ESCUELA POLITÉCNICA NACIONAL

FACULTAD DE INGENIERÍA ELÉCTRICA Y ELECTRÓNICA

CLASIFICACIÓN DE LOS USUARIOS DEL SERVICIO ELÉCTRICO A PARTIR DE LOS DATOS DE TELEMEDICIÓN COMERCIAL E INDUSTRIAL EN EL ÁREA DE CONCESIÓN DE LA EMPRESA ELÉCTRICA AMBATO REGIONAL CENTRO NORTE S.A.(EEASA)

TRABAJO DE TITULACIÓN PREVIO A LA OBTENCIÓN DEL TÍTULO DE INGENIERO EN INGENIERÍA ELÉCTRICA

JEFFERSON ISRAEL GUAMANI MONTA

DIRECTOR: DR. ING. GABRIEL SALAZAR YÉPEZ.

Quito, abril 2023

AVAL

Certifico que el presente trabajo fue desarrollado por Jefferson Israel Guamani Monta, bajo mi supervisión.

DR. ING. GABRIEL SALAZAR YÉPEZ.
DIRECTOR DEL TRABAJO DE TITULACIÓN

DECLARACIÓN DE AUTORÍA

Yo, Jefferson Israel Guamani Monta, declaro bajo juramento que el trabajo aquí descrito es de mi autoría; que no ha sido previamente presentado para ningún grado o calificación profesional; y, que he consultado las referencias bibliográficas que se incluyen en este documento.

A través de la presente declaración dejo constancia de que la Escuela Politécnica Nacional podrá hacer uso del presente trabajo según los términos estipulados en la Ley, Reglamentos y Normas vigentes.

JEFFERSON ISRAEL GUAMANI MONTA

DEDICATORIA

Al Padre por tanto amor incondicional y su fidelidad

A mis padres Raúl y Isabel

A mis tías Elvira, Martha

A mis tíos Washington y Luis

A mis hermanos Diego y Kevin

Israel

AGRADECIMIENTO

Al Padre por siempre ser mi ayudador y proveedor en medio de la tempestad, su mano siempre presente en toda adversidad y dificultad encontrada en este caminar.

A mi Madre María Isabel por su lucha incansable por apoyarme indudablemente desde el primer día de mi vida, por nunca haberse dejado de apoyarme, pese al sin número de errores, por ser esa mamá dura y a la vez por su amor tan invaluable, esta meta es por usted.

A mis tías Elvira y Martha por siempre creer en mí, en cuidarme como un hijo en el largo de mis estudios superiores, por siempre darme cuando menos lo espere y cuando más lo necesite.

A mis tíos Washington y Luis por contribuir con mi crecimiento desde muy temprana edad hasta el grado universitario por apoyarme tanto económicamente como moralmente.

A mis hermanos Kevin y Diego por ser mi luz y mi motivación diaria, por el nuevo legado que han dejado sobre este nuevo comienzo en la familia.

A mis amigos Bryan, Colon, Andrés, Andrés A, Ángel, Jorge, Alejandro que se convirtieron en mi familia en una ciudad desconocida, gracias por su confianza y los momentos vividos en el objetivo de conseguir el tan anhelado título Universitario.

Al Ing. Juan Ramírez por ayudarme incansablemente en el desarrollo de este proyecto y ser el guía en el desarrollo de este proyecto.

Al Dr. Gabriel Salazar por su apoyo desde el primer día en aceptar ser el director de tesis y sus conocimientos impartidos durante la carrera que sin duda alguna contribuyeron a la finalización.

A la Empresa Eléctrica Ambato por facilitar la base de datos para la ejecución y análisis de este proyecto de titulación, en especial al Ing. Bryan Mayorga operador del centro de control.

ÍNDICE DE CONTENIDO

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

RESUMEN

El presente proyecto de titulación consiste en la obtención de curvas de demanda para ser introducidas dentro del ADMS, con el objetivo de mejorar el estimador de estado, para usuarios industriales y comerciales dentro del área de concesión de la EEASA (Empresa Eléctrica Ambato), para esto se utiliza algoritmos de clasificación con lenguaje de máquina para el tratamiento de cada uno de los datos.

La metodología consiste en cuatro etapas. La primera etapa inicia por la descarga de cada perfil de consumo con datos de telemedición de los 184 clientes comerciales e industriales desde el portal de la EEASA, posteriormente estos datos son almacenados en una hoja de Excel. La segunda etapa se basa en el ordenamiento, limpieza y clasificación de los datos históricos de tipo comercial e industrial de consumo eléctrico, para luego iniciar con el preprocesamiento mediante la aplicación del algoritmo K-means que es un algoritmo de clasificación no supervisada (clusterización), de esta manera formando una matriz de datos que posteriormente es exportada a Python. La tercera parte consiste en el uso de PCA's (Principal Component Analysis) para la reducción del gran volumen de datos, inmediatamente estos datos son caracterizados por estacionalidad.

Finalmente, la aplicación del algoritmo Fuzzy-c-means para la obtención de una curva atípica trimestral, obteniendo como resultado las curvas de demanda que servirán como referencia para el comportamiento de los usuarios industriales y comerciales.

PALABRAS CLAVE: Matriz de datos, K-means, PCA's, Fuzzy-c-means, curva de demanda, telemedición, cluster, preprocesamiento, clientes, machine learning, estacionalidad.

ABSTRACT

This titling project consists of obtaining demand curves to be introduced into the ADMS, with the aim of improving the state estimator, for industrial and commercial users within the concession area of the EEASA (Empresa Eléctrica Ambato), to for this, classification algorithms with machine language are used for the treatment of each of the data.

The methodology consists of four stages. The first stage begins by downloading each consumption profile with telemetering data from the 184 commercial and industrial clients from the EEASA portal, later this data is stored in an Excel sheet. The second stage is based on the ordering, cleaning and classification of the historical data of commercial and industrial type of electricity consumption, to then start with the pre-processing by applying the K-means algorithm, which is an unsupervised classification algorithm (clustering). , thus forming a data matrix that is later exported to Python. The third part consists of the use of PCA's (Principal Component Analysis) for the reduction of the large volume of data, immediately these data are characterized.

Finally, the application of the Fuzzy-c-means algorithm to obtain a quarterly atypical curve, obtaining as a result the demand curves that will serve as a reference for the behavior of industrial and commercial users.

KEYWORDS: data, K-means, DBI, PCA, Fuzzy-c-means, load profile, telemetering, cluster, preprocessing, trimester, customers, algorithms

1 INTRODUCCIÓN

La clasificación de usuarios industriales y comerciales en función del consumo de energía en los últimos tiempos ha sido notable, en las últimas décadas el número de usuarios va en crecimiento, los consumidores de electricidad crecen en forma dinámica y pueden desempeñar un papel importante en el apoyo a la decisión y en la definición del comportamiento estratégico del mercado.

Para la medición del consumo eléctrico de los clientes pertenecientes a la Empresa Eléctrica Ambato Regional Centro Norte S.A (EEASA) es empleado el sistema de lectura automática de medidores (AMI), para clientes de facturación masiva, especial y medición de transformadores de distribución [1], que recopilan el consumo y el estado de datos de dispositivos de medición eléctrica y transfieren los datos a bases de datos centrales en intervalos predefinidos. La clasificación de usuarios permite el nuevo diseño de tarifas para cada tipo de usuario como también las curvas de consumo ofrecen soluciones fundamentales que permiten pronosticar el consumo futuro de electricidad, para planificar la cantidad de electricidad que necesita el país y de esta manera no depender de otros países que vendan energía a bajo costo.

El uso de datos históricos en su mayoría sin ningún tipo de preprocesamiento ha desencadenado la presencia de errores con el estimador de estado de distribución (DSE), es la función de potencia básica del ADMS que indica la calidad de las señales. El (DSE) compara el valor que se debería tener en un punto, con el valor que se calcula mediante flujos [1]. La cantidad de datos de cómo y cuándo los consumidores utilizan la electricidad, ha desencadenado la aplicación de técnicas de minería de datos en varios sistemas de energía, el problema de clasificar a los clientes de electricidad en función de sus comportamientos es la gran dispersión de datos. Se utilizan técnicas estadísticas y de agrupamiento para determinar los perfiles de carga, para así obtener los diferentes patrones de consumo, estos patrones están representados por sus perfiles y cada perfil es útil para obtener mejores decisiones de gestión de soporte en términos de planificación de ofertas en mercados de energía en tiempo real, oferta de precios y detectar fallas tempranas dentro del sistema ADMS (Advanced Distribution Management System).

Sobre la base de lo mencionado, se ha propuesto en este Trabajo de Titulación un procedimiento para la obtención de perfiles de consumo para luego ser implementados en el ADMS de la EEASA, este estudio inicia almacenando los datos de demanda de los

usuarios comerciales e industriales del Sistema de Telemedición del portal de la Empresa Eléctrica Ambato. A partir del almacenamiento de los datos, el procedimiento a continuar es con el tratamiento de estos, durante el proceso, la falta de valores se rellena y los valores atípicos se detectan, posteriormente estos datos son suavizados. Una vez que los datos están listos luego del proceso de tratamiento se usan técnicas de minería de datos para la reducción del volumen de datos, para luego ser procesados mediante la utilización de machine learning usando algoritmos K-means, por medio de las técnicas de clusterización y finalmente agrupados mediante el algoritmo Fuzzy-c-means, que intenta medir la afinidad que tiene una muestra de datos con respecto a un clúster, este último proceso sirvió para obtener las curvas de demanda [2].

1.1 OBJETIVO GENERAL

El objetivo general de este Proyecto Técnico es:

Clasificar los usuarios del servicio eléctrico a partir de datos de telemedición comercial e industrial de los días laborables utilizando el análisis de componentes principales (PCA) y conglomerados Fuzzy-c-means para la generación de curvas de demanda que se utilizarán en el estimador de estado del ADMS en el área de concesión de la Empresa Eléctrica Ambato Regional Centro Norte S.A (EEASA).

1.2 OBJETIVOS ESPECÍFICOS

Los objetivos específicos del Proyecto Técnico son:

- Realizar una investigación bibliográfica sobre la temática relacionada a la clasificación y obtención de curvas de demanda diaria en base a datos medidos.
- Preprocesar los datos históricos de la carga provenientes de la plataforma del sistema de telemedición para los clientes especiales de la EEASA en hojas de (Excel) mediante programación en Python.
- Implementar el modelo computacional en lenguaje Python 3.7 que permita manejar la información para una clasificación, tratamiento y conglomeración de los datos históricos una vez procesados mediante algoritmos de aprendizaje de máquina aplicado a la EEASA.
- Comparar los resultados de la clasificación por consumo y por patrón de consumo.

- Obtener las curvas de demanda diaria del servicio eléctrico de los usuarios industriales y comerciales para ser ingresadas en el ADMS.

1.3 ALCANCE

Para el desarrollo del estudio se descargarán del portal de la EEASA los datos históricos de telemedición de los 184 clientes de consumo de tipo comercial e industrial para días laborables de los años (2018-2019), posteriormente estos datos debido a la gran dimensionalidad serán almacenados en una hoja de Excel de tipo .xlsx, con el tipo de día y consumo en un intervalo de 10 minutos dando un total de 144 variables en el dominio del tiempo, posteriormente la base de datos será exportado a Python 3.7. Con la finalidad de ordenar y limpiar los datos se realizará el pre procesamiento de los mismos, que consiste en la limpieza de datos faltantes o no censados, mediante la imputación de datos, este proceso se lo realizará con el uso de la biblioteca sklearn.impute, este algoritmo predecirá los datos atípicos y procederá a realizar una predicción de los datos faltantes con la finalidad de que los datos estén listos para ser tratados. Una vez que la base de datos históricos (data) se encuentra en el entorno de Python y totalmente limpia, se procederá a la clasificación de los usuarios comerciales e industriales por el método de K-means con la utilización de la librería sklearn.cluster con la finalidad de clasificar por estacionalidad invierno y verano, en esta etapa la cantidad de datos manejada es de una alta dimensionalidad, por lo que se usará minería de datos para encontrar un equivalente mediante el uso de las PCA utilizando la librería sklearn.decomposition simplificando el tiempo de ejecución en Python, este proceso generará una matriz equivalente que se encontrará en el dominio de las PCA, esta técnica de reducción de datos proporcionará un uso de datos flexibles y manipulables en el dominio de las PCA, posteriormente esta matriz será introducida en el entorno de Matlab.

Una vez procesados los datos, la matriz simplificada y en el dominio de las PCA será importada a Matlab, el algoritmo a implementarse será Fuzzy-c-meams utilizando la librería [centers,U]=fcm(data,Nc), esto para dar una fiabilidad al proceso que realizará una clasificación por lógica difusa con su respectiva función de pertenencia, el algoritmo evaluará los datos ingresados, entregando las curvas de consumo por estacionalidad invierno y verano, las mismas que luego serán normalizadas por patrón de consumo para ser ingresadas en el sistema ADMS de la EEASA.

Adicionalmente para validar el número óptimo de conglomeración se utilizará el índice de validez Davies-Bouldien que considera el factor distancia entre los centros de cada cluster,

este índice utilizará la librería de Matlab evalclusters. Para finalizar, este proyecto se entregarán las curvas de consumo como producto final y de esta manera mejorar el estimador de estado en el sistema ADMS.

1.4 MARCO TEÓRICO

El estudio presenta las definiciones y principales conceptos de manera coherente para realizar el presente Trabajo de Titulación. En primer lugar, se hace referencia al sistema de lectura automatizada de medidores. Después, se describe al sistema de gerenciamiento avanzado de la distribución ADMs. Luego, se describe las características de la demanda y los principales factores que afectan su comportamiento. Luego, se aborda el estudio de las técnicas de análisis muy utilizadas en el dominio de la clasificación de perfiles de carga y se presenta posibles soluciones para el procesamiento y tratamiento de datos atípicos. También se describe y se presenta el índice de agrupación para escoger el número óptimo de conglomeraciones. Por último, se hace una introducción al lenguaje de programación Python y al aprendizaje automático, describiendo y analizando sus diferentes algoritmos utilizados para diferentes tareas para la obtención de las curvas de demanda.

1.4.1 SISTEMA DE LECTURA AUTOMATIZADA DE MEDIDORES

Los componentes necesarios para realizar la clasificación del perfil de carga se ilustran en la Figura 1.1, los datos recopilados a través de dispositivos AMR (Automatic Meter Reading) se transfieren de forma remota a una base de datos central para el almacenamiento. La gestión de usuarios se ocupa de los datos recopilados durante al menos un año para realizar una clasificación con la finalidad de obtener perfiles de carga representativos de todos los clientes que se encuentran dentro del área de concesión de la empresa distribuidora, de esta manera implementar nuevos estudios y aplicaciones que ayuden a mejorar los perfiles de carga en base a los resultados de clasificación [2].

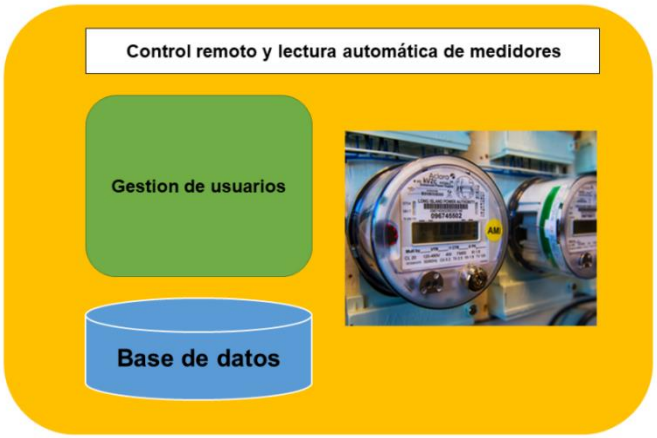


FIGURA 1.1 Componentes para la clasificación de un perfil de carga [2].

1.4.1.1 Sistema de telemedición de la EEASA

Los sistemas de telemedición tienen gran importancia, dada la necesidad de obtener datos de potencia activa, reactiva, aparente, la energía de demanda máxima, curvas de carga, etc. en tiempo real. También, proporciona la información de la demanda diaria de las cámaras de transformación instaladas, con el fin de utilizar la información para una serie de análisis, estudios y aplicaciones. Por este motivo, los sistemas de telemedición se han implementado a nivel mundial para facilitar el trabajo en las empresas distribuidoras de energía eléctrica. Este nuevo sistema representa para las empresas mayor productividad, mejora del servicio, reducción de pérdidas de información y ahorro de gastos. En Ecuador, ha contribuido de manera significa a las empresas de distribución en los procesos de captación, facturación, reconexión, comercialización y corte de la energía eléctrica [36].

Es así, como la Empresa Eléctrica Ambato (EEASA) cuenta con un sistema de telemedición como uno de sus objetivos estrategicos, con motivo de incrementar la innovación con el fin de alcanzar la satisfación de los clientes a través de un servicio de calidad, eficiente, con eficacia operativa y energetica. Por tal motivo, se mantiene contratos para la prestación de servicios de telemedición en clientes con facturación especial [36].

1.4.2 SISTEMA DE GERENCIAMIENTO AVANZADO DE LA DISTRIBUCIÓN, ADMS

El sistema enfocado en el manejo eficiente del control del sistema de distribución eléctrica en tiempo real es el ADMS (Advanced Distribution Management System) que proporciona la solución de gestión en redes eléctricas el mismo que lleva consigo herramientas como: el análisis, monitoreo, planificación y optimización. Esta herramienta esta unificada con el

DMS (Distribution Management System), OMS (Outage Management System), MWM (Mobile Worforce Management System) y SCADA, todo en un solo sistema, por lo tanto, con relación al costo es relativamente bajo presentando un único modelo, lo que implica reducir el tiempo de mantenimiento, una de las ventajas es facilitar su manejo a los operadores y también se debe considerar su alta confiabilidad. La plataforma proporciona un informe de violación de límites, que permite conocer los elementos que se encuentran sobrecargados o por encima de los valores ingresados, una sobrecarga se puede producir por dos razones: inadecuado ingreso de información o sobredimensionamiento del elemento, lo cual provocaría que los fusibles exploten y el accionamiento de alarmas. Las diferentes funciones y herramientas antes mencionadas del ADMS, pueden maximizar sus beneficios en un gran número de dispositivos, redes inteligentes, medición avanzada y energía renovable distribuida. En el Ecuador los siguientes módulos son los que se encuentran en operatividad como se puede apreciar en la Figura 1.2.

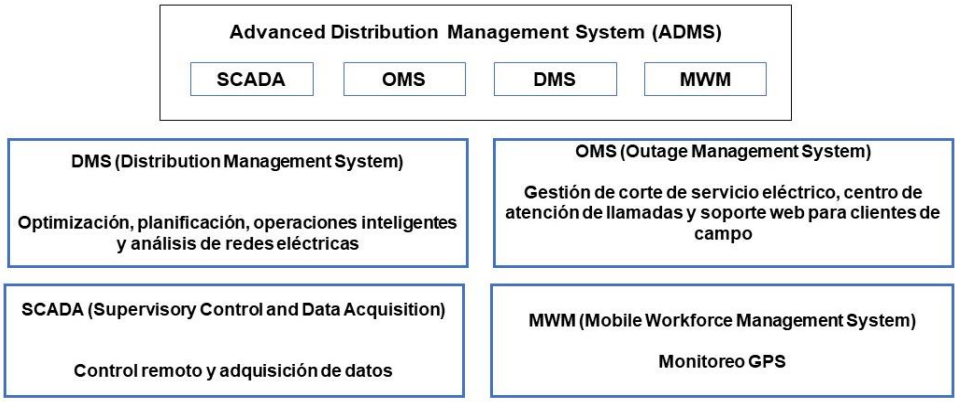


FIGURA 1.2 Módulos del sistema AMDS utilizado en Ecuador [Elaboración propia].

1.4.2.1 SCADA (Supervisory Control and Data Acquisition)

La aplicación SCADA se encarga del control remoto, maniobras y la adquisición de datos en tiempo real, también de los equipos de corte, es un sistema para las actividades operativas, mejora la disposición de los operadores, proporciona toda la funcionalidad necesaria para mejorar la gestión de las redes de distribución de subtransmisión, media y baja tensión. Las potentes funciones de operación que presenta el sistema facilitan la operación diaria segura y eficiente [4]. En la Figura 1.3 se aprecia los módulos que forman parte del SCADA.

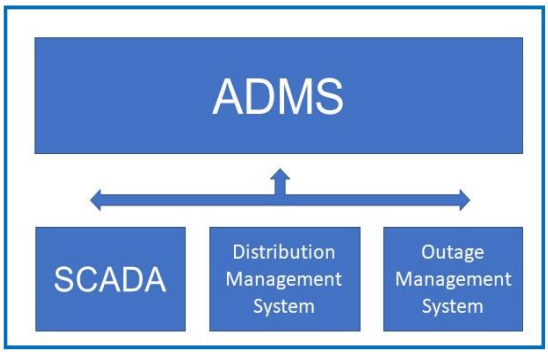


FIGURA 1.3 Centro de control SCADA como parte de los módulos del ADMS [Elaboración propia].

1.4.2.2 OMS (Outage Management System)

Esta herramienta es una ayuda para el operador, permite a los operadores del Centro de Control actuar adecuadamente a los eventos de interrupción detectados por medio del SCADA, PMUs, reconectadores, sectores de falla, datos AMI (Advanced Metering Infraestructure) que provienen de la plataforma de medición de energía. También su función es proporcionar información de usuarios del servicio eléctrico que se reportan mediante la función (Trouble Call System) que se realiza por llamadas o usuarios por mantenimiento programado. En la Figura 1.4, se muestra el navegador de ordenes de trabajo [5].

Navegador de Solicitud de Trabajo

Filtro de visualización

Id	Estado	Alimentador actual	Lugar de trabajo	Plan de maniobras	Propósito	Tipo
WR 316004445	Guardado	05_22_2223	EERCS_UT_9812	SP 199049927	Aplomar Poste No. 505385 se encuentra inclinado con redes en baja tensión.	
WR 316004443	Guardado	05_05_0525	EERCS_UT_18664	SP 199049944	ENSANCHAMIENTO DE VIA	
WR 316004441	Guardado	05_03_0325	EERCS_UT_28330, EERC	SP 199049928	Reubicación de poste con redes en M.V	
WR 316004438	Guardado	05_08_0824	EERCS_UT_29071	SP 199049945	Construcción de redes eléctricas.	
WR 316004436	Guardado	05_08_0824	EERCS_UT_19440	SP 199049940	Construcción de redes eléctricas.	
WR 316004435	Guardado	05_03_0321	EERCS_UT_638	SP 199049942	MONTAJE Y CAMBIO DE TRANSFORMADORES DE DISTRICUCION	
WR 316004433	Guardado	05_23_2312	EERCS_UT_11804	SP 199049917	cierre de puentes en baja tensión para enlazar redes nuevas	
WR 316004432	Guardado	05_23_2312	EERCS_UT_11245	SP 199049901	cerrar puentes de baja tensión para enlazar redes nuevas	
WR 316004428	Guardado	05_23_2312	EERCS_UT_11515	SP 199049904	Tendido de conductor a postes nuevos	
WR 316004425	Guardado	05_08_0824	EERCS_UT_29071	SP 199049924	Construcción de redes eléctricas.	
WR 316004424	Guardado	05_08_0824	EERCS_UT_19440	SP 199049914	Construcción de redes eléctricas.	
WR 316004423	Guardado	05_05_0524	EERCS_UT_2226	SP 199049899	MONTAJE DE ESTRUCTURA CR	
WR 316004422	Guardado	05_07_0723	EERCS_SF_9342		Repotenciación del sistema eléctrico	
WR 316004417	Guardado	05_23_2311	EERCS_SF_14311		CAMBIO DE POSTE EN ENLACE DE ALIMENTADORES 2311	
WR 316004404	Guardado		T2		Construcción de la oficinas y vestidores de la línea energizada, correspondiente a la	
WR 316004401	Guardado	05_15_1521	EERCS_SF_41941	SP 199049198	MONTAJE DE TRANSFORMADOR.	

FIGURA 1.4 Entorno del navegador OMS *[5].*

1.4.2.3 DMS (Distribution Management System)

El sistema general de distribución es la parte primordial e inteligente del software que posee algoritmos permitiendo un desarrollo eficaz, esta herramienta se encarga de la toma de decisiones y operación, englobando el equipamiento instalado en la red, también se encarga del análisis para la operación del sistema eléctrico. La función presta una versatilidad en modo estudio como en tiempo real, esta aplicación es el corazón del ADMS que hace posible la realización de toda la asignación de las tareas técnicas. En la Figura 1.5 se puede apreciar como el DMS es parte del entorno de producción [3] [1].

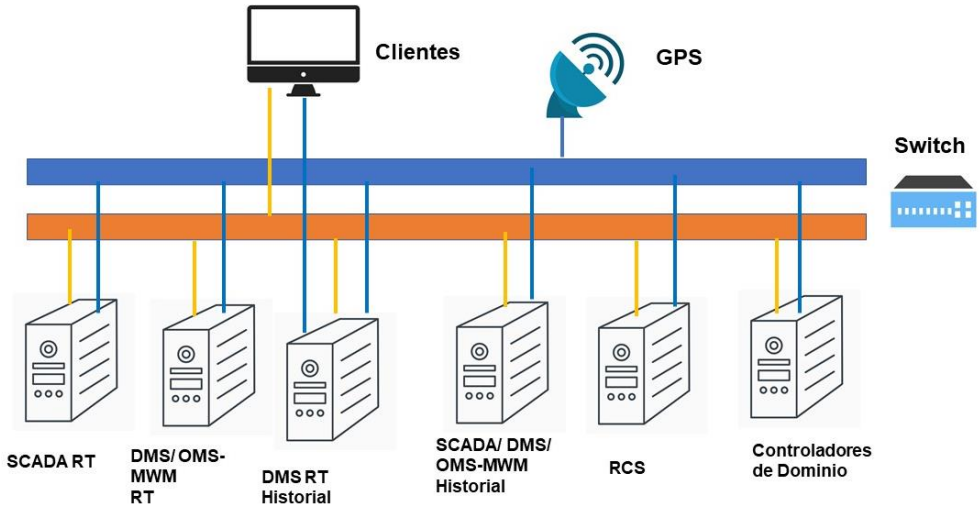


FIGURA 1.5 Entorno de producción [Elaboración propia].

1.4.2.4 MWM (Mobile Worfoce Management)

Esta herramienta minimiza la interrupción al permitir una detección y restauración más rápida a través de una mejor conciencia situacional, automatización y uso efectivo de las cuadrillas de campo. Además, recibe los datos de la ubicación de los vehículos mediante GPS.

1.4.3 CARACTERÍSTICA DE LA DEMANDA

1.4.3.1 Demanda

La Demanda de un sistema o instalación es la medida de potencia que un consumidor en cualquier instante de tiempo (variable tiempo), el consumo de energía eléctrica son funciones no lineales en el tiempo y presentan distintos valores en diferentes puntos

geográficos dentro de la red, esto debido a la naturaleza de los varios usuarios como son: industriales, comerciales y residenciales, cabe destacar que la demanda se puede expresar en kVA, kVAR o kW, por lo tanto, el concepto de demanda es el energía eléctrica consumida E_i en un periodo de tiempo T_i, que se aprecia en la Ecuación 1.1.

$$E_I = \int_0^T P(t)dt \qquad \rightarrow \qquad D_I = \frac{E_i}{T_i} \tag{1.1}$$

En la demanda de potencia los factores asociados están directamente relacionados con los cambios como es la actividad de los clientes que pueden ser industriales, comerciales y residenciales, el comportamiento social, las condiciones climáticas y las conexiones eléctricas conectadas a la red. La demanda eléctrica es alta en el transcurso del día mientras que al anochecer las cargas industriales presentan una demanda alta de potencia, el uso de luminarias es inminente en el horario de la noche a comparación que en el día la luz natural tiende a dominar el panorama, también las personas se encuentran en sus trabajos y la demanda es menor; por lo tanto, la demanda en horas de la tarde hasta antes del mediodía es bajo. Los intervalos en los cuales la demanda es medida o los intervalos T_i en los cuales se registra los datos de facturación a los usuarios son: 60, 30 o 15 minutos [6].

1.4.3.2 Curva de demanda eléctrica

Para la curva de demanda lo más importante es contar con una base de datos históricos (data), los mismos que garantizaran los resultados y proporcionan información sobre el comportamiento de la curva en periodos pasados, con la base de datos mencionada, se analiza el comportamiento de los usuarios. Estas mediciones son obtenidas a través de mediciones con equipos apropiados instalados por la empresa de energía. La forma de la curva de demanda depende de la estacionalidad, factores climáticos y si es una carga de tipo residencial, industrial, comercial y del día de la semana. El objetivo más importante es identificar los eventos que generan irregularidades en el comportamiento que provoca la curva generada por la demanda. La curva se genera de la variación de las demandas en un periodo de tiempo que generalmente se lo representa gráficamente [7].

1.4.3.3 Clases de consumidores de la demanda eléctrica

El producto final hacia donde se destina la energía eléctrica tiene un propósito que es clasificar las cargas, como se puede observar en la Figura 1.6, las empresas distribuidoras a nivel macro en el país clasifican su suministro a menudo en diferentes usuarios que son de tipo:

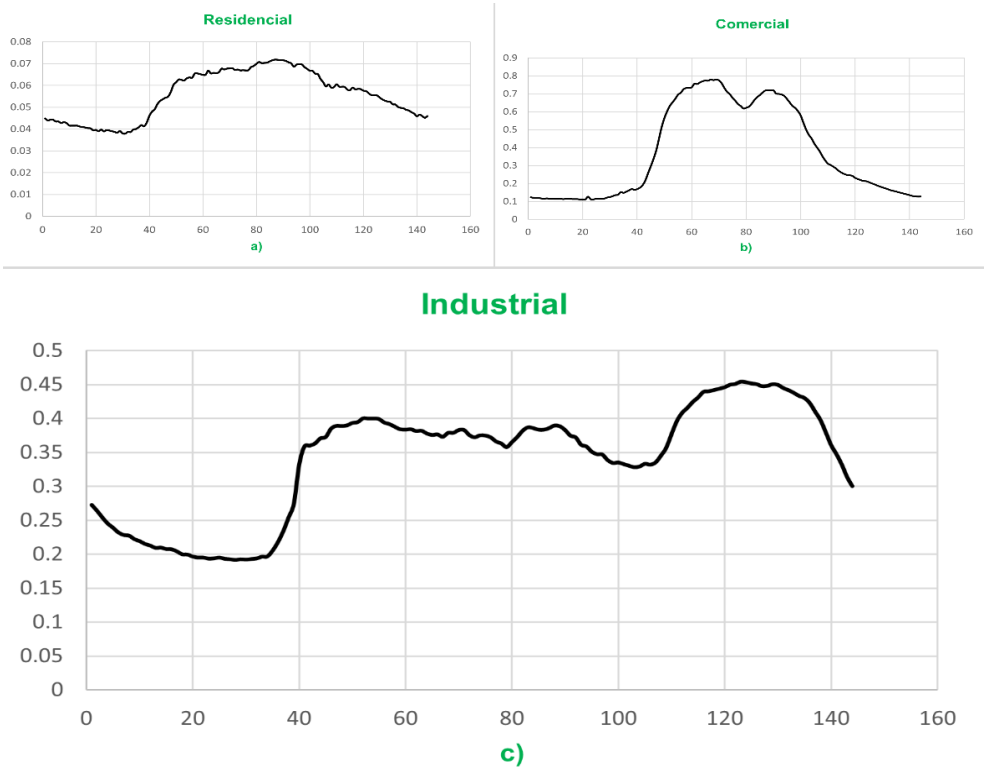


FIGURA 1.6 Curvas de demanda para diferentes consumidores del sector eléctrico [Elaboración propia].

a) Usuarios residenciales: su uso es frecuente en iluminación como en artefactos de consumo eléctrico en el horario nocturno que comprenden urbanizaciones, apartamentos, hogares domésticos su comportamiento es idéntico en la demanda pico y media, en donde la principal característica de este tipo cargas son resistivas (calefacción y alumbrado) y aparatos electrodomésticos de característica reactiva. [2] [3].

b) Usuarios comerciales: el consumo eléctrico en las horas de la mañana es constante, mientras que por la tarde estos usuarios al presentar un horario comercial muestran un descenso aproximadamente en las 13:00 horas, estas

cargas reducen el factor de potencia, en la actualidad predominan las cargas sensibles, estas cargas provocan que se introduzcan armónicos a la red.

c) **Usuarios industriales:** la demanda es proporcional a su nivel de producción, los servicios que se encuentran en la sociedad como los motores instalados están incluidos dentro de este sector. El control a estas cargas del consumo de reactivos es permanente y una gestión de carga por su tarifa que es (baja y alta), esto para evitar que su pico máximo de la curva de consumo coincida con el de la curva de carga residencial

1.4.3.4 Factores que afectan la curva de demanda

La evolución de la demanda eléctrica en el transcurso del tiempo ha identificado los tipos de variables de causa que pueden ser controlables y no controlables, que afectan a la curva de demanda, el patrón que determina la demanda eléctrica es muy similar a nivel nacional [6] y [8]

- Factores Controlables

  - Tarifica eléctrica

  - Voltaje

  - Potencia

  - Frecuencia

Para el control de la demanda de los grandes usuarios los factores controlables son varios identificándolos con el mayor consumo de las horas pico que se encuentren en las zonas del valle ocasionado que el factor de carga del sistema mejore, desencadenando el mejoramiento en las unidades generadoras, estos factores se relacionan directamente con el Ente Regulador, las decisiones políticas, la operación en tiempo real encargado por el Centro de Despacho [7].

- Factores no Controlables

  - Población

  - Condiciones meteorológicas (viento, nubosidad, temperatura, pluviosidad, etc.)

- Información económica (PIB e ingreso per cápita)
- Paros y huelgas generales
- Eventos fortuitos y especiales
- El calendario (hora del día, día de la semana, fiestas y fin de semana)
- Tipos de consumidores (comercial, industrial, residencial, publico, agricultor, etc.)
- Cierres de instalaciones industriales.
- Pandemias mundiales y fenómenos naturales.

Otro de los factores que afectan la curva de demanda en relación directa con el objetivo de cambiar la forma de la curva en ciertos periodos de tiempo, es la gestión de carga del lado de la demanda que afecta al comportamiento de los clientes, basado en cambiar el consumo sustentado en los precios de la electricidad y otros incentivos. El manejo de la carga consiste en el control directo con él envió de señales al cliente, por ejemplo, aire acondicionado, calentador de agua y encendido, estas señales pueden apagar o encender los electrodomésticos de esta manera modificando la demanda mientras que en el área industrial y comercial la empresa de servicios públicos envía señales al cliente, con la información sobre el momento en que necesita la reducción de la carga y ofrece beneficios como descuento de precio [9].

1.4.4 PREPROCESAMIENTO DE DATOS

El preprocesamiento de datos de consumo de energía puede implicar algunos errores e inconsistencias, como la precisión del equipo, la falla de comunicación entre el equipo y el conjunto de datos, la conversión y el almacenamiento de datos. Estos problemas pueden causar valores atípicos, valores duplicados y valores inexistentes. Para minimizar estos errores, se aplica diferentes fases de procesamiento previo al conjunto de datos. Los datos deben convertirse en una unidad en común. Después, los valores que no existen se corrigen usando algunas técnicas como la regresión. Los valores atípicos se pueden eliminar mediante el uso de técnicas de agrupación o análisis estadístico y, al evitarlos, es posible tener una mejor interpretación de los datos y evitar tendencias incorrectas. La reducción de datos es necesaria para disminuir la cantidad de datos y mejorar el análisis de minería de datos para encontrar los patrones relevantes en el conjunto de datos. La

transformación de datos es fundamental en el análisis, ya que se pretende comparar patrones y no cantidad de consumo de energía, por lo que es necesario transformar los datos para comparar los patrones. Para transformar los datos, se pueden aplicar normalizaciones como min-max. Otras normalizaciones que se pueden aplicar a los datos es la división de todos los datos por el valor máximo para tener una escala similar a la normalización min-max, pero en esta conversión, el valor inferior no es cero (a menos que haya valores iguales a cero en la escala), pero algo cercano a cero. Otra normalización es la división de los datos por el promedio del conjunto de datos, obteniendo un valor por unidad de la serie. Esta fase es la más importante y es probablemente la fase de expansión que requiere más tiempo y esfuerzo. Si esta fase no se realiza correctamente, puede conducir al descubrimiento de patrones erróneos, lo que resulta en errores en la interpretación del problema [31]. En la Figura 1.7 se puede observar que el proceso consta de varias fases que son explicadas en los siguientes apartados.

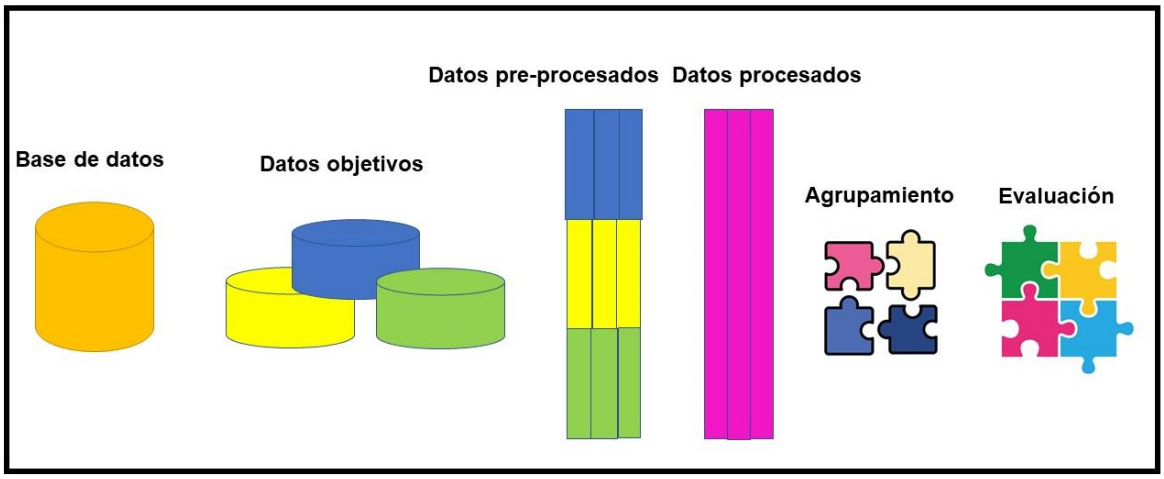


FIGURA 1.7 Proceso de análisis y clasificación de datos [Elaboración propia].

Una de las fases importantes es la minería de datos, después de seleccionar y tratar los datos, los algoritmos de minería de datos se pueden aplicar para descubrir patrones en el conjunto de datos. Se pueden aplicar diferentes técnicas para extraer la información relevante. Las técnicas más utilizadas son el agrupamiento, la clasificación y las reglas de asociación.

1.4.4.1 Preparación de datos.

La preparación de datos es un paso crucial antes de la agrupación y clasificación, implica identificar los datos de entrada a la herramienta computacional, sobre los que operarán las técnicas implementadas, deben estar codificados en un formato específico para luego no

lidiar con los valores de atributos faltantes. La calidad de los datos debe mantener las siguientes métricas: exactitud, precisión, consistencia, completitud, puntualidad, credibilidad e interpretabilidad [10].

1.4.4.2 Selección Y Muestreo

En esta etapa el enfoque es en la selección de variables relevantes en los datos, las variables relevantes que aportaran para el estudio de la base de datos, una vez que las variables hayan sido seleccionadas se aplican técnicas de muestreo, con el objetivo de obtener una muestra de datos que sea lo suficientemente representativa de la población. Una de las ventajas del muestreo es que permite inferir las características o propiedades de la población con un error acotable o medible [10].

1.4.4.3 Exploración

Una vez estructurada la base de datos esta proviene de diferentes fuentes, su exploración se debe ejecutar mediante técnicas de análisis exploratorio para la identificación de valores inusuales, valores desaparecidos, valores de tipo “nan”, discontinuidades, peculiaridades o valores externos, como resultado de la fase de exploración se obtiene si las técnicas de análisis de datos son adecuadas. La fase exploratoria examina las variables individuales y su relación las técnicas de análisis de datos, examinar las variables individuales y su relación, evaluar problemas en la recolección de datos e investigación, es decir, afirma si el formato esta adecuado y listo para su manipulación en el entorno a aplicarse. Además, si son adecuadas las técnicas de análisis de datos, es necesario realizar un análisis previo de la información de que se dispone antes del uso de cualquier técnica. La exploración puede indicar la necesidad de transformar los datos, si la técnica necesita una distribución normal o si se necesita utilizar pruebas no paramétricas, para el análisis exploratorio se cuenta con técnicas formales, técnicas graficas o visuales [10].

1.4.4.4 Valores perdidos

Los valores perdidos se producen debido a errores humanos o del dispositivo que recopila la información, también puede ser al transferir o procesar datos, lo que genera espacios vacíos, que producen problemas en el análisis posterior en un conjunto de datos. El objetivo es obtener un conjunto de datos completos para analizar por la vía de los métodos estadísticos tradicionales, sin embargo, este análisis se complica cuando una matriz de datos está formada por diversas variables sobre la cual se realizan estudios multivariables,

haciéndose necesario la aplicación de métodos estadísticos. Para corregir los valores faltantes se reemplaza por la media truncada de las instancias restantes. A continuación, las Ecuaciones. 1.2 y 1.3

La media truncada de recorte implica recortar P por ciento de observaciones de ambos extremos.

Dadas las observaciones, X_i:

$$n = n\acute{u}mero\ de\ observaciones$$

$$Xi\ desde\ el\ m\acute{a}s\ pequeño\ al\ mas\ grande$$

$$p = \frac{P}{100}; proporción\ recortada \qquad \textbf{(1.2)}$$

$$Calcular\ np$$

$$Si\ np\ es\ un\ uso\ entero\ k = np\ y\ recortar\ k\ observaciones\ en\ ambos\ extremos$$

$$R = observaciones\ restantes = n - 2k$$

$$Media\ recortada = (\frac{1}{R})(X_{k+1} + X_{k+2} + \cdots + X_{n-k}) \qquad \textbf{(1.3)}$$

La media truncada toma el mayor número de datos eliminando las colas de la serie de datos, es decir, eliminando un porcentaje de datos de los extremos, mayores y menores, el porcentaje a eliminar puede ser hasta del 25% en cada uno de los extremos, el objetivo es eliminar los datos que afecten a la media y de esta manera poder calcular una medida de tendencia central con la mayor cantidad de información posible, llegando a la conclusión de que la media truncada es aproximadamente igual a la media aritmética que su característica indica un leve sesgo en la distribución [12].

1.4.4.5 Limpieza

Esta fase es consecuencia del proceso de exploración, el conjunto de datos puede conllevar valores faltantes, atípicos y/o erróneos, la limpieza de los datos es importante para el proceso de exportación al entorno en donde serán tratados cada uno de los datos, esta parte es importante para solventar cualquiera de los inconvenientes en la etapa de ejecución [10].

1.4.4.6 Transformación

Una vez que la base de datos se encuentra sin ningún tipo de dato atípico, lo siguiente que se procede a realizar es transformar los datos, en tipo numérico de tipo Python que permite representar grandes valores de una mejor manera del tipo int de Python. Este tipo de transformación no tiene un tamaño fijo, se puede expandir al valor a tomar, el único límite es la cantidad de memoria de la que dispone el ordenador [11]. La transformación de los datos es necesaria debido a que los datos se encuentran con comas y al pasar a Python se deben transformar a tipo float que almacena números decimales. También se considera que los decimales se escriben con punto en lugar de comas cuando entre las variables existen diferentes escalas o existen demasiadas o pocas variables, entonces se realiza una normalización o una estandarización de los datos mediante técnicas de reducción o aumento de la dimensión, así como el escalamiento simple o multidimensional.

Si en el análisis exploratorio se indica la necesidad de transformar algunas variables, se podrán aplicar algunas de estas cuatro transformaciones:

- Transformaciones lógicas
- Transformaciones lineales
- Transformaciones algebraicas
- Transformaciones no lineales

Estas fases mencionadas en los puntos anteriores constituyen el proceso de pre-procesamiento de los datos. Cuando se tienen datos faltantes, hay varias estrategias que se pueden seguir, sobre todo si se ha comprobado la aleatoriedad de dichos datos ausentes. Una estrategia es usar el método de aproximación de casos completos que consiste en incluir en el análisis solo los casos con datos completos. La alternativa a los métodos de eliminación de datos es la imputación de la información faltante, donde el objetivo es estimar los valores ausentes basados en valores válidos de otras variables o casos. El método de imputación por regresión utiliza el método de la regresión para calcular o estimar los valores ausentes basados en su relación con otras variables del conjunto de datos [10].

1.4.5 NORMALIZACIÓN

La Normalización de valores de atributos es una técnica para diseñar la estructura lógica de los datos de un sistema de información en el modelo relacional. Cada paso de la normalización consiste en reducir sistemáticamente una relación o tabla en una colección de tablas más pequeñas que son equivalentes a la de partida. Uno de los métodos de normalización utilizados en el presente trabajo es la normalización mínima y máxima de un vector que se calcula a continuación en las Ecuaciones 1.4 y 1.5

El método consiste en mantener la proporcionalidad $\frac{V_i}{V_k} = \frac{v_i}{v_k}$ **(1.4)** para todo i y k, con la expresión $v_i = \frac{V_i}{Máxima\, V_i}$ **(1.5)** usando el criterio de maximizar, en la notación se puede llamar $(V_1, \dots, V_i \dots, V_n)$ al vector con variables alternativas correspondiente al grupo de vectores genéricos, el cual puede ser transformado al vector normalizado de una variable $(v_1, \dots, v_i \dots, v_m)$ en escala normalizada $0 < v_i \leq 1$, en donde se supone que $V_i \geq 0$ para todo i. [15].

1.4.6 LENGUAJE DE PROGRAMACIÓN PYTHON

Python fundado por Guadi van Rossum cuyo origen se remontan a la década de 1990 para actividades de propósito general, permitiendo escribir código en diferentes estilos de programación imperativa, una de las características más importantes es el multiparadigma: Programación estructurada, Programación Orientada a Objetos y Programación Funcional, con una sintaxis fácil de ser interpretada como también de fácil lectura. A pesar de que Python es un lenguaje de programación tipo script, su popularidad e implantación de códigos es alta entre los científicos de datos y los ingenieros de Machine Learning, posee tipado dinámico esto quiere decir que una variable puede poseer datos de varios tipos [17]. Su estructura se puede observar en la Figura 1.8.

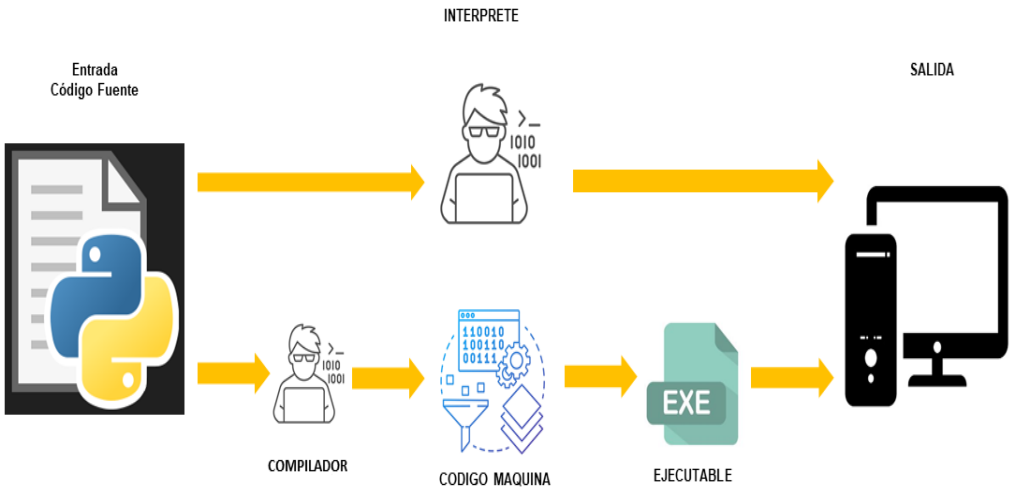


FIGURA 1.8 Mecanismo de ejecución Python *[17].*

En los últimos años, el soporte de biblioteca de Python principalmente Pandas uno de los módulos que posee para tareas como también el módulo Numpy, se han convertido en una sólida alternativa para las tareas de manipulación de datos numéricos, archivos Excel y lectura, operación y creación de matrices con arreglos vectoriales. Además, esta incorporado una amplia gama de algoritmos de aprendizaje automático (Machine Learning) que cada vez mejoran las tecnologías y por consiguiente se desarrollan nuevos algoritmos en la biblioteca scikit-learn [30].

1.4.6.1 Numpy

Una herramienta en la que está basada Scipy, es una colección de algoritmos matemáticos que proporcionan al usuario comandos y clases de alto nivel para la visualización de datos y manipulación, las funciones de Numpy cubren las necesidades básicas para el manejo de matrices y vectores, es el paquete fundamental de Python para la información científica, proporciona herramientas para trabajar con arreglos y entrega una matriz multidimensional objeto de alto rendimiento [19].

1.4.6.2 Scipy

Scipy es una biblioteca de rutinas numéricas para el lenguaje de programación “Python”, que proporciona bloques de construcción fundamentales para modelar y resolver problemas científicos. El equipo de desarrollo como la comunidad actualmente interactúan y operan principalmente en GitHub, una plataforma de control de versiones y

administración de tareas en línea. Más de 110.000 repositorios de GitHub y 6.500 paquetes dependen de Scipy [20].

1.4.6.3 Pandas

La biblioteca proporciona estructuras de datos consistentes y funciones diseñadas para que el trabajo estructurado sea expresivo, rápido y fácil. Esta librería contribuye a que Python sea un entorno de análisis de datos productivo y potente. El objetivo principal de Pandas es el manejo de objetos tipo DataFrame, una estructura de datos bidimensional orientada a columnas con etiquetas de filas y columnas, este puede verse como una hoja de cálculo, en la cual permite realizar transformaciones en distintas formas sea eliminando columnas o filas o a su vez modificando [7].

1.4.6.4 Scikit-learn

Es una biblioteca de aprendizaje de maquina en Python, ofrece herramientas con un alto grado de eficiencia y su simplicidad para tareas comunes en el análisis de datos, tales como, agrupamiento (clustering), clasificación, regresión entre otros. Además, la librería basa su código fuente en las librerías propias de Python como: Scipy, matplotlib, Numpy de esta manera y al igual que toda la comunidad aportante de Python es de código abierto y utilizable comercialmente [18].

1.4.6.5 Método Fit

El método fit va de la mano con el uso de scikit-learn, está diseñado para ajustar los parámetros de un conjunto de datos con rapidez y frecuencia. El módulo está diseñado para aprender de los datos ingresados y predecir nuevas observaciones, este trata de hacer que los datos se ajusten y sean simples [21].

1.4.6.6 Simpleimputer

Los conjuntos de datos comúnmente tienen valores perdidos o faltantes, estos conjuntos de datos son incompatibles con los estimadores de scikit-learn, que asumen que todos los valores de datos en una matriz son numéricos, los valores faltantes se presentan como espacios en blanco, marcadores de posición u “nan”. La librería SimpleImputer proporciona estrategias básicas para imputar valores perdidos, los valores perdidos se pueden inferir de la parte conocida de los datos proporcionados [16]. Debido a esto los valores se reemplazan por la media truncada de la variable.

1.4.7 MACHINE LEARNING

Machine Learning o Aprendizaje Automático es sinónimo de utilizar algoritmos para que puedan desarrollar técnicas, para entregar información importante en base a un conjunto de datos, que permiten a los computadores predecir el comportamiento sin tener que escribir ningún código en específico aplicando algoritmos para el problema a través de métodos numéricos. Es decir, es la ciencia y la aplicación de algoritmos que proporcionan sentido a los datos, es el campo más apasionante e interesante de todas las ciencias de la computación [16]. En la Figura 1.9. se puede observar el proceso de aprendizaje el mismo que está divido en varias etapas que a continuación son explicadas en los apartados siguientes.

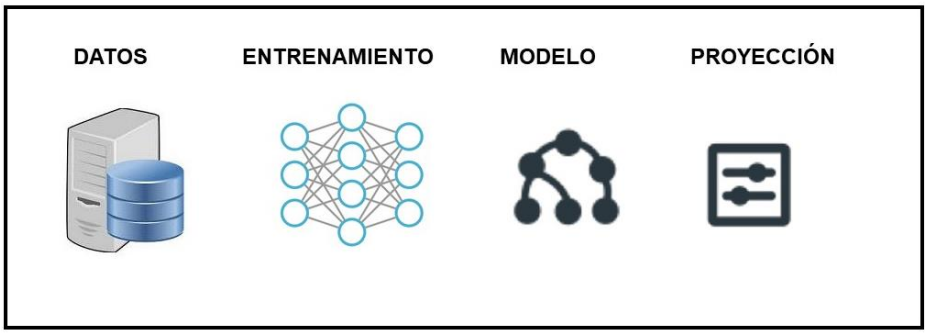


FIGURA 1.9 Esquema general de aprendizaje automático *[7].*

En la actualidad vivimos en una época en la que los datos son abundantes, dado el caso se van utilizando los algoritmos de autoaprendizaje en los tres tipos de aprendizaje automático, como: aprendizaje de refuerzo, aprendizaje supervisado, y aprendizaje no supervisado [16]. En general, combinar los esfuerzos maquina y hombre consiste en que las computadoras realicen el trabajo pesado, y los humanos las tareas creativas. Un sistema de aprendizaje automático requiere un conjunto de entrada de datos, que pueden ser: visuales o audiovisuales, numericos, textuales. Dicho sistema tiene salidas como, un número de coma flotante, por ejemplo, la gravedad de un cuerpo, o el inventario de una empresa; otro tipo de salida puede ser un número entero, una etiqueta categorica, por ejemplo, si una clasficación es buena o mala [7].

1.4.7.1 Aprendizaje No Supervisado

Se denomina “sin supervisión” por qué empieza con datos no etiquetados, es decir, que estos datos no tienen alguna categoría, o no posee alguna descripción adjunta; no siempre

es fácil obtener métricas sobre qué tan bien está funcionando el algoritmo de aprendizaje sin supervisión. El aprendizaje no supervisado se puede utilizar para detectar anomalías en los datos, o para clasificar los datos en grupos que contengan características similares [30][7]

1.4.7.1.1 Agrupación Fuzzy

La teoría de conjuntos difusos (Fuzzy) propuesta por Zadeh en 1965 proporciona una herramienta para el análisis de este trabajo. En general hay dos categorías en el uso de la teoría difusa en el análisis de conglomerados, la primera categoría es la agrupación difusa basada en la matriz de relación como coeficiente de correlación, relación de equivalencia, relación de similitud y relación difusa. Por otro lado, la segunda categoría se basa en funciones objetivas que incluyen FCM [23].

Generalmente el objetivo del análisis por medio de clúster es separar un determinado conjunto de datos u objetos (clases, subconjuntos, grupos). Los objetos se pueden describir como un conjunto de medidas o por relaciones entre objeto. Por lo tanto, puede revelar relaciones y estructura de datos. La mayoría de los algoritmos tradicionales de análisis de conglomerados son nítidos particionando, lo que significa que cada patrón pertenece a un solo racimo. Sin embargo, existen límites estrictos entre los racimos. Además, la mayoría de los objetos tienen atributos ambiguos y puede pertenecer a más de un grupo. [23]

1.4.7.1.2 Fuzzy C-Means

Este método fue originalmente introducido por Bezdek como una mejora de los métodos Fuzzy clustering based on fuzzy relation (FCR), Cosine Amplitud, Max-min Composition, Lambda Cut for Fuzzy relations, uno de los algoritmos para clustering de partición difusa [24], su uso está enfocado en el reconocimiento de patrones, se aplica, en el caso que sea difícil asignar una clasificación de un dato o un grupo de datos que se encuentren cerca de dos clusters. La técnica de minería de datos contribuye a clasificar y manejar grandes cantidades de datos, la técnica FCM se basa en el algoritmo clásico C-Means, en el cual asigna a cada dato un valor de pertenencia dentro de cada cluster, en donde cada punto de datos pertenece a un clúster hasta cierto límite el cual se distingue con una etiqueta, el método asigna a cada dato un grado de pertenencia dentro de cada cluster y por consecuencia un dato puede pertenecer parcialmente a más de un grupo [14] FCM realiza una partición suave y la función objetivo se minimiza en la Ecuación 1.6.

$$J_m = \sum_{i=1}^{N} \sum_{j--=1}^{C} u_{ij}^m \left\| X_i - C_j \right\|^2 \quad \textbf{(1.6)}$$

Donde:

$\mathrm{N} = \text{número de perfil de carga}$

$\mathrm{C} = \text{número de cluster}$

$\mathrm{m} = \text{parametro de ponderación, en general m} = 2$

$\mathrm{u_{ij}} = \text{es el grado de pertencia de } \mathrm{X_i} \text{ en el cluster j}$

$\mathrm{X_i} = \text{es el perfil del i} - \text{esimo alimentador de datos medidos}$

$\mathrm{C_j} = \text{j} - \text{esimo centro del grupo}$

$\|*\| = \text{es cualquier norma que exprese la similitud}$

entre cualquier dato medido y el centro

Se considera un conjunto de N perfiles de carga $X = \{X_1, X_2, \ldots., X_N\}$ para ser agrupados en C clusters $(1 < C < N)$. Los pasos en este algoritmo son los siguientes:

i) Elija C y m, e inicialice la matriz de partición U.

Normalmente el número de cluster en los datos es conocido, de lo contrario es necesario para identificar el valor de C que da el número más razonable de clústeres en los datos. Por otro lado, el parámetro **m** controla la falta de claridad en el proceso de agrupamiento.

La elección del valor del parámetro de ponderación sigue siendo en gran parte heurístico, la mejor elección de **m** es el intervalo [1.5, 2.5] y el intervalo punto medio, m = 2 ha preferido a menudo para muchos usuarios de FCM [14]. Como se aprecia en las Ecuaciones 1.7 y 1.8.

ii) Cálculo del centro de los clusters

$$C_j = \frac{\sum_{i=1}^{N} u_{ij}^m X_i}{\sum_{i=1}^{N} u_{ij}^m} \quad \textbf{(0.7)}$$

iii) Actualizar la matriz para el k-ésimo paso, U:

$$u_{ij} = \frac{1}{\sum_{k+1}^{C} \left(\frac{\left\|x_i - c_j\right\|^2}{\left\|x_i - c_k\right\|^2}\right)^{\frac{2}{m-1}}} \quad \textbf{(0.8)}$$

iv) Si $\left\|U^{(k+1)} - U^{(k)}\right\| < \varepsilon$ entonces el proceso para; de lo contrario vuelve al paso (ii).

En este paso, dos particiones difusas sucesivas son comparadas con un nivel prescrito de precisión, ε para determinar si la solución es consistente y no indeterminada.

El algoritmo comienza con una suposición inicial de los centros de los grupos que probablemente sea inexacta. Posteriormente, se asigna un grado de membresía, es decir, una equivalencia para grupo que pertenece a cada punto de datos. Este proceso se repite mediante la actualización iterativa de los centros de clústeres y las calificaciones de membresía para cada punto de datos. Este proceso se repite mediante la actualización iterativa de los centros de los clústeres y las calificaciones de membresía para cada punto de datos.

Por lo tanto, los centros de los conglomerados se mueven iterativamente a la ubicación correcta dentro del conjunto de datos. Los resultados de FCM son la matriz de membresía final U y los centros de clúster. Las técnicas de lógica difusa permiten manipular los datos de los cuales existe una transición suave entre categorías distintas, es decir, que ciertos datos pueden contar con propiedades de clases diferentes, que pueden pertenecer a más de un grupo de datos con un grado específico de pertenencia [24].

1.4.7.2 Aprendizaje Supervisado

Se comienza el análisis con un conjunto de datos con etiquetas correctamente asociadas, el algoritmo aprenderá la relación entre los datos y sus etiquetas, y aplicará esa relación aprendida para clasificar datos completamente nuevos que la máquina no haya visto antes. El principal objetivo de los algoritmos de este campo es etiquetar nuevos conjuntos de datos que poseen un grupo de salida específico. El termino supervisado se refiere a un conjunto de muestras donde las señales de salida deseadas (etiquetas) ya son conocidas [30][7].

1.4.7.2.1 K-Means

El termino K-means usado por primera vez por MaccQueen en el año de 1967, es un algoritmo no supervisado, utilizado para una gran cantidad de datos sin etiquetar, en el cual los objetos se mueven a través de conjuntos de cluster hasta alcanzar una ajuste deseado

o convergencia. Se denomina k-means debido que cada grupo se representa con la media (o media ponderada de sus puntos), los mismo que son denominados centroides, el objetivo de este algoritmo es reducir la distancia entre cada cluster (k) en lo mínimo posible y su centroide (C), por lo tanto, el centroide es el punto medio de la serie de datos (D) [18].

El algoritmo tiene como objetivo partir de un conjunto de n observaciones en k grupos con características similares o usuarios afines, inicia seleccionando arbitrariamente k objetos en el cual cada observación pertenece al centroide de los k grupos, posteriormente cada elemento es asignado al clúster con su centroide más cercano y se vuelve a calcular la media del grupo, considerando a los nuevos elementos ahora esa media es el nuevo centroide del clúster, nuevamente se calcula la similitud de cada objeto y se asigna al centroide más cercano, recalculando la media y repitiendo el proceso de manera iterativa hasta que se alcance el mínimo deseado de esta manera se logra tener objetos dentro de cada grupo con características similares entre ellos, pero disimiles con elementos de otros grupos. [17]

1.4.8 ÍNDICES DE AGRUPACIÓN

1.4.8.1 Validez del grupo

Uno de los problemas más importantes de la agrupación es decidir el número óptimo de clústeres que se ajusta a un conjunto de datos. Dado que los algoritmos de agrupación en clústeres es un proceso no supervisado, es decir, el número de clusters se descubrirá de forma natural, la partición final de los datos requiere algún tipo de evaluación. La validez de grupo es el término para describir el procedimiento de la evaluación del índice y los índices de validez de grupo se utilizan en la evaluación, para una correcta agrupación.

El número óptimo de clusters es determinado de forma heurística, utilizando índices de validez para los métodos no jerárquicos, con el objetivo de no utilizar denogramas que es una herramienta para los métodos no jerárquicos [25]. El criterio para intuir el número óptimo de clusters depende directamente de la robustez de los datos que serán procesados, existen varios tipos de índices de validez que ayudan a determinar el número adecuado de grupos que se muestran los más utilizados en el análisis de conglomerados en la Figura 1.10.

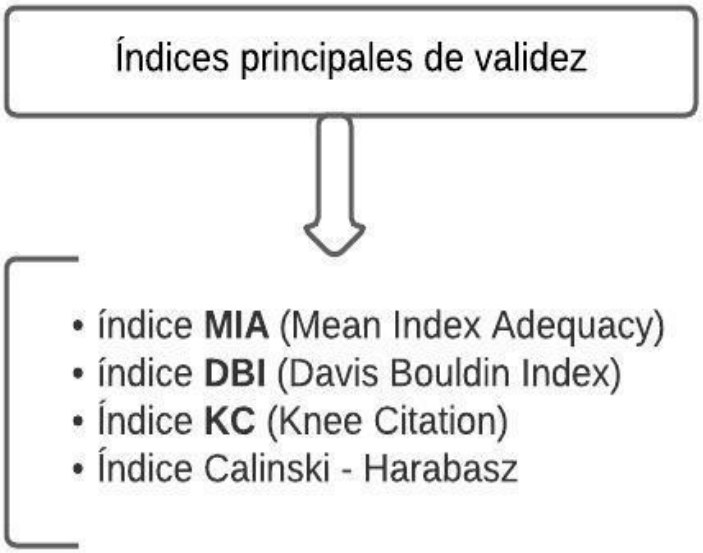


FIGURA 1.10 Grupos más utilizados en el análisis de conglomerados *[25].*

1.4.8.2 Índices Davies Bouldien Index DBI

El uso de este índice compara la distancia de cada individuo del grupo con el centro de su grupo referente, considerando el promedio de aquellas distancias, analizando que coincida la distancia métrica con el esquema de agrupación; entre menor sea el índice DBI mejor será la elección del número de grupos [26]. El índice DBI se define en las siguientes Ecuaciones 1.9 y 1.10.

$$\boldsymbol{DBI = \frac{1}{K}\sum_{\substack{i=1 \\ i=1,\ldots..K \\ i\neq j}}^{K} max\left\{\frac{\sigma_i+\sigma_j}{D_{i,j}}\right\}, K = 2, \ldots. N-1} \quad \textbf{(1.9)}$$

$$\boldsymbol{\sigma_l = (\frac{1}{n^{(k)}}\sum_{v=1}^{n^{(k)}}\left|X^{(v)} - C^{(k)}\right|^2)^{\frac{1}{2}}, k = i, j} \quad \textbf{(0.10)}$$

Donde:

$k = Es\ el\ número\ de\ clústeres$

$\sigma_i = Es\ la\ distancia\ promedio\ entre\ cada\ punto\ en\ el\ clúster\ i\ y\ el\ centroide\ del\ clúster.$

$\sigma_j = Es\ la\ distancia\ promedio\ entre\ cada\ punto\ del\ cluster\ j\ y\ el\ centroide\ del\ clúster.$

$D_{i,j} = Distancia\ entre\ los\ centros\ de\ los\ grupos\ i\ y\ j$

$n^{(k)} = Individuos\ que\ pertenecen\ al\ grupo\ l$

Los valores pequeños para el índice DBI indica clústeres compactos, y cuyos centros se encuentran bien separados los unos de los otros, llegando a la conclusión que el número de clústeres que minimiza el índice DBI se toma como el óptimo

1.4.8.3 Índices de desempeño

Para evaluar la eficiencia de cada algoritmo, se debe utilizar una medida de adecuación. En este trabajo, se emplearon los índices de rendimiento propuestos. Para usar estos índices, considere que el proceso de agrupamiento ha formado K clases de clientes con k=1, K, y cada clase está formada por un subconjunto L (k) de diagramas de carga y r (k) es un patrón asignado a un grupo k [2].

1.4.9 MINERÍA DE DATOS O DATA MINING

La minera de datos se refiere a extraer conocimiento de grandes bases de datos o simplemente extraer patrones inmersos en la base de datos, con el objetivo de determinar patrones que permitan identificar comportamientos característicos de estos datos. Está técnica tiene la capacidad de analizar grandes volúmenes de datos en tiempo real, con el fin de proporcionar datos correctos con el uso de algoritmos computacionales, entre ellos los algoritmos de aprendizaje automático. Los datos recogen un conjunto de hechos (base de datos) y los patrones son expresiones que describen un subconjunto de los datos (un modelo aplicable a ese subconjunto) y por último el descubrimiento del conocimiento que involucra un proceso iterativo e interactivo, tal como se muestra en la Figura 1.11.

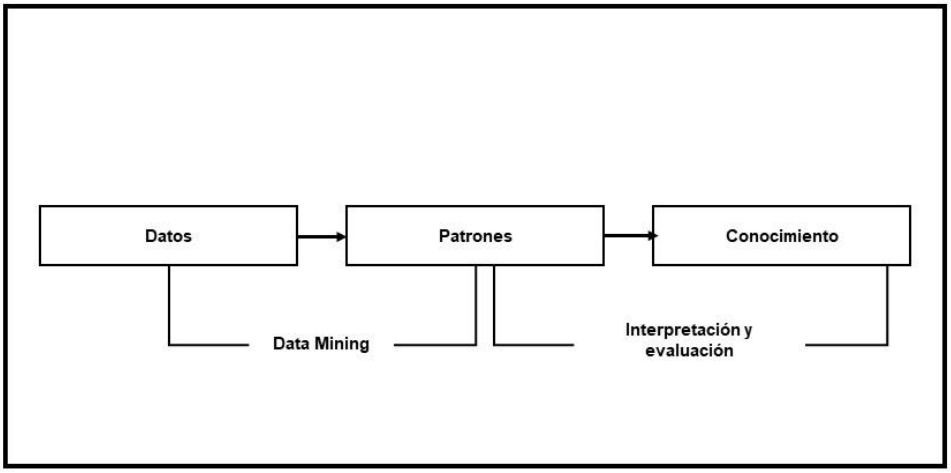


FIGURA 1.11 Data Mining *[propia].*

El objetivo final de todo es incorporar el conocimiento obtenido en algún sistema real y decidir si los modelos obtenidos son útiles o no suele requerir una valoración subjetiva por parte del usuario. Los algoritmos de minería de datos suelen tener tres componentes:

1. El modelo, que contiene parámetros que han de fijarse a partir de los datos de entrada.

2. El criterio de preferencia, que sirve para comparar modelos alternativos.

3. El algoritmo de búsqueda, que viene a ser como cualquier otro programa de inteligencia artificial (IA).

1.4.9.1 Funciones Empíricas Ortogonales (EOF)

El uso de la herramienta EOF es una técnica de minería de datos de tiempo de alta potencia, permiten descomponer una función discreta del tiempo f (t) (como un ángulo de voltaje, magnitud de voltaje o frecuencia), en una suma de un conjunto de funciones de patrones discretos, es decir, que la transformación EOF se utiliza para extraer los componentes individuales más predominantes de una forma de onda de una señal compuesta, esta función permite revelar los principales patrones inmersos en la señal. Uno de los enfoques principales es en el uso del análisis de datos en la ciencia atmosférica espacio – tiempo. Los datos consisten en mediciones de una variable específica, están distribuidos en n ubicaciones espaciales y en p diferentes veces, en donde p representa el intervalo de mediciones y n el número de usuarios. Esta herramienta reducirá la abrupta cantidad de datos, debido a que manejar grandes cantidades de datos no permitirán poder procesarlos y mucho menos poder ejecutar los algoritmos con grandes cantidades de datos, también se debe considerar que al manejar grandes cantidades de datos las respuestas entregadas por el procesamiento de los algoritmos, pueden manejar cierto porcentaje de error, pero al usar esta herramienta con la reducción de datos, el tratamiento de los datos se hace más factible y la deducción de resultados de una más efectiva [25]. A continuación, se presentan las Ecuaciones 1.11 y 1.12.

$$\boldsymbol{F} = \begin{pmatrix} \boldsymbol{f_1(t)} \\ \vdots \\ \boldsymbol{f_n(t)} \end{pmatrix} = \begin{pmatrix} \boldsymbol{X_{11}} & \cdots & \boldsymbol{X_{1p}} \\ \vdots & \ddots & \vdots \\ \boldsymbol{X_{n1}} & \cdots & \boldsymbol{X_{np}} \end{pmatrix} \quad \textbf{(0.11)}$$

La descomposición de valores singulares de una matriz rectangular **F** de dimensiones (n x p) es una factorización de la forma:

$$\boldsymbol{F_{np} = U_{nn}\Lambda_{np}^{\frac{1}{2}}\vee'_{pp}} \quad \textbf{(0.12)}$$

Donde:

U: Es una matriz ortogonal cuyas columnas son los vectores propios ortogonales de FF'

$\vee'$: Es la matriz transpuesta de una matriz ortogonal cuyas columnas son los vectores propios ortogonales FF'

$\Lambda^{\frac{1}{2}}$: Es la matriz diagonal que contiene las raíces cuadradas de los valores propios de **U** o **V** en orden descendente, que se llaman los valores singulares de **F**.

Se establecen las Ecuaciones 1.13 y 1.14.

$$\boldsymbol{F = (U_1 \ldots \ldots U_n)_{nn}} \begin{pmatrix} \boldsymbol{\lambda_1^{\frac{1}{2}}} & \cdots & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \cdots & \boldsymbol{\lambda_p^{\frac{1}{2}}} \\ \mathbf{0} & \cdots & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \cdots & \mathbf{0} \end{pmatrix}_{\boldsymbol{np}} \begin{pmatrix} V_1' \\ \vdots \\ V_p' \end{pmatrix} \quad \textbf{(0.13)}$$

$$\boldsymbol{F = (U_1 \ldots \ldots U_n)_{nn}} \begin{pmatrix} \boldsymbol{\lambda_1^{\frac{1}{2}} V_1'} \\ \vdots \\ \boldsymbol{\lambda_P^{\frac{1}{2}} V_P'} \\ \mathbf{0} \\ \vdots \\ \mathbf{0} \end{pmatrix}_{\boldsymbol{np}} \quad \textbf{(0.14)}$$

Se puede escribir como una suma infinita, como lo muestran las Ecuaciones 1.15 y 1.16 a continuación:

$$\boldsymbol{F = \sum_{i=1}^{p} \lambda_1^{\frac{1}{2}} u_i V_i'} \quad \textbf{(0.15)}$$

$$\boldsymbol{f_k = \left(\sum_{i=1}^{p} \lambda_1^{\frac{1}{2}} u_{ki} v_{1i} + \sum_{i=1}^{p} \lambda_1^{\frac{1}{2}} u_{ki} v_{2i} \ldots . \sum_{i=1}^{p} \lambda_1^{\frac{1}{2}} u_{ki} v_{pi} \right)} \quad \textbf{(0.16)}$$

La anterior serie se puede representar en un formato diferente como lo establece la Ecuación 1.17.

$$\boldsymbol{f_k = \lambda_1^{\frac{1}{2}} u_{k1} V_1 + \lambda_2^{\frac{1}{2}} u_{k2} V_2 + \cdots + \lambda_P^{\frac{1}{2}} u_{kP} V_P} \quad \textbf{(0.17)}$$

La Ecuación mostrada en **(1.12)** representa la descomposición de la función discreta en función del tiempo $\boldsymbol{f}_k$ en una suma de un conjunto de funciones discretas (V_j) que son de naturaleza ortogonal (ya que son los vectores propios ortogonales de F'F), el valor de los coeficientes es el resultado del producto del j-ésimo termino valor singular de **F** por el j-ésimo elemento del vector propio de $\boldsymbol{u}_k$. Por lo tanto, v_j representa el j-esimo termino EOF y el valor del coeficiente es dado por $a_{kj} = \lambda_1^{\frac{1}{2}} \boldsymbol{u}_{kj}$.

Dentro de la herramienta EOF, una de las ventajas posibles es reconstruir la matriz **F** completa (es decir, regresar a los datos originales), utilizando EOF y sus valores correspondientes como la Ecuación 1.18 se establece:

$$\boldsymbol{F} = \sum_{i=1}^{p} \boldsymbol{a_i V_i}' \qquad \textbf{(0.18)}$$

Donde:

$\boldsymbol{a_i}$**:** Es el i-esimo vector cuyos elementos a_{ij}son los valores de toda la serie finita EOF.

Comparando (1.4) y (1.13), se concluye que $a_i = \lambda_i^{1/2} u_i$. Por lo tanto, todos los valores

$\boldsymbol{a_{ij}}$: Pueden ser calculados usando su forma matricial, usando la Ecuación 1.19 de la siguiente manera:

$$\boldsymbol{A_{np}} = \boldsymbol{U_{nn}} \boldsymbol{\Lambda}_{\boldsymbol{np}}^{\frac{1}{2}} \qquad \textbf{(0.19)}$$

Donde:

A: es la matriz de valores EOF

Se puede determinar la Ecuación 1.20.

$$\boldsymbol{F_{np}} = \boldsymbol{A_{np}} \vee'_{\boldsymbol{pp}} \qquad \textbf{(0.20)}$$

Desde la última ecuación, y basado en el hecho de que V es una matriz ortogonal (cuya característica principal consiste en: su transpuesta es igual a su inversa), la matriz A puede calcularse de la siguiente manera como lo indica la Ecuación 1.21.

$$\boldsymbol{A_{np}} = \boldsymbol{F_{np}} \vee_{\boldsymbol{pp}} \qquad \textbf{(0.21)}$$

Donde la matriz v contiene las componentes de EOF de **F** (es decir, los vectores propios de **F'F**)

La suma de los valores singulares de **F** ($\lambda_i^{1/2}$) es equivalente a la varianza total de la matriz de datos y cada valor singular ofrece una medida correspondiente de la variabilidad (EV). Entonces, el número de EOF depende de la variabilidad; así lo establece la Ecuación 1.22.

$$\boldsymbol{E \vee_i = \frac{\lambda_i}{\sum_{j=1}^{p} \lambda_j} * 100} \qquad \textbf{(0.22)}$$

1.4.9.2 ANÁLISIS DE COMPONENTES PRINCIPALES (PCA)

Es un algoritmo matemático cuyo propósito es reducir la dimensión de un conjunto de datos que conserva la mayor variabilidad posible, esto se consigue por la identificación de direcciones llamadas componentes principales, agrupando las variables que tienen correlación entre sí, separándolas de las que no. En donde la variación de los datos es máxima, considerando que en las PCA no se interpreta cada uno de los factores, su interpretación es el agrupamiento de las variables, después de aplicar este método a un conjunto de varias variables y varias observaciones, como resultado se obtiene un nuevo espacio reducido, en donde la interpretación se torna mucho más fácil la interpretación y el análisis de los datos, el nuevo espacio es una combinación lineal de las variables originales [27] [28]. En la Figura 1.12 se puede apreciar la reducción de un conjunto de datos, de dos dimensiones a una sola dimensión.

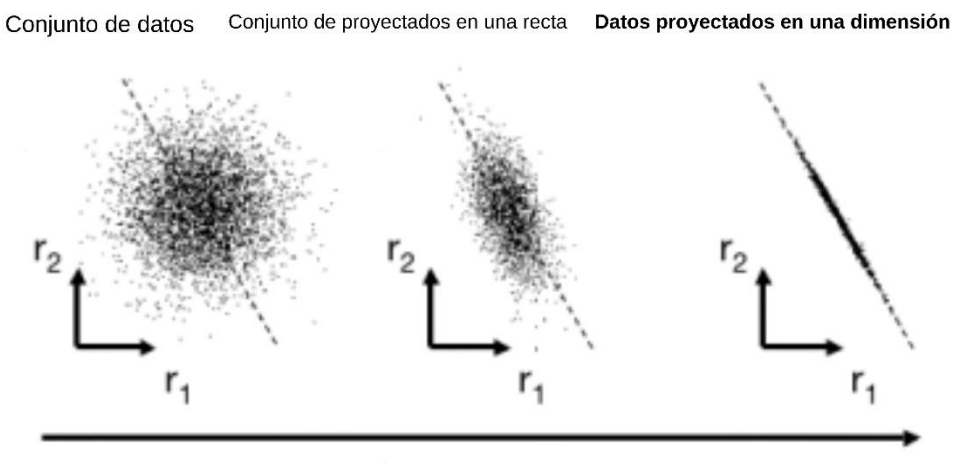


FIGURA 1.12 Aplicación de PCA en un conjunto de datos [propia].

Es una alternativa para la resolución de un problema de aprendizaje no supervisado, conocido como reducción de dimensionalidad, existen varias razones por las cuales se podría hacer una reducción de dimensionalidad, sin embargo, la aplicación de PCA's se resume en dos grandes aplicaciones. La primera es la compresión de datos, que permite aliviar el uso de datos y por lo tanto utilizar menos memoria de la computadora. La segunda aplicación es para relevar características de los datos que no pueden ser apreciadas en datos de alta dimension, en este caso se considera una técnica de visualización [33].

1.4.9.2.1. Algoritmo de PCA

Para iniciar con el algoritmo de PCA, se debe realizar una etapa de pre-procesamiento para un conjunto de datos de entrenamiento como se indica en la Ecuación 1.23 una matriz X de tamaño $m\ x\ n$, donde cada columna representa a una variable, y cada fila una muestra u observación.

$$X = \begin{pmatrix} X_{1,1} & X_{1,2} & X_{1,3}\dots & X_{1,j} & \dots & X_{1,n} \\ X_{2,1} & X_{2,2} & X_{2,3}\dots & X_{2,j} & \dots & X_{1,n} \\ X_{3,1} & X_{3,2} & X_{3,3}\dots & X_{3,j} & \dots & X_{1,n} \\ \dots & \dots & \dots\ \dots & \dots & \dots & \dots \\ \dots & \dots & \dots\ \dots & \dots & \dots & \dots \\ X_{m,1} & X_{m,2} & X_{m,3}\dots & X_{m,j} & \dots & X_{m,n} \end{pmatrix} \quad \textbf{(1.23)}$$

Por lo tanto, se tendrá entonces X_j es un vector columna, que representa a una variable j-ésima de los datos de entrenamiento, a la se la normalizará para que la media sea cero y la escala sea similar a través de un escalamiento de variable. Además $x^{(i)}$ es un vector fila que representa la $i - ésima$ observación de los datos de entrenamiento. Entonces, para la normalización de la media, primero obtenemos la media de cada variable como se indica en la Ecuación 1.24 y a continuación, reemplazamos cada variable X_j con X_j menos su media, (i.e. $X_j = X_j - \mu_j$). Esto hace que cada variable ahora tenga una media de cero.

$$\mu_j = \frac{1}{m}\sum_{i=1}^{m} X_{i,j} \quad \textbf{(1.24)}$$

El escalamiento es necesario ya que las diferentes variables pueden tener escalas muy disntintas. Por ejemplo, si X_1 representa el tamaño de una empresa, y X_2 es el número de trabajadores, es necesario escalar cada variable para tener una gama de valores comparables.

Así, tendriamos que :

$$X_j = \frac{X_j - \mu_j}{S_j} \quad \textbf{(1.25)}$$

donde, S_j es cierta medida de la gama de valores de la variable j (i.e la variabilidad de X_j). Por ejemplo, S_j podría ser el valor máximo menos el mínimo, o la desviación estándar de la variable j [33].

Una vez realizado el pre-procesamiento, se reduce los dato, de n dimensiones a k dimensiones. Con este objetivo, lo primero en el algoritmo de PCA es calcular la matriz de covarianza denotada comúnmente por medio de la letra griega Φ como se indica en la Ecuación 1.26 y definida como:

$$\Phi = \frac{1}{m}\sum_{i=1}^{m}(X^{(i)}X^{(i)})^T = \frac{1}{m}X^T X, \quad \textbf{(1.26)}$$

donde, $X^{(i)}$ corresponde a la $i-ésima$ observación de la matriz X (i.e la $i-ésima$ fila). Con respecto a la matriz de covarianza, es importante mencionar que:

- La matriz de covarianza será una matriz cuadrada de tamaño nxn
- Los valores de la diagonal de la matriz Φ (i.e. $\sum j,j$) representa la varianza de $j-ésima$ variable.
- Los valores fuera de la diagonal representan los valores de la covarianza entre las variables i y j.

A continuación, procederemos a realizar la Descomposición en Valores Singulares (Singular Value Decomposition, SVD) de la matriz de covarianza, de la cual se obtiene una matriz V cuyas columnas contienen los vectores propios de la matriz Φ. Cada vector propio se corresponderá con un valor propio, cuya magnitud indica cuánta variabilidad de los datos se explica por el mismo.

También, dado que V es una matriz cuadrada de tamaño nxn cuyos vectores columna $\boldsymbol{V}_1, \boldsymbol{V}_2, \ldots, \boldsymbol{V}_n$ representa los vectores propios de la matriz Φ, si el objetivo es reducir los datos de n dimensiones hasta k dimensiones, basta con tomar las primeras k columnas de la matriz V [33]. Se toman las primeras k columnas ya que el método SVD retorna los vectores propios en orden descendente con base en la variabilidad de cada uno. Y puesto que la variabilidad es un indicador de cuánta información retiene cada vector propio, al tomar las primeras k columnas de la matriz $\boldsymbol{V}$**,** garantizamos que, a pesar de la reducción

de dimensionalidad, se conserve una gran cantidad de la información original [33]. Las primeras k columnas de $\boldsymbol{V}$ pueden ser agrupadas como una nueva matriz $\boldsymbol{V_{reducida}}$ de tamaño nxk, que es conocida como matriz de rotación, la misma que define un nuevo espacio vectorial, ya que está compuesta integramente por vectores singulares [33].

$$\boldsymbol{V_{reducida}} = \begin{pmatrix} V_{1,1} & V_{1,2} & V_{1,3}V_{1,k} \\ V_{2,1} & V_{2,2} & V_{2,3}V_{1,k} \\ V_{3,1} & V_{3,2} & V_{3,3}V_{1,k} \\ \cdots & \cdots & \cdots \cdots \\ V_{n,1} & V_{n,2} & V_{n,3}V_{n,k} \end{pmatrix} \quad \textbf{(1.27)}$$

Por ulitmo, procedemos a proyectar los datos sobre el nuevo espacio vectorial de baja dimensión a través de la multiplicación vectorial, como se indica en la Ecuación 1.28.

$$\boldsymbol{Z^{(i)}} = \boldsymbol{X^{(i)}}\boldsymbol{V_{reducida}} \quad \textbf{(1.28)}$$

donde el vector fila $\boldsymbol{Z^{(i)}}$ corresponde a la proyección en el nuevo espacio vectorial reducido de la $i - ésima$ observación de la matriz (X). Los vectores $\boldsymbol{Z^{(i)}}$ pueden agruparse dentro de la matriz **Z** que tendrá dimensiones mxk (i.e. $m\ observaciones\ de\ k\ dimensiones$). Esta información reorientada corresponde a las nuevas coordenadas de los datos originales proyectados en baja dimensión, y se la conoce como score o peso.

1.4.10 FORMACIÓN DE CURVAS DE DEMANDA DIARIA ESTACIONALES

Uno de los grandes problemas para las empresas eléctricas de distribución, es el desconocimiento del comportamiento de la demanda de los usuarios comerciales, residenciales e industriales. Las curvas de demanda se dividen aún más por las condiciones de carga de invierno y verano, principalmente durante las épocas de mayor consumo; dentro de la planificación del sistema eléctrico de distribución sería muy importante conocer que grupos de clientes poseen mayor contribución en los picos de demanda durante el día, con el objetivo de enfocar las estrategias y planes a dichos grupos. También, se debe considerar los efectos de factores externos como la temperatura en la electricidad. Se supone que el verano es entre abril y septiembre, mientras que se supone que el invierno es entre octubre y marzo, estos datos son conforme a la variable climática, pero se debe considerar que en nuestro país el clima es invariante día tras día [26].El clima es una condición atmosférica que integra diversas variables meteorológicas que afectan generalmente el consumo eléctrico de los clientes, sin embargo, la demanda eléctrica a nivel de sistema no se ve influenciada por factores como la precipitación o la temperatura, debido a su ubicación geográfica en la línea ecuatorial [34].

El análisis de estacionalidad satisface las necesidades de consumo durante las temporadas más desafiantes del año, la ubicación del Ecuador, sobre la línea ecuatorial, crea dos estaciones bien definidas a lo largo del año; la húmeda o invierno y la seca o verano. La duración de las estaciones varía regionalmente. En la sierra y oriente, el verano dura de junio a septiembre y el invierno de octubre a mayo. En la Figura 1.13 se muestra que el mes con mayor nivel de lluvia mensual acumulada es abril y el menor es agosto [34].

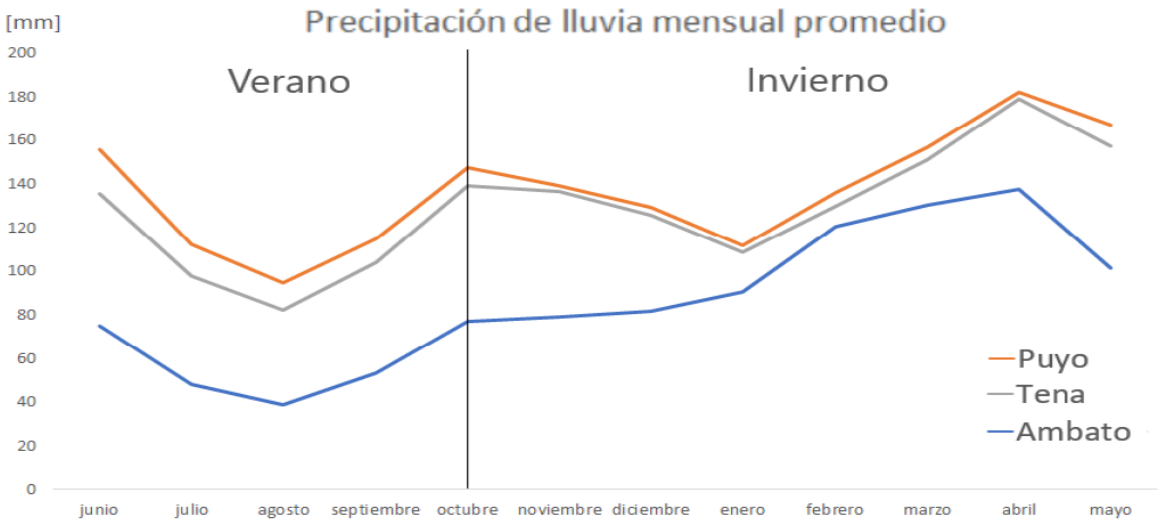


FIGURA 1.13 Promedio mensual de lluvia en las principales ciudades del periodo 2020 – 2021 [35].

La región oriental o amazónica se caracteriza por un clima tropical muy húmedo durante todo el año, con una temperatura promedio de 24°C a 25°C. Mientras que, la región sierra posee un clima frío en las zonas de alta montaña, y calido – seco en los valles, las temperarturas promedio son de 9°C a 20 °C [35].

1.4.10.1 Curvas De Demanda Diaria Estacionales

Un perfil de carga es el equivalente de una forma de carga estimada, para un grupo de clientes que se desarrolla a partir del registro de datos histórico hasta el día actual del consumo de electricidad. Las curvas de carga se dividen aún más por las condiciones de carga de invierno y verano, se forman para los días de semana, fines de semana y festivos. Los efectos de factores externos, como la temperatura en la electricidad afectan en el consumo eléctrico de los clientes. Se supone que el verano es entre abril y septiembre, mientras que se supone que el invierno es entre octubre y marzo, estos datos son conforme a la variable climática, pero se debe considerar que en nuestro país el clima es variante día tras día. Por lo general en el análisis de estacionalidad se suele escoger un día en

específico de la semana para obtener la curva de demanda mensual típica, con la finalidad de sintetizar el análisis [26]

1.4.10.2 Agrupación Y Clasificación Por Estacionalidad De Usuarios Comerciales E Industriales

Los clientes especiales, es decir, los clientes industriales y comerciales también denominados grandes consumidores son aquellos que utilizan la energía eléctrica exclusivamente para consumo propio en sus instalaciones y tienen tratamiento especial, para la facturación. También, tienen atención personalizada con ejecutivos de cuenta y revisores especiales para atender sus solicitudes, temas administrativos y técnicos, asesoramiento en el uso de energía, instalación de nuevos servicios, medición de consumo y potencia, facturación y convenios de pago. Las cargas comerciales se caracterizan por ser resistivas y se focalizan en las áreas céntricas de la ciudad donde la principal actividad es el comercio, edificios de oficinas y centros comerciales que comprenden básicamente los edificios mientras que las cargas industriales su principal componente es la potencia reactiva debido a la gran cantidad de motores instalados tomando en cuenta la calefacción y alumbrado referenciado a la parte industrial del área de concesión. [13] [14]

Sin embargo, deben cumplir con las siguientes consideraciones:

- La capacidad instalada en sistemas monofásicos o trifásicos es mayor o igual a 50 [kVA].
- Los consumos de energía tienen valores superiores a los 4000 [kWh] [36].

La agrupación de perfiles de carga ha sido bien explorada para comprender el comportamiento del cliente, para la gestión del lado de la demanda y para crear arquetipos de clientes para el desarrollo de tarifas y la generación a pequeña escala en el sector eléctrico. Muchos estudios de países desarrollados en el hemisferio norte agrupan poblaciones relativamente homogéneas, donde los expertos de dominio esperan que la variabilidad en la demanda eléctrica está influenciada principalmente por efectos estaciones y de días de semana. Dividir los datos de entrada a lo largo de las dimensiones temporales antes de la agrupación es común, los datos se deben encontrar divididos por estacionalidad, para invierno y verano, para los días de semana y los fines de semana. Según el estudio descubrieron que la clasificación previa a lo largo de una dimensión de demanda del consumidor, primero agrupando perfiles de carga por consumo general y

luego por su forma de carga, este método es eficaz para mejorar los resultados de la agrupación [37]. En la Figura 1.14 se muestra una tabla, donde se valida el procedimiento de agrupación para perfiles de carga.

Abreviación	Algoritmo	Frecuencia de uso	Puntuación
	k-means	19	4
HC	Hierarchical Clustering	12	2
SOM	Self-Organising Maps	7	2
	kmedoids	4	2
MFTL	Modified Follow-The-Leader	4	2
	fuzzy k -means	4	
	SOM+k-means	3	1
	AKM+HC	3	
	fuzzy c-means	2	
AKM	Adaptive kmeans	1	1
WFAKM	Weighted Fuzzy Averages kmeans	1	1
IRC	Iterative Refinement Clustering	1	1
MKM	Modified kmeans	1	
	SAX k-means	1	
	Spherical k-means	1	
AVQ	Adaptive Vector Quantisation	1	
	DBSCAN	1	
FTL	Follow-The-Leader	1	
GMM	Gaussian Mixture Model	1	
	Random Forests	1	
	Voronoi decomposition	1	
	SOM+HC	1	

FIGURA 1.14 Frecuencia de uso y rendimiento de los algoritmos de agrupamiento para agrupar perfiles de carga, según 25 estudios [37].

1.4.10.3 Técnicas de agrupamiento para perfiles de carga

Los índices de forma de carga ofrecen caracterizar Las técnicas de agrupamiento se clasifican en dos categorías: agrupamiento directo y agrupamiento indirecto. Definimos el agrupamiento directo como un método de agrupamiento de los datos recopilados directamente por los medidores inteligentes. Mientras que, si los datos de carga se han procesado mediante técnicas de reducción de dimensiones u otros métodos antes de la agrupación, se clasifica como agrupación indirecta [38].

1.4.10.3.1 Agrupación directa

Investigadores de todo el mundo intentan buscar una mejor manera de realizar patrones de consumo de electricidad, sobre la base de la recopilación y el procesamiento de datos de carga. Estos métodos tradicionales de agrupamiento se utilizan a menudo como punto de referencia para evaluar otros nuevos métodos [38]. En la Figura 1.15 se brinda la clasificación detallada de los métodos de agrupamiento.

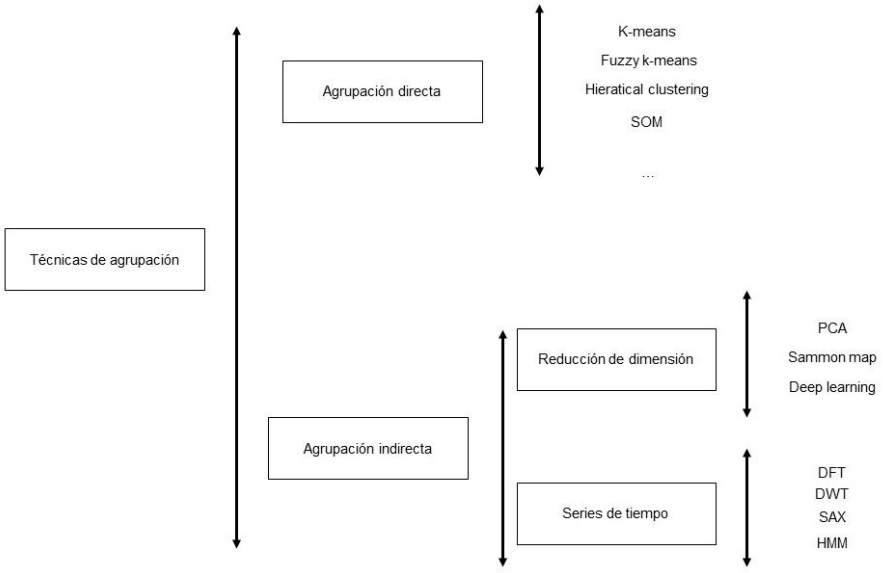


FIGURA 1.15 Clasificación de técnicas de agrupamiento para perfiles de carga [propia].

1.4.10.3.2 Agrupación indirecta

El agrupamiento indirecto significa que los objetos del agrupamiento son las caracteristicas extraidas de los datos de consumo de electricidad en lugar de los datos crudos. Asumiendo que X_{nxm} presenta un conjunto de datos de consumo de electricidad de n datos de clientes con m dimensiones. El procedimiento de extracción de caracteristicas es esencialmente una función de X_{nxm} y cumple con la siguiente Ecuación :

$$Y_{nxk} = f(X_{nxm}) \qquad \textbf{(1.29)}$$

donde Y_{nxk} denota las caracteristicas extraidas de los datos de carga de n clientes con k dimensiones. La extracción de caracteristicas se usa a menudo para reducir la escala de los datos de entrada, de modo que la desigualdad $k < m$ se mantenga generalmente. Las tecnicas de agrupamiento indirecto se pueden clasificar en agrupamientos basados en reducción de dimensiones y basados en series temporales [38].

2 METODOLOGÍA

2.1 ÁREA DE CONCESIÓN DE LOS USUARIOS COMERCIALES E INDUSTRIALES DE LA EEASA

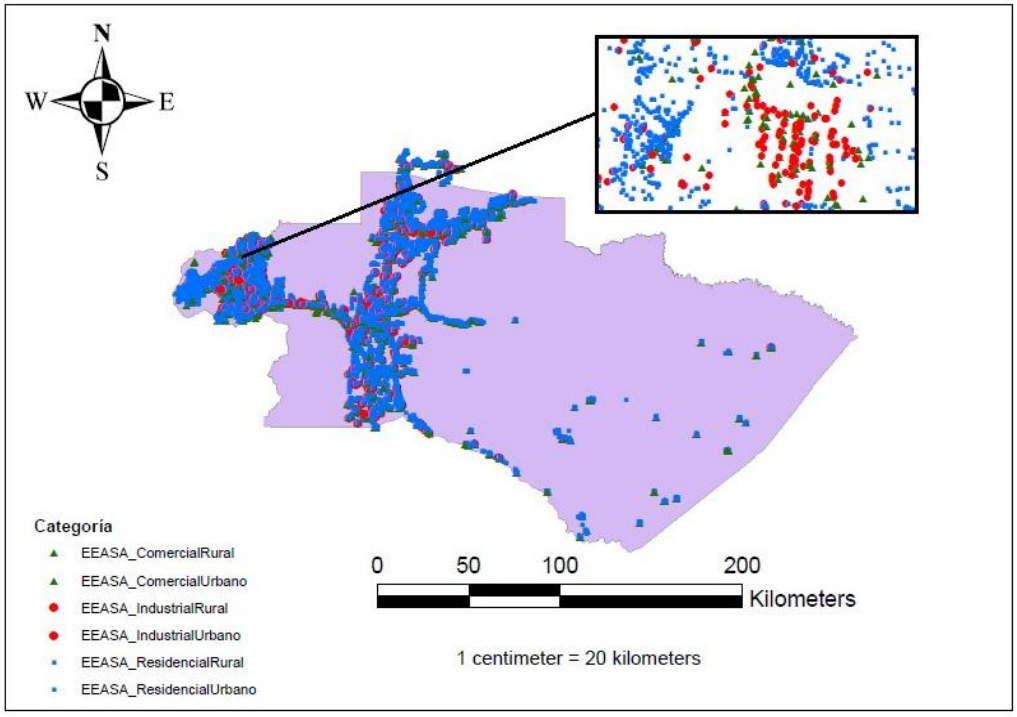


FIGURA 2.1 Área demográfica de los usuarios comerciales e industriales [Elaboración propia].

El área de concesión de los usuarios comerciales, residenciales e industriales de la EEASA abarca la zona central del Ecuador con una superficie de 40 805 km^2 y con un número de 832 075 habitantes, como se observa en la Figura 2.1 que comprenden las provincias de Pastaza y Tungurahua; los cantones que pertenecen son: Huamboya, Palora y Pablo Sexto en la provincia de Morona Santiago, en la parte sur de la provincia de Napo esta su capital Tena y sus cantones Archidona, Carlos Julio Arosemana Tola y Archidona. Los clientes servidos hasta diciembre 2019 son 278 279 clientes [29]. A partir de los datos históricos de los clientes de telemedición industrial y comercial como se observa en la Figura 2.1 para los días laborables de los años (2018-2019) se conformó una matriz de datos, para el caso de estudio no se tomó en cuentan los días festivos, debido a que el comportamiento de la demanda es diferente, en comparación con los días laborables. La clasificación, limpieza y exploración de datos se ejecutó en el entorno de Python, el presente capitulo se divide en 4 parte principales.

Las etapas para el desarrollo del proyecto, para la obtención de las curvas de demanda contribuirán a la clasificación de usuarios del servicio eléctrico de tipo industrial y comercial, el análisis de estudio se aplicará para 184 usuarios entre comerciales e industriales. Estas etapas conllevan un proceso de cuatro etapas sumamente elaborado y dividido en varios procesos, en donde la metodología usada no es especifica, como consecuencia lleva a ser explicado mediante la Figura 2.2, en donde se puede observar detalladamente el proceso de ejecución divido en ocho procedimientos.

En la primera etapa, se procederá a la descarga de los datos de telemedición del consumo eléctrico de los clientes comerciales e industriales, desde el portal de la EEASA, posteriormente un ordenamiento en una hoja de (Excel), con el tipo de día y consumo, en un intervalo de 10 minutos hasta completar el registro de un día, dando un total de 144 registros, luego estos registros se convertirán en variables en el dominio del tiempo.

En la segunda etapa se focaliza directamente en la limpieza, ordenamiento y manipulación de datos históricos de consumo, de clientes de tipo comercial e industrial para los días laborables para los años (2018-2019), debido a la gran dimensionalidad de los datos se realizará un preprocesamiento y clasificación, estos datos forman una matriz de datos, que luego son exportados a Python 3.7. Posteriormente, estos datos serán agrupados mediante clusters utilizando algoritmos computacionales.

En la tercera parte del trabajo, se corrige los datos faltantes o no censados mediante la herramienta de imputación de datos, con esto se realiza una predicción de los datos faltantes, con la finalidad de clasificar por estacionalidad se utilizará el algoritmo de K-means para la clasificación, el mismo que agrupa por el método de conglomeración. Luego se realiza una nueva clasificación que estará divida en tres etapas de consumo anual, estas tres etapas son dadas según el índice de agrupación presentado en la sección 1.4.8, para entonces la cantidad de datos manejada es de una alta dimensionalidad, por lo que se usará minería de datos para encontrar un equivalente, mediante las PCA, simplificando el tiempo de ejecución en Python, esto proporcionará un uso de datos flexibles y manipulables en el dominio de las PCA.

Finalmente, una vez procesados los datos, el algoritmo a implementarse será Fuzzy-c-means en Matlab, esto para dar una fiabilidad al proceso que realiza una clasificación por lógica difusa con su respectiva función de pertenencia, esto con la intención de que cada dato de consumo tenga una asignación más prolija y de esta manera asignar correctamente al grupo que pertenece, se introduce una matriz equivalente que se encontrará en el

dominio de las PCA, posteriormente, el algoritmo evaluará los datos ingresados, entregando las curvas de consumo por estacionalidad las mismas que también, serán normalizadas por patrón de consumo para ser ingresadas en el ADMS de la EEASA. Adicionalmente se considera que para validar el número óptimo de conglomeración se utilizará un índice de agrupación especificado en la sección 1.4.8 que se puede utilizar dentro de las librerías de Matlab.

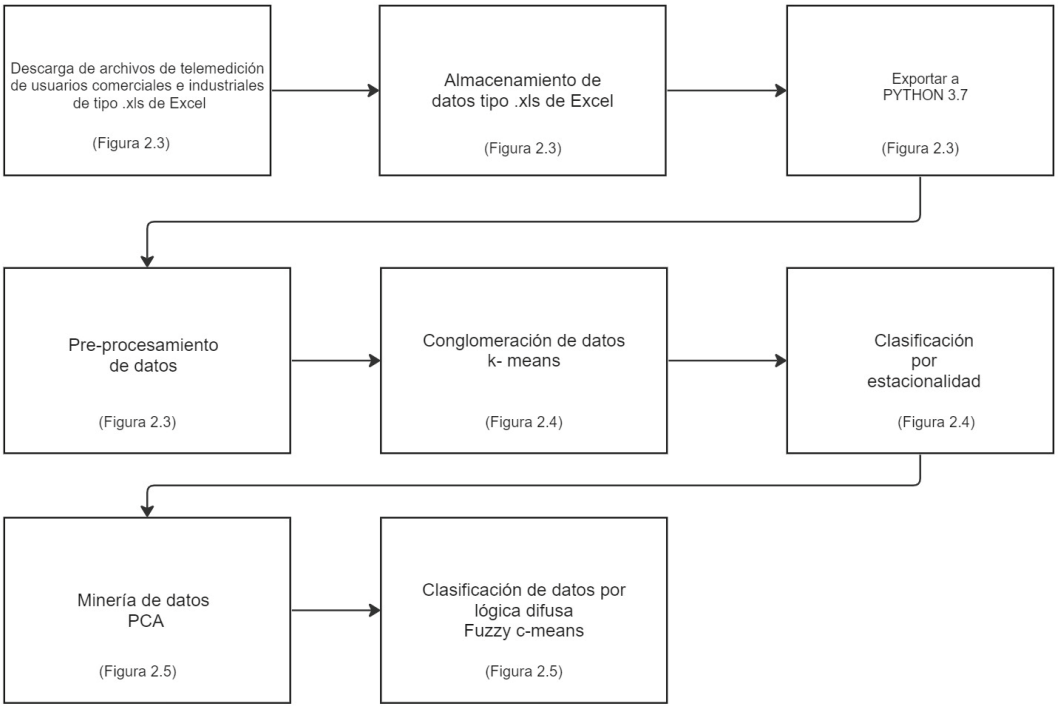


FIGURA 2.2 Proceso de obtención de las curvas de consumo de los datos de telemedición [Elaboración propia].

Las etapas que se pueden observar en la Figura 2.2. especifican los procesos de ejecución para la obtención de las curvas de consumo, en donde se muestra detalladamente cada subproceso. Por lo tanto, en la Figura 2.3, Figura 2.4 y Figura 2.5, se presenta el diagrama de flujo para cada procedimiento especificado en la Figura 2.2.

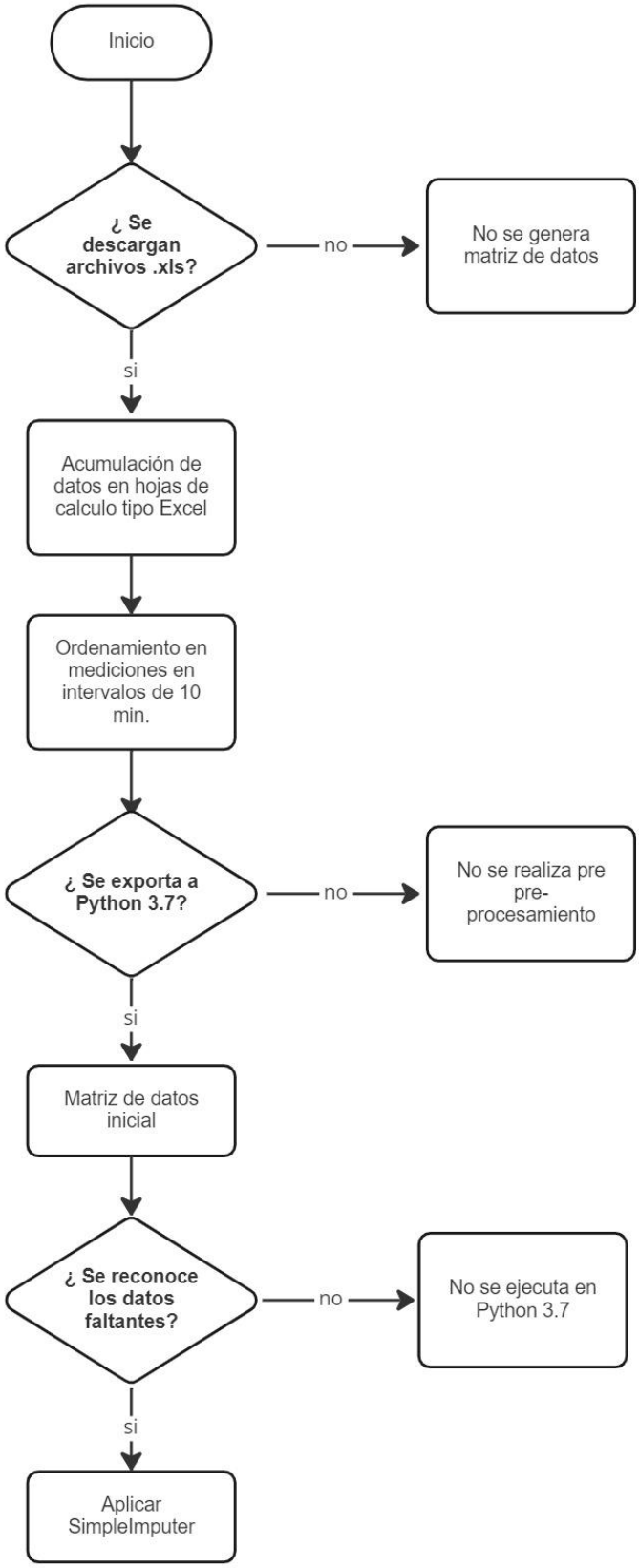


FIGURA 2.3 Diagrama de flujo de proceso para la obtención de curvas para los usuarios industriales y comerciales [Elaboración propia].

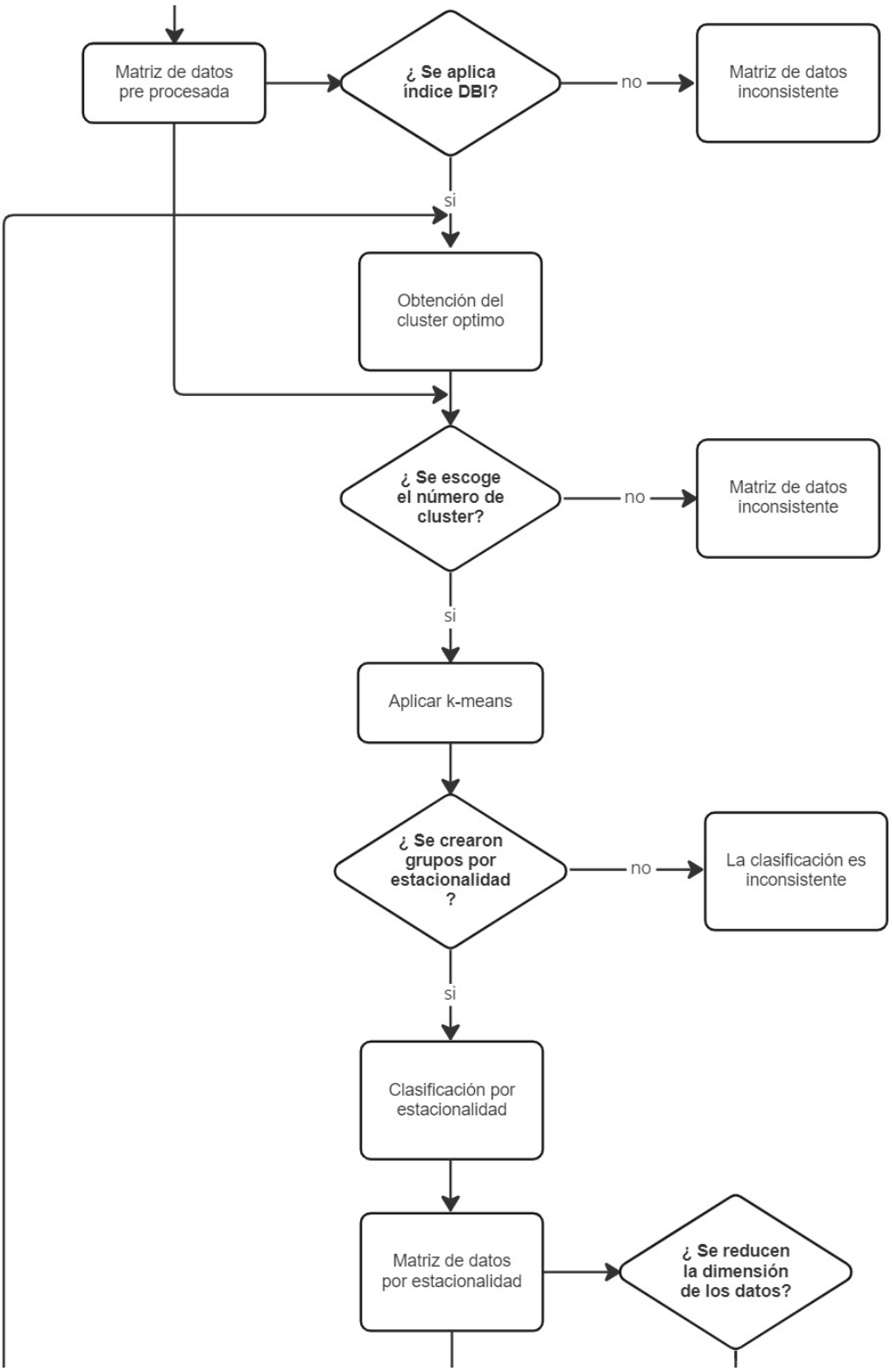


FIGURA 2.4 Diagrama de flujo de proceso para la obtención de curvas para los usuarios industriales y comerciales [Elaboración propia].

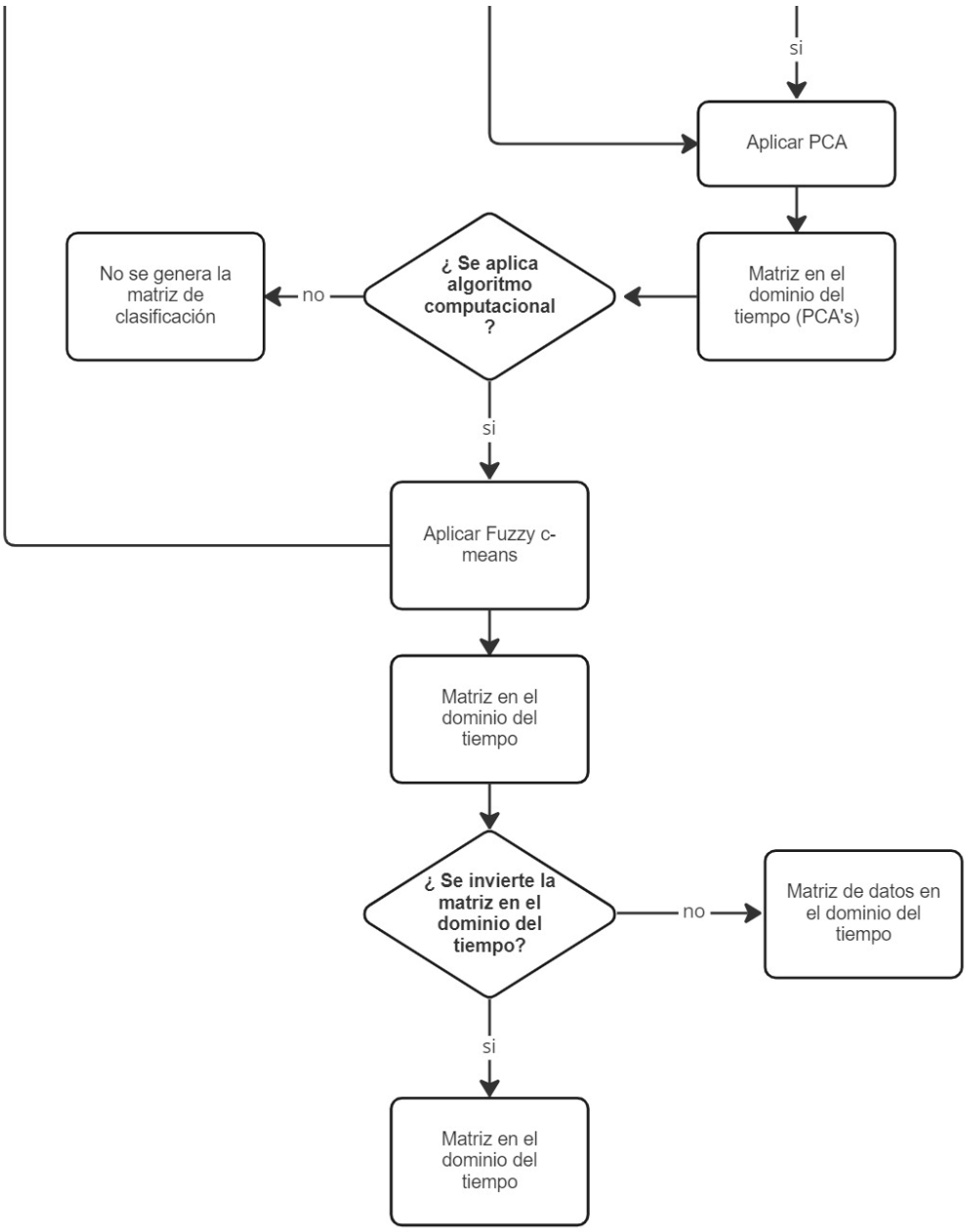


FIGURA 2.5 Diagrama de flujo de proceso para la obtención de curvas para los usuarios industriales y comerciales [Elaboración propia]

2.2 PREPROSECAMIENTO DE LA BASE DE DATOS DE USUARIOS COMERCIALES E INDUSTRIALES

Se utilizan datos históricos de consumo de clientes de tipo comercial e industrial para días laborables de los años (2018-2019), para formar una matriz de datos de extensión *.csv en donde se detalla el número de medidor, fecha de consumo, potencia activa y reactiva todo esto en intervalos de 10 minutos. Posteriormente se realiza una clasificación en un solo

archivo de Excel para así formar una matriz de 38 782 filas con 144 columnas, la cual está compuesta del consumo solo en días laborables de los usuarios comerciales e industriales.

En la Figura 2.6 se presenta la fuente de información que es el portal de la EEASA de donde es descargado el comportamiento anual de cada usuario en un archivo de tipo Excel.

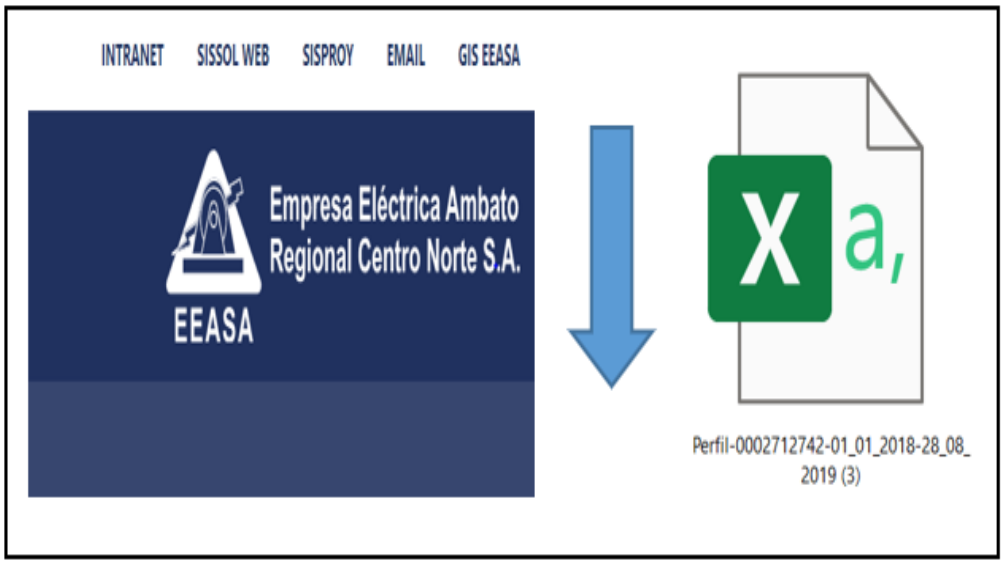
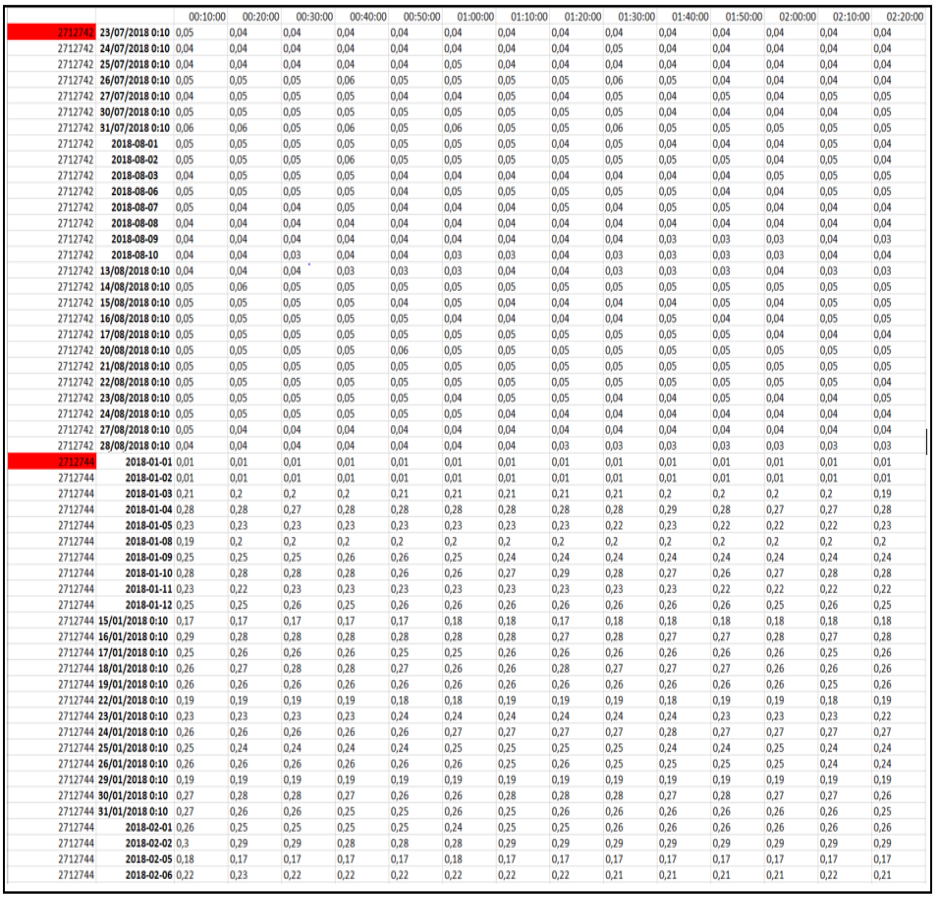


FIGURA 2.6 Proceso de descubrimiento del conocimiento mediante minería de datos [Elaboración propia].

2.3 MUESTREO Y SELECCIÓN

Después de haber descargado el registro de consumo anual en archivo *.csv (archivo de Excel) de cada usuario, estos archivos se encuentran en un gran volumen de datos, inmediatamente se procede a segmentar a cada usuario con su respectivo medidor, y este procedimiento se repite para todos los usuarios con su respectivo medidor, es decir, usuario por usuario, para luego durante la etapa de selección clasificar por días considerando que las variables se encuentran en el dominio del tiempo, esta matriz de datos está compuesta por los días laborables de lunes a viernes sin contar los días festivos, la muestra está conformada por usuarios comerciales e industriales de datos de telemedición, una vez completada la matriz seguidamente se procede a la fase de selección de cada cliente de la EESA con su respectivo medidor, una de las variables que van a aportar la información para el tratamiento de datos es la potencia activa, así como como las fuentes que pueden ser útiles. En la Figura 2.7 se ilustra la selección del medidor y sus datos de consumo.

Una vez seleccionadas las variables se aplican técnicas de muestreo adecuadas, con el fin de obtener una muestra de los datos que sea los suficientemente representativa de la población, la muestra permite inferir las propiedades o características de toda la población.



		00:10:00	00:20:00	00:30:00	00:40:00	00:50:00	01:00:00	01:10:00	01:20:00	01:30:00	01:40:00	01:50:00	02:00:00	02:10:00	02:20:00
2712742	23/07/2018 0:10	0,05	0,04	0,04	0,04	0,04	0,04	0,04	0,04	0,04	0,04	0,04	0,04	0,04	0,04
2712742	24/07/2018 0:10	0,04	0,04	0,04	0,04	0,04	0,04	0,04	0,04	0,05	0,04	0,04	0,04	0,04	0,04
2712742	25/07/2018 0:10	0,04	0,04	0,04	0,04	0,04	0,05	0,04	0,04	0,04	0,04	0,04	0,04	0,04	0,04
2712742	26/07/2018 0:10	0,05	0,05	0,05	0,06	0,05	0,05	0,05	0,05	0,06	0,05	0,04	0,04	0,04	0,04
2712742	27/07/2018 0:10	0,04	0,05	0,05	0,05	0,04	0,04	0,05	0,04	0,05	0,04	0,05	0,04	0,05	0,05
2712742	30/07/2018 0:10	0,05	0,05	0,05	0,05	0,05	0,05	0,05	0,05	0,05	0,04	0,04	0,04	0,04	0,05
2712742	31/07/2018 0:10	0,06	0,06	0,05	0,06	0,05	0,06	0,05	0,05	0,06	0,05	0,05	0,05	0,05	0,05
2712742	2018-08-01	0,05	0,05	0,05	0,05	0,05	0,05	0,05	0,04	0,05	0,04	0,04	0,04	0,05	0,04
2712742	2018-08-02	0,05	0,05	0,05	0,06	0,05	0,05	0,05	0,05	0,05	0,05	0,05	0,04	0,05	0,04
2712742	2018-08-03	0,04	0,05	0,05	0,05	0,04	0,04	0,04	0,04	0,04	0,04	0,04	0,05	0,05	0,05
2712742	2018-08-06	0,05	0,05	0,05	0,05	0,04	0,05	0,05	0,05	0,05	0,05	0,04	0,04	0,05	0,05
2712742	2018-08-07	0,05	0,04	0,04	0,05	0,04	0,04	0,04	0,05	0,04	0,05	0,05	0,04	0,04	0,04
2712742	2018-08-08	0,04	0,04	0,04	0,04	0,04	0,04	0,04	0,04	0,04	0,04	0,04	0,04	0,04	0,04
2712742	2018-08-09	0,04	0,04	0,04	0,04	0,04	0,04	0,04	0,04	0,04	0,03	0,03	0,03	0,04	0,03
2712742	2018-08-10	0,04	0,04	0,03	0,04	0,04	0,03	0,03	0,04	0,03	0,03	0,03	0,03	0,04	0,04
2712742	13/08/2018 0:10	0,04	0,04	0,04	0,03	0,03	0,03	0,04	0,04	0,03	0,03	0,03	0,04	0,03	0,03
2712742	14/08/2018 0:10	0,05	0,06	0,05	0,05	0,05	0,05	0,05	0,05	0,05	0,05	0,05	0,05	0,05	0,05
2712742	15/08/2018 0:10	0,05	0,05	0,05	0,05	0,04	0,05	0,04	0,04	0,04	0,04	0,05	0,04	0,05	0,05
2712742	16/08/2018 0:10	0,05	0,05	0,05	0,05	0,05	0,04	0,04	0,04	0,04	0,05	0,04	0,04	0,05	0,05
2712742	17/08/2018 0:10	0,05	0,05	0,05	0,05	0,05	0,05	0,05	0,05	0,05	0,05	0,05	0,04	0,04	0,04
2712742	20/08/2018 0:10	0,05	0,05	0,05	0,05	0,06	0,05	0,05	0,05	0,05	0,05	0,05	0,05	0,05	0,05
2712742	21/08/2018 0:10	0,05	0,05	0,05	0,05	0,05	0,05	0,05	0,05	0,05	0,05	0,05	0,05	0,05	0,05
2712742	22/08/2018 0:10	0,05	0,05	0,05	0,05	0,05	0,05	0,05	0,05	0,05	0,05	0,05	0,05	0,05	0,04
2712742	23/08/2018 0:10	0,05	0,05	0,05	0,05	0,05	0,04	0,05	0,05	0,04	0,04	0,05	0,04	0,04	0,05
2712742	24/08/2018 0:10	0,05	0,05	0,05	0,05	0,05	0,05	0,04	0,04	0,04	0,04	0,04	0,04	0,04	0,04
2712742	27/08/2018 0:10	0,05	0,04	0,04	0,04	0,04	0,04	0,04	0,04	0,04	0,04	0,04	0,04	0,04	0,04
2712742	28/08/2018 0:10	0,04	0,04	0,04	0,04	0,04	0,04	0,04	0,03	0,03	0,03	0,03	0,03	0,03	0,03
2712744	2018-01-01	0,01	0,01	0,01	0,01	0,01	0,01	0,01	0,01	0,01	0,01	0,01	0,01	0,01	0,01
2712744	2018-01-02	0,01	0,01	0,01	0,01	0,01	0,01	0,01	0,01	0,01	0,01	0,01	0,01	0,01	0,01
2712744	2018-01-03	0,21	0,2	0,2	0,2	0,21	0,21	0,21	0,21	0,21	0,2	0,2	0,2	0,2	0,19
2712744	2018-01-04	0,28	0,28	0,27	0,28	0,28	0,28	0,28	0,28	0,28	0,29	0,28	0,27	0,27	0,28
2712744	2018-01-05	0,23	0,23	0,23	0,23	0,23	0,23	0,23	0,23	0,22	0,23	0,22	0,22	0,22	0,23
2712744	2018-01-08	0,19	0,2	0,2	0,2	0,2	0,2	0,2	0,2	0,2	0,2	0,2	0,2	0,2	0,2
2712744	2018-01-09	0,25	0,25	0,25	0,26	0,26	0,25	0,24	0,24	0,24	0,24	0,24	0,24	0,24	0,24
2712744	2018-01-10	0,28	0,28	0,28	0,28	0,26	0,26	0,27	0,29	0,28	0,27	0,26	0,27	0,28	0,28
2712744	2018-01-11	0,23	0,22	0,23	0,23	0,23	0,23	0,23	0,23	0,23	0,23	0,22	0,22	0,22	0,22
2712744	2018-01-12	0,25	0,25	0,26	0,25	0,26	0,26	0,26	0,26	0,26	0,26	0,26	0,25	0,26	0,25
2712744	15/01/2018 0:10	0,17	0,17	0,17	0,17	0,17	0,18	0,18	0,17	0,18	0,18	0,18	0,18	0,18	0,18
2712744	16/01/2018 0:10	0,29	0,28	0,28	0,28	0,28	0,28	0,28	0,27	0,28	0,27	0,27	0,28	0,27	0,28
2712744	17/01/2018 0:10	0,25	0,26	0,26	0,26	0,25	0,25	0,26	0,26	0,26	0,26	0,26	0,26	0,25	0,26
2712744	18/01/2018 0:10	0,26	0,27	0,28	0,28	0,27	0,26	0,26	0,28	0,27	0,27	0,27	0,26	0,26	0,26
2712744	19/01/2018 0:10	0,26	0,26	0,26	0,26	0,26	0,26	0,26	0,26	0,26	0,26	0,26	0,26	0,25	0,26
2712744	22/01/2018 0:10	0,19	0,19	0,19	0,19	0,18	0,18	0,19	0,19	0,19	0,18	0,19	0,19	0,18	0,19
2712744	23/01/2018 0:10	0,23	0,23	0,23	0,23	0,24	0,24	0,24	0,24	0,24	0,24	0,23	0,23	0,23	0,22
2712744	24/01/2018 0:10	0,26	0,26	0,26	0,26	0,26	0,27	0,27	0,27	0,27	0,28	0,27	0,27	0,27	0,27
2712744	25/01/2018 0:10	0,25	0,24	0,24	0,24	0,24	0,25	0,25	0,25	0,25	0,24	0,24	0,25	0,24	0,24
2712744	26/01/2018 0:10	0,26	0,26	0,26	0,26	0,26	0,26	0,25	0,26	0,25	0,25	0,25	0,25	0,24	0,24
2712744	29/01/2018 0:10	0,19	0,19	0,19	0,19	0,19	0,19	0,19	0,19	0,19	0,19	0,19	0,19	0,19	0,19
2712744	30/01/2018 0:10	0,27	0,28	0,28	0,27	0,26	0,26	0,28	0,28	0,28	0,27	0,28	0,27	0,27	0,26
2712744	31/01/2018 0:10	0,27	0,26	0,26	0,25	0,25	0,26	0,25	0,26	0,26	0,26	0,26	0,26	0,26	0,25
2712744	2018-02-01	0,26	0,25	0,25	0,25	0,25	0,24	0,25	0,25	0,26	0,26	0,26	0,26	0,26	0,26
2712744	2018-02-02	0,3	0,29	0,29	0,28	0,28	0,28	0,29	0,29	0,29	0,29	0,29	0,29	0,29	0,29
2712744	2018-02-05	0,18	0,17	0,17	0,17	0,17	0,18	0,17	0,17	0,17	0,17	0,17	0,17	0,17	0,17
2712744	2018-02-06	0,22	0,23	0,22	0,22	0,22	0,22	0,22	0,22	0,21	0,21	0,21	0,21	0,22	0,21

FIGURA 2.7 Agrupación de los datos de consumo diario de los diferentes clientes de la EEASA [Elaboración propia].

2.4 EXPLORACIÓN

Una vez que los medidores se encuentren identificados y listos para ser pre-procesados el siguiente paso es clasificar hoja a hoja en Excel, tomándose en cuenta que los datos no son etiquetados por lo que el aprendizaje es: el no supervisado y su rendimiento a menudo es subjetivo y especifico del dominio [30]. El número de medidores es el mismo de los usuarios comerciales e industriales que son 184, en donde la variable identificada es la potencia activa que es el consumo diario, estos datos serán importados a Python.

2.5 EXPORTACIÓN DE LA MATRIZ DE DATOS A PYTHON

Una vez completada la matriz de datos de telemedición, se procede a importar el archivo en *.csv a Python 3.7, estos datos serán exportados mediante el comando de lectura: read_excel(“.xls”,sheet_name) a través de la biblioteca Scikit Learn y utilizando Machine Learning se realiza el pre-procesamiento, este procedimiento también prepara a la matriz de datos exportada para interactuar con las bibliotecas numéricas y científicas NumPy y SciPy, siendo el pilar importante en Machine Learning [29]. En la Figura 2.8 se observa las condiciones de inicio para iniciar con el proceso de estructuración de la exportación de datos.

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from sklearn.cluster import KMeans
from sklearn.impute import SimpleImputer

# ==========> Defino data frame
df=pd.read_excel("Medidores_anual.xls",sheet_name='Hoja2',header=None)
```

FIGURA 2.8 Agrupación de los datos de consumo diario de los diferentes clientes de la EEASA [Elaboración propia].

2.6 LIMPIEZA

Una vez que los datos se encuentran cargados en la interface de Python y al desarrollar el código de inicio para cargar la matriz de datos, el siguiente procedimiento es la limpieza de los datos debido a que los medidores pueden haber fallado por calibración o existió algún suceso como puede ser una maniobra dentro del ADMS, al realizar la limpieza de datos se encontraron datos atípicos que se los conoce como “nan”, la matriz de datos importada queda de la siguiente manera y cómo podemos observar en la Figura 2.9. Los datos ausentes se muestran como “nan” (not a number).

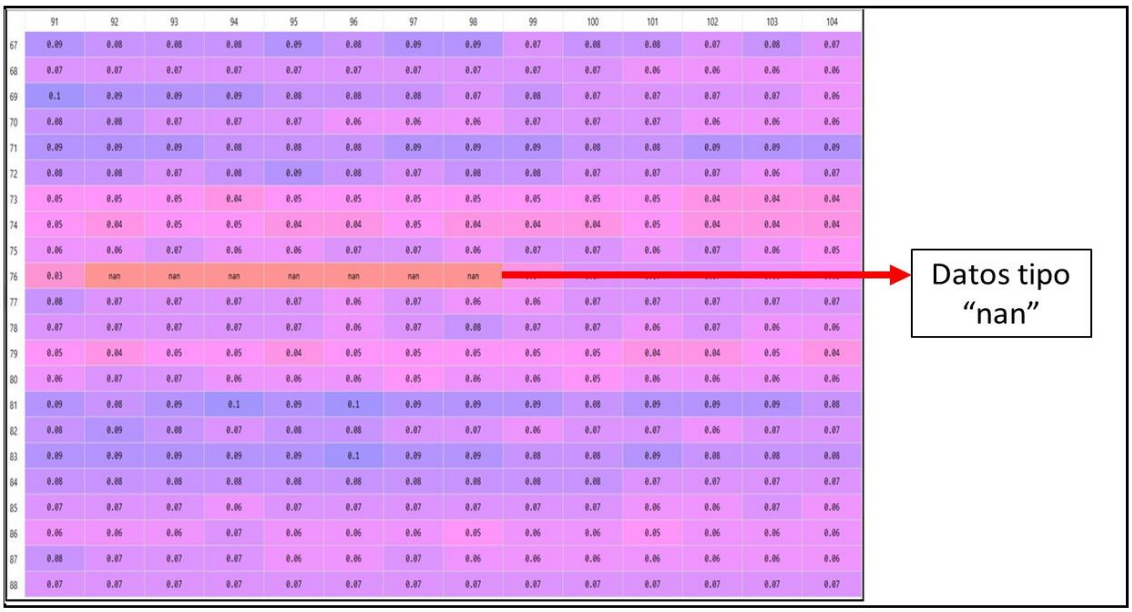


FIGURA 2.9 Agrupación de los datos de consumo diario de los diferentes clientes de la EEASA [Elaboración propia].

Los datos “nan” (not a number) impiden la manipulación y tratamiento, por lo tanto, se procede a corregir mediante algoritmos, por ejemplo, existe el registro faltante en la fila 76 y la columna 92 no tiene registro de consumo de potencia activa, el método aplicado para rellenar los datos faltantes es conforme a la teoría en el punto 1.4.4.4.

2.7 VALORES PERDIDOS

```
msg_dtype if msg_dtype is not None else X.dtype)
```

FIGURA 2.10 Marcador de posición de valores perdidos en el entorno de Python [Elaboración propia].

Se usa el marcador de posición como se muestra en la Figura 2.10 de los valores faltantes missing_values y dentro de este dominio se imputarán los valores faltantes, conforme la teoría en el apartado 1.4.6.6. La variable a completar los valores faltantes es la potencia activa generada por los usuarios comerciales e industriales potencia activa.

```
imputer = SimpleImputer(missing_values=np.nan, strategy='mean')
imputer = imputer.fit(A)
A= imputer.transform(A)
```

FIGURA 2.11 Aplicación del algoritmo SimpleImputer en Python [Elaboración propia].

El algoritmo SimpleImputer se ejecuta en la matriz de datos con valores numéricos totalmente libres de valores de tipo “nan”, con la técnica de la media truncada como se puede observar en la Figura 2.11. En este caso de ejemplo es a la matriz (A), es decir, al cliente número de 1 de 184 clientes, posteriormente el algoritmo es aplicado a las matrices de datos de forma secuencial hasta llegar al cliente 184 con la matriz (DDD_dd). Luego se entrena el modelo con el comando fit y de esta manera se obtiene la matriz para la aplicación de Machine Learning en la segunda etapa del proceso.

2.8 SELECCIÓN DEL ÍNDICE DE VALIDEZ

Se utilizó el software de Matlab para deducir el índice de validez en esta parte del desarrollo de este proyecto debido a su fácil manipulación y entorno de fácil manejo, para la detección del índice de validez correcto, de las técnicas mencionadas con anterioridad en el apartado de índice de validez, no existe un criterio exacto que destaque una de ellas en particular, por lo tanto, cualquier método puede ser aplicado para la deducción del índice adecuado, para el presente trabajo se ha escogido dos técnicas que son K-means y Fuzzy-c-means los mismos métodos necesitan un número óptimo de grupos para el análisis de conglomerados en las 2 técnicas, por tal motivo, la simplicidad aporta de una manera significativa, por esta razón se escogió un límite de 3 grupos con el propósito de no tener demasiados perfiles de demanda que alteren el resultado final a conseguir, que son las curvas de demanda para determinado periodo de tiempo, por lo cual el índice utilizado con frecuencia y con un historial óptimo de resultados por su grado de efectividad es el índice de Davies-Bouldien, este método considera el factor distancia entre los centros de cada cluster relacionando directamente la distancia entre los individuos de clusters distintos, ahora su implementación es un paso crucial dentro del análisis de los grupos, debido a que posee la ventaja de que el índice se encuentra en el software MATLAB dado por la función *evalclusters* que su formato viene ya implementado por la herramienta propia que es mathworks en el siguiente formato:

$$evalcluster(X, kmeans,' DaviesBouldin', [1:n]) \quad \textbf{(0.1)}$$

Dónde:

***X*=** Es la matriz de atributos que servirá para el análisis de conglomerados y en este caso M x n ($184\ clientes\ x\ 144\ mediciones$)

$\boldsymbol{kmeans}$ **=** Algoritmo de clusterización, para el caso de K-means

$\boldsymbol{n}$ **=** es el límite de grupos para el análisis de conglomerados, en este caso se desea obtener el numero optimo entre 1 a 5.

El algoritmo adecuado conforme a la teoría en el punto 1.4.8.2 sirve para evaluar el número óptimo de cluster, el mismo que es implementado en lenguaje de máquina, como se observar en la Figura 2.12. Para el análisis de la matriz de datos el índice preferido para el desarrollo de este proyecto es “DaviesBouldin”, con un rango de [1:5] para escoger el mejor indicador para la aplicación de K-means.

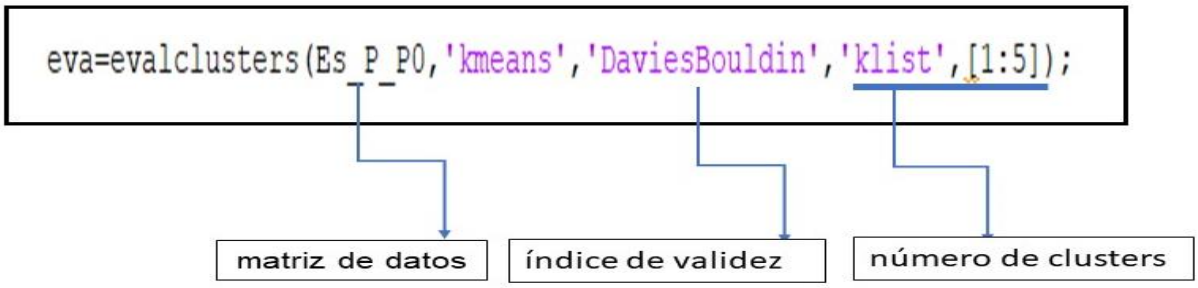


FIGURA 2.12 Implementación del algoritmo del índice de validez [Elaboración propia].

Mientras tanto en la interfaz de Python 3.7 en el desarrollo del código para la aplicación de K-means, en la parte de selección del índice de validez como se observa en la Figura 2.13, la condición del cluster es 3, el mismo que es aplicado a la matriz de datos en este caso servirá de ejemplo la matriz de datos denominada (C) que pertenece a uno de los usuarios comerciales e industriales.

```
kmeans=KMeans(n_clusters=3)
kmeans=kmeans.fit(C)
labels=kmeans.predict(C)
centroids_3=kmeans.cluster_centers_
```

FIGURA 2.13 Aplicación del número de clusters para el proceso de agrupación [Elaboración propia].

2.9 MACHINE LEARNING: K-MEANS

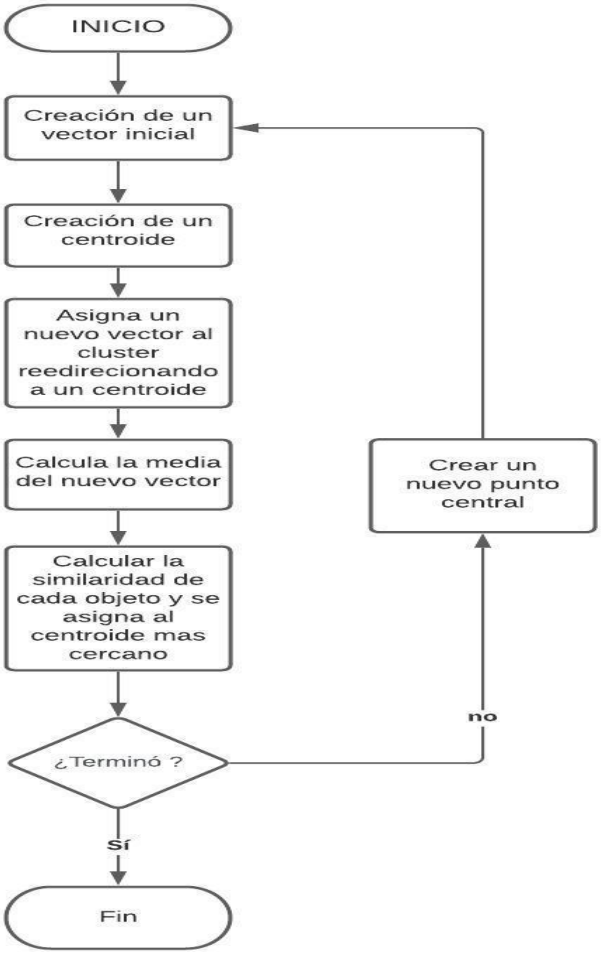


FIGURA 2.14 Algoritmo de agrupamiento aplicado a la clasificación de usuarios residenciales y comerciales de la EEASA [Elaboración propia].

Paso 1: Iniciar seleccionando K vectores aleatoriamente [0; k] del conjunto de n observaciones [$X_1, \ldots, X_n$]. Y posteriormente crear un vector inicial K [$Y_1, \ldots, Y_k$]

Paso 2: Si del nuevo vector inicial X_n esta cerca a Y_i asignar al clúster más cercano X_i. Luego de esto, segmentar el conjunto de datos en K clústeres $[X_1, \ldots., X_k]$.

A continuación, se muestra la Ecuación 2.2.

$$\boldsymbol{X_i = \{ X_n | \, d(X_n, Y_i) \;\; d(X_n, Y_i)\}, j = 1, \ldots. K\}} \qquad \textbf{(0.2)}$$

Paso 3: Para el nuevo centroide obtenido de los nuevos clústeres alcanzados en el paso 2 cada centro se renueva iteradamente de la siguiente forma: $Y_i = c\ (X_i), i = 1, \ldots., K$

Paso 4: Una vez generado los nuevos vectores calcular los centros de los clústeres comúnmente un clúster se calcula tomando el promedio de cada característica de cada punto de datos que le pertenece: $\frac{1}{N}\sum X_i\ , i = 1, \ldots., N$

Paso 5: Se consigue la distorsión total mediante la suma de las distancias desde los centros de los clústeres más cercanos a los datos.

Así lo indica la Ecuación 2.3.

$$\boldsymbol{D = \sum_{n=1}^{N} d(X_n, Y_{i(n)})} \qquad \textbf{(0.3)}$$

Donde:

$\boldsymbol{i(n) = k\ , si\ X_n = X_k}$

Paso 6: Verificar que los valores centrales del nuevo clúster son iguales a los valores que entrega la iteración mayor, detenerse o en caso contrario, repetir los pasos 2-5 con el valor central de un nuevo grupo.

2.10 PROCESO DE CLUSTERIZACIÓN PARA LA CLASIFICACIÓN POR ESTACIONALIDAD

La finalidad del método K-means es clasificar a los clientes por estacionalidad, agrupará cada curva de consumo de potencia activa de los 184 clientes que pertenecen a la EEASA, el procedimiento que ejecutará será el especificado en el apartado 1.4.7.2.1, para esto se utiliza la curva de carga diaria de los clientes comerciales e industriales que se encuentran en la matriz de datos compuesta por los días laborables en intervalos de 10 minutos, cada una de estas curvas de carga diaria entraran en el proceso de clusterizacion. El principal propósito en esta etapa del proyecto es establecer la tendencia predominante de cada una

de las curvas de carga, también contribuye a la detección de errores tempranos en el ADMS.

La fuente de datos para el desarrollo del proyecto es la demanda de los años (2018-2019). Como ejemplo del proceso de ejecución, se ha considerado al usuario con el # de medidor 2712742, para clasificar su curva de demanda que pertenece al segundo trimestre del año como se observa en la Figura 2.15, la curva representa a un cliente de tipo industrial, al igual que este cliente todos los 184 clientes son clusterizados mediante el algoritmo de Machine Learning, este método se replica para los tres escenarios que pertenecen a las 3 estaciones del año, obteniendo como resultado una previa clasificación asignada por el número de cluster.

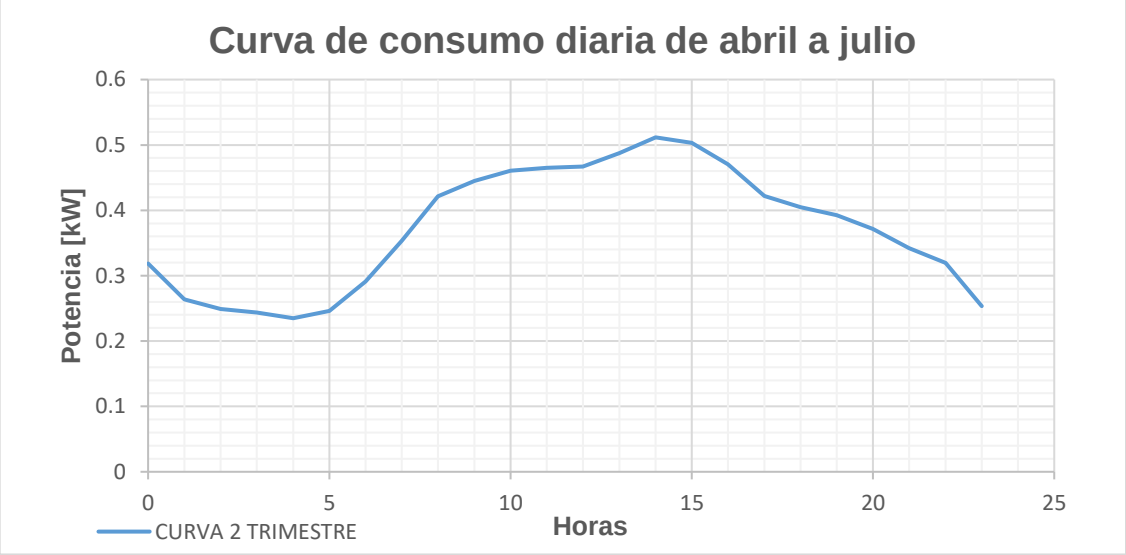


FIGURA 2.15 Perfil de consumo representativo [Elaboración propia].

El proceso se repite para el escenario de los meses de agosto a diciembre, para el mismo cliente del proceso anterior en donde luego del proceso de clusterización, entrega una curva de consumo como se puede ver en la Figura 2.16, el comportamiento de este cliente como se había explicado con anterioridad es industrial por lo que existe una desconexión del servicio al medio, esta curva de consumo diaria es clasificada para posteriormente ser clasificada por su estacionalidad.

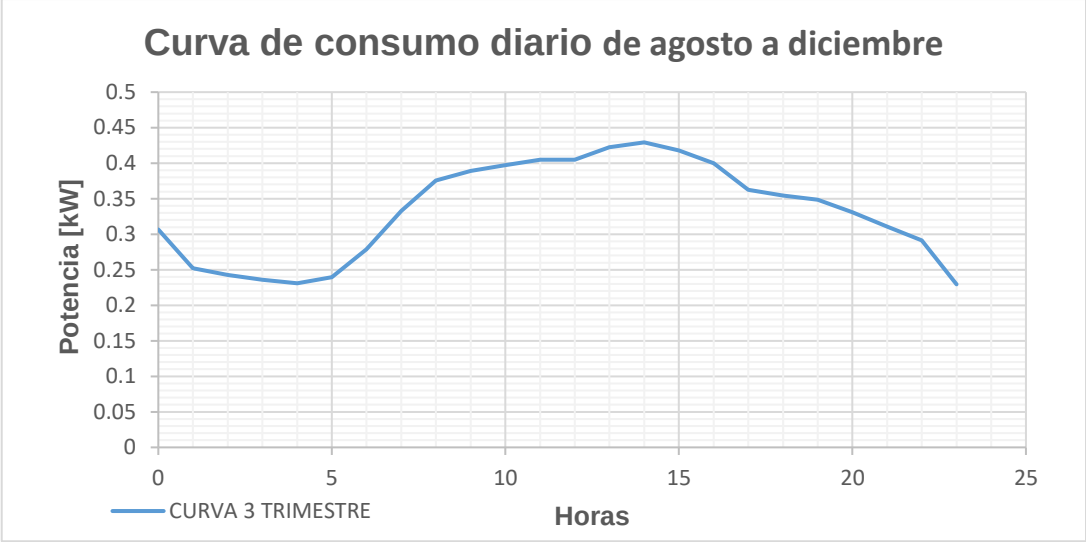


FIGURA 2.16 Perfil de consumo representativo [Elaboración propia].

Por último, el proceso para el escenario de los meses de diciembre a marzo conlleva el mismo tipo de ejecución para los meses de abril a julio y de agosto a diciembre, en donde se puede observar en la Figura 2.17, el comportamiento de este cliente, que su consumo es activo en un horario diferente, debido a que su desconexión no es a la mitad del día, como en el resto del año, llegando a la conclusión que en esta etapa del año su comportamiento es diferente y su desconexión es en las horas de madrugada.

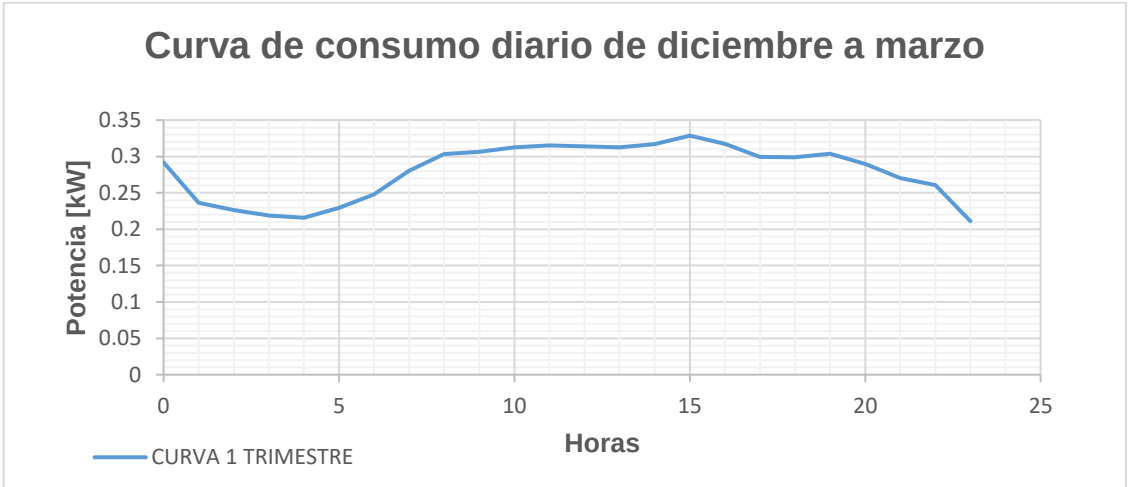


FIGURA 2.17 Perfil de consumo representativo [Elaboración propia].

Después de realizar el proceso anterior y observar el comportamiento del usuario que sirvió como ejemplo, existe una desconexión de parte de los clientes industriales en ciertas etapas del año, pero su desconexión no siempre es en la misma hora en la primera parte del año.

2.11 DETERMINACIÓN DEL PERFIL DE CARGA

Se obtuvieron los datos de consumo utilizando en el análisis de agrupamiento mediante el método de K-means a partir de la medición de las curvas de distribución de consumo diario de 184 medidores pertenecientes a la EEASA. La medición se registró cada 10 minutos, por lo que hay 144 variables en el dominio del tiempo para una curva diaria. En este trabajo, solo se utilizan medidas para días laborables una vez que las curvas de cargan están disponibles, los procedimientos para determinar los perfiles de carga se presentan en la Figura 2.18.

FIGURA 2.18 Diagrama de flujo para la determinación de perfiles de carga [Elaboración propia].

La Figura 2.18. muestra que las acciones de pre procesamientos son necesarias antes de que los conjuntos de datos se puedan utilizar en la agrupación, en este paso se puede utilizar para detectar datos incorrectos que faltan o valores indefinidos, además, se debe considerar que todas las mediciones deben ser normalizadas en valores por unitarios utilizando un factor de normalización adecuado que podría ser la potencia media en un periodo de tiempo determinado o la potencia pico, en este trabajo el normalizado elegido

es la potencia máxima de toda la medición. Después de este proceso, los datos de carga están listos para agruparse, el algoritmo de agrupamiento producirá los perfiles de carga típicos

2.12 ANÁLISIS TRIMESTRAL POR ESTACIONALIDAD

Hay dos tipos de patrones de consumo, los clientes de tipo industrial y comercial, es decir, que están divididos por el tipo de cliente y el área. Durante el procesamiento de los datos de consumo eléctrico realizados en la etapa clasificación mediante la técnica de K-means se obtuvo resultados dependiendo del número de clústeres ejecutados que fueron 3, de lo cual la clasificación se divide en 3 escenarios divididos en las estaciones del año, de lo cual se estudió el comportamiento de los perfiles de consumo diario, posteriormente una clasificación por la curva de consumo diario en diferentes meses del año, que será almacenada en una hoja de Excel, como se puede apreciar en la Tabla 2.1.

TABLA 2.1. Formato de registro por temporada previo al resultado del cluster después del proceso de agrupamiento por clustering [Elaboración propia].

Meses del año	Fecha	Cluster	Potencia [kW]	Potencia [kW]	Potencia [kW]
Febrero	23/02/2018 0:10	1	1.81	1.94	1.81
Abril	2018-04-04	1	1.93	1.63	1.58
Abril	2018-04-05	1	1.99	1.86	1.88
Abril	2018-04-11	1	1.39	1.41	1.41
Abril	2018-04-12	1	1.46	1.61	1.49
Mayo	2018-05-09	1	2.39	2.3	2.08
Mayo	2018-05-10	1	3.46	3.46	3.44
Mayo	2018-05-11	1	2.64	2.7	2.66
Junio	2018-06-07	1	1.68	1.54	1.85
Junio	2018-06-08	1	2.27	2.2	2.2
Junio	22/06/2018 0:10	1	2.09	1.93	2.04
Julio	2018-07-12	1	1.85	1.7	2.06
Julio	13/07/2018 0:10	1	2.15	2.4	2.08
Julio	17/07/2018 0:10	1	1.76	1.73	1.57
Julio	18/07/2018 0:10	1	1.74	1.58	1.45
Julio	19/07/2018 0:10	1	1.44	1.34	1.49
Julio	20/07/2018 0:10	1	1.49	1.46	1.3
Julio	26/07/2018 0:10	1	1.44	1.34	1.16
Agosto	2018-08-10	1	1.3	1.23	1.19
Agosto	22/08/2018 0:10	1	1.61	1.34	1.44
Octubre	2018-10-04	1	1.59	1.59	1.76
Octubre	2018-10-05	1	2.99	2.73	2.52

La agrupación por estacionalidad se realiza después de ejecutado el algoritmo de clustering en el cual se obtiene como resultado, en base al número de clusters aplicados que fueron 3, por lo tanto, se divide en tres escenarios o trimestres para su respectivo análisis y

clasificación, estos tres escenarios dependen directamente del número de clusters, en el proceso se obtiene como resultado el comportamiento por estacionalidad de cada uno de los clientes (184 medidores) en el cual la potencia está dada en [kW], cada comportamiento pertenece a un dato de consumo durante los años 2018-2019, de esta clasificación se encarga el algoritmo K-means, una vez segmentado los datos se procede a una clasificación interna para determinar la tendencia optima dada por la predicción del algoritmo, en el cual se deduce por el número de tendencia más alto, es decir, el mes que predomina para ello hemos utilizado un contador, como se observa en la Tabla 2.2, el mismo que entrega la información en consideración con el número de veces que se repite el mes con más frecuencia dentro del escenario

TABLA 2.2. Conteo para la deducción del mes predominante [Elaboración propia].

Meses del año	Número de veces que se repite
Abril	4
Agosto	2
Febrero	1
Julio	7
Junio	3
Mayo	3
Octubre	2
Total	**22**

Para el respectivo análisis se contabiliza el número de veces que predomina cada medición dada por la fecha en la cual su membrete es el mes, como se muestra en la Tabla 2.3. Una vez realizado el conteo respectivo considerando la fecha se obtiene el dato final, el mismo que sirve para encontrar el comportamiento adecuado dentro de las 3 estaciones del año, de esta manera deducir los meses en los cuales las curvas determinadas en la primera etapa del año son viables para el análisis respectivo. El comportamiento en el invierno muestra que la tendencia de consumo se encuentra desde los meses de diciembre, enero, febrero y marzo.

TABLA 2.3. Resultados de la clasificación dividida en trimestres luego de realizar el clustering [Elaboración propia].

PRIMER TRIMESTRE DEL AÑO		
Meses del año	Frecuencia	Probabilidad %

Enero	35	14.17004049
Febrero	31	12.55060729
Marzo	37	14.97975709
Abril	12	4.858299595
Mayo	14	5.668016194
Junio	14	5.668016194
Julio	15	6.072874494
Agosto	19	7.692307692
Septiembre	5	2.024291498
Octubre	21	8.502024291
Noviembre	20	8.097165992
Diciembre	24	9.71659919
Total	247	100

Luego de haber realizo el clustering se preparan los datos para una clasificación por estacionalidad en base a los resultados obtenidos por el número de cluster's que fueron 3 los mismo que se dividen en tres escenarios, como resultados mostraron que en la primera parte del año que desde el mes de diciembre, enero, febrero hasta el mes de marzo en donde existe la más alta probabilidad o tendencia, razón por la cual las curvas obtenidas en este periodo se recomiendan sean incorporadas al sistema ADMS en cierta temporada del año, esto se realizó mediante una clasificación por estacionalidad según el comportamiento de los consumidores, en la cual cada medidor de los 184 totales (clientes) pertenece a la medición de un mes del año calendario, con una fecha y hora, por lo tanto, se realizó la clasificación cliente por cliente en donde predominaba el mes con más frecuencia y este se agrupa en la Tabla 2.4, en donde se puede observar la frecuencia con los meses en donde existe mayor predominancia, es decir, que el consumo de cada cliente fue más fuerte en los meses que inicia el año, este comportamiento puede ser mostrado

debido a que en la provincia de Tungurahua se celebran fechas conmemorativas que van desde el mes de diciembre cerrando ya su etapa en el mes de marzo, también se debe considerar según la sección 1.4.10, indica que en abril es el mes con mayor afluencia fluvial en Tungurahua, esto implica que el comportamiento de los consumidores sea elevado, debido a que el ser humano por naturaleza busca protección y su hogar es su refugio resguardándose y esto implica que las personas obtén por comprar varios artículos, provocando una alta producción en las industrias.

TABLA 2.4. Resultados de las pruebas realizadas [Elaboración propia].

SEGUNDO TRIMESTRE DEL AÑO		
Meses del año	Frecuencia	Probabilidad %
Enero	37	14.91935484
Febrero	27	10.88709677
Marzo	23	9.274193548
Abril	9	3.629032258
Mayo	22	8.870967742
Junio	13	5.241935484
Julio	23	9.274193548
Agosto	22	8.870967742
Septiembre	7	2.822580645
Octubre	29	11.69354839
Noviembre	16	6.451612903
Diciembre	20	8.064516129
Total	248	100

A continuación, en la Tabla 2.4. se detalla los resultados de la observación por estacionalidad, los meses de abril a julio pertenecen al otoño, estos meses en el calendario

indican ser la mitad del año, dichos meses presentan un número alto de frecuencia con una alta probabilidad, pero no mayor como lo indica en la Tabla 2.5 que son los meses a los cuales fue asignado la estación de verano, la primera parte del año ya fue tomada en cuenta para asignar al comportamiento de invierno, por consecuencia se asignan que estos meses pertenecen al otoño. Los clientes industriales y comerciales entran en una etapa del año en donde su comportamiento va variando en base a la primera etapa del año que es el invierno, donde se realizó una alta producción que abarca la primera parte del año.

En la tabla 2.5. se muestra el análisis por estacionalidad en la parte final del año, entre los meses de agosto a diciembre, estos meses son asignados a la estación de verano, debido a su alto índice de frecuencia y probabilidad. Por lo tanto, se puede concluir que las mejores curvas para verano son de agosto a diciembre. También se debe considerar que el indicador probabilístico sirve de ayuda para analizar el comportamiento de mejor manera.

TABLA 2.5. Resultados de las pruebas realizadas [Elaboración propia].

TERCER TRIMESTRE DEL AÑO		
Meses del año	Frecuencia	Probabilidad %
Enero	49	21.68141593
Febrero	11	4.867256637
Marzo	24	10.61946903
Abril	12	5.309734513
Mayo	11	4.867256637
Junio	7	3.097345133
Julio	15	6.637168142
Agosto	14	6.194690265
Septiembre	10	4.424778761
Octubre	18	7.96460177
Noviembre	17	7.522123894

Diciembre	38	16.81415929
Total	226	100

2.13 NORMALIZACIÓN

Todas las curvas de consumo diarias son normalizadas por su propia demanda máxima como se indica en la Ecuación 1.4, en el apartado 1.4.5. La normalización está dada por una constante global en cada escenario del año de los 184 clientes con datos de telemedición.

En el caso de la manipulación de datos para normalizar cada uno de los escenarios las distancias euclidianas serían los vectores $(V_1, \dots, V_i \dots, V_n)$ los cuales son 144 mediciones es decir que el grupo de vectores es $(V_1, \dots, V_{144} \dots, V_n)$ de los cuales se el valor de máximo será tomado como referencia que toma el valor de $Máxima\ V_i$, este valor se obtiene al utilizar el comando **MAX** que es una herramienta de Excel este valor será el que se encuentre en el denominador de la Ecuación 1.5 y para el cual se dividirán todo el conjunto de vectores tipo $(V_1, \dots, V_{144} \dots, V_n)$ de esta manera obtiene los nuevos $(v_1, \dots, v_{144} \dots, v_m)$ y de esta manera se encontrarán todos los nuevos vectores normalizados.

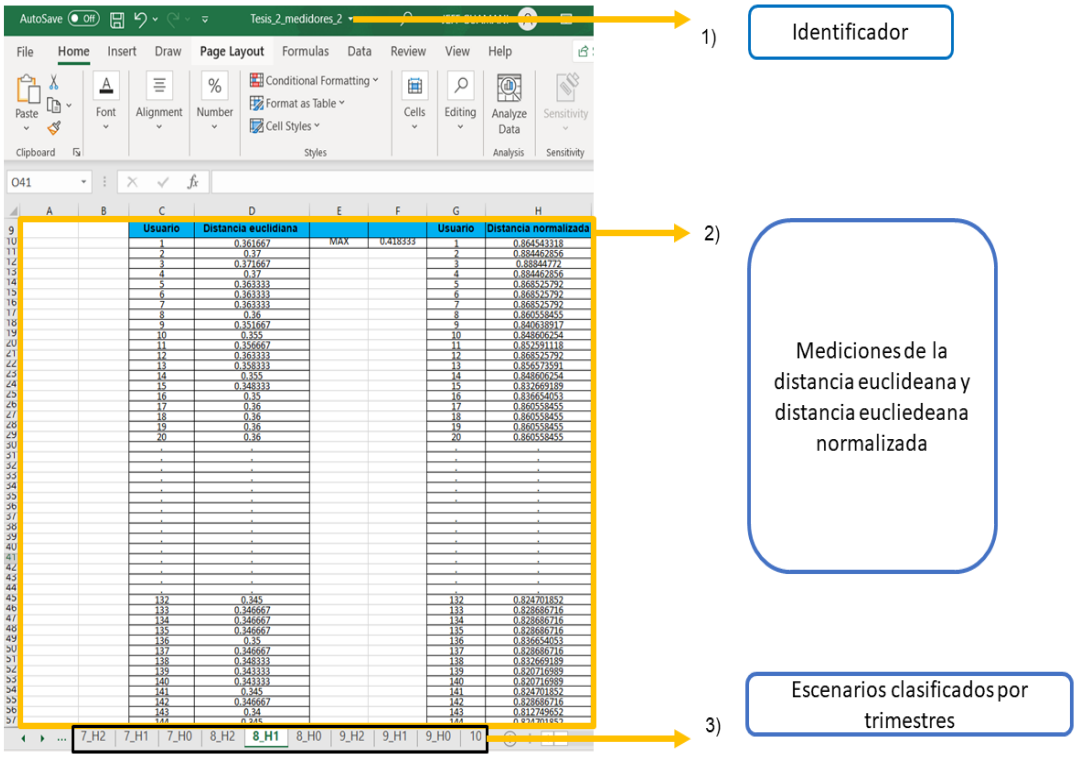


FIGURA 2.19 Archivos Excel de las distancias euclidianas y distancias euclidianas normalizadas [Elaboración propia].

El resultado luego utilizar la herramienta K-means con un argumento de 3 cluster's los mismo que están divididos en 3 escenarios como se muestra en la Figura 2.19. Cada escenario está repartido en una hoja de Excel como H0, H1, H2, en donde la H2 representa a los resultados del cluster número 3, la H1 representa al cluster número 2 y H0 a los resultados del cluster número uno, estos se encuentran agrupados en un archivo con el nombre de Tesis_2_medidores_2, posteriormente se procede a realizar el respectivo al almacenamiento de datos en el formato que se encuentra en la Figura 2.19. en donde a cada medidor le pertenece una distancia euclidiana, luego de realizar esto para las tres estaciones del año, una vez terminado el proceso de clasificación se procede a normalizar a cada uno de los datos, es decir, que se tendrá un total de 552 curvas las mismas que serán normalizas, utilizando el método de proporcionalidad, para ello se obtendrá el valor máximo para luego este valor ser dividido para cada uno de las distancias originalmente encontradas y de esta manera normalizar y posteriormente dividir en 2 grupos que serían las curvas con distancia euclideana sin parametrizar y el otro grupo estaría con las curvas normalizadas, luego de clasificar cada curva con su respectivo usuario se procede a organizar a cada una de estas curvas, posteriormente se procede a reducir la dimensionalidad de los datos con el método de minería de datos PCA.

Los perfiles de consumo presentados en la Figura 2.20 muestran las curvas de consumo de 184 usuarios industriales y comerciales no normalizadas, se observa el comportamiento de cada cliente con respecto al consumo eléctrico, la Figura 2.20 representa el proceso de minería de datos antes de aplicar PCA's, este procedimiento se realiza para la primera estación del año entre los meses de diciembre a marzo, cada uno de los perfiles presenta información del cliente como es su potencia activa, y su función se encuentra en las Ecuaciones 2.2 y 2.3.

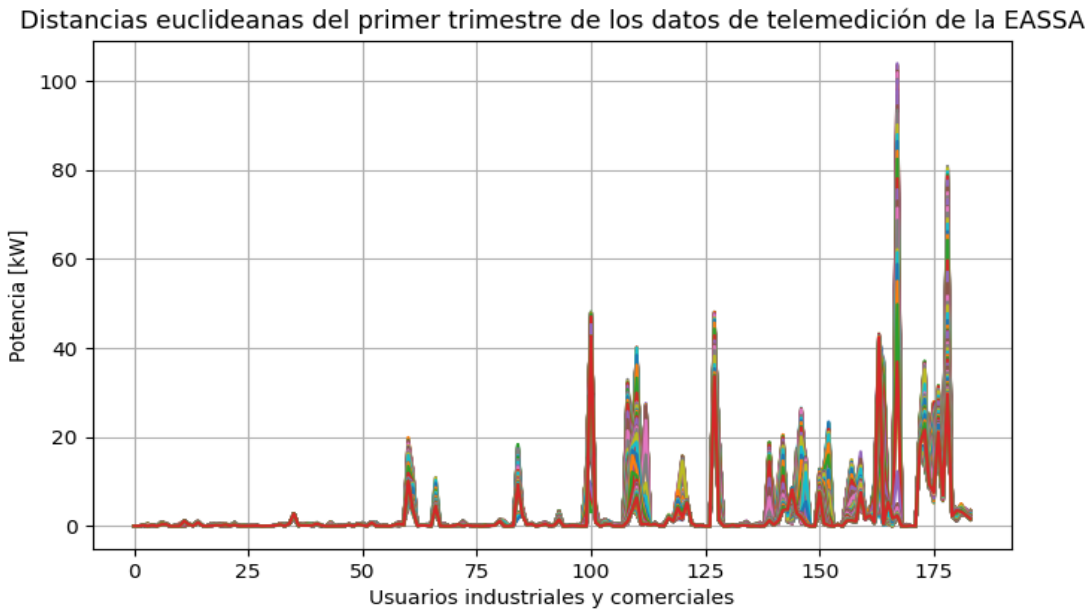


FIGURA 2.20 Perfiles de consumo de la Potencia activa [kW] de los usuarios industriales y comerciales no normalizados [Elaboración propia].

$$\boldsymbol{D_{(t)} = EOF(D(t))} \quad \textbf{(0.1)}$$

$$\boldsymbol{D_{(t)} = \sum_{n}^{i} A_i f_{eof}} \quad \textbf{(0.2)}$$

En la Figura 2.21 se muestra que cada una de las curvas se encuentran alienadas a un patrón, es decir, que cada curva esta alienada a la normalización, en donde su máximo es el unitario, cada perfil de consumo se va ajustando a su máximo parámetro que es 1. El método ajustado es el de proporcionalidad, y se encuentra en el intervalo de los valores de 0 y 1 respectivamente en el rango de {0; 1}, este alineamiento es de utilidad para un mejor tratamiento de los datos.

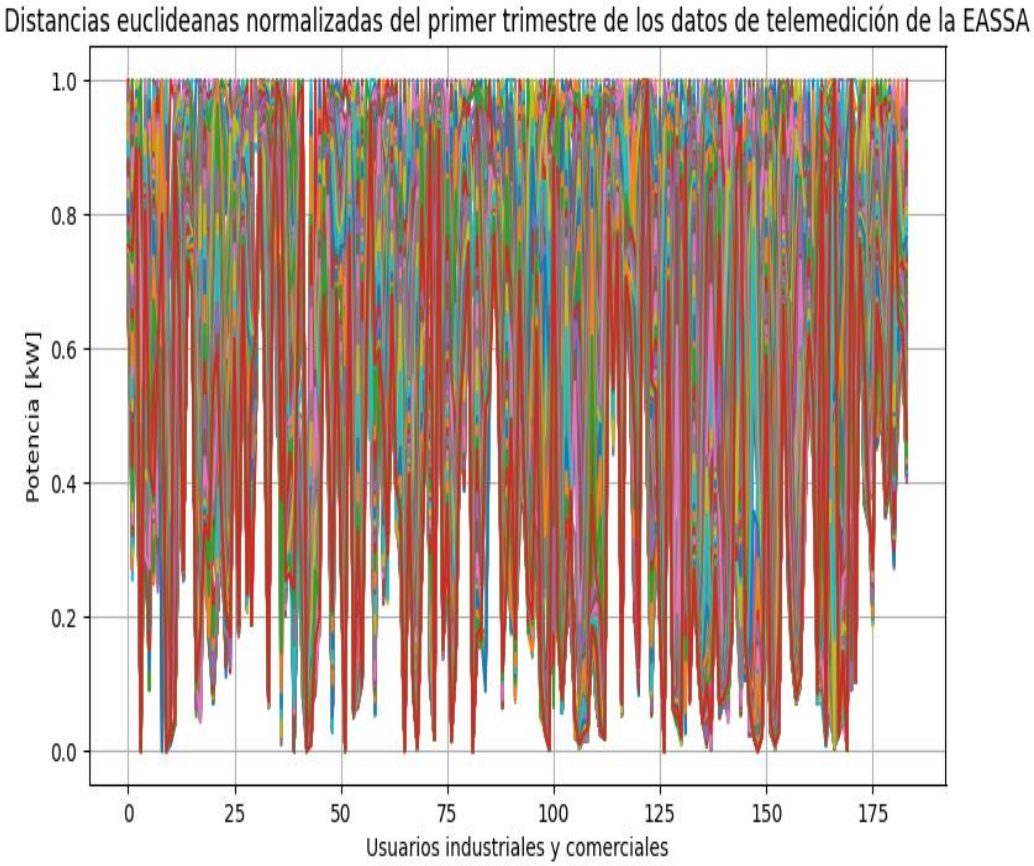


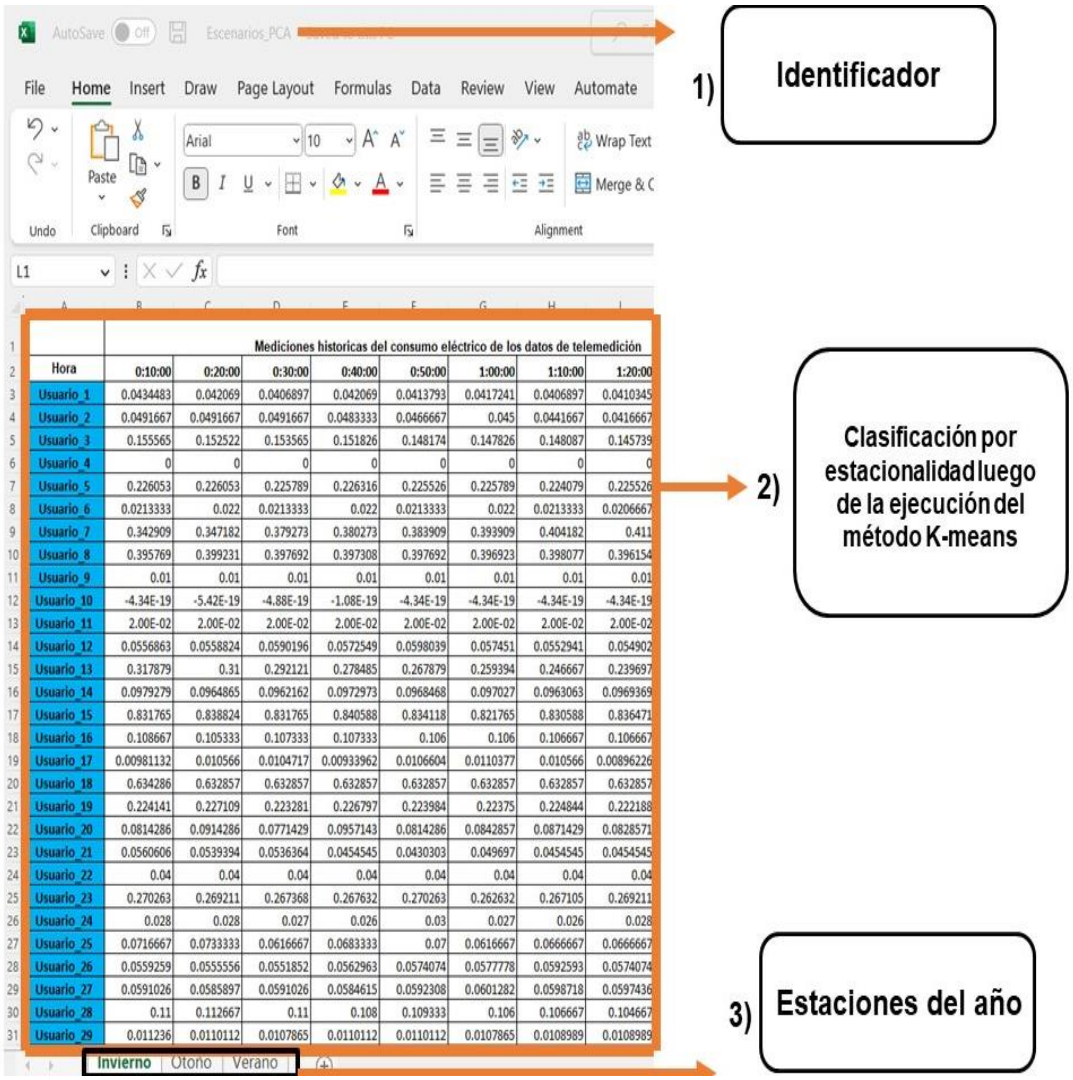
FIGURA 2.21 Perfiles de consumo de la Potencia activa [kW] de los usuarios industriales y comerciales normalizados [Elaboración propia].

2.14 FUNCIONES EMPÍRICAS ORTOGONALES

Los escenarios reducidos mediante este método de minería de datos son las curvas de consumo diario, que se encuentran previamente clasificadas según el apartado 2.11, como consecuencia de un proceso de ejecución anterior de K-means para las tres estaciones del año en los cuales están divididos las curvas de consumo.

Conforme al método explicado en el apartado 1.4.9.2. Usando una matriz de datos (nxp) que pertenece a una función discreta, donde (n) se trata de las telemediciones de consumo, mientras que (p) representa las observaciones en el tiempo. Dado que la matriz es la agrupación que corresponde al número de usuarios comerciales e industriales, de los cuales son 184 los mismos que se encuentran divididos en filas y las columnas son 144, los cuales serán simplificadas en 5 columnas en el dominio del tiempo, durante 1 año estadístico, la matriz F es una matriz rectangular [24].

Luego de realizar la clasificación por estacionalidad acorde al número óptimo de cluster, estas curvas son el resultado de la ejecución del clustering y se encuentran organizadas respectivamente conforme la estación del año en una hoja Excel de tipo .xlrd, estas curvas de carga se encuentran en intervalos de 10 minutos en los cuales han sido censadas durante 1 día, esto quiere decir que en un día se tiene 1440 minutos, los mismos que son divididos para 10 min, dando como resultado 144 intervalos de medición estos intervalos registran todos los usuarios residenciales y comerciales, el formato de acopio de información se encuentra en la Figura 2.22. posteriormente esta nueva base de datos se encuentra lista para realizar la reducción mediante el uso de minería de datos que son las PCA's, el archivo será cargado en Python 3.7 y se utilizará algoritmos que facilitaran su tratamiento:



Mediciones historicas del consumo eléctrico de los datos de telemedición

Hora	0:10:00	0:20:00	0:30:00	0:40:00	0:50:00	1:00:00	1:10:00	1:20:00
Usuario_1	0.0434483	0.042069	0.0406897	0.042069	0.0413793	0.0417241	0.0406897	0.0410345
Usuario_2	0.0491667	0.0491667	0.0491667	0.0483333	0.0466667	0.045	0.0441667	0.0416667
Usuario_3	0.155565	0.152522	0.153565	0.151826	0.148174	0.147826	0.148087	0.145739
Usuario_4	0	0	0	0	0	0	0	0
Usuario_5	0.226053	0.226053	0.225789	0.226316	0.225526	0.225789	0.224079	0.225526
Usuario_6	0.0213333	0.022	0.0213333	0.022	0.0213333	0.022	0.0213333	0.0206667
Usuario_7	0.342909	0.347182	0.379273	0.380273	0.383909	0.393909	0.404182	0.411
Usuario_8	0.395769	0.399231	0.397692	0.397308	0.397692	0.396923	0.398077	0.396154
Usuario_9	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
Usuario_10	-4.34E-19	-5.42E-19	-4.88E-19	-1.08E-19	-4.34E-19	-4.34E-19	-4.34E-19	-4.34E-19
Usuario_11	2.00E-02	2.00E-02	2.00E-02	2.00E-02	2.00E-02	2.00E-02	2.00E-02	2.00E-02
Usuario_12	0.0556863	0.0558824	0.0590196	0.0572549	0.0598039	0.057451	0.0552941	0.054902
Usuario_13	0.317879	0.31	0.292121	0.278485	0.267879	0.259394	0.246667	0.239697
Usuario_14	0.0979279	0.0964865	0.0962162	0.0972973	0.0968468	0.097027	0.0963063	0.0969369
Usuario_15	0.831765	0.838824	0.831765	0.840588	0.834118	0.821765	0.830588	0.836471
Usuario_16	0.108667	0.105333	0.107333	0.107333	0.106	0.106	0.106667	0.106667
Usuario_17	0.00981132	0.010566	0.0104717	0.00933962	0.0106604	0.0110377	0.010566	0.00896226
Usuario_18	0.634286	0.632857	0.632857	0.632857	0.632857	0.632857	0.632857	0.632857
Usuario_19	0.224141	0.227109	0.223281	0.226797	0.223984	0.22375	0.224844	0.222188
Usuario_20	0.0814286	0.0914286	0.0771429	0.0957143	0.0814286	0.0842857	0.0871429	0.0828571
Usuario_21	0.0560606	0.0539394	0.0536364	0.0454545	0.0430303	0.049697	0.0454545	0.0454545
Usuario_22	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04
Usuario_23	0.270263	0.269211	0.267368	0.267632	0.270263	0.262632	0.267105	0.269211
Usuario_24	0.028	0.028	0.027	0.026	0.03	0.027	0.026	0.028
Usuario_25	0.0716667	0.0733333	0.0616667	0.0683333	0.07	0.0616667	0.0666667	0.0666667
Usuario_26	0.0559259	0.0555556	0.0551852	0.0562963	0.0574074	0.0577778	0.0592593	0.0574074
Usuario_27	0.0591026	0.0585897	0.0591026	0.0584615	0.0592308	0.0601282	0.0598718	0.0597436
Usuario_28	0.11	0.112667	0.11	0.108	0.109333	0.106	0.106667	0.104667
Usuario_29	0.011236	0.0110112	0.0107865	0.0110112	0.0110112	0.0107865	0.0108989	0.0108989

FIGURA 2.22 Presentación y formato de la clasificación luego del proceso de K-means [Elaboración propia].

El formato en la Figura 2.22 que se muestra, corresponde al agrupamiento por estacionalidad, cada estación optará por el método de reducción de PCA's, estos datos son los datos de telemedición los mismos que serán preprocesados para el proceso de reducción, cada uno de ellos presenta un comportamiento distinto.

Los perfiles de carga reducidos se muestran la Figura 2.23, luego del proceso de minería de datos estos datos son equivalentes a la gran dimensionalidad de datos, por lo cual es mejor trabajar con un equivalente que servirá y ayudara al tratamiento de los mismos, por esta razón manejar abruptas cantidades de datos puede causar un error en el proceso de tratamiento de datos, de esta manera cuidando que la información se mantenga en las mismas condiciones sin perder algún dato alguno, también se debe considerar que si no se procesa de una manera exacta las curvas resultantes no entregan el producto esperado, el trabajar con una cantidad mínima de datos ayudará al tratamiento del mismo y una mejor operación dentro de los datos a tratar. Los datos se encuentran en el dominio del tiempo luego de haber sido reducidos, obteniendo un equivalente de 5 columnas de 144 columnas.

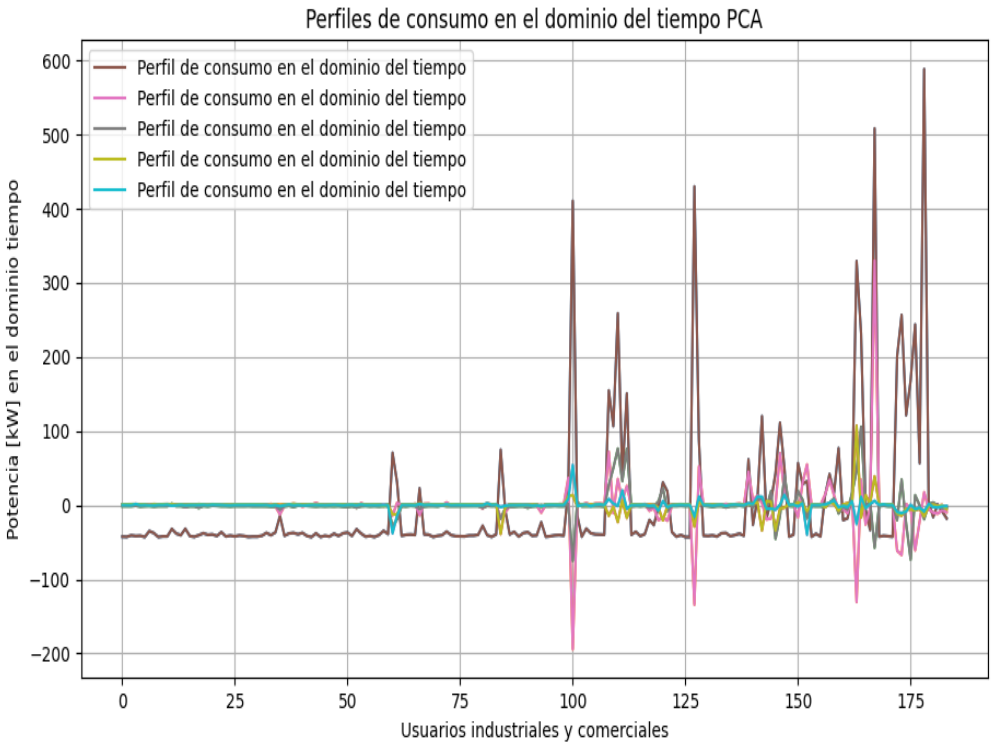
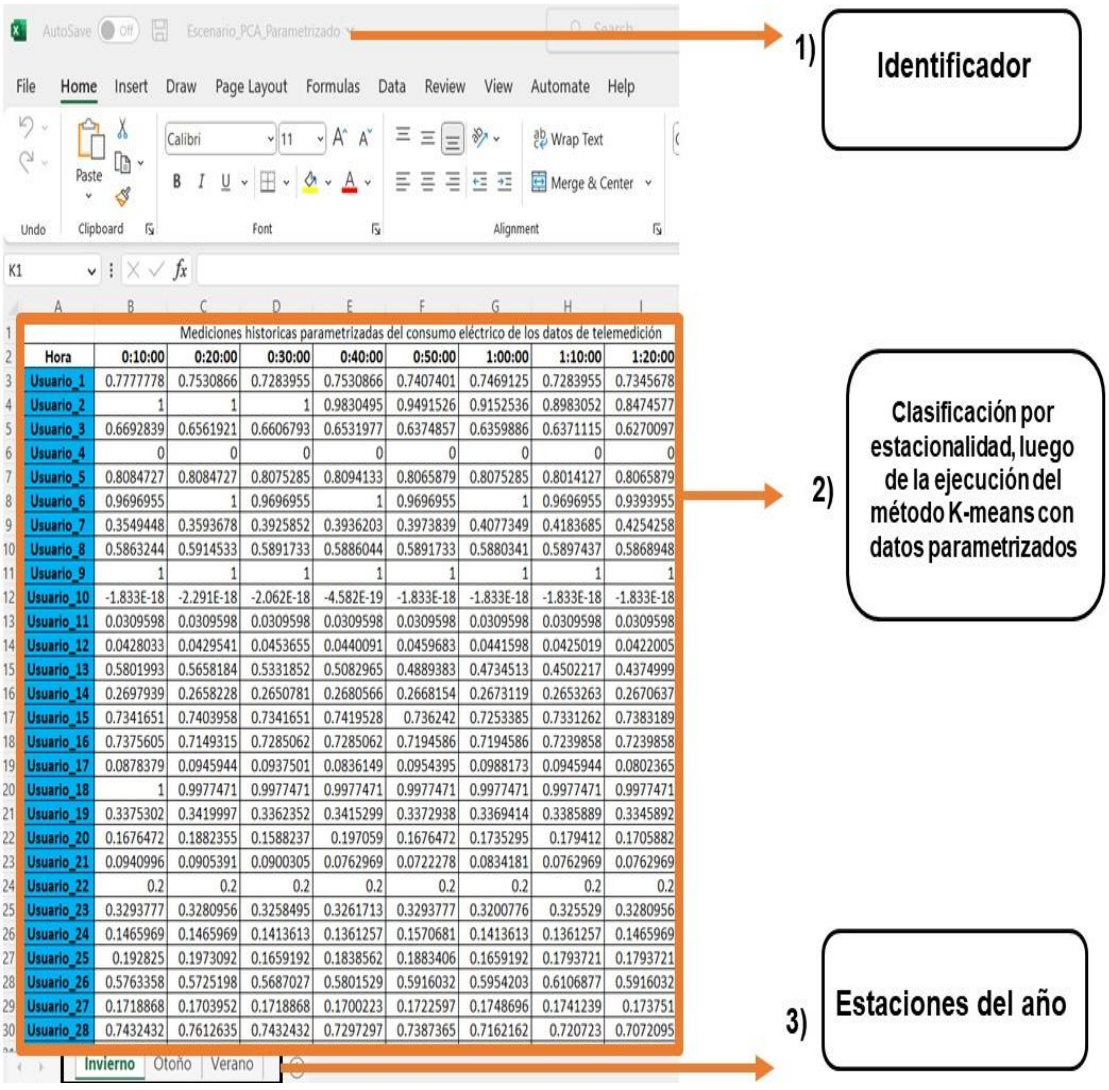


FIGURA 2.23 Datos en el dominio del tiempo a una reducción de 5 componentes [Elaboración propia].

Los datos presentados en la Figura 2.24 son distancias euclideanas normalizadas, que luego servirán para ser procesados de igual manera que el resto de los datos, estos datos parametrizados se obtiene con un alto margen de limpieza, el rango de los datos no supera entre [0-1], debido a su normalización como se pudo observar en la Figura 2.19, se puede observar que cada uno de los datos son divididos para su máximo, este mismo procedimiento se repite para invierno, verano y otoño este, se debe tomar en cuenta que cada escenario es distinto pero el formato de agrupación es el mismo para todos, su comportamiento es reflejado como se observa en la Figura 2.24.



Hora	0:10:00	0:20:00	0:30:00	0:40:00	0:50:00	1:00:00	1:10:00	1:20:00
Usuario_1	0.7777778	0.7530866	0.7283955	0.7530866	0.7407401	0.7469125	0.7283955	0.7345678
Usuario_2	1	1	1	0.9830495	0.9491526	0.9152536	0.8983052	0.8474577
Usuario_3	0.6692839	0.6561921	0.6606793	0.6531977	0.6374857	0.6359886	0.6371115	0.6270097
Usuario_4	0	0	0	0	0	0	0	0
Usuario_5	0.8084727	0.8084727	0.8075285	0.8094133	0.8065879	0.8075285	0.8014127	0.8065879
Usuario_6	0.9696955	1	0.9696955	1	0.9696955	1	0.9696955	0.9393955
Usuario_7	0.3549448	0.3593678	0.3925852	0.3936203	0.3973839	0.4077349	0.4183685	0.4254258
Usuario_8	0.5863244	0.5914533	0.5891733	0.5886044	0.5891733	0.5880341	0.5897437	0.5868948
Usuario_9	1	1	1	1	1	1	1	1
Usuario_10	-1.833E-18	-2.291E-18	-2.062E-18	-4.582E-19	-1.833E-18	-1.833E-18	-1.833E-18	-1.833E-18
Usuario_11	0.0309598	0.0309598	0.0309598	0.0309598	0.0309598	0.0309598	0.0309598	0.0309598
Usuario_12	0.0428033	0.0429541	0.0453655	0.0440091	0.0459683	0.0441598	0.0425019	0.0422005
Usuario_13	0.5801993	0.5658184	0.5331852	0.5082965	0.4889383	0.4734513	0.4502217	0.4374999
Usuario_14	0.2697939	0.2658228	0.2650781	0.2680566	0.2668154	0.2673119	0.2653263	0.2670637
Usuario_15	0.7341651	0.7403958	0.7341651	0.7419528	0.736242	0.7253385	0.7331262	0.7383189
Usuario_16	0.7375605	0.7149315	0.7285062	0.7285062	0.7194586	0.7194586	0.7239858	0.7239858
Usuario_17	0.0878379	0.0945944	0.0937501	0.0836149	0.0954395	0.0988173	0.0945944	0.0802365
Usuario_18	1	0.9977471	0.9977471	0.9977471	0.9977471	0.9977471	0.9977471	0.9977471
Usuario_19	0.3375302	0.3419997	0.3362352	0.3415299	0.3372938	0.3369414	0.3385889	0.3345892
Usuario_20	0.1676472	0.1882355	0.1588237	0.197059	0.1676472	0.1735295	0.179412	0.1705882
Usuario_21	0.0940996	0.0905391	0.0900305	0.0762969	0.0722278	0.0834181	0.0762969	0.0762969
Usuario_22	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2
Usuario_23	0.3293777	0.3280956	0.3258495	0.3261713	0.3293777	0.3200776	0.325529	0.3280956
Usuario_24	0.1465969	0.1465969	0.1413613	0.1361257	0.1570681	0.1413613	0.1361257	0.1465969
Usuario_25	0.192825	0.1973092	0.1659192	0.1838562	0.1883406	0.1659192	0.1793721	0.1793721
Usuario_26	0.5763358	0.5725198	0.5687027	0.5801529	0.5916032	0.5954203	0.6106877	0.5916032
Usuario_27	0.1718868	0.1703952	0.1718868	0.1700223	0.1722597	0.1748696	0.1741239	0.173751
Usuario_28	0.7432432	0.7612635	0.7432432	0.7297297	0.7387365	0.7162162	0.720723	0.7072095

Figura 2.24 Presentación y formato de la clasificación luego del proceso de K-means [Elaboración propia].

Los perfiles normalizados en el rango de {1} también son procesados por el método de reducción mediante el uso de PCAs. Una vez reducido la cantidad de datos de gran

dimensionalidad que son 184 usuarios por 144, como se muestra en la Figura 2.25, muestran que los perfiles del dominio del tiempo cumplen su reducción de gran dimensionalidad, cada una de las curvas se encuentran el dominio del tiempo para luego ser utilizados para el siguiente paso de obtención de las curvas de demanda estacionales.

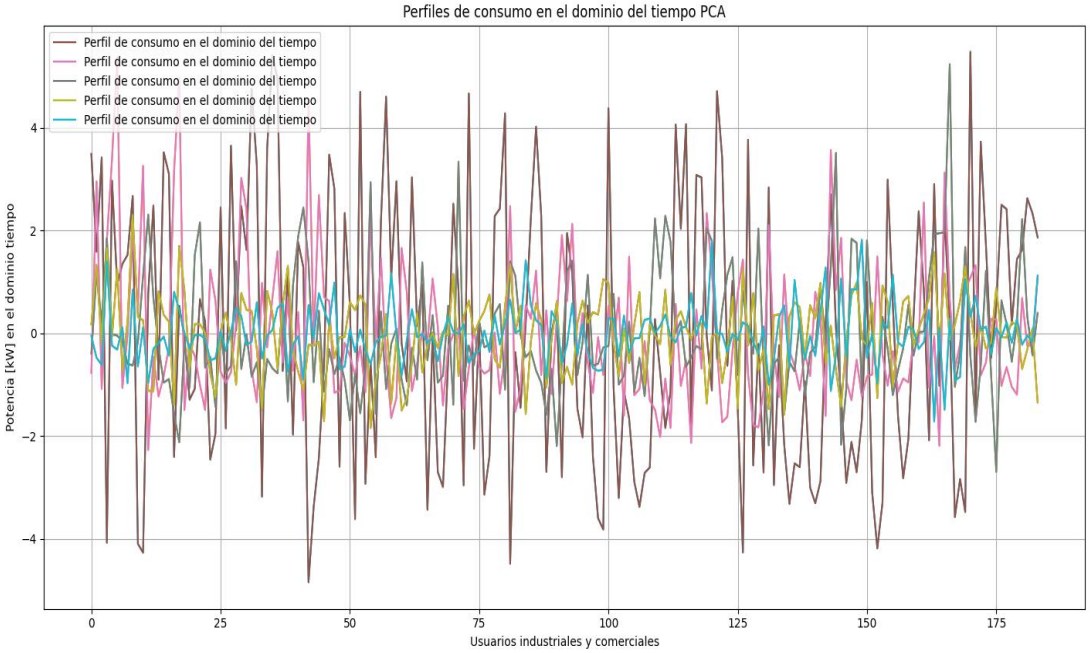


FIGURA 2.25 Datos de forma parametrizados los mismos que han sido normalizados [Elaboración propia].

Una vez realizado la minimización de los datos, cada uno de los datos que se encuentran cercanos al centro de cada cluster, la Figura 2.26 representa la información de un perfil de carga, mientras más cercano del centro se encuentren los datos, es que el cluster tiene menor dispersión y sus datos se son óptimos. Esto quiere decir que la distancia hacia cada cluster es menor y mientras menos disperso son los datos esto quiere decir que el número de cluster es el correcto. Luego de realizar la minimización datos mediante minería de datos se procede a realizar una clusterización con los datos reducidos, estos servirán para realizar una clasificación final que corresponderá al comportamiento final de cada usuario, y como se muestra en la Figura 2.26 cada vez los datos son menos dispersos y se ajustan al centro, este proceso de cluster se dio con un número de 3 como óptimo, esta agrupación muestra que los datos son cada vez menos dispersos y de esta manera comprobando que el método de PCA aún se mantiene efectivo.

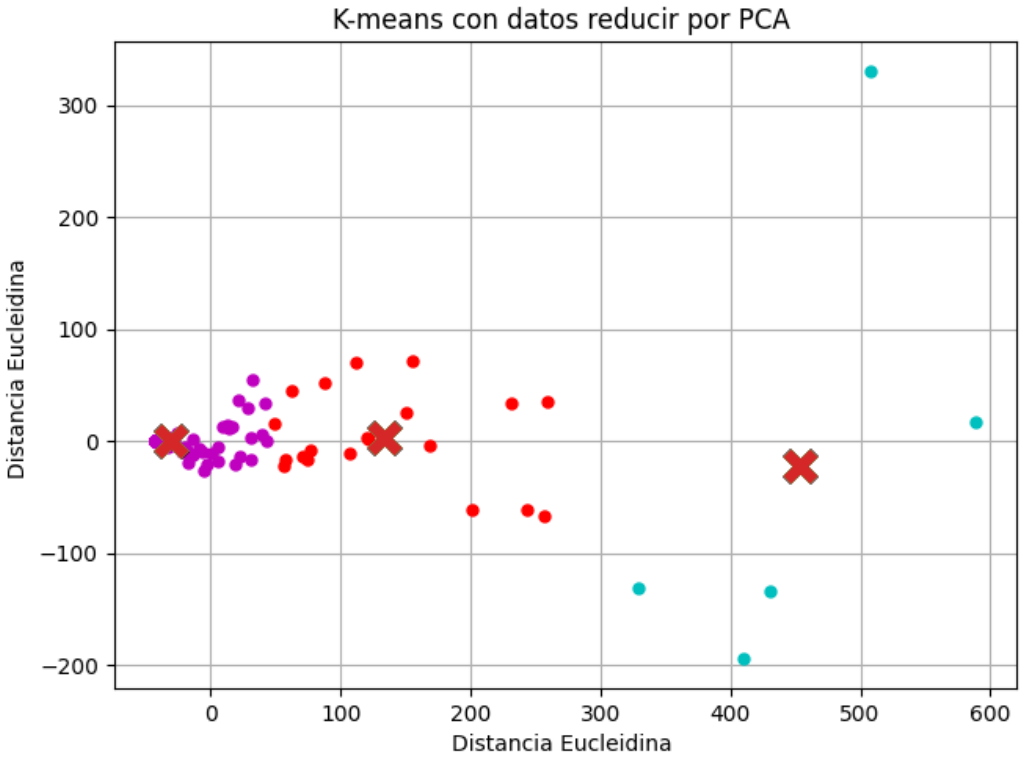


FIGURA 2.26 Presentación y formato de la clasificación luego del proceso de K-means [Elaboración propia].

La Figura 2.27 muestra el tratamiento de los datos y se confirma que cada uno de los datos al realizar el proceso de normalización, los datos son cada vez más limpios y confiables con su tratamiento. Cada uno de los grupos formados se encuentran ligados con cada centro, esto quiere decir que cada dato se encuentra cercano al centro dando un resultado confiable, es decir, que su agrupación es efectiva, cada centro lleva datos uniformes con cada uno de los grupos, el método K-means de los vecinos más cercanos se estaría cumpliendo, deduciendo que sus datos no presentan una alta cantidad de datos dispersos. Se realiza un cluster con los datos normalizados los mismos que muestran la agrupación homogénea, es decir, que el tratamiento de los datos está siendo eficaz de tal manera que cada elemento se junta a su centro, de los cuales son 3 centros los mismos que son 3 cluster's el número optimo del mismo, luego de realizar el proceso de conglomerados cada elemento muestra un comportamiento en el dominio del tiempo, cada punto que se encuentre cercano al centro muestra una efectividad y limpieza de los datos. Los datos por escenarios son distintos ya que cada uno de ellos presenta un comportamiento diferente con el método de conglomerados.

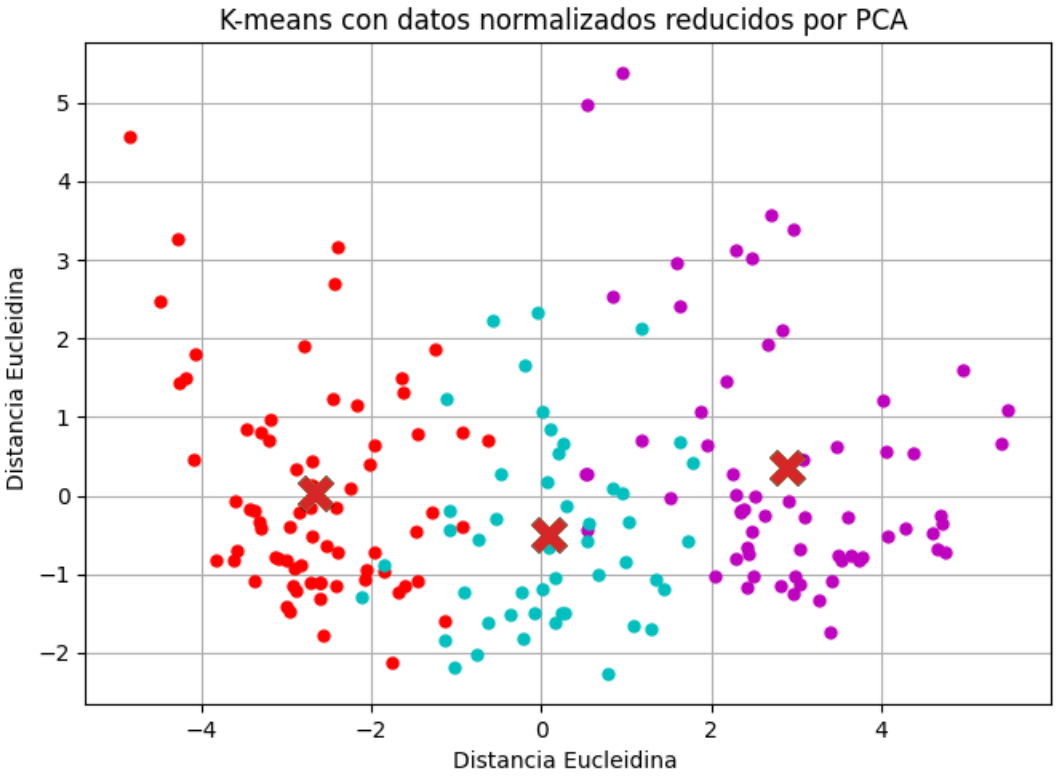


FIGURA 2.27 Presentación y formato de la clasificación luego del proceso de K-means [Elaboración propia].

2.15 FUZZY-C-MEANS

La herramienta Fuzzy-c-means es muy útil para realizar el proceso de clasificación por estacionalidad, dentro del proceso para continuar se la matriz de datos en el dominio del tiempo, reducida mediante PCA's, posteriormente la matriz de datos es exportada a la interfaz de programación MATLAB, el proceso de selección con Fuzzy C-means asigna una ponderación a cada valor, es decir, que coloca un valor mediante la función objetivo y estos son repartidos mediante la condición de inicio.

El proceso de asignación de valores de pertenencia de cada uno de los clientes mediante los valores Fuzzy-c-means, es asignado mediante los valores de la función objetivo, como se observa en la Tabla 2.6, los valores introducidos son los datos luego del proceso de reducción dado por las PCA's dado en 5 componentes principales en función del tiempo, cada uno de estos datos son pre procesados para luego de esto realizar el proceso de retrocesión en el tiempo, para de esta manera obtener las curvas que serán introducidas en el sistema ADMS, el conjunto de datos es de tipo difuso que es una clase de objetos con un grado de pertenencia continuo, tales conjuntos se caracterizan por una función de

pertenencia [23],como se muestran en la Tabla 2.6, cada grado de pertenencia se encuentra resaltado por datos de distinto color, estas etiquetas sirven para identificar el tipo de cliente para un análisis posterior.

TABLA 2.6. Presentación y formato de la clasificación luego del proceso de Fuzzy-c-means [Elaboración propia].

Pertenecia	f_1	f_2	f_3	f_4	f_5
1	1.9105	-0.751414	0.236997	0.0793581	-0.0287466
2	3.8332	-0.499204	-0.976643	0.0524966	0.294954
3	3.44417	-0.685247	0.387495	-0.266253	-0.57224
4	-5.05941	4.44308	0.430873	0.140846	0.491771
5	2.71308	3.55841	-0.688073	0.250499	-0.224817
6	0.590962	5.25252	-0.843383	1.70381	0.124756
7	1.55226	-0.950331	0.296176	-0.768868	-0.0278984
8	2.52203	0.423251	-0.10259	0.935184	-0.335218
9	-0.129768	-0.27009	3.24798	-1.15542	-0.543232
10	-4.13004	0.193697	-0.103364	0.498563	-0.52874
11	-1.51425	1.02168	-0.291811	-0.445289	0.00829913
12	4.18393	-0.383393	-0.681358	0.151001	-0.0916539
13	2.87625	-0.170706	0.38367	-1.01237	-0.412261
14	-0.839452	-1.2694	0.209784	0.843001	-0.130193
15	3.74837	-0.663505	-0.595101	0.349485	-0.00477675
16	-0.832592	-1.27009	-0.0434108	0.460424	-0.0545989
17	-2.16128	1.70376	-2.60631	-1.63611	0.974131
18	0.181756	2.3713	0.64923	-1.33696	0.806016
19	0.00268227	-1.46694	0.390388	0.812546	0.0507422
20	-1.30924	-0.408678	-1.01225	-0.591838	-0.138623
21	-2.74519	-0.503954	-1.31331	0.267131	-0.38539
22	-0.510667	4.72269	-2.42357	2.85308	1.45189
23	-0.0659035	-1.805	0.0888127	-0.300509	0.105277
24	-1.28637	-0.446696	-1.06381	0.137362	-0.380632
25	-1.57377	-1.57382	-0.395468	0.0259416	0.272033
26	2.45291	-0.481613	0.28786	-0.0216461	0.00242569
27	-1.74735	-1.36278	-0.0697095	0.651128	-0.137321
28	2.12213	-1.01675	-0.310681	-0.420344	0.114985
29	1.59821	-0.186448	0.861025	-0.526031	-0.0820072
30	0.416106	0.543135	-1.61296	-0.81275	-0.46379
31	4.59487	-0.338249	-0.552869	0.341359	-0.214608
32	4.58852	-0.549441	-0.484594	0.374427	-0.0278108
33	3.30846	-1.1186	-0.164471	-0.496952	0.5198
34	-3.29371	0.584665	-1.63117	-1.08593	-0.413224
35	2.79335	-0.319624	-0.0541056	0.631671	0.103643

1) Variables en el dominio del tiempo

2) Cada usuario es asignado con un valor de pertenencia dado por la aplicación del algoritmo Fuzzy c-means.

2.15.1 FUZZY-C-MEANS EN MATLAB

En esta etapa la agrupación de datos utiliza diferentes medidas de similitud para colocar elementos en clases, donde la medida de similitud controla cómo se forma las agrupaciones. El método para resolver es usando las líneas de comandos o una interfaz gráfica de usuario, este procedimiento se representa la Figura 2.28, esta ejecución sirve para encontrar los centros de los clústeres en MATLAB fundamentándose en la fcm:

[Center, U, obj_fcm]= fcm (data, cluster_n)

Los argumentos de esta función son:

1) *data – la base de datos de consumo eléctrico mediante la reducción por EOF's*

2) *cluster_ n: el número de clusters (debe ser mayor a uno)*

La función retorna los siguientes parámetros:

1) *center: la matriz de los centros de los conglomerados, donde cada fila contiene las coordenadas del centro de un grupo individual;*

2) *U: matriz resultante*

3) *obj_fcn: el valor de la función objetivo en cada iteración*

A continuación, en la Figura 2.28, se describe el algoritmo aplicado para el tratamiento de datos de los usuarios industriales y comerciales de la EEASA.

```
[Es_P_P2]=xlsread('Es_P_P2.xlsx');% Importamos la base de datos reducida mediante EOFS'S
[centers,U] = fcm(Es_P_P2,2); %Algoritmo proporcionado por Matlab
maxU = max(U);%Matriz resultante
index1 = find(U(1,:) == maxU);%Matriz resultante asigna el grupo mediante FCM
index2 = find(U(2,:) == maxU);%Matriz resultante asigna el grupo mediante FCM
plot(Es_P_P2(index1,1),Es_P_P2(index1,2),'ob')%Grafica del cluster generado
hold on
plot(Es_P_P2(index2,1),Es_P_P2(index2,2),'or')%Grafica del cluster generado

plot(centers(1,1),centers(1,2),'xb','MarkerSize',15,'LineWidth',3)%Grafica de los cluster generados
plot(centers(2,1),centers(2,2),'xr','MarkerSize',15,'LineWidth',3)%Grafica de los cluster generados
hold off
grid on
title('Fuzzy')
xlabel('Distancia Euclideana')
ylabel('Distancia')
```

FIGURA 2.28 Entorno de programación haciendo uso del algoritmo Fuzzy-c-means [Elaboración propia].

Luego de la ejecucion del algoritmo usando lenguaje de maquina se puede apreciar en la Figura 2.29, que cada uno de los datos se agrupa de manera eficaz, es decir, que cada uno de los vecinos mas cercanos conllevan una pertenencia sin mezclarse un dato con otro de cada uno de los centroides, siendo este un indicador del correcto uso del algoritmo Fuzzy-c-means

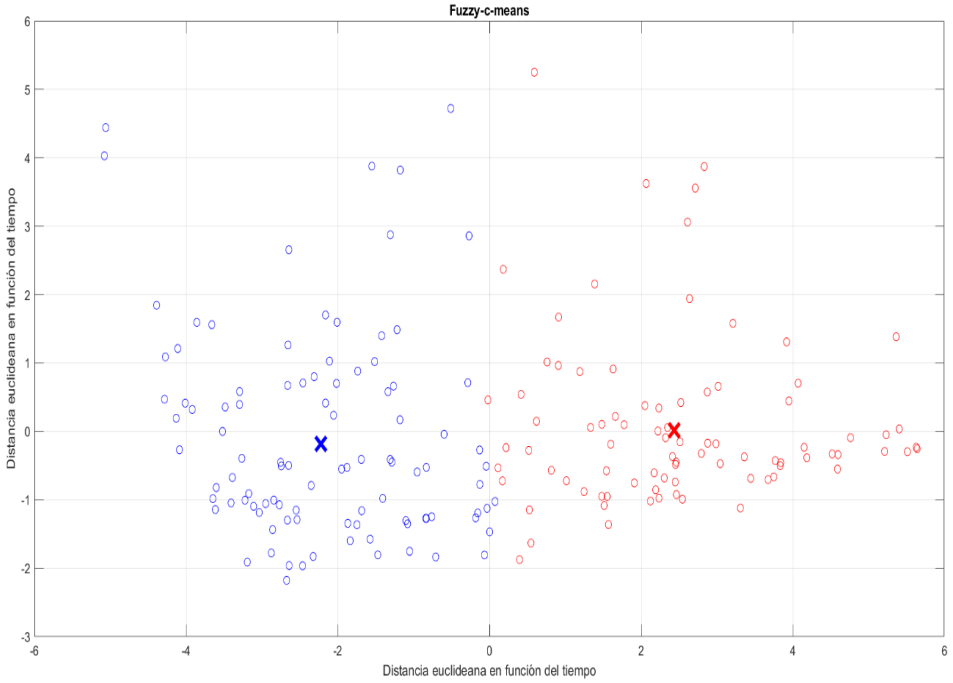


FIGURA 2.29 Datos de dispersión de Fuzzy – C – means [Elaboración propia].

3 RESULTADOS

En esta parte del desarrollo de este proyecto se presenta los resultados y el análisis de los perfiles de carga representativas que servirán para el mejoramiento del sistema ADMS para la detección de fallas tempranas que ayudará al operador a determinar acciones correctas, estas curvas obtenidas provienen de un proceso de tratamiento extenso de aplicación de algoritmos simultáneos, estos tienen como origen los datos de telemedición de los clientes comerciales e industriales de la Empresa Eléctrica Regional Centro Norte S.A. (EEASA).

3.1 GRUPACIÓN DE DATOS DE TELEMEDICIÓN MEDIANTE EL ALGORITMO DE K-MEANS

Los múltiples procesos permitieron el estudio de la base de datos de los usuarios comerciales e industriales, uno de ellos fue el tratamiento mediante el algoritmo K-means implementado en Python, se puede observar en la Figura 3.1 el comportamiento de los primeros 30 clientes que ingresaron bajo el proceso de análisis y cada uno de los clientes pertenece a un grupo de distinto distintivo (color), y lo que el algoritmo realiza es clasificar de acuerdo a su tendencia o similitud, además, sus centroides son asignados al azar y cada uno de ellos va cumpliendo un ordenamiento simultaneo, observando que los resultados son sensibles previo a su inicialización, también, uno de los aspectos relevantes es que al realizar el proceso de agrupación no se encuentran totalmente dispersos y las distancias euclideanas no llevan mucha diferencia, lo que conlleva a un óptimo resultado en la continuación de la ejecución del proceso para encontrar las curvas representativas.

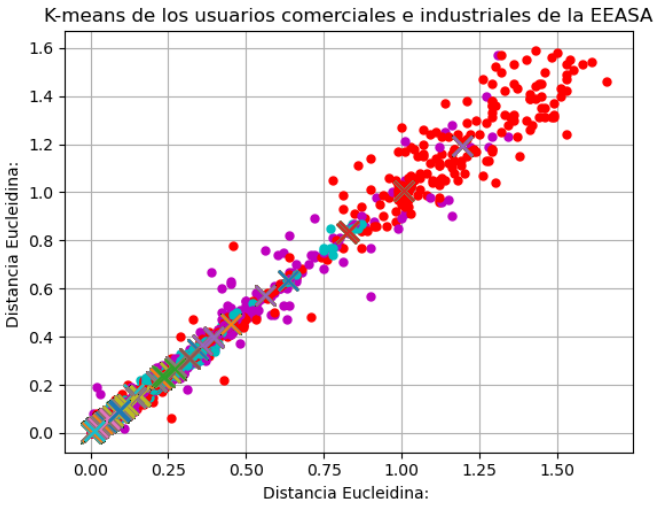


FIGURA 3.1 Ejecución del algoritmo K-means de los datos de consumo diario de los clientes comerciales e industriales de la EEASA [Elaboración propia].

3.1.1 ÍNDICE DE VALIDEZ ADECUADO PARA EL PROCESO DE DETERMINACIÓN DEL PERFIL DE CARGA

El resultado para escoger el número correcto de clusters se basa en el indicador DBI (Davies Bouldin Indice) como se muestra en la Figura 3.2, en el segundo trimestre de los meses de abril, mayo, junio y julio se muestra que el resultado recomendado es 2, mientras que, en el tercer trimestre, en los meses de agosto, septiembre, octubre, noviembre y diciembre el mejor indicador es 4. Por lo tanto, el número de cluster utilizado para el análisis de conglomerados fue de 3 clusters, de esta manera se simplifica el análisis de perfiles de consumo obtenidas.

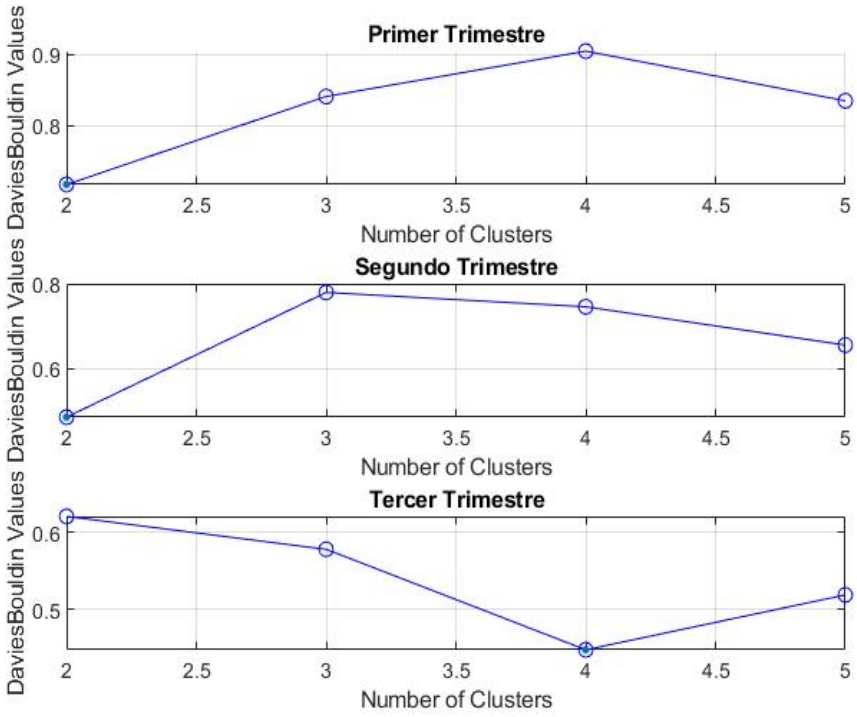


FIGURA 3.2 Algoritmo de agrupamiento aplicado a la clasificación de usuarios residenciales y comerciales de la EEASA [Elaboración propia].

En la Figura 3.3 se muestra el proceso de validación para la selección del cluster óptimo para el agrupamiento de la técnica por K-means, se aprecia que en la primera parte del primer trimestre el cluster recomendado por DBI es 2 el número óptimo, los resultados empeoran y en ciertas ocasiones generan datos dispersos que no están cerca de los

centros, mientras que en la segunda parte del trimestre se muestra que es 2 pero se muestra disperso, por otro lado en la tercera parte del año el número óptimo para el agrupamiento es de 4, en donde se puede apreciar que los centros ocupan la mayor parte de datos, mientras que los otros 2 centros del cluster se encuentran distante de los datos, llegando a la conclusión que el número optimo escogido en base a un resultado experimental es de 3 el número de agrupamiento optimo esto para los datos de consumo de los usuarios sin parametrizar, en donde se puede observar que cada cluster conlleva un agrupamiento óptimo, en donde los datos son menos dispersos y pertenecen o se acercan a cada centro en el cual el agrupamiento es eficiente, de esta manera cada punto que representa un perfil de consumo conlleva datos cercanos conllevando el método aplicado que es el algoritmo de los vecino más cercanos, presentando muy poca dispersión.

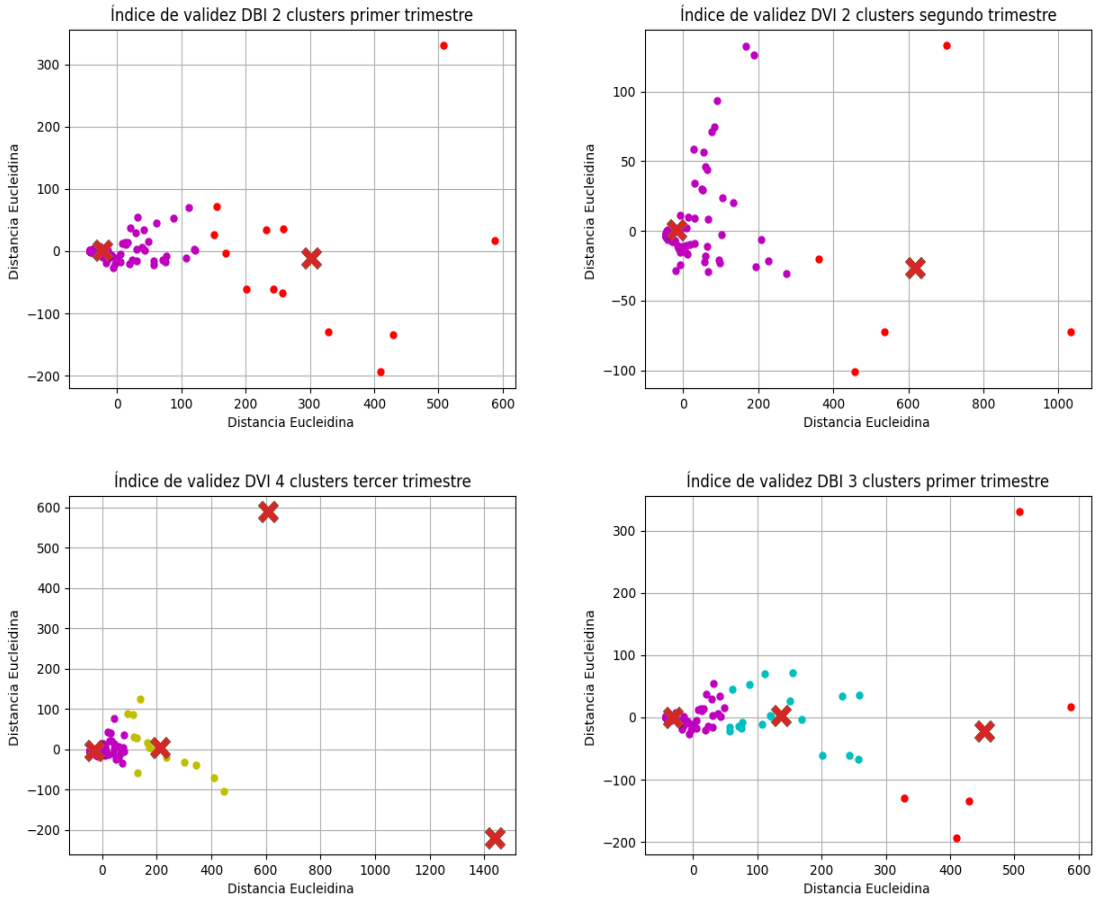


FIGURA 3.3 Índice DBI para la matriz de atributos de las 3 temporadas del año [Elaboración propia].

Se puede observar en la Figura 3.4, que para la selección del número óptimo de los datos normalizados que se encuentran entre el rango de [0-1], conllevan un comportamiento diferente, estos datos servirán para obtener un resultado más limpio en el proceso para las

3 estaciones del año, para el análisis de conglomerados mediante el método K-means es de 3 clusters, de esta manera se simplifica el análisis de perfiles de consumo, esto debido al análisis mediante el índice DBI muestra como resultado que el número óptimo es 2 en el primero trimestre del año, ahora en el segundo trimestre del año el índice indica que el número optimo es de 3, concordando con el análisis general que se realizó en base a los ensayos generados a la vez también corroborando con el número generado por el índice DBI, y en la última etapa del año el indicador muestra que es 2 el número de agrupamiento pero este argumento generaba datos que se alejan del centro, provocando como resultado final un desacertado agrupamiento en el proceso de escoger los perfiles de consumo adecuados, llegando a la conclusión que el número para el agrupamiento optimo es de 3.

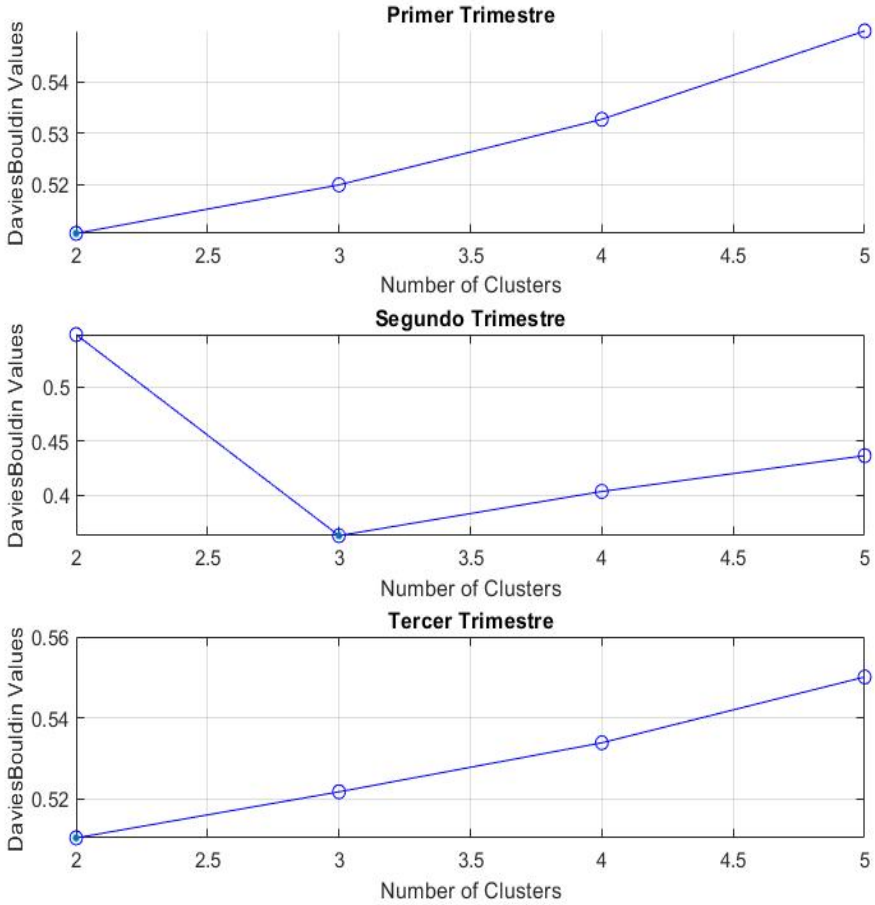


FIGURA 3.4 Índice DBI para la matriz de atributos de las 3 temporadas del año de los datos parametrizados [Elaboración propia].

En la Figura 3.5 se muestra el proceso de selección del cluster óptimo para el agrupamiento de la técnica por K-means, se aprecia que en la primera parte del primer trimestre el cluster recomendado por DBI es 2 el número optimo y se puede observar que los datos se agrupan muy cercanamente al centroide, mientras que en la segunda parte del trimestre se muestra

que el número optimo es 3, es decir, que con el agrupamiento de número optimo 3 mostraría un comportamiento eficiente, se muestra que los datos no se encuentran dispersos no existe una mezcla de datos, como se puede observar en el resto de escenarios todo se encuentra muy simétrico del centro de cada cluster, ahora en la tercera parte del año el número óptimo que recomienda el DBI para este tratamiento es de 2 clusters, en donde se puede apreciar que los centros ocupan la mayor parte de los datos agrupados a su alrededor pero con un alto número de elementos que pertenece a los datos que se encuentran muy dispersos, es decir, que se encuentran lejos de los centros de los clusters, esto provocaría que al seguir con el número óptimo de 2 clusters los datos serian inconsistentes, llegando a la conclusión que el número optimo escogido en base a un resultado experimental es de 3 el número de agrupamiento óptimo, en donde al existir un mayor número de centros conlleva que los datos se puedan agrupar de mejor manera con una pertenencia cercana a cada centro sin provocar que los datos se dispersen, se debe tomar en cuenta que este tipo de datos se encuentran parametrizados logrando obtener un mejor tratamiento de los mismos, que conllevaría a conseguir los perfiles de consumo que serían los más eficientes para seguir al paso de FCM, en donde si el proceso fue el correcto provoca obtener los perfiles de carga deseados presentado curvas de consumo que llevan la forma de una curva de consumo.

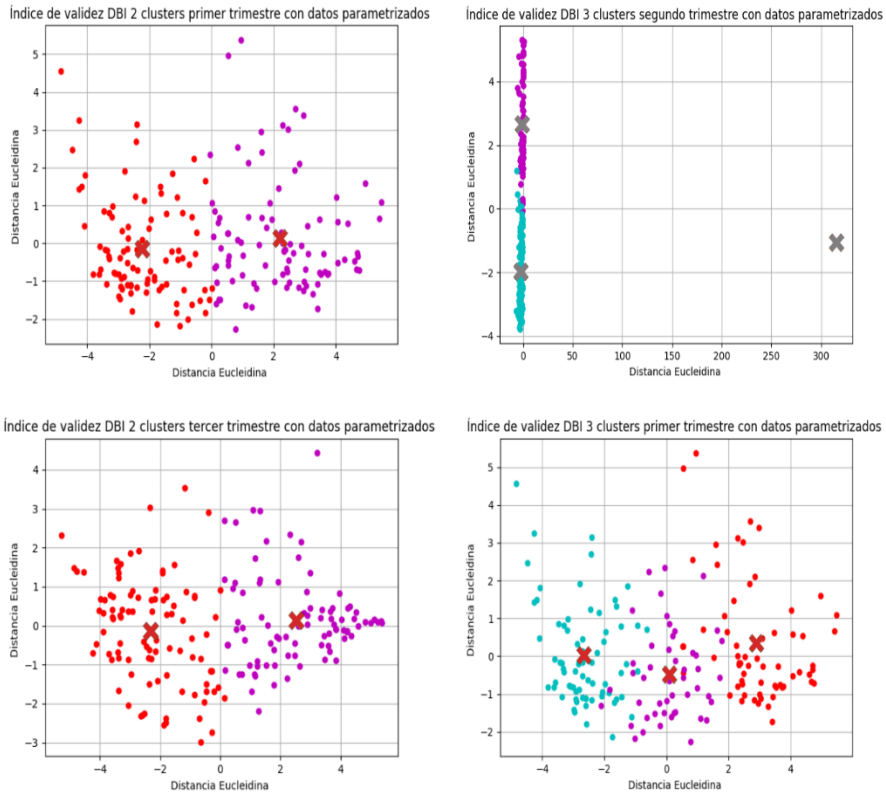


FIGURA 3.5 Índice DBI para la matriz de atributos de las 3 temporadas del año [Elaboración propia].

3.2 AGRUPACIÓN SEGÚN EL GRADO DE PERTENENCIA APLICANDO EL ALGORITMO DE FUZZY-C-MEANS

A razón de ejemplo se selecciona el resultado de la matriz de pertenencia U de los clientes no normalizados, para el segundo periodo del mes de abril a julio, la Tabla 2.7 muestra el resultado del procedimiento aplicado en el numeral 2.15.1, el resultado de las primeras 20 observaciones muestran que cada individuo tiene un valor de pertenencia asignado que como resultado es agruparse de acuerdo con sus atributos.

TABLA 3.1. Asignación del grado de pertenencia [Elaboración propia].

Individuo	Grado de pertenencia
1	0.99831835
2	0.998589963
3	0.998582133
4	0.99858322
5	0.998408641
6	0.998286964
7	0.999439026
8	0.998815728
9	0.998216225
10	0.998321186
11	0.998489721
12	0.999562616
13	0.998726639

14	0.998521898
15	0.999257103
16	0.998628137
17	0.998262342
18	0.998299557
19	0.998912819
20	0.998280871

3.2.1 IDENTIFICACIÓN DE LOS CLIENTES INDUSTRIALES Y COMERCIALES POR FUZZY-C-MEANS.

Después de haber ejecutado el algoritmo Fuzzy-c-means conforme el numeral 2.15.1, se toma como ejemplo, los resultados de las curvas de consumo de los 20 primeros clientes para el periodo de abril a julio de las curvas normalizadas, en la Tabla 2.7 se obtiene el grado de pertenencia juntamente con la asignación de cada cluster para los primeros clientes asignados como ejemplo, se puede observar que el resultado entregado identifica a cada cliente con una etiqueta de pertenencia y esa etiqueta es el cluster, para posteriormente agrupar cada uno de los datos y formas las curvas en el apartado 3.4. El método entrega un grado de pertenencia lo que hace que el resultado proporcione una acertada clasificación de los datos de telemedición.

TABLA 3.2. Asignación de cluster y del grado de pertenencia [Elaboración propia].

Cliente	Cluster	Grado de pertencia
1	2	0.950746888
2	2	0.918129747
3	2	0.943096494
4	1	0.717761382
5	2	0.744581748

6	2	0.540132215
7	2	0.8700477
8	2	0.95219935
9	1	0.512138258
10	1	0.911521277
11	1	0.884272533
12	2	0.916666051
13	2	0.948433393
14	1	0.772654658
15	2	0.930218079
16	1	0.791731177
17	1	0.720526107
18	2	0.535750035
19	1	0.543938154
20	1	0.882855539

3.3 OBTENCIÓN DE LAS CURVAS DE CONSUMO MEDIANTE LA REDUCCIÓN DE DATOS POR PCA'S APLICANDO EL ALGORITMO FUZZY-C-MEANS

Luego de reducir la matriz de datos que contiene la clasificación por estacionalidad con el método del apartado 2.14, el siguiente paso es la aplicación del método Fuzzy-c-means, explicado en la teoría en el numeral 1.4.7.1.2. Cada uno de los escenarios que son 3 trimestres en el año son clasificados mediante el algoritmo Fuzzy-c-means, este tratamiento es dependiente del número de centros en donde el grado de pertenencia es quien prevalece y gracias al mismo valor se logra la agrupación deseada.

Las curvas obtenidas como resultado del proyecto son mostradas a continuación, estas curvas de consumo se adaptan el comportamiento de consumo en los 12 meses del año, cada una de estas curvas es obtenida para los usuarios comerciales e industriales que mejora el análisis empírico antes incorporado por curvas con un estudio analítico.

En la Figura 3.6. se concentran las curvas para los usuarios normalizados, en la curva para los clientes de actividades de usuarios “Comerciales” se puede observar que su demanda en el intervalo de la mañana, específicamente de 10:00 – 11:00 coincide con la demanda de los clientes de tipo “Industrial”, mientras que para los usuarios de tipo comercial presentan una desconexión total al medio día de 12:00-13:00.

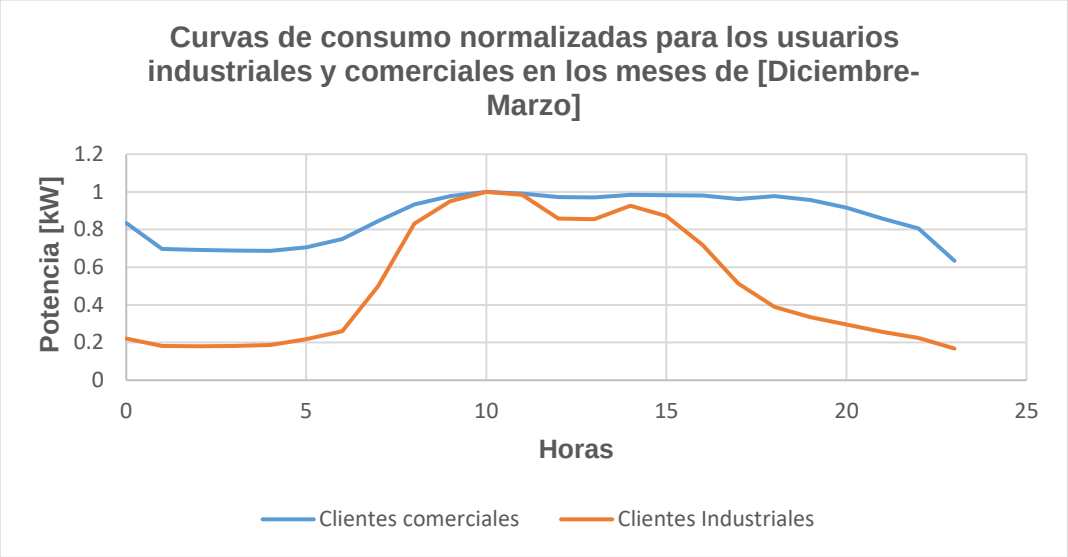


FIGURA 3.6 Perfil de carga representativo para los meses de diciembre-marzo [Elaboración propia].

En la Figura 3.7. se concentran las curvas para los usuarios normalizados, en la curva para los clientes de actividades de “Comercio”, presentan una similitud con la curva de consumo del primer trimestre en los meses de diciembre-marzo, se puede observar esta similitud debido a que los datos están normalizados y son más sensibles a la formación de curvas, siendo mínimo los valores en los cuales la demanda se diferencia un periodo de otro, por otro lado, en los clientes de tipo “Industrial” la desconexión para la segunda etapa del año se mantiene en el horario de 12:00-13:00 , con un pico en su demanda a las 10:00, con un pico concurrente a las 14:00.

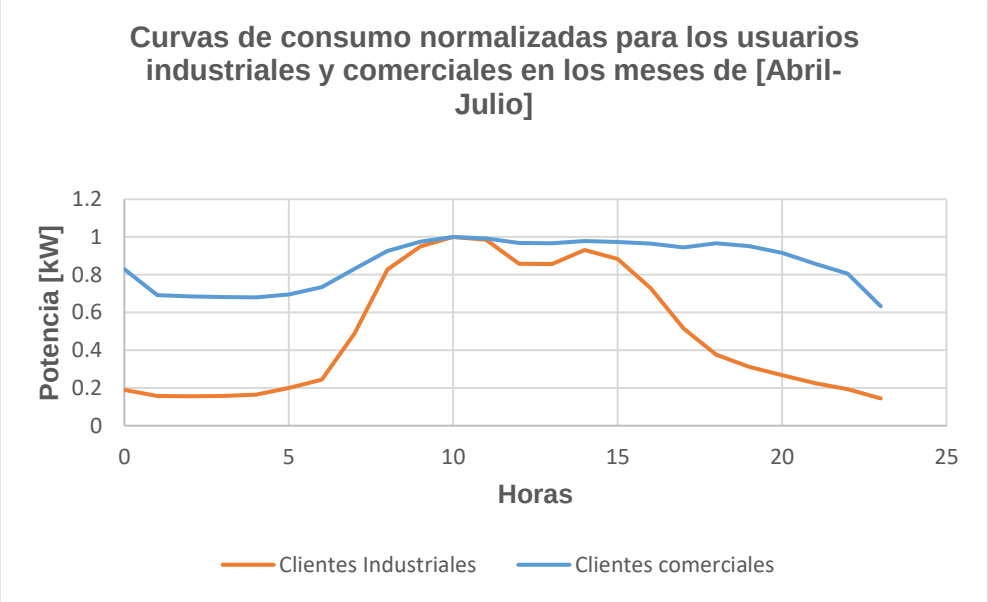


FIGURA 3.7 Perfil de carga representativo para los meses de abril-julio [Elaboración propia].

En la Figura 3.8. se puede observar que la curva para los clientes de actividades de tipo "Comercial" para la última etapa del año, en lo meses agosto-diciembre, indica que la curva al inicio del día del día en las horas de la 01:00 queda con un consumo residual alto del día anterior, mientras, que en las horas de 01:00-05:00, la demanda es relativamente baja en comparación con los otros meses como es en la Figura 3.6 y la Figura 3.7. Ahora con los clientes de tipo "Industrial", la demanda en la hora de la 01:00 es sumamente baja en comparación con el resto de los meses menciones, también, el intervalo de desconexión se mantiene en las horas de 12:00-13:00, este comportamiento se puede observar en la Figura 3.9.

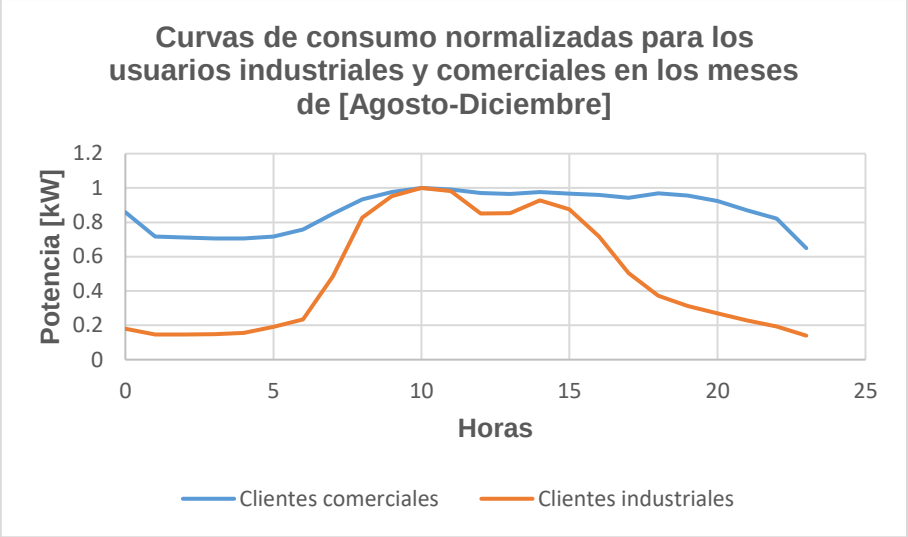


FIGURA 3.8 Perfil de carga representativo para los meses de agosto-diciembre [Elaboración propia].

En la Figura 3.9. se puede observar que la curva para los clientes de actividades de tipo "Industrial" para la primera etapa del año, en los meses de diciembre-marzo, indica que la curva al inicio del día del día en las 00:00 horas, la demanda inicial es baja, mientras desciende hasta las 06:00 para luego crecer exponencialmente presentando su pico más alto a las 11:00 horas y una desconexión total en un solo periodo de tiempo de las 13:00 horas, por lo tanto, en la curva de tipo "Comercial" muestra la típica curva de uso convencional con un uso continuo que inicia a las 00:00 con demanda baja para luego ir creciendo exponencialmente, presentando sus pico más altos de consumo eléctrico a las 11:00 horas y a las 15:00 horas.

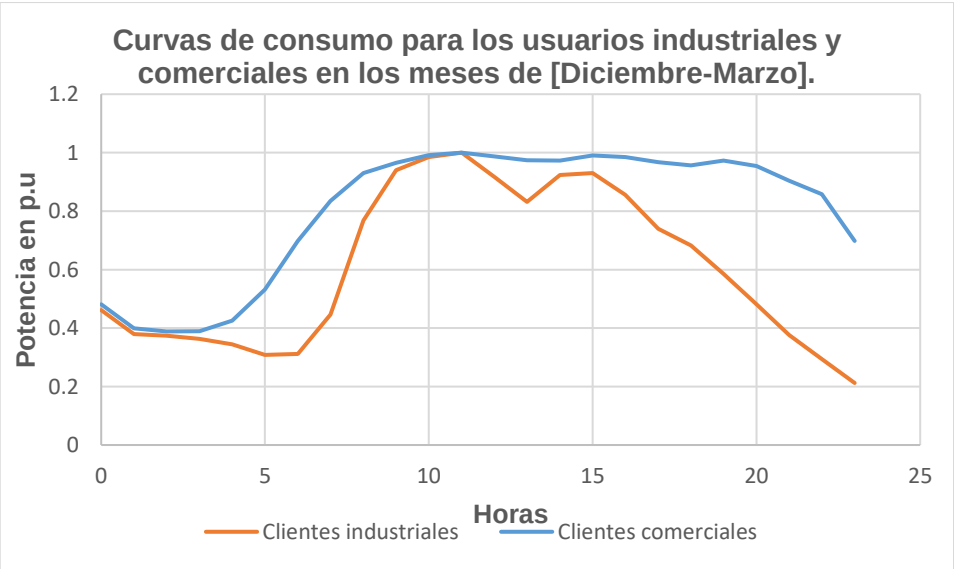


FIGURA 3.9 Perfil de carga representativo para los meses de diciembre-marzo [Elaboración propia].

En la Figura 3.10. se puede observar que la curva para los meses de abril-julio, para los clientes de actividades de tipo “Industrial”, indica que la curva al inicio del día en las 01:00 horas, es el valor con menor consumo de todo el escenario, mientras que la desconexión prevalece con la primera temporada a las 13:00 horas, con su pico más alto de consumo a las 11:00 am, por lo tanto, en la curva de tipo “Comercial” muestra un único pico alto que coincide con la curva de consumo “Industrial”, a las 11:00, se puede observar que en el segundo trimestre del año, no existe un intervalo de desconexión, más bien una hora fija.

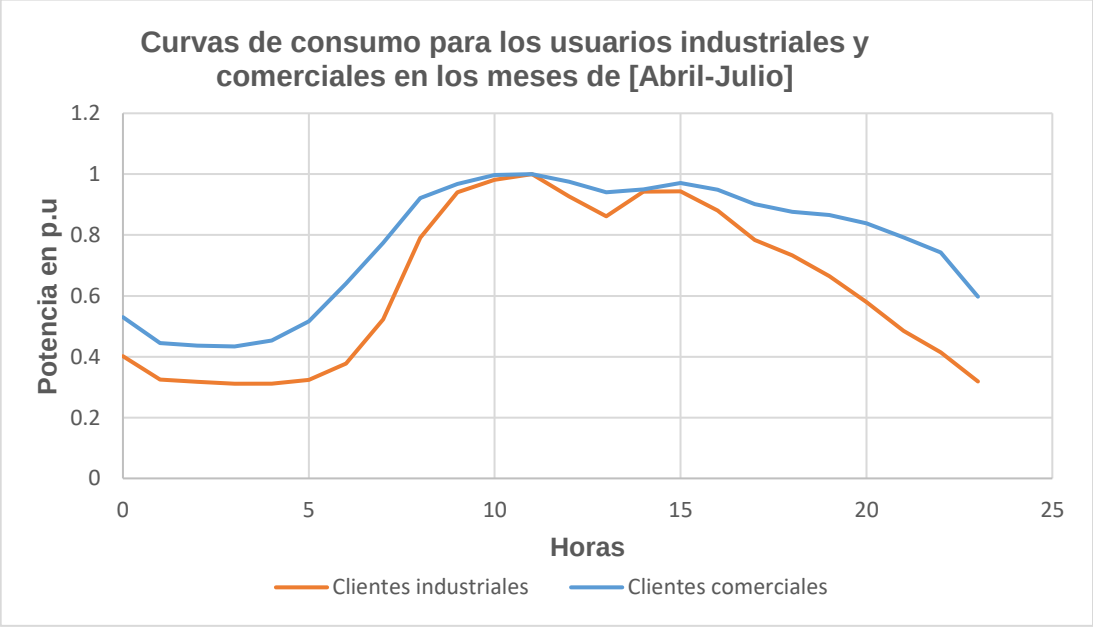


FIGURA 3.10 Perfil de carga representativo para los meses de abril-julio [Elaboración propia].

En la Figura 3.11. se puede observar que la curva para los meses de agosto-diciembre, para los clientes de actividades de tipo “Industrial”, indica que su demanda coincide en el intervalo de las horas de 00:00 – 04:00, mientras que la curva para los usuarios de tipo “Comercial” muestra dos picos en su demanda a las 11:00 horas y 15:00, sin embargo, la demanda de los dos tipos de clientes vuelve a coincidir en el intervalo de 09:00 a 11:00 horas, la desconexión prevalece con las 3 estaciones del año exactamente a las 13:00 horas, por ultimo vuelve a existir una coincidencia en las demandas en una hora fija a las 14:00 horas.

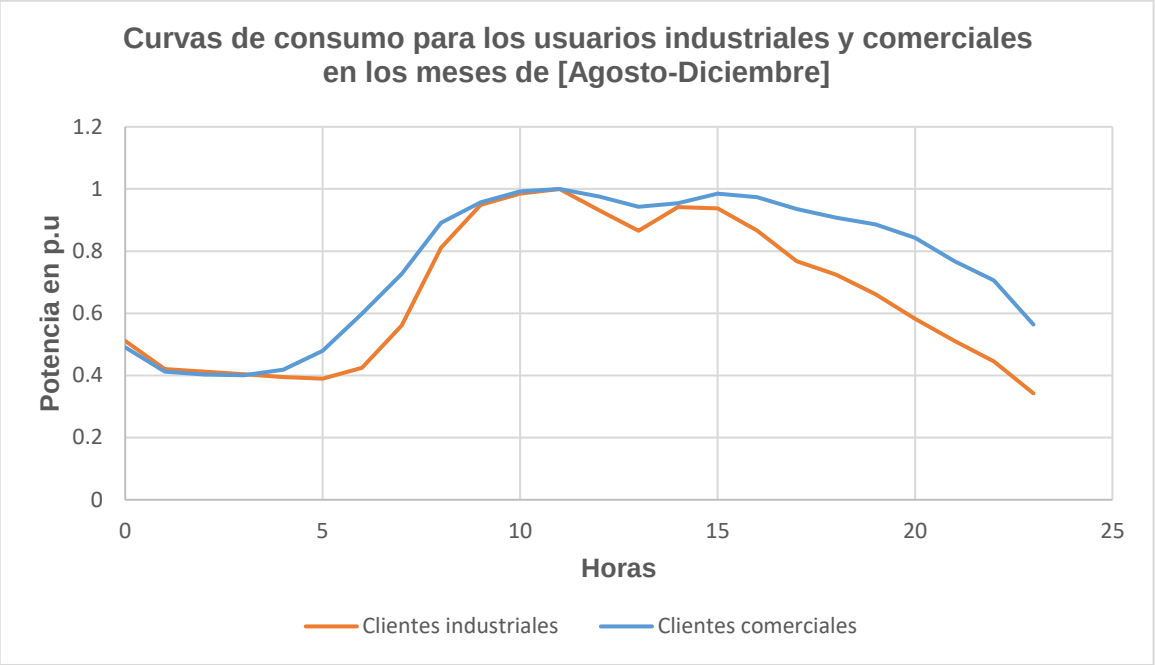


FIGURA 3.11 Perfil de carga representativo para los meses de agosto-diciembre [Elaboración propia].

4 CONCLUSIONES Y RECOMENDACIONES

4.1 CONCLUSIONES

En los meses de diciembre a marzo en el primer periodo del año con los datos normalizados, la curva consumo para los usuarios de tipo “Comercial”, muestra que existe una alta demanda en las horas de la mañana, como también un periodo de desconexión de un lapso de una hora en la curva de tipo “Industrial”, de esta manera las Industrias cuidan la vida útil de sus maquinarias y equipos, este consumo se relaciona directamente a la zona geográfica como es la ciudad de Ambato, debido a que durante el mes de diciembre existe una alta actividad comercial como industrial, que conlleva un lapso de 4 meses, en donde la producción en el parque industrial es relevante para producir una gran parte de productos para todo el año en curso.

Conforme a las curvas entregadas que serán incorporadas en el sistema ADMS, los clientes comerciales tienden a descender su consumo una vez que se encuentren fuera del horario laboral, es decir, que en horas de la noche y madruga su demanda siempre será de un bajo consumo, mientras que para los clientes de tipo “Industrial” presentan una desconexión sin un intervalo de tiempo, más bien en una hora exacta y los tres escenarios del año coinciden a las 13h00 horas, sin un periodo de descanso indicando que en esta temporada del año la producción a nivel industrial es relativamente baja.

La determinación de las curvas para los 3 trimestres del año contribuye a la optimización del DSE, el proceso ejecutado en el presente proyecto conlleva una clasificación con un criterio que obligadamente mejorará la sistematización del ADMS, activando la prevención de alarmas tempranas, en el comportamiento de los usuarios industriales y comerciales, de esta manera el operador de la EEASA obtendrá un criterio más real acerca del comportamiento del consumo eléctrico en todo el año, las curvas encontradas conllevan un estudio bajo varios criterios avanzados de clasificación y no una clasificación bajo ningún criterio experimental.

Se obtuvo como resultado que la demanda para los usuarios industriales durante los 12 meses del año coincide con la curva de los usuarios de tipo comercial, esto debido a que geográficamente en la ciudad de Ambato se encuentra el parque industrial más significativo del país, en donde los usuarios industriales abarca la mayor parte del consumo eléctrico, también, se debe considerar que como resultado se obtiene la curva de tipo comercial

cumple con los parámetros comunes de una curva de este tipo siendo un indicador adecuado para obtener resultados adecuados que mejoraran el estimador de estado.

La metodología actual en la EEASA no permite dar una fiabilidad al planificador a la hora de transferir carga, ya que el ADMS proporciona un informe de violación de límites, que permite conocer los elementos que se encuentran sobrecargados o por encima de los valores ingresados, esta sobrecarga se puede producir por dos razones: inadecuado ingreso de información o sobredimensionamiento del elemento, lo cual provocaría que los fusibles exploten, el accionamiento de alarmas y ese no es el caso. Con la realización de este proyecto el tratamiento de los datos entrego curvas normalizadas y curvas en p.u, estas curvas optimizaran la fiabilidad del estimador de estado mediante el uso de algoritmos de aprendizaje de máquina como fue en la ejecución de este proyecto.

La clasificación de usuarios comerciales e industriales que usan electricidad es esencial para las distribuidoras, y debido a la obtención de perfiles de carga se puede analizar su comportamiento en ciertas estaciones del año, desarrollando tarifas y estrategias en la gestión de la red, este trabajo propone una nueva metodología para agrupar clientes, los resultados presentan una tendencia a la hora de agrupar a los clientes y formar centroides.

4.2 RECOMENDACIONES

Se sugiere realizar un mantenimiento preventivo en los equipos que engloban todo el sistema de telemedición, así como también, el portal de descarga, debido a que a la hora de revisar los archivos Excel que genera esta plataforma, se encontró varias inconsistencias en los datos de la potencia activa, datos faltantes que en su magnitud fueron muchos, estos errores se relacionan con el fallo de los contadores de energía o varios sensores de energía no actuaron en el momento pertinente, estos datos faltantes se presentaron en horarios comúnmente presentados en horas de la madrugada.

Se recomienda seguir con la implementación de del sistema de telemedición, debido a que el uso de esta tecnología proporciona un mejor manejo eficiente del sistema ADMS y de esta manera continuar con el análisis de curvas de consumo para otros tipos de clientes, distintos días como los faltantes en este trabajo como son los días festivos y fines de semana.

Se encomienda realizar un análisis de clasificación preparatorio para que estas nuevas curvas puedan ser ingresadas al sistema ADMS con un porcentaje de asertividad acerca de los meses en los cuales la demanda de los usuarios cambia en donde los factores ambientales, el clima y sobre todo el comportamiento de los consumidores es distinto al ser la ciudad de Ambato un ciudad que se encuentra en la mitad de la región andina, donde muchas industrias dedicas al trabajo en acero necesitan un consumo elevado, se debe considerar que el comportamiento es muy distinto al resto de empresas que se encuentran fuera de este territorio.

Se propone la clasificación constante de las curvas de consumo para mejorar el estimador de estado, de esta manera ayudar al operador a cumplir sus funciones con acciones oportunas y anticipadas frente al comportamiento por temporada de cada uno de los usuarios evaluados en este proyecto, por lo cual, continuar con el análisis y la determinación de las curvas de consumo para los usuarios de tipo residencial, contribuye directamente al sistema de gerenciamiento de la EEASA. La implementación de la tecnología en los sistemas de distribución avanza y la EEASA al ser una pionera en este tipo de manejo de datos, efectivizará la clasificación de usuarios, la automatización con criterios relevantes sobre escoger la curva adecuada para los meses ayudará de gran manera a pronosticar el comportamiento de todos los tipos de usuarios.

5 REFERENCIAS BIBLIOGRÁFICAS

[1] A. Bermeo, «"Sistemas de simulación para la operación de redes eléctricas de Distribución en tiempo real" Trabajo de Titulación, Facultad de Ingeniería Eléctrica, Universidad de Cuenca,» 2018. [En línea]. Available: http://dspace.ucuencua.edu.ec/bitstream/123456789/30359/1/trabajo%20de%20titulacion.pdf.

[2] M. T. İŞYAPAR, «CLASSIFICATION OF ELECTRICITY CUSTOMERS BASED ON REAL CONSUMPTION VALUES USING DATA MINING AND MACHINE LEARNING TECHNIQUES AND ITS CORRESPONDING APPLICATIONS,» 2013.

[3] «"Network Manager SCADA.DMS. Distribution Network Management",» *ABBB,* vol. [En línea]. Available: https://library..e.abb.com/public/d812ff32efa92201852575fa00562955/BR_SCADA_DMS.pdf..

[4] J. C. CASTRO VÁZQUEZ, «INTEGRACIÓN DE SUBESTACIONES AL SISTEMA AVANZADO PARA EL MANEJO DE LA DISTRIBUCIÓN DEL ECUADOR,» *ESCUELA POLITÉCNICA NACIONAL, FACULTAD DE INGENIERÍA ELÉCTRICA Y ELECTRÓNICA,* p. 175, 2019.

[5] J. M. y. I. E. B. A., «"Xxxxiv Seminario Nacional del Sector Eléctrico",» 2019.

[6] V. Hinojosa, Pronóstico de Demanda a corto plazo en Sistemas de Suministro de Energía Eléctrica utilizando Inteligencia Artificial, 2007.

[7] A. Gallo, "Análisis predictivo para Minería de datos y proyección a corto plazo de la demanda de potencia en el Sistema Eléctrico Ecuatoriano",Trabajo de Titulación, Facultad de Ingenieria Eléctrica y Electrónica, Escuela Politécnica Nacional, Quito: https://bibdigital.epn.edu.ec/browse?type=author&value=Gallo+Cruz%2C+Angel+Omar., 2020.

[8] C. González, «"Predicción de la demandaeléctrica horaria mediante redes neuronales artificiales",» pp. 5-28.

[9] K. Kostková, L. Omelina, P. Kycina y P. Jamrich, «An introduction to load management,» *Electric Power Systems Research,* pp. 184-191, 2012.

[10] J. Á. Camacho, «Algoritmos, Inteligencia Artificial, Machine Learning,» 26 Agosto 2020. [En línea]. Available: https://www.jacobsoft.com.mx/es_mx/pre-

procesamiento-de-datos-con-python/. [Último acceso: 2021].. [Último acceso: 2021].

[11] R. Gago y P. Andrio, «"Uso de algoritmo de aprendizaje automático a base de datos genéticos", Trabajo de Titulación, Universitar Oberta de Catalunya,» vol. [En línea].http://openaccsess.uoc.edu/webapps/o2/bistream/10609/65426/6rgagoTFM0617memoria.pdf., Cataluña, http://openaccsess.uoc.edu/webapps/o2/bistream/10609/65426/6rgagoTFM0617memoria.pdf., 2017.

[12] UNAM, «¿Se puede medir de una forma más precisa la acummmulación o tenencia y la varibilidad?,» vol. [En línea]. Available: http://asesorias.cuautitlan2.unam.mx/Laboratoriovirtualdeestadistica/DOCUMENTO/TEMA2/1.MEDIDAS DE TENDENCIA CENTRAL Y DISPERSION.pdf.., 2017.

[13] S. Ramirez, «"Redes de Distribución de Energía", Universidad Nacional de Colombia,» Manizales, Tercera Edición, [En línea]. Disponible: http://blog.espol.edu.ec/econde/files/2012/08/libro-redes-de-distribucion.pdf., 2012.

[14] Z. Zakaria, K. Lo y M. Sohod, «"Application of fuzzy clustering to determine electricity consumers load profiles",» *First Int. Power Energy Conf.,* vol. [En Línea]. Disponible: https://ieeexplore.ieee.org/document/4154471, pp. 2-5, 2006.

[15] J. Rodríguez, Á. Suazo y I. Santelices, «"Análisis por medio de la normalización de variables para un modelo de planificación ambiental hídrica estacional",» *Obras y Proy.,* vol. 20, nº doi: 10.4067/s0718-28132016000200006, pp. 76-85, 2016.

[16] S. Raschka, Second Edition, 2015.

[17] J. S. N. Valenzuela, Aplicaciones prácticas, Ra-Ma 2018th..

[18] W. McKinney, Phton for Data Analysis, O'Reilly..

[19] J. Ávila, «"Procesamiento de Datos con Python",» Vols. %1 de %2[En línea]. Available: https://www.jacobsoft.com.mx/es_mx/pre-procesamiento-de-datos-con-python/.. [Último acceso: 01 Mayo 2021]., 26 de Agosto 2020.

[20] P. Virtanen, «SciPy 1.0 fundamental algorithms for scientific computing in Python,» *Nat. Methods,* vol. 17, nº 3, doi: 10.1038/s41592-019-0686-2, 2020., pp. 261-272

[21] L. J. Reina y R., "Tema 2: Introducción a scikit-learn", 2017.

[22] J. Cepeda, «"De Potencia en tiempo real usando técnicas de minería de datos y tecnología WAMS en el Centro Nacional de Control de Energía Ecuador D.G.

Colome",» de *Instituto de Energía Eléctrica-Universidad Nacional de San Juan-CONICET*, San Juan, pp. 1-8.

[23] J. C. R. González, «"Fuzzy C-Means Distribuido En la Nube",» vol. [Online] Disponible: http://openaccess.uoc.edu/webapps/o2/bitstream/10609/59066/7/ruizjcTFG0117memoria.pdf, p. 77, 2017.

[24] J. Chavarro, «“TECNICAS DE LOGICA DIFUSA APLICADAS A LA MINERIA DE DATOS. Fuzzy Logic Techniques for Data Mining,”,» nº x, pp. 1-6.

[25] J. Cepeda y D. Colome, «“Benefits of empirical orthogonal functions in pattern recognition applied to vulnerability assessment”,» *IEEE PES Transm. Distrib. Conf. Expo,* vol. 2014, 2014.

[26] J. E. Parra Flores, «ANÁLISIS DE CONGLOMERADOS DEL COMPARTIMIENTO DE LA DEMANDA ELÉCTRICA EN CLIENTES RESIDENCIALES UTILIZANDO DATOS DE MEDIDORES ONTELIGENTES,» Mayo 2017. [En línea]. Available: https://bibdigital.epn.edu.ec/bitstream/15000/15361/1/CD-7057.pdf. [Último acceso: 2021].

[27] T. F. Herrera y E. Delahiz, «"Aplicación de Minería de Datos para la CLasificación de Porgramas Universiarios de Ingeniería Industrial Acreditados en Alta Calidad en Colombia",» *Application of Data Mining for the Classification of University rograms Of Insdustrial Engineering Accredited,* vol. 29, nº 3, pp. 89-96, 2018.

[28] A. C. y Q. Juárez, «"Análisis del comportamiento bursátil de las principales bolsas financieras en el mundo usando el análisis multivariado (análisis de componentes principales pca) para el período de 2011 a 2014",» vol. 1, nº 2, pp. 25-36, 2015.

[29] «E. Informe, "Regionl Centro Norte S.a.",» vol. 03, 2021.

[30] L. González, «"Introducción a Machine Learning Ligdi González",» 1era edición, 2018, pp. 8-12.

[31] S. Ramos, J. Soares, S. Cembranel y Z. Vale, Data mining techniques for electricity customer characterization, 14 ed., Portugal: Procedia Computer Science, 2021, pp. 475-488.

[32] V. Vlacárcel, «DATA MININGYEL DESCUBRIMIENTO DEL CONOCIMIENTO,» *Revista de la Facultad de Ingeniería Industrial,* vol. 2, pp. 83-86, 2004.

[33] B. Escobar, Análisis y síntesis de audio espacial usando análisis de componentes principales, Quito, 2019.

[34] P. Pourrut, «LOS CLiMAS DEL ECUADOR . FUNDAMENTOS EXPLICATIVO,» Quito, 1983.

[35] B. Mayorga, «Segmentación de clientes usando datos de medidores inteligentes de electricidad».

[36] P. Laines y M. Oviedo, «METODOLOGÍA PARA CALCULAR LA DEMANDA MÁXIMA UNITARIA DE CLIENTES INDUSTRIALES Y COMERCIALES QUE CUENTEN CON SITEMA DE TELEMEDICIÓN,» Facultad de Ingeniería Eléctrica y Electrónica, Quito , 2021.

[37] W. Toussainta y D. Moodleyb, «Datos pare crear arquetipos que capturten comportamiento doméstico en Sadáfrica,» vol. TBC, nº TBC, 2020.

[38] Y. Wang, Q. Chen, C. Kang, M. Zhang, K. Wang y Y. Zhao, «Load Profiling and Its Application to Demand Response: A Review,» *TSINGHUA SCIENCE AND TECHNOLOGY,* vol. 20, nº 2, pp. 117-129, Abril 2015.

6 ANEXOS

Los anexos descritos a continuación se presentan en formato digital en el CD adjunto:

ANEXO A. Datos descargados del portal de la EEASA en punto .xls.

ANEXO B. Demanda de los 184 clientes en intervalos de 10 mediciones.

ANEXO C. Clasificación por temporadas del año (trimestre).

ANEXO D. Matriz de datos por temporadas sin normalizar.

ANEXO E. Matriz de datos por temporadas normalizadas.

ANEXO F. Resultados de la ejecución del algoritmo K-means.

ANEXO G. Resultados de la ejecución del algoritmo K-means con datos normalizados.

ANEXO H. Matriz reducida por PCA para los meses de abril - julio.

ANEXO I. Matriz reducida por PCA para los meses de agosto - diciembre.

ANEXO J. Matriz reducida por PCA para los meses de diciembre - marzo.

ANEXO K. Matriz reducida por PCA con datos normalizados en los meses de abril - julio.

ANEXO L. Matriz reducida por PCA con datos normalizados en los meses de agosto - diciembre.

ANEXO M. Matriz reducida por PCA con datos normalizados en los meses de diciembre - marzo.

ANEXO N. Matriz de datos con las curvas parametrizadas para los 3 trimestres del año.

ANEXO O. Matriz de datos con las curvas p.u para los 3 trimestres del año

ANEXO P. Código del programa en Python

ANEXO Q. Código del programa en Matlab

ORDEN DE EMPASTADO