

ANÁLISIS DIAGNÓSTICO DE DATOS AMI A PARTIR DE TÉCNICAS DE APRENDIZAJE NO-SUPERVISADO SOBRE SERIES TEMPORALES

Edier Alexis Murillo Narvaez



Escuela de Ingeniería Eléctrica y Electrónica
Facultad de Ingeniería
Programa de Pregrado en Ingeniería Electrónica
Junio
2025

ANÁLISIS DIAGNÓSTICO DE DATOS AMI A PARTIR DE TÉCNICAS DE APRENDIZAJE NO-SUPERVISADO SOBRE SERIES TEMPORALES

Edier Alexis Murillo Narvaez

Documento presentado al
Programa de Pregrado en Ingeniería Electrónica
como requisito para optar por el título de
Ingeniero Electrónico

Director: Wilfredo Alfonso-Morales, PhD.

Codirector: Wilmar Alexander Torres López, MSc.



Escuela de Ingeniería Eléctrica y Electrónica
Facultad de Ingeniería
Programa de Pregrado en Ingeniería Electrónica
Junio
2025

Aceptación

Nota de aceptación:

Firma del Jurado

Firma del Jurado

Wilfredo Alfonso-Morales, PhD., Director

Wilmar Alexander Torres López, MSc., Codirector

Cali, Junio de 2025

Dedicatoria

Dedico este trabajo a mi familia, quienes me han acompañado desde el primer momento y han creído en mí incondicionalmente. Gracias por ser mi fuente de fe, mi motor diario y por darme la fuerza necesaria para continuar. Su valentía, perseverancia y deseo de salir adelante se han convertido en una inspiración constante, guiando mi crecimiento tanto personal como profesional. Hoy soy quien soy gracias a ustedes.

A mis padres:

Flor Elcy Narvaez Mora

Edier Murillo Granda

A mis hermanos:

Mayra Alejandra Murillo Narvaez

Juan Jose Murillo Narvaez

Agradecimientos

Expreso mi más sincero agradecimiento a mi director, Wilfredo Alfonso Morales, y a mi codirector, Wilmar A. Torres López, por su invaluable apoyo, dedicación y enseñanzas a lo largo de este proceso. Su guía constante ha sido fundamental para mi crecimiento académico y personal. Me siento profundamente honrado de haber compartido este camino de aprendizaje con ustedes, quienes con su compromiso y motivación me impulsaron a culminar con éxito mi carrera profesional. Los conocimientos y experiencias adquiridos bajo su orientación han dado un nuevo sentido a mi trayectoria profesional.

También agradezco sinceramente a Cristhian Fernández y a John W. Ortiz por su valioso acompañamiento y por haber compartido sus conocimientos. Aprecio enormemente su disposición para brindar apoyo y dedicar tiempo durante el desarrollo de este trabajo, aportando de manera significativa a su calidad y avance.

A Andrei Urbano, has sido un pilar fundamental en mi vida y en este proceso lleno de desafíos. Agradezco profundamente su apoyo incondicional, su motivación constante incluso en los momentos de mayor incertidumbre y su firme voluntad de no dejar que me rindiera.

Agradezco a mis amigos y amigas Camila Abadía, Valeria Rodríguez, Sebastián Ruiz, Alejandro Escobar y Daniel Valderrama por ser ese refugio incondicional donde siempre encontré apoyo, aliento y la fuerza para seguir adelante.

Finalmente, agradezco a todas las personas que conocí durante mi paso por la universidad. Cada experiencia compartida y desafío superado contribuyó significativamente a mi crecimiento personal y profesional.

Glosario

PC_i i-ésimo componente principal. 44

AMI Infraestructura de Medición Avanzada. VI, VIII, XII, 1–4, 6–9, 11, 12, 20–22, 34

BIRCH Balanced Iterative Reducing and Clustering using Hierarchies. 13

CV Coeficiente de variación. 45–47, 49, 52, 58, 60, 64

DB Davies-Bouldin. 45

DBSCAN Density-based spatial clustering of applications with noise. VII, 13, 14, 17, 19, 38–40, 42, 43, 45–47, 57, 61, 63, 64, 66

DTW Dynamic Time Warping. VII, VIII, XII, XIII, 6, 15–17, 19, 39–41, 46, 47, 64

EDA Exploratory data analysis. VII, 19, 23, 35

eps Epsilon. IX, 40, 43

ETL Extract, Transform and Load. VII, 19, 21, 31, 46

GMM Gaussian Mixture Models. 13

HES Head-End System. 8

IA Inteligencia Artificial. 7

IQR Rango Intercuartil. 34

ML *Machine Learning*. 2

NaN Not a Number - No es un número. 23, 24, 34

PCA Análisis de componentes principales. VI, VII, XII, XIII, 16, 17, 19, 38, 42–44, 46, 47, 61, 63, 64, 66

RMS Root Mean Square - Raíz Cuadrada Media. 46, 47, 49, 52, 57–61, 64

TIC Tecnologías de la Información y Comunicación. 1

Tabla de Contenido

	Pág.
Aceptación	I
Dedicatoria.....	II
Agradecimientos	III
Glosario	IV
Índice General	VI
Lista de Figuras	VIII
Lista de Tablas	X
Resumen	XII
Abstract	XIII
1. Introducción.....	1
1.1. Objetivos	3
1.1.1. Objetivo General	3
1.1.2. Objetivos Específicos	3
1.2. Estructura del Libro	4
2. Datos AMI a partir de técnicas de aprendizaje no supervisado	6
2.1. Transformación de las redes eléctricas.....	6
2.2. Infraestructura de Medición Avanzada (AMI)	7
2.3. Big Data.....	9
2.4. <i>Clustering</i> para análisis de datos AMI	11
2.5. Algoritmos de agrupamiento no supervisados	12
2.6. Técnicas para series de tiempo	15
2.7. Análisis de componentes principales (PCA)	16
2.8. Resumen.....	17
3. Análisis de la información e implementación de técnicas clustering	19

3.1.	Características de la demanda	19
3.1.1.	Determinación y Análisis de las Características de la Base de Datos	20
3.2.	Primer proceso de ETL	21
3.2.1.	Limpieza de datos	22
3.2.2.	Enriquecimiento y preparación de datos de demanda.....	23
3.3.	Primer proceso de Análisis exploratorio de datos (EDA)	23
3.4.	Segundo proceso de ETL	31
3.5.	Proceso de desagregación.....	36
3.5.1.	El promedio de consumo horario en días laborales (lunes a viernes)	36
3.5.2.	El consumo típico de un sábado	37
3.5.3.	El promedio horario entre domingos y días festivos	37
3.5.4.	Registros de 72 horas por municipio	38
3.6.	Análisis de Clustering en el Consumo Energético de Clientes	38
3.6.1.	Clustering por forma	39
3.6.1.1.	Proceso de clustering DBSCAN.....	39
3.6.1.2.	Proceso de clustering K-MEANS	40
3.6.1.3.	Proceso de clustering K-MEDOIDS con DTW	41
3.6.2.	<i>Clustering</i> por magnitud	42
3.6.2.1.	Procedimiento de Aplicación del Análisis de Componentes Principales (PCA).....	42
3.6.2.2.	Configuración de los algoritmos de clustering por magnitud dentro de Implementación	43
3.7.	Métricas de Evaluación y Caracterización de Clusters	45
3.8.	Resumen.....	46
4.	Evaluación y Análisis de Resultados	47
4.1.	Análisis de <i>Clustering</i> para el Municipio de Popayán	48
4.2.	Análisis de <i>Clustering</i> para el Municipio de Morales.....	51
4.3.	Análisis de <i>Clustering</i> para el Municipio de Santander de Quilichao	55
4.4.	Análisis de <i>Clustering</i> para el Municipio de El Tambo	58
4.5.	Reducción de Dimensionalidad con PCA	61
4.6.	Resumen.....	63
5.	Conclusiones y Trabajos Futuros	66
5.1.	Conclusiones Generales.....	66
5.2.	Trabajos futuros	67
	Referencias	69

Lista de Figuras

FIGURA	Pág.
2.1. Diagrama de bloques de la infraestructura de medición avanzada en la red eléctrica inteligente.	8
2.2. <i>Clustering</i> de datos para clasificación.	12
2.3. Comparación de tres familias importantes de algoritmos para agrupamiento.	15
2.4. Ejemplo de tecnica DTW.....	16
2.5. Diagrama de metodología propuesta.....	18
3.1. Muestra de la estructura de la base de datos de archivos AMI.	21
3.2. Muestra de la base de datos de clientes con información geográfica y demográfica.	22
3.3. Distribución de la Demanda por Hora del Día.	24
3.4. Demanda por hora y estrato.	25
3.5. Curva de demanda promedio por hora y municipio.	26
3.6. Curva de demanda promedio por hora y ubicación (Rural o Urbana).	27
3.7. Curva de demanda promedio por hora y día de semana.	27
3.8. Curva de demanda promedio por hora y tipo de día semana con festivos.	28
3.9. Curva de demanda promedio por hora y tipo de día (domingo y festivo juntos).	29
3.10. Curvas de demanda promedio por hora y cliente.	30
3.11. Promedio de consumo 24 horas días de la semana Popayán.	37
3.12. Esquema de doble agrupamiento de clientes por forma y magnitud del consumo energético.	39

3.13. Análisis para determinar parámetro eps para el municipio de Popayán....	40
3.14. Gráfica de la variación de Davies-Bouldin en función del numero de clusters del municipio de Popayán.	41
4.1. Clústeres por Forma, municipio de Popayán.	48
4.2. Clústeres por Forma, municipio de Morales.	52
4.3. Clústeres por Forma, municipio de Santander de Quilichao.	55
4.4. Resultados de Clustering por Forma para El Tambo utilizando DBSCAN, K-Medoids y K-Means.	58
4.5. Metodología: desde la recolección de datos hasta la validación de resultados	65

Lista de Tablas

TABLA	Pág.
3.1. Distribución de usuarios por municipio, ubicación y subcategoría	31
3.2. Distribución de usuarios por municipio, ubicación y subcategoría	32
3.3. Distribución de usuarios por municipio, ubicación y estrato (filtrados)....	33
4.1. Análisis de métricas por cluster DBSCAN para el municipio de Popayán .	49
4.2. Análisis de métricas por cluster para el municipio de Popayán (K-MEDOIDES CON DTW)	49
4.3. Análisis de métricas por cluster para el municipio de Popayán, método K-MEANS	50
4.4. Análisis de métricas por cluster para el municipio de Popayán, método K-MEANS. Continuación.....	51
4.5. Análisis de métricas por cluster DBSCAN para el municipio de Morales. .	52
4.6. Análisis de métricas por cluster K-MEDOIDES para el municipio de Morales.	53
4.7. Análisis de métricas por cluster K-MEANS para el municipio de Morales	54
4.8. Análisis de métricas por cluster DBSCAN para el municipio de Santander de Quilichao.	56
4.9. Análisis de métricas por cluster K-MEDOIDES para el municipio de Santander de Quilichao.	56
4.10. Análisis de métricas por cluster K-MEANS para el municipio de Santander de Quilichao.	57
4.11. Análisis de métricas por cluster DBSCAN para el municipio de El Tambo.	59
4.12. Análisis de métricas por cluster K-MEDOIDES para el municipio de El Tambo	59
4.13. Análisis de métricas por cluster K-MEANS para el municipio de El Tambo	60

4.14. Resumen de componentes principales (PCA) por municipio y método de clustering.	62
---	----

Resumen

La gestión energética moderna enfrenta el desafío de optimizar el suministro eléctrico mediante el análisis de datos granulares de consumo, generados por la Infraestructura de Medición Avanzada (AMI). Sin embargo, el gran volumen de información requiere técnicas avanzadas para identificar patrones, gestionar la demanda y detectar anomalías, como el fraude energético. Para abordar este problema, este trabajo propone un modelo basado en técnicas de aprendizaje no supervisado aplicadas a series temporales, combinando algoritmos de clustering (DBSCAN, K-Means y K-Medoids) con la métrica Dynamic Time Warping (DTW) para capturar similitudes en los patrones de consumo.

La metodología se implementó en dos etapas: primero, se agruparon los usuarios según la forma de sus curvas de consumo mediante técnicas de clustering no supervisado, y luego se segmentaron por magnitud utilizando Análisis de Componentes Principales (PCA). Los resultados revelaron patrones diferenciados en los cuatro municipios estudiados: destacó la marcada heterogeneidad en Popayán y El Tambo, donde se identificaron múltiples perfiles de consumo contrastantes; la notable homogeneidad en Morales, con patrones de consumo muy similares entre usuarios; y en Santander de Quilichao se observó un comportamiento intermedio, con algunos grupos bien definidos pero también presencia significativa de consumos atípicos.

Palabras clave— Infraestructura de Medición Avanzada (AMI), gestión energética, aprendizaje automático, Big Data, clusterización, series de tiempo, patrones de consumo, demanda energética, análisis no supervisado.

Abstract

Modern energy management faces the challenge of optimizing electricity supply by analyzing granular consumption data generated by Advanced Metering Infrastructure (AMI). However, the large volume of information requires advanced techniques to identify patterns, manage demand, and detect anomalies, such as energy fraud. To address this problem, this paper proposes a model based on unsupervised learning techniques applied to time series, combining clustering algorithms (DBSCAN, K-Means, and K-Medoids) with the Dynamic Time Warping (DTW) metric to capture similarities in consumption patterns.

The methodology was implemented in two stages: first, users were grouped according to the shape of their consumption curves using unsupervised clustering techniques, and then they were segmented by magnitude using Principal Component Analysis (PCA). The results revealed distinct patterns in the four municipalities studied: marked heterogeneity stood out in Popayán and El Tambo, where multiple contrasting consumption profiles were identified; notable homogeneity was observed in Morales, with very similar consumption patterns among users; and in Santander de Quilichao, intermediate behavior was observed, with some well-defined groups but also a significant presence of atypical consumption.

Keywords— Advanced Metering Infrastructure (AMI), energy management, machine learning, Big Data, clustering, time series, consumption patterns, energy demand, unsupervised analysis.

Capítulo 1

Introducción

En el contexto de la gestión energética avanzada, la Infraestructura de Medición Avanzada (AMI) se ha consolidado como una herramienta fundamental para monitorear y optimizar el suministro eléctrico. También conocida como medición inteligente, la AMI representa un avance tecnológico que busca potenciar el desempeño del sistema eléctrico mediante las posibilidades que ofrecen las Tecnologías de la Información y Comunicación (TIC), promoviendo nuevos esquemas de participación e interacción entre las empresas proveedoras de energía y sus clientes [1]. Este paradigma genera grandes volúmenes de información sobre los usuarios, registrando su consumo eléctrico en intervalos de tiempo específicos, ya sea diarios, horarios e incluso cada 15 minutos. Esta granularidad permite identificar patrones de consumo a lo largo del tiempo, como picos de demanda, períodos de baja utilización y estacionalidades [2]. Además, facilita la gestión de la demanda, la optimización de la distribución de recursos y la detección de anomalías o comportamientos inusuales en el sistema, como problemas técnicos o posibles fraudes [1].

No obstante, identificar patrones de consumo eléctrico para cada usuario no es una tarea sencilla, ya que implica procesar grandes volúmenes de información, haciendo necesaria la aplicación de técnicas de *Big Data* y algoritmos avanzados de análisis basados en técnicas de aprendizaje automático para descubrir patrones ocultos en los datos de consumo de energía [2]. Entre estas técnicas destacan los algoritmos de *clustering* o agrupamiento, dentro del ámbito del aprendizaje no supervisado, los cuales permiten asociar a los usuarios con patrones de consumo similares, simplificando así la interpretación y el análisis de los datos.

El uso de algoritmos de *clustering* resulta crucial en el análisis de datos AMI, ya que permite identificar patrones de consumo de energía sin la necesidad de supervisión directa. Prescindir de este tipo de técnicas limita la capacidad de las empresas para

comprender y gestionar eficientemente el comportamiento energético de sus usuarios. Asimismo, dificulta la segmentación efectiva de los consumidores, lo que obstaculiza la implementación de estrategias personalizadas de eficiencia energética y gestión de la demanda [2]. Por el contrario, al aplicar clustering, es posible mejorar la comprensión de los patrones de consumo, detectar irregularidades que podrían indicar actividades fraudulentas, como el uso no autorizado del servicio, y fortalecer la eficiencia del sistema eléctrico [2]. Por ejemplo, al detectar irregularidades mediante el análisis de datos AMI se puede garantizar la integridad y eficiencia del suministro eléctrico, donde patrones de consumo inusualmente bajos o nulos durante períodos de ocupación pueden ser señales claras de posibles casos de uso no autorizado del servicio [1].

Por otro lado, los algoritmos tradicionales de *clustering* suelen formar agrupaciones mediante hiperesferas en espacios de alta dimensionalidad, definidos por un centro y un radio. Aunque este enfoque es computacionalmente eficiente, desconoce la diversidad estructural que pueden presentar los patrones de consumo, lo que puede llevar a un desempeño subóptimo al no capturar adecuadamente la complejidad de los datos [3]. Gracias a la evolución de los algoritmos de *Machine Learning* (ML), hoy es posible detectar *clusters* con formas irregulares o no convexas, lo que representa con mayor fidelidad la estructura real de los datos recopilados por los medidores inteligentes. Sin embargo, seleccionar las técnicas más apropiadas exige considerar aspectos como la complejidad del conjunto de datos, su volumen, la interpretabilidad de los resultados y la capacidad de adaptación a cambios temporales en los patrones de consumo de energía [4].

Con base en lo anterior, este trabajo propone desarrollar un modelo para el análisis diagnóstico de datos AMI utilizando técnicas de aprendizaje no supervisado sobre series temporales, con el fin de responder a la siguiente pregunta de investigación:

¿Qué técnicas de aprendizaje no supervisado son las más apropiadas para desarrollar un análisis de datos AMI y comprender eficazmente el consumo de energía eléctrica?

Este trabajo aborda la problemática de la gestión ineficiente del consumo de energía al revelar patrones ocultos y segmentar a los usuarios en grupos con comportamientos

similares, lo cual permite personalizar tarifas y servicios. Las técnicas identificadas no solo facilitan la detección de anomalías y problemas en el sistema de medición, sino que también optimizan la previsión de la demanda y permiten identificar patrones estacionales y cíclicos, además de revelar zonas de alto consumo donde implementar estrategias de eficiencia energética. En conjunto, estas técnicas mejoran la visibilidad del comportamiento energético de los usuarios y fortalecen la toma de decisiones en la gestión de recursos.

1.1. Objetivos

1.1.1. Objetivo General

Desarrollar un análisis diagnóstico de datos de la Infraestructura de Medición Avanzada (AMI) a partir de técnicas de aprendizaje no supervisado sobre series temporales que permitan la comprensión del consumo de energía.

1.1.2. Objetivos Específicos

- Definir marcos de investigación relevantes para la extracción de datos AMI mediante técnicas no supervisadas y la segregación de la información.
- Seleccionar técnicas de *clustering* adecuadas que permitan realizar un análisis diagnóstico de la información contenida en la base de datos asignada.
- Implementar técnicas de *clustering* sobre el conjunto de datos AMI para identificar patrones de consumo de energía en diferentes contextos geográficos y temporales, considerando variaciones en forma y magnitud.
- Evaluar las técnicas de aprendizaje no supervisado aplicadas al análisis de datos AMI por medio de índices de desempeño.

1.2. Estructura del Libro

Este trabajo de grado consiste de 5 capítulos, los cuales buscan responder a cada uno de los objetivos específicos planteados, que se describen a continuación:

- En el Capítulo 1, se exploran los fundamentos y la importancia de la Infraestructura de Medición Avanzada (AMI) en la gestión moderna de la energía. Este capítulo comienza con una introducción al concepto de AMI y su papel esencial en el monitoreo y optimización del suministro eléctrico mediante la recopilación en tiempo real de datos precisos de consumo a través de medidores inteligentes. A continuación, se examina el papel crucial del *clustering*, una técnica de aprendizaje no supervisado, en el análisis de estos datos, permitiendo identificar patrones de consumo y detectar anomalías como el fraude energético. El capítulo también aborda los desafíos asociados con el análisis de grandes volúmenes de datos y cómo los algoritmos de *clustering* tradicionales pueden ser limitados en la captura de estructuras complejas.
- En el Capítulo 2, se presentan los trabajos más relevantes revisados en el estado del arte, junto con una explicación detallada de los conceptos clave necesarios para entender la Infraestructura de Medición Avanzada (AMI). Este capítulo aborda los fundamentos del análisis de datos en el contexto de AMI, explorando técnicas de *clustering* y aprendizaje no supervisado, así como el tratamiento de series temporales. Además, se examinan los mecanismos para la compresión y manejo eficiente del consumo energético en el sector.
- El Capítulo 3 presenta una descripción de la solución desarrollada, comenzando con una explicación de la base de datos utilizada. Luego, se presentan tanto la teoría como la práctica de los métodos aplicados para la extracción de características de los datos. Este capítulo detalla la implementación de técnicas de *clustering* sobre los conjuntos de datos de AMI para identificar patrones de consumo de energía en distintos contextos geográficos y temporales.
- El Capítulo 4 presenta los resultados de las pruebas realizadas con técnicas de aprendizaje no supervisado aplicadas al análisis de datos de AMI. En este capítulo, se detallan los resultados obtenidos a través de diversos índices de

clustering, que se utilizan para evaluar la calidad y la efectividad de los métodos de agrupamiento empleados. Se discuten los índices empleados para evaluar el rendimiento de las técnicas de *clustering* y se analizan los hallazgos en términos de precisión y relevancia en el contexto de los datos de consumo energético.

- Por último, el Capítulo 5 presenta las conclusiones generales del trabajo, proporcionando un análisis de los resultados obtenidos en relación con cada uno de los objetivos establecidos. En este capítulo, se explican cómo se cumplieron los objetivos del estudio. Además, se sugieren posibles direcciones para futuras investigaciones.

Capítulo 2

Datos AMI a partir de técnicas de aprendizaje no supervisado

En este capítulo se lleva a cabo una revisión exhaustiva de la literatura sobre técnicas de *clustering* basadas en la forma y amplitud. También se investiga la Infraestructura de Medición Avanzada (AMI), explorando diversos enfoques para la extracción de datos y el diagnóstico del consumo energético. Se presentan distintos métodos para la extracción de características de los datos, incluyendo modificaciones en la estructura de los modelos, como la incorporación de variables exógenas y efectos calendario, con el objetivo de mejorar la precisión de los pronósticos en diferentes contextos. Además, el capítulo ofrece un análisis de conceptos clave como *Dynamic Time Warping* (DTW), evaluando cómo esta técnicas pueden contribuir a un análisis más detallado y preciso del consumo energético.

2.1. Transformación de las redes eléctricas

El mundo está viviendo un período de intensa y dinámica innovación científica y tecnológica. La automatización ha evolucionado rápidamente hacia la digitalización, la interconexión y la inteligencia, lo que ha dado lugar a nuevas tecnologías, industrias, modelos y formatos a gran escala. Estos desarrollos han impactado profundamente el panorama global de la innovación, el ecosistema industrial y las tendencias económicas. En este contexto, la transformación digital de la economía global se ha convertido en una tendencia predominante [5].

Para las empresas de energía eléctrica, adoptar la transformación digital es fundamental para adaptarse a esta era de cambios aprovechando las tecnologías digitales y nuevos enfoques operativos. Este proceso de transformación abre el camino hacia una nueva era de desarrollo empresarial, permitiendo mejorar la eficiencia

operativa y la calidad del servicio. Para lograrlo, es crucial formular una estrategia de desarrollo digital, establecer organizaciones especializadas para coordinar estas iniciativas y construir conjuntamente la “Internet de la energía”, integrando constantemente nuevas tecnologías y conceptos [5].

La digitalización de la red eléctrica está provocando cambios significativos en los modos de producción y mejorando la eficiencia. En este escenario, las empresas eléctricas enfrentan una competencia de mercado cada vez más intensa y presiones para reducir costos y mejorar la eficiencia. Para abordar estos desafíos, las empresas están buscando urgentemente tecnologías innovadoras centradas en la IA [5].

2.2. Infraestructura de Medición Avanzada (AMI)

La red eléctrica convencional facilita únicamente la comunicación unidireccional y tiene un diseño centralizado. En los medidores tradicionales, los datos de consumo de energía no se pueden almacenar y el uso de electricidad solo se puede rastrear en la factura mensual o trimestral, y solo se cuenta con el lector del medidor de la compañía eléctrica. Dado que los problemas relacionados con fallas se monitorean manualmente, las compañías de servicios de distribución no pueden responder rápidamente a las interrupciones del suministro. La red inteligente es un sistema avanzado para facilitar la generación, distribución y consumo de electricidad que admite la comunicación bidireccional entre proveedores y consumidores. Para monitorear y regular mejor el flujo y el uso de energía, se está integrando una red eléctrica más sofisticada con tecnologías de comunicación e información. La Infraestructura de Medición Avanzada (AMI) es la columna vertebral de la red inteligente. AMI se utiliza para reducir las pérdidas no técnicas mediante un sistema integrado de medidores inteligentes, que recopila y analiza datos mediante comunicaciones bidireccionales y proporciona una gestión inteligente de varias aplicaciones y servicios relacionados con la energía en función de los datos recibidos [6].

La modernización de las redes eléctricas tradicionales mediante la implementación de la Infraestructura de Medición Avanzada implica no solo reemplazar los medidores convencionales con medidores inteligentes que registran datos de consumo de energía

en intervalos cortos y permiten una comunicación bidireccional con la compañía eléctrica, sino también contar con un sistema de comunicación robusto y con plataformas de gestión de medida (HES) que centralicen, procesen y administren la información recopilada para su posterior análisis y toma de decisiones. Esto facilita el monitoreo remoto, la gestión de la demanda, la detección de fallos y la integración de fuentes de energía renovable. Además, AMI mejora la eficiencia operativa al reducir costos, pérdidas de energía y errores humanos, y permite la implementación de tarifas dinámicas que benefician tanto a las empresas de energía como a los consumidores. A pesar de los desafíos en su implementación, como la inversión en infraestructura y la seguridad de los datos, AMI es clave para avanzar hacia redes eléctricas más inteligentes, eficientes y sostenibles. El diagrama de bloques de la Infraestructura de Medición Avanzada en la red eléctrica inteligente se presenta en la Figura 2.1.

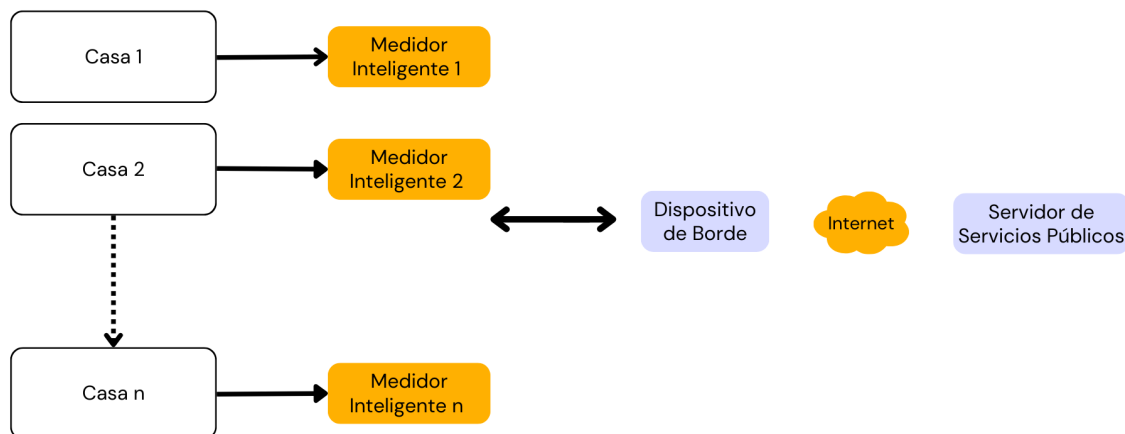


Figura 2.1: Diagrama de bloques de la infraestructura de medición avanzada en la red eléctrica inteligente. Adaptado de [6]

De acuerdo con el artículo titulado “*Identifying susceptible consumers for demand response and energy efficiency policies by time-series analysis and supplementary approaches*” [7], en una infraestructura de medición avanzada, la recopilación y análisis de datos juegan un papel fundamental en la optimización de la gestión de recursos como energía. Estos sistemas permiten la obtención de lecturas detalladas del consumo en tiempo real, que facilita el monitoreo continuo y la identificación precisa de patrones de uso. Los perfiles de carga y demanda proporcionan una visión detallada de cómo varía el consumo a lo largo del tiempo, lo cual es esencial para la planificación y la gestión de recursos. Además, los datos sobre la calidad del

suministro, como el voltaje y la frecuencia, permiten a las empresas asegurar un suministro estable y eficiente, mientras que los registros de eventos, tales como alarmas y anomalías, ayudan a detectar y abordar problemas de manera proactiva.

La infraestructura AMI también gestiona datos relacionados con la comunicación y el estado de los medidores. Esto incluye información sobre la conectividad de los dispositivos y cualquier error de transmisión que pueda ocurrir. Los parámetros de configuración y su historial permiten ajustes precisos y un mantenimiento adecuado de los medidores. Además, los datos sobre tarificación y facturación aseguran una facturación precisa basada en el consumo real registrado. La información del perfil de usuario y los patrones de uso ayudan a personalizar el servicio y mejorar la experiencia del cliente. En conjunto, estos datos permiten una optimización integral de las operaciones, una facturación más exacta y una mejor atención al cliente.

En el estudio realizado en el artículo “*The Power of Big Data and Data Analytics for AMI Data: A Case Study*” [2], la gestión y análisis de datos se han transformado significativamente con la integración de tecnologías de Big Data. Los sistemas AMI generan volúmenes masivos de datos provenientes de lecturas de consumo, perfiles de carga, calidad del suministro, y comunicaciones en tiempo real, que se estructuran en series temporales, registros de eventos y parámetros configurables. Este enfoque de Big Data permite la captura, almacenamiento y procesamiento eficiente de grandes conjuntos de datos, ofreciendo una visión más granular y dinámica del comportamiento de los recursos y el perfil del usuario. Con herramientas de análisis avanzadas, es posible extraer información valiosa para optimizar operaciones, prever patrones de consumo, mejorar la gestión de recursos y ofrecer una facturación más precisa. La capacidad de manejar y analizar estos datos a gran escala no solo mejora la eficiencia operativa, sino que también permite una toma de decisiones más informada y una respuesta más ágil a las demandas del mercado.

2.3. Big Data

Los datos generados por las redes inteligentes (smart grids) requieren tratamientos complejos debido a su naturaleza, distribución y las limitaciones en tiempo real que

imponen ciertas necesidades operativas. En este contexto, las técnicas de Big Data resultan adecuadas para una gestión avanzada y eficiente de estos datos, facilitando aplicaciones innovadoras y decisiones más informadas para las empresas de servicios públicos. Por ejemplo, el análisis de grandes volúmenes de datos permite a estas empresas comprender mejor el comportamiento de los clientes, gestionar la conservación de energía, analizar patrones de consumo y demanda, y realizar un seguimiento más preciso del tiempo de inactividad y los cortes de energía [8].

Sin embargo, esto también presenta desafíos significativos para aquellas empresas que carecen de los sistemas adecuados y las habilidades de análisis necesarias para procesar estos datos. Por lo tanto, el objetivo principal para las empresas de servicios públicos es desarrollar la capacidad de gestionar grandes volúmenes de datos y utilizar análisis avanzados para transformar los datos recopilados en información valiosa, luego en conocimiento, y finalmente en planes de acción efectivos.

En este contexto, las 5 V del Big Data [8] juegan un papel crucial al describir las características principales necesarias para gestionar eficientemente estos datos:

- **Volumen:** Las redes inteligentes generan enormes cantidades de datos, desde mediciones de consumo energético hasta datos de sensores en tiempo real, que pueden alcanzar terabytes, petabytes o incluso exabytes.
- **Velocidad:** Los datos en redes inteligentes deben procesarse rápidamente, a menudo en tiempo real o casi real, para permitir decisiones inmediatas sobre el suministro de energía, la respuesta a fallos, o la optimización del rendimiento de la red.
- **Variedad:** Los datos provenientes de redes inteligentes incluyen diversos tipos, como datos estructurados (lecturas de medidores), no estructurados (imágenes de drones para mantenimiento) y semiestructurados (logs de eventos).
- **Veracidad:** Es fundamental asegurar la precisión y confiabilidad de los datos, ya que decisiones críticas, como la gestión de la red y la respuesta a emergencias, dependen de la calidad de la información.
- **Valor:** El análisis de datos de las redes inteligentes debe enfocarse en extraer información útil que pueda convertirse en conocimiento accionable para mejorar

la eficiencia operativa, optimizar el uso de recursos, reducir costos y mejorar la satisfacción del cliente.

Después de aplicar técnicas de Big Data en la gestión de datos de las infraestructuras de medición avanzada (AMI), se abre la puerta a otras metodologías analíticas que pueden mejorar aún más el uso y el valor de los datos generados. Una de estas técnicas es el *clustering* o agrupamiento, que es un tema importante en la interpretación y el análisis de grandes volúmenes de datos provenientes de AMI.

2.4. *Clustering* para analisis de datos AMI

El artículo “*Time-series clustering and forecasting household electricity demand using smart meter data*” [4], muestra como el *clustering* y el análisis de datos han contribuido significativamente al ámbito del consumo de energía mediante la identificación de patrones de consumo y la optimización de la gestión de la red eléctrica. A través del *clustering*, es posible agrupar diferentes tipos de consumidores según sus patrones de uso de energía, esto permite a las empresas de servicios públicos y a los gestores energéticos comprender mejor las demandas de energía y planificar de manera más eficiente la generación y distribución de recursos. Además, estas técnicas facilitan la previsión de picos de demanda y la identificación de momentos de baja demanda, contribuyendo a equilibrar la carga en la red eléctrica, reducir pérdidas de energía y mejorar la estabilidad del suministro eléctrico.

El análisis de datos también ha permitido el desarrollo de tarifas personalizadas, la detección de anomalías y fraude, y la promoción de la eficiencia energética. Mediante el *clustering*, las empresas pueden crear tarifas adecuadas para diferentes segmentos de consumidores, incentivando el uso eficiente de la energía. La detección de patrones inusuales ayuda a identificar problemas en la red y prevenir fraudes, mientras que el análisis de consumo puede sugerir mejoras en la eficiencia energética tanto para industrias como para hogares. Además, estas técnicas facilitan la integración de fuentes de energía renovable y permiten una mejor planificación y expansión de la infraestructura energética, asegurando que la capacidad futura satisfaga la demanda prevista de manera eficiente y sostenible.

En la Figura 2.2, se ilustra de forma esquemática el proceso de clustering aplicado a datos sin procesar. Este procedimiento permite organizar conjuntos de datos heterogéneos en grupos con características similares mediante algoritmos de aprendizaje no supervisado. Tal como se observa, los datos de entrada (Raw Data) son procesados por un algoritmo que identifica patrones y estructura interna en los datos, generando como salida (Output) grupos diferenciados.

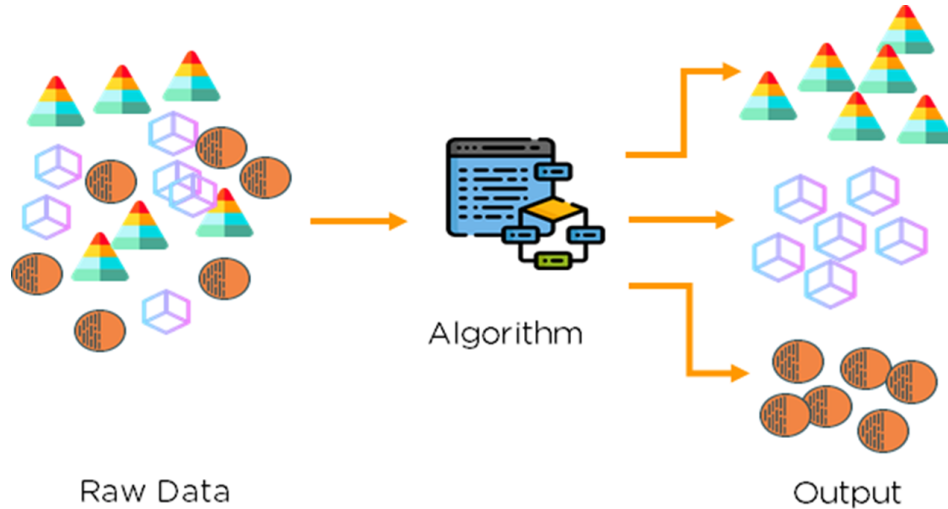


Figura 2.2: *Clustering* de datos para clasificación. Tomado de [4]

El *clustering* se convierte en una herramienta esencial para manejar y analizar la gran cantidad de datos generados por dispositivos de medición, como los medidores inteligentes. Estos dispositivos recopilan continuamente datos sobre el consumo de energía, agua y gas, lo que resulta en grandes volúmenes de información que deben ser procesados para extraer conocimientos significativos. Para manejar estos volúmenes de datos, se recurre a los algoritmos de agrupamiento no supervisado, una subcategoría del clustering que no requiere etiquetas o datos previamente clasificados. Estos algoritmos permiten identificar patrones ocultos y relaciones dentro de los datos sin necesidad de entrenamiento previo.

2.5. Algoritmos de agrupamiento no supervisados

Según el artículo “*Density-based clustering algorithm for associating transformers with smart meters via GPS-AMI data*” [9], los algoritmos de agrupamiento no

supervisado son métodos de aprendizaje automático que se utilizan para agrupar un conjunto de datos en subconjuntos basados en sus características y similitudes. A diferencia del aprendizaje supervisado, no requieren etiquetas o información previa sobre las categorías de los datos; el objetivo es descubrir patrones o estructuras inherentes dentro del conjunto de datos.

Del artículo “*A comparison of unsupervised learning algorithms for gesture clustering*” [10] se encuentra que los algoritmos de agrupamiento no supervisado son métodos de aprendizaje automático que se utilizan para agrupar un conjunto de datos en subconjuntos o “clusters” basados en sus características y similitudes. A diferencia del aprendizaje supervisado, no requieren etiquetas o información previa sobre las categorías de los datos; el objetivo es descubrir patrones o estructuras inherentes dentro del conjunto de datos. Entre los algoritmos más destacados se encuentran DBSCAN, K-Means, BIRCH, Mean Shift, Gaussian Mixture Models (GMM) y Spectral Clustering. Cada uno de estos métodos ofrece diferentes ventajas y es adecuado para tipos de datos y problemas de agrupamiento.

Además, Autoencoders y Isolation Forest también son herramientas importantes en el contexto del clustering. Los autoencoders son redes neuronales que pueden aprender representaciones compactas de los datos, mejorando la calidad del agrupamiento al reducir la dimensionalidad antes de aplicar técnicas de clustering. Por otro lado, el Isolation Forest es útil para detectar anomalías y manejar outliers, lo que puede facilitar la identificación de clusters más significativos al tratar con datos ruidosos o atípicos.

Basado en el artículo “*Improving Clustering Method Performance Using K-Means, Mini Batch K-Means, BIRCH and Spectral*” [11], el algoritmo K-Means es un método de clasificación no paramétrico que se caracteriza por ser efectivo, rápido y robusto. Este algoritmo funciona identificando un número determinado de patrones dentro de los datos de entrenamiento y agrupando aquellos que se encuentran más cercanos entre sí. Posteriormente, la clase de un nuevo dato se determina mediante una técnica de votación basada en estos patrones cercanos. Sin embargo, K-Means presenta algunas limitaciones. Es sensible a características irrelevantes y a la elección del número de vecinos. Además, su precisión puede variar significativamente dependiendo de la distribución de los datos: en algunos casos puede ser alta, pero en otros muy baja.

El artículo “*DBSCAN Clustering Algorithm Based on Density*” [12] hace referencia a que el algoritmo DBSCAN tiene buenos resultados de agrupamiento y es un representante típico del algoritmo de densidad. El algoritmo tiene la característica de encontrar cualquier clase de forma en el conjunto de datos. El algoritmo DBSCAN requiere que el usuario establezca el parámetro constante global de distancia de vecindad y el umbral antes de ejecutarse. La distancia de vecindad establece el radio del rango de vecindad del punto de muestra. El umbral se establece cambiando el número mínimo de puntos de muestra dentro del radio del rango de vecindad que se marcará como un punto central. Vale la pena señalar que la distancia de vecindad y el umbral son constantes. El algoritmo DBSCAN es insensible a los puntos de ruido y se puede aplicar a conjuntos de datos que contienen puntos de ruido. Puede identificar puntos de ruido y excluirllos de los resultados de agrupamiento. Para cada clase en el resultado de agrupamiento, la densidad dentro de la clase es mayor que la densidad en el borde de la clase. La densidad del punto de ruido es menor que la del borde. De acuerdo con las características de distribución de datos, el algoritmo utiliza la diferencia de densidad para identificar diferentes regiones de densidad y marca los resultados de agrupamiento.

La investigación “*Entropy-Based Feature Selection for Data Clustering Using k-Means and k-Medoids Algorithms*” [13], presenta el algoritmo de agrupamiento k-Medoides, un algoritmo iterativo basado en la distancia. Utiliza una estrategia de ascenso de pendientes para seleccionar centroides de clúster a partir de los puntos del conjunto de datos. En este enfoque, se selecciona un nuevo centroide en cada iteración y los demás permanecen sin cambios. Un punto de datos se asigna a un clúster específico cuyo centroide es el más cercano a dicho punto. La iteración continúa hasta que no se reasignan los puntos de datos. Este algoritmo se aplica a conjuntos de datos numéricos y presenta dos ventajas principales: es menos sensible al ruido y a los valores atípicos, y se adapta bien a distintos tipos de características. Sin embargo, una de sus desventajas es que, una vez ejecutado el proceso, no se puede deshacer.

La Figura 2.3 presenta el resumen de los algoritmos de clustering mencionados anteriormente. Se destacan aspectos clave como la detección de densidades en DBSCAN, la optimización basada en distancias en K-Means y la minimización de distancias en K-Medoids, los cuales son fundamentales para comprender las diferencias y aplicaciones de cada método.

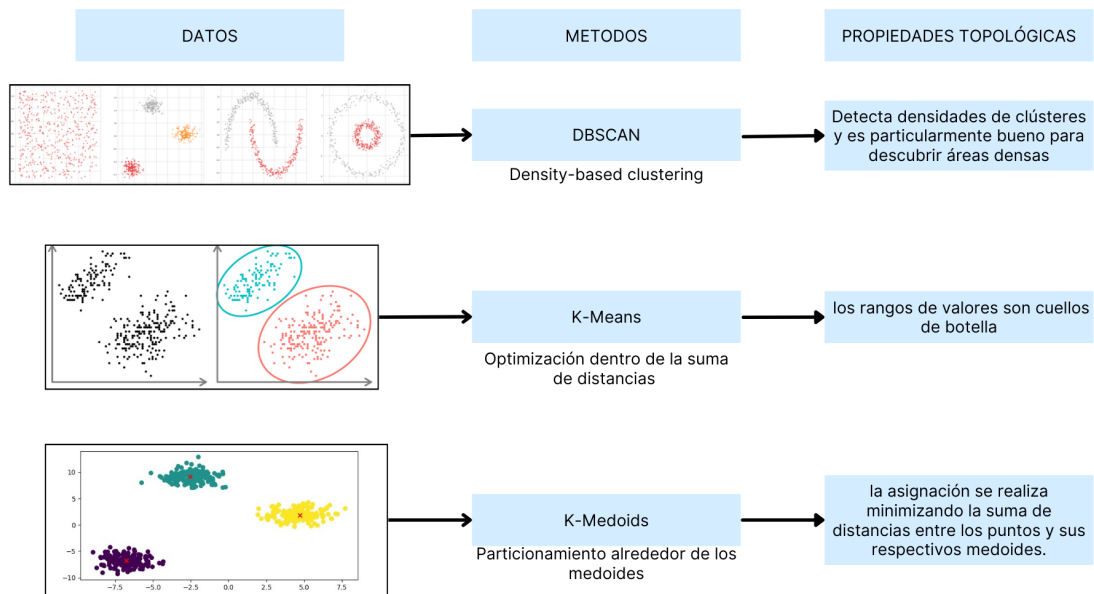


Figura 2.3: Comparación de tres familias importantes de algoritmos para agrupamiento. Adaptado de [9, 13]

El agrupamiento no supervisado se ha utilizado ampliamente en problemas donde se busca identificar patrones en conjuntos de datos estáticos y sincrónicos. Sin embargo, cuando los datos presentan una naturaleza secuencial o temporal, como en las series temporales, se requieren técnicas adicionales que buscan capturar la dinámica temporal de las observaciones.

2.6. Técnicas para series de tiempo

Al igual que los algoritmos de *clustering* buscan descubrir patrones ocultos en datos sin etiquetas, las técnicas para series temporales también intentan identificar patrones de comportamiento, pero con un enfoque específico en la dinámica temporal. En este caso, la técnica *Dynamic Time Warping* (DTW), es una herramienta diseñada para comparar secuencias temporales de diferente longitud o velocidad, capturando su estructura temporal mediante la alineación flexible de los puntos de datos. Esto marca el comienzo de un análisis más detallado de cómo manejar datos que cambian a lo largo del tiempo, utilizando herramientas diseñadas específicamente para captar la estructura temporal de las observaciones.

En el trabajo titulado “*A Time Series Data Augmentation Method Based on Dynamic Time Warping*” [14], la técnica DTW hace que la comparación de la similitud de las series temporales sea más razonable. Para adaptarse a los cambios de fase y amplitud entre series temporales, la deformación temporal dinámica utiliza un algoritmo de programación dinámica para mejorar la alineación de los puntos de datos de las series temporales desde la distancia euclidiana “uno a uno” hacia “uno a muchos”.

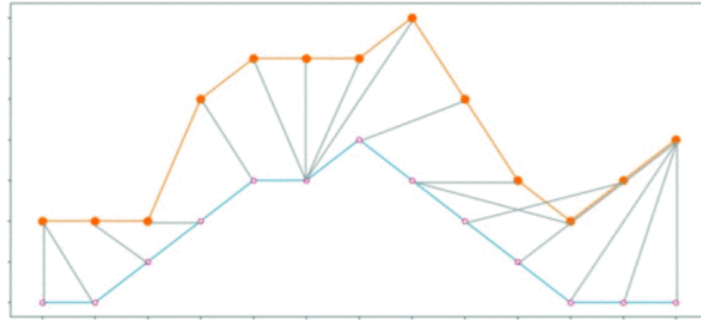


Figura 2.4: Ejemplo de tecnica DTW. Tomado de [14]

Según lo expuesto en “*Multivariate time series clustering based on common principal component analysis*” [15], las técnicas para series temporales, como *Dynamic Time Warping*, y los métodos de reducción de dimensionalidad, como Principal Component Analysis (PCA), pueden complementarse para mejorar el análisis y la clasificación de datos que varían con el tiempo. El uso de PCA en el agrupamiento de series temporales multivariadas puede aumentar la precisión y eficiencia del análisis, y se sugiere que su combinación con técnicas como DTW permite capturar de manera más efectiva las estructuras subyacentes en datos de alta dimensión.

2.7. Análisis de componentes principales (PCA)

En el artículo "*Principal Component Analysis*"[16] se menciona que el análisis de componentes principales (PCA), es una técnica estadística utilizada para reducir la dimensionalidad de los datos de manera que se mantenga la mayor cantidad de varianza posible. El objetivo principal de PCA es simplificar el análisis de grandes volúmenes de datos, eliminando redundancia y conservando las características más significativas, también se discute cómo PCA puede aplicarse en una variedad de

campos, como la visualización de datos de alta dimensión, la mejora del rendimiento de modelos predictivos, la compresión de datos y la reducción del ruido. En el contexto de series temporales, PCA puede utilizarse para capturar las principales tendencias en los datos antes de aplicar técnicas adicionales.

Por otro lado en el artículo “*The analysis and research of clustering algorithm based on PCA*” [17] Se menciona que en el proceso de segmentación de clientes, la aplicación de PCA antes de implementar algoritmos de *clustering* permite reducir la dimensionalidad de los datos, preservando las características más relevantes y eliminando redundancias. Esto es particularmente útil cuando se trabaja con grandes volúmenes de datos que contienen muchas variables, como en el análisis de comportamiento de consumo. Al aplicar PCA, las dimensiones más significativas son extraídas, lo que facilita la identificación de patrones ocultos en el comportamiento de los clientes. Posteriormente, algoritmos de clustering como K-Means o DBSCAN se utilizan para agrupar a los clientes en clusters, donde cada grupo representa un patrón o tendencia de consumo similar.

2.8. Resumen

A partir de la revisión de la literatura y del análisis de las principales características de los enfoques estudiados, este trabajo propone una solución para el desarrollo de un análisis diagnóstico de datos provenientes de la Infraestructura de Medición Avanzada (AMI), utilizando técnicas de aprendizaje no supervisado aplicadas a series temporales. Esta iniciativa responde a la creciente necesidad de interpretar de manera automatizada y precisa los grandes volúmenes de datos generados por los medidores inteligentes, con el objetivo de identificar patrones de comportamiento en el consumo energético, detectar anomalías y segmentar a los usuarios según sus hábitos de uso.

La propuesta metodológica se fundamenta en la aplicación de algoritmos de agrupamiento no supervisado como K-Means, DBSCAN o K-Medoids, complementados con métricas diseñadas para datos temporales, como *Dynamic Time Warping* (DTW). Estos algoritmos fueron seleccionados por su complementariedad y capacidad para abordar distintos retos en los datos AMI. K-Means destaca por su

eficiencia computacional y su simplicidad para identificar clústeres esféricos, mientras que DBSCAN es especialmente útil para detectar clústeres con formas arbitrarias y para manejar ruido y valores atípicos de forma robusta. K-Medoids, por su parte, emplea puntos reales del conjunto de datos como centroides, lo que mejora la interpretabilidad y estabilidad frente a outliers. Esta combinación permite capturar tanto similitudes estáticas como dinámicas en las series de consumo, mejorando la capacidad de diagnóstico y análisis contextual de la red eléctrica, y facilitando la toma de decisiones estratégicas en áreas como la gestión de la demanda, la eficiencia energética, la planificación operativa y la detección temprana de irregularidades.

El proceso metodológico se presenta esquemáticamente en la Figura 2.5, donde se ilustra el flujo de trabajo que será detallado en los capítulos siguientes. Este proceso inicia con una fase de preprocesamiento de datos, que incluye la limpieza, eliminación de valores atípicos y selección de características relevantes. Posteriormente, se aplican los modelos de *clustering* sobre los datos procesados, lo que permite optimizar la segmentación de usuarios y garantizar tanto la calidad de la información como el rendimiento de los algoritmos utilizados en el análisis energético.

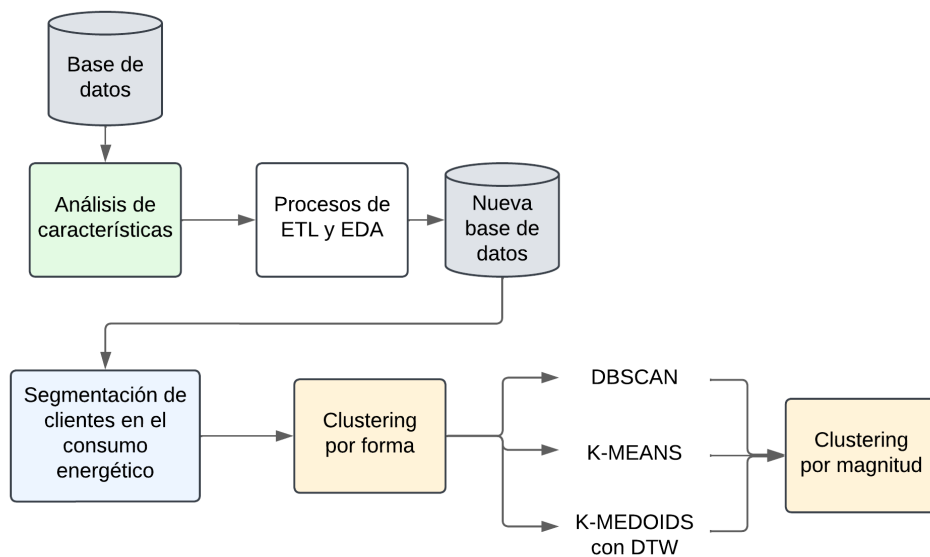


Figura 2.5: Diagrama de metodología propuesta.

Capítulo 3

Análisis de la información e implementación de técnicas clustering

Este capítulo presenta el análisis de los datos de consumo energético mediante la implementación de técnicas de *clustering*. El proceso inicia con la identificación de las características de la demanda y el preprocesamiento de datos a través de un proceso de Extracción, Transformación y Carga (ETL), que incluye tareas de limpieza, filtrado y estructuración de los datos; seguidamente, se lleva a cabo un Análisis Exploratorio de Datos (EDA) con el objetivo de obtener *insights* relevantes sobre los patrones de consumo. Posteriormente, se aplican técnicas de *clustering* no supervisado, como DBSCAN, K-Means y K-Medoids con DTW, para segmentar a los clientes según su comportamiento energético, complementadas con PCA para reducir la dimensionalidad y mejorar la segmentación. Este enfoque integral permite un análisis detallado y eficiente del comportamiento de los usuarios, mejorando la comprensión de sus hábitos de consumo energético y facilitando la toma de decisiones estratégicas.

3.1. Características de la demanda

La evaluación de la demanda energética y el análisis de los patrones de consumo son fundamentales para optimizar el uso y la distribución de la energía y de la infraestructura eléctrica disponible en cualquier sistema, especialmente en un contexto marcado por el crecimiento sostenido de la demanda y el uso intensivo de tecnologías inteligentes. En esta investigación se aprovechan datos detallados provenientes de medidores inteligentes, registros históricos de consumo y perfiles de clientes, con el propósito de construir una visión integral de cómo, cuándo y por qué se produce el consumo energético.

3.1.1. Determinación y Análisis de las Características de la Base de Datos

Como insumo principal para el desarrollo del análisis diagnóstico, se dispone de registros históricos de demanda eléctrica obtenidos a partir de sistemas de medición avanzada (AMI) correspondientes a cuatro municipios del departamento del Cauca. La base de datos está conformada por un conjunto de archivos que contienen información relevante para el análisis del consumo energético, el perfilado de clientes y el monitoreo a través de medidores inteligentes. Estos datos permiten obtener una visión detallada del uso de la energía en distintas zonas y segmentos poblacionales, desde septiembre de 2020 hasta marzo de 2021, facilitando mediante esta ventana temporal la formulación de estrategias orientadas a optimizar la operación de la red de distribución y mejorar la eficiencia en la gestión energética.

El conjunto de datos está compuesto por nueve archivos en formato “.dat”, los cuales almacenan información en formato binario, adecuado para el manejo eficiente de grandes volúmenes de datos. Cada archivo contiene series temporales de consumo energético organizadas por usuario, medidor y zona geográfica. A partir de los datos registrados por los medidores inteligentes, cada archivo se importa como un DataFrame independiente con un delimitador específico. Posteriormente, estos nueve DataFrames se integran en una única estructura consolidada, lo que facilita el análisis conjunto de todos los registros. Una muestra de la estructura final de la base de datos está en la Figura 3.1, donde se observa una estructura unificada que contiene 59 columnas y un total de 5.968.193 registros.

(5968193, 59)

	DT	Meter Number	Customer Account Number	Serial Number	Port	Channel	Conversion Factor	Data Type	Start Date	Start Time	...	Interval#20	QC#21	Interval#21	QC#22	Interval#22	QC#23
0	DT	10000642	850523231	84122107	0.0	1.0	0.032	Active Energy Wh	011221	00:00	...	-2.0	9.0	-2.0	9.0	-2.0	9.0
1	DT	10000642	850523231	84122107	0.0	2.0	0.032	Inductive Reactive Energy Varh	011221	00:00	...	-2.0	9.0	-2.0	9.0	-2.0	9.0
2	DT	10000642	850523231	84122107	0.0	1.0	0.032	Active Energy Wh	011321	00:00	...	-2.0	9.0	-2.0	9.0	-2.0	9.0
3	DT	10000642	850523231	84122107	0.0	2.0	0.032	Inductive Reactive Energy Varh	031721	00:00	...	-2.0	9.0	-2.0	9.0	-2.0	9.0
4	DT	10000642	850523231	84122107	0.0	2.0	0.032	Inductive Reactive Energy Varh	011321	00:00	...	-2.0	9.0	-2.0	9.0	-2.0	9.0

Figura 3.1: Muestra de la estructura de la base de datos de archivos AMI.

3.2. Primer proceso de ETL

Para obtener un análisis preciso del consumo energético, se realizó un estudio preliminar mediante la aplicación de técnicas (ETL), orientado a la caracterización de las curvas de demanda a partir de la base de datos inicial.

Una vez definida la estructura de los datos, se procedió a verificar y eliminar 1.509.780 registros duplicados correspondientes a usuarios con medida bifásica o trifásica, que para los fines de este trabajo se consideraron como duplicados, con el fin de asegurar que el DataFrame contuviera observaciones únicas. Posteriormente, se evaluó la presencia de valores faltantes en cada columna y se realizó un conteo por variable, garantizando así la completitud del conjunto de datos para su posterior análisis.

Luego, se integró la base de datos de clientes, la cual complementa los registros de consumo con información demográfica y de identificación, brindando una perspectiva integral de los patrones de comportamiento del uso energético, permitiendo análisis segmentados por tipo de clientes. La Figura 3.2 muestra de manera parcial la estructura de la base de datos de clientes.

	PERIODO	PRODUCTO	ZONA	MUNICIPIO	UBICACION	CATEGORIA	SUBCATEGORIA
0	112021	225433263	S.O ZONA CENTRO	POPAYÁN	Urbano	INDUSTRIAL	INDUSTRIAL
1	112021	225542012	S.O ZONA CENTRO	POPAYÁN	Urbano	COMERCIAL	COMERCIAL
2	112021	230075437	S.O ZONA SUR	BALBOA	Rural	RESIDENCIAL	ESTRATO 1
3	112021	233480316	S.O ZONA CENTRO	MORALES	Rural	RESIDENCIAL	ESTRATO 1
4	112021	233481103	S.O ZONA CENTRO	MORALES	Rural	RESIDENCIAL	ESTRATO 1

Figura 3.2: Muestra de la base de datos de clientes con información geográfica y demográfica.

Finalmente se combinan dos conjuntos de datos, el DataFrame de Clientes y el DataFrame de medidas AMI, en un solo DataFrame consolidado. Esta fusión permitió asociar a cada registro de consumo su correspondiente información demográfica y geográfica, facilitando el análisis conjunto de los perfiles de consumo energético.

3.2.1. Limpieza de datos

Como primer paso, se filtraron los datos provenientes de los archivos AMI para conservar únicamente las observaciones correspondientes a tipos específicos de energía activa: “Active Energy Wh”, “Secure Energy in Wh”, “Active Energy WH” y “Active Energy”. Estas fueron almacenadas en un nuevo DataFrame denominado *dfkWh*, como se muestra en la Figura 3.1. A continuación, se aplicó una conversión de tipos de datos: la columna ‘Serial Number’ se transformó a tipo `object`, adecuado para identificadores no numéricos, mientras que las columnas de intervalo horario (‘I1’ a ‘I24’) se convirtieron al tipo `float`, apropiado para valores numéricos continuos. Una vez homogeneizados los tipos, se procedió a unificar el DataFrame.

El siguiente paso consistió en filtrar únicamente los registros de clientes residenciales. Para ello, se seleccionaron las filas del DataFrame donde la columna ‘CATEGORIA’ presentaba el valor ‘RESIDENCIAL’. Posteriormente, se reorganizó la estructura del conjunto de datos seleccionando columnas que representan identificadores y metadatos, agrupando las variables de intervalo horario (columnas que comienzan con ‘I’, representando intervalos horarios de consumo de energía) en una estructura más adecuada para análisis longitudinales.

Los datos fueron transformados a formato largo (*long format*), donde cada fila representa una combinación única entre identificadores de cliente, hora específica y valor de demanda de energía. Asimismo, se normalizó el formato de la columna de hora, eliminando caracteres no numéricos para mantener la consistencia.

3.2.2. Enriquecimiento y preparación de datos de demanda

En este punto, es importante destacar las características principales que enriquecen la información temporal relacionada con la demanda de productos. Primero, se combinan los valores de ciertas columnas con el fin de identificar patrones de demanda específicos para cada serie y producto, además de evitar duplicaciones.

Para enriquecer la información temporal relacionada con la demanda, se aplicaron varias transformaciones adicionales. Primero, se combinaron columnas clave para evitar duplicaciones y permitir una identificación precisa de cada serie de demanda por cliente. Luego, a partir de la variable de fecha, se extrajeron componentes temporales como año, mes, día y hora, los cuales fueron almacenados en columnas independientes. También se categorizó cada observación según el tipo de día: laboral, sábado, domingo o festivo, variables útiles para el análisis del comportamiento por contexto temporal.

Finalmente, se identificaron valores de demanda igual a cero, los cuales podrían distorsionar el análisis estadístico. Estos valores fueron reemplazados por NaN mediante la biblioteca NumPy, permitiendo que fueran correctamente manejados en procesos posteriores de imputación o exclusión.

3.3. Primer proceso de Análisis exploratorio de datos (EDA)

Una vez completada la limpieza y estructuración de los datos, se procedió con un Análisis Exploratorio de Datos (EDA, por sus siglas en inglés) para comprender la distribución y el comportamiento de la demanda eléctrica. En este proceso, se utilizaron estadísticas descriptivas y visualizaciones gráficas, como histogramas y diagramas de

caja (*boxplots*), con el fin de examinar la variabilidad del consumo y su relación con variables temporales. Se prestó especial atención a la detección de valores atípicos, identificación de tendencias y patrones horarios característicos.

El *boxplot* mostrado en la Figura 3.3 resume visualmente la dispersión y concentración de los datos en función de la hora del día. Su aplicación facilita la evaluación de la variabilidad horaria en la demanda, la identificación de picos inusuales y la validación de las transformaciones previas, como la exclusión de valores nulos o cero, que podrían introducir sesgos en el análisis.

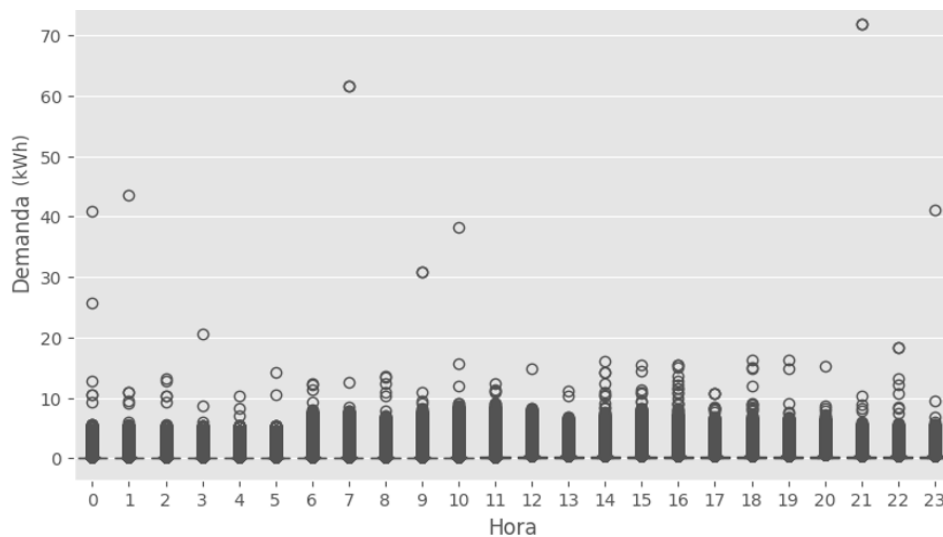


Figura 3.3: Distribución de la Demanda por Hora del Día.

A partir de la gráfica de boxplot se identificó que la mayoría de los valores de demanda se concentran en niveles bajos, aunque existen múltiples valores atípicos en distintas horas, con algunos picos superiores a 60 o 70 unidades. Estos hallazgos sugieren momentos críticos de alta demanda que requieren un análisis detallado. La eliminación de los valores cero, reemplazándolos por NaN, permitió una representación más precisa de la variabilidad real del consumo.

Posteriormente, se analizó la demanda por estrato socioeconómico. La Figura 3.4 evidencia cómo varía el consumo a lo largo del día según el nivel de ingresos. En particular, el estrato 5 presenta un consumo considerablemente mayor, con picos en la mañana y tarde, mientras que el estrato 1 mantiene niveles bajos y estables con un leve incremento nocturno. Los estratos intermedios (2, 3 y 4) muestran patrones

similares al estrato 1, aunque con variaciones en la magnitud del consumo. Por su parte, el estrato 6 presenta el consumo más alto durante casi todo el día, excluyendo al estrato 5, con un pico a las 7:00 horas el cual disminuye un poco a las 8:00 y luego aumenta hasta alcanzar su punto máximo a las 13:00 horas.

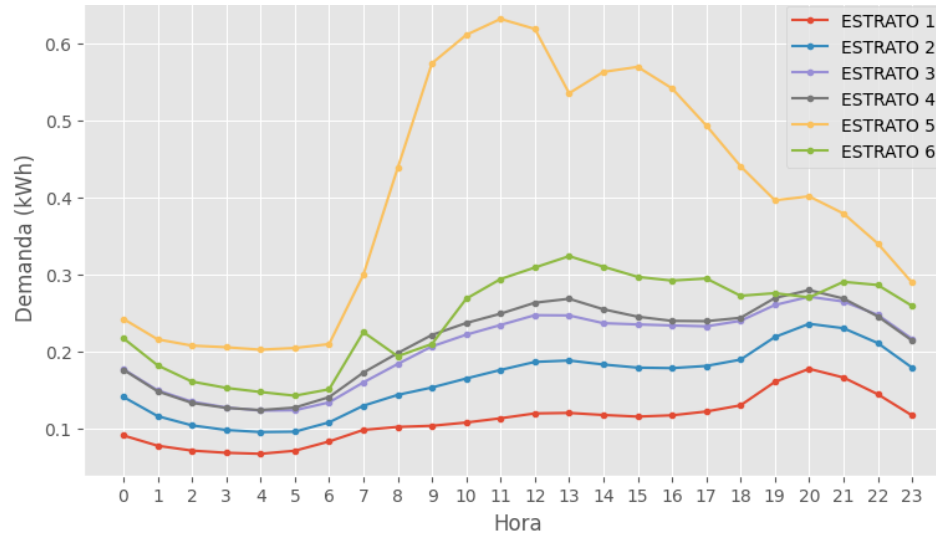


Figura 3.4: Demanda por hora y estrato.

El análisis por municipio permitió evidenciar diferencias en los patrones de consumo. En la Figura 3.5 se observa que Santander de Quilichao tiene el consumo más alto y sostenido, con un pico al mediodía y demanda estable hasta la noche. Popayán sigue un patrón similar, con valores levemente inferiores. Por el contrario, El Tambo y Morales muestran picos de consumo entre las 19:00 y 21:00 horas, lo que sugiere hábitos energéticos nocturnos relacionados con iluminación y uso de electrodomésticos una vez finalizadas las jornadas laborales y escolares. La marcada diferencia en el patrón de consumo respecto a otros municipios indica que el acceso a la energía y los hábitos de uso podrían estar influenciados por factores socioeconómicos, disponibilidad de servicios y estilo de vida de la población.

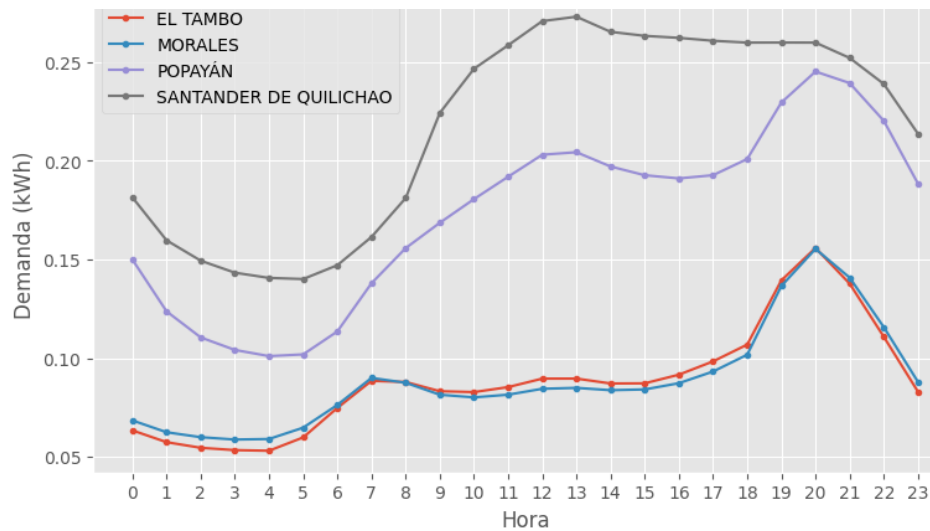


Figura 3.5: Curva de demanda promedio por hora y municipio.

Además de las diferencias en los patrones de consumo, es importante considerar la cantidad de registros disponibles en cada municipio, ya que esto influye en la representatividad de los datos. Popayán posee el mayor número de medidores (7.563), seguido por El Tambo (2.379), Morales (1.328) y Santander de Quilichao (941), lo que sugiere un consumo más elevado por medidor en este último municipio.

Asimismo, se exploraron las diferencias entre zonas rurales y urbanas. La Figura 3.6 muestra que las zonas urbanas tienen un consumo mayor durante todo el día, con un pico entre las 19:00 y 21:00 horas. Las zonas rurales, si bien presentan un consumo más estable, también exhiben un aumento en ese mismo intervalo. Estos patrones de consumo pueden diferir entre zonas debido a factores como densidad poblacional, actividad económica, acceso a infraestructura y hábitos de uso de la energía.

El número de registros disponibles también influye en la representatividad. Las zonas urbanas concentran 8.052 medidores frente a 4.159 en zonas rurales. Sin embargo, ambas presentan patrones definidos, destacando el pico nocturno más pronunciado en la zona rural.

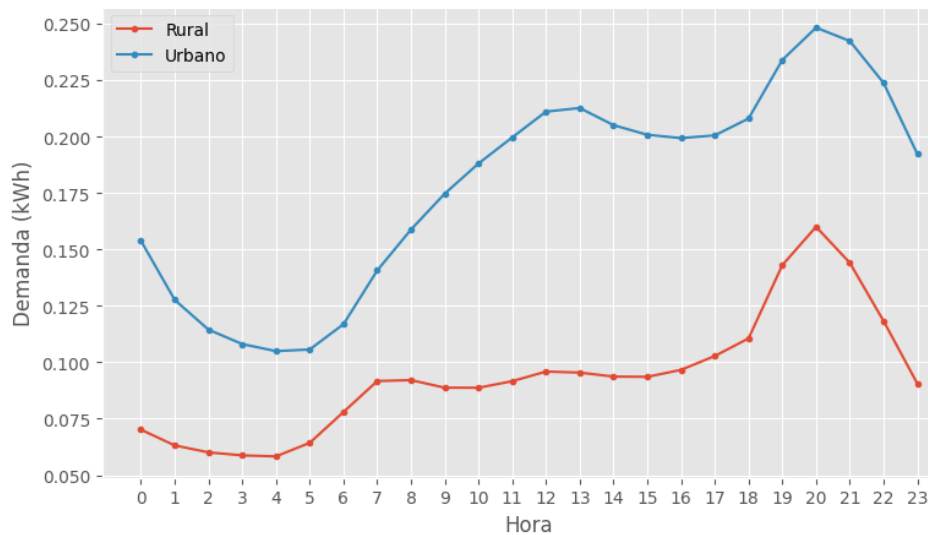


Figura 3.6: Curva de demanda promedio por hora y ubicación (Rural o Urbana).

El consumo de energía no solo varía a lo largo del día, sino también en función del día de la semana. En la Figura 3.7 se representa la curva de consumo promedio por hora y día, donde Se observa un patrón común: baja demanda en la madrugada, aumento progresivo en la mañana y pico entre las 19:00 y 21:00 horas. Sin embargo, los fines de semana presentan niveles más bajos en comparación con los días laborables, reflejando una disminución de la actividad comercial e industrial.

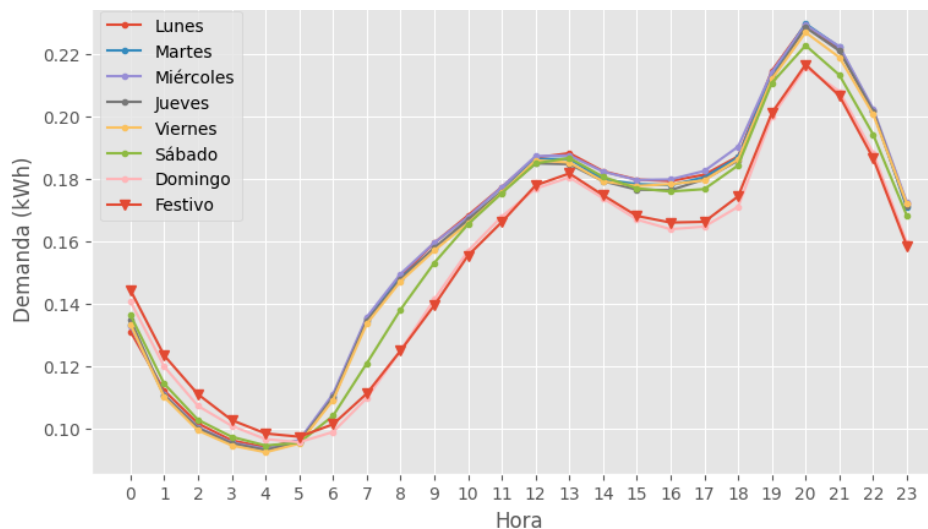


Figura 3.7: Curva de demanda promedio por hora y día de semana.

Se generó también una visualización por tipo de día (laboral, sábado, domingo/festivo). Como se muestra en la Figura 3.8, los días laborables concentran

los mayores niveles de demanda, especialmente en los mismos rangos horarios ya mencionados.

Para evaluar estas variaciones en el consumo energético, se generó también una visualización por tipo de día (laboral, sábado, domingo/festivo). Como se muestra en la Figura 3.8, los días laborables concentran los mayores niveles de demanda, especialmente entre las 19:00 y 21:00. En contraste, los domingos y festivos presentan valores más bajos, lo que sugiere una menor actividad industrial y comercial. Estas diferencias pueden estar influenciadas por horarios laborales, patrones de movilidad, actividades recreativas y uso de electrodomésticos, lo que impacta directamente en la distribución de la carga eléctrica a lo largo del día. Dado que el consumo de energía no solo varía a lo largo del día y entre distintos meses, sino también según la naturaleza del día, es fundamental analizar estas fluctuaciones para ajustar estrategias de distribución de energía y anticipar cambios en la demanda en función del calendario.

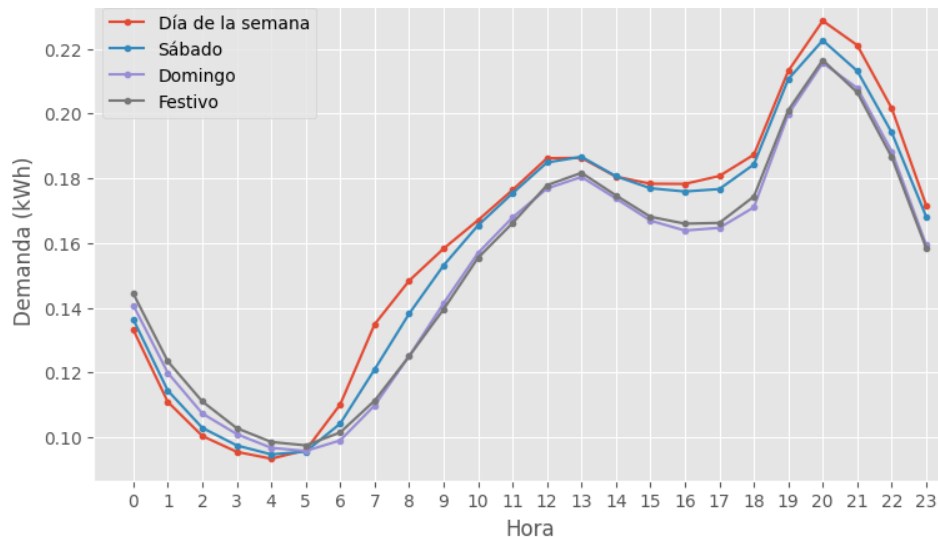


Figura 3.8: Curva de demanda promedio por hora y tipo de día semana con festivos.

La Figura 3.9 agrupa los datos en tres categorías generales: días laborables, sábados y domingos/festivos, lo cual permite una visión más clara de las diferencias en el patrón de consumo de energía.

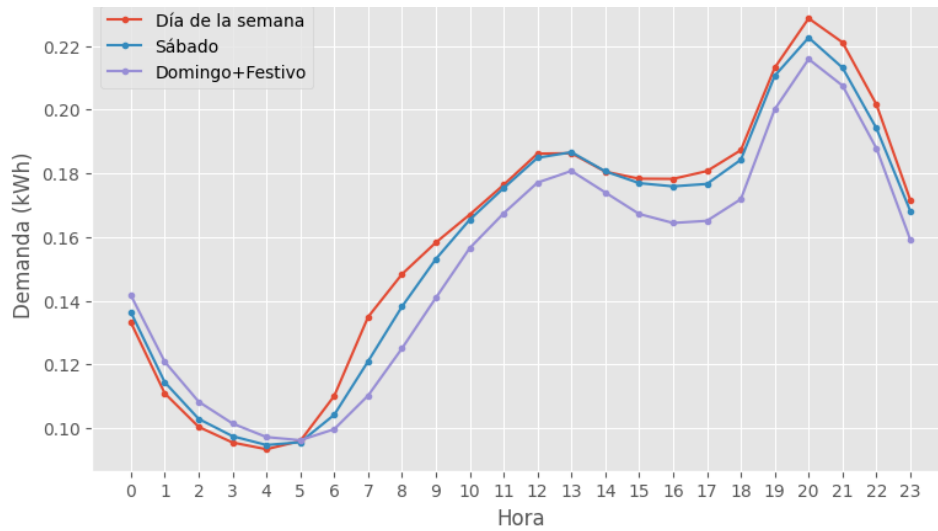


Figura 3.9: Curva de demanda promedio por hora y tipo de día (domingo y festivo juntos).

Finalmente, se agruparon los datos por hora y cliente, calculando el promedio de la demanda energética (kWh) para cada combinación. La Figura 3.10 muestra las curvas de demanda energética promedio por hora (0-23) para distintos clientes, donde se observa que la mayoría de las curvas presentan una concentración de consumo entre 0 y 2 kWh, reflejando un patrón de bajo consumo energético generalizado. Sin embargo, la línea roja destaca por su discontinuidad, lo que sugiere posibles datos faltantes o errores de medición en intervalos específicos. Además, el gráfico revela patrones diferenciados: mientras la mayoría de los usuarios muestran demandas bajas (0-2 kWh), algunos presentan picos residenciales en horas matutinas/nocturnas

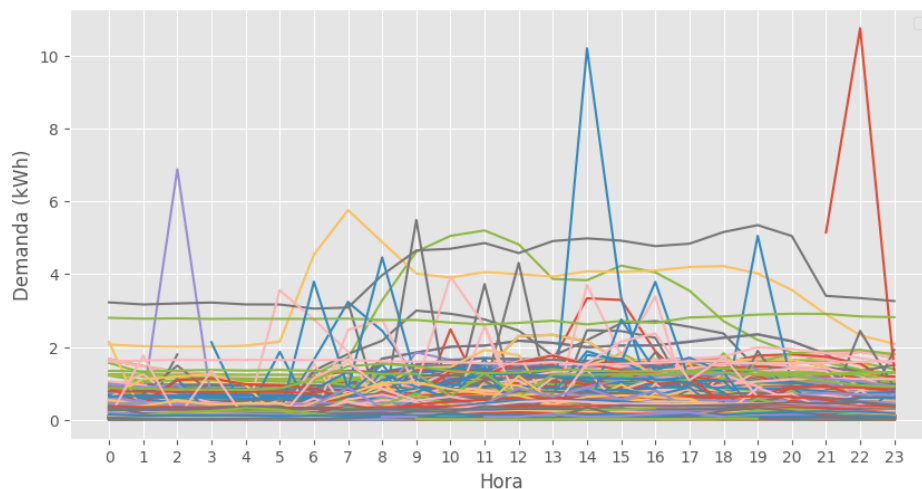


Figura 3.10: Curvas de demanda promedio por hora y cliente.

Para complementar el análisis de la variación en el consumo energético según diferentes criterios temporales, se realizó una caracterización de los usuarios del sistema eléctrico a partir de su ubicación geográfica, tipo de zona (urbana o rural) y estrato socioeconómico.

En la Tabla 3.1 se presenta la distribución de usuarios únicos, resultado del agrupamiento de los datos por municipio, ubicación y subcategoría (estrato socioeconómico), empleando el conteo de valores únicos de la variable `PRODUCTO_SERIAL_TWACS`. Esta información permite identificar la concentración de clientes en determinadas zonas y estratos, lo que es clave para interpretar patrones de consumo y planificar estrategias de gestión de la demanda.

Se observa que los estratos 1 y 2 concentran la mayor cantidad de usuarios en la mayoría de los municipios, mientras que los estratos 5 y 6 presentan una participación mínima, limitada a pocos registros en Popayán urbano.

Tabla 3.1: Distribución de usuarios por municipio, ubicación y subcategoría

MUNICIPIO	UBICACIÓN	SUBCATEGORÍA	Usuarios Únicos
EL TAMBO	Rural	ESTRATO 1	2333
		ESTRATO 2	44
	Urbano	ESTRATO 1	2
MORALES	Rural	ESTRATO 1	1322
		ESTRATO 2	6
POPAYÁN	Rural	ESTRATO 1	388
		ESTRATO 2	59
		ESTRATO 3	6
	Urbano	ESTRATO 1	1320
		ESTRATO 2	3296
		ESTRATO 3	1653
		ESTRATO 4	814
		ESTRATO 5	16
		ESTRATO 6	11
SANTANDER DE QUILICHAO	Rural	ESTRATO 2	1
	Urbano	ESTRATO 1	123
		ESTRATO 2	148
		ESTRATO 3	666
		ESTRATO 4	3

3.4. Segundo proceso de ETL

El segundo proceso de ETL se realiza para adaptar los datos a un formato más adecuado para el análisis específico. Después de la etapa inicial de extracción y transformación básica, este nuevo procesamiento permite reorganizar y agrupar los datos para facilitar comparaciones entre diferentes categorías, como días laborales y fines de semana, o distintos estratos socioeconómicos. Además, permite realizar agregaciones adicionales, tratar valores atípicos o errores en los datos que podrían distorsionar el análisis, y ajustar la granularidad temporal para mejorar la calidad del análisis y las visualizaciones.

El primer paso consiste en calcular la demanda promedio por hora para cada usuario, agrupando los datos por Hora y PRODUCTO_SERIAL_TWACS (identificador del usuario), y luego filtrando aquellos usuarios cuya demanda media supera los 4 kWh. Los registros asociados a estos usuarios son eliminados del DataFrame `df_clear`, permitiendo enfocar el análisis en consumidores más representativos y reduciendo el impacto de valores extremos.

Posteriormente, se filtra el DataFrame `df_clear` para conservar únicamente los registros correspondientes a los estratos 1, 2, 3 y 4. Esta decisión se debe a que los estratos 5 y 6 presentan una cantidad muy reducida de usuarios (16 y 11 registros respectivamente, concentrados únicamente en Popayán urbano), lo que limita su representatividad estadística. Luego, se agrupan los datos por MUNICIPIO, UBICACION y SUBCATEGORIA (estrato), contando la cantidad de usuarios únicos en cada grupo para obtener una visión más detallada de su distribución. Esta distribución de usuarios se presenta en la Tabla 3.2.

Tabla 3.2: Distribución de usuarios por municipio, ubicación y subcategoría

MUNICIPIO	UBICACIÓN	SUBCATEGORÍA	Usuarios Únicos
EL TAMBO	Rural	ESTRATO 1	2333
		ESTRATO 2	44
		Urbano	2
MORALES	Rural	ESTRATO 1	1321
		ESTRATO 2	6
POPAYÁN	Rural	ESTRATO 1	388
		ESTRATO 2	59
		ESTRATO 3	6
	Urbano	ESTRATO 1	1320
		ESTRATO 2	3292
		ESTRATO 3	1653
		ESTRATO 4	811
SANTANDER DE QUILICHAO	Rural	ESTRATO 2	1
	Urbano	ESTRATO 1	123
		ESTRATO 2	148
		ESTRATO 3	665
		ESTRATO 4	3

A continuación, se identifican las combinaciones de municipio, ubicación y subcategoría con al menos 20 usuarios únicos, eliminando aquellas con menor representación. Esta depuración garantiza que el análisis se base en datos significativos y confiables, eliminando combinaciones poco significativas que podrían distorsionar los resultados. De acuerdo a la Tabla 3.3 se puede inferir que ciertos municipios tienen una mayor cantidad de usuarios en estratos bajos, como el estrato 1 en áreas rurales, lo que podría indicar una población de menores recursos. Esta segmentación es útil para la planificación de servicios, ya que permite identificar áreas con mayor demanda y ajustar la asignación de recursos según las características socioeconómicas y geográficas de cada región.

Tabla 3.3: Distribución de usuarios por municipio, ubicación y estrato (filtrados)

MUNICIPIO	UBICACIÓN	SUBCATEGORÍA	Usuarios Únicos
EL TAMBO	Rural	ESTRATO 1	2333
		ESTRATO 2	44
MORALES	Rural	ESTRATO 1	1321
POPAYÁN	Rural	ESTRATO 1	388
		ESTRATO 2	59
		ESTRATO 1	1320
	Urbano	ESTRATO 2	3292
		ESTRATO 3	1653
		ESTRATO 4	811
SANTANDER DE QUILICHAO	Urbano	ESTRATO 1	123
		ESTRATO 2	148
		ESTRATO 3	665

Una vez filtrada la base de usuarios, se aborda la identificación de fechas faltantes debido a que en muchos análisis de series temporales, como el análisis de consumo de servicios, las fechas faltantes pueden representar un desafío importante, ya que la falta de datos para ciertos períodos puede sesgar los resultados o dificultar la interpretación precisa de las tendencias. Debido a lo anterior, se construye un marco de datos completo con todas las combinaciones posibles entre cada cliente y las fechas esperadas, y se aplica una operación *Left Join* con el DataFrame original `df_clear2`. Esto permite detectar y cuantificar los registros faltantes por cliente.

En el análisis de calidad de los datos, se identificó al cliente 277375299_67385223.0 como el que presenta la mayor cantidad de datos faltantes, con un 100 % de ausencia de registros durante el periodo evaluado. Esta ausencia total de información implica que dicho cliente no cuenta con observaciones válidas que puedan ser útiles para el análisis, por lo que se recomienda excluirlo de los estudios posteriores. En contraste, el cliente 554493796_63747474.0 presenta el menor porcentaje de datos faltantes, con tan solo un 1.9 %, es decir, cuenta con una alta disponibilidad de la información. Este bajo nivel de ausencia convierte a este registro en una fuente fiable y representativa. Con base en esta información, se filtran los clientes que cumplen con un umbral de aceptación de hasta 99.9 % de datos faltantes, conservando 11.415 de los 12.157 clientes iniciales (93.9 %), como se muestra a continuación:

Cientes iniciales: 12157
Cientes válidos (Máximo datos faltantes 99.9%): 11415
Porcentaje de clientes válidos: 93.9%

Luego de realizar la selección y filtrado de los clientes con registros válidos, se procede con la detección y tratamiento de valores atípicos, también conocido como *outlier*, la cual es una observación que se encuentra significativamente alejada del resto de los datos, siendo potencialmente resultado de errores de medición, variabilidad natural extrema o condiciones inusuales.

Para la detección de valores atípicos, se aplica una metodología estadística basada en el rango intercuartílico (IQR) aplicado por cliente, tipo de día y hora, donde se calculan los cuartiles Q1 y Q3 y se definen límites de detección con 1.5 veces el IQR. Esta regla es ampliamente utilizada porque logra un adecuado equilibrio entre sensibilidad y robustez al momento de detectar valores alejados del comportamiento típico de los datos. Los valores que se encuentran fuera de estos límites se reemplazan con NaN, indicando explícitamente que tales observaciones no son fiables para el análisis posterior.

El DataFrame resultante se reindexa para mantener su consistencia y claridad estructural tras estos procedimientos de limpieza y preparación, y se prepara para el proceso de imputación. En esta etapa, se trabaja siguiendo los lineamientos del trabajo de grado “*Imputación de datos faltantes usando técnicas estadísticas y de machine learning para los sistemas de infraestructura de medición avanzada (AMI)*” [18], la cual se compone de una muestra con 3559 clientes AMI, excluyendo aquellos con más del 25 % de datos faltantes.

La imputación se realiza mediante el modelo Prophet con regresores (hora del día y tipo de día). El modelo se entrena cliente por cliente, imputando los valores faltantes y conservando la estacionalidad individual y asegurando coherencia con el comportamiento original. El resultado se almacena en la columna `imputation_prophet_regressor` del DataFrame final `imputed_df`. Una vez concluido este proceso, el DataFrame resultante es descargado y almacenado, y sobre el cual se desarrollan las estrategias de clusterización.

Con el DataFrame limpio y estructurado, se procede a una segunda fase de análisis exploratorio de datos (EDA). Esta etapa es fundamental para comprender en profundidad las características estadísticas de los datos imputados y asegurar que la distribución y las relaciones entre variables se mantengan coherentes tras el tratamiento previo. Además, este segundo EDA permite identificar nuevas oportunidades de transformación y evaluar la efectividad de los procesos de limpieza realizados.

PRODUCTO_SERIAL_TWACS	0
Date	0
Start Date	0
ZONA	0
MUNICIPIO	0
UBICACION	0
SUBCATEGORIA	0
Hour	0
Demand	2502539
Year	0
Month	0
Day	0
Year_Month	0
day_of_week	0
Holiday	0
day_of_week_Holi	0
Type_of_day_Holi	0
Demanda_con_Atipicos	3481421
imputation_prophet_regressor	0

La columna `imputation_prophet_regressor` no presenta valores nulos, lo que evidencia que el proceso de imputación con Prophet fue exitoso en cubrir las ausencias de datos de consumo. En contraste, las columnas originales `Demand` y `Demanda_con_Atipicos` aún conservan valores ausentes, con 2.502.539 y 3.481.421 registros faltantes respectivamente. Esta observación refuerza la necesidad de trabajar

con la columna imputada en los análisis posteriores, garantizando así un conjunto de datos completo, coherente y apto para la fase de clusterización.

3.5. Proceso de desagregación

Con el conjunto de datos imputado y libre de valores nulos en la variable de consumo (`imputation_prophet_regressor`), se dispone de una base sólida y consistente para avanzar en los análisis siguientes. A partir de este punto, se da inicio al proceso de desagregación por municipio, con el objetivo de explorar y comparar los patrones de demanda energética a nivel territorial. Este análisis se centrará en los cuatro municipios presentes en la base de datos: Popayán, Morales, Santander de Quilichao y El Tambo. La desagregación seleccionada se realiza con una resolución horaria de 24 puntos por día, distinguiendo tres escenarios representativos:

1. El promedio de consumo horario en días laborales (lunes a viernes).
2. El consumo típico de un sábado.
3. El promedio horario entre domingos y días festivos.

3.5.1. El promedio de consumo horario en días laborales (lunes a viernes)

Para iniciar el análisis desagregado por municipio, se filtran los registros correspondientes exclusivamente a Popayán en días laborales (lunes a viernes), identificados mediante la variable `Type_of_day_Holi`. Posteriormente, se genera una tabla dinámica (pivot table) que permite visualizar el promedio horario de demanda energética por cada medidor (`PRODUCTO_SERIAL_TWACS`). como se presenta en la Figura 3.11. Esta tabla tiene como índice combinado el municipio y el identificador del medidor, y como columnas, las horas del día clasificadas por tipo de día.

		Type_of_day_Holi	Weekday															
		Hour	0	1	2	3	4	5	6	7	8	9	...	14	15	16	17	18
MUNICIPIO	PRODUCTO_SERIAL_TWACS																	
POPAYÁN	275333321_19297356.0	0.181070	0.135473	0.106721	0.089924	0.082628	0.086132	0.098683	0.111208	0.113015	0.102690	...	0.148211	0.132703	0.112126	0.115542	0.158973	
	275338236_19297445.0	0.189720	0.145720	0.138081	0.122255	0.086680	0.065004	0.098219	0.187298	0.284696	0.333493	...	0.187137	0.162730	0.137891	0.154854	0.241353	
	275344087_19297340.0	0.165221	0.122844	0.120825	0.124196	0.115991	0.107886	0.119550	0.152588	0.187350	0.204251	...	0.244553	0.214269	0.183537	0.199017	0.282735	
	275779767_64738314.0	0.045465	0.040216	0.041888	0.049753	0.050741	0.038949	0.030553	0.053407	0.117747	0.197512	...	0.137884	0.199194	0.264375	0.279211	0.227911	
	275789636_64738275.0	0.165219	0.149036	0.132374	0.112368	0.092631	0.079711	0.076599	0.080219	0.085396	0.090703	...	0.151768	0.145785	0.142461	0.153251	0.179113	
	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
	898464476_63746759.0	0.028043	0.028580	0.030230	0.029038	0.026209	0.025659	0.029309	0.035187	0.039973	0.042772	...	0.053701	0.045591	0.041683	0.047665	0.061488	
	898465034_64354401.0	0.066597	0.057930	0.049077	0.042290	0.041794	0.048365	0.056943	0.061039	0.059946	0.060838	...	0.116869	0.095565	0.080964	0.085429	0.106796	
	898466850_63748332.0	0.143673	0.115418	0.098827	0.089531	0.084853	0.086251	0.094761	0.106372	0.113436	0.111224	...	0.101727	0.098384	0.097381	0.110883	0.144155	
	898466926_63748010.0	0.219315	0.183480	0.157468	0.143075	0.135893	0.131966	0.130446	0.132142	0.136820	0.142821	...	0.211038	0.226458	0.234393	0.238638	0.246386	
	898467351_63745254.0	0.050955	0.047539	0.047954	0.038992	0.031036	0.043189	0.078332	0.115199	0.126693	0.106073	...	0.084553	0.068023	0.060521	0.090397	0.157383	

2609 rows × 24 columns

2609 rows × 24 columns

Figura 3.11: Promedio de consumo 24 horas días de la semana Popayán.

Este procedimiento se replica para los demás municipios presentes en el conjunto de datos: Morales, Santander de Quilichao y El Tambo, generando sus respectivas tablas dinámicas con el promedio horario de consumo para los días laborales.

3.5.2. El consumo típico de un sábado

En este caso, se filtran los registros correspondientes exclusivamente a Popayán en sábado, utilizando la variable `day_of_week`. Si bien en otros apartados se emplea la variable `Type_of_day_Holi` para diferenciar entre días laborales, sábados y domingos/festivos, en este análisis es igualmente válido usar `day_of_week`, ya que ambas variables permiten identificar el sábado de forma precisa. La diferencia radica en que `day_of_week` indica directamente el día de la semana (Monday, Tuesday, ..., Saturday, Sunday), mientras que `Type_of_day_Holi` clasifica el día según su naturaleza (Weekday, Saturday, Sunday/Festive). Esta tabla tiene como índice el municipio y el identificador del medidor, y como columnas las horas del sábado. El DataFrame resultante, `df_pivot_sabado`, se ordena cronológicamente para facilitar el análisis comparativo con otros días de la semana. Este mismo procedimiento se aplica a los municipios de Morales, Santander de Quilichao y El Tambo.

3.5.3. El promedio horario entre domingos y días festivos

En este caso, se filtran los registros correspondientes a Popayán en domingos o festivos, utilizando la variable `Type_of_day_Holi`. Se genera una tabla dinámica con el promedio horario de consumo por medidor (`PRODUCTO_SERIAL_TWACS`) en los días mencionados. Esta tabla dinámica tiene como índice combinado el municipio y el

identificador del medidor, y como columnas, las horas del día, organizadas por tipo de día y hora. Este análisis se replica para Morales, Santander de Quilichao y El Tambo, manteniendo la coherencia metodológica en todas las localidades.

3.5.4. Registros de 72 horas por municipio

Una vez generados los DataFrames para los tres tipos de día (laborales, sábados y domingos/festivos), se concatenan los resultados para obtener un solo DataFrame que contenga el comportamiento horario de 72 horas consecutivas. El nuevo DataFrame, `df_72_hours`, se estructura uniendo los 24 valores horarios de cada tipo de día. Posteriormente, se renombran las columnas de 0 a 71 para representar las horas consecutivas del periodo analizado.

Este DataFrame unificado describe el comportamiento de la demanda energética durante tres tipos de días consecutivos y sirve como insumo para la etapa de *clustering*, que permitirá identificar clientes con patrones de consumo similares. En dicha etapa se emplearán los algoritmos DBSCAN, K-Means y K-Medoids, presentados en capítulos anteriores.

3.6. Análisis de Clustering en el Consumo Energético de Clientes

El análisis de clustering tiene como objetivo agrupar a los clientes en segmentos con patrones de consumo energético similares. A través de técnicas de clustering, podemos identificar distintos grupos de clientes con comportamientos de demanda comunes. Este enfoque permite comprender mejor las variaciones en el consumo.

En la metodología propuesta, el proceso de clustering se dividió en dos etapas principales como se puede apreciar en la Figura 3.12: primero, se agrupan los clientes según la forma de sus patrones de consumo utilizando algoritmos como DBSCAN, K-Means o K-Medoids. Posteriormente, dentro de cada grupo formado, se aplica una reducción de dimensionalidad mediante PCA para facilitar un nuevo agrupamiento basado en la magnitud del consumo. Esta estrategia de doble agrupamiento permite

distinguir no solo la forma del comportamiento de consumo, sino también su intensidad, proporcionando una segmentación más precisa y significativa de los clientes.

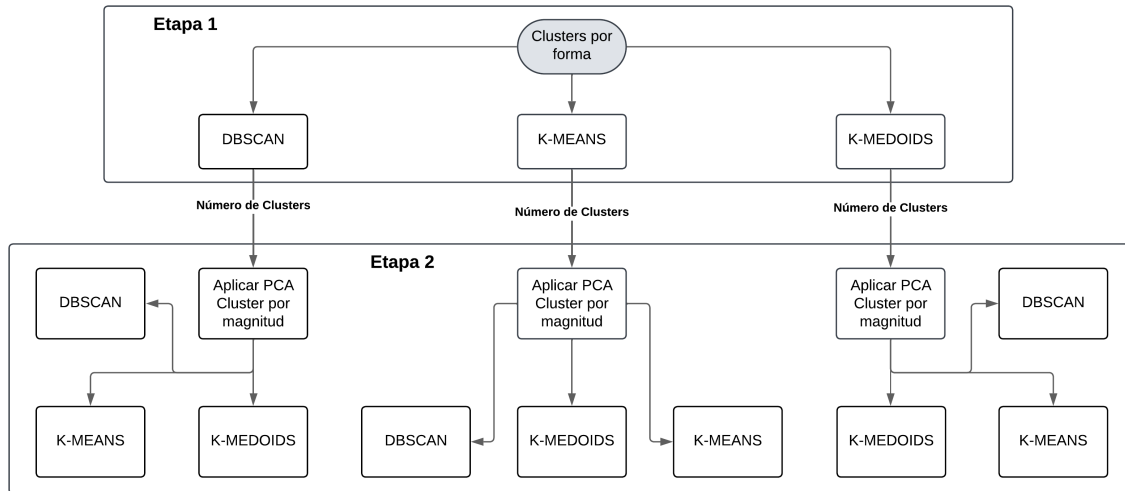


Figura 3.12: Esquema de doble agrupamiento de clientes por forma y magnitud del consumo energético.

3.6.1. Clustering por forma

Volviendo a la Figura 3.12 en la etapa 1, el análisis de *clustering* por forma se enfocó en identificar patrones de consumo energético a través de la estructura temporal de los datos. En lugar de basarse únicamente en características numéricas como el consumo total, este enfoque agrupa los clientes según la similitud en la forma de sus series temporales de demanda, independientemente de la magnitud. Al utilizar técnicas como DBSCAN, K-Means y K-Medoids con distancias específicas, como DTW, se puede identificar grupos de clientes que siguen patrones de consumo similares en términos de la evolución a lo largo del tiempo.

3.6.1.1. Proceso de clustering DBSCAN

Una vez se consolidó el conjunto de datos unificando series temporales de consumo promedio correspondientes a días laborales, sábados y domingos/festivos,

conformando un total de 72 horas continuas, se procedió a normalizar individualmente las series al rango $[0,1]$ para homogenizar las escalas. Posteriormente, se calculó la matriz de distancias entre las series utilizando la métrica *Dynamic Time Warping* (DTW), adecuada para comparar patrones temporales con posibles desplazamientos. A partir de esta matriz, se analizó la distribución de las distancias promedio hacia los tres vecinos más cercanos ($k=3$) para cada serie. Este análisis se graficó con el fin de identificar un valor adecuado para el parámetro *eps*, necesario en el algoritmo de agrupamiento DBSCAN. Dicho valor fue seleccionado observando el punto de inflexión (“codo”) en la curva obtenida vista en la Figura 3.13. Finalmente, se aplicó DBSCAN, logrando agrupar las series en clusters de comportamiento similar y detectando aquellas series que fueron consideradas como ruido.

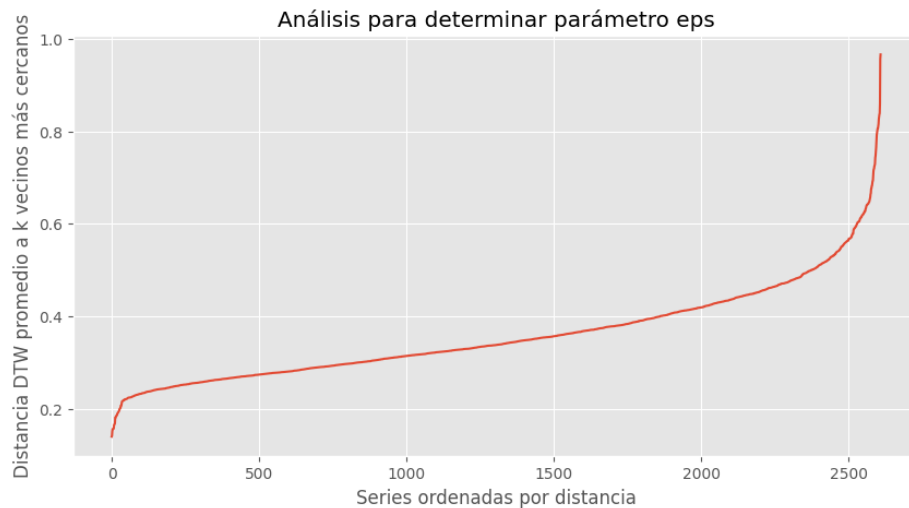


Figura 3.13: Análisis para determinar parámetro *eps* para el municipio de Popayán.

El valor de *eps* se varió entre 0.4 y 0.6, con el objetivo de analizar qué configuración permitía obtener *clusters* con menor cantidad de ruido y una representación adecuada de los datos incluidos.

3.6.1.2. Proceso de clustering K-MEANS

Se aplicó la técnica t-SNE para reducir la dimensionalidad de los datos normalizados a dos componentes principales. Esta reducción permitió representar las series en un espacio bidimensional, facilitando su visualización y posterior agrupamiento.

Con las representaciones bidimensionales obtenidas, se evaluó el número óptimo de *clusters* aplicando el algoritmo KMeans para diferentes valores de `n_clusters` entre 2 y 10. Para cada caso, se calculó el índice Davies-Bouldin, el cual mide la compacidad y separación de los *clusters*, seleccionándose el número de *clusters* (Figura 3.14) que minimizó dicho índice. Una vez determinado el número óptimo de *clusters*, se procedió a realizar la agrupación final mediante KMeans.

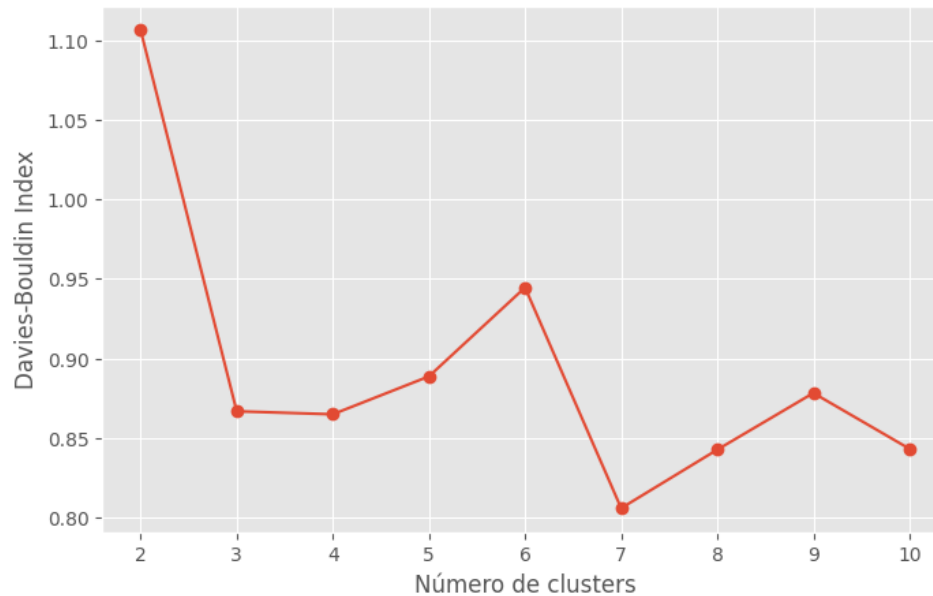


Figura 3.14: Gráfica de la variación de Davies-Bouldin en función del numero de clusters del municipio de Popayán.

3.6.1.3. Proceso de clustering K-MEDOIDS con DTW

Una vez estandarizadas las series, se calculó la matriz de distancias entre todas ellas utilizando la métrica DTW, lo que permitió medir de manera más precisa la similitud entre los diferentes patrones de consumo. A partir de esta matriz de distancias, se aplicó la técnica t-SNE para proyectar los datos en un espacio bidimensional, facilitando tanto el análisis visual como la evaluación de los posibles agrupamientos.

Posteriormente, se exploró el número óptimo de *clusters* mediante la aplicación del algoritmo K-Medoids sobre la matriz de distancias, evaluando cada partición utilizando el índice Davies-Bouldin.

Finalmente, una vez determinado el número óptimo de *clusters*, se aplicó de manera definitiva el algoritmo K-Medoids para formar los grupos de clientes, obteniendo una segmentación adecuada basada en las similitudes de sus perfiles de consumo.

Estos procedimientos fueron aplicados de manera independiente para cada uno de los cuatro municipios considerados en el estudio. En cada caso, se partió de las series temporales de consumo correspondientes, que fueron normalizadas, proyectadas en un espacio bidimensional con t-SNE y finalmente agrupadas utilizando el algoritmo de *clustering*, siguiendo las mismas metodologías descritas. De esta manera, se logró obtener una segmentación específica para cada municipio, respetando las particularidades de sus patrones de consumo.

3.6.2. *Clustering por magnitud*

Volviendo a la Figura 3.12, en la etapa 2, el análisis de *clustering* por magnitud se enfocó en diferenciar a los clientes no por la forma de sus patrones de consumo, sino por la intensidad o magnitud de sus consumos. Para ello, se aplicó una reducción de dimensionalidad mediante PCA sobre cada uno de los *clusters* obtenidos en la etapa 1, extrayendo componentes principales que capturaron las variaciones más relevantes en términos de magnitud. A diferencia de la etapa anterior, en este proceso no se realizó la normalización Min-Max, ya que el objetivo era preservar las diferencias reales en la escala de los consumos. Posteriormente, se realizaron nuevos agrupamientos utilizando técnicas como DBSCAN, K-Means y K-Medoids, esta vez orientadas a segmentar los clientes según el nivel de su consumo energético.

3.6.2.1. *Procedimiento de Aplicación del Análisis de Componentes Principales (PCA)*

Para la etapa 2 del flujo metodológico representado en la Figura 3.12, se aplicó un análisis de componentes principales (PCA) con el objetivo de reducir la dimensionalidad de los datos asociados a la magnitud del consumo. Inicialmente, se realizó una limpieza de los datos, eliminando las columnas no numéricas o irrelevantes, y conservando únicamente aquellas variables pertinentes al consumo energético. Posteriormente, se ejecutó PCA considerando todos los componentes

disponibles, con el fin de evaluar la efectividad del proceso de reducción dimensional.

Como parte del análisis, se calculó la varianza explicada individual y acumulada por cada componente principal. A partir de ello, se construyó un vector de varianza acumulada, lo cual permitió cuantificar progresivamente el porcentaje de la variabilidad total acumulada al incluir sucesivamente cada nuevo componente. El criterio adoptado para la selección fue identificar el número mínimo de componentes k tal que la varianza acumulada cumpliera:

$$\sum_{i=1}^k \lambda_i \geq 0.95 \quad (3.1)$$

Donde λ_i representa la proporción de varianza explicada por el i -ésimo componente principal. Esta estrategia aseguró que al menos el 95 % de la variabilidad total de los datos originales fuera preservada en el nuevo espacio de características, garantizando así una representación eficiente y evitando la pérdida significativa de información relevante [19].

En los casos donde se identificaron patrones de consumo similares entre distintos *clusters* generados previamente, estos fueron unificados antes de aplicar la reducción de dimensionalidad, con el objetivo de mejorar la representatividad interna de los grupos y fortalecer la coherencia del análisis.

3.6.2.2. Configuración de los algoritmos de clustering por magnitud dentro de Implementación

Para realizar el agrupamiento por magnitud, se aplicó el algoritmo de *clustering* DBSCAN sobre los datos proyectados mediante PCA. Inicialmente, se configuró el algoritmo con un parámetro `min_samples` igual a 5, mientras que el valor óptimo de `eps` fue determinado empíricamente mediante la evaluación de distintos puntos en el rango de 0.5 a 10, con incrementos constantes de 0.3. Para cada valor de `eps` considerado, se calculó el índice Davies-Bouldin, seleccionándose aquel valor que minimizó este índice, garantizando así una mejor compactación y separación de los subgrupos formados. Una vez determinado el valor óptimo de `eps`, se aplicó DBSCAN de forma definitiva para

asignar etiquetas de subcluster a cada observación en el espacio PCA. Posteriormente, se incorporaron estas etiquetas como una nueva columna en el conjunto de datos. Además, se generaron matrices de dispersión entre todas las combinaciones de componentes principales para visualizar la distribución de los subclusters identificados, facilitando el análisis visual de la segmentación lograda a partir de las magnitudes de consumo energético.

Para la segmentación basada en magnitud utilizando K-Means, se evaluó inicialmente un rango de posibles cantidades de *clusters*, variando el número de grupos entre 2 y 10. Para cada valor de k considerado, se aplicó el algoritmo K-Means sobre los datos reducidos mediante PCA y se calculó el índice Davies-Bouldin, el cual permitió medir la compactación y separación de los *clusters* formados. El número óptimo de *clusters* fue determinado seleccionando el valor de k que minimizó dicho índice, garantizando una estructura de agrupamiento más robusta. Posteriormente, se aplicó K-Means de forma definitiva utilizando este número óptimo de *clusters* y se asignaron las etiquetas de subcluster a cada observación, incorporándolas como una nueva columna en el conjunto de datos. Además, se generó una curva del índice Davies-Bouldin en función del número de *clusters*.

Para la segmentación por magnitud utilizando el algoritmo K-Medoids, se trabajó directamente con los componentes principales obtenidos mediante PCA, seleccionando aquellas columnas asociadas a las variables PC_i , donde cada PC_i representa un nuevo eje que captura una parte de la variabilidad total de los datos, ordenados de mayor a menor importancia. Se empleó la distancia Manhattan como métrica para medir la similitud entre observaciones, en lugar de la distancia Euclidiana tradicional, con el objetivo de capturar mejor las diferencias absolutas en los consumos. Posteriormente, se evaluaron diferentes cantidades de *clusters*, variando k entre 2 y 10, inicializando aleatoriamente los medoides en cada caso. Para cada valor de k considerado, se aplicó K-Medoids y se calcularon las etiquetas de cluster, evaluándose la calidad del agrupamiento mediante el índice Davies-Bouldin.

Una vez obtenidos los subclusters resultantes de la segmentación por magnitud, se procedió a calcular un conjunto de métricas estadísticas que permiten caracterizar el comportamiento del consumo dentro de cada grupo. Estas métricas fueron calculadas con base en el consumo total mensual por cliente, y se agruparon según las etiquetas

de subcluster generadas por cada método de *clustering*. El análisis incluyó indicadores como la media, mediana, mínimo, máximo y desviación estándar del consumo, con el objetivo de establecer comparaciones entre los subclusters formados por DBSCAN, K-Means y K-Medoids. Esta caracterización estadística constituye la base para la interpretación de los resultados y la validación de la coherencia de los agrupamientos obtenidos.

3.7. Métricas de Evaluación y Caracterización de Clusters

Con el objetivo de evaluar la calidad de los agrupamientos generados en ambas etapas del proceso de *clustering* (por forma y por magnitud), se aplicaron distintas métricas que permitieron validar tanto la cohesión interna de los *clusters* como interpretar sus características particulares.

Para la etapa de agrupamiento basada en la forma de los perfiles de consumo, se utilizó el índice Davies-Bouldin (DB) como principal métrica de evaluación. Este índice cuantifica la compacidad intra-cluster y la separación entre *clusters*, penalizando aquellos agrupamientos en los que los *centroides* se encuentran cercanos o con alta dispersión interna. Un valor más bajo del índice DB indica una mejor estructuración del agrupamiento. [20]

En la etapa de análisis por magnitud, se hizo énfasis en la caracterización interna de los grupos a partir de las propiedades estadísticas del consumo energético mensual. Para ello, se calcularon las siguientes métricas sobre los subgrupos identificados:

- **Promedio (media) del consumo mensual:** Permite estimar el nivel medio de demanda energética dentro de cada *cluster*.
- **Desviación estándar:** Refleja la dispersión de los valores respecto a la media, indicando qué tan homogéneo es el grupo.
- **Coefficiente de variación (CV):** Calculado como la relación entre la desviación estándar y la media, proporciona una medida relativa de dispersión. Se define

como:

$$CV = \frac{\sigma}{\mu} \quad (3.2)$$

donde σ representa la desviación estándar y μ la media del consumo. Valores más bajos del CV indican mayor estabilidad y homogeneidad del grupo. [21]

- **Error cuadrático medio (RMS) de las curvas promedio:** Esta métrica fue empleada para comparar la forma promedio de consumo dentro de cada *cluster* con respecto a una curva representativa. Un RMS bajo sugiere que las curvas individuales son similares a la curva promedio del grupo, lo que respalda la coherencia interna del *cluster*. [22]

3.8. Resumen

El Capítulo 3 presenta el desarrollo metodológico del análisis de la información y la implementación de técnicas de *clustering* para la segmentación del consumo energético de los clientes. Inicialmente, se describen las características de la demanda y se realiza un proceso completo de extracción, transformación y carga (ETL), dividido en dos etapas. En la primera, se llevó a cabo la limpieza, enriquecimiento y preparación de los datos; en la segunda, se profundizó en el análisis exploratorio, desagregando el comportamiento horario del consumo en días laborales, sábados y domingos/festivos, obteniendo curvas de demanda promedio de 72 registros por cliente y municipio.

Posteriormente, se implementaron técnicas de *clustering* para identificar patrones de consumo similares entre los clientes. El proceso se dividió en dos enfoques: *clustering* por forma y *clustering* por magnitud. El primero se centró en la similitud de las curvas de consumo a lo largo del tiempo, aplicando algoritmos como DBSCAN, K-Means y K-Medoids con distancia DTW. El segundo se enfocó en las diferencias de magnitud de consumo, aplicando reducción de dimensionalidad mediante PCA, seguido de la aplicación nuevamente de los mismos algoritmos de *clustering* sobre los componentes principales. Cada técnica fue configurada y evaluada de manera individual utilizando el índice Davies-Bouldin para determinar el número óptimo de agrupamientos.

Capítulo 4

Evaluación y Análisis de Resultados

Este capítulo presenta la evaluación de los resultados obtenidos al aplicar técnicas de *clustering* sobre series temporales de consumo energético. El análisis se centra en identificar patrones de comportamiento de los usuarios, considerando tanto la forma como la magnitud de sus curvas de demanda. Para ello, se implementaron los algoritmos DBSCAN, K-Means y K-Medoids, utilizando la distancia *Dynamic Time Warping* (DTW) como métrica de similitud temporal. Este análisis se realizó de forma independiente para cada uno de los municipios incluidos en el estudio, lo que permitió capturar particularidades territoriales en los patrones de consumo.

El proceso de segmentación se desarrolló en dos etapas: la primera, enfocada en las tendencias de consumo (forma), y la segunda, orientada a la intensidad del consumo energético (magnitud), previa reducción de la dimensionalidad mediante un Análisis de Componentes Principales (PCA), permitiendo obtener agrupamientos más compactos y diferenciables en términos de intensidad de consumo. Los agrupamientos obtenidos se representaron visualmente en espacios de dimensión reducida, y su calidad fue evaluada mediante el índice de Davies-Bouldin. Además, se realizó una caracterización estadística de los subgrupos con base en indicadores como la media, mediana, desviación estándar, RMS y coeficiente de variación (CV). Finalmente, se identifican y discuten los patrones de consumo más relevantes hallados en cada municipio, destacando su utilidad en la caracterización de usuarios y en la formulación de estrategias para una gestión más eficiente de la demanda eléctrica.

A diferencia de los enfoques convencionales reportados en la literatura, como los presentados por [2] en su análisis de big data para AMI y [4] en agrupamiento de series temporales, este trabajo demuestra que la combinación sinérgica de clustering por forma (basado en DTW) y magnitud (mediante PCA) permite una segmentación significativamente más granular de usuarios AMI.

4.1. Análisis de *Clustering* para el Municipio de Popayán

La Figura 4.1 muestra los resultados del agrupamiento de usuarios en el municipio de Popayán, según la forma de sus curvas de consumo energético. Se aplicaron los algoritmos DBSCAN, K-Medoids con DTW y K-Means, y los usuarios fueron proyectados en un espacio bidimensional. Cada punto representa un usuario, y los colores indican la asignación a un *clúster* específico según el método empleado.

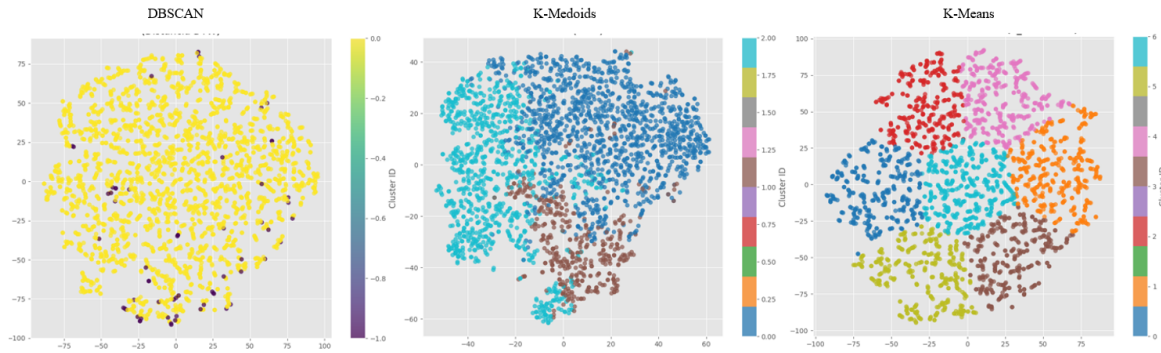


Figura 4.1: Clústeres por Forma, municipio de Popayán.

Popayán muestra una notable diversidad en los perfiles de consumo energético, lo cual se evidencia en los resultados obtenidos mediante el enfoque de agrupamiento en dos etapas: una primera fase basada en la forma de las curvas temporales y una segunda orientada a la magnitud del consumo dentro de cada grupo previamente definido. Las Tablas 4.1, 4.2 y 4.3 resumen los indicadores estadísticos asociados a cada técnica.

En las tablas se destacan algunas filas porque sus valores representan clusters consistentes y accionables, es decir, grupos de clientes que muestran patrones de consumo eléctrico claros y diferenciados, lo que permite utilizarlos como base para la toma de decisiones estratégicas.

Tabla 4.1: Análisis de métricas por cluster DBSCAN para el municipio de Popayán

Forma	DB Index	N° Clúster (Forma)	Cientes (Forma)	Magnitud	N° Clúster	Cientes	Desv. Est.	Media	CV	RMS
DBSCAN -		1	2471	DBSCAN (0)	0	2471	5.31	11.19	0.47	0.16
				DBSCAN (-1)	-1	138	-	-	-	-
				K-MEANS (0)	0	2437	4.81	10.89	0.44	0.16
				K-MEANS (1)	1	172	15.29	38.03	0.40	0.56
				K-MEDOIDES (0)	0	1023	10.66	19.90	0.33	0.29
				K-MEDOIDES (1)	1	1586	2.62	8.03	0.54	0.12

Tabla 4.2: Análisis de métricas por cluster para el municipio de Popayán (K-MEDOIDES CON DTW)

Forma	DB Index	N° Clúster (Forma)	Cientes (Forma)	Magnitud	N° Clúster	Cientes	Desv. Est.	Media	CV	RMS
K-MEDOIDES CON DTW	0,98	0	1295	DBSCAN (0)	0	1274	5.58	11.37	0.49	0.17
				DBSCAN (1)	1	6	1.01	38.60	0.026	0.56
				DBSCAN (-1)	-1	15	-	-	-	-
				K-MEANS (0)	0	1285	6.02	11.58	0.52	0.17
				K-MEANS (1)	1	10	21.40	70.12	0.30	1.02
				K-MEDOIDES (0)	0	516	9.51	18.25	0.52	0.27
				K-MEDOIDES (1)	1	779	2.50	7.91	0.31	0.12
K-MEDOIDES CON DTW	0,98	1	395	DBSCAN (0)	0	377	4.32	11.20	0.38	0.16
				DBSCAN (-1)	-1	18	-	-	-	-
				K-MEANS (0)	0	262	2.48	8.90	0.29	0.13
				K-MEANS (1)	1	10	9.28	42.05	0.22	0.63
				K-MEANS (2)	2	123	3.49	17.02	0.20	0.25
				K-MEDOIDES (0)	0	221	7.04	16.01	0.42	0.16
				K-MEDOIDES (1)	1	174	1.76	7.51	0.23	0.26
K-MEDOIDES CON DTW	0.98	2	919	DBSCAN (0)	0	915	10.13	13.49	0.75	0.20
				DBSCAN (-1)	-1	4	-	-	-	-
				K-MEANS (0)	0	84	14.31	41.01	0.35	0.62
				K-MEANS (1)	1	835	5.40	11.03	0.49	0.16
				K-MEDOIDES (0)	0	83	14.36	41.12	0.35	0.62
				K-MEDOIDES (1)	1	490	2.52	7.32	0.34	0.24
				K-MEDOIDES (2)	2	346	3.85	16.35	0.23	0.11

En particular, la combinación metodológica que ofreció los mejores resultados fue la utilización de K-MEANS durante la etapa de agrupamiento por forma. Como se muestra en la Tabla 4.3, este algoritmo logró una buena compactación y separación entre grupos, alcanzando un índice Davies-Bouldin de 0.8057, inferior al obtenido con K-MEDOIDES (0.98). Esto indica una mayor cohesión interna y mayor distancia entre grupos.

Durante la segunda etapa de agrupamiento, orientada a la magnitud del consumo, K-MEANS volvió a destacar por su capacidad para diferenciar perfiles. Por ejemplo, dentro del *Cluster* por forma 1, se identificaron dos subgrupos claramente diferenciados: uno con un consumo promedio de 67.85, un RMS de 0.99 y un CV de 0.32, frente a otro con una media de 8.62, RMS de 0.17 y CV de 0.26.

Tabla 4.3: Análisis de métricas por cluster para el municipio de Popayán, método K-MEANS

Forma	DB Index	N° Clúster (Forma)	Cientes (Forma)	Magnitud	N° Clúster	Cientes	Desv. Est.	Media	CV	RMS
K-MEANS	0.8057	0	355	DBSCAN (0)	0	334	5.01	12.00	0.42	0.17
				DBSCAN (-1)	-1	21	—	—	—	—
				K-MEANS (0)	0	222	2.75	9.14	0.30	0.13
				K-MEANS (1)	1	121	3.73	18.07	0.21	0.26
				K-MEANS (2)	2	2	0.01	42.44	0.00	0.66
				K-MEANS (3)	3	10	6.38	40.42	0.15	0.59
				K-MEDOIDS (0)	0	170	1.97	11.72	0.17	0.29
				K-MEDOIDS (1)	1	87	1.50	6.26	0.24	0.60
				K-MEDOIDS (2)	2	87	3.70	19.65	0.19	0.18
K-MEANS	0.8057	1	403	DBSCAN (0)	0	389	5.71	12.01	0.48	0.17
				DBSCAN (-1)	-1	14	—	—	—	—
				K-MEANS (0)	0	392	6.07	12.19	0.50	0.18
				K-MEANS (1)	1	11	21.65	67.85	0.32	0.99
				K-MEDOIDS (0)	0	224	2.51	8.21	0.31	0.10
				K-MEDOIDS (1)	1	168	5.34	17.49	0.30	0.39
				K-MEDOIDS (2)	2	11	21.65	67.85	0.34	0.18
K-MEANS	0.8057	2	344	DBSCAN (0)	0	323	4.33	9.01	0.48	0.14
				DBSCAN (-1)	-1	21	—	—	—	—
				K-MEANS (0)	0	323	4.14	8.93	0.46	0.13
				K-MEANS (1)	1	21	14.03	35.19	0.40	0.54
				K-MEDOIDS (0)	0	145	2.67	11.94	0.22	0.08
				K-MEDOIDS (1)	1	31	13.86	29.90	0.46	0.15
				K-MEDOIDS (2)	2	168	1.73	5.75	0.30	0.67
K-MEANS	0.8057	3	335	DBSCAN (0)	0	329	5.29	12.26	0.43	0.18
				DBSCAN (-1)	-1	6	—	—	—	—
				K-MEANS (0)	0	290	3.50	10.84	0.32	0.16
				K-MEANS (1)	1	45	5.94	24.58	0.28	0.36
				K-MEDOIDS (0)	0	136	2.57	15.30	0.17	0.44
				K-MEDOIDS (1)	1	22	4.40	29.66	0.15	0.25
				K-MEDOIDS (2)	2	177	2.23	8.62	0.26	0.17
K-MEANS	0.8057	4	375	DBSCAN (0)	0	370	5.01	9.75	0.51	0.14
				DBSCAN (-1)	-1	5	—	—	—	—
				K-MEANS (0)	0	24	4.76	24.63	0.19	0.37
				K-MEANS (1)	1	216	1.95	6.37	0.31	0.09
				K-MEANS (2)	2	135	2.42	13.32	0.18	0.20
				K-MEDOIDS (0)	0	144	4.96	15.53	0.32	0.23
				K-MEDOIDS (1)	1	231	2.11	6.61	0.33	0.10
K-MEANS	0.8057	5	374	DBSCAN (0)	0	315	5.68	12.89	0.44	0.19
				DBSCAN (1)	1	8	1.01	41.21	0.02	0.60
				DBSCAN (-1)	-1	15	—	—	—	—
				K-MEANS (0)	0	53	13.97	41.95	0.33	0.64
				K-MEANS (1)	1	321	5.57	12.93	0.43	0.19
				K-MEDOIDS (0)	0	317	5.38	12.76	0.42	0.66
				K-MEDOIDS (1)	1	57	14.03	40.88	0.34	0.12

Tabla 4.4: Análisis de métricas por cluster para el municipio de Popayán, método K-MEANS. Continuación.

Forma	DB Index	N° Clúster (Forma)	Cientes (Forma)	Magnitud	N° Clúster	Cientes	Desv. Est.	Media	CV	RMS
K-MEANS	0.8057	6	423	DBSCAN (0)	0	415	4.97	10.68	0.46	0.16
				DBSCAN (-1)	-1	8	—	—	—	—
				K-MEANS (0)	0	95	1.22	4.84	0.25	0.07
				K-MEANS (1)	1	35	1.41	19.38	0.07	0.28
				K-MEANS (2)	2	1	—	68.46	—	1.05
				K-MEANS (3)	3	95	0.97	11.51	0.08	0.17
				K-MEANS (4)	4	3	6.43	37.85	0.17	0.57
				K-MEANS (5)	5	2	0.46	69.35	0.01	1.09
				K-MEANS (6)	6	12	2.27	25.12	0.09	0.37
				K-MEANS (7)	7	60	1.25	14.99	0.08	0.22
				K-MEANS (8)	8	119	1.07	8.67	0.12	0.13
				K-MEANS (9)	9	1	—	57.60	—	0.88
				K-MEDOIDS (0)	0	104	1.28	5.00	0.25	0.08
				K-MEDOIDS (1)	1	146	1.13	9.25	0.12	0.27
				K-MEDOIDS (2)	2	5	10.94	61.83	0.18	0.94
				K-MEDOIDS (3)	3	70	4.17	19.86	0.21	0.16
				K-MEDOIDS (4)	4	98	1.30	12.98	0.10	0.17

Los clusters por magnitud que sean consistentes y accionables se identifican mediante las métricas que son: Media (valor central del consumo), Desviación Estándar (dispersión absoluta), Coeficiente de Variación (CV) (dispersión relativa), y RMS (error de agrupamiento).

4.2. Análisis de *Clustering* para el Municipio de Morales

La Figura 4.2 se presentan de manera comparativa los resultados obtenidos al aplicar tres técnicas de agrupamiento DBSCAN, K-Medoids con DTW y K-Means a un conjunto de usuarios del municipio de Morales, considerando como criterio de análisis la similitud en las formas de sus curvas de consumo energético. Para facilitar la interpretación, los resultados fueron proyectados en un espacio bidimensional, permitiendo observar la distribución y separación de los grupos generados por cada método.

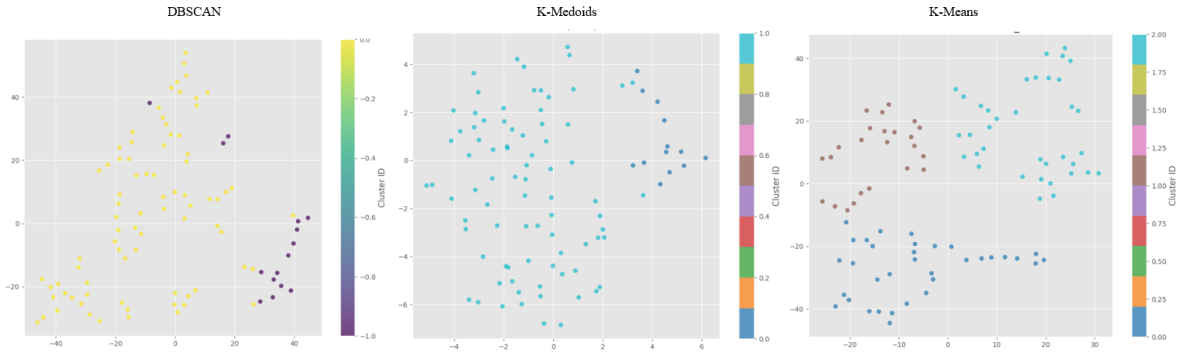


Figura 4.2: Clústeres por Forma, municipio de Morales.

El análisis de segmentación realizado para el municipio de Morales, basado en las métricas presentadas en las Tablas 4.5, 4.6 y 4.7, evidencia un comportamiento energético relativamente más homogéneo y menos disperso que el observado en municipios como Popayán. Esta relativa uniformidad se sustenta en varios indicadores: los consumos promedio por subgrupo no superan los 14 kWh, y en la mayoría de los casos, se encuentran en el rango de 4 a 9 kWh. Las desviaciones estándar son bajas, oscilando entre 1 y 2.6, mientras que los valores de RMS rara vez superan 0.13, lo que sugiere curvas de consumo con magnitud moderada y poca variabilidad. Asimismo, los coeficientes de variación (CV) se mantienen dentro de rangos bajos, sin extremos elevados, lo que refuerza la evidencia de una mayor homogeneidad interna dentro de cada subgrupo. A diferencia de otros municipios donde se observan subgrupos con promedios superiores a 60 kWh y desviaciones superiores a 20, Morales muestra una estructura más compacta, sin presencia de consumidores fuertemente atípicos.

Tabla 4.5: Análisis de métricas por cluster DBSCAN para el municipio de Morales.

Forma	DB Index	N° Clúster (Forma)	Cientes (Forma)	Magnitud	N° Clúster	Cientes	Desv. Est.	Media	CV	RMS
DBSCAN	-	0	79	DBSCAN (0)	0	75	2.03	5.72	0.35	0.09
				DBSCAN (-1)	-1	4	-	-	-	-
				K-MEANS (0)	0	51	1.12	4.58	0.25	0.07
				K-MEANS (1)	1	28	1.98	8.70	0.23	0.13
				K-MEDOIDS (0)	0	29	2.02	8.60	0.23	0.07
				K-MEDOIDS (1)	1	50	1.12	4.55	0.25	0.13

Tabla 4.6: Análisis de métricas por cluster K-MEDOIDS para el municipio de Morales.

Forma	DB Index	N° Clúster (Forma)	Clientes (Forma)	Magnitud	N° Clúster	Clientes	Desv. Est.	Media	CV	RMS
K-MEDOIDS	0.75	0	41	DBSCAN (0)	0	40	6.15	5.41	1.14	0.10
				DBSCAN (-1)	-1	1	—	—	—	—
				K-MEANS (0)	0	2	0.32	7.50	0.04	0.14
				K-MEANS (1)	1	10	0.43	5.23	0.08	0.08
				K-MEANS (2)	2	2	0.67	8.47	0.08	0.13
				K-MEANS (3)	3	1	—	14.18	—	0.21
				K-MEANS (4)	4	1	—	11.74	—	0.20
				K-MEANS (5)	5	10	0.06	5.60	0.01	0.09
				K-MEANS (6)	6	8	0.60	3.84	0.16	0.06
				K-MEANS (7)	7	4	0.72	8.10	0.09	0.13
				K-MEANS (8)	8	1	—	10.86	—	0.16
				K-MEANS (9)	9	2	1.05	10.10	0.10	0.16
				K-MEDOIDS (0)	0	6	1.12	8.81	0.13	0.13
				K-MEDOIDS (1)	1	1	—	14.18	—	0.21
				K-MEDOIDS (2)	2	5	0.36	7.02	0.05	0.11
				K-MEDOIDS (3)	3	25	0.58	5.24	0.11	0.09
				K-MEDOIDS (4)	4	4	1.76	9.92	0.18	0.17
K-MEDOIDS	0.75	1	40	DBSCAN (0)	0	38	1.99	5.37	0.37	0.08
				DBSCAN (-1)	-1	2	—	—	—	—
				K-MEANS (0)	0	13	0.63	5.07	0.12	0.07
				K-MEANS (1)	1	2	0.79	9.01	0.09	0.13
				K-MEANS (2)	2	3	0.26	7.72	0.03	0.13
				K-MEANS (3)	3	11	0.55	3.06	0.18	0.04
				K-MEANS (4)	4	1	—	13.79	—	0.21
				K-MEANS (5)	5	7	0.51	6.47	0.08	0.10
				K-MEANS (6)	6	2	0.89	8.93	0.10	0.13
				K-MEANS (7)	7	1	—	8.06	—	0.17
				K-MEDOIDS (0)	0	8	—	13.79	—	0.21
				K-MEDOIDS (1)	1	1	0.69	8.97	0.08	0.13
				K-MEDOIDS (2)	2	14	—	8.06	—	0.17
				K-MEDOIDS (3)	3	4	0.59	6.60	0.09	0.10
				K-MEDOIDS (4)	4	10	0.17	7.59	0.02	0.13
				K-MEDOIDS (5)	5	2	0.68	4.99	0.14	0.07
				K-MEDOIDS (6)	6	1	0.50	2.98	0.18	0.04

Tabla 4.7: Análisis de métricas por cluster K-MEANS para el municipio de Morales

Forma	DB Index	N° Clúster (Forma)	Cientes (Forma)	Magnitud	N° Clúster	Cientes	Desv. Est.	Media	CV	RMS
K-MEANS	0,97	0	25	DBSCAN (0)	0	16	1,10	4,57	0,24	0,07
				DBSCAN (-1)	-1	6	—	—	—	—
				K-MEANS (0)	0	4	0,33	3,02	0,11	0,04
				K-MEANS (1)	1	1	—	8,45	—	0,12
				K-MEANS (2)	2	10	0,48	4,55	0,11	0,07
				K-MEANS (3)	3	2	0,51	6,39	0,08	0,10
				K-MEANS (4)	4	1	—	9,57	—	0,14
				K-MEANS (5)	5	2	0,13	5,00	0,03	0,09
				K-MEANS (6)	6	3	0,23	5,44	0,04	0,08
				K-MEANS (7)	7	1	—	8,31	—	0,13
				K-MEANS (8)	8	1	—	9,17	—	0,13
				K-MEDOIDES (0)	0	10	0,48	4,55	0,11	0,07
				K-MEDOIDES (1)	1	1	—	6,39	—	0,10
				K-MEDOIDES (2)	2	2	0,51	6,39	0,08	0,08
				K-MEDOIDES (3)	3	4	0,33	3,02	0,11	0,04
				K-MEDOIDES (4)	4	4	0,60	5,73	0,10	0,13
				K-MEDOIDES (5)	5	3	0,46	9,57	0,05	0,14
K-MEANS	0,97	1	36	DBSCAN (0)	0	32	1,86	5,23	0,36	0,08
				DBSCAN (-1)	-1	4	—	—	—	—
				K-MEANS (0)	0	16	1,19	7,08	0,17	0,10
				K-MEANS (1)	1	3	1,82	12,93	0,06	0,20
				K-MEANS (2)	2	17	0,86	3,74	0,23	0,06
				K-MEDOIDES (0)	0	6	1,12	8,14	0,14	0,04
				K-MEDOIDES (1)	1	1	—	13,79	—	0,07
				K-MEDOIDES (2)	2	8	0,41	2,92	0,14	0,11
				K-MEDOIDES (3)	3	1	—	10,84	—	0,10
				K-MEDOIDES (4)	4	8	0,65	6,60	0,11	0,17
				K-MEDOIDES (5)	5	11	0,63	4,70	0,06	0,21
				K-MEDOIDES (6)	6	1	—	14,17	—	0,21
K-MEANS	0,97	2	33	DBSCAN (0)	0	24	1,18	5,26	0,22	0,08
				DBSCAN (-1)	-1	9	—	—	—	—
				K-MEANS (0)	0	3	0,39	7,69	0,05	0,15
				K-MEANS (1)	1	10	0,89	5,77	0,15	0,10
				K-MEANS (2)	2	3	1,11	9,58	0,12	0,14
				K-MEANS (3)	3	12	0,93	4,52	0,21	0,07
				K-MEANS (4)	4	4	0,94	8,05	0,12	0,13
				K-MEANS (5)	5	1	—	11,74	—	0,20
				K-MEDOIDES (0)	0	5	1,05	5,02	0,21	0,08
				K-MEDOIDES (1)	1	3	1,55	8,51	0,18	0,14

La combinación metodológica que ofreció los mejores resultados en Morales fue la aplicación de K-Medoids, tanto en la agrupación por forma como en la segmentación por magnitud, como se presenta en la Tabla 4.7. Esta estrategia permitió una diferenciación más clara y consistente entre perfiles de consumo, siempre que los grupos resultantes contaran con una cantidad suficiente de usuarios. Por ejemplo, en el caso de la agrupación por forma, K-Medoids logró identificar subgrupos bien definidos y con diferencias notorias en sus estadísticas descriptivas, lo que facilitó la distinción entre usuarios de consumo medio y bajo.

En términos de calidad del agrupamiento, K-Medoids obtuvo un índice Davies-Bouldin de 0.75, lo que representa una mejora aproximada del 22 % respecto al valor alcanzado por K-Means (0.97), reflejando clústeres más compactos internamente y mejor separados entre sí. Aunque K-Means también logró generar subgrupos contrastantes en algunos casos, sus valores internos de dispersión fueron mayores, y la separación entre grupos, menos marcada. Por el contrario, en K-Medoids la diferencia de consumo medio y coeficiente de variación entre subgrupos fue suficiente para mantener perfiles diferenciados.

4.3. Análisis de *Clustering* para el Municipio de Santander de Quilichao

La Figura 4.3 muestra la segmentación de usuarios de Santander de Quilichao según la similitud en sus perfiles de consumo eléctrico.

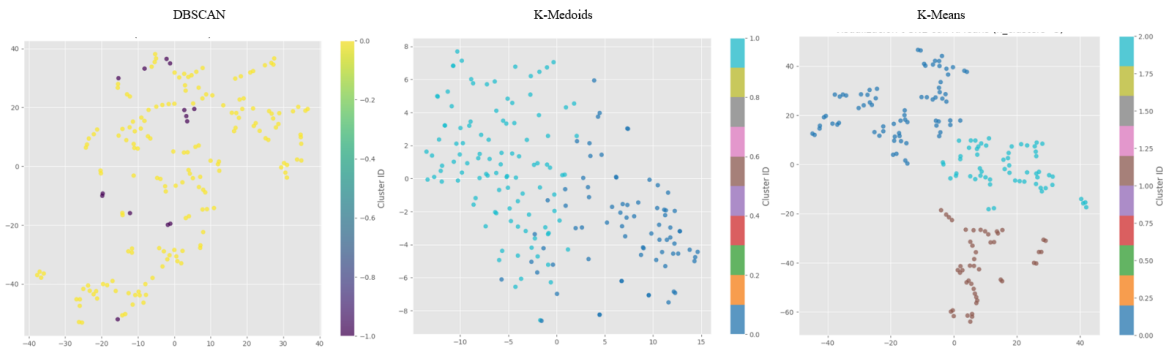


Figura 4.3: Clústeres por Forma, municipio de Santander de Quilichao.

En el municipio de Santander de Quilichao, las métricas presentadas en las Tablas 4.8, 4.9 y 4.10 evidencian la existencia de agrupamientos con patrones de consumo claramente diferenciados, tanto en términos de la forma de las curvas de consumo como en la magnitud del uso energético.

Tabla 4.8: Análisis de métricas por cluster DBSCAN para el municipio de Santander de Quilichao.

Forma	DB Index	N° Clúster (Forma)	Cientes (Forma)	Magnitud	N° Clúster	Cientes	Desv. Est.	Media	CV	RMS
DBSCAN	-	0	167	DBSCAN (0)	0	144	5.94	13.74	0.43	0.20
				DBSCAN (-1)	-1	23	-	-	-	-
				K-MEANS (0)	0	48	5.33	23.77	0.22	0.34
				K-MEANS (1)	1	107	3.12	10.82	0.29	0.15
				K-MEANS (2)	2	5	6.50	56.61	0.11	0.98
				K-MEANS (3)	3	1	-	83.34	-	1.23
				K-MEANS (4)	4	5	6.95	34.55	0.20	0.59
				K-MEANS (5)	5	1	-	88.46	-	1.24
				K-MEDOIDES (0)	0	10	6.71	21.27	0.32	0.40
				K-MEDOIDES (1)	1	44	1.80	7.61	0.24	0.18
				K-MEDOIDES (2)	2	20	5.74	29.81	0.19	0.44
				K-MEDOIDES (3)	3	59	1.40	12.92	0.11	1.01
				K-MEDOIDES (4)	4	7	14.24	61.95	0.23	1.13
				K-MEDOIDES (5)	5	27	3.88	20.90	0.19	0.11

Tabla 4.9: Análisis de métricas por cluster K-MEDOIDES para el municipio de Santander de Quilichao.

Forma	DB Index	N° Clúster (Forma)	Cientes (Forma)	Magnitud	N° Clúster	Cientes	Desv. Est.	Media	CV	RMS
K-MEDOIDES	0.486	0	76	DBSCAN (0)	0	54	7.04	15.78	0.45	0.23
				DBSCAN (-1)	-1	22	-	-	-	-
				K-MEANS (0)	0	69	8.67	18.33	0.47	0.27
				K-MEANS (1)	1	7	14.48	59.52	0.24	0.96
				K-MEDOIDES (0)	0	67	8.53	17.96	0.47	0.26
				K-MEDOIDES (1)	1	9	17.82	53.13	0.33	0.88
K-MEDOIDES	0.486	1	105	DBSCAN (0)	0	96	4.66	12.01	0.39	0.17
				DBSCAN (-1)	-1	9	-	-	-	-
				K-MEANS (0)	0	19	4.24	23.60	0.18	0.34
				K-MEANS (1)	1	51	1.88	13.02	0.14	0.18
				K-MEANS (2)	2	34	1.90	7.48	0.25	0.16
				K-MEANS (3)	3	1	-	83.33	-	1.23
				K-MEDOIDES (0)	0	1	-	83.33	-	1.23
				K-MEDOIDES (1)	1	22	4.41	22.82	0.19	0.33
				K-MEDOIDES (2)	2	32	1.84	7.32	0.25	0.10
				K-MEDOIDES (3)	3	50	1.53	12.61	0.12	0.18

Tabla 4.10: Análisis de métricas por cluster K-MEANS para el municipio de Santander de Quilichao.

Forma	DB Index	N° Clúster (Forma)	Cientes (Forma)	Magnitud	N° Clúster	Cientes	Desv. Est.	Media	CV	RMS
K-MEANS	0.6818	0	74	DBSCAN (0)	0	66	4.38	10.84	0.40	0.15
				DBSCAN (-1)	-1	8	—	—	—	—
				K-MEANS (0)	0	11	5.21	25.91	0.20	0.38
				K-MEANS (1)	1	62	3.29	10.12	0.33	0.14
				K-MEANS (2)	2	1	—	83.34	—	1.23
				K-MEDOIDS (0)	0	32	1.47	12.51	0.12	0.18
				K-MEDOIDS (1)	1	1	—	83.34	—	1.23
				K-MEDOIDS (2)	2	12	5.50	25.23	0.22	0.36
				K-MEDOIDS (3)	3	29	1.96	7.21	0.27	0.10
K-MEANS	0.6818	1	48	DBSCAN (0)	0	30	8.17	17.83	0.46	0.26
				DBSCAN (-1)	-1	18	—	—	—	—
				K-MEANS (0)	0	39	8.49	19.30	0.44	0.29
				K-MEANS (1)	1	8	12.75	48.72	0.26	0.84
				K-MEANS (2)	2	1	—	88.46	—	1.27
				K-MEDOIDS (0)	0	40	9.31	19.94	0.47	0.30
				K-MEDOIDS (1)	1	8	18.78	54.14	0.35	0.91
K-MEANS	0.6818	2	59	DBSCAN (0)	0	49	4.33	13.84	0.31	0.19
				DBSCAN (-1)	-1	10	—	—	—	—
				K-MEANS (0)	0	23	4.26	21.93	0.19	0.32
				K-MEANS (1)	1	36	2.96	11.13	0.27	0.16
				K-MEDOIDS (0)	0	23	2.96	11.13	0.27	0.16
				K-MEDOIDS (1)	1	36	4.26	21.93	0.19	0.32

Durante la etapa de agrupamiento por forma, el algoritmo K-MEDOIDS mostró el mejor desempeño en términos de compacidad, alcanzando un índice Davies-Bouldin de 0.486 (Tabla 4.9), superior al de K-MEANS (0.6818) y al comportamiento general de DBSCAN. Este resultado sugiere que K-MEDOIDS fue más efectivo para conformar *clusters* bien definidos y robustos frente al ruido en esta fase inicial.

Sin embargo, al aplicar el segundo nivel de segmentación orientado a la magnitud del consumo, K-MEANS demostró una mayor capacidad de diferenciación interna entre subgrupos. Por ejemplo, en el *Cluster* por forma 0 generado con K-MEANS (Tabla 4.10), se identificaron subgrupos con perfiles de consumo claramente contrastantes: uno con media de 25.91 kWh, desviación estándar de 5.21 y RMS de 0.38; y otro con media de 10.12, desviación estándar de 3.29 y RMS de 0.14. Estas diferencias también se reflejan en los coeficientes de variación, que fueron de aproximadamente 0.20 y 0.33, respectivamente, indicando mayor estabilidad relativa en el primer subgrupo y mayor dispersión en el segundo.

Por su parte, K-MEDOIDS aplicado a la segmentación por magnitud ofreció una

diferenciación más moderada entre perfiles. En el *Cluster* por forma 1, por ejemplo, un subgrupo presentó una media de 22.82, desviación estándar de 4.41, RMS de 0.33 y CV de 0.19, mientras que otro subgrupo del mismo *cluster* mostró una media de 7.32, desviación estándar de 1.84, RMS de 0.10 y CV cercano a 0.25. Aunque las diferencias son claras, el grado de separación es menos acentuado que el obtenido con K-MEANS, cuya segmentación permite distinguir con mayor claridad entre perfiles de alto y bajo consumo.

En consecuencia, la estrategia más adecuada para este municipio consiste en emplear K-MEDOIDS en la etapa de agrupamiento por forma, aprovechando su capacidad para generar *clusters* compactos y resistentes al ruido, seguido por K-MEANS en la etapa de magnitud, donde resulta ser más efectivo en esta última fase, donde destaca por su habilidad para identificar con mayor precisión subgrupos de consumo diferenciados en intensidad.

4.4. Análisis de *Clustering* para el Municipio de El Tambo

La Figura 4.4 presenta los resultados del proceso de segmentación de perfiles de consumo energético basado en la forma temporal de la demanda. Cada gráfico muestra la proyección bidimensional de los datos transformados mediante técnicas de reducción de dimensionalidad, donde los puntos representan clientes individuales y los colores indican el clúster asignado por cada algoritmo.

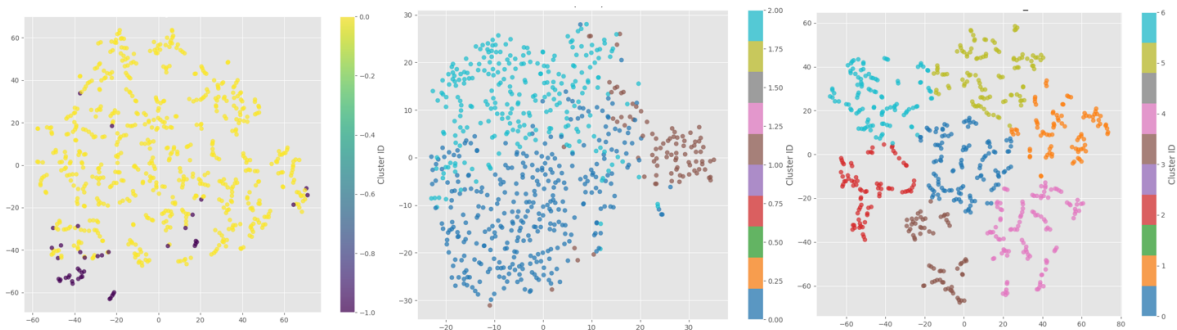


Figura 4.4: Resultados de Clustering por Forma para El Tambo utilizando DBSCAN, K-Medoids y K-Means.

Los indicadores presentados en las Tablas 4.11, 4.12 y 4.13, correspondientes al municipio de El Tambo, evidencian una notable heterogeneidad en los patrones de consumo energético entre los distintos grupos de usuarios identificados mediante algoritmos de *clustering*. Se observan variaciones importantes tanto en la cantidad de clientes por *cluster* como en métricas como el consumo medio, la desviación estándar y el valor RMS.

Tabla 4.11: Análisis de métricas por cluster DBSCAN para el municipio de El Tambo.

Forma	DB Index	N° Clúster (Forma)	Clientes (Forma)	Magnitud	N° Clúster	Clientes	Desv. Est.	Media	CV	RMS
DBSCAN	-	0	632	DBSCAN (0)	0	625	2.58	6.01	0.43	0.09
				DBSCAN (-1)	-1	7	-	-	-	-
				K-MEANS (0)	0	436	1.34	4.65	0.29	0.07
				K-MEANS (1)	1	196	2.29	9.36	0.24	0.14
				K-MEDOIDS (0)	0	346	1.11	4.20	0.26	0.06
				K-MEDOIDS (1)	1	286	2.37	8.41	0.28	0.13

Tabla 4.12: Análisis de métricas por cluster K-MEDOIDS para el municipio de El Tambo

Forma	DB Index	N° Clúster (Forma)	Clientes (Forma)	Magnitud	N° Clúster	Clientes	Desv. Est.	Media	CV	RMS
K-MEDOIDS	0.98	0	332	DBSCAN (0)	0	328	2.75	6.35	0.43	0.09
				DBSCAN (-1)	-1	4	-	-	-	-
				K-MEANS (0)	0	247	1.53	5.11	0.30	0.08
				K-MEANS (1)	1	85	2.67	10.37	0.26	0.15
				K-MEDOIDS (0)	0	176	1.14	4.38	0.26	0.06
				K-MEDOIDS (1)	1	156	2.65	8.79	0.30	0.13
K-MEDOIDS	0.98	1	90	DBSCAN (0)	0	88	2.12	5.66	0.38	0.08
				DBSCAN (-1)	-1	2	-	-	-	-
				K-MEANS (0)	0	55	1.10	4.31	0.26	0.06
				K-MEANS (1)	1	34	1.37	7.97	0.17	0.12
				K-MEANS (2)	2	1	-	29.95	-	0.42
				K-MEDOIDS (0)	0	2	1.56	8.90	0.17	0.15
				K-MEDOIDS (1)	1	38	0.94	6.14	0.15	0.09
				K-MEDOIDS (2)	2	34	0.69	3.59	0.19	0.05
				K-MEDOIDS (3)	3	15	1.14	8.97	0.12	0.13
				K-MEDOIDS (4)	4	1	-	29.95	-	0.42
K-MEDOIDS	0.98	2	253	DBSCAN (0)	0	249	2.58	5.77	0.45	5.35
				DBSCAN (-1)	-1	4	-	-	-	-
				K-MEANS (0)	0	100	0.85	3.54	0.24	0.05
				K-MEANS (1)	1	43	1.58	8.53	0.19	0.12
				K-MEANS (2)	2	82	0.72	5.73	0.13	0.09
				K-MEANS (3)	3	3	0.68	10.85	0.06	0.20
				K-MEANS (4)	4	18	0.72	9.11	0.08	0.15
				K-MEANS (5)	5	7	1.59	13.63	0.12	0.21
				K-MEDOIDS (0)	0	113	2.26	8.13	0.28	0.12
				K-MEDOIDS (1)	1	140	1.09	4.03	0.27	0.06

Tabla 4.13: Análisis de métricas por cluster K-MEANS para el municipio de El Tambo

Forma	DB Index	N° Clúster (Forma)	Cientes (Forma)	Magnitud	N° Clúster	Cientes	Desv. Est.	Media	CV	RMS
K-MEANS	0.7421	0	117	DBSCAN (0)	0	116	2.22	5.85	0.38	0.09
				DBSCAN (-1)	-1	1	—	—	—	—
				K-MEANS (0)	0	80	1.30	4.65	0.28	0.07
				K-MEANS (1)	1	37	1.88	8.72	0.22	0.13
				K-MEDOIDES (0)	0	65	1.01	4.27	0.24	0.07
				K-MEDOIDES (1)	1	52	1.94	8.03	0.24	0.12
K-MEANS	0.7421	1	89	DBSCAN (0)	0	86	2.68	6.64	0.40	0.10
				DBSCAN (-1)	-1	3	—	—	—	—
				K-MEANS (0)	0	36	1.19	4.21	0.28	0.06
				K-MEANS (1)	1	5	4.44	16.38	0.27	0.24
				K-MEANS (2)	2	48	1.46	8.10	0.18	0.12
				K-MEDOIDES (0)	0	23	1.10	9.54	0.11	0.14
				K-MEDOIDES (1)	1	34	0.72	6.65	0.11	0.10
				K-MEDOIDES (2)	2	20	0.64	4.33	0.15	0.06
				K-MEDOIDES (3)	3	4	4.17	17.54	0.24	0.26
				K-MEDOIDES (4)	4	8	0.38	2.48	0.15	0.035
K-MEANS	0.7421	2	185	DBSCAN (0)	0	183	2.78	5.97	0.47	0.09
				DBSCAN (-1)	-1	2	—	—	—	—
				K-MEANS (0)	0	58	2.05	9.48	0.22	0.15
				K-MEANS (1)	1	127	1.34	4.48	0.30	0.07
				K-MEDOIDES (0)	0	82	2.34	8.57	0.27	0.13
				K-MEDOIDES (1)	1	103	1.07	4.32	0.25	0.06
K-MEANS	0.7421	3	174	DBSCAN (0)	0	170	2.27	5.55	0.41	0.08
				DBSCAN (-1)	-1	4	—	—	—	—
				K-MEANS (0)	0	68	0.72	5.39	0.13	0.08
				K-MEANS (1)	1	9	1.36	10.43	0.13	0.16
				K-MEANS (2)	2	55	0.68	3.26	0.21	0.05
				K-MEANS (3)	3	41	1.36	8.19	0.17	0.12
				K-MEANS (4)	4	1	—	29.95	—	0.42
				K-MEDOIDES (0)	0	73	3.14	8.16	0.38	0.12
				K-MEDOIDES (1)	1	101	1.06	4.05	0.26	0.06
K-MEANS	0.7421	4	110	DBSCAN (0)	0	104	2.20	6.12	0.36	0.09
				DBSCAN (1)	1	4	1.45	14.15	0.10	0.20
				DBSCAN (-1)	-1	2	—	—	—	—
				K-MEANS (0)	0	77	1.32	5.08	0.26	0.07
				K-MEANS (1)	1	33	2.64	10.06	0.26	0.15
				K-MEDOIDES (0)	0	68	1.15	4.81	0.24	0.07
				K-MEDOIDES (1)	1	42	2.64	9.43	0.28	0.14

El análisis de los clusters revela patrones claramente diferenciados. En el algoritmo K-MEANS (Cluster 1) por forma, el Subcluster 2 del método K-Means presenta un RMS de 0.12, media de 8.10 y CV de 0.18, mostrando un consumo moderado con estabilidad aceptable. En contraste, el Subcluster 0 del mismo algoritmo registra un RMS más bajo (0.06), media de 4.21 y CV de 0.28, indicando un perfil de bajo consumo con mayor variabilidad relativa.

En la tabla 4.13 en K-Means cluster 1 por forma se que obtiene el Subcluster 1 de K-Medoids presentó el perfil más estable, con un CV de 0.11, una media de 6.65 kWh y un RMS de 0.10. Por su parte, el Subcluster K-Medoids (4) destacó por su

eficiencia, registrando la media más baja (2.48 kWh) y el menor RMS observado (0.035). En contraste, uno de los casos más problemáticos fue cluster 1 K-Means por forma y subcluster K-Means (1), con una media de 16.38 kWh y un RMS de 0.24, una desviación estándar de 4.44 que indica gran variabilidad, reflejando un patrón de consumo considerablemente más elevado y menos uniforme.

De la tabla 4.12 se evidencia que los grupos obtenidos mediante K-Medoids por forma presentan menor variación interna que los de K-Means. En el clúster 2 de K-Medoids por forma, el subclúster 0 del metodo K-Medoids registró un RMS de 0.12 y una media de 8.13, mientras que el subclúster 1 alcanzó un RMS de 0.06 y una media de 4.03.

K-Means se confirma con su mejor índice Davies-Bouldin (0.7421 vs 0.98 de K-MEDOIDS), demostrando mayor capacidad para capturar los extremos del comportamiento energético. Los datos sugieren que una estrategia combinada sería óptima: usar K-MEANS para segmentación granular y DBSCAN para detectar anomalías (como los casos con media=29.95 o usuarios marcados como ruido). Esta aproximación equilibra diferenciación interna (K-MEANS) y detección de atípicos (DBSCAN).

Con DBSCAN se observó una tendencia a formar un único gran cluster y a dejar una proporción relevante de usuarios como ruido: 177 agrupados y 4 (2.2 %) como ruido en Santander de Quilichao; 79 agrupados y 15 (15.9 %) en Morales; y 2471 agrupados y 138 (5.3 %) en Popayán. Esta limitación, sumada a que al no ser un método particional no permite calcular métricas como el índice Davies-Bouldin, muestra su baja capacidad para diferenciar patrones temporales complejos.

4.5. Reducción de Dimensionalidad con PCA

Esta sección presenta los resultados del proceso de reducción de dimensionalidad aplicado a los datos de magnitud mediante Análisis de Componentes Principales (PCA). La cantidad de componentes seleccionadas para representar adecuadamente la variabilidad de los datos en cada municipio, según el método de *clustering* por forma

previamente utilizado, se resume en la Tabla 4.14.

Tabla 4.14: Resumen de componentes principales (PCA) por municipio y método de clustering.

Municipio	Método	N° Clúster	Agrupación de Clusters	PCA
Popayán	DBSCAN	1	Cluster 0	5
	K-MEANS	0	Cluster 0	5
		1	Cluster 1	2
		2	Cluster 2	4
		3	Cluster 3	3
		4	Cluster 4	3
		5	Cluster 5	4
		6	Cluster 6	2
	K-MEDOIDS (DTW)	0	Cluster 0	3
		1	Cluster 1	5
		2	Cluster 2	5
Morales	DBSCAN	1	Cluster 0	4
	K-MEANS	0	Cluster 0	5
		1	Cluster 1	3
		2	Cluster 2	4
	K-MEDOIDS (DTW)	0	Cluster 0	4
		1	Cluster 1	4
El Tambo		2	Cluster 2	3
	DBSCAN	1	Cluster 0	5
	K-MEANS	0	Cluster 0	4
		1	Cluster 1	4
		2	Cluster 2	3
		3	Cluster 3	5
		5	Cluster 5	4
	K-MEDOIDS (DTW)	0	Cluster 0	5
		1	Cluster 1	6
		2	Cluster 2	4
Santander de Quilichao	DBSCAN	1	Cluster 0	3
	K-MEANS	0	Cluster 0	2
		1	Cluster 1	3
		2	Cluster 2	4
	K-MEDOIDS (DTW)	0	Cluster 0	3
		1	Cluster 1	2

El análisis mediante PCA permitió reunir evidencia importante sobre la estructura interna de los datos de consumo eléctrico, revelando diferencias importantes tanto entre municipios como entre métodos de agrupamiento. En primer lugar, la cantidad de componentes principales necesarias para explicar la varianza de los datos varió considerablemente entre los municipios de Popayán, Morales, El Tambo y Santander de Quilichao, lo que sugiere la existencia de distintos niveles de complejidad en los patrones de consumo.

Asimismo, se observó que la cantidad de componentes requeridos depende no solo del municipio, sino también del algoritmo de *clustering* aplicado en la etapa de agrupación por forma (DBSCAN, K-Means o K-Medoids). Esta variabilidad sugiere que la estructura interna de cada *cluster* influye directamente en la dimensionalidad necesaria para describir adecuadamente sus características. Por ejemplo, municipios como Popayán, caracterizados por una mayor heterogeneidad en los patrones de consumo, requirieron un mayor número de componentes principales para conservar la información relevante. En contraste, regiones como Santander de Quilichao presentaron una menor necesidad de componentes, lo que sugiere patrones de consumo más homogéneos.

Finalmente, los resultados permiten inferir que tanto el tamaño como la diversidad interna de los *clusters* influyen en el número de componentes requeridos: agrupamientos más grandes o con mayor diversidad estructural tienden a necesitar una representación de mayor dimensión para preservar la varianza explicada y capturar adecuadamente sus características principales. Estos hallazgos refuerzan la pertinencia de utilizar PCA como paso previo a la segmentación por magnitud, ya que permite optimizar la representación de los datos manteniendo su capacidad descriptiva.

4.6. Resumen

Este capítulo presenta la evaluación de técnicas de *clustering* aplicadas a series temporales de consumo energético en los municipios de Popayán, Morales, Santander de Quilichao y El Tambo. El análisis se realizó en dos etapas: primero, agrupamiento

por la forma de las curvas de demanda utilizando DBSCAN, K-Means y K-Medoids con distancia DTW; segundo, segmentación por magnitud tras aplicar reducción de dimensionalidad con PCA. Los resultados muestran que K-Means fue, en general, el algoritmo más efectivo en ambas etapas, destacando por su capacidad de compactación y diferenciación de grupos, mientras que K-Medoids sobresalió en la agrupación por forma en Santander de Quilichao. DBSCAN, aunque útil para detectar perfiles atípicos, fue limitado como técnica principal de segmentación.

El análisis estadístico empleó métricas como la media, desviación estándar, RMS y coeficiente de variación (CV) para caracterizar los subgrupos. Se evidenciaron patrones de consumo claramente diferenciados, particularmente en Popayán y El Tambo, donde se identificaron tanto usuarios con consumos elevados como grupos más estables y homogéneos. Además, los resultados de PCA revelaron que la cantidad de componentes principales necesarios varió entre municipios y algoritmos, reflejando distintos niveles de complejidad en los datos. Esto confirma la pertinencia de aplicar PCA antes de la segmentación por magnitud, ya que permite mantener la estructura representativa de los datos y optimizar la calidad de los agrupamientos.

A diferencia de estudios previos centrados en clustering unidimensional como el trabajo “A comparative study of clustering techniques for electrical load pattern segmentation” [3], este trabajo evidenció que la integración de técnicas temporales (DTW) y de magnitud (PCA) mejora la identificación de patrones de consumo en escalas geográficas locales. Sin embargo, la ausencia de comparación con algoritmos como GMM sugiere una línea futura de investigación.

Finalmente, para facilitar la comprensión de la metodología aplicada, se incluye un diagrama metodológico Fiugra 4.5 que resume de manera secuencial el proceso desarrollado, desde la recolección y depuración de los datos, pasando por las etapas de análisis exploratorio, reducción de dimensionalidad y aplicación de técnicas de clustering, hasta llegar a la interpretación y validación de los resultados.

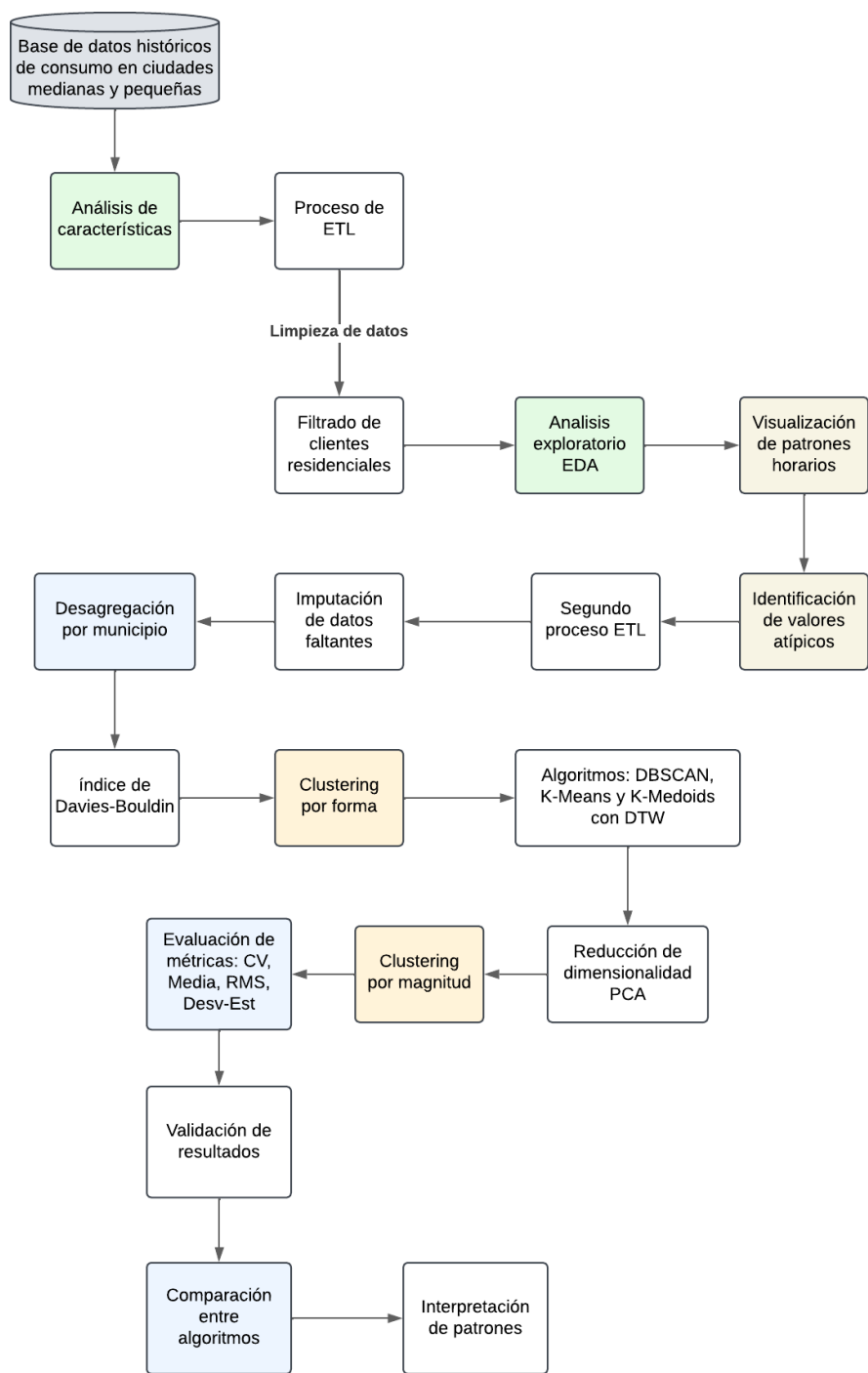


Figura 4.5: Metodología: desde la recolección de datos hasta la validación de resultados

Capítulo 5

Conclusiones y Trabajos Futuros

5.1. Conclusiones Generales

- El tratamiento de datos faltantes mediante técnicas de imputación resultó fundamental para asegurar la integridad y coherencia del conjunto de datos utilizados en el análisis. Este proceso permitió preservar la estructura temporal de las series sin introducir distorsiones significativas en la magnitud del consumo, garantizando así que los resultados posteriores de reducción de dimensionalidad y *clustering* fueran representativos y consistentes entre los distintos grupos analizados.
- El análisis basado en la forma de las curvas de consumo demostró ser eficaz para identificar patrones temporales compartidos entre usuarios, independientemente del nivel de demanda. Entre los algoritmos evaluados, K-Means se destacó por su capacidad para distinguir patrones contrastantes, mientras que DBSCAN y K-Medoids ofrecieron ventajas particulares en escenarios con diferentes niveles de densidad y ruido.
- La reducción de dimensionalidad mediante Análisis de Componentes Principales (PCA) permitió condensar la información asociada a la magnitud del consumo. Esta etapa facilitó una segunda fase de segmentación más manejable tanto en términos computacionales como interpretativos, sin perder las características esenciales del comportamiento energético de los usuarios.
- El análisis comparativo entre algoritmos evidenció que K-Means facilitó una mejor diferenciación entre perfiles de consumo bajo, medio y alto, mientras que K-Medoids generó agrupamientos más homogéneos en términos de variabilidad interna. Por su parte, DBSCAN resultó útil para la detección de

comportamientos atípicos, aunque presentó limitaciones al aplicarse como método principal en conjuntos con alta dispersión o baja densidad relativa.

- Los resultados obtenidos permiten concluir que la segmentación conjunta por forma y magnitud constituye una herramienta sólida para la caracterización del comportamiento de los usuarios del servicio eléctrico. Esta aproximación metodológica puede ser aprovechada como base para diseñar estrategias de gestión operativa, planificación de infraestructura y políticas tarifarias o de eficiencia energética más alineadas con los perfiles reales de consumo. Por ejemplo, en el ámbito de detección de anomalías, los operadores pueden identificar usuarios cuyo consumo difiere significativamente del patrón típico de su cluster, lo que facilita la detección de posibles casos de uso no autorizado del servicio o fallas técnicas en la red. En cuanto a la eficiencia energética, los resultados permiten diseñar programas focalizados, como campañas de reducción del consumo en horas pico para clusters con alta demanda nocturna, o la implementación de incentivos tarifarios dirigidos a grupos de usuarios con potencial de ahorro energético.

5.2. Trabajos futuros

A partir de los resultados obtenidos en esta investigación, se identifican diversas oportunidades para profundizar y ampliar el alcance del presente estudio. En primer lugar, se sugiere incorporar variables contextuales adicionales, tales como factores climáticos, ciclos estacionales o características sociodemográficas y de tipología del usuario (residencial, comercial, industrial), con el fin de enriquecer la caracterización de los patrones de consumo y facilitar un análisis multivariado más integral.

Asimismo, se plantea explorar el comportamiento dinámico de los grupos de consumo en el tiempo, mediante técnicas de análisis temporal multiclase o aprendizaje no supervisado en entornos evolutivos. La implementación de algoritmos de seguimiento de *clusters* (*cluster tracking*) o esquemas con ventanas móviles (*sliding windows*) permitiría identificar transiciones o cambios importantes en los hábitos de consumo de los usuarios a lo largo del tiempo.

Por último, se sugiere avanzar hacia la integración de esta segmentación en modelos predictivos de demanda o en sistemas de alerta temprana para la detección de anomalías. La vinculación con sistemas de soporte a la toma de decisiones (DSS) para empresas de servicios públicos abriría la puerta a una gestión más proactiva, eficiente y personalizada, basada en el conocimiento profundo de los perfiles específicos de los usuarios.

Referencias

- [1] F. M. Q. Mateo and P. L. C. Romero, “Modelo basado en minería de datos para la detección de pérdidas no técnicas de redes de distribución,” *Universidad de Sevilla*, 2017.
- [2] J. S. Guerrero-Prado, W. Alfonso-Morales, E. Caicedo-Bravo, B. Zayas-Pérez, and A. Espinosa-Reza, “The power of big data and data analytics for ami data: A case study,” *Sensors*, vol. 20, no. 11, p. 3289, 2020.
- [3] 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.
- [4] H. Kim, S. Park, and S. Kim, “Time-series clustering and forecasting household electricity demand using smart meter data,” *Energy Reports*, vol. 9, pp. 4111–4121, 2023.
- [5] C. Li, F. Wu, Z. Chen, and D. He, “Application of artificial intelligence technology in digital transformation of power grid,” in *2024 3rd International Conference on Energy and Power Engineering, Control Engineering (EPECE)*, pp. 9–13, IEEE, 2024.
- [6] R. Sriranjani, D. J. S. Raja, N. Hemavathi, A. Parvathy, N. Vinotha, and R. Sharmitha, “Implementation of energy efficient advanced metering infrastructure in smart grid,” in *2023 1st International Conference on Advanced Engineering and Technologies (ICONNIC)*, pp. 316–321, IEEE, 2023.
- [7] M. Kojury-Naftchali, A. Fereidunian, and H. Lesani, “Identifying susceptible consumers for demand response and energy efficiency policies by time-series analysis and supplementary approaches,” in *2016 24th Iranian Conference on Electrical Engineering (ICEE)*, pp. 1130–1135, IEEE, 2016.
- [8] M. H. Rashid, “Ami smart meter big data analytics for time series of electricity consumption,” in *2018 17th IEEE International Conference On Trust, Security And*

Privacy In Computing And Communications/12th IEEE International Conference On Big Data Science And Engineering (TrustCom/BigDataSE), pp. 1771–1776, IEEE, 2018.

- [9] E. Cook, M. B. Saleem, Y. Weng, S. Abate, K. Kelly-Pitou, and B. Grainger, “Density-based clustering algorithm for associating transformers with smart meters via gps-ami data,” *International Journal of Electrical Power & Energy Systems*, vol. 142, p. 108291, 2022.
- [10] A. Ball, D. Rye, F. Ramos, and M. Velonaki, “A comparison of unsupervised learning algorithms for gesture clustering,” in *Proceedings of the 6th international conference on Human-robot interaction*, pp. 111–112, 2011.
- [11] T. Wahyuningrum, S. Khomsah, S. Suyanto, S. Meliana, P. E. Yunanto, and W. F. Al Maki, “improving clustering method performance using k-means, mini batch k-means, birch and spectral,” in *2021 4th International Seminar on Research of Information Technology and Intelligent Systems (ISRITI)*, pp. 206–210, IEEE, 2021.
- [12] D. Deng, “DbSCAN clustering algorithm based on density,” in *2020 7th international forum on electrical engineering and automation (IFEAA)*, pp. 949–953, IEEE, 2020.
- [13] M. K. Dhar, S. N. Hasan, T. R. Otushi, and M. Khan, “Entropy-based feature selection for data clustering using k-means and k-medoids algorithms,” in *2020 Fifth International Conference on Research in Computational Intelligence and Communication Networks (ICRCICN)*, pp. 36–40, IEEE, 2020.
- [14] X. Yang, Z. Zhang, X. Cui, and R. Cui, “A time series data augmentation method based on dynamic time warping,” in *2021 International Conference on Computer Communication and Artificial Intelligence (CCAI)*, pp. 116–120, IEEE, 2021.
- [15] H. Li, “Multivariate time series clustering based on common principal component analysis,” *Neurocomputing*, vol. 349, pp. 239–247, 2019.
- [16] I. T. Jolliffe and J. Cadima, “Principal component analysis: a review and recent developments,” *Philosophical transactions of the royal society A: Mathematical, Physical and Engineering Sciences*, vol. 374, no. 2065, p. 20150202, 2016.

- [17] L. Ni and S. Jinhang, “The analysis and research of clustering algorithm based on pca,” in *2017 13th IEEE International Conference on Electronic Measurement & Instruments (ICEMI)*, pp. 361–365, IEEE, 2017.
- [18] J. W. Ortiz-Villegas, “Imputación de datos faltantes usando técnicas estadísticas y de machine learning para los sistemas de infraestructura de medición avanzada (ami),” 2024.
- [19] C. M. Bishop and N. M. Nasrabadi, *Pattern recognition and machine learning*, vol. 4. Springer, 2006.
- [20] D. L. Davies and D. W. Bouldin, “A cluster separation measure,” *IEEE transactions on pattern analysis and machine intelligence*, no. 2, pp. 224–227, 2009.
- [21] V. Barnett, T. Lewis, *et al.*, *Outliers in statistical data*, vol. 3. Wiley New York, 1994.
- [22] D. Chicco, M. J. Warrens, and G. Jurman, “The coefficient of determination r-squared is more informative than smape, mae, mape, mse and rmse in regression analysis evaluation,” *Peerj computer science*, vol. 7, p. e623, 2021.