Reto Digitalización 3

Santiago Estrada Bernal

INFORME FINAL

CELSIA S.A E.S.P.

Universidad EIA
Envigado, Antioquia
Invention Studio
2025

Contenido

1. Descripción del Problema

Las fallas en redes eléctricas representan uno de los principales desafíos operativos para las empresas del sector energético, ya que pueden generar interrupciones del servicio, pérdidas económicas, daños en la infraestructura y afectaciones significativas para los usuarios. En sistemas eléctricos modernos, los eventos como caídas de tensión, sobrecargas, fallas monofásicas o trifásicas, e incluso comportamientos anómalos asociados a pérdidas no técnicas, pueden evolucionar rápidamente si no se identifican de manera oportuna. Por esta razón, la detección temprana de fallas se ha convertido en un componente crítico para garantizar la continuidad del servicio, mantener la estabilidad del sistema y optimizar la toma de decisiones operativas. Empresas como CELSIA S.A. E.S.P., responsables de la operación y distribución de energía en diversas regiones del país, requieren mecanismos confiables y basados en datos que permitan monitorear el estado de su red eléctrica y anticipar condiciones anómalas que puedan afectar la calidad del suministro.

En este contexto, CELSIA planteó la necesidad de conocer el estado real de su red para actuar de manera preventiva o correctiva cuando sea necesario. Sin embargo, una de las principales restricciones del problema radica en la disponibilidad limitada de información: los datos accesibles para este proyecto corresponden únicamente al consumo energético expresado en kWh, sin variables eléctricas más detalladas como voltaje, corriente o calidad de la energía, que tradicionalmente se emplean para la detección de fallas. Esta limitación implica el reto de identificar patrones asociados a fallos o comportamientos irregulares utilizando exclusivamente datos de consumo, los cuales tienen baja resolución temporal y no capturan directamente el comportamiento eléctrico instantáneo de la red. En consecuencia, el proyecto se centra en explorar enfoques de análisis de datos y técnicas de agrupamiento que permitan extraer información relevante del consumo diario, con el objetivo de inferir estados anómalos de la red y ofrecer a CELSIA una herramienta inicial para la supervisión no intrusiva de su infraestructura.

2. Metodología Implementada

El desarrollo del proyecto se llevó a cabo siguiendo un enfoque estructurado dividido en dos fases principales. En primer lugar, se realizó una revisión exhaustiva de literatura académica, documentos técnicos y casos de uso reportados por empresas del sector energético. Este análisis permitió comprender cómo se aborda actualmente la detección de fallas en redes eléctricas y cuáles son las variables comúnmente empleadas en estos sistemas. El hallazgo más relevante fue que la identificación precisa de fallos requiere típicamente información como voltaje RMS, corriente RMS, frecuencia y potencia, lo que evidenció desde el inicio la dificultad inherente al presente estudio, dado que los datos disponibles se limitan únicamente al consumo energético (kWh). Esta limitación metodológica condicionó el enfoque posterior, reforzando la importancia de explorar técnicas que permitieran extraer valor a partir de un conjunto de datos reducido en complejidad.

En la segunda fase, se procedió a aplicar técnicas de análisis de datos y modelos de agrupamiento utilizando la información suministrada por CELSIA. Para ello se adoptó la metodología ASUM-DM (Analytics Solutions Unified Method for Data Mining), un marco metodológico que guía proyectos analíticos mediante etapas iterativas que incluyen la comprensión del negocio, la exploración y preparación de los datos, la construcción y evaluación de modelos, y por último su implementación o interpretación. Esta metodología permitió mantener un proceso ordenado, documentado y orientado a los requisitos del problema. Una vez entrenados los modelos candidatos y seleccionada la opción que mejor se ajustaba a las restricciones del proyecto, se analizaron sus métricas y comportamiento de agrupamiento con el fin de obtener conclusiones sobre su capacidad para detectar actividades o consumos inusuales dentro de la red eléctrica.

3. Tratamiento de Datos

A continuación, se expondrá el procedimiento que se llevó a cabo para entender, filtrar y transformar los datos originales proveídos por CELSIA para ser usados en el modelo de agrupamiento.

3.1. Conversión de Tipo de Dato

El archivo csv original estaba compuesto por tres columnas: ID, FECHA, KWH_IMPORT. La primera permite identificar a qué medidor pertenece el registro, la segunda, cuándo se tomó el dato y la tercera, el consumo registrado. Dada la naturaleza de cada variable, fue imperativo asegurar que FECHA fuese tipo *datetime* y KWH_IMPORT, *float*.

3.2. Entendimiento de KWH_IMPORT

La variable de mayor interés, KWH_IMPORT, tiene un comportamiento incremental con posibilidad de saturación. Esto significa que cada registro será mayor al anterior, salvo que se llegue al punto de saturación y se reinicie. Para asegurarse que los valores usados en el futuro modelo tuviesen sentido, se organizaron los datos de cada medidor según su fecha y se verificó crecimiento positivo, permitiendo concluir que ninguno de los 401 medidores llegó al punto de saturación ni tuvo un registro erróneo.

3.3. Consumo Neto

Se añadió una columna de consumo neto que permita determinar la cantidad de energía real consumida durante un período de tiempo de un día. Esta se obtuvo de la resta entre registros adyacentes de KWH_IMPORT. Evidentemente, el primer registró de cada medidor no tiene un consumo neto asociado por la falta de valores pasados. Por esta razón, fueron eliminados. Adicionalmente, se verificó que no existiesen consumos negativos y se eliminaron los registros de medidores que registraron un valor de KWH_IMPORT constante (consumo neto 0)., resultando en el retiro de ocho medidores.

3.4. Entendimiento de FECHA

Los registros corresponden a las mediciones tomadas durante un año. Consecuentemente, el año en la variable fecha es igual para todas las filas del dataset. Se

decidió desagregar los valores de la columna FECHA en DÍA, MES y AÑO, ignorando la última por su falta de valor agregado. Además, se tuvo en cuenta la naturaleza cíclica de las variables temporales restantes, transformando las variables unidimensionales en bidimensionales con funciones sinusoidales que permitan expresar numéricamente la cercanía equivalente entre, por ejemplo, el día 7 del 1 y el día 7 del 6. Así, se obtienen las columnas DÍA_sin, DÍA_cos, MES_sin y MES_cos con el comportamiento expuesto en la Figura 1.

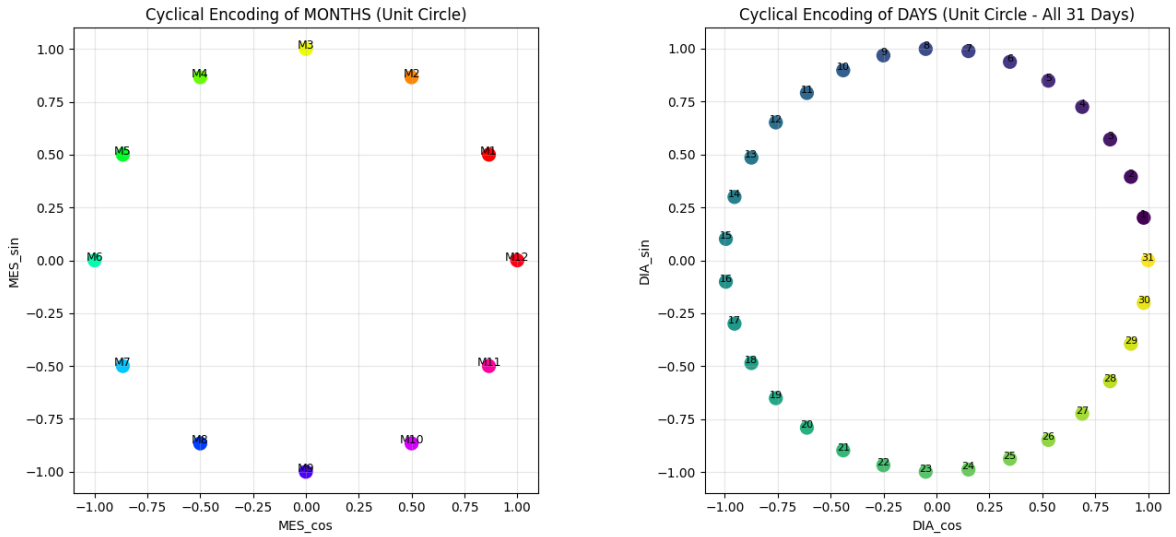

Figura 1. Variables sinusoidales DÍA y MES

3.5. Variables de Contexto

Con el fin de proveer al modelo con un contexto de cada registro, se añadieron columnas adicionales con información relevante de cada medidor. De esta forma aparecen METER_AVG_CONSUMO, METER_STD_CONSUMO y CONSUMO_DEVIATION, que representan el consumo neto promedio del medidor, la desviación estándar del medidor y la desviación del registro con respecto al promedio del medidor. Así, los registros pertenecientes a un mismo medidor tendrán el mismo valor para las primeras dos variables, pero cada registro de CONSUMO_DEVIATION es diferente.

3.6. Exploración Univariada

Una vez se establecieron todas las variables numéricas y categóricas, se comenzó la exploración univariada para determinar cuáles eran realmente relevantes. Los histogramas de las variables numéricas confirmaron que el AÑO es el mismo para todos los registros.

Además, las variables FECHA, DÍA y MES ya fueron tenidas en cuenta en otras columnas. El ID del medidor no es relevante para la identificación de fallas en la red eléctrica (sería relevante si el objetivo fuese hallar medidores defectuosos).

3.7. Exploración Multivariada

Se analizó el factor de colinealidad de las variables numéricas entre sí para evitar redundancias. Esto permitió hallar que la variable de contexto METER_AVG_CONSUMO es innecesaria, pues su información es capturada en CONSUMO_DEVIATION y presenta colinealidad con CONSUMO_NETO.

3.8. Preparación de los Datos

Es esencial transformar todos los datos a tipos interpretables por máquina y/o de bajo costo computacional. Dado que no se cuentan con variables categóricas, basta con hacer una transformación de numérico a numérico para facilitar el manejo de recursos dado el dataset. Los histogramas del análisis univariado permitieron visualizar que ninguna de las variables presentaba un comportamiento normal. Por consiguiente, la transformación se llevó a cabo con el escalador MinMaxScaler que toma cada característica y la lleva a un rango de 0 a 1. El resultado final es un dataset curado con columnas CONSUMO_NETO, MES_sin, MES_cos, DIA_sin, DIA_cos, METER_STD_CONSUMO y CONSUMO_DEVIATION y 140628 filas totales.

4. Construcción de Modelos

En miras de asegurar el mejor resultado posible, la metodología ASUM-DM propone el diseño y entrenamiento de varios modelos que satisfagan la necesidad impuesta para luego ser comparados. Por esta razón, se presentan a continuación todos los modelos entrenados junto con una breve explicación de los hiperparámetros seleccionados para cada uno.

4.1. K Means

El modelo de K Medias fue diseñado a partir del gráfico del codo que permite observar la inercia de un modelo como muestra la Figura 2. La estrategia consiste en buscar el punto de inflexión en el que la inercia cambia de concavidad para obtener un comportamiento asintótico.

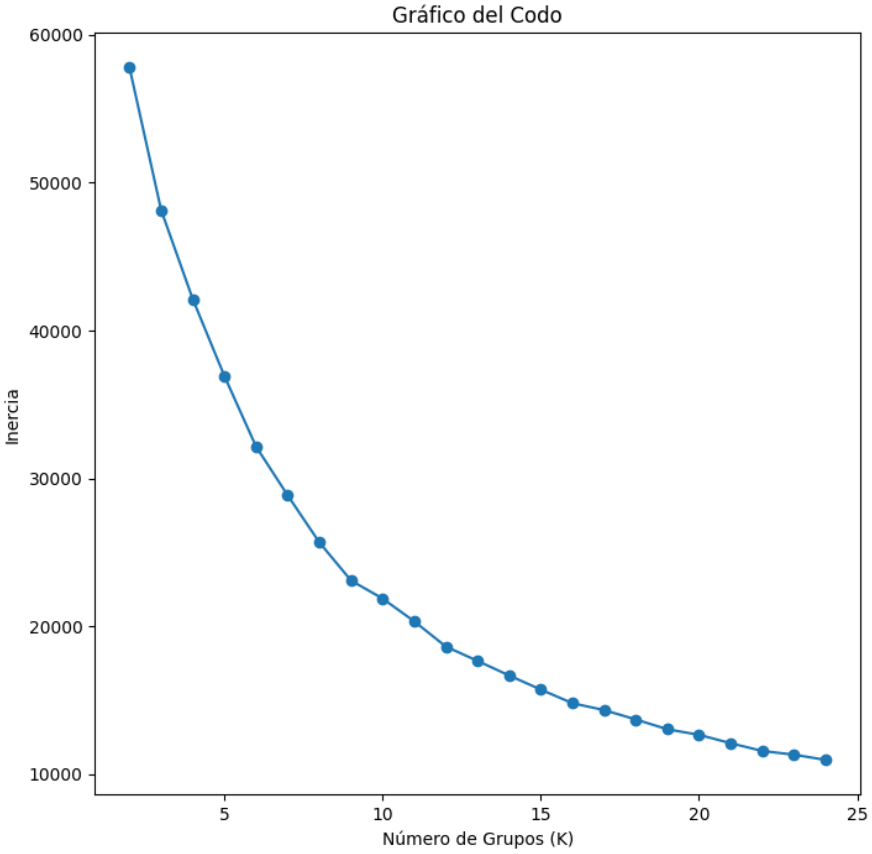

Figura 2. Gráfico del codo K Means

El intervalo en el que esto ocurre fue identificado como [5, 10]. A partir de allí, se entrenaron varios modelos con la cantidad determinada de grupos y fueron comparados a partir de su índice de silhouette. Dados los resultados inconclusos iniciales, se decidió hacer

gráficos de silhouette para los intervalos [2,5), [5, 10) y [10,15), lo que permitió determinar que el modelo con mejor índice es aquel con 12 clústeres. No obstante, siguiendo las recomendaciones de literatura, se decidió entrenar un segundo modelo de K medias con k = 5 para ser comparado con los demás modelos.

4.2. MiniBatchKMeans

El modelo mini batch k means es una versión computacionalmente menos costosa que el K means convencional por su uso de lotes para el entrenamiento. El método para encontrar el número de clústeres cambió a uno por dendograma propio del modelo jerárquico aglomerativo (que no fue usado por su alto costo computacional) de la Figura 3. De esta manera, se entrenó el modelo con 4 grupos y lotes de tamaño 1000.

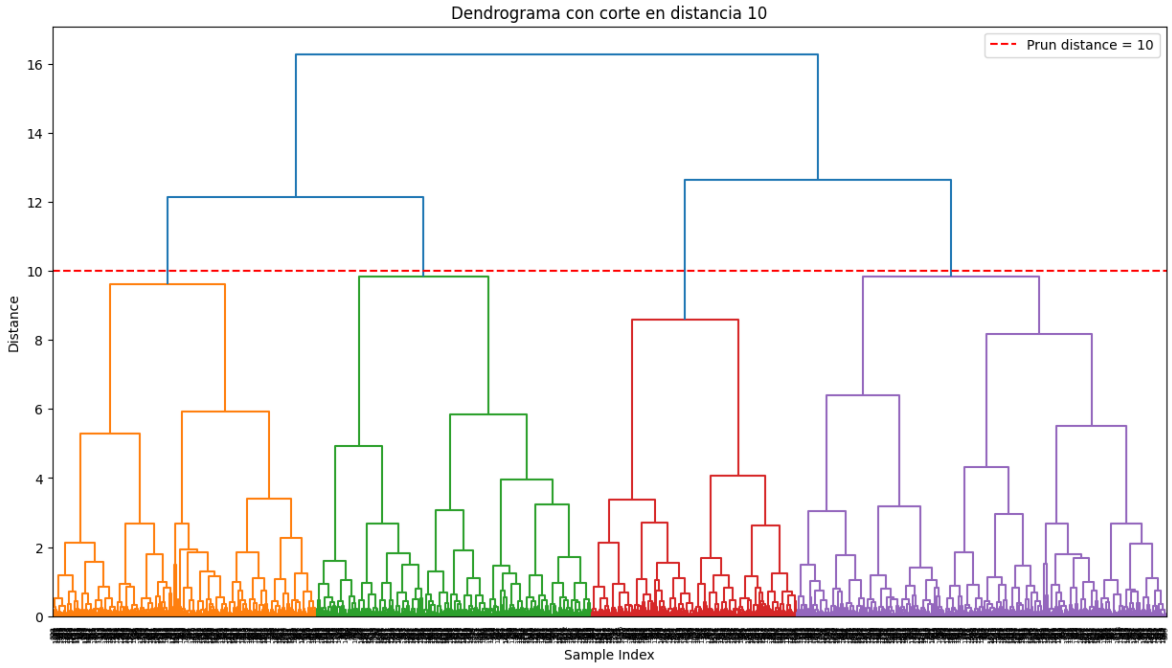

Figura 3. Dendograma

4.3. Birch

El modelo de agrupación Birch tiene tres hiperparámetros principales, número de clústeres, umbral y factor de ramificación. Este último fue dejado con valor por defecto de 50. La cantidad de grupos fue determinada por el mismo dendograma usado para el modelo descrito en 4.2. El umbral se determinó con una búsqueda exhaustiva entrenando varios modelos y seleccionando aquel que aseguró 4 clústeres con el mayor índice de silhouette.

4.4. DBScan

El modelo DBSCAN se construyó siguiendo una estrategia de ajuste de hiperparámetros basada en tres etapas. Primero, para reducir la carga computacional asociada a la exploración exhaustiva de parámetros, se seleccionó una muestra representativa de 5.000 observaciones del conjunto de datos preprocesado, manteniendo su estructura estadística y garantizando reproducibilidad mediante una semilla fija. Posteriormente, se realizó un análisis de distancias k-nearest (k=3) para estimar valores razonables del parámetro epsilon mediante el gráfico de k-distances, identificando visualmente (Figura 4) el punto de inflexión que separa regiones densas de regiones dispersas. Esta etapa permitió acotar rangos adecuados de exploración y evitó la selección arbitraria de epsilon, alineando la búsqueda con la geometría real de los datos.

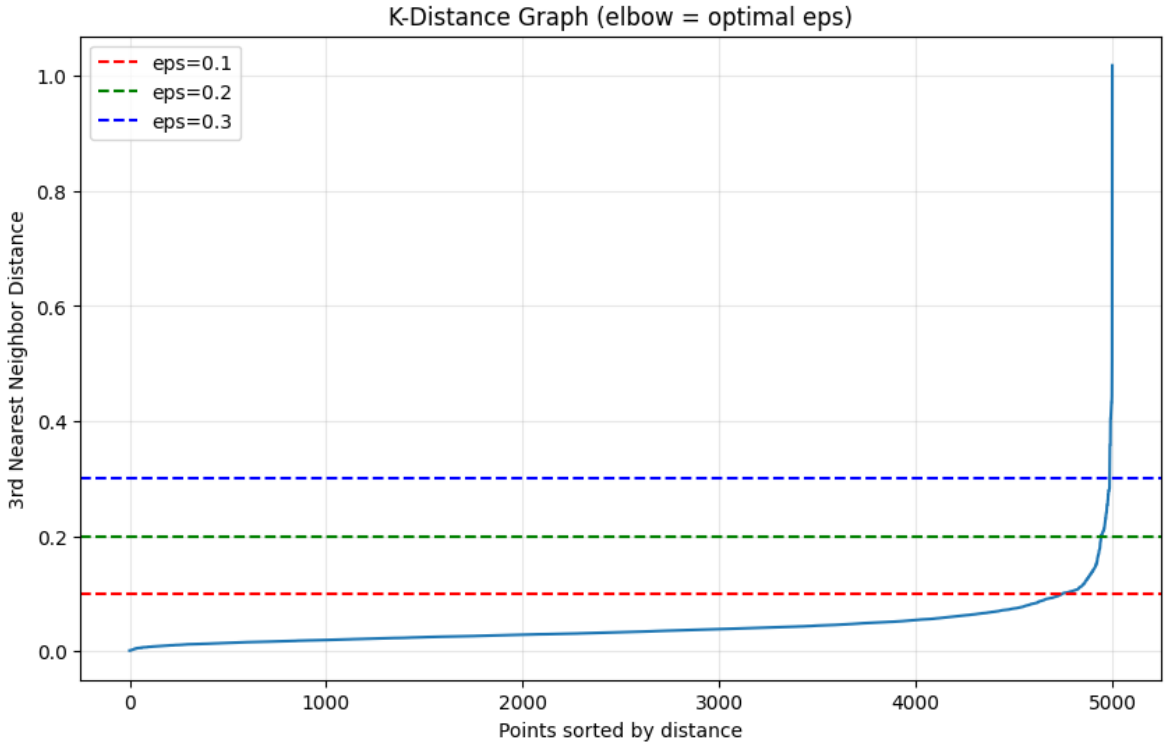

Figura 4. Distancia K

Con estos rangos delimitados, se ejecutó una búsqueda sistemática combinando valores de epsilon entre 0.1 y 0.5 e intervalos de min_samples entre 2 y 10. Para cada combinación se calcularon métricas clave: cantidad de clústeres, porcentaje de ruido y puntaje de silhouette, este último únicamente cuando la partición producía al menos dos clústeres y menos del 90 % de ruido. La selección final se basó en el mayor índice de

silhouette bajo la restricción adicional de mantener el ruido por debajo del 20%. Una vez identificados los mejores hiperparámetros, se entrenó el modelo final sobre el conjunto completo usando paralelización para optimizar tiempos de cómputo.

4.5. Gaussian Mixture Models

El modelo GMM se basó en la maximización de verosimilitud que, a su vez, implica la minimización del valor BIC (Bayes Information Criterion). Así, se entrenaron modelos de entre 2 y 40 clústeres de cuatro formas diferentes, calculando el valor BIC para cada uno y permitiendo seleccionar visualmente los valores que generan la menor cantidad de grupos posibles con el mejor desempeño. Como muestra la Figura 5, este modelo fue el GMM con tipo *full* de covarianza y 15 grupos.

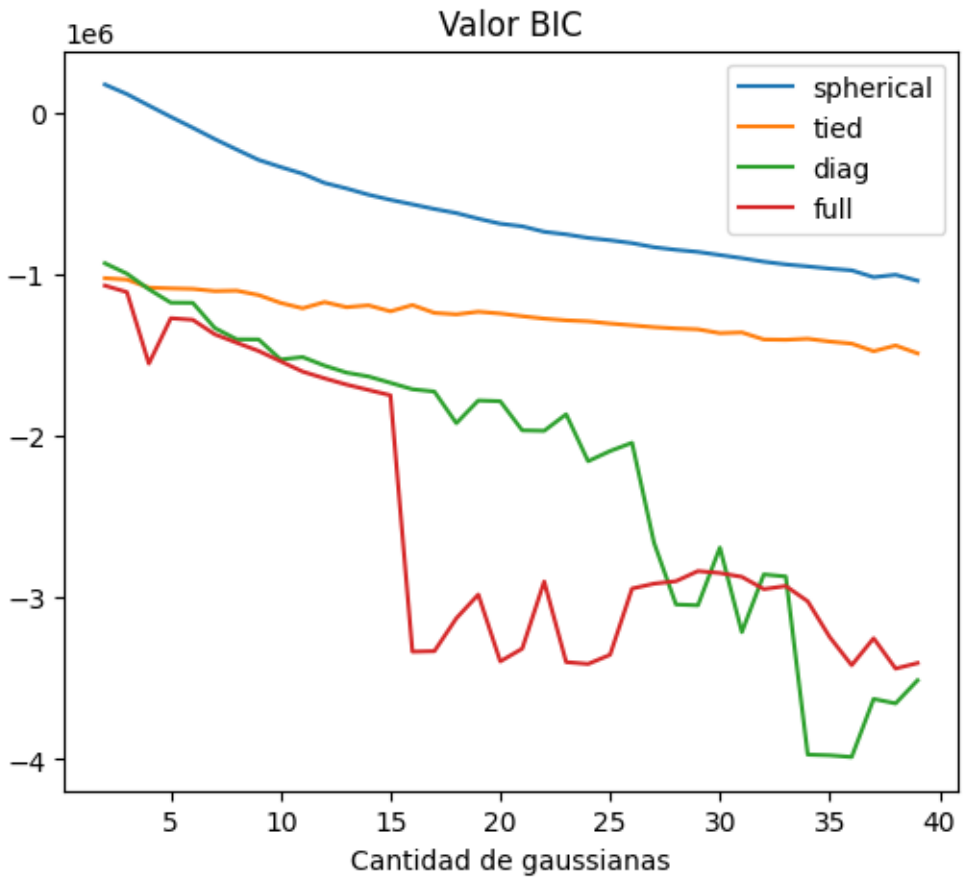

Figura 5. Valor BIC para diferentes modelos GMM

5. Análisis de Resultados

Una vez culminado el entrenamiento de los mejores modelos de cada tipo, se llevó a cabo una comparación numérica fundamentada en el índice de silhouette que arrojó los resultados contenidos en la Tabla 1.

Tabla 1. Resumen de índice de Silhouette para todos los modelos

Modelo	**Índice de Silhouette**	**Número de Grupos**
K-Means (k = 12)	0.277	12
K-Means (k = 5)	0.236	5
MiniBactch (k =4)	0.232	4
Birch (k = 4)	0.188	4
DBScan	0.240	415
GMM	0.256	15

Se evidencia que el mejor desempeño fue demostrado por los modelos K-Means con 12 clústeres, seguido de GMM y DBScan. En primera instancia, se podría optar por alguno de estos modelos por su habilidad superior de agrupar los registros. Sin embargo, se debe tener en cuenta el contexto del problema. Dado que se buscan fallas eléctricas, conviene, y arguye la literatura, tener un número reducido de clústeres para poder identificar en ellos categorías como, por ejemplo, normal, falla individual, falla de sector y dato anómalo.

Se analizó la cantidad de grupos generados por el modelo como factor determinante, eliminando aquellos con más de diez clústeres. De esta forma se concluye que el mejor modelo es K-Means con k = 5, obteniendo un resultado similar a los vistos en literatura.

5.1. Patrones de Consumo

A partir del modelo seleccionado, se realizó una visualización del patrón de consumo correspondiente a un medidor representativo de cada clúster, tal como se muestra en la Figura 6. En esta comparación inicial se observa que los clústeres 1 y 2 presentan consumos

ligeramente inferiores a su media durante los primeros registros, mientras que los demás grupos mantienen valores más cercanos a su comportamiento promedio.

Figura 6. Patrones de Consumo

Si bien este análisis preliminar permite identificar tendencias generales, un estudio más profundo revela que la variabilidad interna y la presencia de numerosos valores atípicos —como se evidencia posteriormente en la Sección 5.2 mediante diagramas de cajas y bigotes— limitan la utilidad de basarse únicamente en un medidor representativo por clúster.

Esto sugiere que estrategias de análisis que consideren el rango completo de la distribución, en lugar de un único caso, pueden ofrecer una caracterización más robusta y precisa de los patrones de consumo asociados a cada grupo.

5.2. Comparación por Grupos

La Figura 7 presenta una comparación conjunta entre la distribución del consumo energético por clúster y sus estadísticas descriptivas principales. En los diagramas de caja se evidencia una presencia considerable de valores atípicos en todos los grupos, lo cual sugiere una alta dispersión del consumo y dificulta la identificación de un medidor verdaderamente representativo dentro de cada clúster. Cada punto corresponde a una entrada (no a un medidor), y esta granularidad permite apreciar que el comportamiento fuera de los rangos típicos no está concentrado en un conjunto reducido de usuarios, sino que aparece de forma amplia y recurrente en todos los grupos.

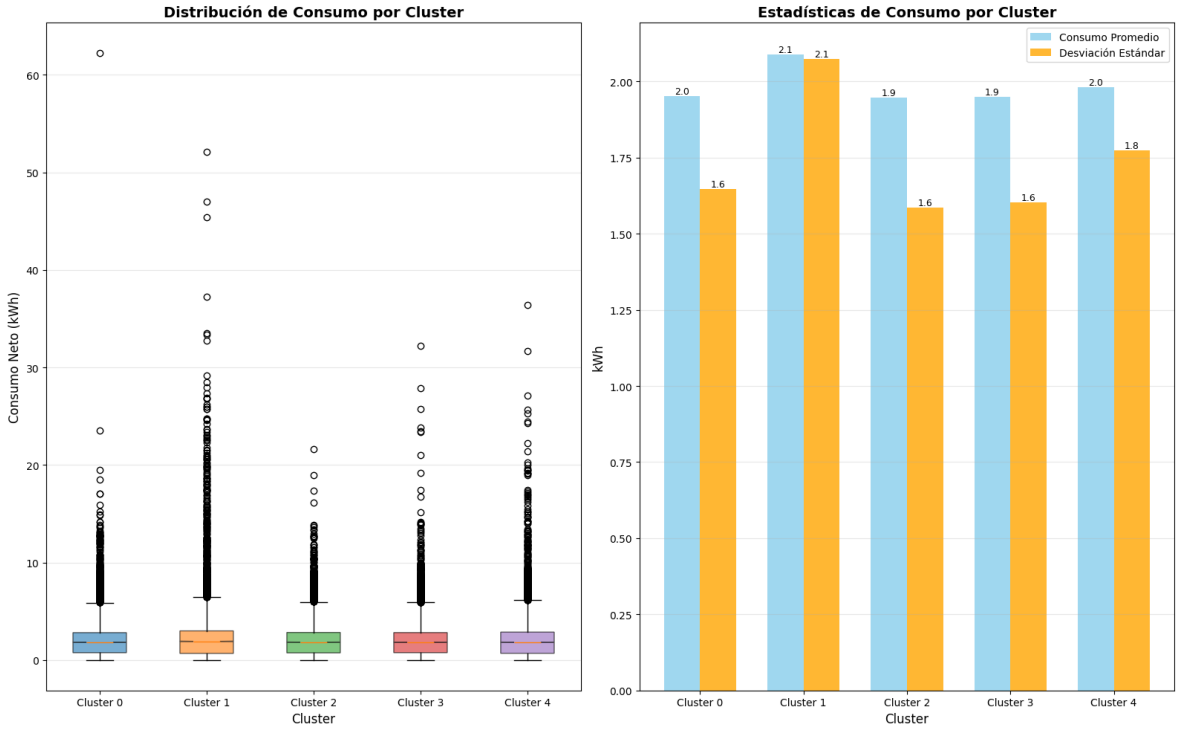

Figura 7. Diagrama de Cajas y Bigotes y Diagrama de Barras por Clúster

El clúster 0 presenta una nube de outliers muy densa cerca del extremo superior de la caja, además de un valor extremo significativamente alejado del resto, indicando casos puntuales de consumos inusualmente altos. Por su parte, el clúster 2 contiene la mayor concentración de valores atípicos, lo cual sugiere una mayor variabilidad interna y un

comportamiento menos homogéneo en su población de usuarios. Los clústeres 3 y 4 muestran distribuciones similares entre sí, con una presencia moderada de outliers que, aunque menos densa que en el clúster 2, sigue siendo suficientemente amplia para afectar la interpretación de patrones típicos. En conjunto, este comportamiento evidencia que la estrategia basada en seleccionar un único medidor representativo por clúster —como se explora en la sección previa— no es la más adecuada para capturar la complejidad del consumo real. De hecho, este panorama sugiere que enfoques alternativos, como la aplicación de umbrales para clasificar consumos extremadamente altos como sospechosos, podrían resultar más efectivos, tal como lo propone parte de la literatura consultada.

El gráfico de barras complementa este análisis mostrando las medias y desviaciones estándar del consumo por clúster. Las medias son muy similares entre todos los grupos, con diferencias marginales que no permiten distinguir clústeres claramente distintos en términos de nivel de consumo. Sin embargo, la desviación estándar aporta información más reveladora: el clúster 1 presenta la mayor variabilidad, lo que concuerda con su clasificación como un grupo de consumidores de alto consumo y comportamiento más irregular. En contraste, los clústeres 0, 2 y 3 tienen desviaciones estándar prácticamente idénticas, lo que sugiere perfiles de consumo más estables y homogéneos. El clúster 4 muestra una variabilidad levemente mayor que estos últimos, aunque inferior a la de clúster 1.

6. Conclusiones

Los indicadores de desempeño muestran una imagen mixta sobre la calidad del agrupamiento. El Silhouette Score = 0.2355 indica una separación entre clústeres baja a moderada: los valores cercanos a 0 sugieren que muchas observaciones quedan en la frontera entre grupos o que las particiones no están claramente definidas desde la perspectiva de distancia interna/externa. Esto concuerda con el Davies–Bouldin Index = 6.2939, que al ser relativamente alto apunta a una mayor similitud entre clústeres (peor discriminación), y con la observación práctica de que las medias de consumo por clúster son muy parecidas. En contrapartida, el Calinski–Harabasz Score = 18328.84 es elevado, lo que normalmente denota una separación entre grupos relativamente buena en función de la varianza explicada entre/ intra-clústers; sin embargo, este índice es sensible al tamaño del conjunto y a la estructura global de varianzas, por lo que debe interpretarse junto con las otras métricas y no de forma aislada. En conjunto, las métricas sugieren que, si bien el modelo consigue particionar el espacio de consumo en cinco grupos consistentes, la calidad de esa partición no es cristalina: la separación es débil y el agrupamiento refleja más variaciones sutiles en patrones de consumo que categorías bien definidas y separadas.

La distribución proporcional de los grupos (cada grupo concentra entre 18.1% y 21.5% de las muestras) refuerza la hipótesis de que K-Means está capturando segmentos poblacionales homogéneos en términos de consumo promedio global, pero no detecta subpoblaciones pequeñas o eventos raros que normalmente caracterizan fallas. En aplicaciones de detección de fallos, esperaríamos que los consumos anómalos o patrones asociados a eventos eléctricos (picos súbitos, caídas de tensión, corrientes inversas, etc.) conformen grupos minoritarios o puntos claramente separados; la aproximación observada aquí —clústeres proporcionales y medias de consumo muy cercanas: 1.95–2.09 kWh— indica que la segmentación está dominada por diferencias leves en nivel y variabilidad de consumo más que por la presencia de modos de falla evidentes. Por tanto, desde la perspectiva operativa de CELSIA, este resultado significa que el modelo proporciona una herramienta útil de perfilado de clientes (e.g. distinguir consumidores típicos de consumidores de mayor carga y de consumidores estables de los variables), pero no constituye por sí solo una solución fiable para la detección de fallos en la red.

La interpretación de cada clúster respalda esta conclusión. Los grupos difieren principalmente en el nivel promedio de consumo energético y en la variabilidad temporal de dicho consumo. Algunos clústeres corresponden a consumidores de bajo consumo y comportamiento estable, característicos de usuarios residenciales regulares, mientras que otros reflejan consumos más elevados y con mayor variabilidad, perfil más asociado a usuarios comerciales o residenciales con cargas intermitentes. Sin embargo, las diferencias absolutas entre estos niveles son relativamente pequeñas, lo que limita la capacidad del modelo para identificar comportamientos significativamente desviados que puedan asociarse a una falla eléctrica. La designación de medidores representativos por clúster permitió verificar visualmente estas características, observando que las curvas de consumo dentro de cada grupo mantienen patrones coherentes, pero sin anomalías manifiestas.

Estas conclusiones deben interpretarse en el contexto de la restricción fundamental del proyecto: la disponibilidad de únicamente datos de consumo energético agregados en kWh, sin variables eléctricas de mayor resolución como voltaje, corriente o frecuencia, que son los indicadores comúnmente utilizados para diagnosticar fallas en tiempo real. La naturaleza agregada del consumo suaviza variaciones súbitas y limita la detección de eventos de corta duración, lo que reduce significativamente la sensibilidad del modelo para identificar irregularidades asociadas a fallas. Por ello, aunque el modelo logra revelar patrones de uso y segmentar usuarios según su comportamiento energético, sus capacidades para la detección de fallos en la red son intrínsecamente limitadas.

A partir de estos resultados, se pueden extraer varias recomendaciones y líneas de trabajo futuro. La más determinante es la necesidad de complementar el consumo en kWh con información eléctrica de mayor granularidad, lo cual permitiría detectar fenómenos físicos propios de eventos de falla. Adicionalmente, se sugiere integrar modelos de detección de anomalías que no dependan exclusivamente de la formación de clústeres balanceados, tales como algoritmos basados en aislamiento, cambios bruscos en series temporales o ventanas deslizantes de energía. Asimismo, contrastar los agrupamientos obtenidos con registros reales de fallas o inspecciones de campo permitiría evaluar la capacidad del enfoque para apoyar decisiones operativas.

En síntesis, K-Means k=5 produce una segmentación interpretable y útil para caracterizar perfiles de consumo, pero las métricas y la distribución observada indican que no es suficiente como detector de fallos en la red bajo las condiciones actuales de datos. El resultado constituye una base de trabajo valiosa para priorizar inspecciones y diseñar hipótesis operativas (por ejemplo, enfocar muestreos de campo en los clústeres variables de mayor consumo), pero alcanzar una detección de fallas fiable exigirá datos eléctricos más completos, técnicas complementarias de detección de anomalías y una validación cruzada con eventos reales en la red.

7. Recursos Complementarios

Para complementar el presente informe se tienen dos recursos complementarios. El primero es un documento PDF titulado Investigación_NuevoEnfoque.pdf. El segundo, el material en el repositorio de GitHub https://github.com/santy-estrada/electric-grid-project.git.