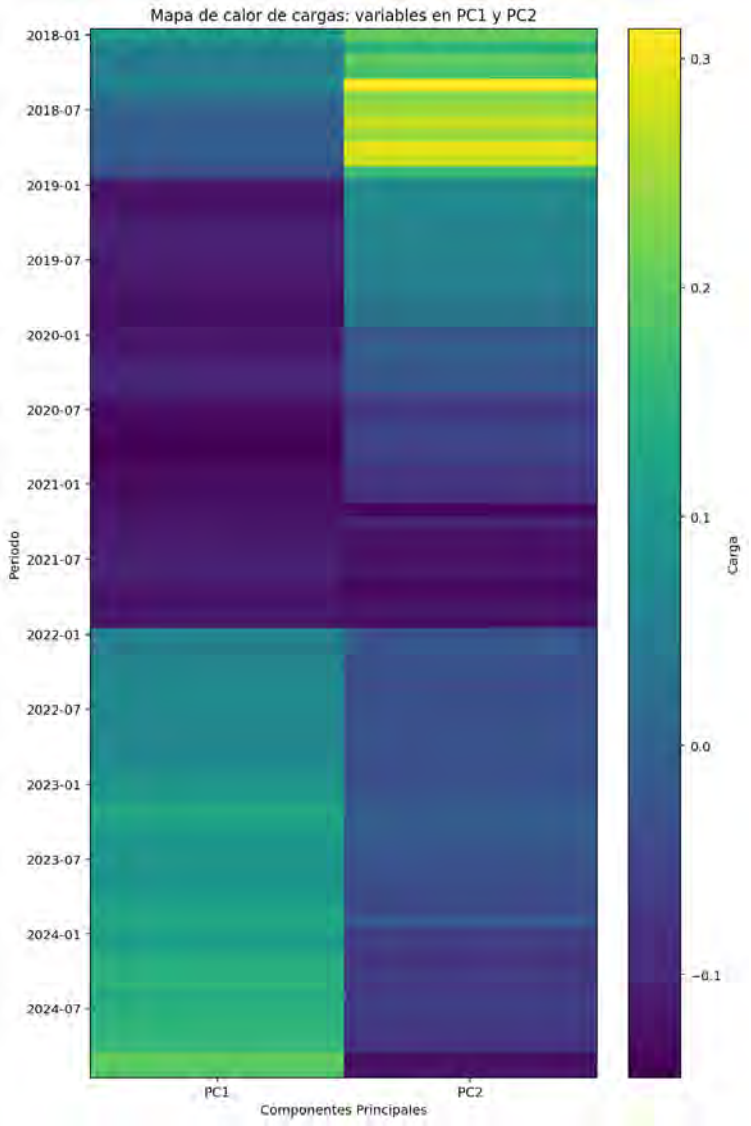

Figura 3.26: Mapa de calor de cargas variables en PC1 y PC2.

Fuente: Elaboración propia.

Para visualizar los clusters se creó un gráfico de dispersión en el espacio de las dos primeras componentes principales PC1 y PC2, con cada cliente coloreado según su cluster y los centroides destacados con una X grande y contorno negro. Esto te permite visualizar claramente la separación y la posición de los centroides respecto a los puntos de los clientes.

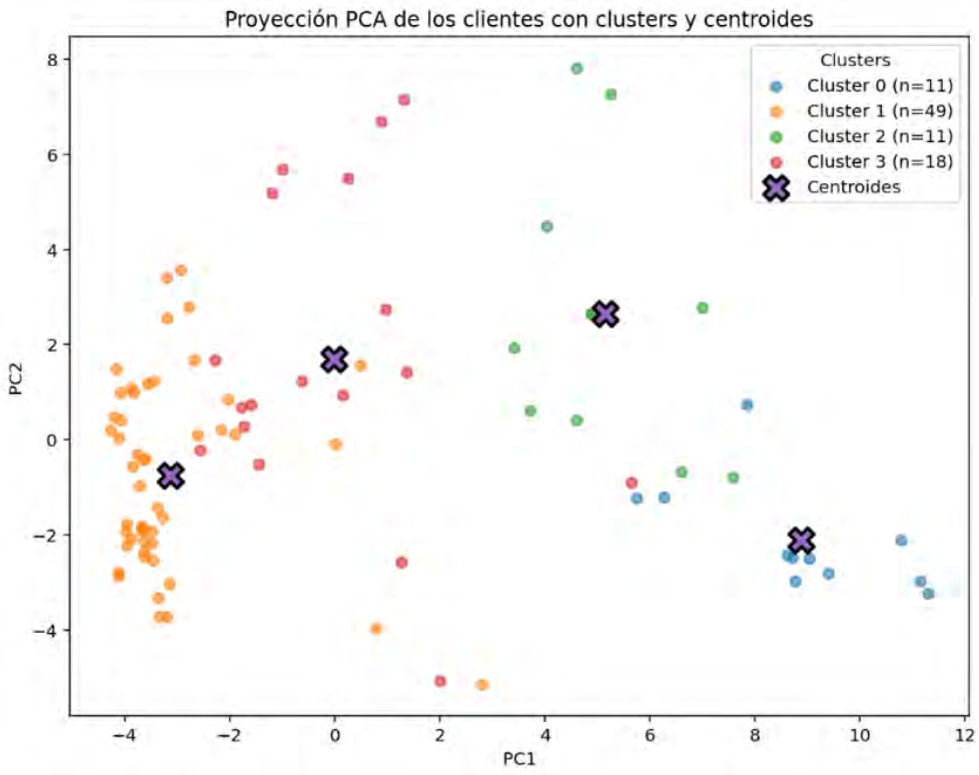

Figura 3.27: Proyección PCA de los clientes con sus cluster y centroides

Fuente: Elaboración propia.

Esta segmentación permite diferenciar claramente grupos con comportamientos de consumo homogéneos.

Tabla 3.13: *Número de clientes por clúster*

Cluster	**Número de clientes**
0	11
1	49
2	11
3	18

Fuente: Elaboración propia

3.2.5.4. Segmentación de clientes por consumo y sector económico.

Se procede con asociar el sector económico a cada cluster para entender qué tipo de clientes caen en cada patrón.

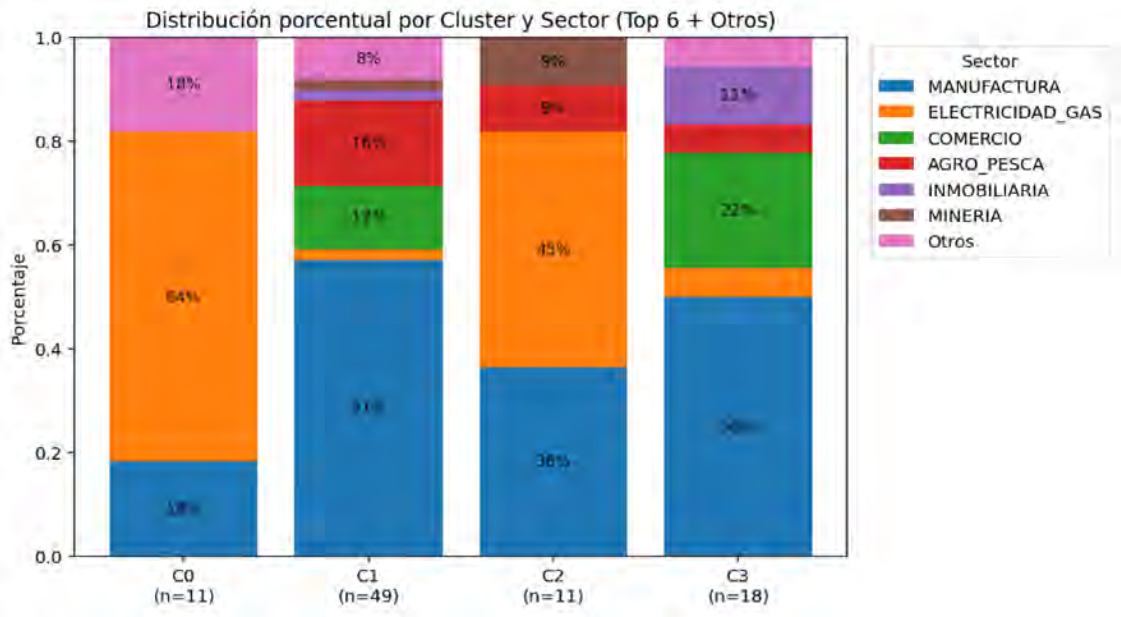

Figura 3.28: Distribución porcentual por cluster y sector

Fuente: Elaboración propia.

Sector Económico

- **Cluster 0:** Muy dominado por **ELECTRICIDAD_GAS** (64 %), resto repartido equitativamente entre **MANUFACTURA** (18 %) y **Otros** (18 %), prácticamente sin presencia de **COMERCIO**, **AGRO_PESCA** o **INMOBILIARIA**, sectores con picos en 2023–2024.
- **Cluster 1:** Predomina la **MANUFACTURA** (57 %), sectores secundarios: **AGRO_PESCA** (16 %), **COMERCIO** (12 %), minoritarios: **Otros** (8 %), **INMOBILIARIA** (2 %) y **ELECTRICIDAD_GAS** (2 %), reflejando un declive fuerte

de la industria en 2022.

- **Cluster 2:** Combinación de **ELECTRICIDAD_GAS** (45 %) y **MANUFACTURA** (36 %), resto equilibrado entre **AGRO_PESCA** (9 %) y **Otros** (9 %), muy escasa o nula presencia de **COMERCIO** e **INMOBILIARIA**, con consumos altos en 2018 y 2023.

- **Cluster 3:** Mayoritario **MANUFACTURA** (50 %), segundo **COMERCIO** (22 %), sectores menores: **INMOBILIARIA** (11 %), **ELECTRICIDAD_GAS** (6 %), **AGRO_PESCA** (6 %) y **Otros** (6 %), alta demanda durante 2019–2020, descenso sostenido a partir de 2021.

3.2.5.5. Segmentación de clientes por consumo y departamento.

Se procede con asociar el departamento a cada cluster para entender qué tipo de clientes caen en cada patrón.

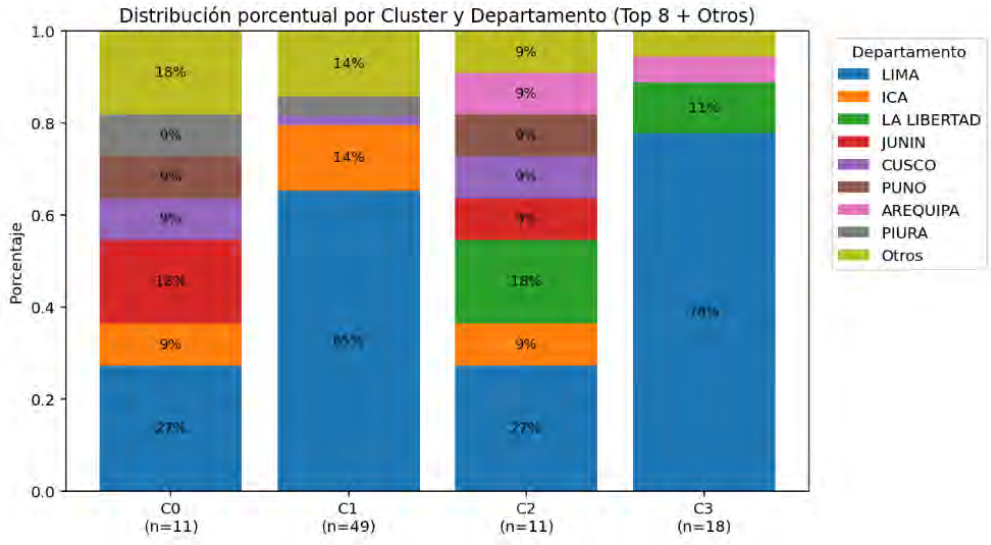

Figura 3.29: Distribución porcentual por cluster y departamento

Fuente: Elaboración propia.

Distribución Geográfica

- **Cluster 0:** Muy disperso ("Otros" 18 %), con aportes moderados de **LIMA** (27 %), **JUNIN** (18 %), **CUSCO** (9 %) e **ICA** (9 %), hasta 2021 por debajo de la media, pero desde mediados de 2023 un fuerte crecimiento sostenido.
- **Cluster 1:** Fuertemente concentrado en **LIMA** (65 %), con un segundo grupo en "Otros" (14 %) e **ICA** (14 %), subida pronunciada hasta 2021 seguida de un colapso en 2022 y estancamiento bajo desde entonces.
- **Cluster 2:** Equilibrado entre **LIMA** (27 %), "Otros" (9 %) y **LA LIBERTAD** (18 %), más pequeñas cuotas de **JUNIN** (9 %) y **CUSCO** (9 %), muy volátil en 2018, luego estable por debajo de la media, con un pico puntual en 2023.
- **Cluster 3:** Dominado por el departamento de **LIMA** (78 %), con el resto en "Otros" (5.5 %) y **LA LIBERTAD** (11 %), pico temprano (2019), descenso progresivo desde entonces hasta estabilizarse muy por debajo de la media.

3.2.6. Presentación y automatización

- **Aplicación de Web Scraping:** Se detalla el script desarrollado en Python que automatiza la descarga de archivos Excel del portal del COES, reduciendo los tiempos y errores de la recolección manual.
- **Implementación del Data Lake:** Se almacenó los datos brutos provenientes del COES en la capa Bronce. Luego, se describe el flujo de limpieza, transformación y conformación de tablas estructuradas en la capa Plata, finalizando con la capa Oro donde los datos están listos para el análisis.

- **Aplicación del algoritmo K-Means:** Se utilizó el método del codo y el índice de Silhouette para determinar la cantidad óptima de grupos, lo cual llevó a identificar patrones de comportamiento similares del consumo de energía entre distintos periodos.

- Finalmente, los hallazgos se integran en un informe final, facilitando la toma de decisiones en la Gerencia Comercial. Además, garantizando que las técnicas de Big Data e IA implementadas se conviertan en un recurso operativo y estratégico dentro de EGEMSA.

Capítulo 4

Resultados y Discusión

El análisis de los patrones de consumo eléctrico de los clientes de EGEMSA en el periodo 2018–2024, realizado mediante técnicas de Big Data e Inteligencia Artificial (IA), permitió segmentar a la cartera en cuatro clusters con características diferenciadas. La clasificación se obtuvo a través de clustering y fue validada por el Jefe del Departamento de Comercialización de EGEMSA, quien corroboró que los grupos reflejan realidades contractuales y operativas observadas en la gestión comercial.

Los resultados no solo ofrecen una descripción del comportamiento de los clientes, sino que también constituyen un insumo para la toma de decisiones estratégicas de la Gerencia Comercial, al identificar oportunidades de crecimiento, riesgos de fuga y dinámicas de mercado sectoriales.

La mejora de la oferta comercial requiere información objetiva del comportamiento del mercado y de las transacciones reales. En el contexto del SEIN, el COES provee datos públicos de las transacciones de energía (inyecciones, retiros y valorizaciones) que describen cómo se valorizan y liquidan las transferencias entre participantes, información indispensable para analizar tendencias, perfilar clientes y sustentar condiciones contractuales competitivas. En un mercado donde los usuarios libres pueden negociar y elegir al proveedor con mejores con-

diciones, disponer de datos del COES organizados y analizados permite identificar patrones y oportunidades de negocio de manera ágil y confiable, fortaleciendo la posición comercial de EGEMSA.

4.1. Resultados generales del clustering

El análisis mostró que los clusters difieren tanto en número de clientes como en el volumen de consumo que representan.

Tabla 4.1: *Segmentación de clientes de EGEMSA mediante clustering (2018–2024)*

Cluster	**Denominación asignada**	**Clientes**	**Consumo total (MWh)**	**Consumo mensual promedio (MWh)**	**Participación (%)**
0	Expansión sostenida	11	1,927,597	22,948	43.2
1	Reducción abrupta	49	1,121,445	13,351	25.1
2	Volatilidad contractual	11	920,901	10,963	20.6
3	Declive sostenido	18	492,788	5,867	11.0

Fuente: Elaboración propia a partir de resultados del clustering de clientes de EGEMSA

Se observa que:

- El **Cluster 0** (*Expansión sostenida*), con apenas 11 clientes, muestra casi la mitad del consumo total (43 %).

- El **Cluster 1** (*Reducción abrupta*) concentra la mayor cantidad de clientes (49), pero con menor peso relativo en el consumo.
- El **Cluster 2** (*Volatilidad contractual*) presenta consumos elevados pero inestables.
- El **Cluster 3** (*Declive sostenido*) refleja pérdidas graduales de consumo por no renovación de contratos o coyunturas externas.

4.2. Validación con el experto (Jefe del Departamento de Comercialización)

La clasificación fue revisada y validada por el Jefe del Departamento de Comercialización de EGEMSA, quien aportó explicaciones cualitativas que complementan el análisis cuantitativo:

- **Cluster 0 – Expansión sostenida:**

  Se verifica consumo creciente de los que ya eran clientes:

  - El cliente *EMP REG DE SERV PUBLICO DE ELECTRICIDAD (ELECTRO ORIENTE)* tuvo un crecimiento fuerte, ya que en el último año (2024) se firmó un nuevo contrato de 90 MWh, anteriormente se tenía un contrato promedio de alrededor de 8 hasta 10 MWh, lo que vendría a ser un crecimiento sustancial de 900 %.
  - El cliente *EPS. SEDACUSCO S.A.* debido a la escasez de agua de los últimos periodos, realizó más bombeos de agua, por ende, el consumo eléctrico de las máquinas encargadas de dicho bombeo fue en crecimiento.
  - El cliente *CIRION TECHNOLOGIES PERU S.A.* es una empresa encargada de administrar servidores de datos, brindando soporte a otras compañías, entre ellas

empresas dedicadas a las redes sociales; por lo que mientras más demanda de redes sociales, mayor consumo eléctrico de los servidores.

- *CONSORCIO ELECTRICO DE VILLACURI SAC* es un cliente distribuidor de energía eléctrica que atiende a la zona agroindustrial de ICA; en los últimos años la exportación de productos agropecuarios ha ido en crecimiento.
- En general, los clientes que son empresas distribuidoras de energía eléctrica tienen lo que se llama un crecimiento vegetativo, que es lo normal en estos clientes (la población crece, por ende, su consumo eléctrico aumenta).

Se verifica incremento de nuevos clientes:

- *ENEL DISTRIBUCION PERU S.A.A.* es un nuevo cliente.
- *COMPANIA PESQUERA DEL PACIFICO CENTRO SA*: en general, en macroeconomía, el rubro de pesquería ha ido en crecimiento.

- **Cluster 1 – Reducción abrupta:**

  Se verifica culminación de contrato de respaldo que abarcaba varios clientes:

  - *ATRIA* es un comercializador de energía eléctrica (se dedica a la compra y venta, ya que su generación es muy pequeña), atiende a una bolsa de clientes pequeños (entre 50 a 70 clientes).
  - Como cliente de EGEMSA, *ATRIA* atendía a su vez a un gran número de clientes pequeños; por lo que, cuando terminó su contrato en 2022, se refleja una caída fuerte del consumo eléctrico.

- **Cluster 2 – Volatilidad contractual:**

  Se verifican contratos de corto plazo (menos de 1 año):

  - *SOCIEDAD MINERA CERRO VERDE S.A.A.*: en el año 2023, una vez terminado su contrato, ya no es cliente.

- **Cluster 3 – Declive sostenido:**

  Se verifica culminación de contratos y ausencia de renovación:

  - *MOLINERA SUDAMERICA S.A.C.* no renovó su contrato.
  - *PLASTICOS BASICOS DE EXPORTACION S.A.C.* no renovó su contrato.
  - *MIRAGE HOLDING SOCIEDAD ANONIMA CERRADA* no renovó su contrato.

  Por la pandemia del COVID-19 hubo bajo consumo, ya que los clientes fueron fuertemente afectados:

  - *LIMA GOLF CLUB*, al dedicarse al rubro del entretenimiento, perdió afiliaciones o membresías, lo que impactó directamente en el consumo eléctrico.

  Bajo consumo a causa de la veda en la pesca:

  - *SEABOARD OVERSEAS PERU S.A.* redujo su consumo eléctrico en ciertos periodos.

- **Sobre los gráficos de barras:**

  - **Sector económico:** el porcentaje refleja realmente la venta de energía principalmente en dos sectores: Manufactura y Electricidad; los demás sectores se reparten en menor medida.
  - **Departamento:** el porcentaje refleja que en Lima ha ido creciendo el consumo de las industrias.

La validación confirma que los *clusters* son coherentes con la dinámica real de la cartera, reforzando la utilidad práctica del análisis.

4.3. Justificación de la denominación de los clusters

La denominación asignada a cada cluster busca representar de forma precisa los patrones de consumo identificados:

- **Cluster 0 – Expansión sostenida:** Clientes con crecimiento estructural y continuo, asociados a sectores en expansión (agroindustria, distribución eléctrica, digitalización).
- **Cluster 1 – Reducción abrupta:** Clientes que presentan caídas súbitas en su consumo, explicadas principalmente por el término de contratos de respaldo.
- **Cluster 2 – Volatilidad contractual:** Consumos inestables vinculados a contratos de corta duración (< 1 año), con ingresos puntuales pero no sostenibles.
- **Cluster 3 – Declive sostenido:** Reducción progresiva del consumo por no renovación de contratos y choques externos (COVID-19, vedas pesqueras).

Esta nomenclatura facilita la comunicación interna y orienta la definición de estrategias diferenciadas por grupo.

4.4. Discusión de hallazgos

1. **Alta concentración en pocos clientes:** El 43 % del consumo total depende de solo 11 clientes del Cluster 0, lo que representa un riesgo de concentración, pero también una oportunidad de consolidar relaciones estratégicas.
2. **Volatilidad y caídas bruscas:** Los Clusters 1 y 2 demuestran que la finalización de contratos puede provocar variaciones súbitas en el consumo, afectando la estabilidad de ingresos.

3. **Clientes en riesgo:** El Cluster 3 agrupa clientes con descensos sostenidos, que representan riesgos de pérdida definitiva si no se implementan estrategias de recuperación.

4. **Factores externos sectoriales:** La escasez hídrica, la digitalización de servicios, la agroindustria, la pandemia y las restricciones ambientales (vedas) influyen directamente en el comportamiento del consumo eléctrico.

5. **Dimensión sectorial y geográfica:** La mayor parte del consumo se concentra en los sectores Manufactura y Electricidad, y geográficamente en Lima y regiones agroexportadoras como Ica.

4.5. Implicancias estratégicas para la Gerencia Comercial

- **Cluster 0 Expansión sostenida:** Priorizar acuerdos de largo plazo, programas de fidelización y acompañamiento estratégico.
- **Cluster 1 Reducción abrupta:** Reducir la dependencia de comercializadores, establecer alertas tempranas de vencimiento contractual.
- **Cluster 2 Volatilidad contractual:** Incorporar contratos cortos como "capas tácticas", aplicando primas por flexibilidad y mecanismos de retención.
- **Cluster 3 Declive sostenido:** Implementar estrategias de recuperación selectiva y prevención de fuga mediante analítica predictiva.

4.6. Comparativa antes y después

Sin Big Data/IA: Procesos manuales y subjetivos, carentes de segmentación sólida.

Con Big Data/IA: Análisis automatizado, segmentaciones claras y un repositorio centralizado que soporta consultas rápidas.

Tabla 4.2: *Comparación de Procesos con y sin Big Data/IA*

Procesos	**Sin Big Data/IA**	**Con Big Data/IA**
Recolección de datos	Descarga manual desde el portal del (COES); múltiples archivos Excel dispersos. Tiempo típico: 6 horas por actualización.	Ingesta automática (Web Scraping). Tiempo: 20 minutos (no supervisado).
Consolidación de datos	Copia/pega y uniones manuales; versiones duplicadas; riesgos de error humano. Tiempo: 8 h por ciclo.	Data Lake (SQL Server) + ETL, con jobs programados; control de calidad (esquema medallón y validaciones). Tiempo: 1h; trazabilidad.
Análisis de datos	Tablas dinámicas, descubrimientos limitados a la experiencia del analista; difícil repetir; segmentación débil. Tiempo: 5 días para un informe.	ML/IA para segmentación (K-means) y patrones de consumo; gráficos actualizados. Tiempo: 2h para visualización y evaluación de resultados.

Fuente: Elaboración propia

El análisis muestra que la adopción de un Data Lake y la aplicación de ML resultaron en mejoras sustanciales en la velocidad de procesamiento, la precisión de hallazgos y la capacidad de respuesta de la Gerencia Comercial frente a las fluctuaciones del mercado eléctrico.

Conclusiones

1. Las conclusiones en base al objetivo general son:

   Los resultados evidencian una **reducción sustancial en tiempos operativos**, una **mejora en la calidad y confiabilidad de los datos**, y una **evolución significativa en la capacidad de análisis y segmentación**, consolidando el impacto positivo de la adopción de **Big Data** e **Inteligencia Artificial** en la Gerencia Comercial de EGEMSA.

2. Las conclusiones en base al primer objetivo específico son:

   Recolección de datos: Se pasó de descargas manuales de aproximadamente 6 horas y múltiples archivos Excel dispersos a una ingesta automática mediante *Web Scraping* de 20 minutos sin supervisión, logrando un proceso más eficiente y confiable.

3. Las conclusiones en base al segundo objetivo específico son:

   Consolidación y organización: Mediante un Data Lake en SQL Server y procesos ETL automatizados, se reemplazaron tareas manuales (copiar, pegar y unir datos) que demandaban hasta 8 horas y eran propensas a errores, por una arquitectura tipo medallón (Bronce, Plata y Oro) con controles de calidad, reduciendo la consolidación a 1 hora, con trazabilidad completa y sin duplicados.

4. Las conclusiones en base al tercer objetivo específico son:

   Análisis avanzado: La integración de técnicas de IA y *Machine Learning* (*K-Means*) permitió automatizar la segmentación y detección de patrones de consumo, reducien-

do el tiempo de análisis de 5 días a 2 horas, con visualizaciones y actualización en tiempo real, fortaleciendo la identificación de oportunidades y la toma de decisiones estratégicas.

Recomendaciones

Se recomienda poner más énfasis en la limpieza de los datos, este es el paso más importante en el análisis de datos: dado que errores, valores atípicos, registros duplicados o inconsistencias en la identificación de los clientes pueden distorsionar la formación de los clústeres y conducir a conclusiones equivocadas para la Gerencia Comercial.

Se recomienda aplicar técnicas automáticas para la identificación del nombre de los clientes para luego asignar su RUC correspondiente, ya que hacerlo manualmente consume bastante tiempo.

Se recomienda incluir datos adicionales como Contratos, Factor de carga, demanda coincidente, etc. para el análisis y aplicar más algoritmos de Machine Learning para tener una perspectiva más amplia de los clientes.

Se sugiere la implementación de Apache Airflow u otra plataforma de orquestación para programar tareas de Web Scraping y ETL de forma periódica. Esto permitirá mantener actualizados los clústeres, asegurando que EGEMSA disponga de información reciente al momento de planificar estrategias comerciales.

Se recomienda desarrollar una API (RESTful o GraphQL) que reciba los datos de un cliente en tiempo real y retorne el clúster al que pertenece. De esta manera, el área comercial podrá personalizar ofertas, pronósticos o estrategias de venta basadas en la clasificación automática.

Referencias bibliográficas

A. Soto, Paúl; R. Castro, J. M. R. R. D. C. T. (2023). Partición de una red eléctrica de distribución aplicando algoritmos de agrupamiento k-means y dbscan. *Revista Latinoamericana de Computación*. Disponible en: https://doi.org/10.37116/revistaenergia.v20.n1.2023.572.

Alvarez Rubio, Andrea Mercedes; García Juarez, H. D. S. C. V. P. I. J. M. (2024). *Inteligencia Artificial y ciencia de datos en metodología de la investigación científica*. Editorial Académica, Lima, Perú.

Anaconda, Inc. (2025). *Anaconda Distribution*. Disponible en: https://www.anaconda.com/.

Banco Central de Reserva del Perú (2023). Proyecciones económicas y de demanda eléctrica en el perú. Disponible en: https://www.bcrp.gob.pe/docs/Publicaciones/Reporte-Inflacion/2023/diciembre/reporte-de-inflacion-diciembre-2023-recuadro-4.pdf.

Chester, D. and Maecker, H. (2015). *Hierarchical Clustering Methods for Data Analysis*. Springer.

Chino Espinoza, Hebert; Lavilla Alvarez, V. (2019). Aplicación de técnicas de minería de datos para identificación de patrones de comportamiento de las variables de proceso de

generación y distribución de energía eléctrica, para la empresa egemsa. Disponible en: https://repositorio.unsaac.edu.pe/handle/20.500.12918/4179.

Cielen, D., Meysman, A., and Ali, M. (2016). *Introducing Data Science: Big Data, Machine Learning, and more, using Python tools*. Manning Publications, Shelter Island, NY. Disponible en: https://www.manning.com/books/introducing-data-science.

COES (2025). Comite de operaciones económica del sis sistema nacional interconectado. Disponible en: https://www.coes.org.pe/Portal/mercadomayorista/liquidaciones.

Curo Martínez, J. (2022). Pronóstico de la demanda eléctrica diaria del sein para reducir la demanda coincidente de un usuario libre del sector industrial. Disponible en: https://repositorio.uncp.edu.pe/bitstream/handle/20.500.12894/8492/T010_47298964_T%202_removed.pdf?sequence=1.

Curto, J. (2016). Introducción al business intelligence. Citado en (Joyanes Aguilar, 2019).

Dammert Lira, Alfredo; Molinelli Aristondo, F. C. N. M. A. (2011). *Fundamentos técnicos y económicos del sector eléctrico*. Disponible en: https://cdn.www.gob.pe/uploads/document/file/607310/Libro_Fundamentos_Tecnicos_Economicos_Sector_Electrico_Peruano.pdf?v=1587651119.

de Mántaras, R. L. (2017). *Inteligencia Artificial*. Editorial UOC, Barcelona, España.

EGEMSA (2022). Memoria-egemsa-2022. Disponible en: https://www.egemsa.com.pe/publicaciones/memoria-egemsa-2022.

EGEMSA (2024). Manual de organización y funciones. Disponible en: https://transparencia.egemsa.com.pe/static/archivos/bMOF_2024.pdf.

EGEMSA (2025). Información corporativa de la empresa de generación eléctrica machupicchu s.a. Disponible en: https://web.egemsa.com.pe/quienes-somos.

Fang, H. (2015). *Managing Data Lakes in Big Data Era.* Disponible en: https://ieeexplore.ieee.org/document/7288049.

Figueroa Gallardo, L. B. (2021). Web scraping, visualización y análisis de bases de datos de la operación del sistema eléctrico chileno. Disponible en: https://repositorio.uchile.cl/handle/2250/181581.

García Fernández, L. B. (2021). Modelamiento del pronóstico de la demanda eléctrica diaria del sistema eléctrico interconectado nacional utilizando técnicas de machine learning. Disponible en: https://repositorio.uni.edu.pe/handle/20.500.14076/21832.

gob.pe (2021). Consultar el estado de hasta 100 números de ruc. Disponible en: https://www.gob.pe/13397-consultar-el-estado-de-hasta-100-numeros-de-ruc.

gob.pe (2023). Ley de concesiones eléctricas y reglamento. Disponible en: https://www.gob.pe/institucion/minem/informes-publicaciones/4768546-ley-de-concesiones-electricas-y-reglamento.

Guiraldes Deck, D. A. (2020). Segmentación y caracterización de clientes libres del sistema eléctrico nacional para modelar demanda flexible. Disponible en: https://repositorio.uchile.cl/bitstream/handle/2250/173933/cf-guiraldes_dd.pdf?sequence=1&isAllowed=y.

INEI (2025). Clasificación industrial internacional uniforme revisión 4. Disponible en: https://www.inei.gob.pe/media/MenuRecursivo/publicaciones_digitales/Est/Lib0883/Libro.pdf.

Joyanes Aguilar, L. (2019). *Inteligencia de negocios y analítica de datos.* Editorial RA-MA, Madrid, España.

Lasse Petteri, R. (2018). *Inteligencia artificial 101 cosas que debes saber hoy sobre nuestro*

futuro. Disponible en: https://proassetspdlcom.cdnstatics2.com/usuaris/libros_contenido/arxius/40/39308_Inteligencia_artificial.pdf.

Marr, B. (2016). Big data. la utilización del big data, el análisis y los parámetros smart para tomar mejores decisiones y aumentarăelărendimiento. Citado en (Joyanes Aguilar, 2019).

McCarthy, J. (2007). What is artificial intelligence? Disponible en: http://jmc.stanford.edu/articles/whatisai/whatisai.pdf.

Microsoft (2024). learn.microsoft.com arquitectura medallon. Disponible en: https://learn.microsoft.com/es-es/azure/databricks/lakehouse/medallion.

Microsoft (2025). *Microsoft Documentation*. Disponible en: https://learn.microsoft.com/.

Microsoft Corporation (2025). *Visual Studio Code: Code editing. Redefined*. Disponible en: https://code.visualstudio.com/.

Nargesian, Fatemeh; Zhu, E. . J. M. R. (2019). Data lake management: Challenges and opportunities. In *Nombre de la conferencia o revista*. Disponible en: https://www.vldb.org/pvldb/vol12/p1986-nargesian.pdf.

Palma, B. (2022). El mercado eléctrico peruano, un breve repaso de su estructura. Disponible en: https://fri.com.pe/blog/content/el-mercado-electrico-peruano-un-breve-repaso-de-su-estructura.

Pandas (2025). *Pandas: Python Data Analysis Library*. Disponible en: https://pandas.pydata.org/.

Pedregosa, Fabian; Varoquaux, G. G. A. (2011). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12:2825–2830.

Python Software Foundation (2025). *Python: A Programming Language*. Disponible en: https://www.python.org/.

Ramírez, Y. (2022). Uso de algoritmos de aprendizaje automático para analizar datos de energía eléctrica facturada. caso: Chile 2015 - 2021. Disponible en: https://revistas.utp.ac.pa/index.php/id-tecnologico/article/view/3678/4305.

Rodríguez, P. (2023). *Métodos de Clustering y sus Aplicaciones*. Editorial Académica.

scikit-learn (2025). Clustering. https://scikit-learn.org/stable/modules/clustering.html. Accedido: 10 de agosto de 2025.

Sociedad Nacional de Mineria Petroleo y Energía (2024). Legislación del sector eléctrico. Disponible en: https://snmpe.org.pe/energia/legislacion.html.

Spyder Project Contributors (2025). *Spyder: The Scientific Python Development Environment*. Disponible en: https://www.spyder-ide.org/.

SUNAT (2025). Consulta múltiple de ruc. Disponible en: https://e-consultaruc.sunat.gob.pe/cl-ti-itmrconsmulruc/jrmS00Alias.

Yajure Ramírez, C. A. (2022). Uso de algoritmos de machine learning para analizar los datos de energía eléctrica facturada en la ciudad de buenos aires durante el período 2010 - 2021. *Ciencia, Ingenierías y Aplicaciones*, 5(2):7–37.

Anexos

Se presentan documentos y datos adicionales relevantes para la investigación, los cuales complementan el análisis principal y permiten una mejor comprensión de la metodología aplicada.

- **Anexo A: Código fuente del script de Web Scraping.** Contiene el procedimiento utilizado para la recolección automatizada de datos desde fuentes externas. Disponible en GitHub: https://github.com/sticonab/egemsa-web-scraping Ver Anexo A.
- **Anexo B: Scripts de base de datos y procedimientos almacenados en SQL Server.** Reúne las rutinas implementadas en la base de datos para la gestión, transformación y consulta de datos. Disponible en GitHub: https://github.com/sticonab/egemsa-data-lake Ver Anexo B.
- **Anexo C: Código de análisis K-means.** Incluye el algoritmo aplicado para la segmentación de clientes, así como las configuraciones empleadas en el modelado. Disponible en GitHub: https://github.com/sticonab/egemsa-machine-learning Ver Anexo C.

- **Anexo D: Cronograma de Actividades.**

Figura 4.1: Diagrama de Gantt

Fuente: Elaboración propia

- **Anexo E: Validación con el especialista del departamento de comercialización de EGEMSA.**

CONSTANCIA DE VALIDACIÓN DE DATOS Y RESULTADOS DE TESIS

La **Empresa de Generación Eléctrica Machupicchu S.A. - EGEMSA** con RUC N° 20218339167, hace constar por la presente que los tesistas:

- **Grover Moreano Briceño** – DNI N° 43612546
- **Saul Waldemar Ticona Bejar** – DNI N° 47405354

han desarrollado, en coordinación con nuestra institución, el trabajo de investigación titulado:

"APLICACIÓN DE TÉCNICAS DE BIG DATA E INTELIGENCIA ARTIFICIAL PARA MEJORAR LA CAPACIDAD ANALÍTICA DE EGEMSA"

El proyecto se llevó a cabo en las instalaciones de la empresa durante el periodo febrero/2025 – agosto/2025, cumpliendo con las etapas de planificación, ejecución y cierre del estudio.

Los tesistas realizaron las siguientes **actividades principales**:

- Diagnóstico de la situación actual de la empresa.
- Levantamiento, procesamiento y análisis de información interna.
- Propuesta y validación de un modelo/metodología de mejora aplicable a los procesos de la organización.
- Elaboración de informe final de resultados obtenidos.
- Presentación de conclusiones y recomendaciones prácticas para la toma de decisiones.

Los **profesionales de la empresa** que acompañaron el desarrollo de la investigación manifiestan que:

- Las actividades de los tesistas fueron ejecutadas con responsabilidad, ética y rigurosidad técnica.
- La investigación aportó un análisis objetivo de la problemática, generando evidencia útil para la gestión empresarial.
- Las propuestas y recomendaciones planteadas son viables, pertinentes y alineadas a la estrategia de la empresa.
- El trabajo de los tesistas constituye un aporte valioso que fortalece las capacidades técnicas y la toma de decisiones de la organización.

En mérito a lo expuesto, se expide la presente constancia a solicitud de los interesados, para los fines que estimen convenientes.

Cusco, 29 de diciembre de 2025.

Ing. John Pável Triveño Ramos
Jefe del Departamento de Comercialización
egemsa