UNIVERSIDAD POLITÉCNICA SALESIANA

SEDE CUENCA

CARRERA DE ELECTRICIDAD

CLASIFICACIÓN DE PERFILES DE CARGA EN CONSUMIDORES COMERCIALES MEDIANTE ANÁLISIS DE CONGLOMERADOS

Trabajo de titulación previo a la obtención del título de Ingeniero Eléctrico

AUTORES: BRYAN STEVEN CASTILLO PÉREZ
JOHN PATRICIO ENRIQUEZ LOJA
TUTOR: ING. JOHNNY XAVIER SERRANO GUERRERO, PhD.

Cuenca - Ecuador
2023

CERTIFICADO DE RESPONSABILIDAD Y AUTORÍA DEL TRABAJO DE TITULACIÓN

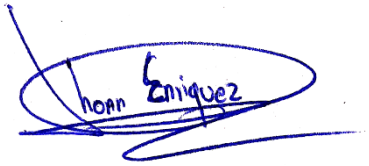
Nosotros, Bryan Steven Castillo Pérez con documento de identificación N° 0302131248 y John Patricio Enriquez Loja con documento de identificación N° 0106245137; manifestamos que:

Somos los autores y responsables del presente trabajo; y, autorizamos a que sin fines de lucro la Universidad Politécnica Salesiana pueda usar, difundir, reproducir o publicar de manera total o parcial el presente trabajo de titulación.

Cuenca, 10 de noviembre de 2023

Atentamente,

Bryan Steven Castillo Pérez
0302131248



John Patricio Enriquez Loja
0106245137

CERTIFICADO DE CESIÓN DE DERECHOS DE AUTOR DEL TRABAJO DE TITULACIÓN A LA UNIVERSIDAD POLITÉCNICA SALESIANA

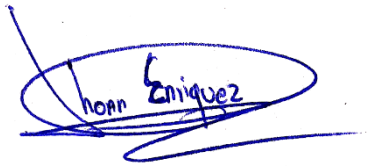
Nosotros, Bryan Steven Castillo Pérez con documento de identificación N° 0302131248 y John Patricio Enriquez Loja con documento de identificación N° 0106245137, expresamos nuestra voluntad y por medio del presente documento cedemos a la Universidad Politécnica Salesiana la titularidad sobre los derechos patrimoniales en virtud de que somos autores del Proyecto técnico: "Clasificación de perfiles de carga en consumidores comerciales mediante análisis de conglomerados", el cual ha sido desarrollado para optar por el título de: Ingeniero Eléctrico, en la Universidad Politécnica Salesiana, quedando la Universidad facultada para ejercer plenamente los derechos cedidos anteriormente.

En concordancia con lo manifestado, suscribimos este documento en el momento que hacemos la entrega del trabajo final en formato digital a la Biblioteca de la Universidad Politécnica Salesiana.

Cuenca, 10 de noviembre de 2023

Atentamente,

Bryan Steven Castillo Pérez
0302131248



John Patricio Enriquez Loja
0106245137

CERTIFICADO DE DIRECCIÓN DEL TRABAJO DE TITULACIÓN

Yo, Johnny Xavier Serrano Guerrero con documento de identificación N° 0104983382, docente de la Universidad Politécnica Salesiana, declaro que bajo mi tutoría fue desarrollado el trabajo de titulación: CLASIFICACIÓN DE PERFILES DE CARGA EN CONSUMIDORES COMERCIALES MEDIANTE ANÁLISIS DE CONGLOMERADOS, realizado por Bryan Steven Castillo Pérez con documento de identificación N° 0302131248 y John Patricio Enriquez Loja con documento de identificación N° 0106245137, obteniendo como resultado final el trabajo de titulación bajo la opción Proyecto técnico que cumple con todos los requisitos determinados por la Universidad Politécnica Salesiana.

Cuenca, 10 de noviembre de 2023

Atentamente,

Ing. Johnny Xavier Serrano Guerrero, PhD.
0104983382

Clasificación de Perfiles de Carga en Consumidores Comerciales Mediante Análisis de Conglomerados

Bryan Steven Castillo Pérez & John Patricio Enriquez Loja
bcastillop1@est.ups.edu.ec & jenriquezl1@est.ups.edu.ec
Universidad Politécnica Salesiana
Cuenca - Ecuador

Resumen

Este trabajo expone métodos de segmentación de perfiles de carga eléctrica de usuarios comerciales y permite elegir la más pertinente de acuerdo a los tipos de datos. La identificación y comprensión de los perfiles de consumo resultan útiles para analizar la eficiencia energética, ajuste tarifario, planificación y la toma de decisiones. Por ello, este estudio presenta un método para evaluar el desempeño de las técnicas de agrupación *K-means*, Agrupación Jerárquica, *Fuzzy C-means*, Mapa Autoorganizado, Modelo de Mezcla Gaussiana y Árbol de Decisiones, aplicado a los factores de forma de los clientes. Los resultados indican que los algoritmos Mapa Autoorganizado y *Fuzzy C-means* presentan una homogeneidad en su categorización. En los tres casos de estudio, los resultados son evidentes debido a su bajo porcentaje de valores atípicos.

Palabras clave: *cluster, típicos, atípicos, perfil de carga, consumo energético, algoritmos de clasificación, aprendizaje no supervisado*

1. Introducción

Las empresas del sector eléctrico encargadas de operar y gestionar el suministro de energía eléctrica, disponen de una sólida base de datos con un

amplio registro de mediciones. Esto se logra a través de la implementación de avanzados sistemas de comunicación y medición, los cuales proporcionan información en un intervalo de tiempo. Los proveedores de energía se esfuerzan por garantizar un suministro eléctrico confiable y estable, con el objetivo de brindar a los usuarios un servicio de calidad.

En los últimos años, Ecuador ha experimentado un incremento significativo en el número de usuarios. Este crecimiento ha provocado importantes desafíos para las empresas distribuidoras eléctricas, los cuales incluyen la clasificación de patrones de consumo, gestión energética, detección de anomalías, formulación de tarifas, previsión de la demanda eléctrico futura, optimización de la red, entre otros aspectos. Es por eso que, en [1], agrupan los datos para construir intervalos de predicción para los valores de carga eléctrica, analizando el espectro singular.

El perfil de carga brinda información del consumidor a lo largo del tiempo [2], permitiendo visualizar patrones y fluctuaciones. Comprender estos cambios es fundamental, para identificar áreas potenciales de mejora. En consecuencia, la clasificación de usuarios se ha realizado tradicionalmente teniendo en cuenta los perfiles de carga diarios o los factores de forma de carga [3].

En un entorno en el que la disponibilidad de información no es una limitante, el verdadero desafío radica en la capacidad de procesar y analizar adecuadamente los datos. En este sentido, las técnicas de agrupamiento han demostrado ser valiosas para comprender y descifrar los patrones ocultos que revelan valiosa información en los perfiles de carga eléctrica. Las técnicas de agrupamiento se pueden dividir en supervisadas y no supervisadas, entre ellas las más utilizadas se puede citar [4], [5].

Las técnicas de aprendizaje supervisado y no supervisado son aplicadas para manejar distintas problemáticas, se presentan resultados de una investigación minuciosa y exhaustiva. Se ha encontrado numerosos esfuerzos por parte de diferentes autores en aspectos comparativos de agrupamiento de series de tiempo [6]. para conocer su estructura. En [7], analiza y compara el rendimiento de tres técnicas de agrupación, las cuales son árbol de decisiones, máquina de suporte vectorial (SVM) y redes neuronales artificiales (ANN) para la minería de datos. Se considera que cada algoritmo ofrece un enfoque único para abordar diferentes escenarios de estudio.

En 1957, Stuart Lloyd propuso el algoritmo de *K-means* [8]. Desde entonces, es ampliamente utilizado en varios campos, entre ellos esta la clasificación perfiles de carga de clientes no residenciales en función de su consumo energé-

tico [9]. Otros propusieron métodos que estiman el numero óptimo de *clusters* para el mejor funcionamiento [10], [11] implementa el método de deducción de factores de forma (SFs) para mejorar el rendimiento computacional, [12] realiza un análisis comparativo con tres medidas de distancia diferente y [13] analiza los puntos débiles e implementa el algoritmo *K-means++*.

Desde su introducción por primera vez en 1973 por Dunn, el método de *Fuzzy C-means* (FCM) [14] ha sido objeto de análisis y mejora. Además, investigaciones como la de [15] han examinado diversas medidas de distancias en su aplicación. En [16], se llevó a cabo una agrupación de datos para abordar los desafíos asociados a este proceso, explorando sus aplicaciones en campos como la medicina y la ingeniería.

El algoritmo *Gaussian Mixture Model* (GMM) se menciona por primera vez en 1977 [17], presenta un método de maximización para estimar los parámetros, constituyendo el fundamento del mismo algoritmo. Por esta razón, en [18] verifica la precisión del algoritmo en el proceso de categorización, proponiendo una mejora en la precisión y eficiencia de agrupación en comparación con otros métodos. Además, en [19] detalla el uso del modelo en diversos campos, como la segmentación de imágenes y clasificación de datos. Otros estudios, como [20] y [21], hacen uso de la técnica para agrupar.

Teuvo Kohonen presentó el *algoritmo Self-Organizing Maps* (SOM) en 1982 [22], ha dado lugar a numerosos estudios que exploran su uso y aplicaciones específicas. En [23], lo emplea para la clasificación de carga eléctrica en entornos de redes inteligentes, conocidas como *Smart Grids*. Mientras que, el autor Rajabi A. en [24], llevó a cabo un análisis comparativo de diversas técnicas de agrupación para la segmentación de patrones de carga eléctrica, se evaluó la eficacia y aplicabilidad del algoritmo SOM en comparación a las otras metodologías. Por otro lado, en [25], hace uso para examinar datos biológicos, explora patrones complejos y la identificación de relaciones entre variables biológicas.

Agrupación jerárquico es otra poderosa técnica que ha demostrado su eficacia en la agrupación de perfiles de carga, tal como se evidencia en las investigaciones. En [26], hace uso para correlacionar los precios de perfiles eléctricos diarios con el período calendario. En cambio, [27] segmenta y obtiene patrones representativos de consumidores de carga diurna. Además, se ha encontrado aplicaciones relevantes en el campo médico como lo plantean en [28] para el ámbito de la lipidómica.

Entre las diversas técnicas de aprendizaje supervisado, destaca la integración del árbol de decisiones. Esta técnica es de gran utilidad para la cla-

sificación y predicción de datos, lo que la convierte en una opción valiosa para el análisis de conjuntos de datos etiquetados. En la literatura científica, existen publicaciones que hacen uso del método y lo contrastan con otras técnicas, para evaluar su eficacia en la clasificación. A continuación, se destacan algunas de estas investigaciones aplicadas en diversos campos [29], [30], [31].

Es importante destacar que algunos autores han propuesto ideas innovadoras en el campo de la clasificación y predicción de datos. Por ejemplo, [32] ha desarrollado un método novedoso para la clasificación del tráfico de red con técnicas de aprendizaje automático. En [33], usa redes neuronales y la metodología *Kaastra-Boy* con el propósito de determinar las variables de entrada significativas para RNN, eliminando parámetros redundantes a través de la conversión de datos. Además, las técnicas de aprendizaje supervisado han sido aplicadas en diversos campos, como se evidencia en el estudio de Hammami [34], donde se utiliza SVM y *K-means* para la categorización de registros de llamadas. Por otro lado, [35] ha desarrollado un método avanzado para el agrupamiento de patrones de carga residencial, que consta de dos etapas y se basa en la optimización metaheurística.

Para ampliar aún más la perspectiva sobre las diferentes técnicas utilizadas en investigaciones relacionadas, se exploraron diversos artículos que abordan enfoques distintos al presente estudio. En [36] realiza un nuevo algoritmo de bucle cerrado que permite hacer un *clustering* cerrado. Por otra parte, [37] realiza la clasificación de los perfiles de carga pero agregando la predicción para analizar y comparar los resultados, en [38] también realiza la clasificación y predicción pero en lugar de utilizar técnicas no supervisadas implementa con técnicas supervisadas como lo son SVM y KNN. Otros trabajos, como es el de [39] se ha enfocado en aplicar las técnicas de agrupación para mejorar la eficiencia y la rentabilidad de las empresas de electricidad. [40] crea una nueva técnica la cual es la *matrix profile based*, siendo una técnica de minería de datos que se utiliza para la extracción de características y la identificación de patrones en series temporales.

En la literatura mencionada, se identifica una brecha en el análisis y la comparación de los perfiles de carga para diferentes tipos de consumidores. Además, se observa la ausencia de un análisis exhaustivo en la obtención de parámetros para los métodos revisados. Es importante mencionar que la aplicación de estos algoritmos en Lationamerica es escaza. Es por ello que, explorar la aplicación de estos algoritmos resulta de gran interés y relevancia para el avance en el estudio de este campo.

Esta investigación se centra en la implementación de diversas técnicas

de agrupación, entre las que se incluyen *K-means*, Agrupación Jerárquica, FCM, SOM, GMM y Árbol de decisión. El objetivo principal es analizar y comparar el rendimiento de estas técnicas en la clasificación de perfiles de carga diarios del sector comercial. Esta comparativa permitirá evaluar el desempeño relativo de cada técnica de agrupación y determinar cuál es la más efectiva en este contexto.

Sin embargo, es fundamental resaltar que, a pesar de la amplia utilización de técnicas de agrupación en diversos campos, existe una notable carencia de investigaciones específicas centradas en su aplicación para el análisis de perfiles de carga diarios. Por lo tanto, este estudio busca contribuir a la literatura existente al proporcionar una evaluación exhaustiva de estas técnicas en este campo.

A continuación, se describe la estructura del documento. En la Sección 2, se realiza un análisis del contexto y antecedentes relevantes, mientras que en la Sección 3 se describe detalladamente el proceso de la metodología empleada en la investigación, incluyendo el diseño experimental y las técnicas utilizadas. Finalmente, en la Sección 4, se presentan los resultados obtenidos a través del análisis de los datos recolectados, brindando una interpretación precisa respaldada por análisis estadísticos y cualitativos pertinentes.

2. Antecedentes

Según la Agencia Internacional de la Energía [41] la evolución del sector eléctrico global, ha experimentado una serie de cambios significativos debido a diversos factores. En los últimos años se observó un crecimiento constante en la capacidad de generación eléctrica. La electricidad se ha convertido en el núcleo del desarrollo de la sociedad, esta se encuentra presente en nuestras vidas diarias casi todo el tiempo

2.1. Perfil de Carga

Un patrón de carga típico representa la distribución temporal del consumo eléctrico diario del usuario final [42]. Para evaluar adecuadamente la carga de los distintos usuarios, es necesario realizar un análisis estadístico de su consumo eléctrico.

Un perfil de carga se define como una curva que representa el consumo de energía eléctrica de un cliente concreto en función del tiempo, que puede

ser diario, semanal, mensual o incluso anual. El perfil de consumo depende de diversos factores, por ejemplo, los hábitos de consumo de los usuarios, los niveles socio-económicos, los aspectos socioculturales, el clima de la región, el sector productivo y el tamaño de la empresa, entre otros [9].

El análisis de agrupación puede revelar patrones de consumo eléctrico ocultos en los perfiles históricos de un usuario. Los usuarios pueden tener diferentes patrones según el período de tiempo analizado. Cada usuario puede pertenecer a diferentes conglomerados, y se selecciona el conglomerado con mayor presencia en los perfiles del usuario como su patrón de consumo característico.

2.2. Selección de datos

El primer paso en el proceso de selección de datos es definir el problema. Después, es posible seleccionar los datos para aplicar algoritmos y resolver el problema. Por ejemplo, para explicar los patrones de consumo eléctrico en una zona determinada es necesario identificar a los clientes más relevantes, como los consumidores residenciales, comerciales o industriales. Otra posible selección de clientes es por sus parámetros contractuales, como son, los niveles de voltaje, ya que no tiene sentido una comparación con los patrones de consumo de energía entre clientes de baja y alta tensión [39].

2.3. Concepto de conglomerado

El conglomerado es una técnica de minería de datos no supervisada que permite determinar patrones intrínsecos en conjuntos de datos. El objetivo principal del conglomerado es dividir los datos, denominados objetos de un conjunto en una serie de grupos, denominados *clusters* que agrupan a los objetos más similares entre sí.

2.4. Técnicas de aprendizaje no supervisado

2.4.1. *K-means*

El método de *K-means* es una técnica no supervisada utilizada para la división de un conjunto de datos en grupos o *clusters* homogéneos. *K-means* es un algoritmo iterativo que comienza seleccionando k centroides aleatorios para luego representarlos en k *clusters*. Luego, cada punto de datos se asigna

al *cluster* cuyo centroide está más cercano. Finalmente, en la siguiente iteración se recalculan los centroides de cada *cluster* utilizando la media de los puntos asignados a cada uno de ellos, este paso se repite hasta que el resultado de las iteraciones no varié, en otras palabras, el centroide no cambie [43], [44], [45].

2.4.2. *Cluster* Jerárquico

La Agrupación Jerárquica es un método que consiste en construir un árbol de fusión binario utilizando los elementos de datos almacenados en las hojas. Cada hoja representa un único elemento. El proceso comienza fusionando de a pares los subconjuntos "más cercanos" almacenados en los nodos, hasta llegar a la raíz del árbol, que contiene todos los elementos de X. La distancia entre dos subconjuntos arbitrarios de X se denota como $\Delta(X_i, X_j)$ y se conoce como distancia de enlace. Este enfoque también se conoce como agrupación jerárquica aglomerativa, ya que se parte de las hojas que contienen elementos individuales (los X_i) y se fusionan iterativamente los subconjuntos hasta llegar a la raíz [46], [47].

2.4.3. *Fuzzy C-means*

El método conocido como *C-Means* se utiliza para realizar *clustering*, donde çrepresenta el número de clases o *clusters*. Si se utiliza la técnica difusa para definir las clases, se le conoce como FCM. El enfoque FCM asigna un grado de pertenencia difusa a cada clase. En el *clustering* difuso, la importancia del grado de pertenencia es similar a la probabilidad de píxel en un modelo de mezclas. La ventaja del FCM radica en su capacidad para formar nuevos *clusters* a partir de los puntos de datos que tienen valores de pertenencia cercanos a las clases existentes. El método FCM se basa en tres operadores fundamentales: la función de pertenencia difusa, la matriz de partición y la función objetivo [48], [49], [50].

2.4.4. *Gaussian Mixture Model*

El modelo de mezcla gaussiano se puede describir como una composición de K modelos gaussianos individuales, los cuales actúan como variables ocultas dentro de este modelo híbrido. En términos generales, un modelo mixto puede emplear diversas distribuciones de probabilidad. Sin embargo, en este caso específico, se utiliza el modelo de mezcla gaussiana debido a las ventajas

que ofrece la distribución gaussiana en cuanto a sus propiedades matemáticas y eficiencia computacional. Esta elección se respalda en el hecho de que la distribución gaussiana proporciona una sólida base teórica y permite un cálculo eficiente, lo cual contribuye a un mejor rendimiento del modelo en cuestión [19].

La elección de la distribución gaussiana en el GMM se basa en sus propiedades matemáticas y su capacidad para capturar eficazmente la estructura subyacente de los datos. La distribución gaussiana es conocida por su forma de campana y su flexibilidad para modelar diferentes tipos de distribuciones de datos [21].

2.4.5. *Self-Organizing Map*

SOM es un método de clasificación automática no supervisada, también conocido como *clustering*. Este enfoque clasifica directamente utilizando un mapa auto-organizativo, y su proceso de aprendizaje se basa en la competencia entre las neuronas o nodos del mapa. Es importante destacar que el comportamiento auto-organizativo está estrechamente vinculado con el método de aprendizaje competitivo. En el caso específico de SOM, el mapa consiste en una grilla regular de neuronas o nodos, donde cada uno de ellos representa una clase o *cluster*. El algoritmo de los mapas auto-organizados de Kohonen guarda similitudes con el método de *k-medias (k-means)*. Ambos utilizan la idea de agregación alrededor de centros móviles, una inicialización aleatoria, la asignación de instancias a los centros más cercanos y la búsqueda de mínimos locales. No obstante, la diferencia radica en que el algoritmo SOM conduce a una representación plana, es decir, a un orden espacial donde se preservan las relaciones entre los datos [51], [52].

2.5. Técnicas de aprendizaje supervisado

2.5.1. Árbol de decisión

Los árboles de decisión son algoritmos estadísticos que se utilizan para la predicción y clasificación en diversas áreas. Estos mecanismos han evolucionado desde sus primeras implementaciones electrónicas en las últimas décadas del siglo XX. En la actualidad, son herramientas computacionales versátiles y ampliamente aplicadas en campos como la inteligencia artificial,

el aprendizaje automático, la minería de datos y el descubrimiento de conocimientos [53].

La característica clave de los árboles de decisión es su capacidad para dividir los datos en subconjuntos más pequeños y homogéneos, basándose en los valores de los atributos predictores. Cada partición representa una hoja o nodo, y dentro de ella se encuentran valores objetivo similares. A medida que se desciende en el árbol, las hojas contienen valores objetivo cada vez más diferentes entre sí, lo que permite la predicción y clasificación efectiva de nuevos datos [54].

2.6. Distancia Euclidiana

La distancia euclidiana se utiliza para medir el espacio en geometría euclidiana, en otras palabras es una forma de geometría plana basada en los principios matemáticos de Euclides [55]. La ecuación para calcular la distancia euclidiana es la longitud de una línea recta que conecta dos puntos en un espacio n-dimensional de geometría euclidiana.

Se elige la distancia euclídea para la medida de similitud utilizada en el algoritmo de *clustering*. La distancia euclídea entre dos vectores n-dimensionales x e y esta determinada por la siguiente ecuación [3]:

$$d_E(x, y) = |x - y| = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \tag{1}$$

2.7. Distancia de la Transformada Wavelet Discreta

La distancia wavelet es una métrica empleada en el análisis de señales y datos en los dominios del tiempo y la frecuencia. Esta medida se basa en la transformada wavelet discreta (DTW), que descompone una señal en diversas escalas y frecuencias discretas. La distancia wavelet se utiliza para evaluar la similitud o diferencia entre dos señales o series de tiempo en distintas escalas y frecuencias, permitiendo un análisis detallado de las características presentes en los datos [56]. La DTW se implementa utilizando filtros pasa altos y pasa bajos, que permiten obtener los coeficientes de detalle (Dx) y aproximación (Ax) en una escala de 2x respectivamente. Estos coeficientes proporcionan información valiosa sobre las características de las señales y series de tiempo en diferentes escalas y frecuencias [57]

2.8. Escalado multidimensional (MDS)

El trabajo [58], presenta una técnica llamada MDS, que utiliza el análisis de conglomerados para reducir el costo computacional en el pre-procesamiento y la visualización de datos. En resumen, MDS convierte las observaciones de datos brutos de múltiples dimensiones en puntos en un espacio de menor dimensión, utilizando una matriz de distancias para capturar las relaciones entre los puntos. Luego, se proyectan los puntos en un espacio de menor dimensión, logrando que puedan ser comparados mediante la disimilitud entre las matrices de distancias original y reducida. Esta técnica permite una eficiente reducción de dimensionalidad de los datos de entrada y el análisis de conglomerados Esto se lo define mediante la ecuación:

$$Stress = \sqrt{\frac{\sum_{i=1,j=1}^{N}(d'_{i,j} - d_{i,j})^2}{\sum_{i=1,j=1}^{N} d_{i,j}^2}} \tag{2}$$

Siendo $i \neq j$, tanto para el numerador como para el denominador

2.9. Normalización de Datos

La normalización es importante para garantizar que los datos sean comparables y no se vean afectadas. Se realiza la conversión del consumo eléctrico horario de kWh en un rango entre 0 y 1 [2].

Hay dos tipos de normalización: la primera implica convertir los datos a un vector con una longitud de 1, y la segunda consiste en calcular las desviaciones estándar de los elementos del vector, conocidas como "puntuaciones Z"[59].

2.10. Índices de validación de conglomerados

2.10.1. Método del Codo

Este método es utilizado para especificar el número óptimo de conglomerados de un conjunto de datos mediante una técnica visual. El gráfico se obtuvo a partir del cálculo de la Suma Cuadrada de Errores (SSE). El número del conglomerado se determinó observando la posición del punto en el brazo “ Codo ” [60].

La premisa fundamental de este concepto es que, a medida que se incrementa el número de grupos k, la concentración de cada grupo aumentará progresivamente, lo cual dará lugar a una reducción natural en la suma

de los errores cuadráticos (SSE), lo que incide en una mayor eficacia en el agrupamiento [61]. La ecuación para la obtención en la suma de los errores cuadráticos la propone en [62]:

$$SSE = \sum_{k=1}^{K-1} \sum_{x_i \in S_k} |X_i - C_k|_2^2 \tag{3}$$

2.10.2. Método de la Silueta

En [63] realiza el calculo del coeficiente de silueta medio de todas las muestras. El coeficiente de silueta se calcula teniendo en cuenta la distancia media A y la distancia media al *cluster* más cercano B de cada punto de datos. La ecuación, para la obtención del coeficiente de silueta esta dada se define como: [64]:

$$\frac{b-a}{\text{máx}(a,b)} \tag{4}$$

2.10.3. Distribución Normal

La distribución normal, también conocida como la “Campana de Gauss", se atribuye originalmente a Abraham de Moivre (1667-1754), un matemático francés. Sin embargo, fue Carl Friedrich Gauss (1777-1855) quien profundizó en esta distribución y formuló la ecuación de la curva. Esta distribución se caracteriza por dos parámetros clave: la media (μ) y la desviación estándar (σ), que determinan completamente la forma de la curva para una variable normal. Utilizando esta notación, la función de densidad de la distribución normal se expresa mediante la siguiente ecuación [65]:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\} \quad ; \quad -\infty \leq x \leq \infty \tag{5}$$

3. Metodología

En este trabajo, se emplea un enfoque analítico que inicia con la recopilación de datos, su pre-procesamiento, clasificación y, finalmente con el análisis de resultados obtenidos. Estos pasos se ilustran de manera detallada en la Figura 1.

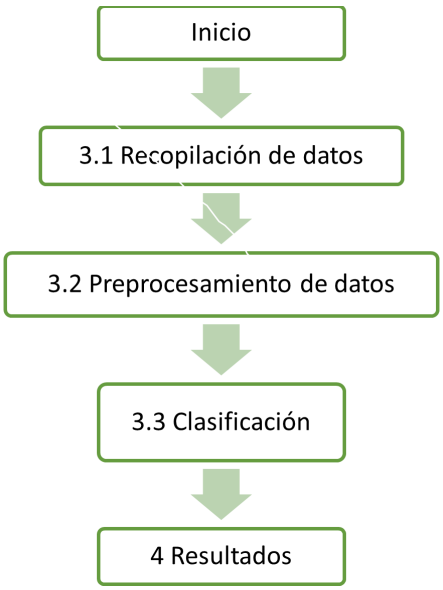


Figura 1: Diagrama de flujo del proceso de clasificación.

La metodología se compone de cuatro etapas principales, con la finalidad de descifrar los perfiles de carga. Además, cada etapa consta de subsecciones, que se describe en la Figura 1.

Como se menciono anteriormente, esta investigación parte desde la recopilación de datos, específicamente del consumo eléctrico. En la segunda etapa se lleva a cabo el pre-procesamiento, que implica el filtrado, la limpieza, la trasformación y normalización de datos. Con el objetivo de reducir el rendimiento computacional en el uso de las técnicas de agrupación que serán implementadas.

En la etapa de la clasificación, es importante determinar parámetros iniciales, con el fin de maximizar y mejorar los resultados. Es por ello que, se parte con la obtención del número óptimo de *clusters*, seguido de la implementación de las técnicas de agrupación. Posteriormente, efectúa la validación de las técnicas, todo este proceso se encuentra en un bucle que busca asegurar buenos resultados. Finalmente, con base en los resultados obtenidos, se realiza un análisis estadístico para profundizar en la interpretación de la segmentación.

3.1. Recopilación de datos

3.1.1. Selección de Datos

Las bases de datos utilizadas en esta investigación contienen registros de potencia activa tomados en intervalos de 15 minutos, donde se presenta la columna del tiempo junto con información adicional como la fecha y hora correspondientes. Esta característica temporal y estructura de datos se encuentra presente en todas las bases de datos analizadas.

La información recopilada respeta rigurosamente la forma en que fue recopilada inicialmente, es decir, sin manipular. Sin embargo, estos datos serán sometidos a un etapa de pretratamiento.

3.2. Preprocesamiento de datos

La adquisición de datos del consumo energético puede incorporar algunos errores e incoherencias, debido a la precisión de los equipos, fallos de comunicación entre dispositivos y el conjunto de datos, conversión de datos y almacenamiento. Estos problemas pueden causar perdida de información, valores erróneos o duplicados. Además, ocasiona distorsión o ser inconsistente, consiguiendo confusión en la calidad del escrutinio de la segmentación.

3.2.1. Filtrado y limpieza de datos

Los datos brutos, contiene información sobre la demanda de potencia para un momento específico dentro del periodo de tiempo seleccionado [66]. Con la recopilación de datos, en la columna del tiempo se realizar una conversión en un formato de fecha y hora. Pero, es común encontrar valores faltantes, en las cuales se aplica la interpolación para estimar valores desconocidos, adicional a esto si se presenta un carácter o símbolo, se lo sustituye por el valor de cero. De esta manera, los datos quedaron listos para el análisis sin distorsiones.

3.2.2. Transformación de escalado multidimensional

MDS, es fundamental en el análisis de datos que busca transformar los valores de múltiples dimensiones o variables a un rango específico y estandarizado. Además, reduce la dimensionalidad de los datos de entrada, con el objetivo de disminuir significativamente el costo computacional.

La transformación conserva información importante del registro de la demanda, lo cual resulta útil para el análisis de la segmentación [67]. Es por ello que, en el presente estudio, adquiere relevancia al abarcar múltiples años de información. Este proceso utiliza la información en función del tiempo, es decir, consta de 24 puntos de datos correspondientes a 24 horas del día. Como resultado, se obtiene una matriz que representa los perfiles de carga diario para cada consumidor comercial.

3.2.3. Normalización

La técnica de normalización es utilizada con el propósito de ajustar las magnitudes de las variables a un rango estandarizado. Esta acción es crucial para eliminar cualquier sesgo causado por las diferencias en las escalas de las mediciones [68]. El proceso busca ajustar cada registro del perfil de carga a una media de cero y una desviación estándar de uno, logrando así una representación estandarizada de los datos. Matemáticamente, esta normalización implica una operación que utiliza el valor original ("x"), su media específica en el conjunto de datos ("μ") y su desviación estándar ("σ"). El resultado final es el valor escalado de la característica ("z"), que refleja la relación entre el valor original y la variabilidad presente en el conjunto de datos. Este enfoque de preprocesamiento es fundamental para garantizar una comparación justa y un análisis preciso de las diferentes características en el conjunto de datos.

3.2.4. Factores de forma de carga (SFs)

Los SFs son parámetros utilizados para analizar las características de una señal eléctrica, especialmente en relación con la potencia activa. Estos factores brindan información valiosa sobre la variabilidad y la distribución de la potencia dentro de una señal determinada. En [11], propone una forma de tratar los perfiles de carga que se muestra en la Tabla 1.

Basado en el escalado multidimensional previamente realizado, se plantea ahora reducir el costo computacional, para mejorar la eficiencia y resultado de los algoritmos. Esto con el objetivo de agilizar el proceso de clasificación.

Tabla 1: Factores de forma de carga.

SFs	**Ecuación**	**Periodo**
Factor de carga	$f_1 = \frac{P_\mu}{P_{\max}}$	Todo el día
Coeficiente de no uniformidad	$f_2 = \frac{P_{\max}}{P_\mu}$	
Desviación estándar	P_σ	
Potencia media	P_μ	
Potencia máxima	$P_{\max}$	
Potencia mínima	$P_{\min}$	

3.3. Clasificación

Con la implementación de algunas técnicas supervisadas y no supervisadas, a partir de las etapas empleadas en el pre-procesamiento, se entrena cada algoritmo para cada base de datos de los usuarios comerciales. En el contexto de la sistematización, permite identificar fortalezas y debilidades de cada método, para diferentes escenarios o casos de estudios planteados. Con el propósito de identificar el procedimiento más apropiado y lograr una menor dispersión de datos en cada conglomerado.

3.3.1. Número óptimo de conglomerados

Un número apropiado de agrupaciones, asegura que cada perfil este bien distribuido y cada grupo contenga patrones significativos y representativos de la matriz de perfiles de carga. Por lo tanto, esta elección precisa evita problemas de sobreajuste, por lo que, se presentan dos métodos importantes en el análisis de *clustering*, tales como el método de la Silueta y el método del Codo.

Para los dos métodos, se establecen un vector con un rango de *clusters* como dato de entrada, con la intensión de evaluar la cohesión de cada agrupación en comparación a otros grupos. El valor que de como resultado por

cada método, son los intervalos a implementar en la clasificación como se observa en la Figura 4.

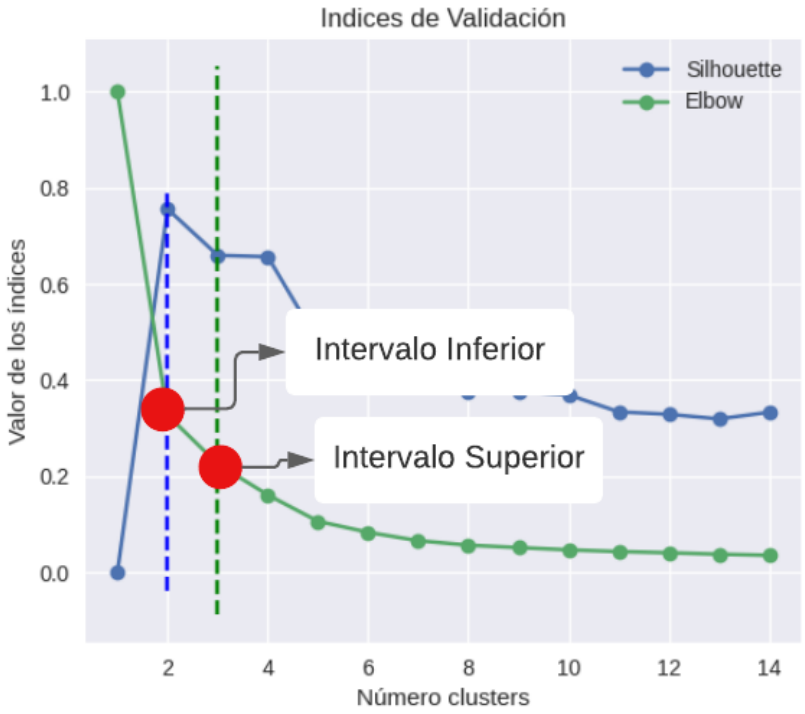


Figura 2: Rango del número óptimo de *clusters*

Es importante destacar que este proceso se aplica a las técnicas de aprendizaje no supervisado, debido a que las técnicas de aprendizaje supervisado utilizan etiquetas. Además, el proceso se realiza individualmente para cada consumidor comercial. Es relevante mencionar que el número óptimo de *clusters* obtenido puede variar para cada método utilizado.

3.3.2. Técnicas de clasificación

Con base en la investigación minuciosa, se elige las siguientes técnicas a implementar, las cuales son:

- **K-means:** se conoce de antemano el rango óptimo del número de conglomerados. Se aplica 10000 iteraciones para converger el algoritmo hacia los *clusters* finales.

- **Clustering Jerárquico:** con el intervalo óptimo del número de agrupaciones conocido. Para esta técnica emergen diversas alternativas en cuanto a criterios de distancia y vinculación susceptibles de implementación. El análisis efectuado en la cita [2] ha arrojado la deducción de que, para lograr una agrupación más homogénea y equilibrada de los perfiles de carga eléctrica (PCE), resulta pertinente la adopción del criterio de vinculación de *Ward* junto con el criterio de distancia DTW. Es por ello, que se utilizo dichos criterios para las diferentes bases de datos.

- **SOM:** se determinaron dos parámetros fundamentales, como el tamaño del mapa de salida, con el fin de obtener una configuración óptima. Luego, se llevó a cabo el entrenamiento del modelo SOM con un número de iteraciones cuidadosamente seleccionado. La elección de estos parámetros se basó en un enfoque de optimización, donde se realizaron pruebas sistemáticas y se utilizaron métodos de ajuste para garantizar un rendimiento óptimo del modelo.

- **FCM:** se aplicó la técnica a diversos conjuntos de datos con el propósito de asignarlos a diferentes agrupaciones de manera difusa, con una matriz que contiene las probabilidades de pertenencia de cada perfil de carga de los *clusters*. La elección adecuada de los parámetros, como el exponente *fuzzy*, tuvo un impacto significativo en la formación de los grupos y en la convergencia del algoritmo. Por esta razón, se buscó el número óptimo de *clusters* que permitiera obtener la mejor agrupación posible.

- **GMM:** se tiene conocimiento del intervalo de posibles conglomerados, dentro del cual se aplica este método probabilístico. La cual consta de tres etapas principales:

    1. El algoritmo elige aleatoriamente las ubicaciones iniciales y las matrices de covarianza.
    2. Con un método se ajusta los parámetros del modelo, es decir, las ubicaciones y matrices de covarianza de las distribuciones gaussianas.
    3. Se aplicara 10 veces este proceso, la cual GMM selecciona el resultado con la mejor verosimilitud.

- **Árbol de Decisiones:** se preparan los datos para la construcción y evaluación del modelo de árbol de decisiones. A continuación, se seleccionan las características relevantes para el modelo y la variable objetivo. Luego, se dividen los datos en conjuntos de entrenamiento y prueba para evaluar el rendimiento del modelo en datos no vistos, siendo los valores del 80 y 20 por ciento. El modelo se prepara utilizando los datos de entrenamiento y se utiliza para predecir las etiquetas de *cluster* en los datos de prueba. La precisión del modelo se calcula comparando las etiquetas predichas con las etiquetas reales. Además de que se coloco un valor óptimo para la semilla que garantiza la división correcta de la toma de decisiones.

3.3.3. Validación de técnicas

A continuación, se expone a detalle el procedimiento metodológico diseñado para la evaluación minuciosa del desempeño de las técnicas previamente mencionadas en la sección anterior, este proceso se puede observar en la Figura 3.

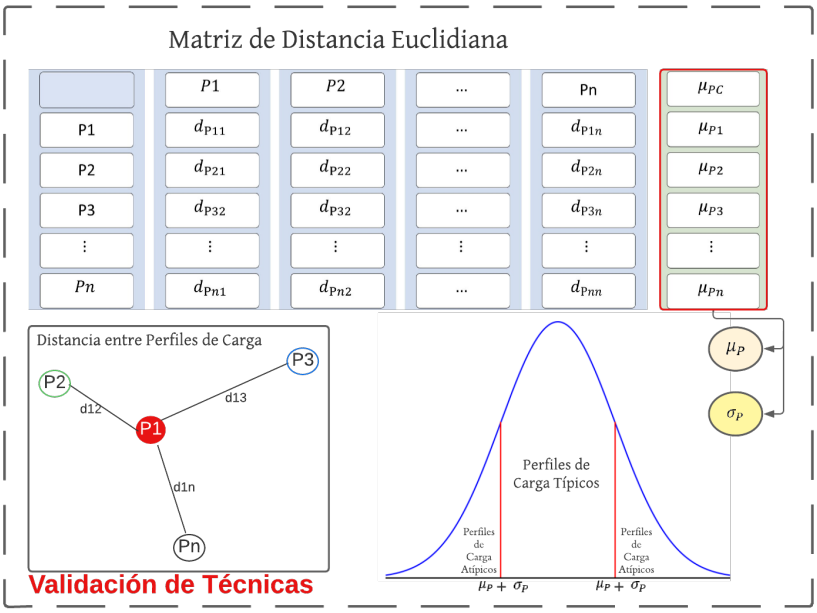


Figura 3: Representación del proceso de validación de las técnicas de clasificación.

El proceso del método consta de 5 etapas, las cuales se explica a continuación:

1. Se calcula las distancias euclidianas para formar una matriz $n \times n$, que refleja la relación entre los perfiles de carga en función de su número.

2. Se obtiene el vector μ_{PC}, que representa el cálculo de la media de cada fila de la matriz que contiene las distancias euclidianas.

3. Se calcula la media μ_P y la desviación estándar σ_P del vector μ_{PC}, donde el punto de simetría de la curva de la distribución de probabilidad es μ_P.

4. Se propone el intervalo de confianza del 68.26 % para la distribución, lo que conduce a la identificación de los puntos de inflexión de la curva en $[\mu_P - \sigma_P, \mu_P + \sigma_P]$. Los perfiles de carga contenidos en el intervalo son valores típicos, en contraste, aquellos que se hallen fuera de dicho rango son identificados como valores atípicos.

5. Finalmente, se representa porcentualmente los valores típicos y atípicos de cada *cluster*.

La metodología detallada es representada en la Figura 3. En adición, se implementa para la evaluación de los resultados obtenidos mediante las diversas técnicas mencionadas, permitiendo un análisis de su desempeño. La técnica que exhiba una menor proporción de valores atípicos será considerada como aquella que mejor se adapta y ajusta de forma óptima a las características fundamentales de la base de datos.

4. Análisis de Resultados

Esta sección se centra en el análisis meticuloso y a la discusión de los resultados obtenidos mediante las técnicas de aprendizaje automático aplicadas al registro de consumo del sector comercial.

Para dos casos de estudio por abordar, se emplean diversas bases de datos extraídos de [69], seleccionados con criterios específicos que atienden a la naturaleza y alcance de la investigación en cuestión. A partir de esta información, se realiza el pre-procesamiento expuesto en la metodológica, una

vez concluida esta etapa, se procede a efectuar la clasificación. Para las técnicas no supervisadas es indispensable conocer el número de conglomerados. Por ende, en la Figura 4, presenta posibles agrupaciones para cada usuario comercial.

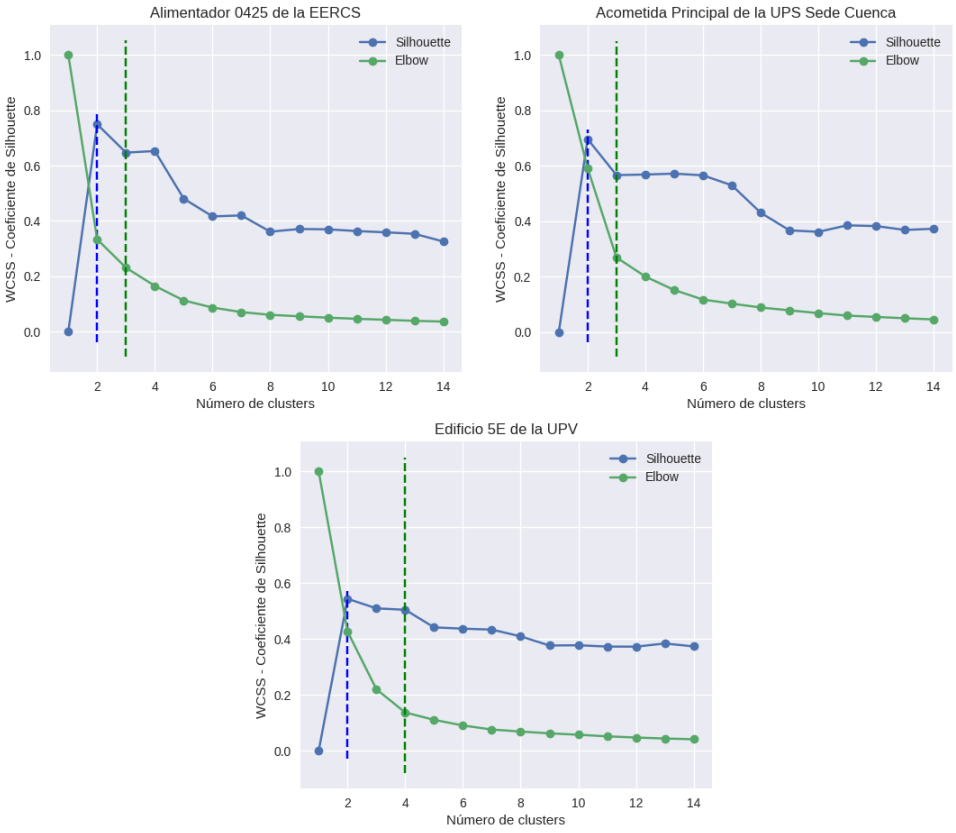


Figura 4: Número de agrupaciones para la base de datos del sector comercial.

En los tres casos de estudio, los resultados de clasificación proporcionaron una comprensión más profunda de las características de cada conjunto de datos y permite evaluar el desempeño de la técnica implementada.

4.1. Caso de estudio 1: Alimentador 0425 de la Empresa Eléctrica Regional Centro Sur (EERCS)

En el presente escenario, se interpreta los perfiles de carga eléctrica correspondientes al alimentador 0425 de la Subestación 4, perteneciente a la EERCS. La Subestación tiene 44.5 MVA de potencia instalada y una tensión de 22kV que alimenta a una zona industrial. Los registros de mediciones son desde el 1 de enero a las 0:00 horas hasta las 23:45 del 31 de diciembre del 2017, y constan de un total de 35041 registros correspondiente a 365 días [69].

La técnica FCM evidencia una destacada capacidad para agrupar y distribuir de los PCE, en comparación con otros algoritmos. Esta se comprueba en el bajo porcentaje de PCE atípicos en la segmentación, lo que se traduce en la formación de grupos homogéneos en términos de patrones de consumo eléctrico.

El conjunto de datos ha sido segregado en tres agrupaciones, la cual se puede visualizar en la Figura 5. Cada agrupación tiene un porcentaje significativo de valores atípicos, el *cluster* 1 alcanzando un 35.48 % del total de datos en dicho grupo, el *cluster* 2, tiene el 1.97 % y el *cluster* 3, se registra un porcentaje del 8.75 %.

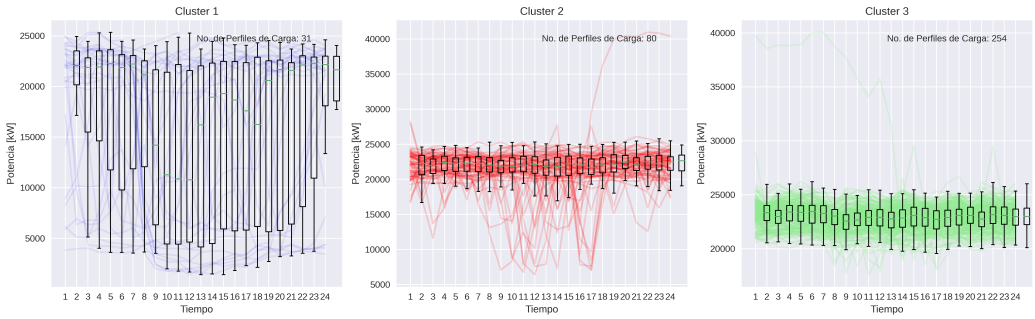


Figura 5: Clusters del Alimentador 0425 perteneciente al EERCS

Es posible verificar visualmente el porcentaje de valores atípicos en cada *cluster*, lo cual se manifiesta en la presencia de algunas curvas que se extienden más allá del rango representado por el diagrama de bigote.

En el análisis de cada agrupación, se pueden observar patrones distintivos en el comportamiento de los PCE. Estas características permite identificar y

comprender de mejor manera las particularidades y variaciones presentes en cada grupo, se puede visualizar respecto al día en la Figura 6.

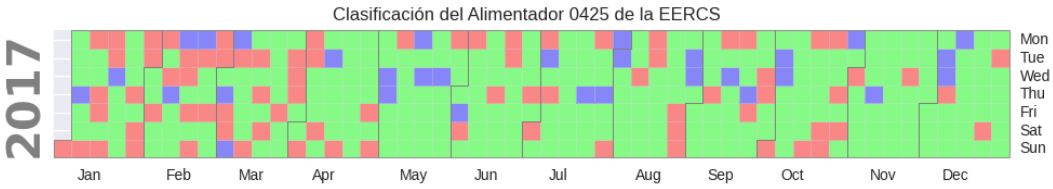


Figura 6: Calendario de calor del Alimentador0425 perteneciente al EERCS

Durante el año 2017, se identifica un total de 31 días de mayor demanda de consumo energético, también conocidos como días anómalos, los cuales se encuentran agrupados en el *Cluster* 1. Es importante destacar que estos patrones se presentan de manera recurrente en todos los meses del año. El consumo eléctrico es directamente proporcional a la producción de la empresa.

Además, se debe considerar aquellos PCE en los que se observan variaciones considerables de la demanda a lo largo de un día. Estas fluctuaciones pueden estar relacionadas con hábitos del personal, meses de producción, días laborables y no laborables, estas características están agrupadas en el *Cluster* 2. Por otro lado, el *Cluster* 3 representa un comportamiento común con PCE típicos, son aquellos días donde la empresa no presenta cambios en su producción, hábitos, etc.

La Figura 7, ofrece una perspectiva extensa de las agrupaciones, lo que permite visualizar la dispersión de los PCE en cada conjunto. Además, esta representación gráfica posibilita la identificación de los perfiles anómalos, es decir, aquellos datos que se encuentran significativamente alejados del centroide del grupo,

En la Tabla 2 se presentan los resultados para cada uno de los métodos de clasificación utilizados. Al examinar la tabla, el método "*Fuzzy C-means*", cuyas desviaciones estándar exhibe valores bajos, Este indicativo, expresa una marcada tendencia a la cohesión, y mínima dispersión intragrupal. Además, siendo estos resultados, los que proporcionan una visión detallada y comparativa de cómo se agrupan y distribuyen los patrones de consumo de energía en cada método de clasificación.

Tabla 2: Tabla de los factores de forma de cada técnica.

Técnicas de clasificación	Cluster	Alimentador 0425			
		P_{μ} [kW]	P_{min} [kW]	P_{max} [kW]	P_{σ} [kW]
K-means	1	22786.55	14886.0	40047.0	154.04
	2	14281.62	1402.0	25339.0	612.78
	3	21499.67	5174.0	40952.0	1163.41
Agrupación Jerárquica	1	22692.39	7748.0	29534.0	183.49
	2	19704.83	3763.0	25279.0	1573.28
	3	13562.34	1402.0	40952.0	747.81
Árbol de Decisiones	1	17410.41	3435.0	25279.0	715.63
	2	11624.54	2725.0	23660.0	0.0
	3	10308.73	1402.0	25339.0	2255.25
	4	8308.02	1413.0	23477.0	3125.43
	5	22578.47	4977.0	29534.0	365.15
	6	28120.62	13498.0	40952.0	0.0
	7	30201.04	20869.0	40047.0	0.0
Mezcla Gaussiana	1	22748.91	15436.0	26665.0	71.54
	2	17823.12	1402.0	25339.0	923.33
	3	22846.27	10143.0	40952.0	403.76
Fuzzy C-means	1	21826.46	6383.0	40952.0	715.36
	2	22949.47	17594.0	40047.0	161.59
	3	16025.52	1402.0	25339.0	761.63
Mapa Autoorganizado	1	15257.38	1402.0	25339.0	687.16
	2	21145.32	5174.0	27713.0	1672.46
	3	22789.90	12989.0	40952.0	103.15

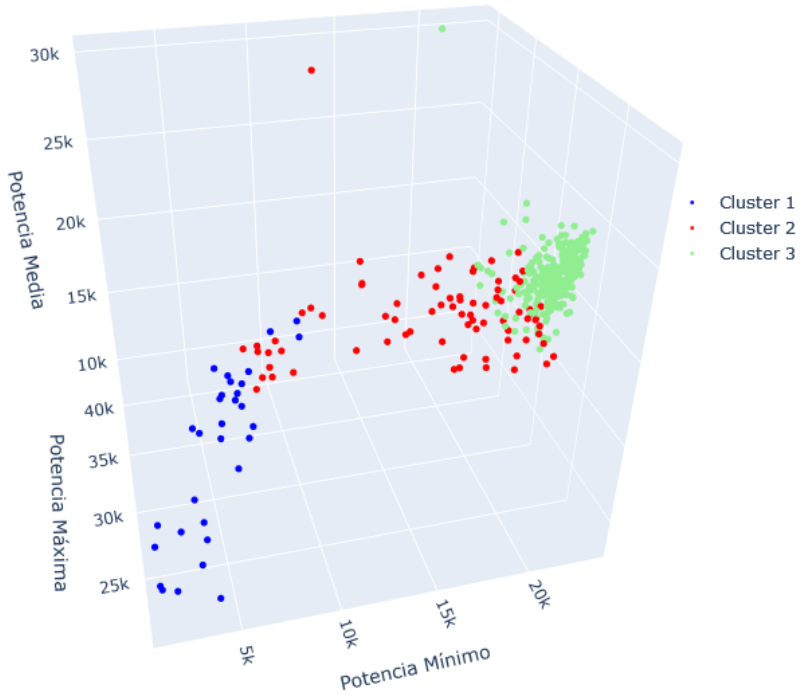


Figura 7: Características de las agrupaciones del Alimentador 0425

4.2. Caso de estudio 2: Edificio 5E de la Universidad Politécnica de Valencia (UPV) en España

Los registros de consumo eléctrico adquiridos corresponden a la UPV, edificio 5E (España), inician a las 0:00 horas del 1 de julio del año 2014 hasta las 18:00 horas del 28 de noviembre del 2016, con un total de 84650 registros [70] [71].

La categorización para este escenario implica un desafío considerable debido a su vasta recopilación de mediciones. Entre las diversas metodologías empleadas, el SOM se ha destacado como la más idónea en términos de adaptación y segregación a los cambios bruscos de la demanda eléctrica demostrando su eficacia para la clasificación.

La Figura 8 ilustra las agrupaciones finales. No obstante, se aprecia un porcentaje de valores atípicos intragrupo. En el *cluster 1*, estos PCE constituyen de 17.36 % del total de datos en dicho grupo. En el *cluster 2*, aumenta a un 21.74 %. Mientras que, en el *cluster 3* es el 18.57 % y el *cluster 4* con

solo el valor de 9.68 %.

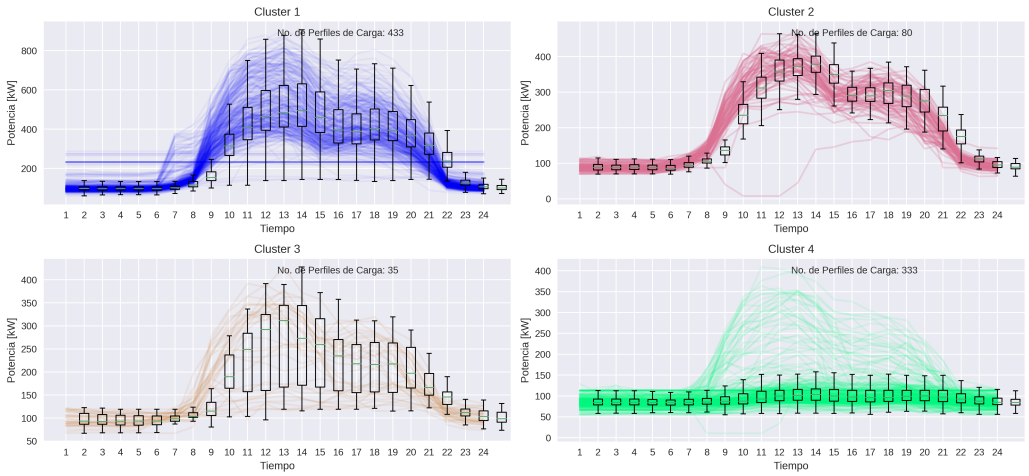


Figura 8: Clusters del Edificio 5E.

Los cuatro grupos presentan una curva de forma similar, caracterizada por un aumento de consumo desde las primeras horas de la mañana, hasta alcanzar un pico en algún momento del día. No obstante, el *cluster 4* muestra este mismo patrón de consumo únicamente con ciertos valores atípicos, mientras que en la mayoría de los perfiles su consumo se mantiene dentro de rangos normales. De manera adicional, al analizar la Figura 8, se puede notar que las curvas presentan un comportamiento más ajustado, lo que indica una mayor uniformidad en las agrupaciones. Esta cohesión en las agrupaciones facilita la identificación de patrones y tendencias relevantes para el análisis.

Los comportamientos de consumo presentan singularidades a lo largo de cada año, teniendo una clasificación más autentica en cuanto a los consumos, distribuyéndolos de manera más precisa, lo cual podría atribuirse a las actividades llevadas a cabo en el recinto. La Figura 9 otorga una percepción más precisa de estas fluctuaciones, posibilitando así una planificación anual más efectiva del consumo eléctrico.

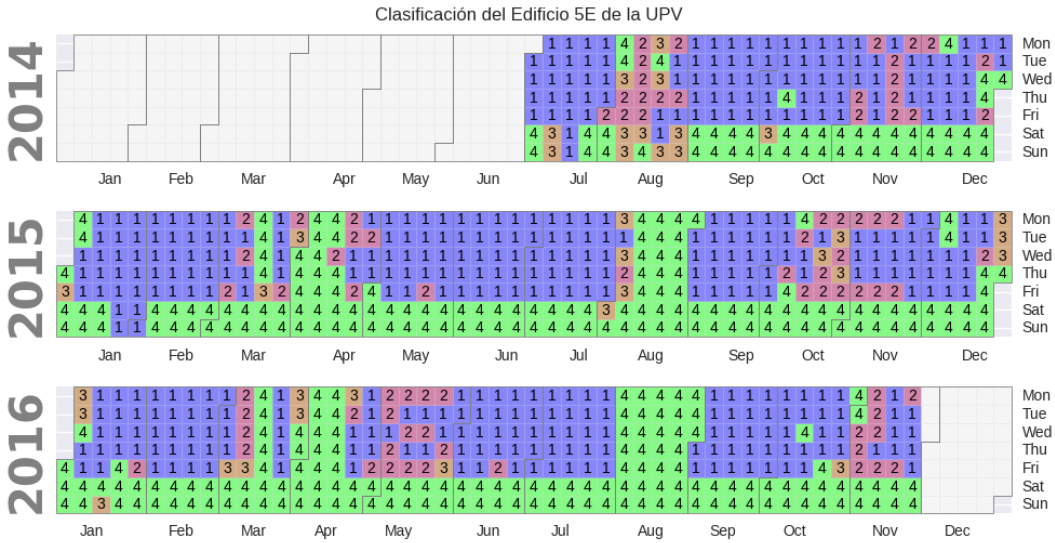


Figura 9: Calendario de calor del Edificio 5E.

Durante el análisis, se han identificado resultados claros debido a la buena organización y el bajo número de datos atípicos fuera de rango. Los resultados revelaron que los *clusters* pueden clasificarse según sus actividades diarias y feriados. El *cluster 2* y el *cluster 3* representan agrupaciones distintas que separan los días festivos o valores atípicos. El *cluster 3* abarca los fines de semana, con ciertas fechas excluidas, mientras que el *cluster 2* representa el perfil de carga típico presente en la mayoría de los días de la semana. Cabe destacar que, al tratarse de una institución de educación superior, es posible que sus costumbres, horarios de funcionamiento y fechas de feriados afecten la demanda de energía de manera particular. Estos eventos y ciclos de actividad pueden generar patrones de consumo específicos, lo que a su vez explica la aparición de los patrones anómalos en los datos recopilados.

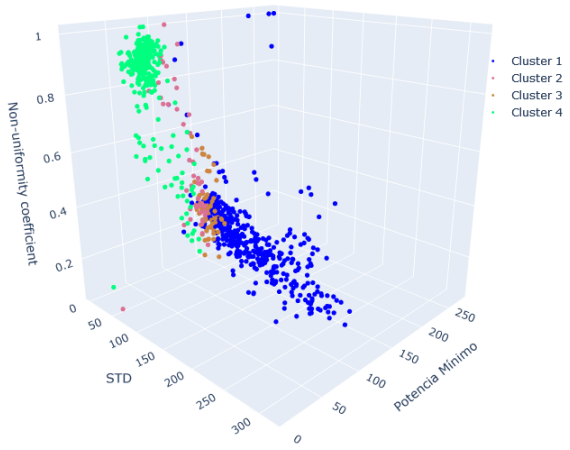


Figura 10: Características de las agrupaciones del Edificio 5E.

En la Figura 10, se aprecia la distinción de las características de los PCE. Los factores específicos, como la potencia mínima, el coeficiente de no uniformidad y la desviación estándar, son seleccionados debido a que demostraron una agrupación más significativa y reveladora en comparación con el enfoque basado únicamente en los límites de potencia.

En la Tabla 3, se observa que el método de SOM exhibe desviaciones estándar con valores relativamente bajas, indicando una notable inclinación hacia una mayor cohesión y una reducida dispersión intragrupal entre los conglomerados identificados.

Tabla 3: Tabla de los factores de forma de cada técnica.

Técnicas de clasificación	Cluster	Edificio 5E			
		P_μ [kW]	P_{min} [kW]	P_{max} [kW]	P_σ [kW]
K-means	1	208.43	7.84	649.56	41.69
	2	316.299	59.16	878.8	34.03
	3	316.20	68.32	906.52	67.42
	4	95.29	10.56	351.52	16.09
Agrupación Jerárquica	1	294.86	7.84	878.8	43.73
	2	178.163	10.56	572.88	46.28
	3	310.70	68.32	906.52	70.37
	4	87.31	53.83	216.16	4.83
Árbol de Decisiones	1	83.53	54.28	733.96	31.31
	2	193.59	7.84	811.96	68.81
	3	347.02	72.96	906.52	57.31
	4	341.47	166.52	670.36	60.30
Mezcla Gaussiana	1	93.45	68.48	160.64	0.61
	2	260.16	7.84	878.8	79.83
	3	314.43	68.32	906.52	64.60
	4	206.36	53.83	792.31	52.20
Fuzzy C-means	1	314.49	59.16	878.8	35.59
	2	207.89	7.84	622.72	41.22
	3	315.84	68.32	906.52	67.78
	4	95.29	10.56	351.52	16.09
Mapa Autoorganizado	1	279.74	59.16	906.52	42.24
	2	151.42	7.84	427.44	37.27
	3	206.71	62.96	475.4	14.66
	4	98.43	10.56	408.72	19.67

4.3. Caso de estudio 3: Acometida Principal de la Universidad Politécnica Salesiana (UPS) de la Sede de Cuenca

Los datos de la Acometida Principal de la UPS, están comprendidas con un total de 42913 mediciones que empieza desde el 8 de marzo de 2017 a las 00:00 horas hasta el 27 de mayo de 2018 a las 23:45 horas, con 446 PCE.

En el proceso de validación, se pudo constatar que el SOM demostró un rendimiento excepcionalmente preciso para identificar peculiaridades, facilitar la visualización y segregación de grupos con características comunes. Esta técnica, al igual que los casos anteriores se evidencia porcentajes favorables de valores típicos asignados adecuadamente a sus respectivos grupos.

La segmentación resulto en cuatro conglomerados, la cual se observa en la Figura 11. Estas agrupaciones muestran una tendencia uniforme con menos valores atípicos, con porcentajes que oscilan entre 8 y 9 % para los *cluster 1, 2* y *3*, variando únicamente en valores decimales. Sin embargo, el *cluster 4* exhibe un cambio significativo en los valores atípicos, presentando un porcentaje del 42.59 %.

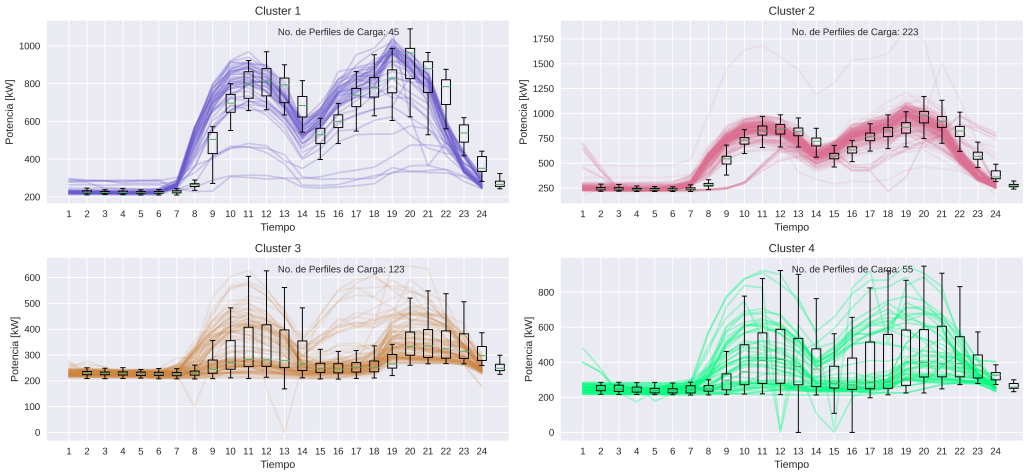


Figura 11: *Clusters* de la base de datos de la Acometida Principal

Los conglomerados presentan particularidades comunes, pero se pueden observar valores irregulares que se encuentran fuera de rango del diagrama de bigote. La Figura 12, destaca una caracteriza en los PCE a lo largo del tiempo, esto debido a la identificación de patrones coherentes brindando una comprensión de los hábitos de consumo.

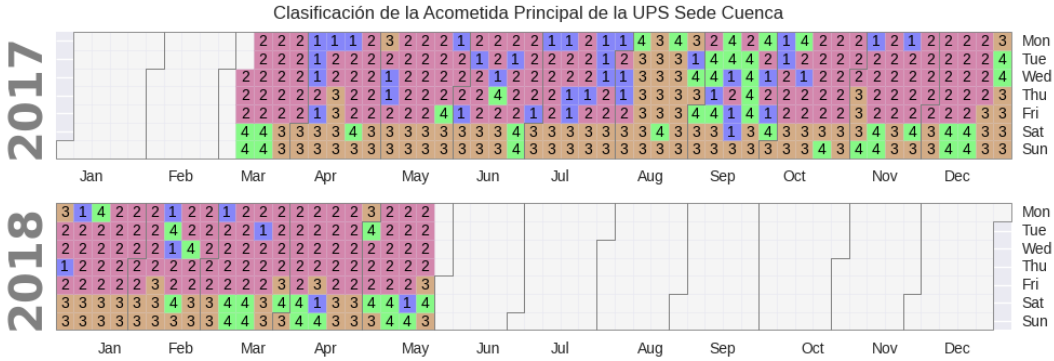


Figura 12: Calendario de calor de la Acometida Principal

En particular, el *cluster 3* se destaca por su presencia significativa durante los fines de semana, pero con algunas fechas excluidas, lo cual sugiere que los patrones de consumo energético en esos días son distintos a los días de semana. En adición, es importante enfatizar que en el mes de agosto el personal laboral como los estudiantes se encuentran en un período vacacional. Por ende, se caracteriza por tener un consumo energético bajo.

El *cluster 2* corresponde a días hábiles de lunes a viernes con un consumo típico, en las cuales se llevan a cabo actividades laborales y académicas. Por otro lado, los *clusters 1* y *4* representan los días no laborables, festivos y otros eventos, presentando días irregulares de consumo alto.

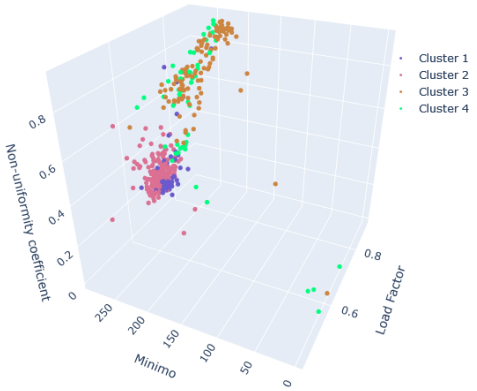


Figura 13: Características de las agrupaciones de la Acometida Principal de la UPS

La Figura 13 muestra los conglomerados con sus características, destacando la dispersión de cada PCE con respecto a la potencia mínima, coeficiente de no uniformidad y factor de forma. Estos parámetros son fundamentales para proporcionar información clave sobre el desempeño y la eficiencia de la técnica empleada.

En la Tabla 4 se presentan los resultados y la eficiencia de todos los algoritmos implementados, resaltando la distribución de potencia respecto al número de agrupaciones.

Tabla 4: Resultado de los factores de forma

Técnicas de clasificación	Cluster	Acometida Principal			
		P_{μ} [kW]	P_{min} [kW]	P_{max} [kW]	P_{σ} [kW]
K-means	1	568.82	179.2	1212.32	27.46
	2	292.78	0.0	775.52	42.70
	3	543.71	193.28	1850.88	92.03
	4	339.57	0.0	1044.16	68.34
Agrupación Jerárquica	1	565.05	179.2	1505.6	33.76
	2	544.22	193.28	1850.88	103.02
	3	298.49	0.0	948.32	45.03
Árbol de Decisiones	1	251.64	98.08	466.72	11.00
	2	483.47	0.0	1172.16	88.85
	3	606.65	239.84	1505.6	49.08
	4	1043.05	193.28	1850.88	0.0
Mezcla Gaussiana	1	294.37	178.72	736.48	40.73
	2	573.74	209.76	1140.8	23.30
	3	458.04	0.0	1505.6	91.61
	4	544.22	193.28	1850.88	103.02
Fuzzy C-means	1	451.82	0.0	1115.52	99.06
	2	291.98	0.0	775.52	42.35
	3	572.24	221.76	1212.32	25.88
	4	543.71	193.28	1850.88	92.03
Mapa Autoorganizado	1	527.09	179.2	1038.4	57.62
	2	568.43	193.28	1850.88	42.38
	3	283.23	0.0	775.52	34.07
	4	339.73	0.0	948.32	64.53

5. Conclusiones

La capacidad de discernimiento al momento de elegir el algoritmo resulta complejo, este trabajo contribuye con un método para evaluar el desempeño de las técnicas de *clustering*, con el objetivo de identificar los grupos con menor incidencia de valores atípicos.

Los algoritmos son capaces de proporcionar una segmentación aceptable. Sin embargo, la técnica que sobresale en la clasificación es el SOM, logra adaptarse de manera efectiva a las particularidades de la base de datos de las instituciones educativas, a pesar de la diversidad entre los dos países.

En el análisis llevado a cabo, se encuentra que la técnica del árbol de decisiones no demuestra ser efectiva para la clasificación de datos. Estos resultados negativos podrían explicar la escasa presencia de literatura destacada que respalde la utilidad de esta técnica en dicho campo.

Con base al estudio del edificio 5E, se ha demostrado que ciertas curvas anómalas en los *clusters* pueden ser mejor integradas con otros grupos, lo cual se atribuye a sus características distintivas. Sin embargo, en el caso de la clasificación de la Acometida Principal de la UPS junto con la del edificio 5E de la UPV, se observa una similitud en el comportamiento anual. Esta coincidencia podría atribuirse a su función común, dado que ambas pertenecen a una institución de educación superior.

Referencias

[1] F. Jiang, Q. Zhu, J. Yang, G. Chen, and T. Tian, "Clustering-based interval prediction of electric load using multi-objective pathfinder algorithm and elman neural network," *Applied Soft Computing*, vol. 129, p. 109602, 2022.

[2] M. Gunsay, C. Bilir, and G. Poyrazoglu, "Load profile segmentation for electricity market settlement," in *2020 17th International Conference on the European Energy Market (EEM)*. IEEE, 2020, pp. 1–5.

[3] A. Mutanen, M. Ruska, S. Repo, and P. Jarventausta, "Customer classification and load profiling method for distribution systems," *IEEE Transactions on Power Delivery*, vol. 26, no. 3, pp. 1755–1763, 2011.

[4] B. Mahesh, "Machine learning algorithms-a review," *International Journal of Science and Research (IJSR).[Internet]*, vol. 9, no. 1, pp. 381–386, 2020.

[5] G. Bonaccorso, *Machine learning algorithms*. Packt Publishing Ltd, 2017.

[6] S. Aghabozorgi, A. S. Shirkhorshidi, and T. Y. Wah, "Time-series clustering–a decade review," *Information systems*, vol. 53, pp. 16–38, 2015.

[7] C. A. U. Hassan, M. S. Khan, and M. A. Shah, "Comparison of machine learning algorithms in data classification," in *2018 24th International Conference on Automation and Computing (ICAC)*. IEEE, 2018, pp. 1–6.

[8] S. Lloyd, "Least squares quantization in pcm," *IEEE transactions on information theory*, vol. 28, no. 2, pp. 129–137, 1982.

[9] L. García-Santander, J. San Martín-Ayala, F. Ulloa-Vásquez, D. Carrizo, V. Esparza, J. Rohten, and C. Mejias, "Classification of behavior profiles for non-residential customers considering the variable of electrical energy consumption: Case study—saesa group sa company," *Energies*, vol. 15, no. 18, p. 6634, 2022.

[10] G. Chicco, R. Napoli, and F. Piglione, "Comparisons among clustering techniques for electricity customer classification," *IEEE Transactions on power systems*, vol. 21, no. 2, pp. 933–940, 2006.

[11] M.-A. Milton, C.-O. Pedro, S.-G. Xavier, and E.-E. Guillermo, "Characterization and classification of daily electricity consumption profiles: shape factors and k-means clustering technique," in *E3S Web of Conferences*, vol. 64. EDP Sciences, 2018, p. 08004.

[12] L. F. Ugarte, E. Lacusta Jr, M. C. de Almeida *et al.*, "Characterization of load curves in a real distribution system based on k-means algorithm with time-series data," in *Congresso Brasileiro de Automática-CBA*, vol. 2, no. 1, 2020.

[13] Z. Zhao, J. Wang, and Y. Liu, "User electricity behavior analysis based on k-means plus clustering algorithm," in *2017 International Conference*

on Computer Technology, Electronics and Communication (ICCTEC). IEEE, 2017, pp. 484–487.

[14] J. C. Dunn, "A fuzzy relative of the isodata process and its use in detecting compact well-separated clusters," 1973.

[15] J. Arora, K. Khatter, and M. Tushir, "Fuzzy c-means clustering strategies: A review of distance measures," *Software Engineering: Proceedings of CSI 2015*, pp. 153–162, 2019.

[16] S. Deng, "Clustering with fuzzy c-means and common challenges," in *Journal of Physics: Conference Series*, vol. 1453, no. 1. IOP Publishing, 2020, p. 012137.

[17] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the royal statistical society: series B (methodological)*, vol. 39, no. 1, pp. 1–22, 1977.

[18] J. Janouek, P. Gajdo, M. Radeckỳ, and V. Snáel, "Gaussian mixture model cluster forest," in *2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA)*. IEEE, 2015, pp. 1019–1023.

[19] P. Sarang, "Gaussian mixture model: A probabilistic clustering model for datasets with mixture of gaussian blobs," in *Thinking Data Science: A Data Science Practitioner's Guide*. Springer, 2023, pp. 197–207.

[20] L. Scrucca, M. Fop, T. B. Murphy, and A. E. Raftery, "mclust 5: clustering, classification and density estimation using gaussian finite mixture models," *The R journal*, vol. 8, no. 1, p. 289, 2016.

[21] Y. Li, J. Zhang, P. Tang, and L. Tian, "Clustering in the wireless channel with a power weighted statistical mixture model in indoor scenario," *China Communications*, vol. 16, no. 7, pp. 83–95, 2019.

[22] T. Kohonen, "Self-organized formation of topologically correct feature maps," *Biological cybernetics*, vol. 43, no. 1, pp. 59–69, 1982.

[23] S.-l. Yang, C. Shen *et al.*, "A review of electric load classification in smart grid environment," *Renewable and Sustainable Energy Reviews*, vol. 24, pp. 103–110, 2013.

[24] A. Rajabi, M. Eskandari, M. J. Ghadi, L. Li, J. Zhang, and P. Siano, "A comparative study of clustering techniques for electrical load pattern segmentation," *Renewable and Sustainable Energy Reviews*, vol. 120, p. 109628, 2020.

[25] O. Akman, T. Comar, D. Hrozencik, and J. Gonzales, "Data clustering and self-organizing maps in biology," in *Algebraic and Combinatorial Computational Biology*. Elsevier, 2019, pp. 351–374.

[26] D. Roberts and S. F. Brown, "Identifying calendar-correlated day-ahead price profile clusters for enhanced energy storage scheduling," *Energy Reports*, vol. 6, pp. 35–42, 2020.

[27] S. Cen, J. H. Yoo, and C. G. Lim, "Electricity pattern analysis by clustering domestic load profiles using discrete wavelet transform," *Energies*, vol. 15, no. 4, p. 1350, 2022.

[28] G. C. Mecatti, M. C. F. Messias, and P. de Oliveira Carvalho, "Lipidomic profile and candidate biomarkers in septic patients," *Lipids in Health and Disease*, vol. 19, no. 1, pp. 1–9, 2020.

[29] J. L. Viegas, S. M. Vieira, R. Melício, V. Mendes, and J. M. Sousa, "Classification of new electricity customers based on surveys and smart metering data," *Energy*, vol. 107, pp. 804–817, 2016.

[30] A. Singh, N. Thakur, and A. Sharma, "A review of supervised machine learning algorithms," in *2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom)*. Ieee, 2016, pp. 1310–1315.

[31] F. Osisanwo, J. Akinsola, O. Awodele, J. Hinmikaiye, O. Olakanmi, J. Akinjobi *et al.*, "Supervised machine learning algorithms: classification and comparison," *International Journal of Computer Trends and Technology (IJCTT)*, vol. 48, no. 3, pp. 128–138, 2017.

[32] M. Shafiq, X. Yu, A. A. Laghari, L. Yao, N. K. Karn, and F. Abdessamia, "Network traffic classification techniques and comparative analysis using machine learning algorithms," in *2016 2nd IEEE International Conference on Computer and Communications (ICCC)*. IEEE, 2016, pp. 2451–2455.

[33] X. Serrano-Guerrero, R. Prieto-Galarza, E. Huilcatanda, J. Cabrera-Zeas, and G. Escrivá-Escrivá, "Election of variables and short-term forecasting of electricity demand based on backpropagation artificial neural networks," in *2017 IEEE International Autumn Meeting on Power, Electronics and Computing (ROPEC)*. IEEE, 2017, pp. 1–5.

[34] S. E. Hammami, H. Afifi, M. Marot, and V. Gauthier, "Network planning tool based on network classification and load prediction," in *2016 IEEE Wireless Communications and Networking Conference*. IEEE, 2016, pp. 1–6.

[35] K. Li, X. Cao, X. Ge, F. Wang, X. Lu, M. Shi, R. Yin, Z. Mi, and S. Chang, "Meta-heuristic optimization-based two-stage residential load pattern clustering approach considering intra-cluster compactness and inter-cluster separation," *IEEE Transactions on Industry Applications*, vol. 56, no. 4, pp. 3375–3384, 2020.

[36] C. Zhang and R. Li, "A novel closed-loop clustering algorithm for hierarchical load forecasting," *IEEE Transactions on Smart Grid*, vol. 12, no. 1, pp. 432–441, 2020.

[37] A. Satre-Meloy, M. Diakonova, and P. Grünewald, "Cluster analysis and prediction of residential peak demand profiles using occupant activity data," *Applied Energy*, vol. 260, p. 114246, 2020.

[38] M. R. Haq and Z. Ni, "Classification of electricity load profile data and the prediction of load demand variability," in *2019 IEEE International Conference on Electro Information Technology (EIT)*. IEEE, 2019, pp. 304–309.

[39] S. Ramos, J. Soares, S. S. Cembranel, I. Tavares, Z. Foroozandeh, Z. Vale, and R. Fernandes, "Data mining techniques for electricity customer characterization," *Procedia Computer Science*, vol. 186, pp. 475–488, 2021.

[40] C. Nichiforov and M. Alamaniotis, "Load-based classification of academic buildings using matrix profile and supervised learning," in *2021 IEEE PES Innovative Smart Grid Technologies Europe (ISGT Europe)*. IEEE, 2021, pp. 01–05.

[41] I. E. Agency. (2022) Electricity market report - july 2022. [Online]. Available: https://www.iea.org/reports/electricity-market-report-july-2022

[42] J. Yang, J. Zhao, F. Wen, and Z. Dong, "A model of customizing electricity retail prices based on load profile clustering analysis," *IEEE Transactions on Smart Grid*, vol. 10, no. 3, pp. 3374–3386, 2018.

[43] S. A. Azad, A. S. Ali, and P. Wolfs, "Identification of typical load profiles using k-means clustering algorithm," in *Asia-Pacific World Congress on Computer Science and Engineering*. IEEE, 2014, pp. 1–6.

[44] T. Zhang, G. Zhang, J. Lu, X. Feng, and W. Yang, "A new index and classification approach for load pattern analysis of large electricity customers," *IEEE Transactions on Power Systems*, vol. 27, no. 1, pp. 153–160, 2011.

[45] J. Wang and X. Su, "An improved k-means clustering algorithm," in *2011 IEEE 3rd international conference on communication software and networks*. IEEE, 2011, pp. 44–46.

[46] F. Nielsen and F. Nielsen, "Hierarchical clustering," *Introduction to HPC with MPI for Data Science*, pp. 195–211, 2016.

[47] F. Murtagh and P. Contreras, "Methods of hierarchical clustering," *arXiv preprint arXiv:1105.0121*, 2011.

[48] J. Nayak, B. Naik, and H. Behera, "Fuzzy c-means (fcm) clustering algorithm: a decade review from 2000 to 2014," in *Computational Intelligence in Data Mining-Volume 2: Proceedings of the International Conference on CIDM, 20-21 December 2014*. Springer, 2015, pp. 133–149.

[49] C. Ramos-Palencia, D. Mújica-Vargas, and M. Mejía-Lavalle, "Análisis comparativo de la calidad de agrupamiento del algoritmo intuitionistic fuzzy c-means." *Res. Comput. Sci.*, vol. 149, no. 8, pp. 365–378, 2020.

[50] J. C. Ruiz González, "Fuzzy c-means: Distribuido en la nube," 2017.

[51] M. Albornoz and C. Alfaraz, "Redes de conocimiento: construcción, dinámica y gestión," 2016.

[52] T. Kohonen, S. Kaski, K. Lagus, J. Salojarvi, J. Honkela, V. Paatero, and A. Saarela, "Self organization of a massive document collection," *IEEE transactions on neural networks*, vol. 11, no. 3, pp. 574–585, 2000.

[53] B. De Ville, "Decision trees," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 5, no. 6, pp. 448–455, 2013.

[54] S. B. Kotsiantis, "Decision trees: a recent overview," *Artificial Intelligence Review*, vol. 39, pp. 261–283, 2013.

[55] M. M. Selvam, R. Gnanadass, and N. P. Padhy, "Fuzzy based clustering of smart meter data using real power and thd patterns," *Energy Procedia*, vol. 117, pp. 401–408, 2017.

[56] J. I. Marín Hurtado and A. López Parrado, "Implementación en dsps de técnicas de comunicación basadas en wavelets." Armenia, 2004.

[57] P. Chaovalit, A. Gangopadhyay, G. Karabatis, and Z. Chen, "Discrete wavelet transform-based time series analysis and mining," *ACM Computing Surveys (CSUR)*, vol. 43, no. 2, pp. 1–37, 2011.

[58] K. Li, Z. Ma, D. Robinson, and J. Ma, "Identification of typical building daily electricity usage profiles using gaussian mixture model-based clustering and hierarchical clustering," *Applied energy*, vol. 231, pp. 331–342, 2018.

[59] H. Abdi, L. J. Williams *et al.*, "Normalizing data," *Encyclopedia of research design*, vol. 1, 2010.

[60] H. Humaira and R. Rasyidah, "Determining the appropiate cluster number using elbow method for k-means algorithm," in *Proceedings of the 2nd Workshop on Multidisciplinary and Applications (WMA) 2018, 24-25 January 2018, Padang, Indonesia*, 2020.

[61] J. Zeng, J. Wang, L. Guo, G. Fan, K. Zhang, and G. Gui, "Cell scene division and visualization based on autoencoder and k-means algorithm," *IEEE Access*, vol. 7, pp. 165 217–165 225, 2019.

[62] D. Marutho, S. H. Handaka, E. Wijaya *et al.*, "The determination of cluster number at k-mean using elbow method and purity evaluation on headline news," in *2018 international seminar on application for technology of information and communication*. IEEE, 2018, pp. 533–538.

[63] K. R. Shahapure and C. Nicholas, "Cluster quality analysis using silhouette score," in *2020 IEEE 7th international conference on data science and advanced analytics (DSAA)*. IEEE, 2020, pp. 747–748.

[64] S. Yilmaz, J. Chambers, and M. K. Patel, "Comparison of clustering approaches for domestic electricity load profile characterisation-implications for demand side management," *Energy*, vol. 180, pp. 665–677, 2019.

[65] S. Pértegas Díaz and S. Pita Fernández, "La distribución normal," *Cad Aten Primaria*, vol. 8, pp. 268–274, 2001.

[66] M. Bourdeau, P. Basset, S. Beauchêne, D. Da Silva, T. Guiot, D. Werner, and E. Nefzaoui, "Classification of daily electric load profiles of non-residential buildings," *Energy and Buildings*, vol. 233, p. 110670, 2021.

[67] R. Li, Z. Wang, C. Gu, F. Li, and H. Wu, "A novel time-of-use tariff design based on gaussian mixture model," *Applied energy*, vol. 162, pp. 1530–1536, 2016.

[68] L. Xiqiao, W. Wanlu, Z. Bo, Y. Xu, H. Shuai, and Q. Lijuan, "Analysis of large-scale electricity load profile using clustering method," in *2018 IEEE 15th International Conference on Networking, Sensing and Control (ICNSC)*. IEEE, 2018, pp. 1–5.

[69] J. X. Serrano Guerrero, "Caracterización de la demanda de energía mediante patrones estocásticos en las redes eléctricas inteligentes," Ph.D. dissertation, Universitat Politècnica de València, 2020.

[70] G. Escrivá Escrivá, "Nuevas herramientas para facilitar la respuesta activa de consumidores en mercados eléctricos liberalizados: Implementación y retribución," Ph.D. dissertation, Universitat Politècnica de València, 2010.

[71] X. Serrano-Guerrero, G. Escrivá-Escrivá, and C. Roldán-Blay, "Statistical methodology to assess changes in the electrical consumption profile of buildings," *Energy and Buildings*, vol. 164, pp. 99–108, 2018.