ESCUELA SUPERIOR POLITÉCNICA DEL LITORAL

Facultad de Ingeniería en Electricidad y Computación

Proyecto de grado

Previo a la obtención de título de:

INGENIERO EN ELECTRICIDAD

COMPROBACIÓN DE TÉCNICAS DE MINERÍA DE DATOS PARA DETECCIÓN DE PÉRDIDAS NO TÉCNICAS UTILIZANDO MEDIDORES INTELIGENTES

AUTOR: Steven Agustín Quinde Jiménez

TUTOR: Johnny Wladimir Rengifo Santana

Máster en Ingeniería Eléctrica

Guayaquil - Ecuador
2020

Resumen

Las pérdidas no técnicas es uno de los principales desafíos que afrontan las empresas eléctricas. Sin embargo, la detección de este fenómeno es sumamente complicado debido a la variedad de factores que las producen. En este trabajo, se implementó métodos de minería de datos no supervisado para detectar pérdidas no técnicas presentes en redes de distribución. La metodología consistió, primero, en generar artificialmente un conjunto de datos de registros de medidores inteligentes mediante modelos ocultos de Markov (HMM). Luego, se aplicó un modelo que altera los perfiles de carga diario de los usuarios para simular distintos tipos de pérdidas. Después, se procedió a la fase de entrenamiento donde se comparó tres algoritmos de agrupamiento: K-medias, K-medianas y Jerárquico Aglomerativo Ward para seleccionar el que mayor cantidad de pérdidas detecte, a través de los indicadores de validación de grupos MIA (Indice de Adecuación de la Media) y Silueta. Además, se definió una serie de criterios para caracterizar a los grupos que diferenciaban a usuarios benignos de los usuarios con pérdidas. Los resultados obtenidos en la fase de prueba evidencian que es posible detectar a usuarios que presentan irregularidades en su patrón de consumo, reconociendo aproximadamente al 68 % de las pérdidas aplicadas. Por tanto, es de suma importancia contar con métodos computacionales inteligentes y eficaces para resolver esta problemática.

Palabras claves: Minería de Datos, HMM, Perfiles de Carga, Algoritmos de Agrupamiento

Agradecimientos

Agradezco a Dios por brindarme paciencia, voluntad y salud para culminar mis estudios universitarios.

Agradezco a mis padres por todo el esfuerzo realizado y el apoyo dado durante tantos años para lograr esta meta.

Agradezco a mi abuelita Martha Mendoza, mi segunda madre, por los valores inculcados en mi niñez y por enseñarme que con esfuerzo los sueños se cumplen.

Agradezco a mi novia, a profesores y amigos que de alguna u otra manera aportaron con conocimientos, enseñanzas o consejos para la obtención de este título.

Finalmente, un especial agradecimiento a los ingenieros Fernando Vaca y Johnny Rengifo por el seguimiento y por la ayuda brindada para el desarrollo de este proyecto.

A todos muchas gracias.

Dedicatoria

A mis hermanos, que los amo con todo mi corazón.

"Aunque nada cambie, si yo cambio, todo cambia"

- Marcel Proust

Declaración expresa

"La responsabilidad y la autoría del contenido de este trabajo de titulación, me corresponde exclusivamente; y doy mi consentimiento para que la ESPOL realice la comunicación pública de la obra por cualquier medio con el fin de promover la consulta, difusión y uso público de la producción intelectual"

Steven Agustín Quinde Jiménez

Evaluadores

M.Sc. Fernando Vaca	**M.Sc. Johnny Rengifo**
PROFESOR DE LA MATERIA	PROFESOR TUTOR

Índice general

Índice de figuras

Índice de tablas

Acrónimos

ALC	América Latina y el Caribe
AMI	Advanced Metering Infraestructure
AMR	Automatic Meter Reading
ARCONEL	Agencia de Regulación de Control de Electricidad
BID	Banco Interamericano de Desarrollo
BPL	Broadband over Power Lines
CNEL	Corporación Nacional de Electricidad
GHMM	Gaussian Hidden Markov Models
GIS	Sistema de Información Geográfica
GPRS	General Packet Radio Service
HAN	Home Area Network
HMM	Hidden Markov Models
MDMS	Measurement Data Management System
MEER	Ministro de Electricidad y Energía Renovables
MIA	Mean Index Adecuacy
NAN	Neighborhood Area Network
OLADE	Organización Latinoamericana de Energía
PLC	Power Line Communication
PME	Plan Maestro de Electricidad
WAN	Wide Area Network
WIFI	Wireless Fidelity

Capítulo 1

PLANTEAMIENTO Y ESTRUCTURA DE LA TESIS

1.1. Planteamiento del proyecto

Entre los principales problemas que afrontan las empresas eléctricas están las pérdidas no-técnicas o comerciales presentes en las redes de distribución, vinculadas especialmente con el consumo de electricidad del usuario final que no es parcial o totalmente facturada debido a la ejecución de prácticas fraudulentas como: alteración del sistema de medida, conexiones directas, o por la manipulación de las instalaciones. Este problema produce pérdidas económicas por los consumos no facturados que afectan en el desarrollo de la empresa eléctrica. Así mismo origina un rápido deterioro de las redes e instalaciones, comprometiendo la fiabilidad del sistema de distribución.

En las subregiones de América Latina y el Caribe (ALC), las pérdidas de electricidad son un problema importante. Según indica el Banco Interamericano de Desarrollo (BID) [1], la tasa promedio de pérdidas por subregión variaron entre el 17 % y 19 % en el año 2012. Las pérdidas se concentran mayormente en el sistema de distribución, y su causa principal es debido a los factores no técnicos mencionados.

El Plan Maestro de Electricidad 2016-2025 (PME) emitido por el Ministerio de Electricidad y Energía Renovables (MEER) de Ecuador, señala que las pérdidas de energía eléctrica han tenido una atención prioritaria. Dicho plan contempla índices de pérdidas a nivel nacional del 12,2 % equivalente a 2690,94 GWh, en el año 2016. Además, establece como meta para el año 2025, una reducción de pérdidas en el país del 8,79 % mediante lineamientos y estrategias. Una de las directrices a seguir en el PME, es la implementación de sistemas de medición inteligente a nivel de red de distribución, centros de transformación y usuario para alcanzar niveles óptimos de pérdidas de energía en el sistema de distribución.

De manera concreta, se estima que Guayaquil en el año 2018 tuvo 609 MWh/año (11,10 %) de energía perdida, correspondiendo a un aproximado de 60 millones de dólares, de acuerdo con los datos de la Agencia de Regulación y Control de Electricidad (ARCONEL). Para este mismo año en Guayaquil, el total de pérdidas comerciales fueron alrededor de 184 MWh/año (3,34 %), valor que representa cerca de 18 millones de dólares. Para hacer frente a esta problemática, la Corporación Nacional de Electricidad (CNEL) ha aplicado según lo establecido en el PME, el cambio y modernización de los

equipos de medición presentando grandes mejoras en cuanto a la reducción de pérdidas comerciales. En la actualidad, Guayaquil cuenta con aproximadamente 700 mil clientes registrados, de los cuales 100 mil disponen de medidores inteligentes [2].

La implementación de una infraestructura de medición avanzada conlleva a la generación de grandes cantidades de datos medidos y recolectados de los consumidores, que con un adecuado procesamiento y análisis pueden convertirse en información útil para la toma de decisiones. El manejo de grandes datos en el sistema energético requiere la integración de técnicas y herramientas específicas para la adquisición y almacenamiento de datos, análisis de correlación de datos, control de datos de múltiple fuentes y visualización de datos. Entre ellos la adquisición y almacenamiento de datos son los elementos principales.

Las técnicas para el procesamiento y análisis de los grandes volúmenes de información involucran el desarrollo de algoritmos avanzados basados en modelos estadísticos, inteligencia artificial y aprendizaje automático. La minería de datos engloba estas tres disciplinas y su objetivo es explorar las bases de datos en búsqueda de patrones repetitivos, tendencias o reglas que permitan extraer información valiosa del conjunto de datos para predecir resultados. Con el uso de la información histórica de consumo de carga se puede determinar patrones de consumo que reflejan el comportamiento de los usuarios, los patrones se pueden aglomerar para crear perfiles de consumo que nos permitiría identificar anomalías en el consumo de manera automática.

1.2. Objetivos del proyecto

1.2.1. Objetivo General

Evaluar técnicas de minería de datos aplicado a registros de medidores inteligentes para detectar fuentes de pérdidas no técnicas en redes de distribución.

1.2.2. Objetivos específicos

- Identificar los diversos algoritmos de minería de datos para procesar la información recolectada de los medidores inteligentes y obtener los patrones de consumos de los usuarios.
- Examinar técnicas o métodos para la detección de usuarios fraudulentos a partir de sus patrones de consumo.
- Analizar y comparar el nivel de pérdidas detectadas en el conjunto de datos.

1.3. Justificación del proyecto

Uno de los objetivos que toda empresa distribuidora de energía eléctrica se plantea año tras año es la reducción de las pérdidas no técnicas, de modo que, desarrollan estrategias, métodos o técnicas viables para mejorar el control y reducción de pérdidas. Sin embargo, las contravenciones para evadir el registro y la facturación del consumo eléctrico son cada vez más refinadas y técnicamente realizadas, de distinta naturaleza y de mucha creatividad; por ello, las medidas que tienen que tomar las empresas eléctricas para enfrentar esta problemática deben ser igual de creativas y eficaces. Por esto, un

estudio que permita la identificación de usuarios que consumen la energía, la cual no les sea parcial o totalmente facturada, representa una mejora en los ingresos económicos de las empresas distribuidoras, garantiza un servicio mas confiable, seguro y continuo al usuario. Finalmente, se obtienen beneficios técnicos-económicos tanto para la empresa como para los consumidores.

1.4. Estructura del proyecto

El contenido del trabajo está conformado en los siguiente capítulos:

- **Capítulo 2: Pérdidas de energía eléctrica e Infraestructura de Medición Avanzada**
  En este capítulo se realiza una introducción acerca de los fundamentos de las pérdidas eléctricas y su clasificación. Se presenta con mayor énfasis las de tipo no técnicas, sus posibles causas y consecuencias. Además, se presenta una breve descripción respecto a la estructura, tecnologías y beneficios que conlleva la implementación de una infraestructura de medición avanzada.

- **Capítulo 3: Minería de datos**
  Este capítulo aborda los conocimientos generales del proceso de minería de datos. Adicionalmente, se describe los métodos, técnicas o estrategias para el análisis de grupos.

- **Capítulo 4: Generación sintética de patrones de consumo de energía y modelo de pérdidas no técnicas**
  En este apartado se expone los conceptos relacionados a los modelos de Markov para su posterior desarrollo en la generación sintética de patrones de consumo eléctrico. De la misma manera, se detalla el modelo empleado para la simulación de usuarios con pérdidas no técnicas.

- **Capítulo 5: Detección de pérdidas no técnicas utilizando métodos de agrupamiento**
  En este capítulo se realiza una selección de algoritmos y criterios basados en el análisis de agrupamiento que permitan detectar o identificar grupos de usuarios con pérdidas no técnicas.

- **Capítulo 6: Conclusiones y líneas futuras**
  Por último, este apartado presenta las resoluciones de los capítulos anteriores y se sugiere líneas futuras de implementación.

Capítulo 2

PÉRDIDAS DE ENERGÍA ELÉCTRICA E INFRAESTRUCTURA DE MEDICIÓN AVANZADA

Las pérdidas eléctricas que se producen en el transporte de la energía a lo largo del sistema de potencia representan la diferencia entre la energía suministrada a la red y la energía entregada al consumidor final. El concepto de pérdidas incluye también la energía eléctrica consumida por el usuario final pero no facturada por la empresa eléctrica, que se deriva a pérdidas financieras de la misma [3].

Por otro lado, la continua expansión de la demanda en el sector eléctrico requiere un cambio en la forma que se maneja la energía a una más eficiente y óptima. La evolución del sistema de distribución hacia redes inteligentes proporciona mayor inteligencia al sistema mediante una infraestructura de medición avanzada, incorporando las tecnologías de la información y la comunicación a fin de mejorar la gestión de la energía eléctrica. [4]. La implementación de esta nueva infraestructura es un paso vital para la modernización del sistema. Desde el punto de vista tecnológico, esta infraestructura facilita las funciones de comunicación y control necesarias para activar servicios críticos de gestión de energía, como esquemas de precios detallados, lectura automática de medidores, respuesta a la demanda y gestión de la calidad de la energía [5].

Actualmente, las empresas distribuidoras de energía eléctrica se inclinan hacia la implementación de un sistema de medición avanzada, debido a que brinda muchos beneficios como la reducción del impacto técnico y económico que ocasionan las altas pérdidas de energía en las empresas eléctricas y por ende, la mejora en los ingresos por la comercialización de la misma.

En este capítulo se presenta una definición precisa de las pérdidas eléctricas, en particular, las de tipo no-técnicas. Para entrar más en contexto se presenta el origen de estas pérdidas y su clasificación. Así mismo, se detalla sus consecuencias tanto para el bienestar de la empresa, como para la sociedad y para el sistema eléctrico como tal. Además, se presenta los métodos efectivos que se aplican para su reducción y mitigación. Para finalizar, se aborda las definiciones acerca de una infraestructura de medición avanzada, su estructura y beneficios de su aplicación en los sistemas eléctricos de distribución.

2.1. Pérdidas de energía eléctrica

2.1.1. Definición

Las pérdidas de energía en el flujo del sistema eléctrico se definen como la diferencia de la electricidad inyectada a las redes y la electricidad suministrada a los usuarios finales, y es una imagen del índice de eficiencia así como de la sostenibilidad financiera del sector eléctrico. Las pérdidas eléctricas se presentan principalmente como calor disipado, pero no se aprovecha la energía eléctrica originadas de ellas pese a que forma parte de la energía generada en el sistema. En líneas generales, las pérdidas se refieren a la cantidades de energía disponible en las redes de transmisión y distribución, que los usuarios no pagan [6]. La Figura 2.1 señala las pérdidas que se producen en cada etapa del sector eléctrico.

En la etapa de generación se consideran las pérdidas durante la transformación de la energía que representan aproximadamente dos tercios del total de insumos. Cabe señalar, que las pérdidas debido al consumo propio de la central depende de su nivel de eficiencia, su factor de planta y su antigüedad.

A continuación, las pérdidas en la transmisión se determinan primordialmente por aspectos técnicos y condiciones externas como geográficas y climatológicas . Los sistemas de distribución, en cambio, responden a pérdidas de tipo no técnicas. Puesto que, el suministro y comercialización de la electricidad requiere otras funciones como la conexión, medición y facturación del servicio [1]

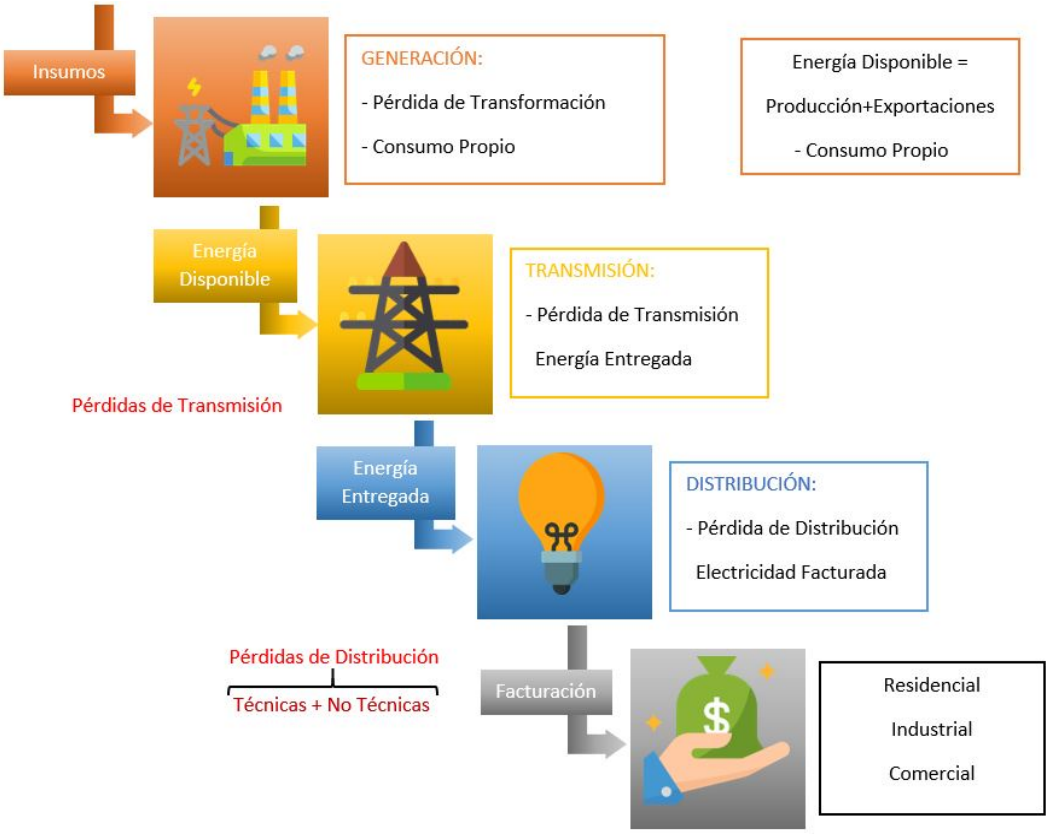


Figura 2.1 – Pérdidas en cada subector del sistema eléctrico -generación, transmisión y distribución-.

Fuente: Tomado de [1]

2.1.2. Clasificación de pérdidas en sistemas eléctricos

El primer tipo de pérdidas de un sistema eléctrico correspondientes a fenómenos físicos se califican como pérdidas técnicas. Estas pérdidas ocurren en las líneas de transmisión y distribución debido a las limitantes propias de conducción y transformación de la energía.

El segundo tipo son las pérdidas no técnicas que se refiere a la cantidad total de pérdidas menos las pérdidas técnicas. Estas pérdidas se produce principalmente en el suministro de la energía al usuario final. La Figura 2.2 muestra en esquema la clasificación de las pérdidas.

2.1.3. Pérdidas técnicas

Las pérdidas técnicas están relacionadas a las pérdidas a causa de las condiciones propias del uso y conducción de la electricidad. Estas pérdidas se vinculan mayormente al estado de las propiedades físicas del sistema y sus partes. Existen muchas maneras de clasificar las pérdidas técnicas, sin embargo, la Organización Latinoamericana de Energía (OLADE) las clasifica en función del componente y las causas que las originan. También se pueden dividir en fijas y variables como se detalla a continuación [6].

Por la función del componente

- Pérdidas por transporte
  - En líneas de transmisión
  - En líneas de subtransmisión
  - En redes de distribución primaria
  - En redes de distribución secundaria
- Pérdidas por transformación
  - En transmisión/subtransmisión
  - En subtransmisión/distribución
  - En transformadores de distribución

Por causa de pérdidas

- Pérdidas por efecto corona
- Pérdidas por efecto Joule
- Pérdidas por corrientes de Foucault e histéresis

Las pérdidas técnicas totales del sistema es la sumatoria de las pérdidas mencionadas anteriormente.

2.1.3.1. Pérdidas técnicas fijas

Son aquellas pérdidas cuya cantidad es aproximadamente independiente de la demanda del sistema. También, denominadas pérdidas en vacío, son proporcionales a las variaciones de voltaje. Se originan en máquinas eléctricas y transformadores, debido a las corrientes parásitas y ciclos de histéresis resultantes de las corrientes de excitación. Además, estas pérdidas se producen en las líneas de transmisión por el efecto corona [6].

Puesto que las variaciones de voltaje en un sistema eléctrico son bajas, las pérdidas de vacío son tratadas como una constante y se calcula en función de la variación de la tensión de la siguiente manera:

$$P_L^{jv} = P_L^{iv}(V^j/V^i)^2 \tag{2.1}$$

donde:

P_L^{iv} : Pérdidas en vacío (W) a un valor de tensión V^i (V).
V^j : Valor de tensión al cual se desea conocer las pérdidas (V).

2.1.3.2. Pérdidas técnicas variables

Estas pérdidas responden a las variaciones de demanda designadas como pérdidas óhmicas. Son causadas por las corrientes que circulan en los componentes relacionados al transporte de la energía como cables, líneas y transformadores. Son pérdidas que se disipan en calor (efecto Joule) y son proporcionales al cuadrado de la corriente.

$$P_L = I^2R \tag{2.2}$$

P_L : Pérdidas en el componente del sistema (W).
I : Corriente que circula por el componente (A)
R : Resistencia en el elemento (Ω)

2.1.4. Pérdidas no técnicas

La pérdidas no técnicas se estiman como la diferencia entre las pérdidas totales de una red eléctrica y las pérdidas técnicas calculadas para la misma [6]. Este tipo de pérdidas son originadas, en particular, en las redes de distribución como consecuencia del aprovechamiento ilícito de la energía, que se convierten en pérdidas financieras directas para la empresa distribuidora. Por este motivo, las pérdidas no-técnicas se designan regularmente como pérdidas comerciales, puesto que la apropiada medición y facturación de la electricidad forma parte integral de la gestión comercial y administrativa de la empresa eléctrica.

Las pérdidas no técnicas o comerciales se definen, también, como la energía entregada al usuario final y que no es facturada por parte de la empresa que la distribuye a causa de alteraciones, manipulaciones y/o anomalías en las instalaciones de los consumidores, en casos singulares son producto de errores administrativos de la propia empresa. A continuación se muestra las potenciales variables, pero no limitantes, que determinan las pérdidas no-técnicas [7]:

- Manipulación del medidor para no registrar el total de la energía consumida.

- Evadiendo los medidores abasteciendo energía directamente de la línea.
- Lecturas de medidores falsas, arregladas sobornando a los lectores de medidores.
- Medidores defectuosos o rotos.
- Suministro no medido.
- Fallas técnicas y humanas en lecturas de medidores, procesamiento de datos y facturación.

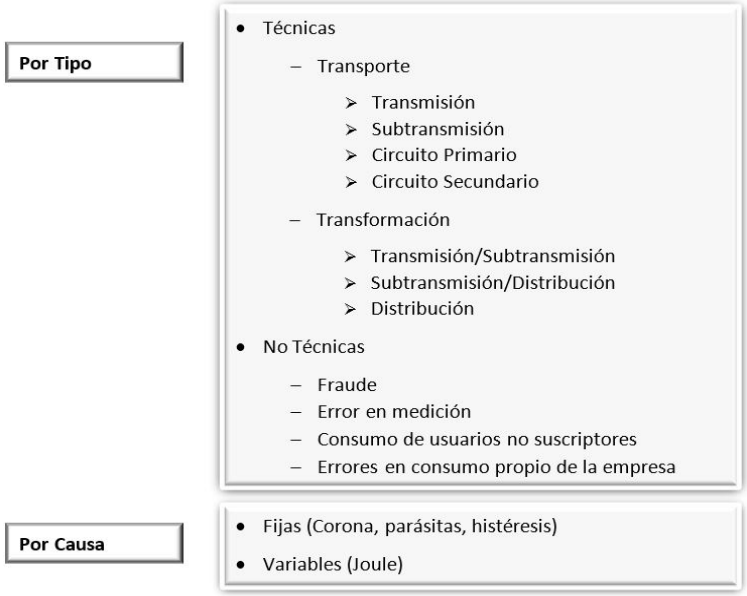


Figura 2.2 – Clasificación de pérdidas eléctricas.
Fuente: Tomado de [6]

2.1.5. Fuentes de pérdidas no técnicas

Las pérdidas no técnicas son producto de factores externos a los sistemas eléctricos. El alcance de estas pérdidas dependerá de una variedad de factores, desde los sociales hasta la manera en que se gestiona la empresa de energía. En todas las redes de distribución predominan cuatro tipo de fuentes de pérdidas no técnicas [8]:

- **Fraude:** se refiere cuando el consumidor intencionadamente trata de engañar a la empresa eléctrica. Una práctica habitual es la modificación de los equipos de medición para que se refleje una lectura de uso de energía mas baja que la real.
- **Hurto:** se refiere cuando el consumidor se apropia ilícitamente de la energía mediante conexiones directas a la red sin pasar por un medidor.
- **Irregularidades de facturación:** se presenta cuando la empresa eléctrica presenta deficiencias en el registro efectivo del consumo de electricidad. Pueden ser irregularidades deliberadas por parte de los propios integrantes de la organización, por ejemplo, aceptando sobornos para arreglar facturas mucho más bajas que por la energía realmente utilizada.

- **Facturas sin pagar:** algunos usuarios y organizaciones no pagan lo que deben por la electricidad. Consumidores residenciales o comerciales pueden haber abandonado la ciudad o una empresa quebró.

2.1.6. Clasificación de pérdidas no técnicas

Las pérdidas no técnicas se pueden clasificar según su fuente de la siguiente manera [9]:

- **Pérdidas administrativas:** Es equivalente a la energía no contabilizada por errores administrativos por parte de los empleados de esta área.
- **Pérdidas fraudulentas:** Son pérdidas generadas por la manipulación de los medidores o por el hurto de energía, son las comunes en los sistemas eléctricos de países en desarrollo.
- **Pérdidas no identificadas:** Incluyen pérdidas por conexiones indebidas en un horario o un transcurso de tiempo del día que no se puedan detectar, normalmente durante la noche.
- **Pérdidas accidentales:** Implica errores en los medidores por defectos de los mismos, falta de mantenimiento o envejecimiento del equipo.
- **Pérdidas por consumo estimado:** Estas pérdidas ocurren cuando los consumidores se encuentran suscritos en la empresa distribuidora y no cuentan todavía con un equipo de medición.

2.1.7. Consecuencias de las pérdidas no técnicas

Los niveles de pérdidas no técnicas constituyen una importante medida de eficiencia operativa de una empresa distribuidora y comercializadora de electricidad; por lo que, es indispensable comprender y evaluar las consecuencias que generan en diferentes aspectos [10], que se presentan enseguida.

2.1.7.1. En la gestión técnica-económica de la empresa

La falta de regulaciones y políticas sobre las pérdidas produce varios efectos en la gestión empresarial, tales como:

- Mayor daño de las redes e instalaciones, presionando a las empresas eléctricas a realizar fuertes gastos en ampliaciones como en renovaciones que posteriormente no rinden correctamente, ya que son sobredimensionadas para tolerar el incremento excesivo de los consumos.
- Pérdidas comerciales por el consumo de energía no facturada, con un aumento de costos de producción o entrega de la electricidad, igual que los de aprovechamiento, causado por el incremento del mantenimiento a las redes e instalaciones debido a las prácticas ilícitas sobre las mismas.

Cuando no existen las debidas medidas regulatorias para controlar y reducir los índices de pérdidas, comúnmente se presenta una mayor aceleración de su tasa de crecimiento que el impacto de las medidas que se toman. De cierta manera esto provoca un grado de impotencia en la organización de la empresa que con el tiempo no es de interés, que favorece la degeneración de los procedimientos y controles, lo cual conduce a un fuerte deterioro funcional dando lugar a:

- Una formación deficiente en los responsables de control y supervisión.
- Encubrimiento de acciones ilegales por parte del mismo personal de la empresa así sea en beneficio propio o de terceros, que afectan económicamente a la corporación.
- Origen de un sentido de impunidad entre los usuarios, que conlleva a un constante crecimiento del fraude o robo de la energía.

2.1.7.2. En el aspecto social

Muchos factores como la crisis económica de un país, la falta de regulaciones adecuadas, la ausencia de acción policial correctiva, la falta de resoluciones administrativas, la baja predisposición a pagar por servicios públicos, y otros contextos han creado una cultura de impunidad total en la sociedad consumidora ante las prácticas ilícitas para apropiarse de la energía eléctrica.

Esto acarrea a que los consumidores que cumplen regularmente con sus responsabilidades y pagos, incitados por aquellos que realizan acciones fraudulentas, opten también por emularlos y procedan a no pagar las facturas, para luego conectarse directamente a la red, o se motivan a alterar los equipos de medición evitando los registros reales, generalizándose así las actividades ilegales.

2.1.7.3. En el aspecto ético y moral

El robo de electricidad mediante conexiones directas ilícitas en las redes y las manipulación de las mediciones para registrar falsas lecturas, realizado de manera indiscriminada y con una elevada impunidad, a más de generar perjuicios económicos a la empresa distribuidora del servicio, influye en gran medida sobre la ética y la moral de la población.

En ciertas zonas marginales puede ser común que se apropien de la energía sin pagar por mínima que sea, aunque no es justificable, pero por ser personas de escasos recursos no cuentan con los medios para pagar las tarifas impuestas incluso podría deberse a la incapacidad de conexión debido a falta de títulos de propiedad. No lo es tanto para zonas residenciales, comerciales e industriales, los cuales si disponen de los recursos necesarios, que convierte a este accionar en una contravención totalmente sancionable, pues persigue fines de lucro, alentando la competencia desleal y la evasión fiscal incidiendo sobre la sociedad.

2.1.7.4. En los aspectos de seguridad

Para apropiarse de la energía eléctrica ilícitamente, los usuarios ocasionan graves agresiones sobre las instalaciones, lo cual conduce a un acelerado daño de la mismas con verdaderas repercusiones para la seguridad pública. Por ejemplo, el procedimiento

para conectarse ilegalmente a la red presenta riesgos de seguridad para la persona que lo realiza, como electrocutarse e incluso la muerte.

2.1.8. Métodos para reducir las pérdidas No Técnicas

Las siguientes métodos suelen ser efectivos para reducir las pérdidas no técnicas [8], en especial para afrontar el conjunto de problemas que adjudica al fraude eléctrico.

2.1.8.1. Métodos técnicos/ de ingeniería

Las innovaciones que se producen en el sector eléctrico permiten instalar y mantener sistemas muy eficientes. Muchos sistemas energéticos dedican recursos y esfuerzos inadecuados a los sistemas de transmisión y distribución, pero no utilizan las últimas tecnologías. La inversión necesaria para reducir las pérdidas incluyen la renovación de líneas eléctricas, transformadores, sistemas de monitoreo de tecnología de la información e instalación y mantenimiento de sistemas de medición modernos que se encuentran en la interfaz de la organización y los consumidores de electricidad.

Se ha producido un avance tecnológico significativo en la medición. Dado que gran parte del robo se deben a la manipulación del medidor, es indispensable reemplazar los antiguos medidores y fáciles de manipular. Los nuevos medidores sellados de alta tecnología que no se pueden alterar de ninguna forma y se pueden leer automáticamente son costosos, sin embargo pueden reducir el robo cuando lo requieran los usuarios moderados a intensos. La inversión en medición de alta tecnología requiere una infraestructura sólida y compleja para que el sistema funcione de manera eficaz.

2.1.8.2. Métodos administrativos

Las corporaciones de energía son entidades muy grandes que operan como burocracias, aunque muchas son organizaciones del sector privado. La combinación de fuertes mejoras técnicas con un programa antirrobo inteligente y activo puede resultar en mejoras significativas [8].

El monitoreo y la inspección de los usuarios de energía en periodos regulares es esencial para reducir este tipo de pérdidas, dando especial atención a aquellas zonas o instalaciones que presentan mayor potencial en lo que respecta al hurto de energía.

La corrupción es una de las problemáticas más difíciles de enfrentar para las empresas eléctricas, pues el robo de electricidad se desarrolla con la complicidad de los empleados de la empresa. Es de suma importancia detectar y sancionar a aquellos elementos corruptos, incluso si estos pertenecen a puestos gerenciales de la organización. Una medida para evitar la corrupción es mantener bien salariado a los empleados para que no tengan que recurrir a sobornos.

Para afrontar el robo de energía diversos países han renovado leyes obsoletas por nuevas leyes, recibiendo atención especial y logrando que las sanciones o penalizaciones sean mucho más fáciles de implementar para impedir futuros robos.

Por otro lado, una cultura de impagos puede ser un problema que compromete la viabilidad financiera de la empresa. La contratación de la recaudación de facturas a una agencia privada puede promover cierta eficacia en la recaudación de ingresos.

Para reducir este inconveniente es preciso promocionar métodos alternativos como facilidades de pago, convenios o concesiones, mayor accesibilidad a agencias de pago. No obstante, el problema de los atrasos o impagos no tiene soluciones sencillas. En casos

particulares, las empresas estatales son las que deben más dinero, y la recaudación puede enfrentar obstáculos legales y políticos.

2.2. Infraestructura de Medición Avanzada

Un sistema con infraestructura de medición avanzada AMI (por sus siglas en inglés, Advanced Metering Infraestructure) se encarga de la medición, almacenamiento, recopilación y transferencia remota de los datos relacionados al consumo energético de los usuarios; para su posterior presentación, análisis, gestión y toma de decisiones [11]. Este sistema evita efectuar exhaustivos procedimientos como: la medición en el lugar de consumo, acciones de conexión y desconexión del servicio así como la gestión de rehabilitación del servivio. AMI permite implementar un control directo en la gestión de la demanda y responder rápidamente a sus variaciones con el monitoreo en tiempo real de los precios de la energía. AMI es la evolución de los sistemas AMR (Automatic Meter Reading). En síntesis, el Instituto de Energía Eléctrica (EPRI, Electric Power Research Institute) define un sistema AMI como:

> "Un sistema completo de medición y recopilación que incluye medidores inteligentes (Smart Meters) en el predio del cliente; redes de comunicación entre el cliente y el proveedor de servicios; y un sistema de recepción y gestión de datos que facilite la información para el proveedor de servicios."

2.2.1. Arquitectura de un AMI

La estructura de un sistema AMI es jerárquico y está compuesto por hardware y software de avanzada tecnología la cual combina la medición de la información en intervalos con las comunicaciones remotas [12]. Generalmente, se constituye de tres principales componentes: medidores inteligentes, redes de comunicación y sistemas de gestión de datos de medición (MDMS). La Figura 2.3 presenta la arquitectura de un sistema AMI.

En la arquitectura AMI los usuarios están equipados de un medidor electrónico de estado sólido que recopila los datos del consumo energético en tiempo real. Estos medidores pueden transmitir la información recopilada mediante redes fijas continuamente disponibles, como banda ancha sobre línea eléctrica (BPL), comunicaciones por línea eléctrica, radiofrecuencia fija; al igual que redes públicas teléfonos fijos, celulares y buscapersonas. Los datos de consumo medidos son receptados por el sistema host de AMI. Por último, se remite al centro de control (MDMS) que administra el almacenamiento y análisis de datos y brinda la información de manera eficaz para la empresa eléctrica [14]. Cabe recalcar que, la comunicación de un sistema AMI es bidireccional.

2.2.2. Equipos de recolección de datos

Estos equipos poseen la capacidad de registrar la información de consumo de los usuarios, así como la capacidad de emitirlos a un nivel jerárquico superior del sistema. Están compuesto por los medidores inteligentes y los concentradores de datos.

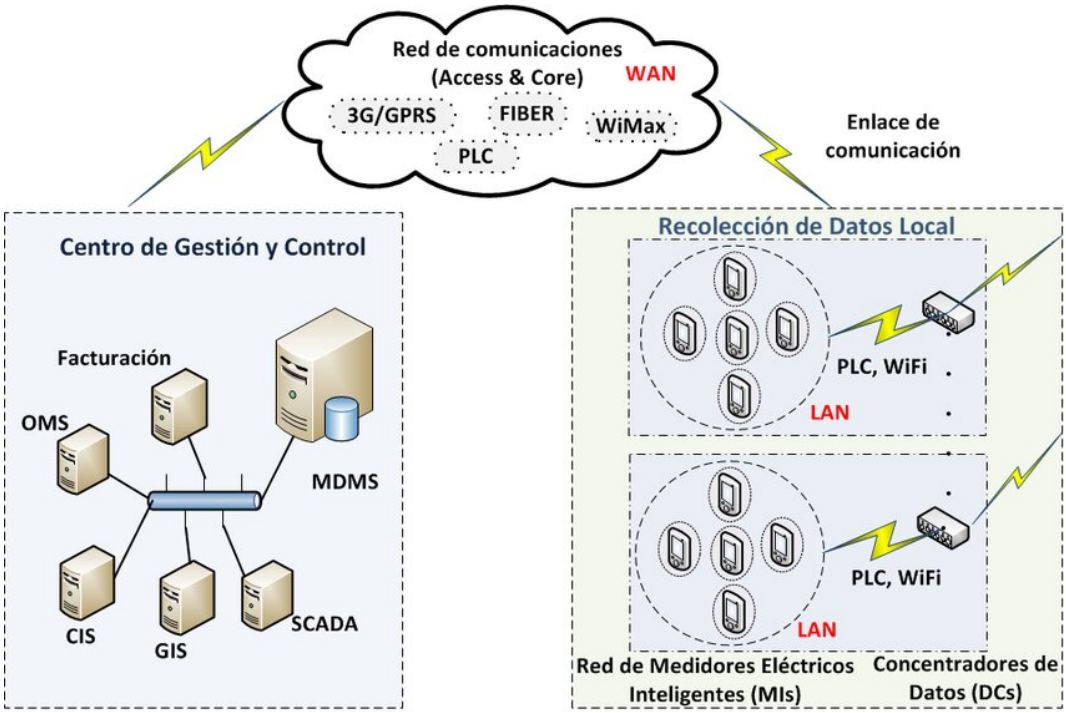


Figura 2.3 – Arquitectura general de un sistema AMI.
Fuente: Tomado de [13]

2.2.2.1. Medidores Inteligentes

El medidor inteligente (ver Figura 2.4) es el equipo que realiza la medición, registro y almacenamiento de la información de consumo de energía, verifica el estado del suministro del servicio, además registra eventos y datos de variables eléctricas; todos estos datos son transmitidos en tiempo real hacia concentradores y seguidamente al centro de gestión de la empresa distribuidora, a fin de llevar a cabo un análisis de ingeniería que faciliten la optimización de procesos y la gestión operativa. Hay varias funcionalidades básicas que los medidores deben tener independientemente del tipo o cantidad de su medida. Estas funcionalidades comprenden [14]:

- **Medición cuantitativa:** el medidor debe ser capaz de medir con exactitud la cantidad de consumo utilizando diferentes principios físicos, topologías y métodos.
- **Control y calibración:** aunque varíe según el tipo, por lo regular, el medidor debería lograr nivelar las pequeñas variaciones.
- **Comunicación:** enviar información almacenada y recibir comandos operativos, al igual que la cualidad de recibir actualizaciones de firmware.
- **Gestión de energía:** si la fuente de energía primaria falla o se apaga, el sistema debe ser capaz de mantener su funcionalidad.
- **Pantalla:** los usuarios son capaces de ver la información del medidor, ya que esta información es la base para la facturación. También se necesita una pantalla, ya que la gestión de la demanda en el extremo del usuario no será posible sin el conocimiento del consumo del usuario en tiempo real.

- **Sincronización:** la sincronización del tiempo es primordial para una fiable transmisión de datos al concentrador central u otros sistemas de recopilación para el análisis de datos y la facturación. La sincronización de tiempo es aún más relevantes cuando se trata de comunicación inalámbrica.

Las características más importantes de los medidores inteligentes eléctricos se especifican a continuación [12]:

- Precios basados en tiempo.
- Datos de consumo para consumidores y empresa distribuidora.
- Medición neta.
- Notificación de pérdida de energía (y restauración).
- Operaciones de encendido/apagado remoto .
- Limitación de carga para propósitos de respuesta de demanda o de "mala paga".
- Pago anticipado de energía.
- Supervisión de la calidad de la energía.
- Detección de fraude y hurto de energía.
- Comunicaciones con otros dispositivos inteligentes en el hogar.

Figura 2.4 – Un medidor inteligente de estado sólido moderno.
Fuente: Tomado de [15]

2.2.2.2. Concentradores de datos

Los concentradores de datos cuentan con mayor capacidad de almacenamiento de información y proveen la infraestructura de conectividad entre los medidores y el software de gestión de lectura dispuestos en el centro de control. Por lo general, estos equipos se comunican con los medidores inteligentes vía radio frecuencia o la tecnología PLC (Power Line Communication), y normalmente se conectan al lado secundario de los transformadores de distribución. El esquema de la Figura 2.5 presenta la ubicación de este equipo dentro de una estructura de medición avanzada. Los principales servicios que brindan los concentradores de datos son los siguientes [16]:

- Detección automática de medidores.
- Detección de fase del equipo de medición.
- Adquisición de los datos de facturación.
- Mantenimiento de los ajustes de tiempo de los medidores.
- Adquisición de informes de la manipulación de medidores.
- Detección de fallos (pueden ser errores individuales de los equipos o cortes de línea).
- Detección de inversión de fase del medidor.
- Control remoto de medidores (conexión, desconexión, limitación de la capacidad temporal, etc).
- Adaptación automática a cambios en la topología de la red.
- Indican automáticamente el conjunto óptimo de medidores que deben actuar como repetidores contadores para asegurar la fiabilidad de comunicación completa con todos los medidores.
- Seguridad de datos: encriptación y validación.
- Soportar varias tecnologías y medios de comunicación.

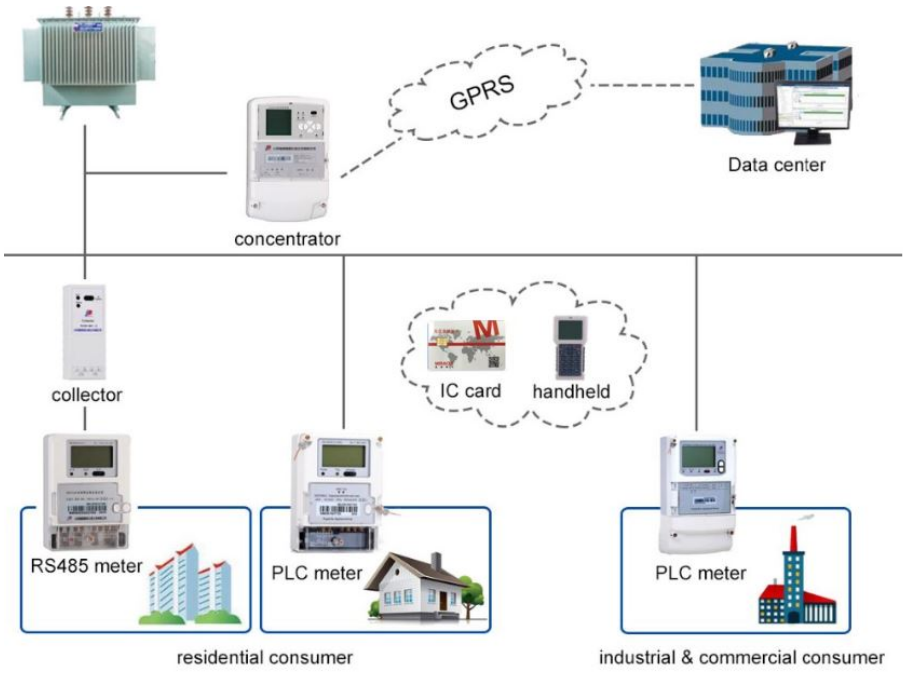


Figura 2.5 – Localización del concentrador de datos.
Fuente: Tomado de [17]

2.2.3. Redes de comunicación

La comunicación es uno de los desafíos tecnológicos primordiales para un sistema AMI. Todo medidor es capaz de transmitir de manera segura y fiable los datos recolectados a un receptor central, no obstante, considerando los diversos lugares y entornos en los que se ubican los medidores, se muestran distintos inconvenientes, de modo que, se optan por soluciones tecnológicas de comunicación tales como: comunicaciones por línea eléctrica (PLC), banda ancha sobre líneas eléctricas (BPL), cobre o fibra óptica, celular, WiMax, bluetooth, servicio general de radio por paquetes (GPRS), internet, satélite, Peer-to-Peer, Zigbee, etc., existiendo variedades de configuraciones de red [18]. Adicionalmente, se requiere que los sistemas de comunicación sean altamente confiables para transferir los grandes volúmenes de datos considerando la cantidad de usuarios y medidores.

Un sistema AMI tiene una estructura jerárquica compuesta por varias redes diferentes que se comunican entre sí, como se visualiza en la figura 2.6, y estas redes se describen enseguida: red de área doméstica (HAN), red de área vecinal (NAN) y red de área extensa (WAN)[19].

La red de área doméstica es la red de sensores que se conectan a los electrodomésticos en los predios del usuario y se comunican con las puertas de enlace de los usuarios o directamente con los medidores inteligentes en áreas residenciales e industriales. Las tecnologías sugeridas para esta red son aquellas con pila de protocolo Zigbee y Wifi.

La red de área vecinal es una red compuesta por medidores inteligentes cercanos que se comunican con los nodos colectores para transferir los datos recolectados a un nivel superior. La tecnología de comunicación que se utiliza habitualmente para esta red es el WiFi, pero se pueden incluir otras como WiMax y 3G/4G celular.

La red de área extensa es una red multipropósito que proporciona conectividad a los colectores NAN con las cabeceras en el centro de control de la empresa. Dada la gran cantidad de datos que se transfieren, se requiere de un amplio ancho de banda, muy alta confiabilidad y usualmente se constituye de tecnologías como fibra óptica, WiMax, celular, sateltal, Metro Ethernet, y comunicación por línea eléctrica.

La selección y el diseño del adecuado sistema de comunicación es un riguroso proceso que debe cumplir con los siguientes requerimientos fundamentales:

- Gran cantidad de transferencia de datos.
- Restricción en el acceso de datos.
- Confidencialidad de datos sensibles.
- Mostrar el estado de la red.
- Representar la información. completa del consumo del cliente.
- Autenticidad de los datos y precisión en la comunicación con el dispositivo de destino.
- Rentabilidad.
- Soporte a la expansión futura.
- Capacidad para alojar funciones modernas más allá de los requisitos AMI.

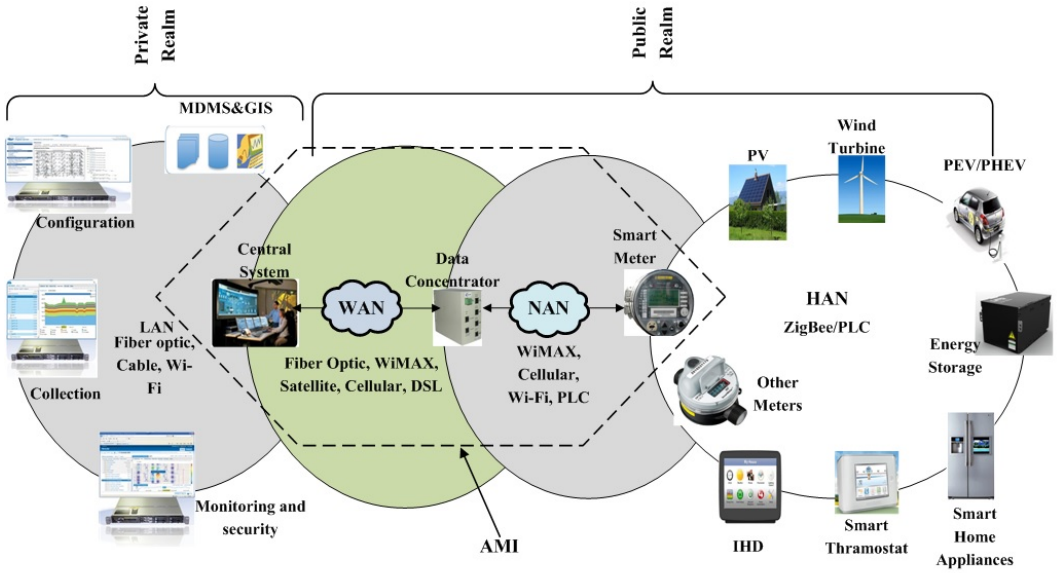


Figura 2.6 – Infraestructura típica de comunicación en sistemas AMI.
Fuente: Tomado de [20]

2.2.4. Sistema de Gestión de Datos

En el centro de control de la empresa proveedora del servicio se establece un sistema de gestión de datos para almacenar y analizar los datos recolectados de los medidores inteligentes con fines de facturación. Además, el manejo de la respuesta de la demanda (DR), perfiles de consumo y reacciones en tiempo real a los cambios e imprevistos en la red. Este sistema por su diversas capacidades de análisis permite la incorporación de otros sistemas de servicios públicos, tal como se muestra en la Figura 2.7 y se enlistan a continuación [14]:

- Sistema de gestión de datos de medición (MDMS).
- Sistema de información al consumidor (CIS), sistema de facturación y sitio web de la empresa de servicios públicos.
- Sistema de gestión de interrupciones (OMS).
- Planificación de recursos empresariales (ERP), gestión de la calidad de la energía y sistemas de pronóstico de carga.
- Gestión de la fuerza de trabajo (MWM).
- Sistema de información geográfica (SIG).
- Gestión de carga de transformadores (TLM).

El MDMS se considera el módulo central del sistema de gestión con las herramientas analíticas necesarias para la comunicación con otros sistemas integrados en el. De igual manera, tiene la responsabilidad de realizar la validación, edición y estimación de los datos de AMI para garantizar un flujo preciso y completo de información desde el cliente a los sistemas de gestión ante posibles interrupciones en las capas inferiores.

Los datos recopilados son enormes y se denominan "Big Data" cuando la recopilación de datos en AMI son en intervalos de 15 min, están en el orden de los terabytes.

La puesta en marcha de un sistema de gestión de datos tendrá que estar sincronizado con la implementación del sistema AMI y tendrá que ejecutarse en el inicio o cerca, con el objetivo de establecer debidamente las interfaces con otros sistemas y lograr la eficiencia operativa de las funcionalidades de AMI al gestionar las grandes cantidades de datos de las mediciones.

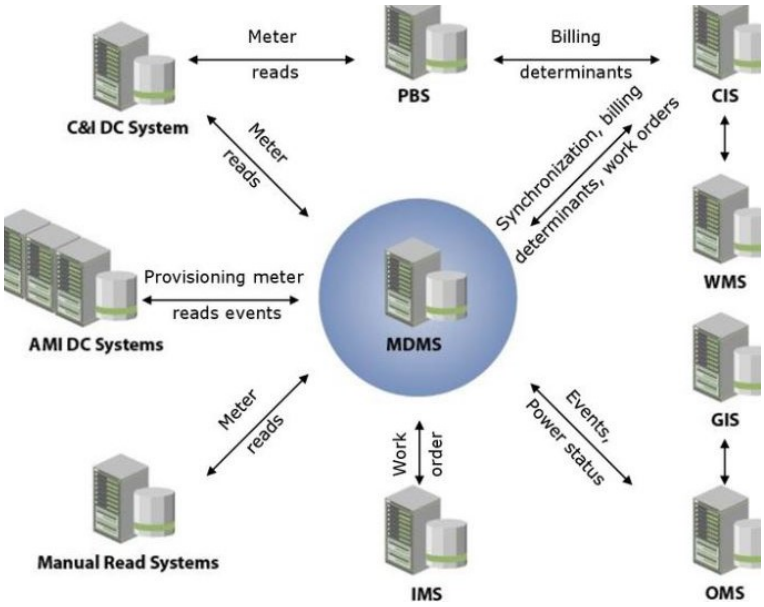


Figura 2.7 – Integración de MDMS en un sistema AMI con otros sistemas empresariales como: CIS, GIS, OMS, etc.
Fuente: Tomado de [21]

2.2.5. Beneficios de los sistemas AMI

El sistema AMI proporciona beneficios a los consumidores, las empresas de servicios públicos, y la sociedad en su conjunto [12].

2.2.5.1. Beneficios para el consumidor

Para el consumidor, esto representa más alternativas sobre precios y servicios, menos irregularidades y más información para gestionar el uso de la energía, los gastos y otras decisiones. Además significa una mayor confiabilidad, mejor calidad de la energía y una facturación más rápida y precisa. También, AMI ayudará a mantener bajos los costos de los servicios públicos, de modo que los precios de la energía serán bajos. Así, como miembros de la sociedad, los consumidores obtienen todos los beneficios que se acumulan para la sociedad en general, como se describe más adelante.

2.2.5.2. Beneficios para la empresa distribuidora

Los beneficios para la empresa de servicios públicos se puede categorizar principalmente en facturación y operatividad.

AMI ayuda a la empresa a evitar lecturas estimadas en el predio del usuario, proporcionar facturas precisas y oportunas, operar de manera más eficiente y confiable, y ofrecer un servicio al consumidor significativamente mejor. AMI elimina el vehículo, la capacitación, el seguro médico y otros gastos generales de la lectura manual del medidor, mientras que el tiempo más corto de lectura para pagar avanza el flujo de efectivo de la empresa, creando un beneficio único.

Desde la perspectiva operativa, con AMI, la empresa sabe de inmediato cuándo y dónde ocurre una interrupción para poder enviar equipos de reparación de una manera más adecuada y eficiente. La información de interrupción y restauración a nivel de medidor acelera el proceso de restauración de interrupciones, que incluye notificar a los consumidores sobre cuándo es probable que vuelva la energía.

Con AMI, la empresa puede recibir importantes beneficios al poder administrar las cuentas de los clientes de manera más rápida y eficiente, empezando con la capacidad de conectarse y desconectar el servicio de forma remota sin tener que enviar personal al sitio del cliente. De manera similar, muchos problemas de mantenimiento y servicio al cliente se pueden resolver de forma más rápida y rentable a través de diagnósticos remotos.

AMI también proporciona grandes cantidades de información sobre el uso de energía y el estado de la red que los consumidores pueden utilizar para tomar decisiones de consumo más informadas y las empresas para tomar mejores decisiones sobre las mejoras del sistema y las ofertas de servicios.

En lugar de depender de estimaciones aproximadas, los ingenieros armados con el conocimiento detallado de AMI sobre las cargas de distribución y la calidad eléctrica pueden dimensionar con precisión los equipos y los dispositivos de protección, y comprender mejor el comportamiento del sistema de distribución. Este enorme aumento de información valiosa ayuda a la empresa a:

- Evaluar el estado del equipo.
- Maximiza la utilización y vida de los activos.
- Optimizar los gastos de capital, operación y mantenimiento.
- Identificar problemas de cuadrícula.
- Mejorar la planificación de la red.
- Localizar e identificar problemas de calidad de energía.
- Detectar y reducir el hurto de energía.

2.2.5.3. Beneficios para la sociedad

La sociedad, en general, se beneficia de AMI de muchas formas. Una manera es mediante una mayor eficiencia en el suministro y uso de energía, produciendo un impacto ambiental favorable. Puede acelerar el uso de la generación distribuida, lo que a su vez puede fomentar el uso de fuentes de energía renovables. Y es probable que las capacidades de registro y medición detalladas de AMI permitan el comercio de emisiones.

Un beneficio importante de AMI es que facilita la respuesta a la demanda y tarifas energéticas innovadoras. Durante los periodos de alta demanda de energía, una pequeña

reducción de la demanda produce una reducción relativamente grande en el precio de mercado de la electricidad. Y una demanda reducida puede evitar apagones continuos.

También existe un problema de equidad social que aborda AMI. El despliegue completo de AMI da como resultado la eliminación de medidores electromecánincos viejos y obsoletos que tienden a disminuir a medida que envejecen. Los medidores inteligentes AMI mantienen su precisión a lo largo del tiempo, lo que resulta en una situación más equitativa para todos los consumidores. Además, los medidores inteligentes son automonitorizados, lo que facilita la identificación de mediciones inexactas, instalaciones incorrectas y, especialmente, hurto de energía eléctrica.

Capítulo 3

MINERÍA DE DATOS

El constante crecimiento de los datos en los últimos treinta años es resultado de la digitalización de la sociedad y la continua evolución de potentes herramientas de recopilación y almacenamiento de datos. Cada día se generan enormes cantidades de datos, que van desde el orden de terabytes hasta petabytes almacenados en grandes y complejas bases de datos. Ante la inminente situación de: riqueza en la información pero pobreza en conocimiento, es necesario el desarrollo de nuevos métodos científicos y herramientas especializadas para que el procesamiento y análisis de datos sea mucho más efectivo y productivo, con el fin de obtener información valiosa y constructiva para la toma de decisiones.

La minería de datos está compuesta de herramientas poderosas y versátiles que sirven para descubrir automáticamente información útil de los grandes volúmenes de datos y convertir dichos datos en conocimientos organizados [22]. Hoy en día toda empresa o entidad requiere de la utilización de métodos de minería de datos para analizar eficientemente las grandes cantidades de datos almacenados, en vista de que resulta imposible hacerlo con los métodos clásicos. Las técnicas de minería de datos desarrollan modelos y algoritmos sofisticados que se basan en campos de estudios como estadísticas, inteligencia artificial y aprendizaje automático; se aplican sobre el vasto conjunto de datos reconociendo patrones o tendencias para así extraer información relevante que permita implementar estrategias de decisión según las necesidades de la empresa.

Los dos principales objetivos de la minería de datos suelen ser la descripción y la predicción. La descripción, se concentra en descubrir patrones que representen los datos que pueden ser analizados por humanos. La predicción, por otra parte, involucra el empleo de algunas variables del grupo de datos para predecir valores futuros o desconocidos. Estos objetivos se logran con la ayuda de diversas técnicas de minería de datos, a fin de cumplir las siguiente tareas primarias de mínería de datos [23]:

- *Clasificación:* encontrar un modelo de aprendizaje predictivo que clasifique un elemento de datos en una de varias clases predeterminadas.
- *Regresión:* encontrar un modelo de aprendizaje predictivo que asigne un elemento de datos a una variable de predicción de valor real.
- *Agrupación:* una común tarea descriptiva en la cual se busca reconocer un conjunto finito de grupos o categorías para describir los datos.

- *Resumen:* una tarea descriptiva adicional que incluye métodos para encontrar una descripción compacta para un conjunto de datos.
- *Modelado de dependencia:* encontrar un modelo local que describa dependencias significativas entre variables o entre los valores de un atributo en un conjunto de datos o en una parte de un conjunto de datos.
- *Detección de cambios:* encontrar los cambios más significativos en el conjunto de datos.

3.1. Proceso de Minería de Datos

La minería de datos es un proceso iterativo de descubrimiento de varios modelos, resúmenes y valores derivados de una determinada colección de datos, mediante métodos automáticos, semiautomáticos o manuales. El procedimiento general aplicado a los problemas de minería de datos implica los siguientes pasos:

1) **Planteamiento del problema y formulación de la hipótesis.** Es posible que se formulen varias hipótesis para un solo problema en esta etapa.
2) **Recolección de datos:** En este paso se generan y se recopilan los datos.
3) **Preprocesamiento de datos.** Transformación de la data bruta en datos que presenten un formato más sencillo de utilizar.
4) **Estimación del modelo.** La tarea principal en esta fase es la selección e implementación de la técnica apropiada de minería de datos.
5) **Interpretación del modelo.** Los modelos deben ser interpretables para que sean útiles en la toma de decisiones.

Cada de una de las fases del proceso, en particular, la fase de extracción de datos, son altamente iterativos, como se ilustra en la Figura 3.1. Si importar cuan poderoso y complejo llegue a ser el método de extracción de datos, empleado en el paso 4, si los datos no son recopilados y preprocesados debidamente o si el planteamiento del problema no es significativo, el modelo resultante no será valido.

El proceso de minería de datos implica la integración de métodos y técnicas interdisciplinarias como tecnologías de base de datos, aprendizaje de máquina, estadística, reconocimiento de patrones, computación de alto desempeño, visualización de datos, procesamiento de imágenes, recuperación de la información, entre otros [24].

Por último, cabe destacar la considerada importancia que está tomando la minería de datos en los sistemas de información y base de datos, a la par, es una de las ciencias multidisciplinarias más sobresaliente en las tecnologías de la información.

3.2. Análisis de grupos

El análisis de grupos es el estudio formal de métodos y algoritmos para el agrupamiento natural, o *clustering*, de objetos de datos con similares características, en los que se desconoce el número de grupos y sus formas. La forma del grupo se refiere

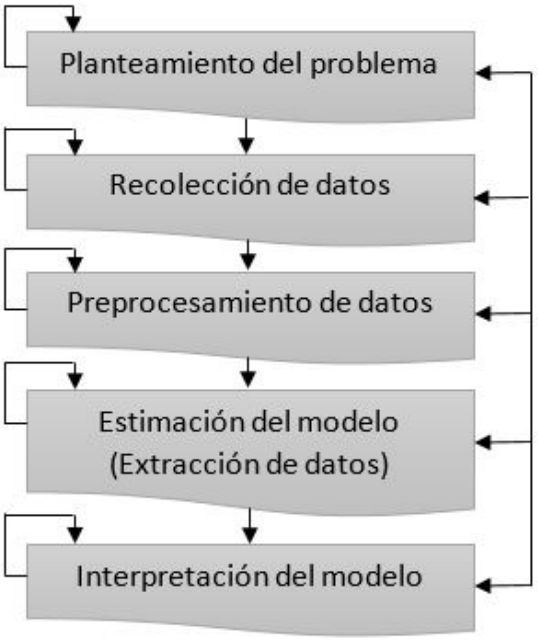


Figura 3.1 – Proceso de minería de datos.
Fuente: Tomado de [23]

a los parámetros de las agrupaciones, es decir, a sus medias, varianzas y covarianzas específicas que también tienen una interpretación geométrica.

A diferencia de la clasificación y la regresión, que analizan conjuntos de datos etiquetados por clase (entrenamiento), la agrupación analiza los objetos de datos sin consultar las etiquetas de clase. En muchos casos, es posible que los datos con etiqueta de clase simplemente no existan al principio. La agrupación en *clusters* se puede utilizar para generar etiquetas de clase para un grupo de datos. Los objetos se agrupan o se agrupan según el principio de maximizar la similitud intraclase y minimizar la similitud interclase. Es decir, los grupos de objetos se forman de modo que los objetos dentro de un grupo tienen una gran similitud entre sí, pero son bastante diferentes a los objetos de otros grupos. Cada grupo así formado puede verse como una clase de objetos, de los cuales se pueden derivar reglas. La agrupación también puede facilitar la formación de taxonomías, es decir, la organización de observaciones en una jerarquía de clases que agrupan eventos similares.

En este sentido, diversas técnicas de agrupamiento pueden generar diversas agrupaciones en el mismo conjunto de datos. La partición la realizan los algoritmos de agrupamiento, mas no los humanos. Por lo tanto, la agrupación en clusters es útil porque puede conducir al descubrimiento de grupos previamente desconocidos dentro de los datos. Cabe mencionar que no existe un método de agrupamiento que sea generalmente aplicable para descubrir la variedad de estructuras presentes en conjuntos de datos multidimensionales. Los mejores criterios para seleccionar el método adecuado estarán dados por la compresión del usuario del problema y los tipos de datos correspondientes. En teoría, hay dos puntos de vista para la mayoría de algoritmos en la metodología de agrupación en clusters [25]:

- Agrupamiento jerárquico

- Agrupamiento no jerárquico/particional

El agrupamiento jerárquico presenta aglomeraciones sucesivas utilizando aglomeraciones previamente determinadas, formando así una jerarquía de agrupaciones (produciendo un dendrograma o un diagrama de árbol) y no solamente una simple partición de objetos. El número de grupos no se requiere como condición de entrada del algoritmo, mientras que se puede utilizar una condición preestablecida para finalizarlo. Por ejemplo, un número predeterminado de grupos. Por tanto, se puede obtener el número deseado de grupos "cortando" el dendrograma al nivel adecuado. *El agrupamiento no jerárquico* consiste en la división inicial de objetos en subconjuntos no superpuestos, por lo que cada objeto pertenece a un solo grupo. Los algoritmos no jerárquico intentan obtener la partición que minimiza la dispersión dentro del grupo o maximiza la dispersión entre grupos.

El procedimiento de agrupamiento depende del tipo de medida de similitud escogido para la segmentación de objetos. Por consiguiente, se pueden agrupar de varias formas, considerando el tipo de similitud entre ellos. Tomando en cuenta las anteriores consideraciones, los pasos principales a examinar en el proceso de agrupamiento son los siguientes:

1. Formulación de problemas - la selección de objetos para agruparlos
2. Selección del modelo de agrupamiento
3. Selección del número de grupos
4. Ilustración gráfica e interpretación de grupos (sacar conclusiones)
5. Evaluar la validez y solidez del modelo utilizando varios métodos, tales como:
    - Repetición del proceso utilizando otras medidas de similitud correspondientes al contexto
    - Repetir el proceso utilizando otras técnicas de agrupación apropiadas
    - Repetir el proceso varias veces, pero ignorando en cada iteración uno o más objetos

3.2.1. Medidas de similitud

El primer paso efectivo para el análisis de grupos es la medición de la similitud entre objetos. Al respecto del proceso de agrupamiento, la similitud es una medida de correspondencia que indica que tan similares son dos objetos. No obstante, con frecuencia, en vez de usar la similitud, se puede considerar la disimilitud puesto que es más conveniente para la idea de medir la distancia entre objetos [25]. La similitud entre objetos puede ser medida de muchas maneras, pero resaltan tres métodos para aplicaciones de agrupamiento: medidas de correlación, medidas de asociación y medidas de distancia [26].

Los algoritmos de agrupamiento que se desarrollan en este proyecto se basan en medidas de distancia para observar la similitud entre objetos, a continuación explicaremos en detalle.

- **Medidas de distancia:** Es la medida de similitud más utilizada en el análisis de grupos, que indica la similitud como la adyacencia de los objetos con relación a los otros objetos del grupo. Son mediciones de disimilitud, entre más alto valores que se tengan mayor será la distancia entre los objetos [26].

  Enseguida se presenta algunas de las medidas de distancia más populares, que se aplican en casi todos los casos. Pero antes, tenemos que especificar que, para medir la similitud entre dos objetos, consideraremos el par de vectores: $x = (x_1, x_2, ..., x_n)$, $y = (y_1, y_2, ..., y_n)$, con la misma dimensión n por términos de simplicidad [25].

  1. *Distancia Minkowski*

  $$d_p(x, y) = \left(\sum_{i=1}^{n} |x_i - y_i|^p \right)^{1/p}, \quad p \in N \tag{3.1}$$

  2. Para $p = 1$ se obtiene la *distancia Manhattan*

  $$d_{cb}(x, y) = \sum_{i=1}^{n} |x_i - y_i| \tag{3.2}$$

  3. Para $p = 2$ se obtiene la *distancia Euclidiana*

  $$d_E(x, y) = \left(\sum_{i=1}^{n} |x_i - y_i|^2 \right)^{1/2} \tag{3.3}$$

  4. Para $p = \infty$ se obtiene la *distancia Chebychev*

  $$d_C(x, y) = \max_i |x_i - y_i| \tag{3.4}$$

  Cuando se aplica estas distancias resulta una matriz diagonal llamada también matriz de similaridad, donde las entradas de la matriz son las distancias entre cada objeto y los demás.

3.2.2. Métodos de Agrupamiento

Hay una gran diversidad de métodos desarrollados para realizar las tareas de agrupamiento, sean jerárquicos o particional. Con el fin de lograr los objetivos, en este proyecto se han seleccionado tres algoritmos por su sencillez pero a su vez por su eficiente funcionalidad para el agrupamiento de datos, según la revisión literaria. El más conocido de ellos es el algoritmo K-medias, el siguiente es una variación del antes mencionado, K-medianas, y por último se utilizó el método jerárquico de Ward.

3.2.2.1. K-medias

Tal y como se mencionó anteriormente, el algoritmo K-medias es uno de los algoritmos más simples, comparados y más conocidos, que se aplica principalmente para resolver problemas de agrupamiento. Es un algoritmo iterativo donde se establece previamente K, el número de grupos o clusters deseado. Cada objeto pertenece a un solo

grupo. Intenta hacer que los objetos de un mismo grupo sean lo más similares posibles y al mismo tiempo mantiene los objetos inter-cluster los mas diferente (separado) posible. El algoritmo pretende que la suma de la distancia al cuadrado (distancia euclidiana) entre los objetos de un grupo y su centroide sea mínima. Entre menos variación haya dentro de los grupos, más homogéneos (similares) serán los puntos de datos dentro del mismo grupo. El proceso (flujograma en Figura 3.2) del algoritmo K-medias es el siguiente:

1. Especificar el número K de grupos.
2. Inicializar los centroides barajando primero el conjuntos de datos y luego seleccionando aleatoriamente K objetos para los centroides sin reemplazo.
3. Seguir iterando hasta que no haya cambios en los centroides, es decir, la asignación de los objetos a los grupos no está cambiando.
    - Calcular la suma de la distancia al cuadrado entre los objetos y todos los centroides.
    - Asignar cada objeto al grupo más cercano (centroide).
    - Calcular los centroides de los grupos tomando el promedio de todos los puntos de datos que pertenecen a cada grupo.

Aunque el agrupamiento de K-medias sigue siendo de los algoritmos más utilizados, hay ciertas limitaciones asociadas que incluyen (a) no existe método eficiente y universal para identificar las particiones iniciales y el número optimo de grupos K, y (b) K-medias es sensible a los valores atípicos y al ruido. Incluso si un objeto está bastante lejos del centroide del grupo, todavía es forzado a formar un grupo y, por lo tanto, distorsiona las formas del grupo [27].

3.2.2.2. K-medianas

El algoritmo K-medianas es una variante de K-medias, presenta el mismo procedimiento para asignar los objetos a K grupos. La diferencia se halla en la medida de similitud. K-medianas utiliza la distancia Manhattan. Además, para asignar el centro de un grupo respecto a los objetos del mismo, se reemplaza el cálculo de la media por el cálculo de la mediana, lo cual permite al algoritmo ser más robusto ante valores atípicos [28]. La figura 3.3 muestra el diagrama de flujo que se usó para desarrollar la función *kmedians* en este proyecto.

3.2.2.3. Jerárquico Aglomerativo - Método de Ward

El agrupamiento aglomerativo es el tipo más común de agrupamiento jerárquico utilizado para agrupar objetos en grupos en función de su similitud. El algoritmo empieza por tratar cada objeto como un grupo único. Luego, los pares de grupos se fusionan sucesivamente hasta que todos los grupos se fusionan en un un grupo grande que contiene todos los objetos. El resultado es una representación basada en árboles de los objetos, denominada dendrograma. La Figura 3.4 muestra un ejemplo de un dendrograma resultante de un proceso de agrupamiento jerárquico.

La mayoría de los algoritmos de agrupamiento jerárquico aglomerativos son variantes de los algoritmos de enlace único o de enlace completo. Estos dos algoritmos básicos

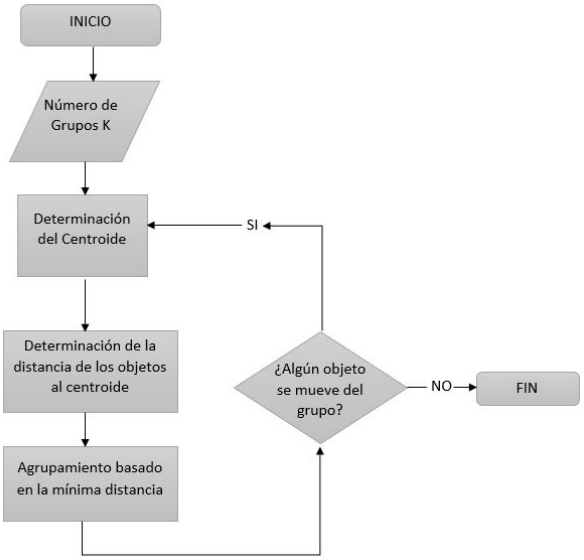


Figura 3.2 – Flujograma general de K-medias.
Fuente: Realizado por el autor.

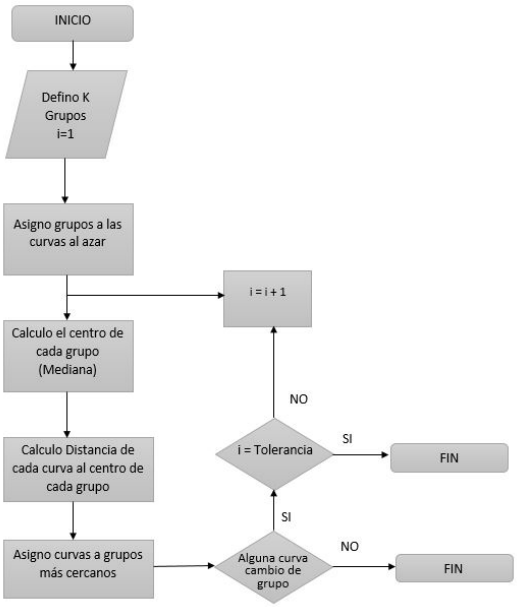


Figura 3.3 – Flujograma de la función *kmedians*.
Fuente: Tomado de [28]

difieren solo en la forma en que caracterizan la similitud entre un par de grupos. En el método de enlace único, la distancia entre dos conglomerados es el mínimo de las distancias entre todos los pares de muestras extraídas de los dos conglomerados (un elemento del primer conglomerado y el otro del segundo). En el algoritmo de enlace completo, la distancia entre dos grupos es el máximo de todas las distancias entre todos los pares extraídos de los dos grupos [23].

El método de Ward es otra variante de métodos de agrupamiento aglomerativo que se basa en un criterio clásico de suma de cuadrados, produciendo grupos que minimizan la dispersión dentro del grupo en cada fusión binaria. Es un enfoque que minimiza la varianza y, en este contexto, es similar a la función objetivo de K-medias, pero se plantea con un enfoque jerárquico aglomerativo. Además, el método de Ward es interesante porque busca grupos en el espacio euclidiano multivariado [29].

3.2.3. Evaluación de grupos

Uno de los más importantes desafíos que presenta el análisis de conglomerados es determinar el número óptimo de grupos que se ajuste al conjunto de datos. Puesto que los algoritmos de agrupamiento es un procedimiento no supervisado, es decir, el número de grupos se descubrirá de forma natural, la partición final de los datos precisa de algún tipo de evaluación. La validación del agrupamiento es el proceso para describir esta evaluación y los índices de validez de grupos son utilizados para la evaluación. Las investigaciones [31; 32] proponen varios índices del cual destacaremos el índice MIA para el fin de este proyecto. Además, utilizaremos el índice Silueta planteado en[33; 22; 34]. A continuación se expone cada uno de ellos.

- **Índice MIA**
  Para obtener este índice, considere que el proceso de clustering ha formado K

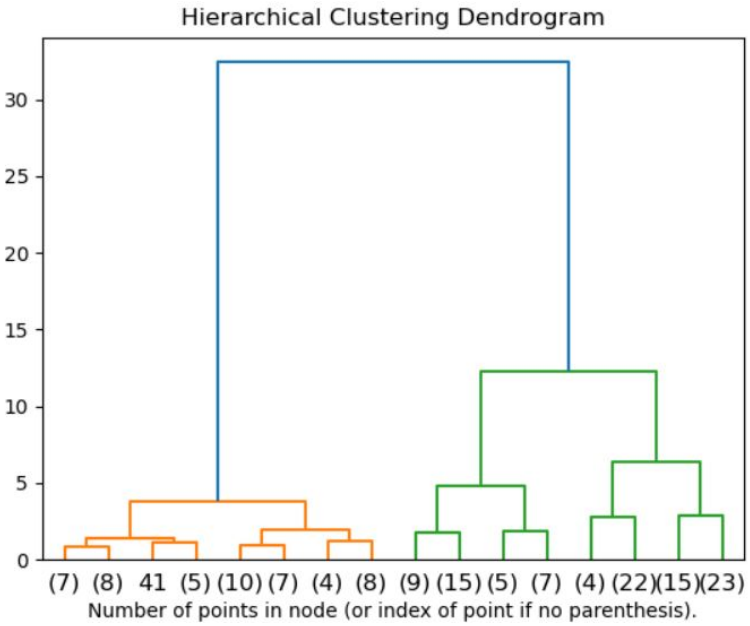


Figura 3.4 – Dendrograma de una agrupación jerárquica.
Fuente: Tomado de [30]

grupos de usuarios con $k = 1, ..., K$, y cada grupo está formado por un $L^(k)$ subconjunto de curvas de carga y $r^(k)$ es un patrón asignado al grupo k. MIA se obtiene del promedio de las distancias medias entre cada patrón asignado al grupo y su centro. Se define como:

$$MIA = \sqrt{\frac{1}{K}\sum_{k=1}^{K} d^2(r^{(k)}, C^{(k)})} \tag{3.5}$$

Este valor muestra que tan diferentes son los grupos entre sí. Mientras más pequeño sea, más eficaz es el algoritmo. En aplicaciones que se requiera grupos bien diferenciados, este índice es de gran utilidad para seleccionar el algoritmo con mejores resultados.

- **Índice de Silueta**
  El análisis de silueta sirve para determinar la cantidad de grupos dentro de un conjunto de datos. Suponga que los datos se han agrupado en k grupos y para cada objeto i, $a(i)$ es la disimilitud promedio de i con otros objetos dentro del mismo grupo. También $b(i)$ es la menor disimilitud promedio de i con cualquier otro grupo. El valor de silueta, $s(i)$ se define como:

$$s(i) = \frac{b(i) - a(i)}{max\{a(i), b(i)\}} \tag{3.6}$$

  El promedio de $s(i)$ sobre todas los objetos dentro de un grupo muestra qué tan cerca están los objetos en el grupo, y el promedio de todo el conjunto de datos muestra cuán correctamente están agrupados los datos.

  Los coeficientes de silueta tienen un rango de $[-1, 1]$, los valores cercanos 1 indica que la muestra está lejos de los grupos vecinos. Un valor de 0 indica que la

muestra está muy cerca del límite de decisión entre dos grupos vecinos y los valores negativos indican que esas muestras podrían haber sido asignado al grupo incorrecto.

Este índice sirve para tener información a cerca de la homogeneidad de los grupos formados.

Por lo dicho antes, cabe recalcar que la validación es una de las tareas principales a realizar en todo proceso de agrupamiento, y determinar el número óptimo de K de grupos es la cuestión mas relevante. Por lo tanto, aplicar estos dos indicadores nos permitirá reconocer que tan diferentes y compactos serán los grupos entre sí, para luego seleccionar un algoritmo que mejores resultados presente. Para ambos índices se puede utilizar un gráfico que contenga los valores de MIA y de Silueta para una serie sucesiva de números de grupos $K = [1, .., N]$, el resultado será una curva de amalgamiento; para el caso de MIA es conocido como curva del codo utilizado con frecuencia para establecer el número ideal de grupos en K-medias, y para el caso del Silueta se conoce como gráfico de siluetas. Ambas gráficas nos ayudará a determinar el número óptimo de grupos mediante la apreciación visual, que consiste en observar el punto donde la curva comience a variar de forma menos significativa. La Figura 3.5 representa la curva característica del índice MIA que se obtuvo en este estudio para la validación del algoritmo K-medianas aplicado al conjunto de datos de la empresa de servicio eléctrico de New Yersey [35]. La Figura 3.6 muestra el gráfico de silueta para el mismo algoritmo y conjunto de datos.

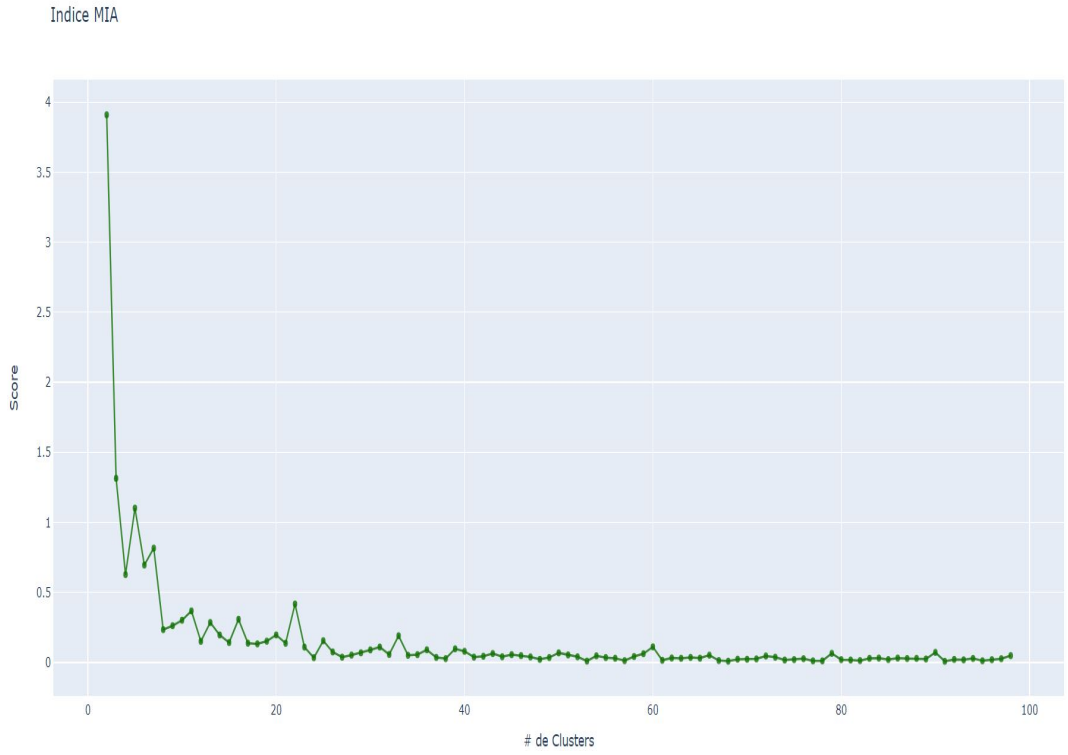


Figura 3.5 – Curva del codo representativa del índice MIA para $K = [2, 98]$ grupos. Fuente: Realizado por el autor.

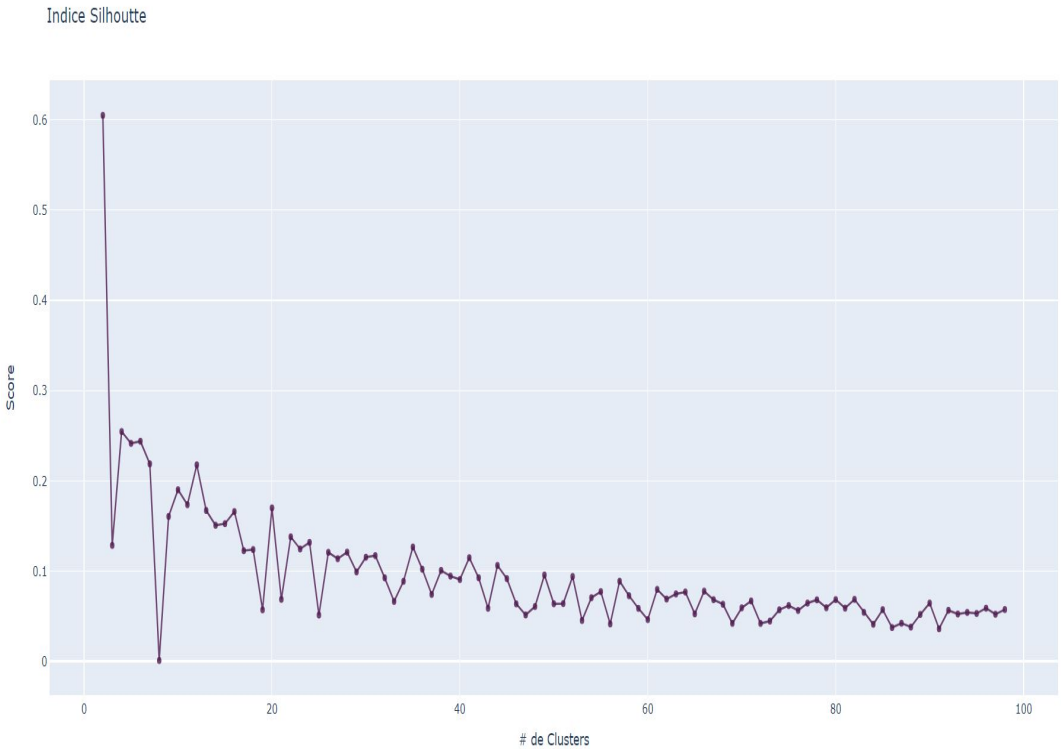


Figura 3.6 – Curva representativa del índice Silueta para $K = [2, 98]$ grupos.
Fuente: Realizado por el autor.

Capítulo 4

GENERACIÓN SINTÉTICA DE PATRONES DE CONSUMO DE ENERGÍA Y MODELO DE PÉRDIDAS NO TÉCNICAS

Los datos son la materia prima en todo proceso de análisis de datos. Sin embargo, se presenta grandes inconvenientes cuando no se cuenta con los suficientes datos para revelar información significativa, de manera que resulta poco efectivo aplicar técnicas o algoritmos de minería de datos. Los datos sintéticos o artificiales son una representación de eventos reales generados de forma algorítmica a partir de información histórica y tienen como propósito servir de soporte para pruebas de modelos estadísticos, matemáticos , y progresivamente, para entrenar modelos de aprendizaje automático.

En el sector eléctrico, los datos de consumo de energía cuentan con una política de privacidad y confidencialidad por parte de la empresa distribuidora del servicio y de los usuarios, que dificulta el libre acceso a estos datos. Por consiguiente, en este proyecto se han desarrollado modelos probabilísticos generativos basados en cadenas de Markov, llamados también modelos ocultos de Markov (HMM), para generar datos sintéticos. Por otro lado, la ausencia de medidores inteligentes y el hecho de que las pérdidas no técnicas son difíciles de detectar ha ocasionado una carencia de muestras de esta naturaleza. Ante la falta de muestras resulta imprescindible simular las pérdidas mediante un modelo de alteración de los datos "benignos". En este capítulo se explica en detalle la ejecución de estos modelos en el proyecto. Primero se entra en contexto con los conceptos relacionados con el generador de datos, los modelos ocultos de Markov y su aplicación mediante la librería *hmmlearn* de Python. Por último, se detalla la estructura conceptual a seguir para modelar las pérdidas. El desarrollo de estos modelos son de suma importancia ya que permitirán el desenvolvimiento eficiente de los algoritmos propuestos más adelante.

4.1. Procesos de Markov

4.1.1. Cadenas de Markov

Una cadena de Markov es un proceso estocástico y una extensión del autómata finito determinista donde las transiciones representan las probabilidades de ir de un estado a otro y dado el estado actual del proceso, los estados futuros son independientes de los estados pasados. En otras palabras, no se afectan las probabilidades de cambiar de un estado actual al siguiente sino sabemos como se llegó a este estado, es suficiente con tener la información del estado presente del proceso. Formalmente, una cadena de Markov se define como [36]:

Una cadena de Markov de N estados es un triplete (S, A, π), donde

- Conjunto de N estados
$$S = \{S_1, S_2, ..., S_N\} \tag{4.1}$$
- Distribución de probabilidad estacionaria sobre los estados S
$$\pi = \{\pi_1, \pi_2, ..., \pi_N\} \tag{4.2}$$
El cual cumple con la siguiente propiedad
$$\sum_{i=1}^{N} \pi_i = 1 \tag{4.3}$$
- Distribución de probabilidad de transición de estado
$$A = \{a_{ij}\}, \quad a_{ij} = P[q_t = S_j | q_{t-1} = S_i], \quad 1 \leq i, j \leq N \tag{4.4}$$
Donde los coeficientes de la matriz de transición tienen las propiedades de
$$a_{ij} \geq 0 \tag{4.5}$$
$$\sum_{j=1}^{N} a_{ij} = 1 \tag{4.6}$$
Además,
$$\pi A = \pi \tag{4.7}$$

Entonces, se cumple con la propiedad de que la probabilidad de estar en un estado i depende solamente del estado anterior $i - 1$. A esta propiedad se la conoce como propiedad markoviana o de Markov [37]:

$$P[q_t = S_j | q_{t-1} = S_i, q_{t-2} = S_k, ...] = P[q_t = S_j | q_{t-1}] \tag{4.8}$$

Este proceso estocástico se lo denomina modelo observable de Markov, puesto que la salida del proceso es el conjunto de estados en cada instante de tiempo, en el que cada estado corresponde a un evento observable. Además, resulta útil para calcular la probabilidad de una secuencia de eventos observables.

4.1.2. Modelos ocultos de Markov (HMM)

Los modelos ocultos de Markov son una extensión del proceso antes mencionado, que incluye el caso en el cual las observaciones se consideran causales del modelo probabilístico, dicho de otra manera, este modelo intenta establecer la relación entre el proceso estocástico donde los eventos son no observables (estados ocultos), pero que se pueden observar mediante otro proceso que producen la secuencia de observaciones [36]. La definición completa de un HMM necesita de los siguientes cinco elementos:

1. Conjunto de N estados del modelo $S = \{S_1, S_2, ..., S_N\}$

2. Conjunto de M símbolos observables

$$V = \{v_1, v_2, ..., v_M\} \tag{4.9}$$

3. Distribución de probabilidad de transición de estado $A = \{a_{ij}\}$, donde

$$a_{ij} = P[q_{t+1} = S_j | q_t = S_i], \quad 1 \leq i, j \leq N \tag{4.10}$$

   Siendo q_t el estado actual. Esta matriz es equivalente a lo definido en cadenas de Markov.

4. Distribución de probabilidad de observación de símbolo en el estado j, $B = \{b_j(k)\}$, donde

$$b_j(k) = P[o_t = v_k | q_t = S_j], \quad 1 \leq j \leq N, \quad 1 \leq k \leq M \tag{4.11}$$

   Se denota v_k como el k-ésimo símbolo del alfabeto, y o_t el actual vector de observaciones.

5. Distribución de probabilidad del estado inicial $\pi = \{\pi_i\}$, donde

$$\pi_i = P[q_1 = S_i], \quad 1 \leq i \leq N \tag{4.12}$$

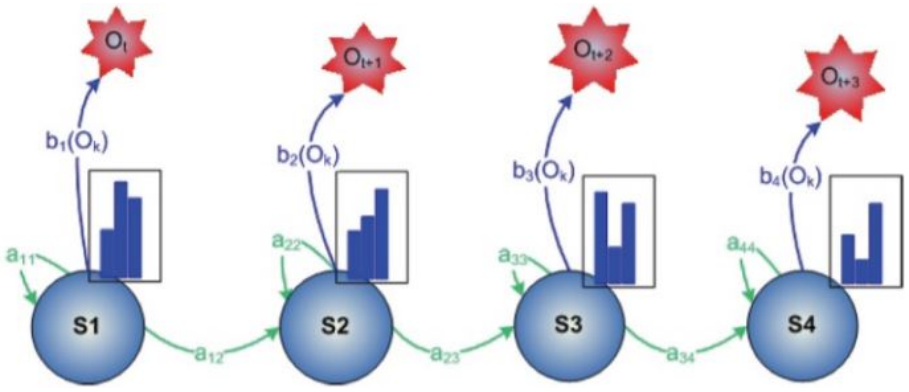


Figura 4.1 – Ejemplo de un modelo oculto de Markov.
Fuente: Tomado de [38]

El modelo de Markov se denota con frecuencia como:

$$\lambda = (A, B, \pi) \tag{4.13}$$

Teniendo los valores del modelo λ se puede generar una observación de secuencia

$$O = \{O_1, O_2, ..., O_T\} \tag{4.14}$$

(donde cada observación $O_t$es uno de los símbolos de V, y T es el número de observaciones en la secuencia) de la siguiente manera [36]:

1. Elegir un estado inicial $q_1 = S_i$ según el vector de probabilidad de estado inicial π.
2. $t = 1$.
3. Elegir una observación $O_t = v_k$ según el vector de probabilidad de observación del símbolo en el estado S_i. Por ejemplo, $b_i(k)$.
4. Cambio a un nuevo estado $q_{t+1} = S_j$ según la matriz de probabilidad de transición de estado S_i. Por ejemplo, a_{ij}.
5. Ajustar $t = t + 1$. Si t <T regresar al paso 3. Caso contrario el procedimiento finaliza.

4.1.3. Problemas básicos de HMM y algoritmos de solución

Existen tres problemas fundamentales que se presentan en un modelo HMM al tratar de implementarlo en aplicaciones reales, a continuación se mencionan estos problemas y su algoritmo de solución:

Problema 1: Evaluación. Dado el modelo $\lambda = (A, B, \pi)$ y una secuencia de observaciones $O = \{O_1, O_2, ..., O_T\}$, calcular cuál es la probabilidad $P(O|\lambda)$ de que O haya sido generado por el modelo λ.
Algoritmo de solución: Para resolver este problema se aplica el algoritmo Backward-Forward.

Problema 2: Estimación. Dado el modelo λ y una secuencia de observaciones O, estimar cuál es la secuencia de estados ocultos más probable. Es decir, se trata de descubrir la parte oculta del modelo.
Algoritmo de solución: Este problema se puede resolver computacionalmente con el algoritmo de Viterbi.

Problema 3: Aprendizaje. Dada la se secuencia de observaciones O, determinar cual debe ser el mejor ajuste de los parámetro del modelo λ para maximizar $P(O|\lambda)$.
Algoritmo de solución: Esto tiene solución aplicando el algoritmo Baum-Welch, o llamado como algoritmo de Espezanza-Maximización (EM).

Para más detalle acerca del desarrollo de los algoritmos que solventan estos problemas, puede referirse a [36; 39].

4.1.4. Modelos Gaussianos ocultos de Markov (GHMM)

Las observaciones de un modelo continuo son una función de densidad continua, en vez de un conjunto de probabilidades discretas. Y lo que se especifica es el conjunto de parámetros de la función de densidad, aproximada como una sumatoria de M distribuciones Gaussianas (GHHM) [38],

$$b_j(o_t) = \sum_{m=1}^{M} c_{jm} N(\mu_{jm}, \sum_{jm}, o_t) \tag{4.15}$$

Donde los parámetros c_{jm}, μ_{jm}, $\sum_{jm}$ son los pesos, los vectores de medias, y las matrices de convarianza, respectivamente. Se observa que el coeficiente c_{jm} debe satisfacer las restricciones estocásticas $c_{jm} \geq 0, \quad 1 \leq m \leq M$ y

$$\sum_{m=1}^{M} c_{jm} = 1, \quad 1 \leq j \leq N \tag{4.16}$$

Para el modelo GHMM las observaciones presentan una distribución multivariante Gaussiana de dimensión L, diferente para cada estado. Es decir, si tenemos K estados diferentes, cada uno de ellos se relaciona a una distribución de parámetros $(\mu_i, \sum_i)$, $1 \leq i \leq K$; donde μ_i es el vector de medias y $\sum_i$ la matriz de covarianzas para el estado i.

Como se trata de un modelo continuo hay una infinidad de símbolos existentes ($V = \Re$). La probabilidad de la observación o_t dado un estado i se denota como $b_j(o_t) = P(o_t|q_t = i), \quad 1 \leq j \leq N, \quad 1 \leq t \leq M$.

Se debe tener en cuenta que la función de densidad respecto al estado i es:

$$N(x, \mu_i, \sum_i) = \frac{1}{(2\pi)^{L/2} |\sum_i|^{1/2}} e^{-0.5(x-\mu_i)\sum_i^{-1}(x-\mu_i)^t}, \tag{4.17}$$

de modo que,

$$b_j(o(t)) = \sum_{k=1}^{K} c_{jk} N(x, \mu_j, \sum_j) \tag{4.18}$$

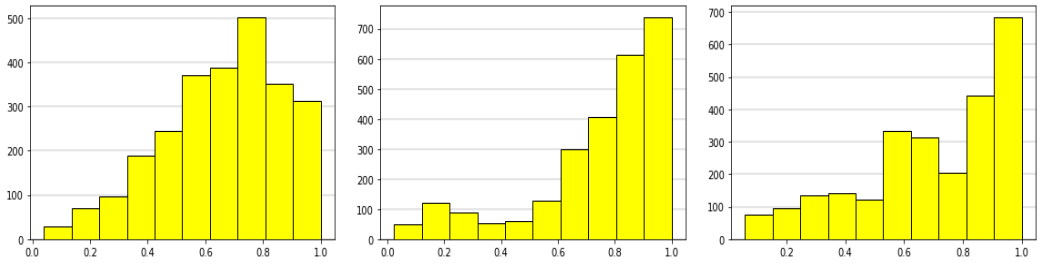
4.2. Generación de patrones de consumo de energía utilizando la librería *hmmlearn*

Para la generación artificial de los patrones de consumos de energía se utilizó la librería *hmmlearn* de Python, la cual tiene disponible los modelos HMM. En particular, para este proyecto, se implementó el modelo oculto de Markov con emisiones Gaussianas *hmmm.GaussianHMM*, dado que los datos de mediciones de energía son continuos. Cabe recalcar, que los algoritmos de Viterbi y EM vienen preestablecido en el modelo.

4.2.1. Selección y preparación de la muestra

El modelo generativo que se ha diseñado replica el comportamiento de un conjunto de datos, es decir, a partir de estos datos se generarán datos análogos. Así pues, los datos

requeridos se han obtenido de la data pública y de libre acceso de la empresa eléctrica PSEG con frecuencia horaria. Los datos representan el perfil de consumo diario de los usuarios durante un año; en varios sectores, sean residencial, comercial e industrial. Sin embargo, se recopiló la muestra como un conjunto completo de estudio para emular una mayor densidad poblacional. En todo caso, la secuencia de observaciones está dado por la medición de potencia de los datos, en cambio, se determina los estados ocultos como los posibles niveles de potencia a alcanzar por cada elemento del modelo en cierto intervalo de tiempo.



(a) 8am **(b)** 17pm **(c)** 22pm

Figura 4.2 – Histograma del dataset PSEG a tres horas distintas del día. Eje x: Kilowatts por unidad. Eje y: Número de usuarios
Fuente: Realizado por el autor.

4.2.2. Desarrollo del modelo

El primer paso para desarrollar este algoritmo es definir el modelo como un objeto, el cual se le debe ingresar los parámetros de entrada para que funcione correctamente. Empezamos con determinar el número de componentes (estados), por lo que se utiliza el mismo procedimiento realizado en [28], el cual consiste en escoger el número de estados basado en el mismo procedimiento que sirve para establecer el número óptimo de grupos mencionado en el capítulo anterior. Por lo tanto, se utiliza la curva del codo del índice MIA resultante para el algoritmo K-medias aplicado al conjunto de datos, que agrupa los niveles de potencia repetidos con mayor recurrencia; tal como se muestra en la Figura 4.3. De esta manera se estableció el número de estados para el modelo, el cual se seleccionó 70 posibles estados o niveles de potencia para la secuencia dada.

Lo siguiente a la estimación de los estados es fijar los demás parámetros de entrada para establecer el modelo. La información a cerca de estos parámetros se encuentra en la librería *hmmlearn* y la mayoría vienen preestablecidos o por "default". La Figura 4.4 muestra la estructura del modelo.

Una vez establecido el modelo se pasa a la fase de entrenamiento mediante la función *.fit()* cuyo argumento de entrada es el conjunto de datos. Mediante esta función se logra la estimación de los parámetros $\lambda(A, B, \pi)$ de los datos observados. Por definición se emplea el algoritmo Esperanza-Maximización (EM).

Cuando el modelo está entrenado se obtiene la mejor representación de la secuencia de observaciones dada a través del algoritmo de Viterbi predefinido en el modelo. Así, se puede obtener a la vez la matriz de transición de estados vista en la Figura 4.5 .

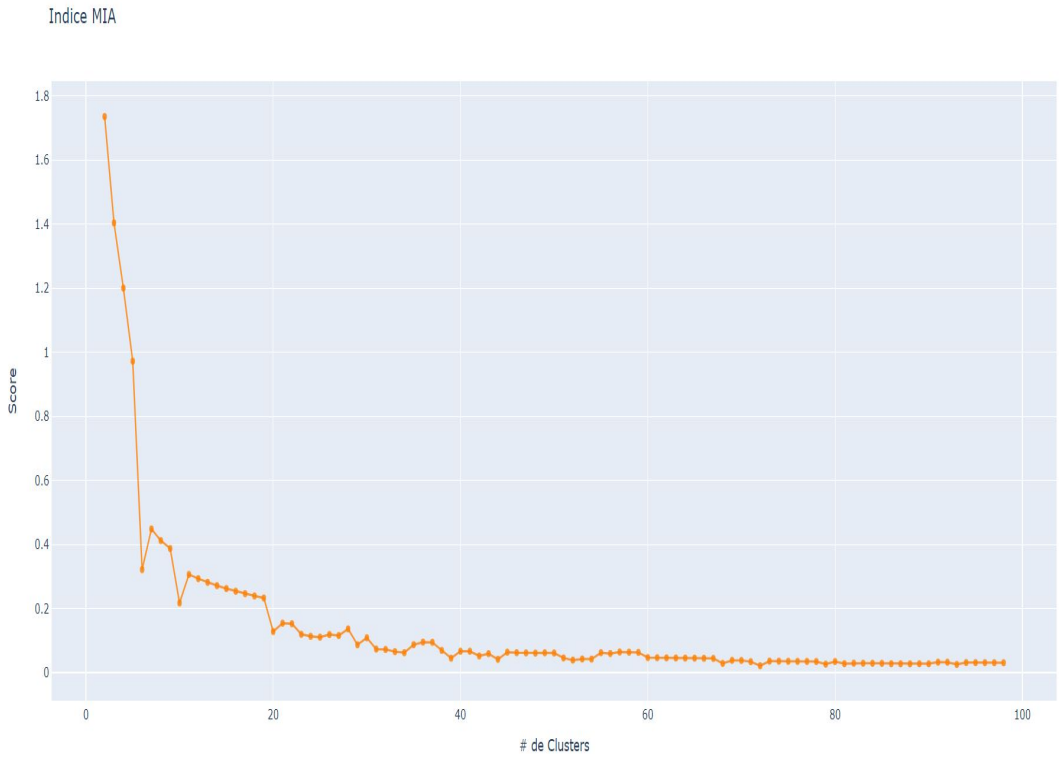


Figura 4.3 – Curva del codo respectiva a la secuencia de observaciones.
Fuente: Realizado por el autor.

```
model = hmm.GaussianHMM(n_components=70, covariance_type='diag',
                        min_covar=0.001,startprob_prior=1.0, transmat_prior=1.0,
                        means_prior=0, means_weight=0, covars_prior=0.01,
                        covars_weight=1, algorithm='viterbi', random_state=None,
                        n_iter=10, tol=0.01, verbose=False, params='stmc',
                        init_params='stmc')
```

Figura 4.4 – Modelo Gaussiano de Markov para generar data sintatica.
Fuente: Realizado por el autor.

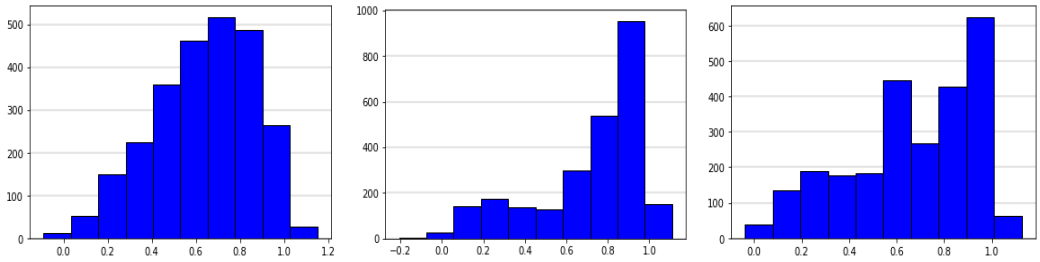
```
In [9]: matrixTrans
Out[9]:
array([[3.57266123e-001, 0.00000000e+000, 0.00000000e+000, ...,
        0.00000000e+000, 1.95017578e-202, 0.00000000e+000],
       [0.00000000e+000, 1.03931413e-001, 0.00000000e+000, ...,
        2.08952303e-104, 0.00000000e+000, 2.01913278e-001],
       [0.00000000e+000, 0.00000000e+000, 6.68122043e-001, ...,
        0.00000000e+000, 5.24025874e-003, 0.00000000e+000],
       ...,
       [0.00000000e+000, 4.29157493e-083, 0.00000000e+000, ...,
        1.01515608e-060, 0.00000000e+000, 1.48573163e-032],
       [9.27558674e-232, 0.00000000e+000, 4.47004330e-036, ...,
        0.00000000e+000, 4.02166738e-001, 0.00000000e+000],
       [0.00000000e+000, 7.52553013e-002, 0.00000000e+000, ...,
        3.71538003e-012, 0.00000000e+000, 8.34195837e-002]])
```

Figura 4.5 – Parte de la matriz de transición de estados obtenida.
Fuente: Realizado por el autor.

4.2.3. Generación artificial

Por último, ya definido y entrenado el modelo, para generar una muestra aleatoria de patrones de consumo de energía basta con utilizar el atributo *sample()*, el cual se le indica la cantidad de curvas a generar. Para el caso, se le atribuye a los intervalos de tiempo en que se caracteriza la medición.

A modo de comparación, se observa cierta similitud entre los histogramas del conjunto de datos original (Figura 4.2) y de los datos generados con el modelo Gaussiano de Markov (Figura 4.6) correspondiente a tres horas distintas del día. Los valores de las mediciones de potencia de la data original se encuentran normalizados, pero en el caso de la data artificial tenemos que se han generado pequeños valores fuera de los rangos de normalización $[0, 1]$; para algunos casos hay valores negativos de potencia. De todas formas, esto significa que el modelo generó valores cercanos a 0 y valores cercanos a 1 y al tratarse de un modelo probabilístico, existe esta posibilidad. Posteriormente, para efectos de este estudio los valores negativos los convertimos en positivos mediante una función de valor absoluto.



(a) 8am **(b)** 17pm **(c)** 22pm

Figura 4.6 – Histograma de la muestra aleatoria generada con *hmmlearn*. Eje x: Kilowatts por unidad. Eje y: Número de usuarios
Fuente: Realizado por el autor.

4.3. Modelo de Pérdidas No Técnicas

El modelo de pérdidas no técnicas aplicado simula los consumos del usuario final que no le es parcial o totalmente medido. Es decir, se representa a usuarios fraudulentos que de alguna manera alteran o evitan los registros de los equipos de medición. Los usuarios fraudulentos tienden a reducir la cantidad de electricidad facturada.

El método para representar este tipo de pérdidas es a través de la manipulación de los datos de consumos registrados por usuario. Por lo tanto, el modelo se define por muestras "maliciosas" generadas a partir de la modificación de una muestra "benigna" aleatoria del conjunto de datos, para el caso, el conjunto de datos PSEG con datos recolectados desde medidores inteligentes. Este método se sustenta en [34; 40; 41]. En síntesis, los autores desarrollan seis tipos generales de pérdidas que, en pocas palabras, son seis diferentes maneras de alterar el patrón de consumo de energía de los usuarios. Si tenemos que el registro diario de consumo eléctrico de cada usuario está dado por, $x = \{x_1, ..., x_{24}\}$, por motivos de simplicidad en este proyecto se han adoptado los siguientes cuatro tipo de pérdidas para $t = 1, ..., 24$ [34]:

- **Tipo 1:** $p_1(x_t) = \alpha x_t, \quad \alpha = rand(0.2, 0.8) \quad \forall t$
- **Tipo 2:** $p_2(x_t) = \beta x_t$

$$\beta = \begin{cases} rand(0.2, 0.8), & \text{si } t_{inicial} < t < t_{final} \\ 1, & \text{sino} \end{cases}$$

$t_{inicial} = rand(0, 23)$
$\Delta_t = rand(1, 24 - t_{inicial})$
$t_{final} = t_{inicial} + \Delta_t$

- **Tipo 3:** $p_3(x_t) = \gamma_t x_t$

$$\gamma_t = \begin{cases} 0, & \text{si } t_{inicial} < t < t_{final} \\ 1, & \text{sino} \end{cases}$$

- **Tipo 4:** $p_4(x_t) = \theta_t x_t, \quad \theta_t = rand(0.2, 0.8) \quad \forall t$

Dicho lo anterior, $p_1()$ es un factor multiplicativo aleatorio que reduce el consumo del consumo de manera constante para todo el tiempo. Con $p_2()$ también se tiene un multiplicador, en cambio, se disminuye el consumo continuamente durante un intervalo de tiempo definido de forma aleatoria. Usando $p_3()$ se emula el caso en que el usuario evita el medidor inteligente y éste registra un consumo igual a cero, igualmente para un periodo aleatorio. Por último, $p_4()$ multiplica cada lectura del medidor por un valor aleatorio diferente.

Para desarrollar el modelo, primero se selecciona una muestra de usuarios del conjunto de datos donde suponemos que en su totalidad están compuestos de usuarios benignos. Para este proyecto se escogió al 10 % de la población. Después, se aplica aleatoriamente algún tipo de pérdidas de las mencionadas a cada uno de los usuarios, que los convierte en usuarios con pérdidas (fraudulentos), de mayor o menor intensidad depende de la clase de pérdidas designada. Finalmente, se reemplaza la muestra modificada en el conjunto original de datos. La Figura 4.7 muestra el gráfico de la cantidad de usuarios asignado a cierto tipo de pérdidas. De igual manera, las Figuras 4.8 y 4.9 muestran el patrón de consumo de un usuario normal antes de aplicar las pérdidas y luego de tener los cuatros tipos de pérdidas, respectivamente.

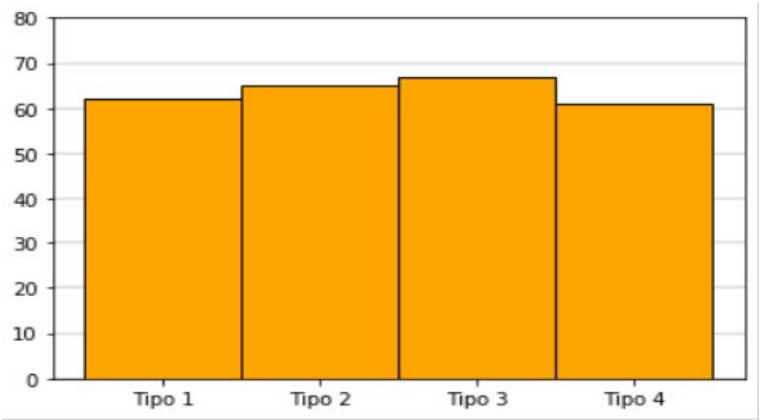


Figura 4.7 – Gráfico de barras de usuarios con algún tipo de pérdidas.
Fuente: Realizado por el autor.

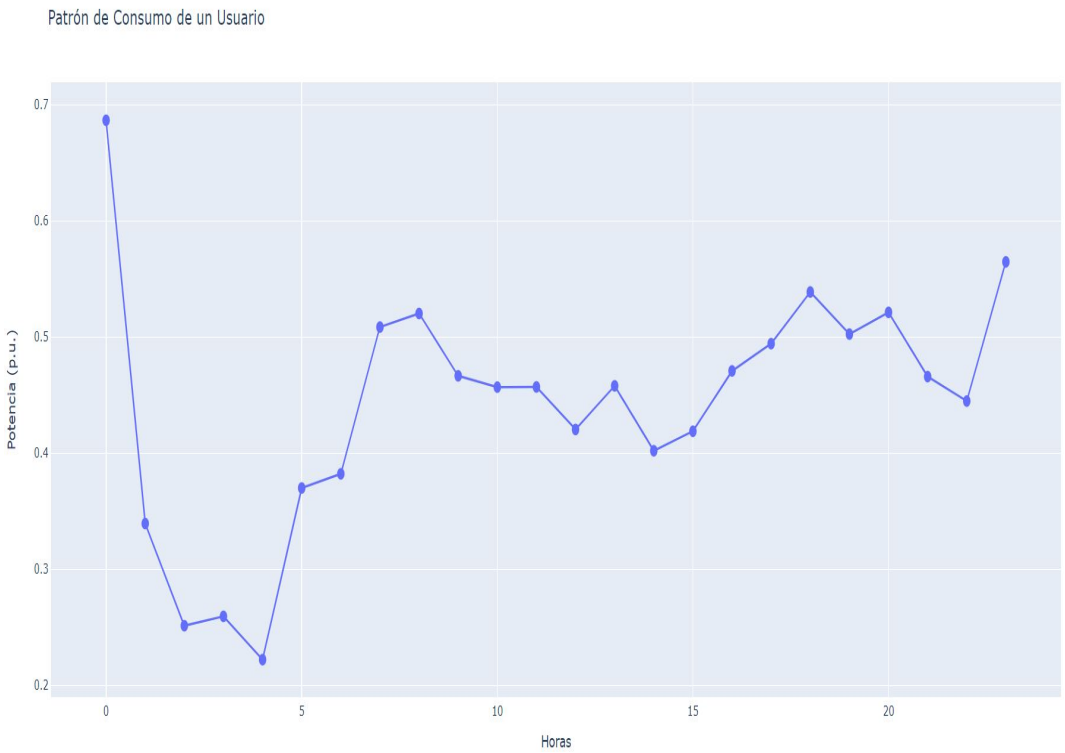


Figura 4.8 – Patrón de consumo eléctrico de un usuario sin pérdidas.
Fuente: Realizado por el autor.

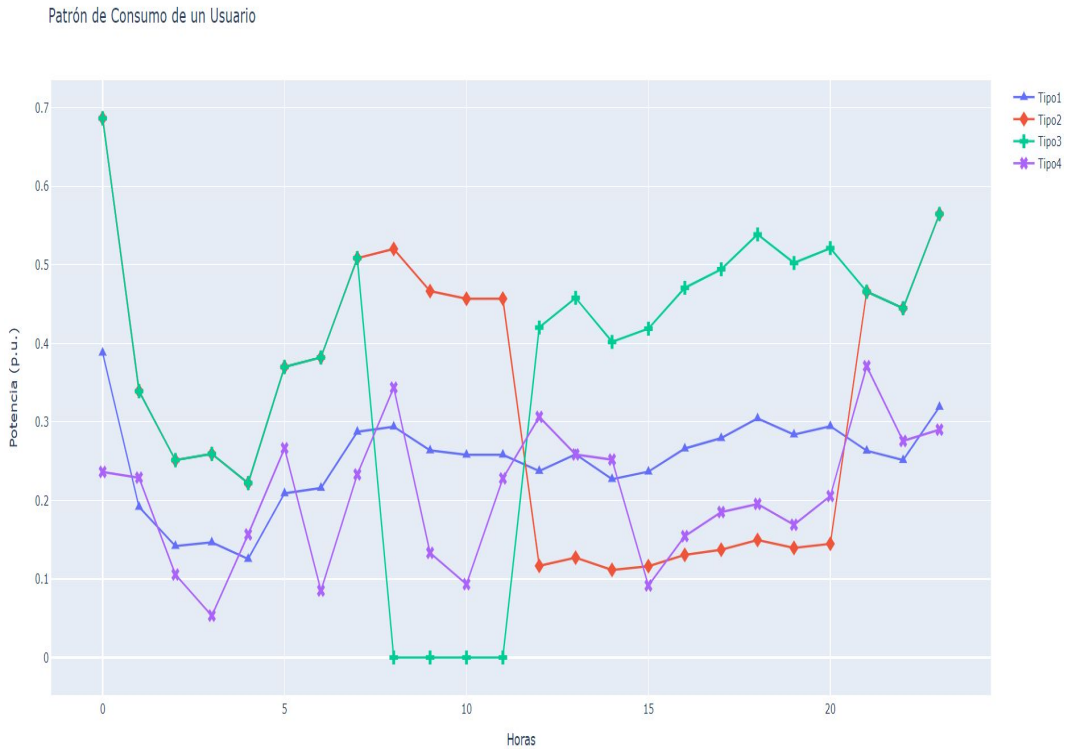


Figura 4.9 – Patrones de consumo eléctrico de un usuario con varios tipos de pérdidas.
Fuente: Realizado por el autor.

Capítulo 5

DETECCIÓN DE PÉRDIDAS NO TÉCNICAS UTILIZANDO MÉTODOS DE AGRUPAMIENTO

El método más factible de identificar a los consumidores fraudulento es por medio del análisis de los datos o los parámetros del sistema eléctrico que representan una situación en el confín del usuario. El consumo instantáneo de energía es el parámetro más significativo que mejor representa las irregularidades o anomalías en en el consumo de energía del usuario. Además, analizar los patrones de consumo de energía es uno de los mecanismos esenciales que llevan a cabo las empresas eléctricas para comprender claramente el estado de la red en el extremo de los clientes.

En este estudio se presenta un método de agrupamiento con el fin de obtener la mayor cantidad de usuarios fraudulentos agrupados. Este trabajo se desarrolló en el lenguaje de programación de software libre *Python 3.7*. Las librerías y herramientas de visualización que se utilizaron son las siguientes:

- pandas
- numpy
- hmmlearn
- sklearn.cluster
- sklearn.metrics
- sklearn.model_selection
- plotly.graph_objs: Scatter, Pie, Box, Figure
- plotly.offline:plot
- matplotlib.pyplot

El procedimiento para descubrir patrones de usuarios con posibles pérdidas no técnicas se puede dividir en 2 fases: fase de entrenamiento y fase de prueba.

El algoritmo desarrollado empieza con la normalización y posterior aplicación del modelo de pérdidas no técnicas sobre los datos generados en el capítulo anterior para así pasar a la fase de entrenamiento. En esta etapa el conjunto de datos se fracciona en un 50 % para la evaluación de los tres diferentes métodos de agrupamiento mencionados en el Capítulo 2. El algoritmo seleccionado es aquel que agrupe con mayor concentración los consumidores irregulares. Así pues, se identifica las características distintivas de este grupo respecto a los demás para formar criterios a partir de ellas. Por último, en la fase de prueba se examinan los criterios propuestos sobre la data restante y se validan para verificar la detección de patrones de consumos con pérdidas. El algoritmo planteado se presenta en la Figura 5.2.

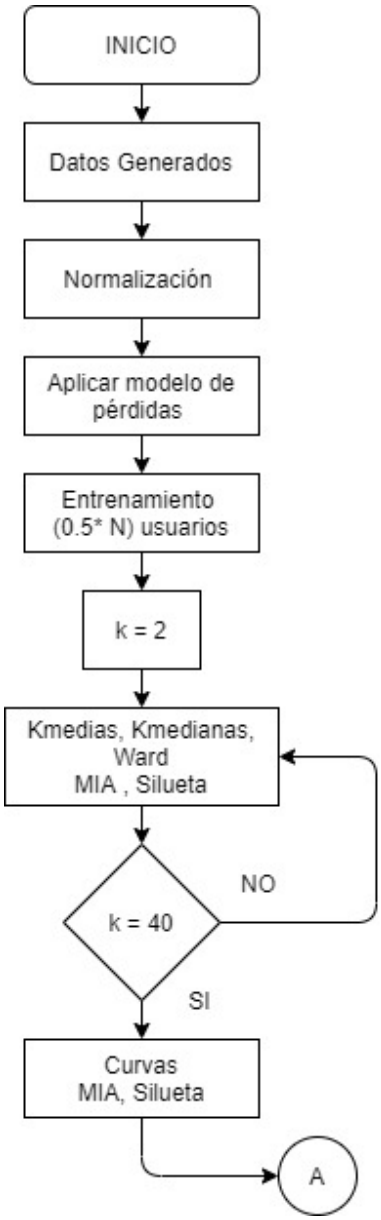


Figura 5.1 – (a) Flujograma de detección de pérdidas no técnicas
Fuente: Realizado por el autor.

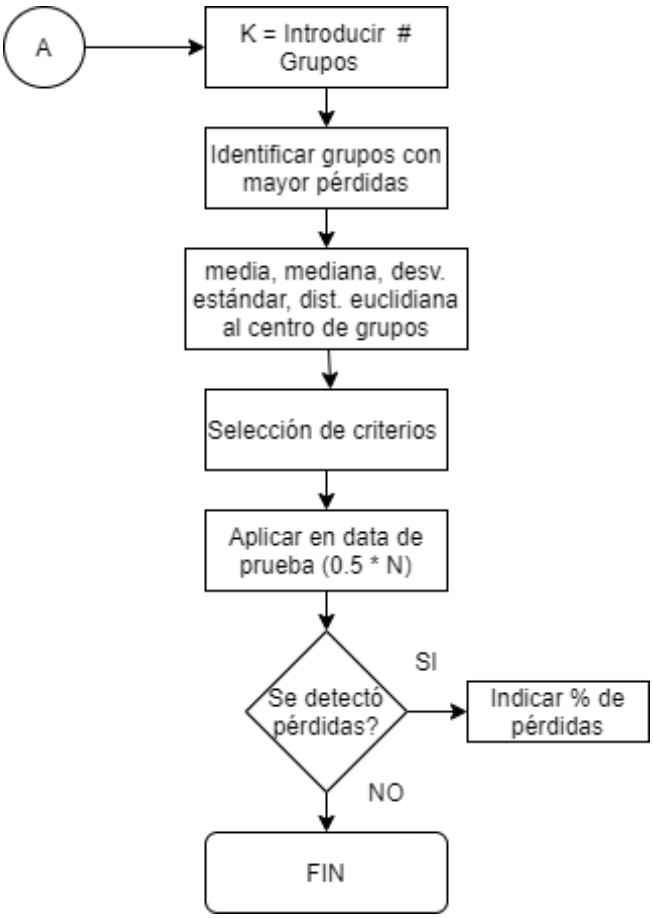


Figura 5.2 – b) Flujograma de detección de pérdidas no técnicas
Fuente: Realizado por el autor.

Antes se mencionó que un paso necesario para procesar el algoritmo es la normalización de los datos. Esto se establece para tener los valores de potencia medida en un rango entre 0 y 1, siendo la potencia máxima de cada usuario el valor de normalización. Además, permite el análisis de los patrones, de lo contrario, si se utiliza las medidas de similitud; estas medirían la diferencia de energía entre los usuarios en lugar de la distancia entre ellos, que los relaciona en función de la energía que consumen sin tomar en cuenta el perfil que se relaciona a la labor que ejercen. El siguiente paso fundamental a realizar es la agregación de los tipos de pérdidas a una muestra del conjunto de datos, que en porcentaje representa el 10 % de la población. La finalidad del algoritmo será detectar los usuarios con pérdidas antes aplicada, a modo de evaluación, en el gran conjunto de datos.

5.1. Entrenamiento

La fase de entrenamiento inicia con el proceso iterativo de hallar los valores de índice MIA y Silueta para cada $k = [2, 40]$ y para cada distinto método de agrupamiento: K-medias, K-medianas y Ward; los resultados se muestran en las tablas comparativas 5.1 y 5.2. Al final de las iteraciones se visualizan las curvas de los indicadores como en las Figuras 5.3 y 5.4.

En líneas generales, de manera gráfica en las curvas representativas de los indicadores se puede observar que a partir de la aglomeración de 8 grupos las variaciones

son menos significativas (codo), además hay mucha similitud entre la eficiencia de los métodos de agrupamiento. No obstante, K-medianas es el algoritmo que a simple vista presenta mayor variación respecto a los demás, tanto en el gráfico MIA y Silueta.

El caso importante a analizar es en el que un algoritmo aglomere o agrupe de manera homogénea y compacta los usuarios con pérdidas en interés, mas no que distorsione las aglomeraciones tomando a a las pérdidas como valores atípicos. Debido a esto, la K-medianas resalta más cambios en diferentes valores de k pues como ya se ha dicho, K-medianas es un algoritmo robusto ante estos valores. Por otro lado, K-medias y Jerárquico Ward por su propiedad aglomerativa realiza grupos bien diferenciados entre sí, y al encontrar un usuario con valores atípicos lo asocie un único estrato que es precisamente lo deseado, desagregar grupos de usuarios normales a grupos específicos de usuarios con pérdidas.

Tabla 5.1 – Valores de índice MIA, de 8 a 20 grupos

	MIA		
K Grupos	**K-medias**	**K-medianas**	**Ward**
8	0.14456259	0.215631128	0.34034709
9	0.087434	0.279469735	0.26167691
10	0.08802213	0.149528993	0.21038816
11	0.12507975	0.043768182	0.17178601
12	0.0948285	0.091456542	0.12172772
13	0.09110828	0.245991773	0.1427455
14	0.08756286	0.117893078	0.17301505
15	0.07828465	0.022935545	0.143423
20	0.08314391	0.047970593	0.08246893

Tabla 5.2 – Valores de índice Silueta, de 8 a 20 grupos

	Silueta		
K Grupos	**K-medias**	**K-medianas**	**Ward**
8	0.26954412	0.089875704	0.26795862
9	0.2572513	0.074421387	0.27505189
10	0.22356781	0.074783922	0.27906376
11	0.22183078	0.094571968	0.28305261
12	0.22806884	0.083355512	0.27391376
13	0.23063788	0.022550366	0.27343636
14	0.2301308	0.108385309	0.19942923
15	0.20298579	0.090345978	0.19249233
20	0.18884809	0.037859353	0.1623031

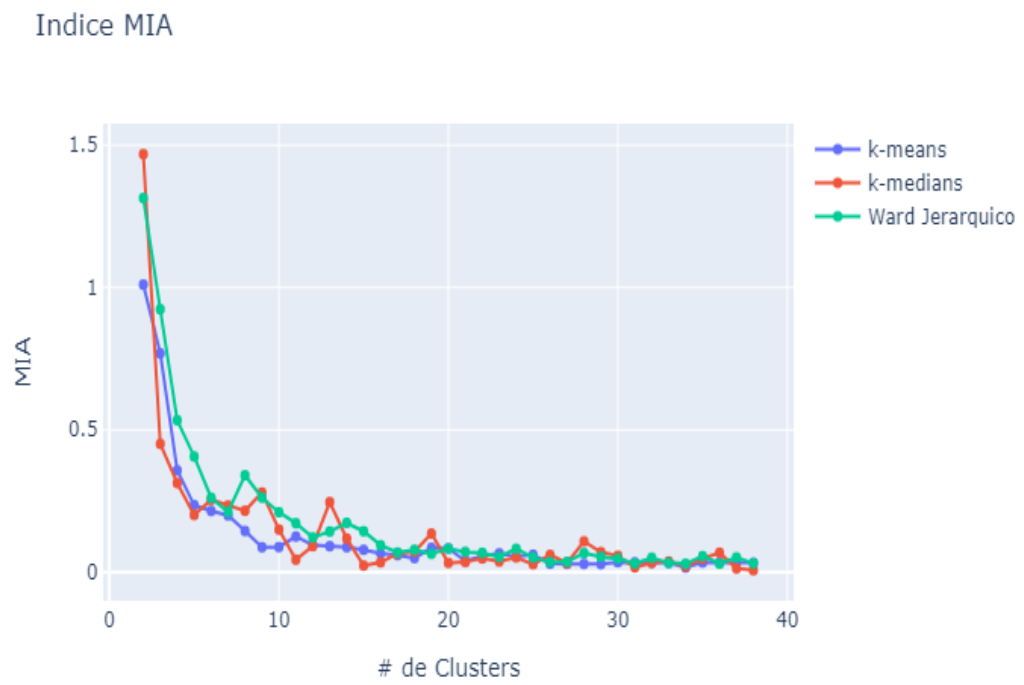


Figura 5.3 – Curva de índice MIA para data de entrenamiento
Fuente: Realizado por el autor.

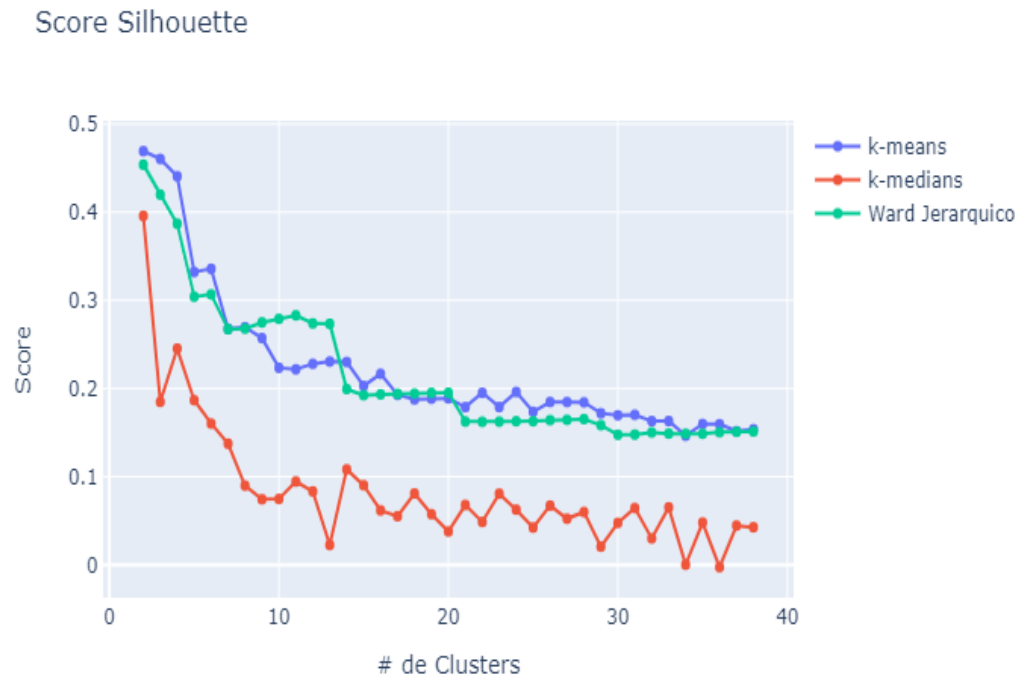


Figura 5.4 – Curva de índice Silueta para data de entrenamiento
Fuente: Realizado por el autor.

Con los índices se tiene un punto de partida de que algoritmo es más eficaz entre que valores de k, sin embargo, falta destacar la cuestión más importante y es: ¿Qué valor de K grupos es el ideal para detectar mayor cantidad de pérdidas?. Esto se responde con los resultados de la Tabla 5.3 que muestra que los mejores algoritmos para agrupar usuarios con pérdidas es efectivamente K-means y Ward, siendo este último el que mayor cantidad de usuarios fraudulentos aglomera con $K = 11$ grupos con $74,14\,\%$. El resto de usuarios fraudulentos se estratifican con los usuarios normales debido a que no presentan algún tipo de pérdidas significativa en su patrón de consumo, posiblemente porque la intensidad y la duración del fraude son muy bajas.

Tabla 5.3 – Porcentaje de pérdidas

	Porcentaje de pérdidas		
K Grupos	**K-medias**	**K-medianas**	**Ward**
8	59.48	56.03	62.07
9	62.93	52.59	68.97
10	65.52	63.79	71.55
11	59.48	41.38	74.14
12	64.66	43.10	74.14
13	72.41	57.76	74.14
14	73.28	56.90	74.14
15	73.28	43.97	74.14
20	69.83	48.28	81.03

Cabe indicar que el criterio para detectar las pérdidas fue obtenido mediante la relación del número de usuarios con pérdidas y el número total de usuarios por grupos, si esta relación era igual o superior al $80\,\%$ entonces se considera a este grupo un estrato de alta concentración de fraude.

Para el valor de grupos $K = 11$ y el algoritmo Ward seleccionados como mejor opción para detección de usuarios con pérdidas en esta fase de entrenamiento se logró determinar cinco grupos exclusivamente de usuarios con fraude. Estos resultados se presentan en los siguientes figuras. Por ejemplo, la Figura 5.6 muestra tres tipos de gráficos; primero el diagrama de caja o llamado boxplot que muestra la distribución y la distorsión de los datos para cada hora. La segunda gráfica utiliza la varianza como índice de medida para observar que tan cerca o lejos se encuentran los valores de la media del subconjunto de datos que se ha formado. Y la tercera gráfica, presenta una curva central que representa el patrón de consumo promedio o central característico del grupo donde las capas superior e inferior que sombrean alrededor de la curva es producto de la desviación estándar $\pm\sigma$. Otro resultado que se presenta es el de cantidad de usuarios con pérdidas detectados en cada grupo (Figura 5.11).

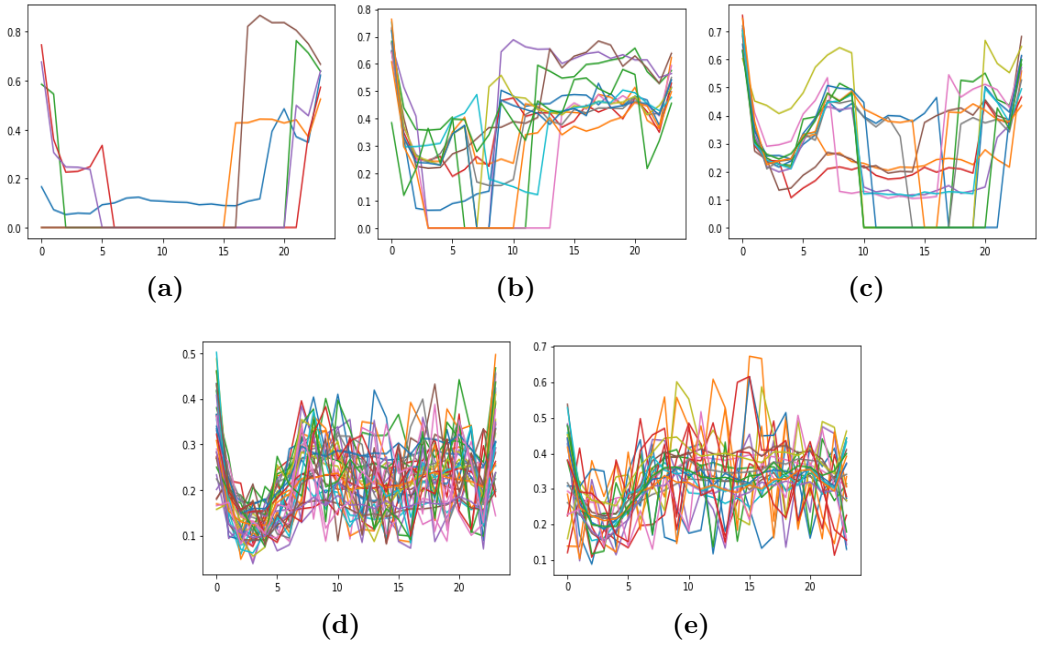


Figura 5.5 – Perfiles de demanda de usuarios con pérdidas no técnicas por grupos.
Fuente: Realizado por el autor.

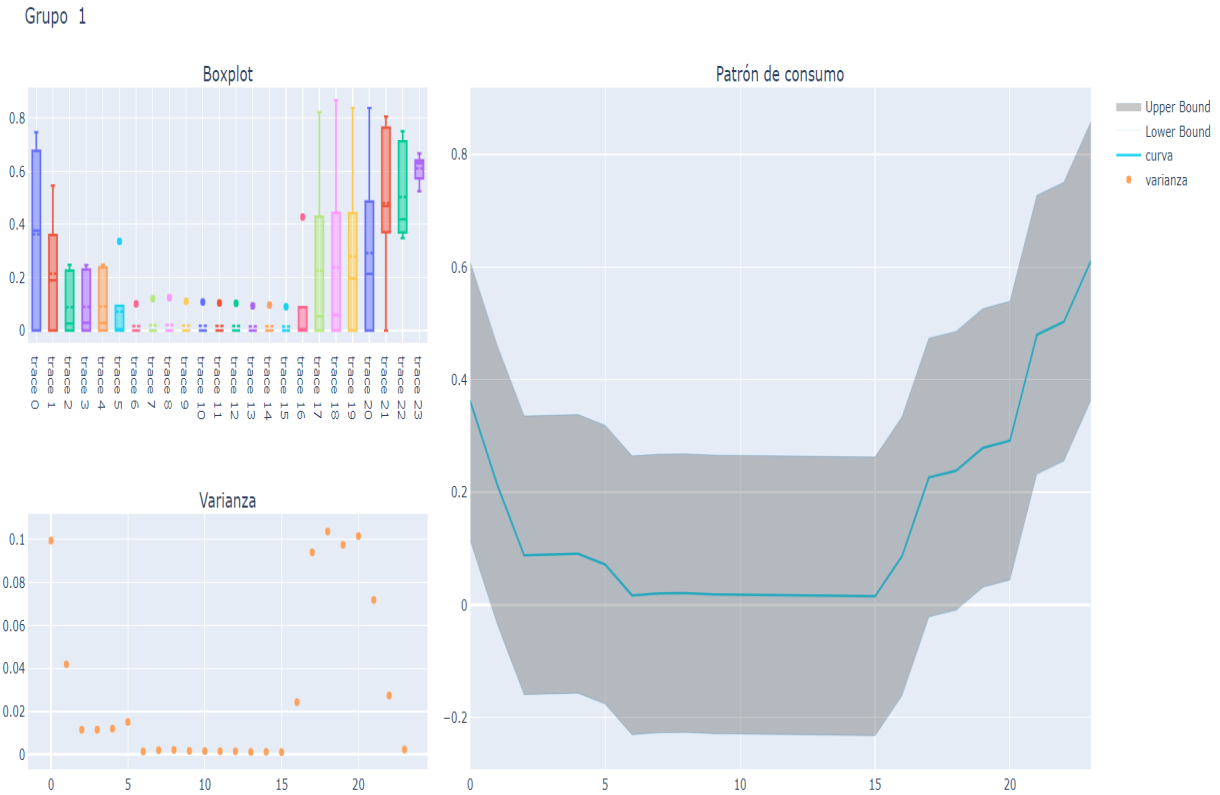


Figura 5.6 – Centro de Grupo 1.
Fuente: Realizado por el autor.

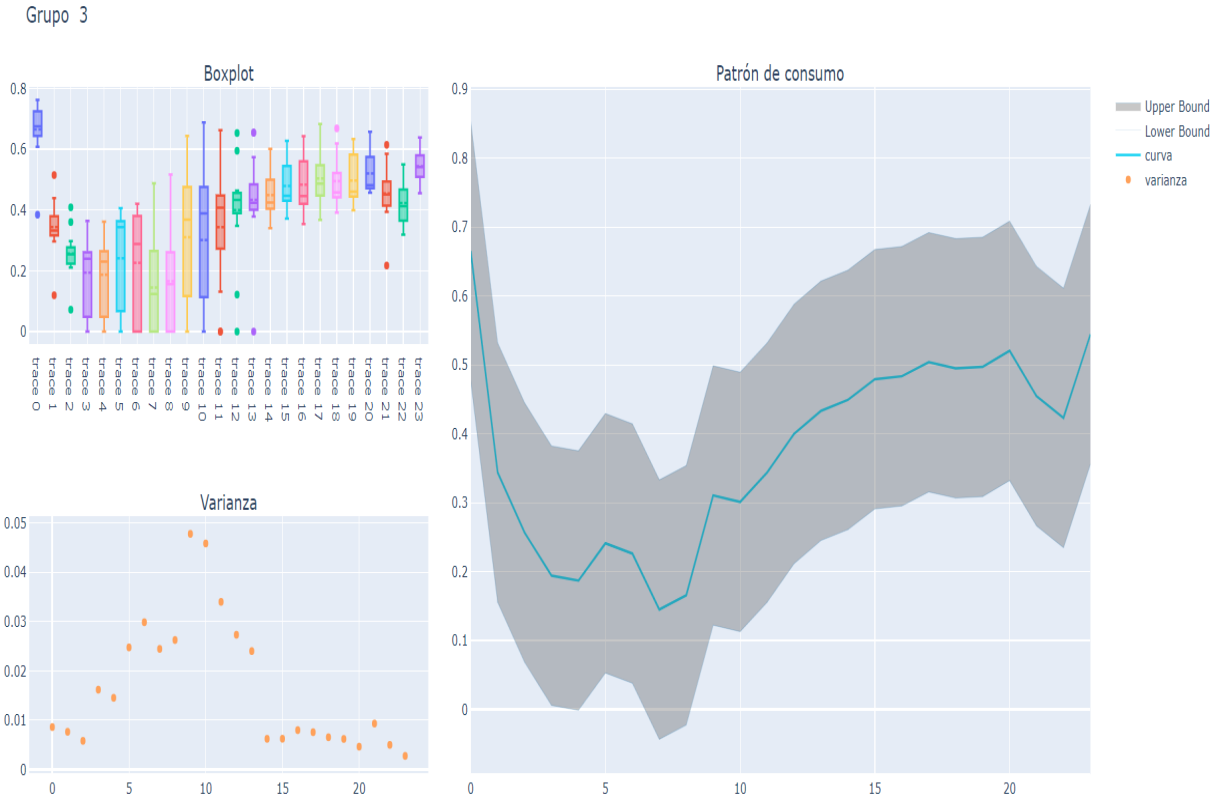


Figura 5.7 – Centro de Grupo 3.
Fuente: Realizado por el autor.

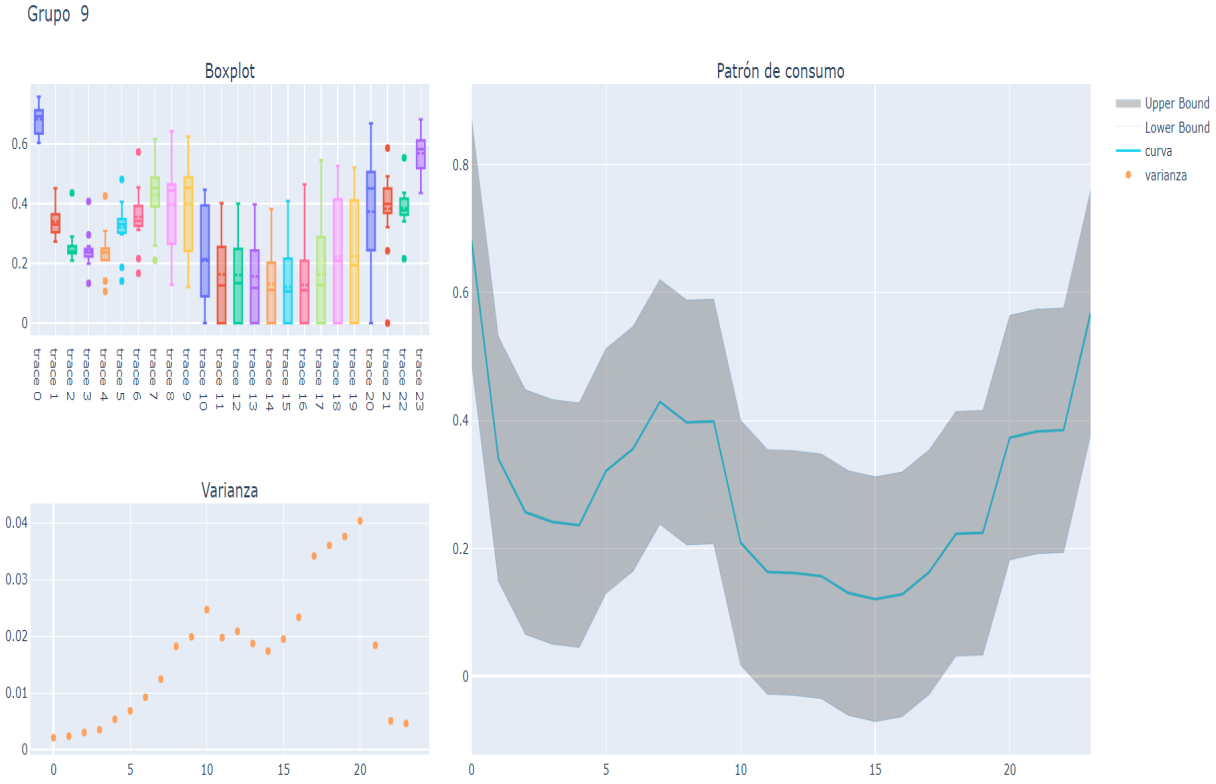


Figura 5.8 – Centro de Grupo 9.
Fuente: Realizado por el autor.

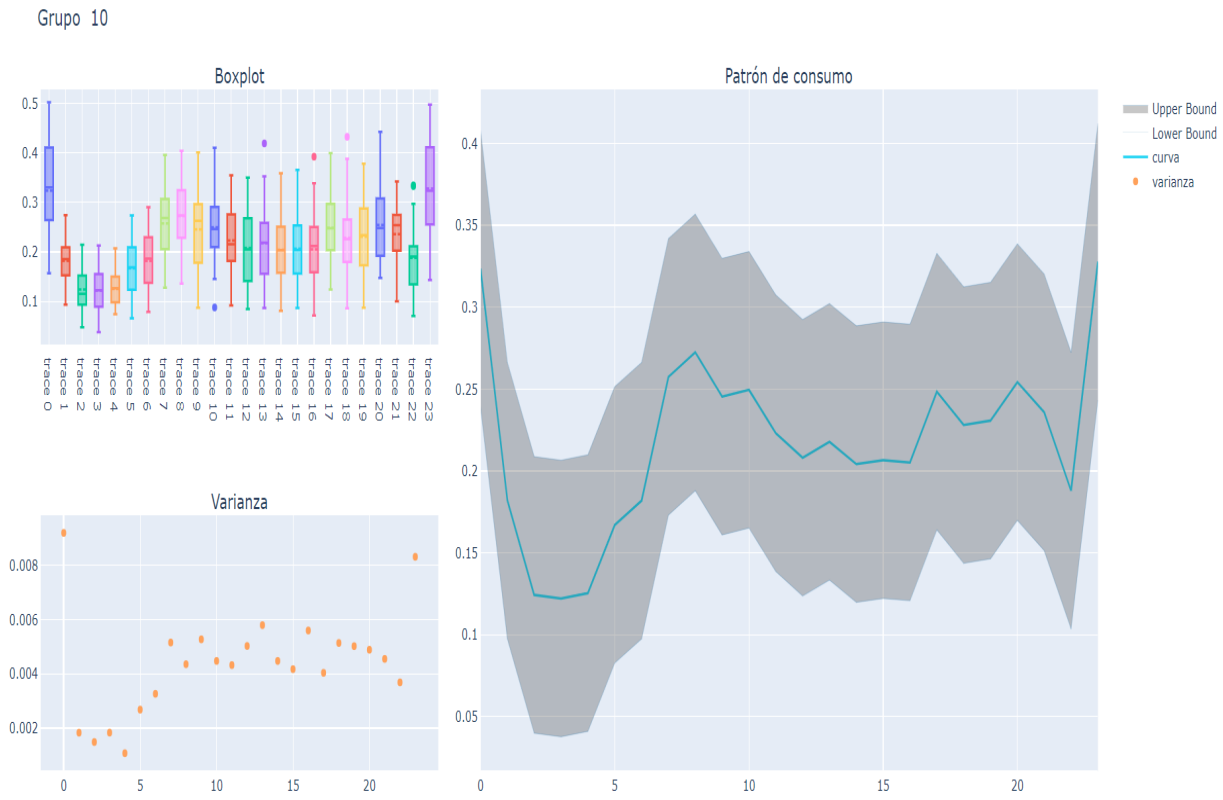


Figura 5.9 – Centro de Grupo 10.
Fuente: Realizado por el autor.

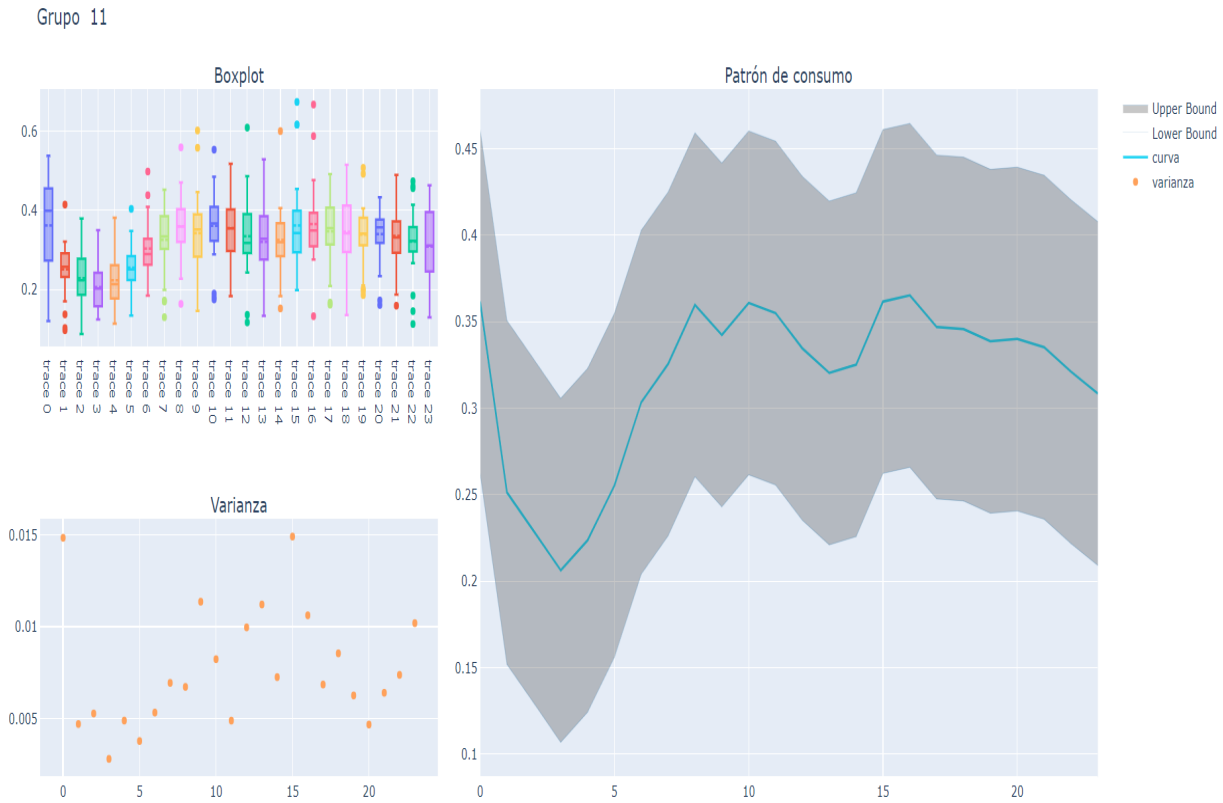


Figura 5.10 – Centro de Grupo 11.
Fuente: Realizado por el autor.

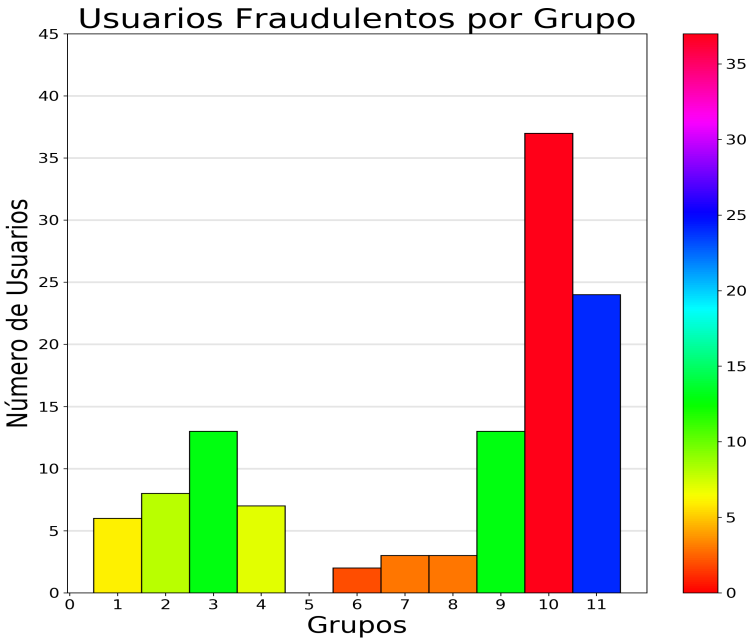


Figura 5.11 – Usuarios fraudulentos por grupos.
Fuente: Realizado por el autor.

En la figura de arriba se observa que aunque hay usuarios con pérdidas además de los ya mencionados y a excepción del grupo 5 que no presenta ningún usuario con perdidas, estos usuarios no representan ni el 10 % de los usuarios total por grupo, así que no se han considerado como grupos fraudulentos.

En esta última instancia de la fase de entrenamiento se intenta diferenciar los grupos fraudulentos de los grupos benignos mediante medidas descriptivas estadísticas como la media, mediana, desviación estándar de la distribución de grupos. También se utiliza la distancia educlidiana entre el centro y la varianza del grupo que pertenecen para limitar a los usuarios que se encuentre entre este espectro. Si bien cierto hemos definido los criterios para 11 grupos, tendremos en cuenta criterios también para valores ± 1, es decir para $K = 10$ y $K = 12$, esto para darle más soporte al algoritmo en el momento de decidir por la opción más óptima. Con esto se definen y establecen criterios que puedan detectar usuarios con las mismas cualidades de los grupos ya formados y así ser aplicado en la posterior fase. El resumen de criterios resultantes se muestran en la Tabla 5.4.

Tabla 5.4 – Valores de índices y medidas descriptivas, de 10 a 12 grupos.

	WARD		
	K=10	**K=11**	**K=12**
MIA	0.21039	0.17179	0.12173
Silueta	0.27906	0.28305	0.27391
Media	127.8	116.18	106.5
Mediana	97	58	29
Desv. Estándar	132.74	130.91	128.65

5.2. Prueba

En esta fase se realiza la evaluación de los criterios establecidos en la fase de entrenamiento sobre la data de prueba. El algoritmo se desarrolla de tal manera que no se base solamente en el proceso iterativo de hallar las curvas de amalgamiento para escoger el número óptimo de K, sino también utilizando las siguientes pautas. El índice MIA está determinado para valores menores 0.21039 hasta 0, recordemos que el valor de MIA entre más pequeño indica mayor heterogeneidad entre grupos. El índice Silueta, por lo contrario, si el valor es cercano a 1 representa mayor homogeneidad de los elementos de un grupo, por eso se plantea este valor desde 0.27391 hasta 1. Las medidas estadísticas media y desviación estándar no aportan mucho al análisis, pero, en cambio el análisis de la mediana proporciona información útil como el que los grupos formados con tamaño menor al valor de la mediana son grupos con usuarios fraudulentos. El último valor de criterio es la distancia euclidiana que permite la integración a un grupo solo usuarios con pérdidas que se encuentre más cercanos al centro del grupo. La Tabla 5.5 presenta las pautas principales para seleccionar K.

Tabla 5.5 – Criterios principales para seleccionar K

	WARD		
	K=10	**K=11**	**K=12**
MIA	0.15030	0.12014	0.16444
Silueta	0.26666	0.26345	0.24725
Mediana	98.5	82.0	79.0

Finalmente, definidos los criterios el algoritmo presenta como resultado el valor de K para el agrupamiento que obtenga mayor porcentaje de pérdidas detectadas e imprime los grupos que en efecto representan a los usuarios con pérdidas. Las figuras de la parte inferior exponen los resultados para el algoritmo que ha seleccionado 11 grupos y ha extraído los grupos con posible fraude. Se detectó en total 67.96 % de usuarios con pérdidas en grupos bien diferenciado tal como en el caso de entrenamiento.

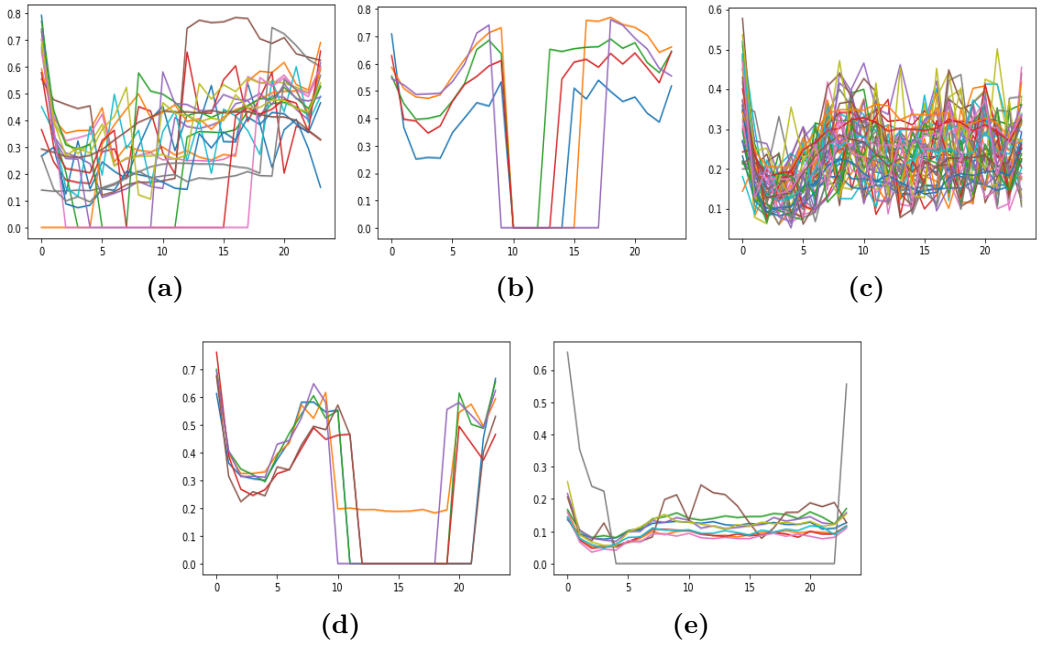


Figura 5.12 – Perfiles de demanda de usuarios con pérdidas no técnicas por grupos detectados en fase de prueba.
Fuente: Realizado por el autor.

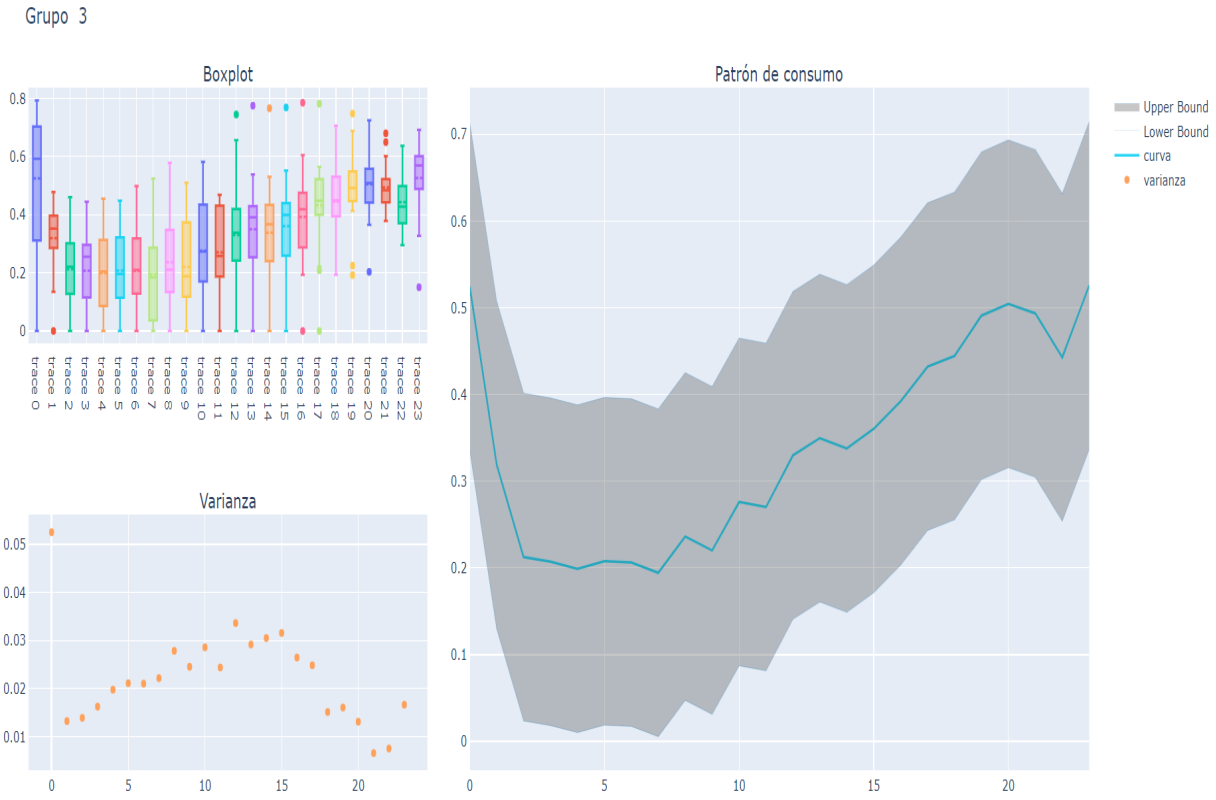


Figura 5.13 – Centro de Grupo 3. Fase de prueba.
Fuente: Realizado por el autor.

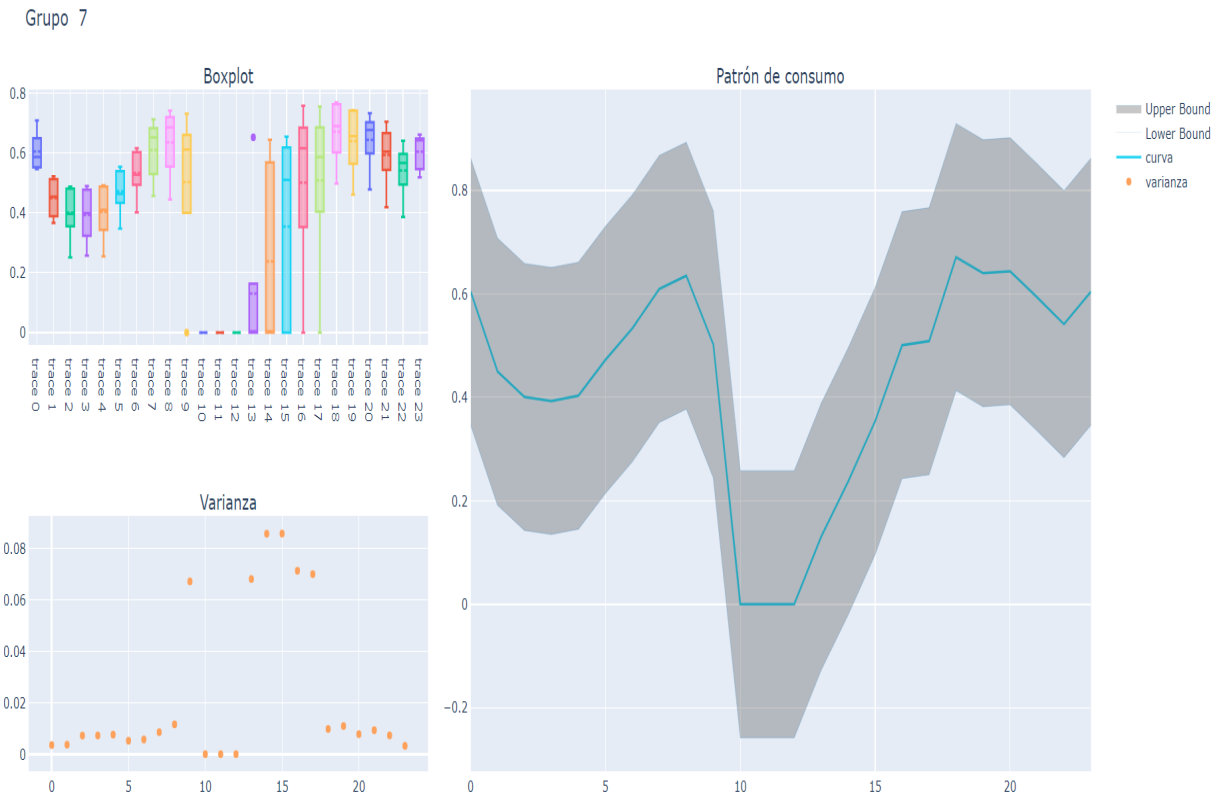


Figura 5.14 – Centro de Grupo 7. Fase de prueba.
Fuente: Realizado por el autor.

Figura 5.15 – Centro de Grupo 9. Fase de prueba
Fuente: Realizado por el autor.

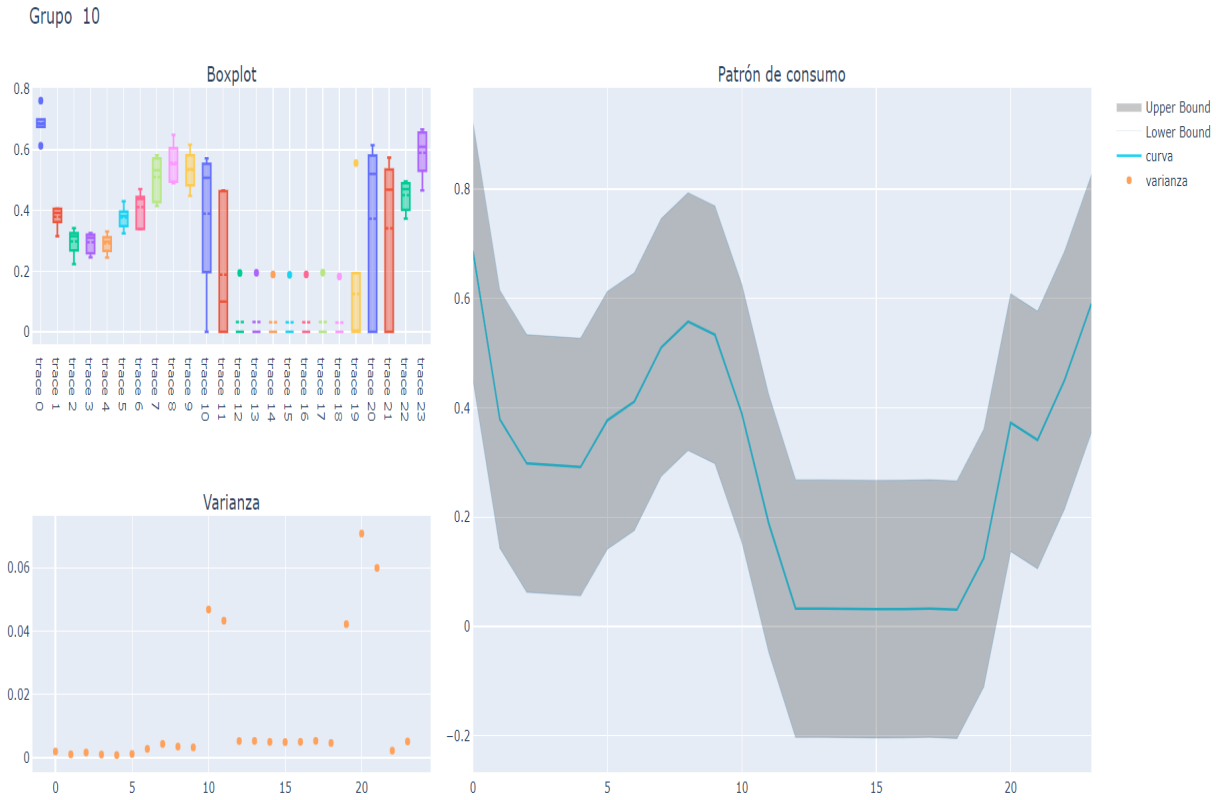


Figura 5.16 – Centro de Grupo 10. Fase de prueba.
Fuente: Realizado por el autor.

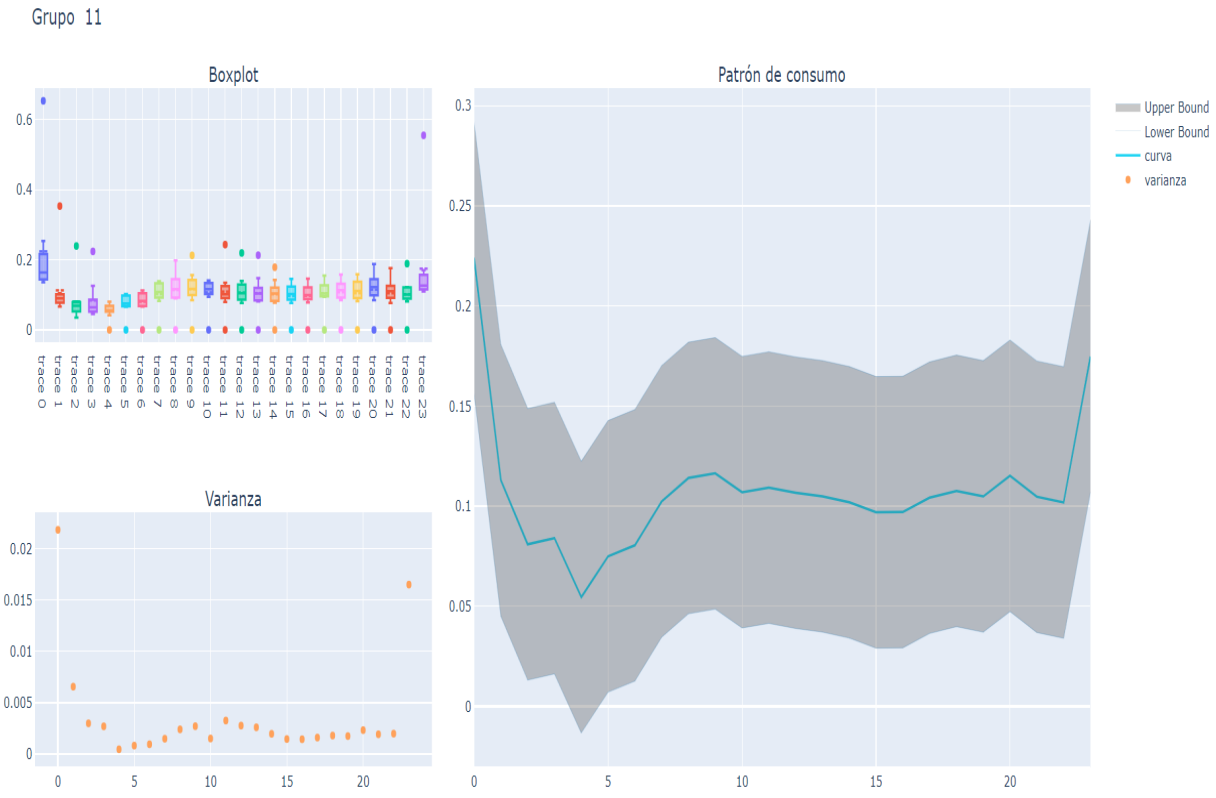


Figura 5.17 – Centro de Grupo 11. Fase de prueba.
Fuente: Realizado por el autor.

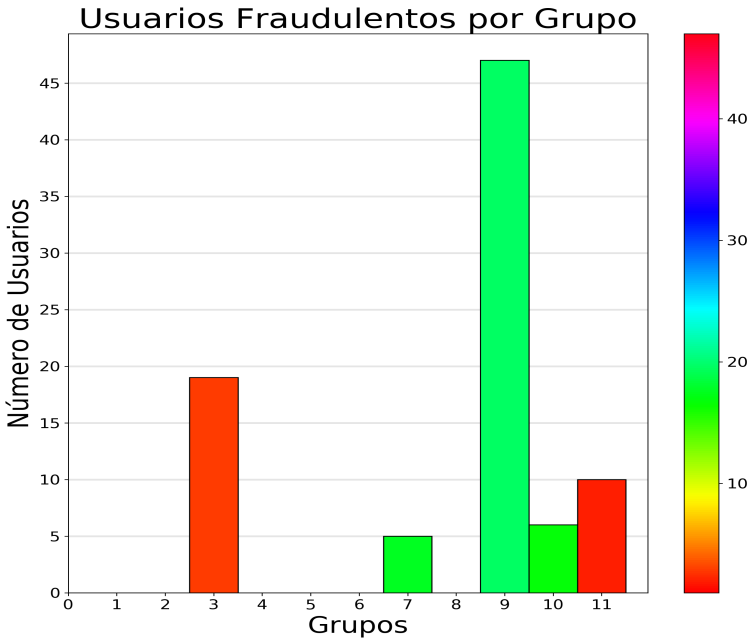


Figura 5.18 – Usuarios fraudulentos por grupos. Fase de prueba.
Fuente: Realizado por el autor.

Capítulo 6

CONCLUSIONES Y LÍNEAS FUTURAS

6.1. Conclusiones

- La metodología propuesta en este trabajo permite la detección de pérdidas no técnicas mediante los registros de consumo eléctrico de los usuarios. Este propósito se llevó a cabo empleando algoritmos de agrupamiento. Los resultados obtenidos en la fase de prueba evidencian que es posible detectar a usuarios que presentan irregularidades en su patrón de consumo, reconociendo aproximadamente al 68 % de las pérdidas aplicadas. Por tanto, es de suma importancia contar con metodologías alternativas computacionales para la solución de problemas, para el caso de estudio, la detección de perdidas no técnicas en los sistemas de distribución.

- Se examinaron diferentes algoritmos de agrupamiento y se plantearon varios criterios de evaluación para detectar uno o varios grupos de usuarios con pérdidas no técnicas, de los cuales sobresale Jerárquico-Ward ya que mostró alta diferenciación entre los grupos de usuarios benignos y con pérdidas.

- El conjunto de datos aleatorio generado a partir de modelos de Markov presentó un muy buen rendimiento frente a los objetivos perseguidos, logrando resolver la problemática de la insuficiencia de datos.

- Debido a la ausencia de registros de patrones de consumo con pérdidas no técnicas se desarrolló un modelo para simular el comportamiento de usuarios con irregularidades o anomalías. Su aplicación fue fundamental para probar la metodología propuesta.

- La detección de un grupo de usuarios con posibles pérdidas no técnicas permitirá a la empresa eléctrica manejar de mejor manera los recursos tanto técnicos como económicos, por ejemplo, optimizando las campañas de inspecciones en sectores específicos.

- El gran problema que manifiesta esta metodología es su baja adaptabilidad, es decir, se tiene que crear diferentes criterios para cada conjunto distinto de datos. Lo cual disminuye su nivel de eficiencia.

- Otro problema que se presenta es la falta de automatización en los procesos de agrupamiento, puesto que requiere de un experto para el análisis de los datos y para la obtención de resultados satisfactorios.

6.2. Líneas futuras

- En el transcurso de este trabajo se ha logrado identificar diferentes técnicas de minería de datos que pueden ser aplicadas a los registros de los sistemas eléctricos de distribución. Una de las propuestas para continuar esta línea de investigación es desarrollar algoritmos más sofisticados y adaptables basados en aprendizaje supervisado que permitan la detección y predicción del fraude energético utilizando otras variables de los usuarios junto a los perfiles de carga como: tipo de contrato, ubicación geográfica, número de inspecciones realizadas en el predio, etc. De esta manera se caracteriza al usuario no tan solamente por su nivel de potencia consumida.
- Otro aspecto que se puede agregar a este trabajo es la utilización de un conjunto de datos con registro de medidores inteligentes con frecuencias de muestreo más bajas, por ejemplo, mediciones de consumo eléctrico cada cinco o quince minutos y durante un periodo superior a seis meses. El empleo de data real en su totalidad proporciona mayor validez a los modelos implementados.

Bibliografía

[1] R. Jiménez, T. Serebrisky, and J. Mercado, "Dimensionando las pérdidas de electricidad en los sistemas de transmisión y distribución en américa latina y el caribe," *BI desarrollo, Ed.) ELECTRICIDAD PERDIDA*, 2014.

[2] J. Arévalo, David y Bowen, "Planificación del control de pérdidas no técnicas del sistema de distribución de energís eléctrica de guayaquil." Proyecto de Grado, 2019.

[3] Y. Al-Mahroqi, I. Metwally, A. Al-Hinai, and A. Al-Badi, "Reduction of power losses in distribution systems," *International Journal of Computer and Systems Engineering*, vol. 6, no. 3, pp. 312 – 322, 2012.

[4] A. G. P. Sevilla and F. O. A. Fernández, "Evolución de las redes eléctricas hacia smart grid en países de la región andina," *Revista Educación en Ingeniería*, vol. 8, no. 15, pp. 48–61, 2013.

[5] S. McLaughlin, D. Podkuiko, and P. McDaniel, "Energy theft in the advanced metering infrastructure," in *International Workshop on Critical Information Infrastructures Security*, pp. 176–187, Springer, 2009.

[6] Banco Interamericano de Desarrollo, *Manual latinoamericano y del Caribe para el control de pérdidas eléctricas*. Documento OLADE, OLADE, Organización Latinoamericana de Enérgia, 1990.

[7] P. Glauner, J. Meira, P. Valtchev, R. State, and F. Bettinger, "The challenge of non-technical loss detection using artificial intelligence: A survey," *International Journal of Computational Intelligence Systems*, vol. 10, pp. 760–775, 02 2017.

[8] T. B. Smith, "Electricity theft: a comparative analysis," *Energy Policy*, vol. 32, no. 18, pp. 2067 – 2076, 2004.

[9] G. M. Mazzini, "Propuesta de control y reducción de pérdidas no técnicas de energía eléctrica implementando modelo de telegestión con sistemas inteligentes. empresa eléctrica pública de guayaquil, ep.," 2015.

[10] A. Tama Franco, "Las pérdidas de energía," 12 2013.

[11] J. Gómez, R. Castán, J. Montero, J. Meneses, and J. García, "Aplicación de tecnologías de medición avanzada (ami) como instrumento para reducción de pérdidas," *Boletín IIE*, vol. 39, no. 4, pp. 180–191, 2015.

[12] N. M. G. Strategy, "Advanced metering infrastructure," *US Department of Energy Office of Electricity and Energy Reliability*, 2008.

[13] J. Aranda, "(spanish) design and performance evaluation of a communication network for smart metering in network simulator?2," *INGENIERIA*, vol. 20, pp. 21–35, 04 2015.

[14] R. R. Mohassel, A. Fung, F. Mohammadi, and K. Raahemifar, "A survey on advanced metering infrastructure," *International Journal of Electrical Power & Energy Systems*, vol. 63, pp. 473–484, 2014.

[15] S. E. International, "Global trends in smart metering," Jun 2019.

[16] 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,"

[17] L. o. A. MIRACLE Product Details from Wenzhou Maihong Electric Technology Co., "Carrier wave and gprs energy meter dcu data concentrator for remote smart meter reading system, view concentrator."

[18] J. Alvarado, "Servicios de medición avanzada (ami) para redes inteligentes y su adaptabilidad en el marco de la legislación ecuatoriana," *Universidad de Cuenca*, 2010.

[19] N. Beigi Mohammadi, J. Mišić, V. B. Mišić, and H. Khazaei, "A framework for intrusion detection system in advanced metering infrastructure," *Security and Communication Networks*, vol. 7, no. 1, pp. 195–205, 2014.

[20] M. A. Faisal, Z. Aung, J. R. Williams, and A. Sanchez, "Securing advanced metering infrastructure using intrusion detection system with data stream mining," in *Pacific-Asia Workshop on Intelligence and Security Informatics*, pp. 96–111, Springer, 2012.

[21] A. Hawkins, "Meter data management agenda what is mdm? where does mdm fit?," CPS Energy, 2018.

[22] J. Han, J. Pei, and M. Kamber, *Data mining: concepts and techniques*. Elsevier, 2011.

[23] M. Kantardzic, *Data mining: concepts, models, methods, and algorithms*. John Wiley & Sons, 2011.

[24] J. Nagi, "An intelligent system for detection of non-technical losses in tenaga nasional berhad (tnb) malaysia low voltage distribution network," *Vasa*, 2009.

[25] F. Gorunescu, *Data Mining: Concepts, models and techniques*, vol. 12. Springer Science & Business Media, 2011.

[26] J. F. Hair, R. E. Anderson, R. L. Tatham, W. C. Black, *et al.*, *Análisis multivariante*, vol. 491. Prentice Hall Madrid, 1999.

[27] R. Xu and D. Wunsch, "Survey of clustering algorithms," *IEEE Transactions on neural networks*, vol. 16, no. 3, pp. 645–678, 2005.

[28] L. Sotillo, "Desarrollo de modelo de demanda aplicable al sistema electrico nacional." Proyecto de Grado, 3 2017.

[29] F. Murtagh and P. Legendre, "Ward?s hierarchical agglomerative clustering method: which algorithms implement ward?s criterion?," *Journal of classification*, vol. 31, no. 3, pp. 274–295, 2014.

[30] scikit learn, "Plot hierarchical clustering dendrogram."

[31] N. Mahmoudi-Kohan, M. Moghaddam, and S. Bidaki, "Evaluating performance of wfa k-means and modified follow the leader methods for clustering load curves," in *2009 IEEE/PES Power Systems Conference and Exposition*, pp. 1–5, IEEE, 2009.

[32] S. Bidoki, N. Mahmoudi-Kohan, and S. Gerami, "Comparison of several clustering methods in the case of electrical load curves classification," in *16th Electrical Power Distribution Conference*, pp. 1–7, IEEE, 2011.

[33] P. J. Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," *Journal of computational and applied mathematics*, vol. 20, pp. 53–65, 1987.

[34] P. Jokar, N. Arianpoo, and V. C. Leung, "Electricity theft detection in ami using customers? consumption patterns," *IEEE Transactions on Smart Grid*, vol. 7, no. 1, pp. 216–226, 2015.

[35] PSEG, "Historical load profiles."

[36] L. R. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[37] A. Crespí Tobeña, "Modelos ocultos de markov para el etiquetado de texto," 2018.

[38] J. Tornero Lucas, "Machine learning: modelos ocultos de markov (hmm) y redes neuronales artificiales (ann)," 2017.

[39] P. Dymarski, *Hidden Markov Models: Theory and Applications*. BoD–Books on Demand, 2011.

[40] M. Zanetti, E. Jamhour, M. Pellenz, M. Penna, V. Zambenedetti, and I. Chueiri, "A tunable fraud detection system for advanced metering infrastructure using short-lived patterns," *IEEE Transactions on Smart grid*, vol. 10, no. 1, pp. 830–840, 2017.

[41] J. L. Viegas and S. M. Vieira, "Clustering-based novelty detection to uncover electricity theft," in *2017 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pp. 1–6, IEEE, 2017.