UNIVERSIDAD POLITÉCNICA SALESIANA

SEDE CUENCA

CARRERA DE INGENIERÍA ELÉCTRICA

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

PROYECTO TÉCNICO CON ENFOQUE GENERAL:

METODOLOGÍA PARA LA IDENTIFICACIÓN DE SISTEMAS DE MEDICIÓN DE ENERGÍA ELÉCTRICA CON ERRORES DE REGISTRO DE CONSUMO DENTRO DE SISTEMAS DE DISTRIBUCIÓN

AUTORES:

CARLOS EDUARDO ARIAS MARÍN
DARÍO XAVIER GÓMEZ BRAVO

TUTOR:

MGS. PABLO ALEJANDRO MÉNDEZ SANTOS

CUENCA - ECUADOR

2019

CESIÓN DE DERECHOS DE AUTOR

Nosotros, Carlos Eduardo Arias Marín con documento de identificación N° 0104692504 y Darío Xavier Gómez Bravo con documento de identificación N° 0105361661, manifestamos nuestra voluntad y cedemos a la Universidad Politécnica Salesiana la titularidad sobre los derechos patrimoniales en virtud de que somos autores del trabajo de titulación: **METODOLOGÍA PARA LA IDENTIFICACIÓN DE SISTEMAS DE MEDICIÓN DE ENERGÍA ELÉCTRICA CON ERRORES DE REGISTRO DE CONSUMO DENTRO DE SISTEMAS DE DISTRIBUCIÓN,** mismo que se ha desarrollado para optar por el título de: *Ingeniero Eléctrico*, en la Universidad Politécnica Salesiana, quedando la Universidad facultada para ejercer plenamente los derechos cedidos anteriormente.

En aplicación a lo determinado en la Ley de Propiedad Intelectual, en nuestra condición de autores nos reservamos los derechos morales de la obra antes citada. En concordancia, suscribimos este documento en el momento que hacemos entrega del trabajo final en formato impreso y digital a la Biblioteca de la Universidad Politécnica Salesiana.

Cuenca, octubre del 2019

Carlos Eduardo Arias Marín

C.I.: 0104692504

Darío Xavier Gómez Bravo

C.I.: 0105361661

CERTIFICACIÓN

Yo, declaro que bajo mi tutoría fue desarrollado el trabajo de titulación: **METODOLOGÍA PARA LA IDENTIFICACIÓN DE SISTEMAS DE MEDICIÓN DE ENERGÍA ELÉCTRICA CON ERRORES DE REGISTRO DE CONSUMO DENTRO DE SISTEMAS DE DISTRIBUCIÓN,** realizado por Carlos Eduardo Arias Marín y Darío Xavier Gómez Bravo, obteniendo el *Proyecto Técnico con Enfoque General*, que cumple con todos los requisitos estipulados por la Universidad Politécnica Salesiana.

Cuenca, octubre del 2019

Mgs. Pablo Alejandro Méndez Santos

C.I.: 0102660578

DECLARATORIA DE RESPONSABILIDAD

Nosotros, Carlos Eduardo Arias Marín con documento de identificación N° 0104692504 y Darío Xavier Gómez Bravo con documento de identificación N° 0105361661, autores del trabajo de titulación: **METODOLOGÍA PARA LA IDENTIFICACIÓN DE SISTEMAS DE MEDICIÓN DE ENERGÍA ELÉCTRICA CON ERRORES DE REGISTRO DE CONSUMO DENTRO DE SISTEMAS DE DISTRIBUCIÓN,** certificamos que el total del contenido del *Proyecto Técnico con Enfoque General*, es de nuestra exclusiva responsabilidad y autoría.

Cuenca, octubre del 2019

Carlos Eduardo Arias Marín

C.I.: 0104692504

Darío Xavier Gómez Bravo

C.I.: 0105361661

DEDICATORIA

Con mucho cariño para mis papitos Julio y Esperanza por haberme brindado su confianza, cariño y apoyo en este largo proceso de estudio.
A mis abuelitos Tarquino y Piedad por haber sido quienes a base de consejos me dieron fuerzas para afrontar las barreras presentadas.
A Karen por ser mi apoyo, compañera de vida e inspiración.
A mis hermanos David y Juan Diego por ser mis mejores amigos y acompañantes en aquellas locuras de vida.

Carlos Eduardo Arias Marín

Dedico este proyecto a Dios, ya que sin Él nada en esta vida fuera posible realizarse, lo dedico a mis padres que a pesar de todas las diferentes pruebas que hemos tenido que superar, han estado siempre brindándome su apoyo incondicional, lo dedico a mis hermanas Adriana y Eva, a mi sobrina Valentina.

Darío Xavier Gómez Bravo

AGRADECIMIENTOS

Primero, quiero dar gracias a Dios por las bendiciones recibidas y por permitirme compartir estas alegrías con mi familia y amigos.

Mi enorme agradecimiento a mis padres, porque gracias a su esfuerzo, trabajo y amor, pude culminar mis estudios.

De igual manera a los Ingenieros Marco Toledo, Santiago Machado y Pablo Méndez por su guía y consejos para el desarrollo de este proyecto.

Agradezco a las personas que pertenecen al departamento de Control de la Medición de la E. E. CENTROSUR, por haber compartido sus experiencias en el campo, estas fueron de gran ayuda para conocer más del tema de las pérdidas no técnicas.

Agradezco a mi compañero Darío Gómez por su esfuerzo, colaboración y apoyo para el desarrollo de este proyecto.

Finalmente, quiero agradecer a los docentes de la Universidad Politécnica Salesiana, quienes supieron instruirme y guiarme para cumplir con esta meta de convertirme en Ingeniero Eléctrico.

Carlos Eduardo Arias Marín

Ante todo, quiero agradecer a Dios, Él cual me ha sabido brindar fortaleza en todo este camino de formación universitaria, ante cualquier caída me supo levantar y dar la fuerza para seguir adelante.

Quiero agradecer a las dos personas más importantes en mi vida que han sido los pilares fundamentales, mi padre Ángel Gómez el cual fue mi ejemplo a seguir, el cual me demostró con perseverancia y trabajo duro, se puede salir adelante en las condiciones más adversas posibles, a mi madre Imelda Bravo que me supo aconsejar y apoyar en los momentos que más lo necesitaba. Agradezco a Dios por la bendición de tenerlos a mi lado y seguir contando con su apoyo incondicional. Agradezco también a mis dos hermanas, Adriana Gómez que ha sido como una segunda madre para mí, que ha estado ahí siempre brindándome su apoyo; a mi hermana Eva Gómez y mi sobrina Valentina Espinoza que me han motivado a ser una buena persona como guía para ellas. A mis primos Christian Bustamante y Vanesa Qhizhpe por su apoyo. A mi mascota y amigo, mi perro Rocko que a pesar de las circunstancias siempre está para sacarme una sonrisa.

Quiero agradecer a todos los miembros de mi familia que de una u otra manera han estado ahí para brindarme su apoyo en todo momento.

Quiero agradecer a mis amigos, Paul Quezada, Gilberth Ramón, que más que unos amigos han sido mis hermanos, gracias por ser mis compañeros de aventuras.

Agradezco a la Ing. Cristina Mejía por su apoyo incondicional y desinteresado durante este proyecto.

Quiero agradecer a mis amigos y compañeros de Universidad, Ing. Luis Aguilar, Ing. Juan Pablo Torres que me supieron apoyar y brindar su amistad incondicional.

Agradezco de una manera muy especial al Ing. Pablo Méndez, Ing. Marco Toledo, y al Ing. Santiago Machado, tutores de este proyecto que supieron brindarnos su conocimiento y apoyo desinteresado, ya que sin su ayuda no hubiera sido posible la culminación de este.

Agradezco al Departamento de Control de la Medición de la Empresa Eléctrica Regional "CENTROSUR" quienes brindaron su ayuda en el transcurso del desarrollo de este proyecto. De una manera especial a la Lcda. Maribel Ríos, Ing. Marcelo León, Sr. Patricio Mendoza, Sr. Jorge Astudillo, Sr. Luis Guanuquiza, Sr. Carlos Maza, Sr. Jorge Cabrera, Ing. Boris Trelles, Sr. Carlos Guamán, Ing. Eddy Bravo.

Agradezco a Ing. Geovanny Mosquera, Ing. Juan Sánchez, Ing. Claudio Jara, y a la Srta. Karla Reinoso que nos supieron acoger en su despacho y fueron un apoyo incondicional para el desarrollo de este proyecto.

Agradezco a mis compañeros de Universidad, que conjuntamente pudimos afrontar las diferentes pruebas y obstáculos que se nos presentaron durante la carrera de Ing. Eléctrica; Daniel Molina, Juan Samaniego, Luis Aguilar, Andrés Argudo, Paul Paucar, Agustín Yubi.

Finalmente agradezco a mi compañero de tesis Carlos Arias, que sin su apoyo y colaboración no hubiera sido posible terminar todo este proceso que ha sido la vida universitaria.

Darío Xavier Gómez Bravo

RESUMEN

Es un problema común para las empresas distribuidoras de energía eléctrica, las pérdidas no técnicas; de ahí que estas empresas tengan que realizar un constante control orientado a mitigar esas pérdidas, que en definitiva representan menoscabo económico. Es por esto que las empresas distribuidoras de energía eléctrica, están constantemente analizando y aplicando técnicas de control y supervisión de los sistemas de medición. El presente proyecto tiene como objetivo diseñar una metodología que permita identificar, dentro de un sistema de distribución, los mecanismos para detectar los sistemas de medición anómalos, que provocan incremento en las pérdidas de energía, estudiando las técnicas que se emplean para el control de este problema, como los métodos orientados a datos: supervisados y no supervisados. Con ellos se analizan los datos técnicos y comerciales que se encuentran distribuidos en diferentes bases de datos dentro del sistema de comercialización de la Empresa Eléctrica Regional Centro Sur C. A. Información que analizada, depurada y segmentada crea una base de datos general que se denominó "matriz base". Con dicha información, se implementaron varias técnicas de agrupación y clasificación en MATLAB (software matemático), entre ellas: K-Medias, Red Neuronal, Árbol de decisión y K-Vecinos, con el propósito de identificar cuál o cuáles de estas técnicas son las que más se acoplan a los datos de la matriz base, mediante una evaluación con "métricas". Análisis que nos permite plantear una metodología y generar listas de sistemas de medición con alta probabilidad de ser causantes de pérdidas no técnicas, los cuales posteriormente son revisados en sitio. El resultado final fue convincente, encontrando algunas novedades en las inspecciones realizadas.

ABSTRACT

Non-technical losses are considered as a common problem for electricity distribution companies, hence, these companies have to carry out a constant control aimed at mitigating those losses, which ultimately represent economic impairment. That is why electricity distribution companies are constantly analyzing and applying control and monitoring techniques for measurement systems. The objective of this project is to design a methodology that allows identifying, within a distribution system, the mechanisms to detect anomalous measurement systems that cause an increase in energy losses. For this, it is necessary to study the techniques used to control this problem, such as data-oriented methods which can be supervised and unsupervised. These methods analyze the technical and commercial data that is distributed in different databases within the marketing system of the *Empresa Eléctrica Regional Centro Sur C. A.* This information is analyzed, debugged and segmented in order to create a general database called "base matrix". With this information, several grouping and classification techniques were implemented in MATLAB (mathematical software), including: K-Medias, Neuronal Network, Decision Tree and K-Neighbors, with the purpose of identifying which of these techniques are the ones that best fits the data of the base matrix, through an evaluation with "metrics". This analysis allows us to propose a methodology and generate lists of measurement systems with a high probability of causing non-technical losses, which are subsequently reviewed on-site. The final result was substantial, allowing us to find some news in the performed inspections.

INTRODUCCIÓN

La presencia de pérdidas de energía eléctrica en las redes de distribución es una situación cotidiana y regular en las empresas de distribución; las pérdidas de energía en esta etapa, son el resultado de restar la cantidad de energía disponible del sistema, de la que es medida y facturada por la empresa distribuidora a sus clientes o usuarios finales [1]. Su clasificación las agrupa en pérdidas técnicas y pérdidas no técnicas. Las pérdidas técnicas son efectos naturales que se deben a fenómenos físicos tales como la disipación de calor en los componentes eléctricos por el paso de la corriente, este tipo de pérdida se reduce mejorando la calidad de los equipos y la ingeniería en cuanto al diseño de las instalaciones [2], [3].

Por otra parte, las pérdidas no técnicas resultan de la diferencia entre las pérdidas totales y las pérdidas técnicas [1]. Este tipo de pérdidas no se originan en el proceso físico del transporte y/o transformación de la energía sino más bien tienen causas sociales por un lado y comerciales por otro. Como causas sociales se tienen: el hurto de energía, realizando conexiones directas a la red de la empresa distribuidora sin que el flujo de energía atraviese un medidor de energía, manipulaciones mal intencionadas en el medidor para que este registre menos cantidad de la energía otorgada, etc. Las causas comerciales, se relacionan directamente con una deficiente gestión administrativa por parte de la empresa distribuidora, errores en la medición, facturación o incluso una mala estimación y/o cálculo de las pérdidas técnicas [2], [4], [5].

Las pérdidas no técnicas causan un impacto económico en los procesos de distribución y consumo de energía, anualmente las empresas distribuidoras pierden millones de dólares, costo que generalmente es cubierto por los servicios públicos, o por parte de los consumidores legítimos que sin darse cuenta pagan por más energía que la suministrada [2], [3].

Para mitigar estas pérdidas no técnicas, los operadores de las empresas distribuidoras generalmente realizan inspecciones masivas de los sistemas de medición, siendo este un método que, si bien permite mantener una supervisión constante y bajo cierto control los niveles de pérdidas, muchas veces su proceso de planificación aún adolece de no contar con una estructura de análisis formales y sólidos que permitan alcanzar resultados más eficaces y eficientes.

Esta es la razón por la que se justifica el interés de aplicar técnicas de análisis de datos y definir una metodología que contribuya a la detección oportuna y eficaz de sistemas de medición que tengan una alta probabilidad de ser causantes de pérdidas no técnicas de energía [2], [4], [5].

La Empresa Eléctrica Regional Centro Sur C. A. es la distribuidora y comercializadora de energía eléctrica para las provincias de Azuay, Cañar y Morona Santiago, territorio que comprende aproximadamente el 11,79% del territorio nacional [6]. La empresa está organizada y estructurada administrativamente en varios departamentos, entre los cuales se encuentra el departamento de Control de la Medición, dependencia que tiene la responsabilidad de realizar la supervisión del

sistema de medición de usuarios finales, así como de tomar las acciones y/o gestiones comerciales que permitan mitigar las pérdidas no técnicas.

Este proyecto de grado, tiene como finalidad plantear una metodología de planificación y análisis de datos técnicos y comerciales de la Empresa distribuidora, que permita determinar listados de sistemas de medición con mayor probabilidad de contribuir a las pérdidas no técnicas; y con esto mejorar los procesos que actualmente se utilizan en la Empresa para el control de pérdidas.

En el Capítulo 1, se analizan los índices y estadísticas de las pérdidas eléctricas, se presentan datos porcentuales del estado situacional actual de la Empresa Distribuidora, respecto de las pérdidas de energía; además se estudia y describe el estado del arte respecto de las metodologías usadas para la mitigación de pérdidas no técnicas y finalmente se da una breve descripción de la metodología usada por el Departamento de Control de la Medición.

El Capítulo 2, describe el proceso de minería de datos y el estudio desarrollado a los datos técnicos y comerciales de la empresa distribuidora.

Finalmente, en el Capítulo 3, se propone una metodología creada en base al estudio realizado en el Capítulo 2 y se analizan los resultados obtenidos, por último, se propone una planificación para la aplicación de la metodología definida.

Al final del documento, se presentan las conclusiones y las recomendaciones que ayudarán a una mejora futura de la metodología planteada.

ÍNDICE GENERAL

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

1. CAPÍTULO 1 – PÉRDIDAS ELÉCTRICAS Y REVISIÓN DE METODOLOGÍAS UTILIZADAS PARA EL CONTROL DE PÉRDIDAS NO TÉCNICAS

Este capítulo hace una revisión del "estado del arte[1]" respecto de las pérdidas de "energía eléctrica[2]", su clasificación, situación actual en el Ecuador y una revisión de las metodologías utilizadas para la mitigación de las pérdidas no técnicas, para finalmente presentar una breve descripción de la propuesta para la Empresa Eléctrica Regional Centro Sur.

La cadena de suministro para proveer energía eléctrica y satisfacer la demanda, inicia con la etapa de producción en la cual a través de diferentes fuentes de energía primaria se genera energía eléctrica, las fuentes de energía son muy variadas pudiendo ser éstas de tipo renovable o no, como por ejemplo fuentes hídricas, solares, eólicas o térmicas de combustible fósil. La siguiente etapa es la de transporte en la cual la energía es transmitida a través de la red de transmisión, en esta etapa la conexión de la carga no es muy común existiendo generalmente solo la entrega de la energía a la etapa de distribución en la cual finalmente se conecta la carga de los usuarios finales del suministro eléctrico y en la cual se realiza la actividad de comercialización de energía. [7]

Todo este proceso implica el uso de elementos eléctricos (transformadores, líneas de tensión, cables, medidores, etc.) que deben estar relacionados y organizados en un sistema eléctrico de potencia grande y complejo, por lo que es normal que se originen pérdidas de energía, tanto en la etapa de transmisión, cuanto en la de distribución, siendo en esta última donde se encuentra el mayor porcentaje de pérdidas. [7]

Las pérdidas de energía eléctrica, se determinan por la diferencia que se presenta entre la energía producida en la etapa de generación y la energía medida y facturada al usuario final en la etapa de distribución. [1]

Las pérdidas de energía se clasifican en dos grupos:

- Pérdidas técnicas; y
- Pérdidas no técnicas.

1.1.PÉRDIDAS TÉCNICAS

Son pérdidas que se producen naturalmente debido a las condiciones propias del sistema eléctrico, provocadas por la circulación de corriente eléctrica a través de los elementos de las redes de transmisión y distribución. [1]

[1] Estudio documental que presenta los avances notables que se han logrado dentro de un área específica.
[2] Movimiento de cargas eléctricas (electrones) en el interior de materiales conductores.

Estas pérdidas no se pueden eliminar pues son inherentes a todo sistema eléctrico, sin embargo, se pueden reducir mejorando la calidad en equipos y/o estructura del sistema eléctrico; por otro lado, la cantidad de pérdidas dependerá del diseño de la red, niveles de voltaje, longitud y características de las líneas, etc., reflejando la eficacia del estado e ingeniería en cuanto a instalaciones eléctricas. [1], [3]

Existe muchos criterios para clasificar estas pérdidas [7], siendo la más común: [1], [7]

A. Pérdidas técnicas variables;
B. Pérdidas técnicas fijas;
C. Pérdidas técnicas de red.

1.1.1. Pérdidas técnicas variables

Todo elemento eléctrico tiene una resistencia interna y al momento que existe flujo de corriente por esta, se produce disipación de calor que varía dependiendo de la magnitud de la corriente. Estas pérdidas se denominan "pérdidas variables" o "pérdidas óhmicas" y se debe al "efecto Joule[3]". [1], [7], [8]

1.1.2. Pérdidas técnicas fijas

Se presenta en los componentes eléctricos y no necesariamente entregando energía, por el simple hecho de estar energizados producen pérdidas causadas por disipación de calor. Estas pérdidas son conocidas como pérdidas fijas [1]; También denominadas pérdidas en "vacío" y dependen de la variación de tensión en transformadores y máquinas eléctricas, debido a los "ciclos por histéresis[4]" y a las "corrientes parásitas de Foucault[5]"[7]; Las pérdidas que surgen en las imperfecciones del aislamiento eléctrico conocidas como "pérdidas por fuga de corriente" también forman parte de este tipo [1]; Las pérdidas por el "efecto corona[6]", es otro caso de este tipo de pérdidas y se producen en alta tensión. [1], [7]

1.1.3. Pérdidas técnicas de red

A lo largo del "sistema eléctrico[7]" están instalados componentes para la medición y el control, que son propensos de mal funcionamiento e ineficiencias que producen pérdidas eléctricas. [1], [8]

[3] Calentamiento que sufre el conductor por el paso de corriente eléctrica. Esto se da por el principio de conservación de la energía.
[4] Fenómeno eléctrico que generalmente se da en máquinas de electricidad (generadores, transformadores, etc.). Este fenómeno implica disipación de energía, puesta de manifiesto por el calentamiento que sufre el material.
[5] Fenómeno eléctrico que se produce cuando un conductor atraviesa un campo magnético variable o viceversa.
[6] Fenómeno eléctrico que consiste en la ionización del aire que rodea a los conductores de las líneas eléctricas produciendo una luz en forma de corona alrededor del conductor.
[7] Conjunto de instalaciones de centrales eléctricas generadoras, líneas de transmisión, subestaciones primarias y líneas de distribución, interconectadas entre sí, que permite generar, transportar y distribuir energía eléctrica.

Por lo general, las pérdidas técnicas variables constituyen el mayor porcentaje de pérdidas, en aproximadamente dos tercios y las pérdidas fijas en un tercio. [1]

1.2.PÉRDIDAS NO TÉCNICAS

Las pérdidas no técnicas se conocen como aquella energía que es consumida pero que no es facturada por parte de la empresa distribuidora. Son también conocidas como "pérdidas negras" o "pérdidas comerciales". [1], [3]

El valor de estas pérdidas, se obtiene de determinar la diferencia que se presenta entre las pérdidas totales del sistema y las pérdidas técnicas medidas y/o calculadas. [1], [8]

Este tipo de pérdidas caen en el ámbito social y comercial, están directamente relacionadas con errores y/o deficiencias administrativas de las empresas distribuidoras, o por acciones mal intencionadas por parte de los usuarios finales del suministro de electricidad. Este tipo de acciones, causan enormes pérdidas económicas, por esta razón, el control y mitigación de las mismas es de suma importancia para empresas de distribución y comercialización de energía eléctrica. [1], [3]

Las pérdidas no técnicas dependen de factores como: [1]

A. Problemas en los sistemas de medición;
B. Problemas de información de la red (Ubicación);
C. Problemas de procesamiento de datos de energía (Lectura).

1.2.1. Problemas en los sistemas de medición

En este caso, los factores de incidencia son amplios y se puede clasificar de la siguiente manera:

1.2.1.1. Hurto de energía

El hurto de energía se constituye por cualquier tipo de conexión ilegal por parte de los consumidores, ya sea conectando carga directamente a la red eléctrica de forma que el consumo no se registra en un medidor de energía o realizando una alteración y/o manipulación del equipo de medición. [1], [9]

El hurto de energía en muchos países es penalizado por la ley y puede ser castigado desde una multa hasta cumplir con una sanción de privación de libertad. [1], [8], [9]

En el Ecuador, según la LOSPEE[8] [10] el hurto y el fraude de energía se consideran como una infracción grave y el artículo 68 de infracciones graves dice "*son aquellas que afectan gravemente la provisión del servicio público y estratégico de energía eléctrica*". Las sanciones cuando se suscitan estos casos, según la regulación "*Modelo de contrato de suministro de energía eléctrica*" [11, p. 21] son:

[8] Ley Orgánica del Servicio Público de Energía Eléctrica

1. *Suspensión del servicio de energía eléctrica. (Art. 71 LOSPEE)*
2. *Pago por reparación o reposición de las instalaciones, equipos y materiales propiedad de la Distribuidora. (Art. 69 LOSPEE)*
3. *Cinco por ciento (5%) de un Salario Básico Unificado (1 SBU), su reincidencia equivaldrá a diez por ciento (10%) de un (1) SBU.*

El hurto de energía representa la mayor parte de pérdidas no técnicas y es complejo determinar la cantidad exacta que representa, debido a que no es fácil detectar y ubicar los sistemas de medición que adolecen de este defecto o intervención, incluso pudiendo presentarse casos en los cuales los mismos usuarios no permiten o dificultan las revisiones técnicas del sistema de medición, es por estos motivos que determinar de forma científica las pérdidas de energía constituye un gran desafío para las empresas distribuidoras y así conseguir mitigar este tipo de pérdidas. [1]

1.2.1.2. Errores de medición

Errores de medición hace referencia: [1]

- Equipos que presentan daños y/o defectos de funcionamiento que impiden una correcta medición de la energía consumida por parte del usuario.
- Errores en la toma de lectura y posterior procesamiento en el sistema de comercialización de la empresa.
- Inadecuada configuración del equipo de medición.
- Errores de instalación del sistema de medición.
- Fallas de medición debido a medidores fuera de vida útil y con falla debido a su obsolescencia o no haberse efectuado labores de mantenimiento.

1.2.2. Problemas de información de red

Estos problemas se originan cuando se registra incorrectamente el consumo de energía en la base de datos del sistema de comercialización de la empresa de distribución, generando pérdidas no técnicas.[1]

Las razones típicas para estos problemas son los siguientes: [1]

- **Sistemas de Medición no registrados:** Cuando un medidor no está registrado en el sistema de comercialización, por lo tanto, el consumo no es facturado. [1]
- **Ubicación incorrecta de los sistemas de medición:** Puede surgir cuando las coordenadas del equipo de medición son incorrectas o no se encuentran registradas en el sistema. [1]
- **Mala aplicación de tarifa:** Se presenta cuando se aplica una mala tarifa al usuario de la energía eléctrica. Por ejemplo, pueden existir casos en los que un equipo de medición tiene asignada una tarifa comercial, pero el uso de energía es residencial. [1]

1.2.3. Problemas de procesamiento de datos de energía

Estos fallos son debido a los errores en la estimación de energía generada o consumida, cuyas causas pueden ser: [1]

1.2.3.1. Estimación de consumos no medidos

Otro inconveniente en el control de pérdidas no técnicas, se debe a que por diversas circunstancias no todos los consumos de usuarios pertenecientes a una empresa de distribución son medidos. Existen medidores que su ubicación y/o emplazamiento dentro de inmuebles, impide la toma de lectura mensual, en cuyo caso es necesario realizar una estimación mensual del consumo de energía, la cual es realizada por el sistema de comercialización en función de parámetros de potencia, factores de coincidencia de carga y tiempos medio de uso. Esto puede ocasionar errores, debido a que cada consumo no leído podría no ser coincidente con la estimación realizada por el sistema. [1]

1.2.3.2. Error en la estimación de pérdidas técnicas

El cálculo de las pérdidas no técnicas, es el resultado de la diferencia entre las pérdidas totales y las pérdidas técnicas del sistema eléctrico, por lo tanto, cualquier imprecisión en el cálculo de las pérdidas técnicas provocará que sea contabilizado como pérdida no técnica. [1]

1.3.ESTADO ACTUAL DE PÉRDIDAS ELÉCTRICAS EN EL ECUADOR

Para la elaboración y análisis presentado a continuación, se obtuvieron datos de la página oficial del "ARCONEL[9]", el cual publica anualmente un módulo de estadísticas en [12]. Los datos se obtuvieron de módulos exhibidos desde el año 2007 hasta el año 2018.

1.3.1. Pérdidas eléctricas en Ecuador

En el plan maestro de electricidad 2016 – 2025, presentado por el "MEER[10]" se expone que el estado ecuatoriano ha tenido como objeto "*mejorar y fortalecer la gestión de las empresas eléctricas*". [13]

Este progreso se ve reflejado en las pérdidas de electricidad, en la Figura 1.1, se observa que el Ecuador presentó en el año 2007 un porcentaje de pérdidas de 21.42% y en el año 2018 un porcentaje de 11.39%, con una reducción de aproximadamente 10 puntos porcentuales para un período de 11 años.

[9] Agencia de Regulación y Control de Electricidad.
[10] Ministerio de Electricidad y Energía Renovable.

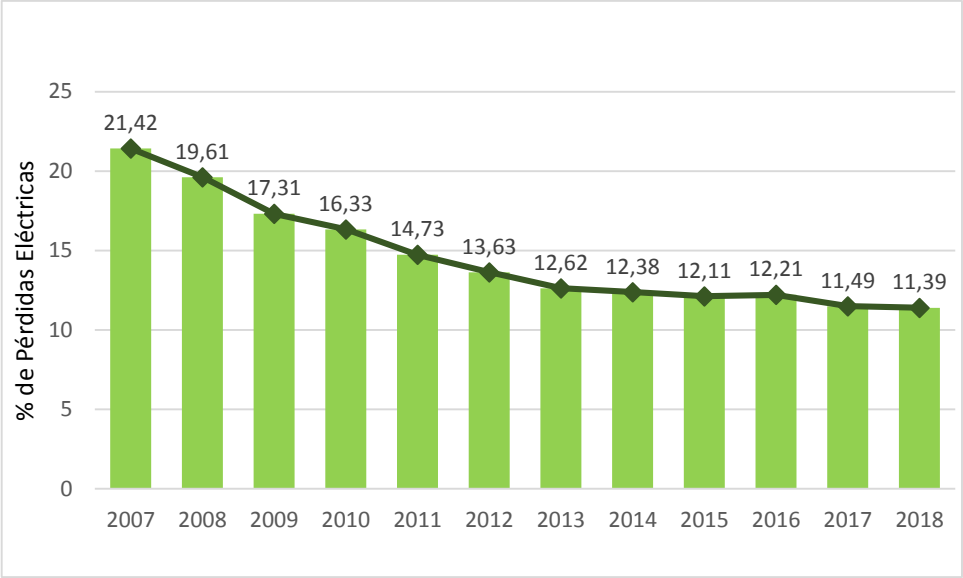


Figura 1.1, Porcentaje de pérdidas eléctricas, período 2007-2018
Datos Obtenidos de Módulos de estadística anual y multianual del sector eléctrico ecuatoriano desde el año 2007 hasta 2018 – ARCONEL [12]

Esta reducción representa un ahorro para el estado ecuatoriano de USD 1.200 millones; en comparación con los países de Latinoamérica, Ecuador logró ubicarse por debajo del promedio que es de 15.53%. Figura 1.2. [13]

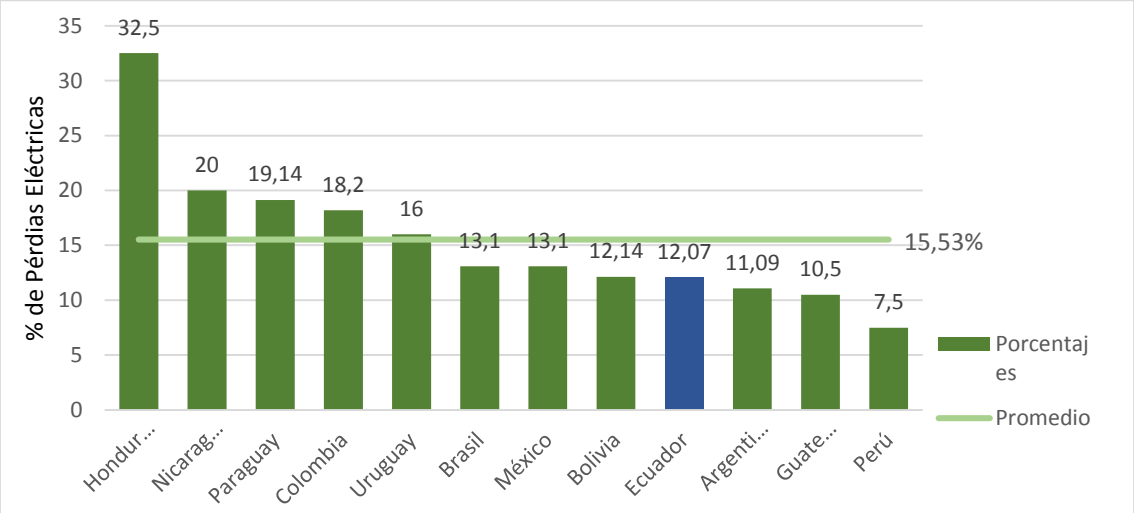


Figura 1.2, Pérdidas de electricidad en los países de la Región.
Fuente: [13, p. 62]

En la Figura 1.3, se puede ver el progreso que ha tenido el país en temas de reducción de pérdidas técnicas, durante el año 2007 tuvo un porcentaje de 9.26% con un aumento a 9.38% en 2009, sin embargo, posteriormente el porcentaje muestra una tendencia de reducción hasta llegar a un porcentaje de 7.20% de pérdidas en 2018.

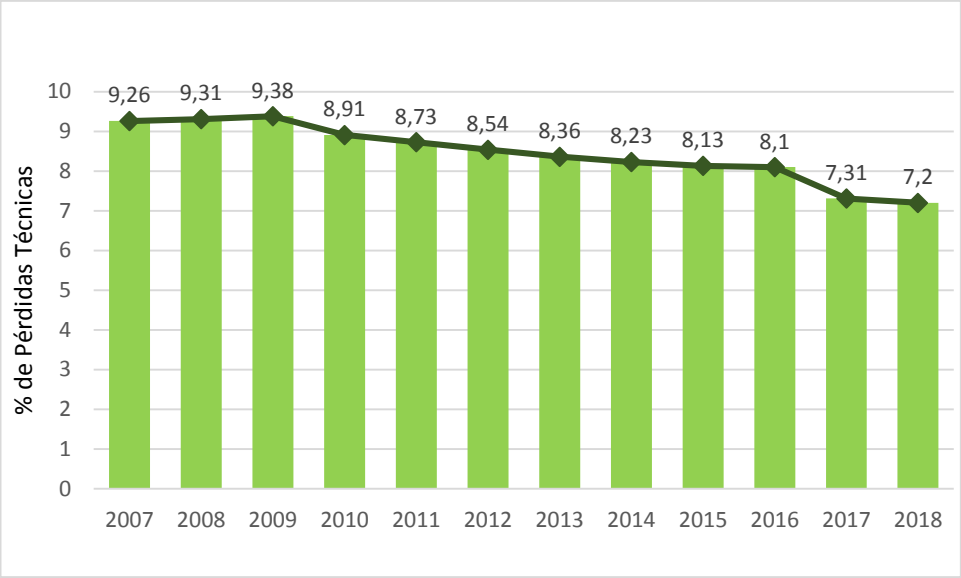


Figura 1.3, Porcentaje de pérdidas técnicas, período 2007-2018
Datos Obtenidos de Módulos de estadística anual y multianual del sector eléctrico ecuatoriano desde el año 2007 hasta 2018 – ARCONEL [12]

En la Figura 1.4, se presenta la disminución de pérdidas no técnicas, en el año 2007 el Ecuador mantuvo un porcentaje de pérdidas no técnicas de 12.16%, mostrando una tendencia importante hacia la baja llegando a 4.19% en el año 2018. Sin embargo, se observa un leve aumento de pérdidas no técnicas desde el año 2016.

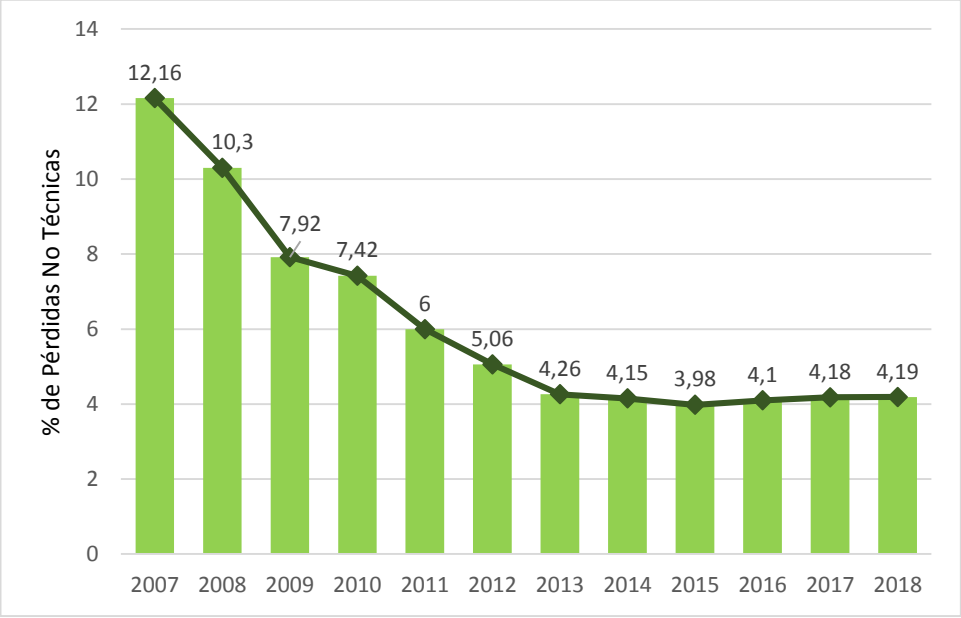


Figura 1.4, Porcentaje de pérdidas no técnicas, período 2007-2018
Datos Obtenidos de Módulos de estadística anual y multianual del sector eléctrico ecuatoriano desde el año 2007 hasta 2018 – ARCONEL [12]

1.3.2. Pérdidas de energía en empresas de distribución del Ecuador

En la Tabla 1.1, se muestra el porcentaje de pérdidas de energía de las empresas de distribución del Ecuador, en donde, se puede ver que el problema persiste. Las empresas CNEL-Manabí y CNEL-Esmeraldas presentan el mayor porcentaje de pérdidas (en comparación con las demás empresas) con 22.81% y 21.79% respectivamente.

A más de esto, se puede observar que 7 de las 21 empresas de distribución aumentaron el porcentaje de pérdidas totales desde el año 2017 al año 2018, entre ellas la E. E. Centro Sur que aumento 0.79% en 2018 en relación de 2017.

Tabla 1.1, Pérdidas de electricidad en las empresas de distribución
Datos Obtenidos de Módulos de estadística anual y multianual del sector eléctrico ecuatoriano desde el año 2007 hasta 2018 – ARCONEL [12]

	Pérdidas Totales		
Empresa	**Año 2017 [%]**	**Año 2018 [%]**	**Diferencia [%]**
CNEL-Guayaquil	10,34	11,1	-0,76
CNEL-Manabí	23,63	22,81	0,82
CNEL-Guayas Los Ríos	15,1	13,93	1,17
CNEL-El Oro	15,64	14,86	0,78
CNEL-Esmeraldas	22,6	21,79	0,81
CNEL-Milagro	15,77	15,15	0,62
CNEL-Sta. Elena	15,19	14,59	0,6
CNEL-Los Ríos	17,27	13,93	3,34
CNEL-Sto. Domingo	11,35	11,21	0,14
CNEL-Sucumbíos	12,44	8,21	4,23
CNEL Bolívar	7,91	7,71	0,2
E. E. Quito	5,41	5,72	-0,31
E. E. Centro Sur	6,25	7,04	-0,79
E. E. Norte	9,28	9,26	0,02
E. E. Cotopaxi	8,65	9,18	-0,53
E. E. Riobamba	10,25	8,53	1,72
E. E. Ambato	5,58	5,62	-0,04
E. E. Sur	10,19	8,72	1,47
E. E. Galápagos	7,96	8,63	-0,67
E. E. Azogues	4,56	5,3	-0,74

1.3.3. Pérdidas de electricidad en la Empresa Eléctrica Regional Centro Sur

La CENTROSUR a diciembre del año 2018, cuenta con 393.960 clientes [12], constituyendo una de las empresas de distribución con un índice estable de porcentajes de pérdidas de electricidad más bajos en el país.

En la Figura 1.5, se presenta las estadísticas de pérdidas que ha tenido la CENTROSUR, observándose que en estos últimos años la Empresa ha presentado pequeñas variaciones de

pérdidas; teniendo un pico máximo en el año 2014: 7.96%; y, un mínimo de 6.02% en el año 2009. Para el año 2018, el porcentaje es de 7.04%.

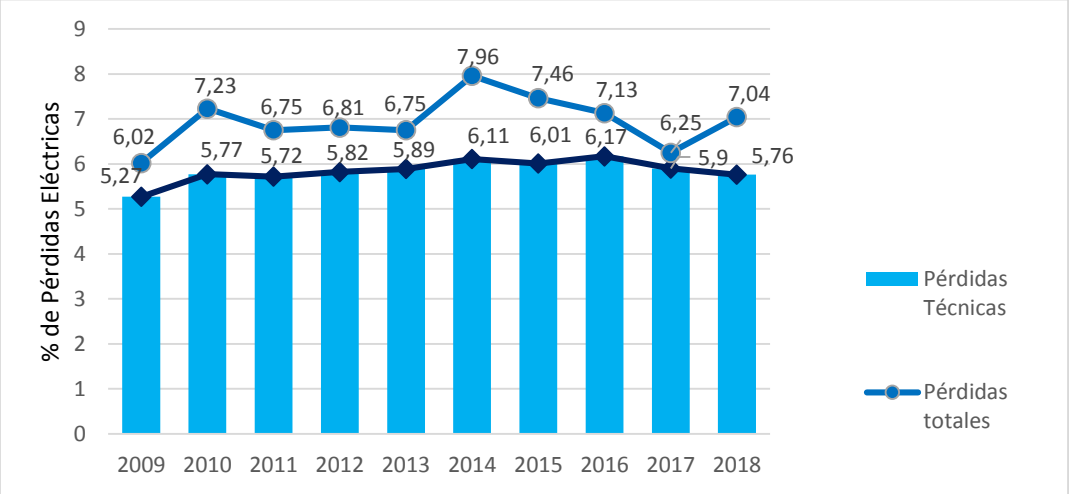


Figura 1.5, Porcentaje de pérdidas eléctricas Totales y Técnicas en la CENTROSUR, período 2009-2018
Datos Obtenidos de Módulos de estadística anual y multianual del sector eléctrico ecuatoriano desde el año 2007 hasta 2018 – ARCONEL [12]

Las pérdidas no técnicas mostraron variaciones apreciables como se observa en la Figura 1.6. Desde el año 2009 hasta el año 2013, las pérdidas no técnicas están dentro de los parámetros de la media; en tanto que en el año 2014 se presenta un aumento de esas pérdidas, llegando a un pico máximo de 1.85%. Incremento debido a que, en aquel año, el sistema de distribución La Troncal que pertenecía a CNEL-Milagro, pasa a integrar la E. E. CENTROSUR, con el consiguiente aumentando de la carga para la empresa y con ello un incremento de pérdidas tanto técnicas, cuanto no técnicas. Desde el 2014 hasta el año 2017, disminuye el porcentaje de pérdidas no técnicas hasta llegar a tener 0.35% y en el año 2018 aumenta a 1.27%; variación que se explica, por el cambio en los procesos administrativos de la E. E. CENTROSUR, referidos a la metodología en el cálculo para la determinación de la energía facturada.

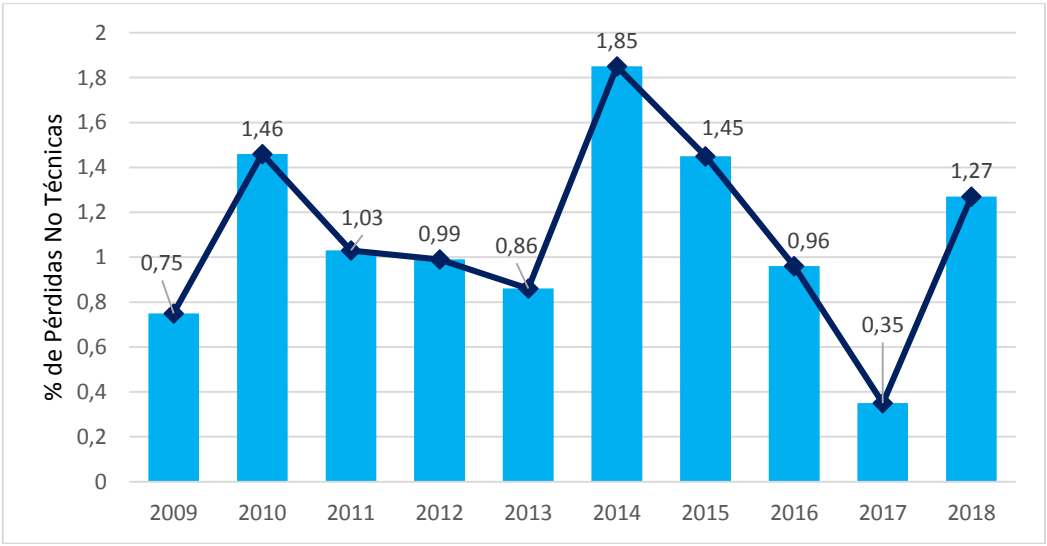


Figura 1.6, Porcentaje de pérdidas no técnicas en la CENTROSUR, período 2009-2018
Datos Obtenidos de Módulos de estadística anual y multianual del sector eléctrico ecuatoriano desde el año 2007 hasta 2018 – ARCONEL [12]

Téngase en cuenta que, si bien aparecen variaciones en los porcentajes de pérdidas no técnicas por los cambios mencionados y que esos porcentajes son bajos, la cantidad de energía que se pierde al año, es grande. Como se observa en la Figura 1.7, la cantidad de energía perdida desde el año 2014 hasta 2018 en promedio ronda los 14 GWh/año; específicamente en el año 2018 la E. E. CENTROSUR perdió 14,84 GWh.

Figura 1.7, Pérdida de Energía No Técnica en la CENTROSUR, período 2009-2018
Datos Obtenidos de Módulos de estadística anual y multianual del sector eléctrico ecuatoriano desde el año 2007 hasta 2018 – ARCONEL [12]

El menoscabo económico que representan estas pérdidas de energía es representativo. La Figura 1.8 muestra las pérdidas económicas que causan a la E. E. CENTROSUR, donde se aprecia que en el año 2018 ha perdido 1.38 millones de dólares. Razón por la que el control de las pérdidas no técnicas de energía, constituye una preocupación que requiere atención permanente; ni que decir de aquellas empresas distribuidoras de energía, que según se informó, presentan pérdidas mayores.

Figura 1.8, Pérdidas económicas por pérdidas no técnicas, período 2009-2018
Datos Obtenidos de Módulos de estadística anual y multianual del sector eléctrico ecuatoriano desde el año 2007 hasta 2018 – ARCONEL [12]

1.4. ESTADO DEL ARTE - METODOLOGÍAS UTILIZADAS PARA EL CONTROL DE PÉRDIDAS NO TÉCNICAS

Las pérdidas no técnicas son una preocupación constante para las empresas distribuidoras de energía, de ahí la necesidad de un método científico que permita su detección y mitigación, es un tema de interés actual.

A continuación, se presenta una descripción general de los métodos y técnicas que se han utilizado en las investigaciones para la detección de pérdidas no técnicas.

Generalmente las investigaciones sobre este tema, presentan la siguiente estructura: [2], [4]

- **Categoría:** Categoría de la solución a la que pertenece la investigación. [2], [4]
- **Algoritmo o tipo:** Algoritmo o tipo de solución propuesto. [2], [4]
- **Datos:** Información requerida para la solución propuesta. También se detalla el tamaño del conjunto de datos requeridos. [2], [4]
- **Criterios:** Los criterios utilizados para la detección de las pérdidas no técnicas. No existe criterios específicos, generalmente se plantean estos criterios en base a la experiencia, lo que se conoce como "criterio del experto". [2], [4]
- **Métricas:** Utilizadas para evaluar el rendimiento de los métodos de detección de pérdidas no técnicas. [4]

1.4.1. Categoría

Los sistemas de detección de fraude se clasifican de la siguiente manera: [2], [4], [14], [15]

- Estudio teórico;
- Métodos orientados a datos;
- Métodos orientados a red (Enfoque de medición; enfoque de sensores; procesamiento de señales);
- Híbridos.

1.4.1.1. Estudio teórico

Estos métodos analizan los aspectos sociales relacionados con el hurto de energía, es decir, determinan cuáles son los factores sociales que ayudan a revelar una pérdida no técnica. [2]

En estos estudios se utilizan técnicas estadísticas para encontrar aspectos que relacionen el historial de hurtos con variables socioeconómicos y sociodemográficos. Tiene la ventaja de presentar listas potenciales para revisión, pero estos métodos no presentan casos específicos de hurto o fallas en la medición. [2]

1.4.1.2. Métodos orientados a datos

Los métodos orientados a datos se enfocan en el análisis de datos referentes con el consumidor (consumo de energía, datos demográficos, características del consumidor, etc.), mediante técnicas de minería de datos para la localización de pérdidas no técnicas. [2], [4], [14], [15]

Estas técnicas utilizan algoritmos para la clasificación del comportamiento de los consumidores, estos algoritmos se pueden categorizar en: supervisados y no supervisados. [2], [4], [14], [15]

- **Aprendizaje supervisado:** Se da cuando el algoritmo es desarrollado para aprender mediante ejemplos, proporcionándole datos de entradas y salidas deseadas. [16] Es decir, cuando el método utiliza ambas etiquetas (positivo/fraude y negativo/no fraude). [4], [14] Para esto la empresa distribuidora deberá tener información de calidad, debido a la integración de una gran cantidad de datos verificados de fraude y no fraude. [14]
- **Aprendizaje no supervisado:** El algoritmo analiza los datos para identificar patrones y trata de obtener experiencia en base a los datos disponibles. [16] Estos métodos se utilizan cuando se tienen datos no etiquetados o cuando una de las muestras (ej. fraude) no tiene la cantidad suficiente de muestras. [4], [15]

1.4.1.3. Métodos orientados a red

Los métodos orientados a red se sustentan en la adquisición de datos mediante el manejo de hardware en la red eléctrica que permita la identificación o estimación de pérdidas no técnicas. [2]

Los métodos orientados a la red abarcan las siguientes técnicas:

- **Enfoque de medición:** Es un método excelente para la detección de pérdidas no técnicas en un área y se trata de la instalación de medidores en el lado de bajo voltaje del transformador de distribución (llamado medidor de observación). La medición obtenida del medidor observador, se puede comparar con la suma de mediciones de los medidores convencionales y teniendo en cuenta el cálculo de pérdidas técnicas, se puede verificar el porcentaje de diferencia entre la energía producida y consumida. Si el porcentaje es grande, mayor es la probabilidad de pérdida no técnica. [4], [14]

  Esta técnica tiene la ventaja que detectar rápidamente pérdidas no técnicas. La desventaja es que el método indica el área en la que está instalado el medidor y no a los consumidores específicos, además el costo de instalación y equipos para el control es alto. [2]
- **Enfoque de sensores:** Comprende la instalación de sensores. En esta metodología se encarga del cálculo de la cantidad y posicionamiento de los sensores para que ayuden con la detección de pérdidas no técnicas. [4]
- **Procesamiento de señales:** Es un método poco estudiado y propuesto por la literatura. Se trata de la generación de señales armónicas. Después de la desconexión de medidores legales, se introduce señales que perturba a equipos conectados a la red, dañando equipos de aquellos consumidores ilegales. [2]

1.4.1.4. Métodos híbridos

Es una combinación de métodos orientados a datos y orientados a red para poder revelar pérdidas no técnicas con mayor precisión. [4]

Por ejemplo, J. Pulz *et al.* [17], propone la instalación de medidores observadores (método orientado a redes). Estos medidores como ya se explicó, permiten estimar las pérdidas no técnicas. En un área sin perdidas no técnicas, la energía consumida debe ser igual a la suma de las pérdidas técnicas y la energía cargada en los clientes. En caso de no coincidir, indicará que cierta área presenta pérdidas no técnicas. Para dichas áreas que presentan pérdidas no técnicas se aplica un modelo de clasificación (método orientado a datos) que ayuda a detectar clientes con alta probabilidad de fraude.

Con esta combinación de métodos lo que se intenta es aumentar la eficiencia en la búsqueda de consumidores maliciosos. La desventaja es el costo por la implementación de dichos equipos.

1.4.2. Algoritmos

Los algoritmos que se presentan a continuación, hacen referencia a métodos orientados a datos y son los encargados de deducir indicadores binarios o probabilidad de pérdida no técnica. [2]

Habitualmente consiste en las siguientes etapas:

1) Procesamiento de datos;
2) Ajuste de modelo de clasificación de datos;
3) Valoración de desempeño del modelo;
4) Marcha del modelo. [2]

Los algoritmos son:

1.4.2.1. Métodos supervisados

- **Máquina de Soporte Vectorial (SVM):** Este algoritmo realiza una clasificación binaria que separa los datos en dos elementos. [16] Es una técnica confiable para detectar pérdidas no técnicas, pero tiende a ser de difícil desarrollo y lento para obtener resultados; razón por la cual, este algoritmo se combina con otros clasificadores para obtener mejores resultados. [4] Este es el algoritmo más común de los métodos supervisados. [14]
- **Red Neuronal Artificial (ANN):** Está inspirada en la neurona del cerebro humano, trata de capas enlazadas que van tomando forma de una neurona y relacionan datos de entrada con datos de salida. [16] Una red neurona artificial, tiene la capacidad de aprender de datos, buscar patrones, clasificar datos y predecir futuros eventos. El algoritmo para el control de pérdidas no técnicas, obtiene datos de entrada (generalmente datos del consumidor), para posteriormente realizar una clasificación binaria para pronosticar posibles clientes que presenten dichas pérdidas. [4]
- **Árboles de decisión:** Es un diagrama o flujo de proceso que muestra los resultados probables de una serie de decisiones conectadas. [16] El árbol de decisión es un clasificador, que generalmente inicia con un único nodo, para posteriormente en base a una serie de reglas (definidas por expertos) se va "ramificando" en posibles resultados

que son usados para la detección de pérdidas no técnicas. Es un algoritmo poco usado para esta aplicación. [2], [4]

- **Bosque de la ruta óptima – Optimun Path Forest (OPF):** Es un clasificador fundamentado en grafos que utiliza la salida de varios árboles de decisión preparados en diferentes grupos de datos de prueba para determinar respuestas probables. [16]

  A diferencia de los anteriores métodos que buscan una clasificación óptima que separa dos clases, OPF divide a manera de grafo dos o más árboles (árboles de decisión óptima) cada uno representado por una clase. Este método es capaz de operar clases superpuestas y no requiere de mucho tiempo de entrenamiento, permitiendo el aprendizaje en línea del sistema en detección de pérdidas de energía. [4]

  J. L. Viegas *et al.* [2] indica que es un método poco común en la literatura y es mejor que SVM y ANN.

- **Regla de inducción:** Es un método que utiliza reglas de "SI-ENTONCES-DE OTRO MODO" (IF-THEN-ELSE) que permiten distinguir a consumidores maliciosos. Estas reglas se pueden definir a través del estudio estadístico y el conocimiento del experto.

  Muchas de las veces estas reglas no son suficientes para lograr un buen reconocimiento de malos clientes, sin embargo, este método se puede combinar con otros clasificadores para lograr un mejor rendimiento. [4]

- **El vecino más cercano – Nearest Neighbor (k-NN):** Es un método que clasifica elementos (objetos, personas, etc.), basándose en su similitud a otros casos. [16]

  Se ingresa al campo de las características una muestra y se determina la clase más común entre sus vecinos más cercanos por "voto mayoritario". Necesitan de un único parámetro denominado "*k*", que determina el número de vecinos más próximos a examinarse. [4]

  G. M. Messinis y N. D. Hatziargyriou [4], indican que este método es el más sencillo y simple para detectar pérdidas no técnicas y generalmente es utilizado para comparar con otros algoritmos.

- **Clasificadores Bayesianos:** Existen dos clasificadores bayesianos: Nayve Bayes y Red bayesiana.

  Nayve Bayes es un clasificador probabilístico que necesita de datos estadísticos que se pueden obtener de registros de la empresa distribuidora. Se limita por la necesidad de conocimiento previo de probabilidad y por tener que buscar datos complejos que normalmente se utilizan para detectar pérdidas no técnicas, razón por la cual, se necesita tener una base sólida de datos para que el algoritmo sea eficiente. [2], [4]

  Red bayesiana es un clasificador que presenta de manera gráfica la probabilidad conjunta de un grupo de variables. Es fácil de interpretar, permitiendo al usuario entender que características están más afectadas por las pérdidas no técnicas. [4]

1.4.2.2. Métodos no supervisados

- **Mapa de auto organización (Self Organizing Map - SOM):** Es un tipo especial de red neuronal. [4], [15], [16] SOM no es un método clasificador, sin embargo, es utilizado como parte de uno. [15] Genera una representación visual 2D, que es de fácil comprensión para los usuarios y el resultado final es un clúster que debe ser evaluado por el conocimiento experto o complementar con un algoritmo de agrupación simple, para decidir sobre el grupo de medidores que deben ser inspeccionados. [4], [15]
- **Algoritmos de agrupamiento:** Son métodos utilizados para agrupar consumidores según la proximidad o atributos similares. El principal problema es la cantidad de grupos que se deben formar y cómo interpretarlos. Básicamente se debería definir dos grupos: fraude y no fraude; en el que, fraude será el grupo más pequeño y homogéneo. El problema es que puede existir varios tipos de fraude, razón por la cual, la cantidad de grupos a formarse se debe elegir minuciosamente y debe ser valorado por el experto. [4], [15], [16]
    - **K-medias:** Es un algoritmo de agrupamiento, que tiene como entradas la base de datos (grupo de variables) y un número entero que indica la cantidad de grupos a descubrirse. Este es un método que agrupa clientes en función de una proximidad de similitud de conjunto definido por puntos denominados "*K*" o "centroide". La ventaja de este método es la eficiencia al momento de manejar grandes conjuntos de datos, pero con la desventaja que es necesario saber la cantidad de grupos que deben formarse. Otra desventaja de este algoritmo es la sensibilidad al ruido, pues al calcular los grupos refiriéndose a un centro, cualquier dato atípico podría alterar este centroide, por lo tanto, una mala formación de los grupos puede perjudicar la respuesta. [4], [5], [16]
    - **K-menoides:** Igual que el algoritmo anterior, los datos de entrada para este algoritmo son la base de datos y el número de la cantidad de grupos a buscar. Este algoritmo escoge como centro el objeto con menor diferencia media, denominado "miedo". La diferencia con el anterior método es el centro de agrupamiento, es decir, K-menoides son los propios objetos de la base de datos, mientras k-medias se calcula a partir de componentes de la partición.

      La ventaja de este método es que no es sensible a ruido y la desventaja es en el procesamiento de datos, ya que puede llegar a ser costoso. [5]
- **Sistema experto:** Es un sistema que se fundamenta en reglas definidas por el experto en materia, estos pueden ser el personal técnico que se dedica a rastrear pérdidas no técnicas, las reglas pueden ser definidas por varios medios, pero generalmente se describen modelos de cómo se expresa una pérdida no técnica. El problema con este método es que las reglas no pueden ser sencillas de diseñar. [4], [15]

- **Factor local atípico (Local Outlier Factor – LOF):** Este método obtiene la densidad local de un punto de datos y la compara con la densidad local de sus vecinos. Este parámetro se calcula para todos los clientes, para luego clasificarlos. Un gran número de clientes aparecerán como fraude, pero deben ser analizados solamente aquellos consumidores que tengan un alto porcentaje de este parámetro. Como es obvio este método tiene la desventaja de presentar listas de clientes que no presentan fraudes. [15]
- **Control estadístico:** Se trata de un control por medio de gráficos de series de tiempo, mediante el cual se monitorea un consumo individual, pudiendo detectar posibles fraudes cuando se visualiza una inconsistencia en una serie de tiempo. El problema es que el propósito general de este método es detectar cambios y, no cualquier cambio hace referencia a fraude, ya que puede ser por otro tipo de cambio de consumo, produciendo falsos positivos. [4]

1.4.3. Métricas

Las métricas sirven para evaluar el rendimiento de las metodologías para detección de pérdidas no técnicas. [4]

Toda técnica produce una lista de posibles clientes maliciosos y los parámetros principales que usa una métrica para calificar a la técnica son:

- Verdadero positivo (TP): Cuando un consumidor comete fraude y la técnica lo clasifica como tal;
- Verdaderos negativos (TN): Los casos correctamente citados como no fraude;
- Falso positivo (FP): Cuando un consumidor no comete fraude y la técnica lo clasifica como fraude;
- Falsos negativos (FN): Cuando un consumidor comete fraude y la técnica lo clasifica como no fraude. [4], [14]

El objeto de toda metodología de detección de pérdidas no técnicas es aumentar los verdaderos positivos y reducir los falsos positivos, evitando así, inspecciones que son costosas y encontrando todos los casos de fraude como sea posible. Las métricas más utilizadas son precisión, razón de falsos positivos (FPR), verdadera razón positiva (TPR) y tasa de detección o exactitud. [14] Estas se describen a continuación:

- **Tasa de detección o exactitud:** Revela si una técnica clasifica correctamente las muestras positivas y negativas. No obstante, cuando en la muestra existe un conjunto de datos desnivelado, esto es, cuando existe más ejemplos positivos que negativos o viceversa, esta métrica no es suficiente para medir el desempeño de la técnica de clasificación. [4]

$$\text{Exactitud} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \quad (1)$$

- **Razón de verdaderos positivos:** Indica si una técnica de clasificación se desempeña correctamente, enunciando la proporción de muestras catalogadas como pérdida no técnica correspondiente al número total de pérdidas no técnicas dentro de un grupo de datos. Generalmente esta métrica se considera conjuntamente con precisión para determinar la eficacia de una técnica. [4]

$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (2)$$

- **Precisión o confianza:** Muestra el nivel de confianza que tiene una técnica de clasificación, es decir, que tan bien clasifica los datos. Un valor alto de precisión indica que los ejemplos que son calificados como positivos (pérdida no técnica), son positivos reales. [4]

$$\text{Precisión} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (3)$$

- **Razón de falsos positivos:** Señala la relación que existente entre las falsas alarmas (muestras que son señaladas falsamente como fraude) con el número total de negativos. Generalmente, se busca que esta métrica tenga valores bajos para evitar inspecciones innecesarias. [4]

$$\text{FPR} = \frac{\text{FP}}{\text{TN} + \text{FP}} \quad (4)$$

G. M. Messinis y N. D. Hatziargyriou en [4] presentan una tabla, en la que se puede encontrar estas y más métricas que son utilizadas en la literatura.

1.5.METODOLOGÍA UTILIZADA POR LA EMPRESA CENTROSUR PARA EL CONTROL DE PÉRDIDAS NO TÉCNICAS

El Departamento de Control de la Medición de la E. E. CENTROSUR cuenta con tres grupos de revisiones en campo; dos de ellos se dedican a revisar equipos de medición “masivos convencionales” y el tercero a revisar equipos de medición “especiales”; la revisión de equipos de medición “masivos convencionales” comprende a aquellos sistemas de medición generalmente conectados a nivel de baja tensión que no requieren de la conexión de transformadores de medida entre el punto de suministro y el punto de medición, sistemas que muy comúnmente corresponden a clientes de tipo residencial, o comerciales e industriales a nivel de baja tensión.

En el caso de las revisiones de sistemas de medición “especial”, estas comprenden sistemas de medición cuya conexión requiere el uso de transformadores de medida (TP’s y/o TC’s), los cuales

generalmente se utilizan en usuarios de categorías comerciales y/o industrial conectados a nivel de baja, media o alta tensión, estas revisiones requieren de software y equipos de calibración especial en la mayoría de los casos.

Según datos estadísticos del Departamento de Control de la Medición, en 2018 los tres grupos de revisión en campo, completaron un total de 5.476 revisiones, es decir en promedio 456 revisiones por mes, recuperando un total 191.312 kWh de energía, y recuperando un monto total de ingreso de 34.661,27 USD.

Teniendo en cuenta estos antecedentes, el Departamento de Control de la Medición utiliza el conocimiento de expertos para determinar los listados de revisiones para el control de pérdidas no técnicas, criterios o reglas que son definidas por el personal de planificación del Departamento en base a la experticia que sobre el tema se tiene. Según el personal del Departamento el proceso de generación de las listas de revisión toma entre una hora y media a dos horas en cada mes.

Como se revisó, este método ha contribuido con el control y mitigación de las pérdidas no técnicas, sin embargo, es factible mejorar el proceso y sobretodo establecer una metodología que sirva de base para la planificación de las actividades en el futuro, razón por la cual es de importancia y necesario el planteamiento de una metodología que permita mejorar la eficiencia de las actividades de control de pérdidas que ejecuta el Departamento de Control de la Medición.

2. CAPÍTULO 2 – MINERÍA DE DATOS

En este capítulo se realiza el análisis, consolidación y procesamiento de la información técnico – económica del proceso comercial de la Empresa "EERCS[11]". Primero se realizará una recopilación y pre-procesamiento de los datos, para posteriormente aplicar técnicas de minería de datos con la finalidad de agrupar y clasificar la información relevante; esto permitirá obtener como resultado listados de sistemas de medición que presenten características que los hagan candidatos de revisión en sitio a fin de verificar su correcto funcionamiento.

La investigación efectuada en el presente trabajo, da inicio con el análisis de diferentes técnicas de minería de datos, efectuar la determinación de la técnica más adecuada para su aplicación a los datos del proceso comercial de la distribución de energía eléctrica, todo esto con el objetivo final de estructurar una metodología adecuada que mejore el proceso del control de las pérdidas no técnicas en los sistemas de distribución de energía.

El capítulo se divide en dos partes: en la sección 2.1 se define y explica el proceso de "minería de datos", en la sección 2.2 se expone la minería de datos aplicada a los datos de la empresa distribuidora.

2.1.MINERÍA DE DATOS

Existen varias definiciones de minería de datos y no es fácil dar un concepto exacto, por este motivo en este caso se puede decir que, minería de datos es un proceso orientado hacia obtener información o conocimiento útil a partir de grandes cantidades de datos sin pérdida de información. [18]–[20]

En la actualidad, las empresas disponen de voluminosas bases de datos, que pueden ser procesadas mediante técnicas informáticas especializadas que abarca el nombre de minería de datos, es decir, que mediante estas técnicas se busca adquirir conocimiento de manera automática. [19]

Previo a explicar el proceso de minería de datos, es necesario entender los siguientes conceptos:

2.1.1. Conceptos Generales de Datos Multivariantes

2.1.1.1. Variables o atributos

Las variables o atributos son los distintivos o características que un individuo (de una población), puede tener. Por ejemplo, el cliente "A" que forma parte de una empresa distribuidora de energía puede tener como atributos el consumo de energía mensual, dirección, número telefónico, sexo, etc. [21]

Las variables pueden ser cuantitativas o cualitativas:

[11] Empresa Eléctrica Regional Centro Sur C. A.

- Cuantitativas: Cuando la variable se puede expresar numéricamente (energía consumida, demanda eléctrica, etc.). [21]
- Cualitativas: Cuando la variable es un dato particular y específico de un individuo (sexo, tarifa, etc.). [21]

Para el análisis, las variables cualitativas se pueden codificar numéricamente. Por ejemplo, si un cliente es de la tercera edad, se pondrá con "1", de no ser así "0".

No en todos los casos las variables cualitativas son binarias (0 o 1); si se tiene más variables, se puede codificar, pero se necesitará de un tratamiento especial. [21]

2.1.1.2. Matriz de datos

La matriz de datos, es un vector que representa la información de n cantidad de individuos de cierta población y cada uno de estos está descritos por p cantidad de variables. Entonces, la matriz de datos, será de $[nxp]$ dimensiones. [21], [22]

La matriz se puede representar de la siguiente manera:

$$\boldsymbol{Matriz\ Base} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{12} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} = \begin{bmatrix} x'_1 \\ x'_2 \\ \vdots \\ x'_n \end{bmatrix} = \left[x_{(1)} x_{(2)} \cdots x_{(p)}\right] \qquad (5)$$

Donde:

- x_{ij} es el valor de una variable escalar j en el individuo i; [21]
- x'_i es un vector fila de $px1$, que representa los valores de las p variables univariantes sobre el individuo i; [21]
- $x_{(j)}$ es un vector columna de $nx1$, que simboliza la variable escalar x_j de las n observaciones. [21]

2.1.2. Proceso de minería de datos

A continuación, se explica cuál es el proceso para realizar una minería de datos.

1. Requerimiento de análisis

Este es el paso inicial del proceso de minería de datos; aquí es en donde los investigadores de minería de datos se trazan objetivos en base al requerimiento del usuario final. Para esto, los objetivos planteados tienen que ser claros para que la investigación sea efectiva. [18], [20], [23]

2. Recopilación e integración de datos

Con los objetivos planteados, el siguiente paso es identificar y estudiar la información que se tiene a disposición. Este paso incluye realizar una recopilación e integración de datos para la adquisición de conocimiento. [18], [19]

Esta fase es muy importante porque las técnicas de minería de datos se entrenan y revelan nuevo conocimiento a partir de los datos que se tiene a disposición, entonces es necesario analizar y encontrar aquellos datos que aportarán con información importante acorde al requerimiento u objetivos planteados en el paso 1. [19], [20]

Además, cuando se realiza una recopilación de información de varias fuentes, es normal que estos se contengan en diferentes formatos, por lo que, al integrar esta información se tiene que conformar en un formato común y se debe eliminar cualquier redundancia e inconsistencia encontrada. [24]

Una vez cumplida con esta fase, se debe tener una matriz de datos parecida a la matriz base mostrada en la ecuación (5).

3. Pre-procesamiento de datos

El propósito del pre-procesamiento de datos es corregir cualquier inconsistencia encontrada en la calidad de los datos de tal manera que el procesamiento (aplicación de técnicas de minería de datos) sea el más óptimo. Para cumplir con esto se tiene que realizar un reconocimiento, limpieza y transformación de los datos. [20], [24] A continuación se detalla cada uno de estos procesos:

A. Reconocimiento de los datos

El reconocimiento de los datos se realiza para analizar y conocer de mejor manera las variables o atributos. Para cumplir con este fin se puede utilizar estadística descriptiva, técnicas de análisis exploratorio de los datos y la matriz de correlaciones.

- **Estadística descriptiva:** Se aplica el cálculo de índices de estadística descriptiva conformando una tabla que muestra los índices estadísticos de las diferentes variables. [18] Por ejemplo, la tabla puede contener lo siguiente:
    - **Datos Nulos:** Muestra la cantidad de datos en blanco encontrados por variable.
    - **Valor de la Media:** Presentará la media o el promedio de los datos.
    - **Valor de la Mediana:** Muestra la mediana de los datos.
    - **Valor de la Moda:** Muestra el valor que más frecuencia tiene de aparición.
    - **Valor Máximo:** Valor numérico más alto de los datos.
    - **Valor Mínimo:** Valor numérico más bajo de los datos.
    - **Desviación estándar:** Valor que indica la dispersión existente de los datos con respecto a la media.
    - **Coeficiente de variación:** Valor que indica la dispersión porcentual de los datos.

- **Análisis exploratorio de los datos:** Este paso es importante porque ofrece información de manera gráfica de las variables o atributos, permitiendo identificar ciertos patrones, como la dispersión, simetría, datos atípicos, etc. [18], [21], [24]
  Para lograr este objetivo, se pueden realizar algunos diagramas como:
  - Dispersión de datos: Es un diagrama que muestra la dispersión existente entre los datos y la relación existente entre las variables. [21]
  - Diagrama de caja o de bigotes: Es un gráfico que representa una caja rectangular, en donde en los extremos (bigotes) estarán los valores máximos y mínimos encontrados en la variable, y la caja estará formada por los rangos intercuantiles. Este gráfico es muy útil para identificar datos atípicos. [21]
  - Histogramas: Gráfico que presenta la distribución en frecuencias de los datos. [21]
- **Matriz de correlaciones:** Es una herramienta estadística muy útil para identificar correlación entre las variables. Es una matriz lineal, que está formada por 1 en la diagonal principal y fuera de esta los coeficientes de correlación. Aquellos coeficientes cercanos a 1, indican una estrecha correlación entre las variables. [21]

B. Limpieza de los datos:

La mayoría de las técnicas de aprendizaje artificial, requieren que la data se encuentre "limpia", es decir, que no existan datos duplicados, datos "en blanco" o inexistentes y datos erróneos, esto con el propósito de que el proceso de "aprendizaje" del algoritmo y la obtención de información sea óptima y adecuada. [18]

- **Datos inexistentes**
  Son datos que no están registrados en el sistema por razones desconocidas. [18]
  Estos datos se pueden tratar de la siguiente manera:
  - **Ignorar:** Se deja pasar, ya que hay algoritmos que no se ven afectados por estos datos. [18][19]
  - **Aproximación de datos:** Se obtiene un promedio de datos y se reemplaza el valor nulo por un valor aproximado o mediante una técnica de regresión lineal, se calcula un valor por el que se pueden reemplazar. El problema de este tratamiento, es que puede alterar información real, pues si bien el dato reemplazado es un aproximado a los demás datos, el valor calculado no es real. [19]
  - **Suprimir la fila o la columna:** Se elimina la fila (muestra), en donde se encuentran datos nulos, o a su vez si se encuentra columnas (variables) que contienen muchas casillas con datos nulos, se procede a eliminar la variable. [18][19]

- **Reemplazar valor:** En algunos casos, se elige colocar un cero en donde se encuentren estos datos. Hay que tener cuidado si se elige esta opción de tratamiento, debido a que, al hacer esto la media de los datos se puede alterar. [18]

- **Datos erróneos**

  Se conoce como dato erróneo o también denominado "atípico" a aquella observación distinta o diferente al resto de datos o que no cumple con un patrón parecido. [18], [19], [21]

  Es importante identificar estas observaciones, porque pueden llegar a distorsionar la media de los datos, la desviación estándar y/o la correlación existente entre variables. [21]

  C. Pérez, en [19], categoriza a los datos atípicos de la siguiente manera:

  - **Error por procedimiento:** Son valores atípicos que se produjeron por errores de registro, por ejemplo, una mala codificación, error al ingresar el dato al sistema, etc. Es importante detectar estos datos a fin de que sean eliminados. [19]
  - **Evento extraordinario (Uno):** Observaciones distintas a las demás, pero que tienen una explicación de manifestación. Este tipo de casos se mantienen en la muestra, salvo que el investigador considere eliminarlos. [19]
  - **Evento extraordinario (Dos):** Observaciones distintas a las demás, pero que no tienen una explicación de existencia. Generalmente estos datos son eliminados. [19]
  - **Observaciones fuera de rango:** Son valores extremos dentro de la muestra de una variable y generalmente se considera eliminarlos si no representan importancia para la población. [19]

C. Transformación o normalización de los datos:

Por "Normalizar los datos" se entiende centrar o escalar los datos, de tal manera, que los mismos se encuentren en un mismo rango de valores. [18]

Existen técnicas de minería de datos que son robustas para grandes diferencias en rangos de valores, como el árbol de decisión o las reglas de inducción. Sin embargo, existen otras técnicas en las que es preciso realizar una normalización. [18]

Se han propuesto varias maneras de normalizar los datos. Las normalizaciones más comunes y las que se aplicarán posteriormente para el análisis son:

- **Normalización máximos – mínimos**

  Es una normalización comúnmente utilizada en la literatura. Normaliza los datos en valores entre 0 y 1. [18] Esta normalización utiliza la siguiente ecuación:

$$v' = \frac{v - min}{max - min} \quad (6)$$

  Donde:

  - v': Es el nuevo valor.
  - v : Es el valor a normalizar.
  - min: Es el valor mínimo de los datos.
  - max: Es el valor máximo de los datos.

- **Normalización Z-Score**

  Otra es aquella denominada "z-score" y centraliza los datos "suavizando" grandes valores. Generalmente esta técnica de normalización se utiliza cuando se desconoce los valores máximos y mínimos y hay existencia de ruidos. [5]
  La normalización z-score se realiza con la siguiente ecuación:

$$v' = \frac{v - mean}{std} \quad (7)$$

  Donde:

  - v': Es el nuevo valor.
  - v : Es el valor a normalizar.
  - $mean$: Es el promedio de los datos.
  - std: Es la desviación estándar de los datos.

4. Procesamiento de datos

El procesamiento de datos es la aplicación de técnicas informáticas de minería de datos; para esto la información debe estar en "óptimas condiciones" una vez que ha pasado por revisión previa, es decir, tener una base de datos sólida con los datos "limpios", sin datos inexistentes, sin datos atípicos y los datos deben estar transformados. [18]

Con el conocimiento de la información disponible y ya con la base de datos sólida, se debe elegir una técnica de minería de datos que más se adecue a dicha base.

A continuación, se profundiza las técnicas (revisadas en el capítulo 1) que se utilizan posteriormente para el análisis de minería de datos.

A. Métodos no supervisados

- **Técnicas estadísticas: Coeficiente de Pearson**

  El Coeficiente de Pearson es una técnica estadística que se utiliza para identificar la relación existente entre dos variables. [25]

  El coeficiente de correlación de Pearson, se calcula mediante la siguiente ecuación:

$$-1 \leq \frac{Cov(X,Y)}{S_X S_Y} = \frac{\sum_{t=1}^{n}(X_t - \bar{X}) * (Y_t - \bar{Y})}{\sqrt{\sum_{t=1}^{n}(X_t - \bar{X})^2} * \sqrt{\sum_{t=1}^{n}(Y_t - \bar{Y})^2}} \leq 1 \qquad (8)$$

  Donde:

  - $Cov(X,Y)$ es la covarianza entre X y Y.
  - S_X es la desviación estándar de X.
  - S_Y es la desviación estándar Y.

  En la ecuación (8), el coeficiente de correlación de Pearson solamente toma valores entre -1 y 1. El valor de 1, significa que todos los puntos están en una línea y que existe correlación positiva entre las variables, es decir que mientras crece X aumenta Y. Caso contrario, un valor de -1, indica que todos los puntos están en la misma línea, pero, tiene correlación negativa, mientras aumenta X disminuye Y. En caso de ser 0, indica que las variables no están correlacionadas y existe dispersión de datos. [25]

- **Agrupamiento K-Medias**

  Como se manifestó en el capítulo 1, el agrupamiento K-Medias es un algoritmo no supervisado, que necesita como parámetros de ingreso: los datos a agrupar y la cantidad de grupos a concentrar denominado "K".

  El algoritmo K-Medias es el siguiente: [5], [26]

Algoritmo K-Medias
1. Ingresar aleatoriamente un valor de K. 2. Formar K agrupaciones, estableciendo cada dato al centroide más cercano. 3. Reajustar los centroides K, que será el promedio del grupo establecido en el paso 2. 4. Repetir pasos 2 y 3 hasta que no exista reajuste de centroides.

  Para comprender este algoritmo, se notará en la Figura 2.1, que están datos generados al azar y en la Figura 2.2, la asignación de cada dato a un grupo. Como se observa, el valor de K es dos, es decir los datos están formados en dos grupos.

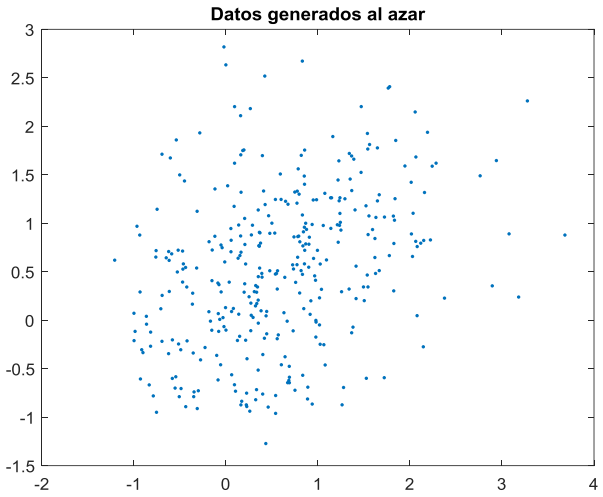


Figura 2.1, Datos generados al azar para agrupaciones

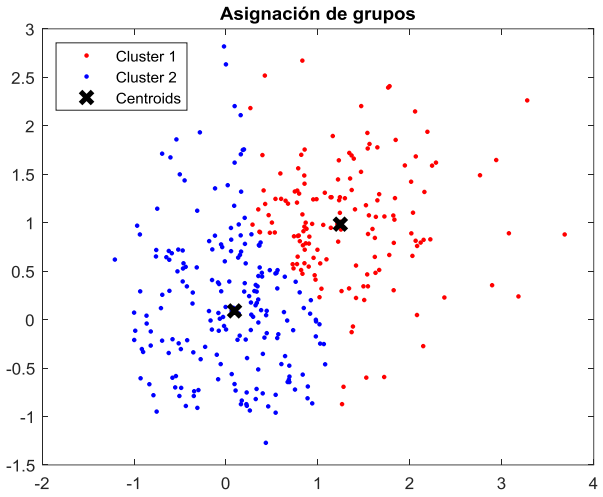


Figura 2.2, Asignación de grupos. K=2.

El inconveniente de esta técnica, es que se requiere conocer la cantidad de grupos "K" a desarrollarse. Para superar este inconveniente hay varios métodos que ayudan a elegir el número de "K" agrupaciones; por ejemplo, el método del codo o "elbow method" en inglés, que es un método que analiza el porcentaje de varianza como una función. Otro, el método de la brecha (GAP) que es semejante al método del codo, pero este encuentra la mayor diferencia entre los grupos existentes, además se dan otros métodos. [5], [26] Sin embargo, ningún método señala cual es el número exacto de agrupaciones que deben desarrollarse; generalmente se elige la cantidad del número de grupos a prueba y error, siempre a criterio del investigador. [5]

B. Métodos supervisados

- **K-Vecinos**

El algoritmo K-Vecinos es un método supervisado, que tiene como único parámetro el valor K, que es el número de vecinos cercanos a considerar. Lo que hace este algoritmo es dar una clase a los nuevos elementos acorde la información proporcionada de los datos de entrenamiento. [4]

El algoritmo es sencillo de aplicar y lo que hace es calcular la distancia entre los nuevos elementos con el conjunto de entrenamiento y, dependiendo del valor K, da una clase a los nuevos elementos. [23], [27]

El algoritmo K-Vecinos es el siguiente:

Algoritmo K-Vecinos

1. Ingresar datos de entrenamiento $E = (X_1, Y_1) \dots (X_n, Y_n)$.
2. Ingresar datos a clasificar $C = (X_1, \dots, X_n)$.
3. Ingresar el valor de K vecinos a considerar.
4. Para todo objeto clasificado calcular la distancia con los datos a clasificar.
5. Quedarse con los K datos de entrenamiento más cercanos a los datos a clasificar.
6. Asignar a X la clase más frecuente.

En la Figura 2.3, se puede ver un ejemplo de ejecución de este algoritmo, en donde el valor K es 3, los datos nuevos están representados de forma cuadricular y los datos de entrenamiento de forma circular.

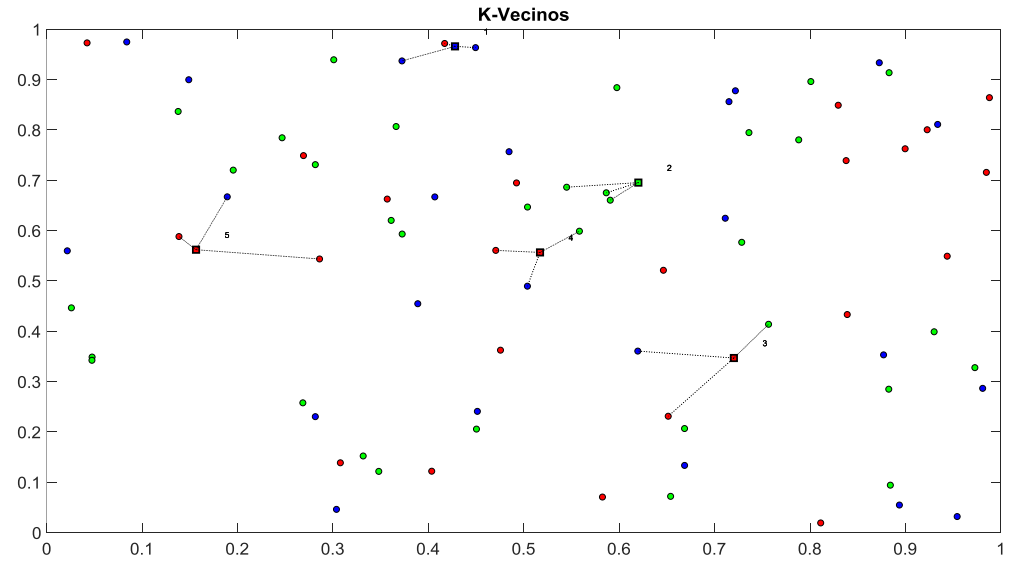


Figura 2.3, Funcionamiento del algoritmo K-Vecinos

Igual que el algoritmo K-Medias, K-Vecinos tiene la desventaja de no saber la cantidad exacta del valor K y no existe un método que ayude a calcular este valor, generalmente K es elegido a prueba y error. [27]

- **Árbol de decisión**

  Una tarea adecuada para el árbol de decisión es el de clasificación. Es un método supervisado, ya que, para la clasificación de los datos, utiliza un conjunto de datos predefinido con clases según los valores de las variables. [18], [20]

  El árbol de decisión, inicia con un nodo y a partir de este surgen nuevos nodos. En base a estos nodos conectados, se realiza la clasificación, por lo que, cada nodo representa una variable del conjunto de entrenamiento.

  Existen varios algoritmos para la creación de los árboles de decisión, entre ellos: ID3, C4.5, CART, CHAID, etc. Cada uno de ellos se distingue por los criterios de división. [20] MATLAB® utiliza como algoritmo de "partición" el árbol de decisión CART. [28] CART es un algoritmo que desarrolla árboles binarios, por lo que cada nodo del árbol estará dividido en dos nodos salientes. [20]

- **Red Neuronal**

  La red neuronal es un algoritmo de aprendizaje, inspirado en la neurona del cerebro humano.

  Es un método supervisado, que se entrena por medio de ejemplos. Las redes neuronales se pueden utilizar para realizar clasificaciones, aproximaciones o predicciones. [29]

  Existe muchos tipos de redes neuronales, pero para los casos de clasificación, usualmente se utiliza el Perceptron Multicapa que utiliza una técnica supervisada denominada backpropagation. [29] La Figura 2.4 muestra la estructura básica de esta red neuronal y como se puede apreciar, esta consta de tres capas:

  - Capa de entrada;
  - Capa Oculta;
  - Capa de salida.

  Salvo los nodos de entrada, cada nodo de la capa oculta y la capa de salida son neuronas que utilizan una función de activación. [29]

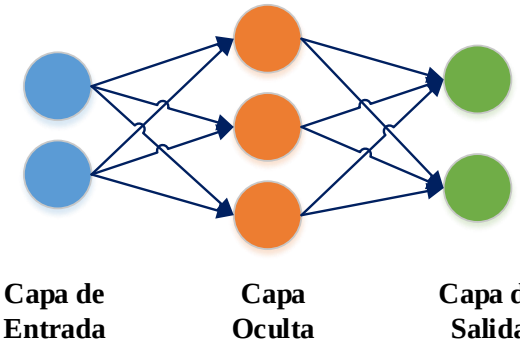


Figura 2.4, Estructura básica de la Red Neuronal Perceptrón Multicapa

El establecimiento de una red neuronal consta de tres etapas:

1. Etapa de entrenamiento: es la etapa de aprendizaje, en donde, se añaden los atributos de entrada (entrada de la red) y se compara con el conjunto objetivo (etiqueta o target). [29]
2. Etapa de Validación: Esta etapa se ejecuta conjuntamente con la etapa de entrenamiento y se realiza para evitar un sobre entrenamiento de la red. [29]
3. Etapa de Prueba: Esta etapa se realiza después de la etapa de entrenamiento y consiste en usar un conjunto de datos distintos a la de entrenamiento y validación para investigar que tan bien aprendió la red al final del proceso. [29]

7. Evaluación e interpretación de los datos

Esta es la fase final del proceso de minería de datos; aquí se analizan los resultados obtenidos y se evalúan comparando los resultados con los objetivos propuestos en la fase 1. Los datos se tienen que mostrar de tal manera que se puedan interpretar para poder hacer uso de esta información extraída. [18], [20]

Hay que tener en cuenta que el objetivo del proceso de minería de datos no es solamente la identificación y la aplicación de una técnica de minería de datos o la integración de un modelo, si no todo el proceso que conlleva y con eso los resultados obtenidos. [23]

Finalmente, ya cumplido con las pruebas necesarias y la corrección del modelo (si es necesario), se realiza el despliegue del mismo mediante una planificación que debe ser organizada por los desarrolladores conjuntamente con los usuarios. [23]

2.2. MINERÍA DE DATOS APLICADA A LOS DATOS DE LA EMPRESA DISTRIBUIDORA

2.2.1. RECOPILACIÓN E INTEGRACIÓN DE DATOS

La empresa distribuidora cuenta actualmente con un sistema de información llamado SAP por sus siglas en inglés "Systems, Applications, Products", dentro del que están diferentes reportes, mismos que son explicados más adelante. También se encontraron otras bases de datos las cuales no están en el sistema SAP, son: base de datos del laboratorio de medidores y base de datos GIS.

Los reportes dentro del sistema SAP, engloba a todos los abonados de la CENTROSUR que (de acuerdo con el corte al 14 de mayo del 2019) son 393.960. Este número de abonados se encuentran en los datos maestros que permite la identificación y ubicación geográfica de cada uno (Número de Cédula, Nombre, Coordenadas).

En la empresa, existen departamentos que utilizan esa información de acuerdo con su funcionalidad, como: Consumo en kWh, demanda en kW, deudas pendientes, por ejemplo, el reporte CLI014 da un histórico de consumo, se va actualizando cada mes y va acumulando los meses anteriores, con un total de 12 meses de consumo kWh de los abonados.

Otros reportes que no están dentro del sistema SAP, como la base del laboratorio de medidores y la base GIS, cuentan también con datos maestros, pero además contienen información sobre marca de medidor, año de fabricación, número de transformador, alimentador, etc.

La información con la que cuenta la empresa distribuidora se encuentra en diferentes bases de datos, esta depende de la tarea y responsabilidad del departamento que administra la información, de tal forma que posteriormente se integre en una sola base de datos para formar una matriz como se presentó en la ecuación (5); a esta matriz se denominará "matriz base".

La matriz base se integra al juntar toda la información considerada importante de acuerdo con el "criterio del experto".

Criterio del experto hace referencia a razonamientos y reglas que son dictadas por un experto en el tema, que ayudan a resolver un problema. En este caso para la recopilación, análisis e integración de datos, estos criterios son desarrollados en base a experiencias de un experto, indicando que datos pueden indicar pérdidas no técnicas.

La Figura 2.5, muestra el proceso realizado para la recopilación de datos.

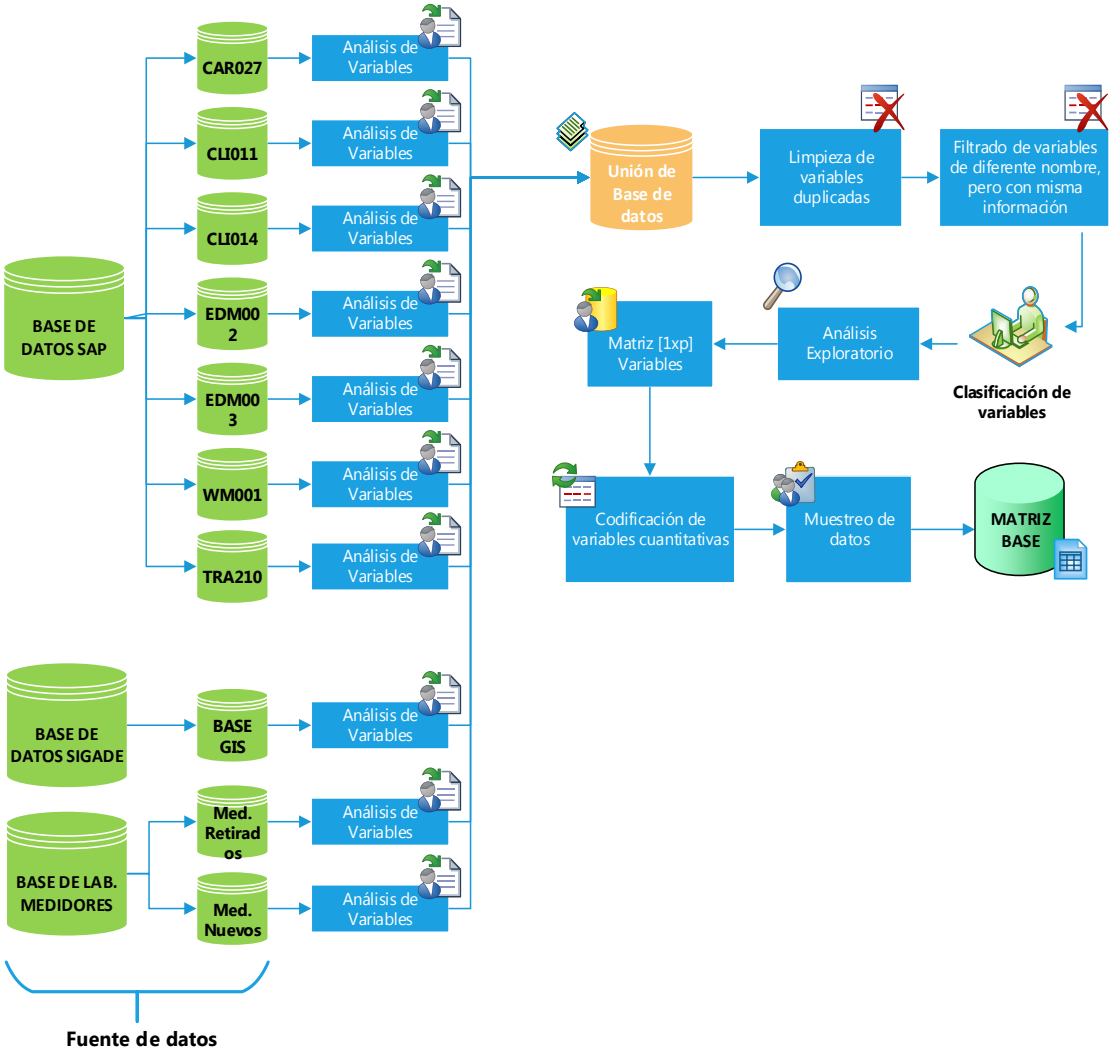


Figura 2.5, Recopilación de datos

El proceso de integración de la "matriz base", se detalla a continuación:

1. **Integración del conjunto de datos:** En este paso se realiza la búsqueda de datos más importantes o relevantes, que nos ayuden a encontrar pérdidas no técnicas.

   Mediante el "criterio del experto", se determina que las siguientes bases de datos son las que tienen información importante:

   - **Base de datos "SAP":** "SAP" significa Sistemas, Aplicaciones y Productos especializados en proceso de datos (En inglés: Systems, Applications and Products). Es el sistema de base de datos que utiliza actualmente la empresa distribuidora, en donde se encuentran datos de los clientes en distintos reportes o bases de datos para diferentes fines. Los siguientes reportes se obtuvieron de la base de datos SAP presentado en la Figura 2.5:
     - **CAR027:** Gestión de "cartera[12]" mensual.

[12] Término financiero, que indica aquellos clientes que se deben tener en cuenta.

        - **CLI011:** Catastro de clientes.
        - **CLI014:** Catastro de clientes – Históricos de consumo.
        - **EDM002:** Listado de órdenes de trabajo.
        - **EDM003:** Listado de órdenes de trabajo "PERC[13]".
        - **WM001:** Listado de órdenes de trabajo.
        - **TRA210:** Catastro mensual de consumos correspondiente a los clientes.
    - **Base de datos del Sistema de Información Geográfica**
        - **Base de datos GIS:** Base de datos de información geográfica (coordenadas, redes eléctricas, transformadores).
    - **Base de datos del laboratorio de medidores**
        - **Medidores retirados:** Datos y resultados de pruebas a medidores retirados del sistema.
        - **Medidores nuevos:** Datos y resultados de pruebas realizados en medidores nuevos.
2. **Análisis de variables:** Se analizan las variables de cada reporte anteriormente presentado, entendiendo que describe y que tipo de información contiene cada variable.
3. **Unión de reportes:** Se unen los reportes, llegando a tener un total de 424 variables.
4. **Limpieza de variables duplicadas:** Se eliminan variables con el mismo nombre.
5. **Filtrado de variables con diferente nombre, pero con la misma información:** Se borra las variables que contienen la misma información. Este paso se realiza para eliminar redundancia de información. Cumplido este paso se tiene trecientas dieciocho variables.
6. **Clasificación de variables:** Se clasifican las variables de la siguiente manera:
    - **Información:** Aquellas variables que aporten con información del cliente, como: "CuentaContrato", "Cuen", "Nombre", "Cédula", etc.
    - **Geográfico:** Variables que indiquen la ubicación geográfica del medidor del cliente, como: "Codparr", "Provincia", "Cantón", etc.
    - **Económico:** Variables que enseñen la relación económica entre el cliente y la EERCS, como: "Fechaultpago", "Meses Adeudados", "Deuda", etc.
    - **Social:** Variables que indiquen un aspecto social con respecto al cliente como: "población".
    - **Técnico:** Variables técnicas, como: "TipoConsumo", "Tensión", "Consumo kWh/mes", etc.
7. **Análisis exploratorio:** Se analizan nuevamente las variables resultantes de los pasos 4, 5, y 6.

[13] Pérdidas comerciales

8. **Reducción de variables por medio del "criterio del experto":** Con el "criterio del experto", se revisa minuciosamente cada variable y se determinan aquellas variables que aportarán información relevante para el control de pérdidas no técnicas.
9. **Matriz de [1xp]:** Ejecutado el paso 8, se determina que la matriz de variables es de $[1 \times 68]$, es decir, 68 variables.

   Estos pasos son importantes y necesarios realizar, como se observó anteriormente, de 424 variables que se tenía al empezar el análisis, la cantidad se redujo a 68 variables, disminuyendo aproximadamente el 84%.
10. **Cuantificación de datos:** Como las variables que integran la matriz de $[1 \times 68]$ se obtuvieron de diferentes reportes, estas no tienen un mismo formato, por lo que, en este paso se codifica algunas variables cuantitativas para el análisis. Esta codificación se muestra posteriormente.
11. **Matriz de datos [X] o matriz base:** Ya con el análisis previamente realizado, se pueden agregar los n abonados y se obtiene la matriz de datos [**X**] o matriz base de tamaño $[n \times p]$. Donde n representa la cantidad de clientes, los cuales al corte de 14 de mayo del 2019 ascienden a 393.960, sin embargo por practicidad de análisis, se ha escogido un universo de 5.615 clientes, correspondientes a las parroquias "Checa", "Chiquintad" y "Octavio Cordero"; concluyendo que el tamaño de la matriz de datos base es de $[5615 \times 68]$.

En la Tabla 2.1, se describen los atributos de la matriz base, el origen de los datos, la forma de vinculación entre los diferentes reportes y el tipo de información que contiene.

Tabla 2.1, Descripción de variables

#	Variable	Reporte	Vinculación	Tipo de Dato	Descripción
V1	CuentaContrato	CLI014	Base	Numérico	Indica el número de cuenta contrato del abonado, el cual sirve como identificación del mismo.
V2	Cuen	CLI014	Base	Numérico	Indica el número de cuenta del abonado.
V3	Cedula	CLI014	Base	Numérico	Muestra el número de cédula del abonado, permite la identificación del mismo.
V4	Nombre	CLI014	Base	Texto	Nombre del abonado.
V5	NMedidor	CLI014	Base	Numérico	Número de medidor del abonado.
V6	Instalacion	CLI014	Base	Numérico	Presenta el número de instalación del abonado.
V7	Direccion	CLI014	Base	Texto	Proporciona la dirección de ubicación del medidor asignado al abonado.
V8	FechaNacimiento	CLI014	Base	Fecha	Muestra la fecha de nacimiento del abonado.
V9	Fecha Pec	CLI011	Nmedidor	Fecha	Indica la fecha de contrato de las cocinas de inducción.
V10	Num. Serie	CLI011	Nmedidor	Alfanumérico	-
V11	Ultima Emisión Por Cliente	CAR027	CuentaContrato	Fecha	Indica la fecha de emisión de la última factura por cliente.
V12	Codparr	CLI014	Base	Numérico	Presenta el código de la parroquia donde se encuentra el medidor asignado al abonado.
V13	Provincia	CLI014	Base	Texto	Indica la provincia donde se encuentra el medidor asignado al abonado.

V14	Canton	CLI014	Base	Texto	Indica el cantón donde se encuentra ubicado el medidor asignado al abonado.
V15	Parroquia	CLI014	Base	Texto	Muestra la parroquia donde se encuentra el medidor asignado al abonado.
V16	Agencia	CLI014	Base	Texto	Muestra la agencia a la que pertenece el abonado.
V17	Coordenada X	CLI014	Base	Numérico	Coordenada X de ubicación del medidor.
V18	Coordenada Y	CLI014	Base	Numérico	Ccoordenada Y de ubicación del medidor
V19	Rurralidad	CLI011	Nmedidor	Texto	Indica si el medidor asignado al abonado está en una zona rural o en una zona urbana.
V20	TarifaCod	CLI014	Base	Alfanumérico	Indica la tarifa del medidor asignado al abonado.
V21	ALIMENTADOR	SIGADE	CuentaContrato	Alfanumérico	Indica el alimentador desde el cual recibe suministro el medidor, el cual permite también tener la ubicación del mismo.
V22	TRAFO	SIGADE	CuentaContrato	Numérico	Muestra el transformador en el cual se encuentra el medidor asignado al abonado.
V23	Partner	CLI011	Nmedidor	Numérico	-
V24	Secuencia	CLI011	Nmedidor	Numérico	Presenta la secuencia de los medidores, acorde a su ubicación geográfica.
V25	CALDES	SIGADE	CuentaContrato	Texto	Inidca la dirección donde se encuentra el medidor asignado al abonado.
V26	USOCOD	SIGADE	CuentaContrato	Texto	Muestra el uso que se da al consumo de la energía.
V27	TerceraEdad	CLI014	Base	Numérico – Codificado	Indica si el abonado es beneficiario de la ley del anciano.
V28	Bdh	CLI014	Base	Numérico - Codificado	Indica si el abonado es beneficiario del bono de desarrollo humano.
V29	TipoConsumo	CLI014	Base	Numérico - Codificado	Indica el tipo de consumo del abonado (privado o público).
V30	Tension	CLI014	Base	Numérico - Codificado	Indica el nivel de tensión del abonado (baja, media o alta).
V31	Fabricante Medidor	CLI014	Base	Numérico - Codificado	Indica el fabricante o la marca comercial de medidor del abonado.
V32	TipMedicion	LAB. MED	Nmedidor	Numérico - Codificado	Indica el tipo de medición del medidor (directo, semi directo e indirecto).
V33	Grupo de consumo	CLI011	Nmedidor	Numérico – Codificado	Indica el grupo de consumo al que pertenece el medidor del abonado (Residencial, Comercial, Industrial, Alumbrado Público u Otros)
V34	Fases	CLI011	Nmedidor	Numérico	Indica el número de fases del medidor asignado al abonado.
V35	Meses Adeudados	CLI014	Base	Numérico	Muestra el número de meses que adeuda el abonado.
V36	Deuda	CLI014	Base	Numérico	Muestra el monto de la deuda del abonado.
V37	Promedio Fact. 6 ult. Meses	CLI014	Base	Numérico	Indica el promedio pagado en los últimos 6 meses por el abonado.
V38	Valor Ultima Factura	CLI011	Nmedidor	Numérico	Muestra el monto de la última factura del abonado.
V39-V51	Variables de consumo	CLI014	Base	Numérico	Consumo actual y consumo de hasta doce meses antes.
V52	Consumo Pro.	CLI014	Base	Numérico	Promedio del consumo actual y consumo de los doce meses anteriores.
V53-V65	Variables de demanda	CLI014	Base	Numérico	Demanda actual y demanda de doce meses anteriores.
V66	Pec por cliente	CLI014	Base	Numérico	Monto a descontar por motivo de tarifa PEC (consumo de inducción, ducha eléctrica o ambas)
V67	Estrato geográfico	CLI014	Base	Numérico - Codificado	Estrato geográfico por consumo promedio del abonado.
V68	Año Fabricación	CLI014	Base	Numérico	Año en el que se fabricó el medidor del abonado.

Como se mencionó anteriormente, algunas variables cualitativas tienen que ser codificadas para el análisis. La codificación de variables se presenta en la Tabla 2.2.

Tabla 2.2, Codificación de variables

#	Atributo	Codificación
V27	TerceraEdad	"X" se cambia por "1" "0" se mantiene
V28	Bdh	"X" se cambia por "1" "0" se mantiene
V29	TipoConsumo	"Privado" se cambia por "1" "Público" se cambia por "0"
V30	Tension	"Baja" se cambia por "1" "Media" se cambia por "2" "Alta" se cambia por "3"
V31	Fabricante Medidor	"ABB" se cambia por "1" "AEM" se cambia por "2" "CIECSA" se cambia por "3" "CONTELECA" se cambia por "4" "ELSTER" se cambia por "5" "FAE" se cambia por "6" "GENERAL ELECTRIC" se cambia por "7" "HEXING" se cambia por "8" "HIKING" se cambia por "9" "HOLLEY" se cambia por "10" "INTECH" se cambia por "11" "ISKRA" se cambia por "12" "KRIZIK" se cambia por "13" "LANDIS" se cambia por "14" "LINTIN" se cambia por "15" "NANSEN" se cambia por "16" "PAFAL" se cambia por "17" "RUSSO" se cambia por "18" "SCO" se cambia por "19" "SCHLUMBERGER" se cambia por "20" "SIEMENS" se cambia por "21" "SIN MARCA CONOCIDA" se cambia por "22" "SONGHE" se cambia por "23" "STAR" se cambia por "24" "SUNRISE" se cambia por "25" "XILI" se cambia por "26" "AEG" se cambia por "27" "DENGLI" se cambia por "28" "ESICO" se cambia por "29" "FUJI" se cambia por "30" "GALILEO" se cambia por "31" "ITRON" se cambia por "32" "JIUMAO" se cambia por "33" "SANXING" se cambia por "34" "SEDCO" se cambia por "35" "WESTINGHOUSE" se cambia por "36" "ACTARIS" se cambia por "37" "UHER" se cambia por "38" "SHENZHEN" se cambia por "39" "GANZ" se cambia por "40" "OSAKI" se cambia por "41" "ION" se cambia por "42" "SCHNEIDER" se cambia por "43"
V32	TipMedicion	"MONOFASICO 1F2H DIRECTO" se cambia por "1" "MONOFASICO 1F3H DIRECTO" se cambia por "2" "MONOFASICO 1F3H INDIRECTO" se cambia por "3" "MONOFASICO SIN REGISTRO" se cambia por "4" "BIFASICO 2F3H DIRECTO" se cambia por "5" "BIFASICO SIN REGISTRO" se cambia por "6" "TRIFASICO 3F4H DIRECTO" se cambia por "7" "TRIFASICO 3F3H INDIRECTO" se cambia por "8" "TRIFASICO 3F4H INDIRECTO" se cambia por "9" "TRIFASICO SIN REGISTRO" se cambia por "10" "SIN MEDIDOR" se cambia por "0"

V33	Grupo de Consumo	"Residencial" se cambia por "1" "Comercial" se cambia por "2" "Industrial" se cambia por "3" "Alumbrado Público" se cambia por "4" “Otros” se cambia por “5”
V67	Estrato Geográfico	Promedio de consumo: 1 - 60 kWh/mes – “E” se cambia por "5" 61 - 110 kWh/mes – “D” se cambia por "4" 111 – 180 kWh/mes – “C” se cambia por "3" 181 – 310 kWh/mes – “B” se cambia por "2" >310 kWh/mes – “A” se cambia por "1"

2.2.2 PRE-PROCESAMIENTO DE DATOS

2.2.2.1. Reconocimiento de los datos

Una vez integrada la matriz base, se realiza un reconocimiento de datos a través de estadística descriptiva. Como se revisó, desde la variable 1 (V1) hasta la variable 26 (V26), son variables que aportan información del cliente (Nombre, cuenta contrato, teléfono, etc.) y desde la variable 27 (V27 - TerceraEdad) en adelante son variables con datos técnicos (consumos, demandas, valores facturados, etc.). Por lo tanto, no es necesario realizar un reconocimiento de datos estadísticos de las variables de información, por ende, solamente se realiza el análisis estadístico de variables con datos técnicos, es decir, desde la variable V27 en adelante.

Tabla 2.3, Resumen estadístico de las variables

Atributo	Nulos	Media	Mediana	Moda	Máximo	Mínimo	Desv. Est.	Coef. Variac.
TerceraEdad	266	0,05	0	0	1	0	0,22	430%
Bdh	266	0,01	0	0	1	0	0,11	889%
TipoConsumo	266	1,01	1	1	2	1	0,10	10%
Tension	266	1,00	1	1	2	1	0,05	5%
FabricanteMedidor	266	11,10	8	8	26	1	6,76	61%
TipMedicion	19	3,67	5	5	10	1	1,89	52%
GrupoDeConsumo	197	1,11	1	1	5	1	0,56	51%
Fases	197	1,66	2	2	3	1	0,48	29%
MesesAdeudados	266	1,10	1	1	30	0	1,91	174%
Deuda	266	11,16	3,84	0	581,83	-5,94	29,08	261%
PromedioFact_6Ult_Meses	266	10,58	6,38	3,56	321,96	0	16,92	160%
ValorUltimaFactura	197	10,48	5,98	3,73	337,64	1,67	17,05	163%
ConsumoKWhActual	266	66,20	46	0	2666,28	-88	115,57	175%
ConsumoKWh1MesAntes	266	57,68	41	0	2169,54	0	91,34	158%
ConsumoKWh2MesesAntes	266	63,34	45	0	2064,48	-142	105,89	167%
ConsumoKWh3MesesAntes	266	65,91	48	0	2006	0	99,64	151%
ConsumoKWh4MesesAntes	266	70,18	49	0	2520,42	0	119,21	170%
ConsumoKWh5MesesAntes	266	59,85	43	0	2360,28	-187	107,83	180%
ConsumoKWh6MesesAntes	266	71,43	51	0	2437,8	-177	125,55	176%
ConsumoKWh7MesesAntes	266	57,49	40	0	2186,88	0	104,28	181%
ConsumoKWh8MesesAntes	266	69,99	47	0	2482,68	-61	127,29	182%
ConsumoKWh9MesesAntes	266	68,68	44	0	2986,56	-36	139,45	203%

ConsumoKWh10MesesAntes	266	65,58	42	0	2098,14	-543	121,77	186%
ConsumoKWh11MesesAntes	266	69,15	46	0	2827,44	-242,5	128,50	186%
ConsumoKWh12MesesAntes	266	60,08	40	0	1911	-193	103,31	172%
ConsumoPro_	266	64,95	46,83	0	2258,705	-15,33	108,13	166%
DemandaActualKW	266	0,01	0	0	8,16	-0,68	0,25	2578%
DemandaKW1MesAntes	266	0,01	0	0	8,16	0	0,27	2540%
DemandaKW2MesesAntes	266	0,01	0	0	8,16	0	0,27	2540%
DemandaKW3MesesAntes	266	0,01	0	0	8,16	-1,4	0,31	2533%
DemandaKW4MesesAntes	266	0,01	0	0	8,16	0	0,24	2625%
DemandaKW5MesesAntes	266	0,01	0	0	8,16	0	0,23	2718%
DemandaKW6MesesAntes	266	0,01	0	0	8,16	0	0,23	2718%
DemandaKW7MesesAntes	266	0,01	0	0	8,16	-6	0,23	2718%
DemandaKW8MesesAntes	266	0,01	0	0	10,2	0	0,28	2807%
DemandaKW9MesesAntes	266	0,01	0	0	13,26	0	0,37	2873%
DemandaKW10MesesAntes	266	0,01	0	0	10,2	0	0,30	2660%
DemandaKW11MesesAntes	266	0,01	0	0	11,22	0	0,30	3491%
DemandaKW12MesesAntes	266	0,01	0	0	13,26	0	0,35	3533%
PecPorCliente	266	0,13	0	0	72,37	0	2,59	1991%
EstratoGeografico	266	4,91	5	5	10	0	2,05	42%
A_oFabricacion	19	1998,09	2013	2015	2017	0	167,25	8%

En la Tabla 2.3, se señalan los datos estadísticos de las variables mencionadas, en donde se puede notar lo siguiente:

- Todas las variables presentaron una cierta cantidad de datos en blanco o nulos.
- En las variables de consumo se observan grandes diferencias entre los valores máximos y mínimos e incluso se presentan porcentajes altos de los coeficientes de variación; esto debido a que la matriz base analizada contiene sistemas de medición que pertenecen a las parroquias de “Checa”, “Chiquintad” y “Octavio Cordero”, en donde existen clientes residenciales, comerciales e industriales, por lo cual, el consumo de estos varía considerablemente.
- También se puede observar que el valor mínimo de las variables “Deuda”, variables de consumo y de demanda son valores negativos, esto se puede presentar por las refacturaciones de consumo realizadas por la empresa distribuidora con origen en errores de lectura y/o mala aplicación tarifaria.
- Otra importante observación es el valor de cero para la moda, situación que se origina en la existencia de sistemas de medición de difícil acceso para la toma de lecturas, casas de zonas rurales que se encuentran abandonadas por parte de sus propietarios, o casos de sistemas de medición dañados sin registro de consumo, siendo éste último de los casos el que reviste de mayor interés para la empresa de distribución.

- También se aprecia una alta diferencia entre valores máximo y mínimo, lo cual puede causar inconvenientes al momento de aplicar las técnicas de clasificación o agrupamiento, como por ejemplo un mal desarrollo de los grupos. Para evitar este tipo de problemas se emplean técnicas de normalización de datos revisadas anteriormente.

2.2.2.2. Limpieza de datos

A. Datos Inexistentes

Para el tratamiento de datos inexistentes o nulos de la matriz base se utiliza Microsoft EXCEL y MATLAB®, con los cuales se verifica lo siguiente:

- EXCEL reconoce los datos faltantes como N/A.
- MATLAB® reconoce como NaN (No es un número) a los datos inexistentes.

La Tabla 2.4, presenta el tratamiento dado a estos datos.

Tabla 2.4, Tratamiento de datos nulos

Variable	Atributo	Tratamiento de dato faltante
V27	TerceraEdad	Se coloca 0 en celdas con datos nulos
V28	Bdh	Se coloca 0 en celdas con datos nulos
V29	TipoConsumo	Se elimina las filas con presencia de datos nulos
V30	Tension	Se elimina las filas con presencia de datos nulos
V31	FabricanteMedidor	Se elimina las filas con presencia de datos nulos
V32	TipMedicion	Se elimina las filas con presencia de datos nulos
V33	GrupoDeConsumo	Se elimina las filas con presencia de datos nulos
V34	Fases	Se elimina las filas con presencia de datos nulos
V35	MesesAdeudados	Se elimina las filas con presencia de datos nulos
V36	Deuda	Se elimina las filas con presencia de datos nulos
V37	PromedioFact_6Ult_Meses	Se elimina las filas con presencia de datos nulos
V38	ValorUltimaFactura	Se elimina las filas con presencia de datos nulos
V39-V51	Variables de consumo	Se elimina las filas con presencia de datos nulos
V52	ConsumoPro	Se elimina las filas con presencia de datos nulos
V53-V65	Variables de demanda	Se elimina las filas con presencia de datos nulos
V66	PecPorCliente	Se coloca 0 en celdas con datos nulos
V67	EstratoGeografico	Se elimina las filas con presencia de datos nulos
V68	AñoFabricacion	Se coloca 0 en celdas con datos nulos

De la matriz base muestreada se encontraron 266 sistemas de medición con datos nulos, los cuales fueron eliminados de la lista, esto con la finalidad de evitar generar inconsistencias en el desarrollo de las técnicas de minería de datos. Este proceso se realiza en MATLAB®, el código desarrollado se puede observar en ANEXO A2.

B. Datos atípicos

La identificación de datos atípicos de la matriz base, se realiza mediante un análisis exploratorio de datos. A través del comando "*scatter*" de MATLAB®, se realiza la dispersión existente de datos, los resultados de la identificación se pueden observar en la Figura 2.6.

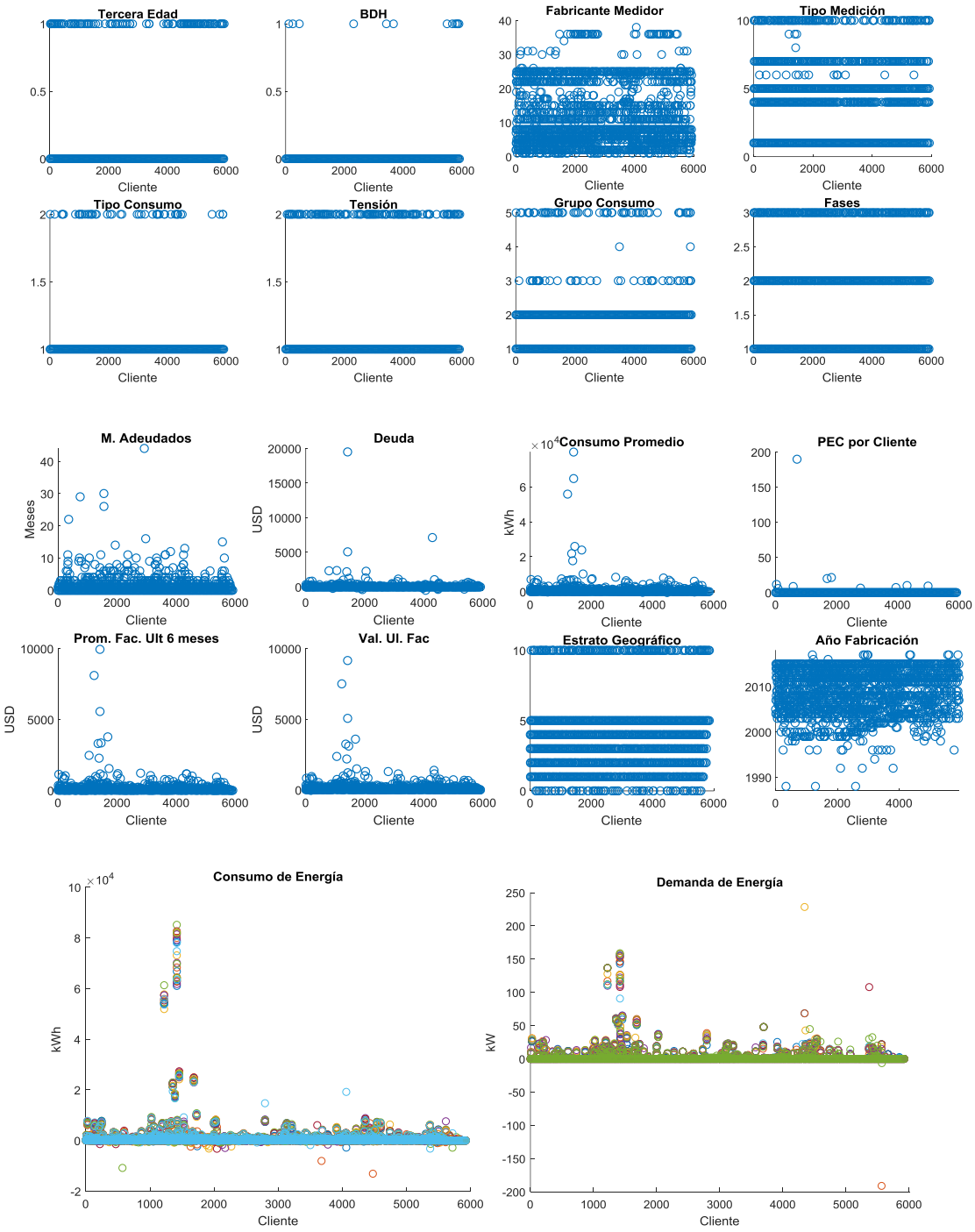


Figura 2.6, Dispersión de datos

Con el análisis exploratorio de los datos, se determina el siguiente tratamiento para los datos atípicos:

Tabla 2.5, Tratamiento a datos atípicos

#	Atributo	Tratamiento de datos atípicos
V27	TerceraEdad	• Se reconoce como dato atípico aquel que es distinto a 0 o 1.
V28	Bdh	• Se reconoce como dato atípico aquel que es distinto a 0 o 1.
V29	TipoConsumo	• Se reconoce como dato atípico aquel que es distinto a 1, 2, o 3.
V30	Tension	• Se reconoce como dato atípico aquel que es distinto a 1, 2, o 3.
V31	FabricanteMedidor	• Se reconoce como dato atípico aquel que es distinto a 1, 2, …, o 43.
V32	TipMedicion	• Se reconoce como dato atípico aquel que es distinto a 0, 1, 2,…, o 10.
V33	GrupoDeConsumo	• Se reconoce como dato atípico aquel que es distinto a 0, 1, 2, 3, 4 o 5.
V34	Fases	• Se reconoce como dato atípico aquel que es distinto a 1, 2, o 3.
V35	MesesAdeudados	• Se observa datos alejados, pero no se considera eliminar.
V36	Deuda	• Se considera dato atípico a los valores menores a 0, es decir, no se considera a aquellos clientes que tienen refacturación.
V37	PromedioFact_6Ult_Meses	• Se observa datos alejados, pero no se considera eliminar de la lista, porque puede tratarse de clientes de consumo alto.
V38	ValorUltimaFactura	• Se observa datos alejados, pero no se considera eliminar de la lista, porque puede tratarse de clientes de consumo alto.
V39-V51	Variables de consumo	• Se observa datos alejados, sin embargo, se considera eliminar solamente a clientes que tienen consumos negativos.
V52	ConsumoPro	• Se observa datos alejados, pero no se considera eliminar de la lista.
V53-V65	Variables de demanda	• Sistemas de medición con estas demandas negativas son eliminados de la lista.
V66	PecPorCliente	• Se observa datos alejados, pero no se considera eliminar de la lista, ya que, son pocos y no son datos atípicos.
V67	EstratoGeografico	• Se considera dato atípico a valores distintos a 1, 2, 3, 4 o 5.
V68	AñoFabricacion	• Se elimina datos negativos.

Se encontraron en total 86 sistemas de medición que presentaron datos atípicos, los cuales fueron eliminados de la lista para evitar errores en la aplicación de las técnicas de minería de datos. El análisis se realiza en MATLAB® y el código desarrollado se muestra en ANEXO A2.

C. Matriz de correlaciones

Se construye la matriz de correlaciones para identificar la relación o dependencia existente entre las variables.

En la matriz de correlaciones (ANEXO A1) se observa alta correlación desde la variable V37 (PromedioFact_6Ult_Meses), hasta la variable V65 (Demanda 12 meses antes).

Sin embargo, pese a esa apreciación, no se pueden descartar aquellas variables que no se encuentran correlacionadas, pues podrían ser requeridas posteriormente, estas variables pueden contener información relevante para el proceso de control de pérdida de energía, que hace necesario recurrir al "criterio del experto".

2.2.3. PROCESAMIENTO DE DATOS

2.2.3.1. Métodos No Supervisados

1) Identificación de clientes anómalos mediante el Coeficiente de Pearson

Este modelo fue implementado por *I. Monedero, et al*. [25] Es un método no supervisado que se basa en la identificación de sistemas de medición "sospechosos" de patrones de consumo mediante técnicas estadísticas (coeficiente de correlación de Pearson), pues las variaciones en el consumo pueden corresponder a variaciones naturales por un cambio de patrón de consumo o demanda del servicio, por variaciones con origen en fallas del equipo de medición, o por modificaciones voluntarias del consumo del cliente, etc.

Para identificar este tipo de variaciones, se aplica el coeficiente de correlación de Pearson, que en estadística indica la relación existente entre dos variables. [25]

Se ejecutan los algoritmos realizados en [25] por *I. Monedero, et al*. Para el tema de estudio, en donde X es la variable de tiempo (generalmente medido de forma mensual), y la variable Y que es el valor del consumo que registra el equipo de medición del cliente.

Los algoritmos son:

a) Algoritmo para la detección de caída drástica con posterior estabilización

Algoritmo
1. Dividir el consumo en tres ventanas de tiempo: Ventana 1: Primeros 6 valores de consumo. Ventana 2: Dos valores intermedios de consumo. Ventana 3: Últimos 5 valores de consumo. 2. Girar 45° en el eje del tiempo a las ventanas 1 y 3. 3. Calcular el coeficiente de Pearson para las ventanas 1 y 3. 4. Aplicar: Si $CoeffPear_{V1} > 0.8\ y\ CoeffPear_{V3} > 0.8\ y\ Promedio\_V1 > 4 * Promedio\_V3$ Entonces: El sistema de medición es sospechoso Caso contrario: Sistema de medición regular

Donde:

$CoeffPear_{V1}$: Es el coeficiente de Pearson de la ventana 1 (girado 45°).

$CoeffPear_{V3}$: Es el coeficiente de Pearson de la ventana 3 (girado 45°).

$Promedio\_V1$: Es el promedio de la ventana 1.

$Promedio\_V3$: Es el promedio de la ventana 3.

En la Figura 2.7, se puede ver el patrón de consumo de cuatro clientes que se detectaron mediante este algoritmo.

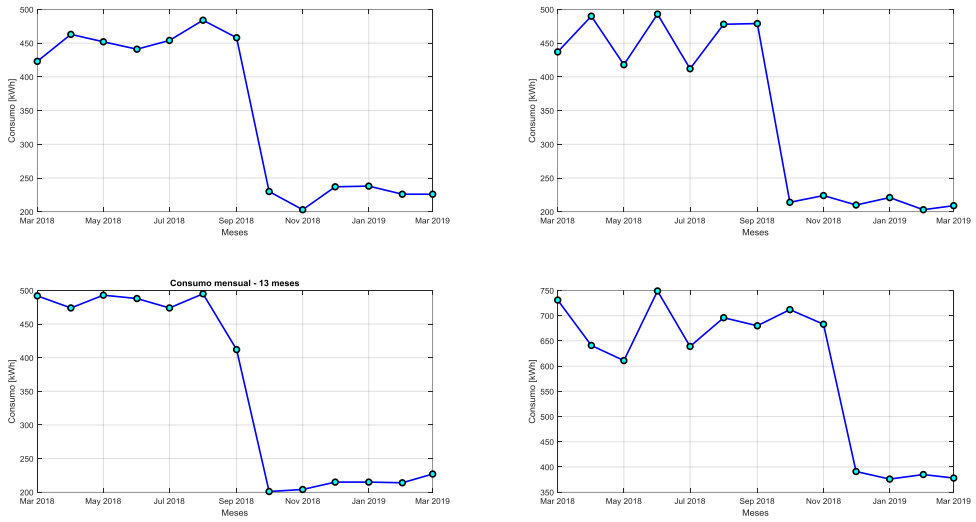


Figura 2.7, Consumos de los sistemas de medición detectados con el algoritmo a

b) Algoritmo para la detección de caída progresiva con posterior estabilización

Algoritmo
1. Dividir el consumo en dos ventanas de tiempo: Ventana 1: Primeros 7 valores de consumo. Ventana 2: Últimos 6 valores de consumo. 2. Calcular el coeficiente de Pearson para la ventana 1. 3. Girar 45° en el eje del tiempo a la ventana 2. 4. Calcular el coeficiente de Pearson de la ventana 2, girado 45°. 5. Aplicar: Si $CoeffPear_{V1} < -0.75\ y\ CoeffPear_{V2} > 0.75\ y\ Promedio\_V1 > 2 * Promedio\_V2$ Entonces: El sistema de medición es sospechoso. Caso contrario: Sistema de medición regular.

Donde:

$CoeffPear_{V1}$: Es el coeficiente de Pearson de la ventana 1.

$CoeffPear_{V2}$: Es el coeficiente de Pearson de la ventana 2 (girado 45°).

$Promedio\_V1$: Es el promedio de la ventana 1.

$Promedio\_V2$: Es el promedio de la ventana 2

Ejemplos del patrón de consumo que detecta este algoritmo, se pueden observar en Figura 2.8.

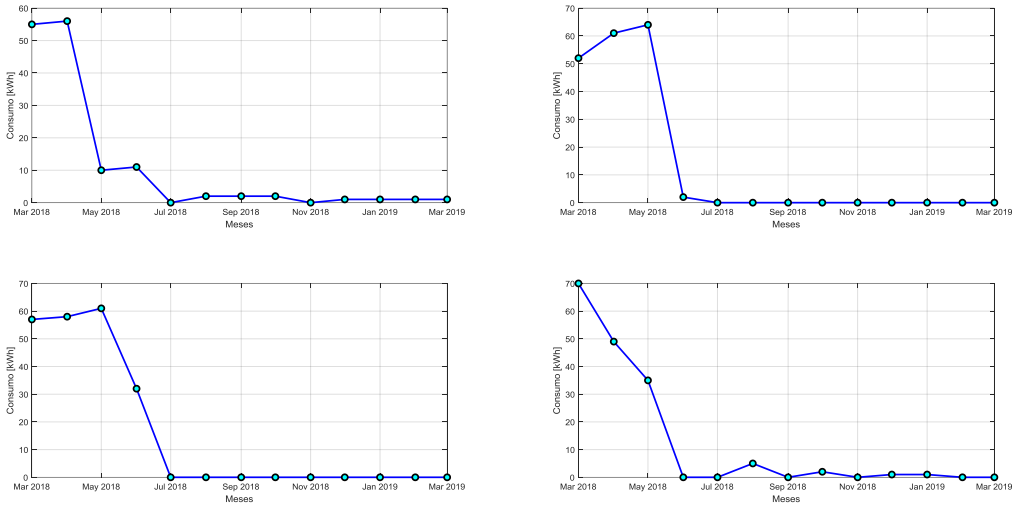


Figura 2.8, Consumos de los sistemas de medición detectados con el algoritmo b

En estos algoritmos, se considera un período de seis meses para la ventana final, debido a que es un tiempo considerable para la estabilización y un tiempo justo para la detección de cualquier anomalía. Cabe recalcar además que si el tiempo de esta ventana es menor (por ejemplo 3 meses) el período será muy corto para considerar la estabilización, esta caída de consumo puede deberse a otras razones, como inmuebles abandonados, negocios cerrados, etc. [25]

2) Agrupamiento K-Medias

Para la aplicación de este algoritmo, se utilizan las variables que tuvieron un alto grado de correlación (véase el ANEXO A1), las variables utilizadas para el agrupamiento son:

- PromedioFact_6Ult_meses;
- ValorUltimaFactura;
- Variables de consumo.

Además, para el agrupamiento en lugar de las variables de consumo, se añaden atributos estadísticos obtenidos de las mismas. Los atributos añadidos son:

- Promedio: Promedio de consumo del mes actual y doce meses antes;
- Desviación estándar: Desviación del consumo mensual de energía;
- Coeficiente de variación: es el porcentaje de desviación estándar dividido entre el promedio;
- Max: Máximo consumo mensual;
- Min: Mínimo consumo mensual;
- Rango: Diferencia entre el máximo consumo y el mínimo consumo.

En total se tienen 9 variables para el agrupamiento.

Como la técnica K-Medias se basa en similitudes para el agrupamiento, los datos tienen que estar normalizados y se utiliza la ecuación (7) para la normalización.

Además de la normalización, previo el agrupamiento con la técnica K-Medias, se realiza una pre-agrupación de sistemas de medición. Para esto se considera agrupar por tipos de tarifas, es decir, se agrupa en consumos de tipo residencial, comercial, industrial y otros.

El algoritmo K-Medias se aplica con el Toolbox de MATLAB® y en la Figura 2.9 se observa un ejemplo de las agrupaciones. En el capítulo 3 se realiza una descripción más detallada del funcionamiento del algoritmo.

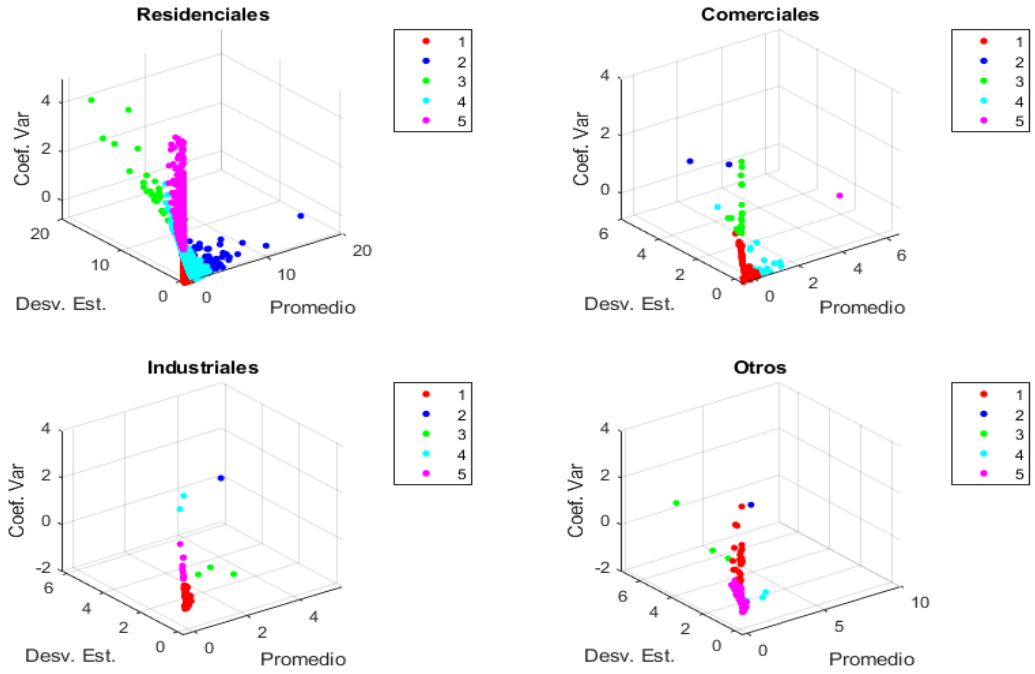


Figura 2.9, Agrupamientos con K-Medias

El código realizado en MATLAB® se puede observar en el ANEXO A2.

2.2.3.1 Métodos Supervisados

A diferencia de las técnicas anteriores, los "métodos supervisados", se entrenan en base de ejemplos, que tienen que estar correctamente etiquetados. Esto puede ser un problema, en el caso de no tener una base de datos con suficientes datos etiquetados. [14]

Para el entrenamiento, de los siguientes métodos supervisados, se utiliza una base de datos con únicamente variables de consumo (trece valores de consumo) de 2062 ejemplos. Entre ellos 1031 datos son etiquetados con "1" (fraude) y 1031 datos son etiquetados con "0" (no fraude).

Los métodos supervisados que se implementan son:

1) **K-Vecinos**

Para la implementación del algoritmo, se toma en consideración los siguientes atributos:

- Promedio: Media de consumo de 13 meses.
- Desviación estándar.
- Coeficiente de variación.
- Min: Mínimo consumo en el período de 13 meses.
- Max: Máximo consumo en el período de 13 meses.
- Rango: Diferencia entre el máximo valor de consumo y el mínimo valor de consumo.

En total, se consideran 6 variables para la clasificación. Para la ejecución de la técnica, los datos son normalizados con la ecuación (6).

El algoritmo se implementa con el Toolbox de MATLAB® y el código desarrollado se puede apreciar en ANEXO A2.

En el capítulo 3 se explica más detalladamente este algoritmo.

2) **Árbol de decisión**

De manera similar, para el entrenamiento de este algoritmo, se utilizan las 6 variables utilizadas en el algoritmo de vecinos más cercanos, es decir, promedio, desviación estándar, coeficiente de variación, mínimo consumo, máximo consumo y rango de consumo.

Previo al entrenamiento y clasificación, los datos son normalizados mediante la ecuación (6).

Para el desarrollo del árbol de decisión se utiliza el Toolbox de MATLAB®. El código desarrollado se observa en el ANEXO A2.

Como se mencionó anteriormente, MATLAB® utiliza como algoritmo de "partición" el árbol de decisión CART.

El árbol de decisión generado se puede apreciar en ANEXO A3.

3) **Red Neuronal**

Para el entrenamiento de la red, se utilizan las mismas variables utilizadas en los métodos mencionados anteriormente, es decir, promedio de consumo, desviación estándar, coeficiente de variación, mínimo, máximo y rango de consumo. Estos datos tienen que estar normalizados y para ello se utiliza la ecuación (6).

La creación y el entrenamiento de la red neuronal se realiza mediante el Toolbox de MATLAB®, en el que se utiliza la Red Neuronal tipo Perceptron Multicapa. La Red Neuronal implementada se presenta en la Figura 2.10, en la cual se observa que está constituida, en la capa de entrada, por las 6 variables mencionadas; la capa oculta está formada por 10 neuronas y la capa de salida cuenta con 1 neurona para la clasificación.

El algoritmo de entrenamiento es el Leveberg-Marquardt backpropagation y la función de activación es la sigmoidal.

Los datos fueron fraccionados aleatoriamente en 3 partes: 70% para el entrenamiento, 15% para la validación y 15% de prueba.

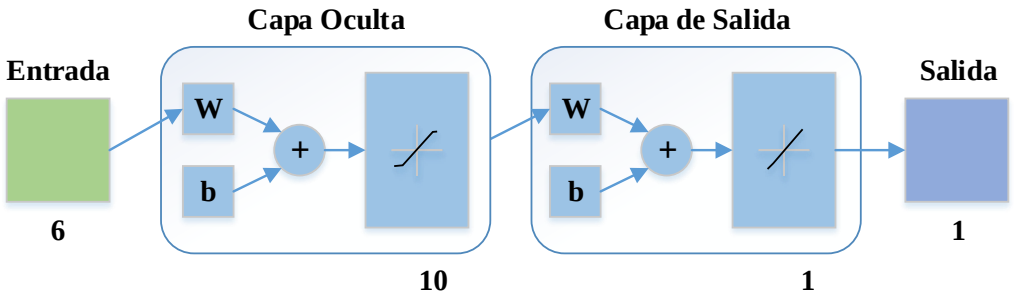


Figura 2.10, Estructura de la Red Neuronal

El proceso de entrenamiento de la red neuronal se puede ver en ANEXO A4.

2.3. Resumen de la minería de datos aplicado a los datos de la empresa distribuidora

En general, la Figura 2.11 muestra el resumen del proceso de minería de datos realizada a los datos de la empresa distribuidora.

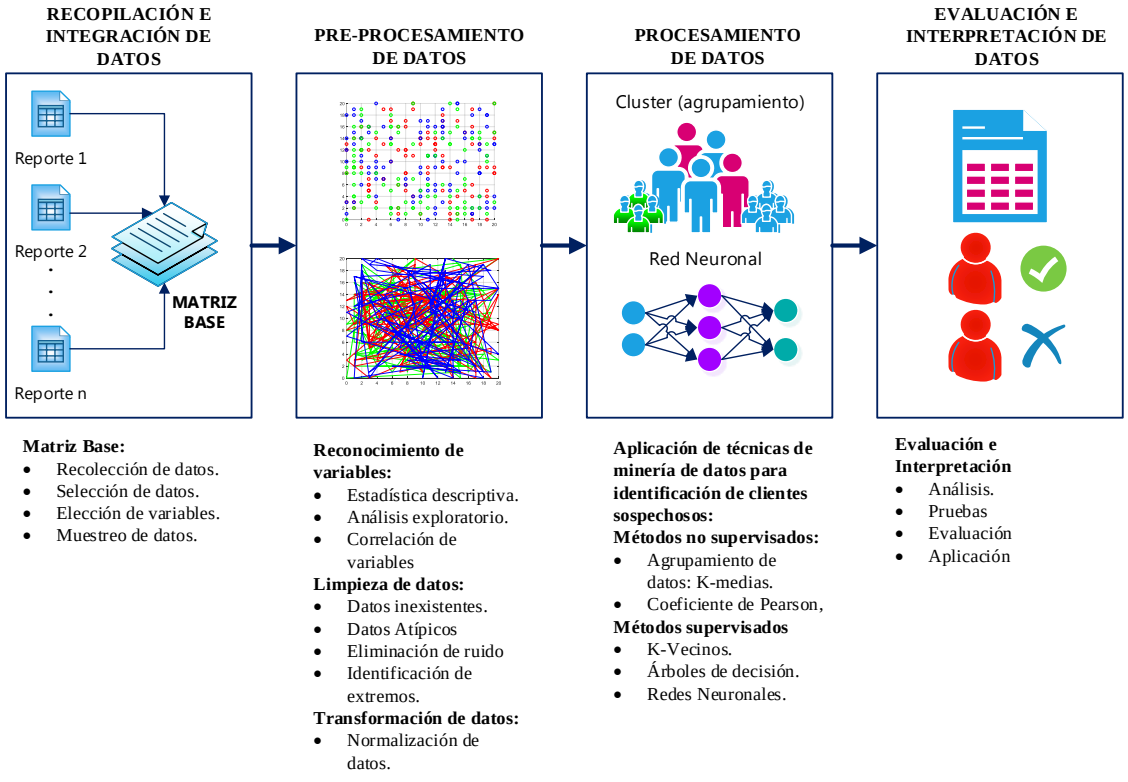


Figura 2.11, Proceso de minería de datos

- **Etapa 1: Recopilación e integración de datos:** En esta etapa se realiza una recopilación de datos de la empresa distribuidora, reconocimiento de variables y un muestreo de datos; el objetivo es contar con una base de datos que ayude con la identificación de patrones en aquellos clientes causantes de pérdidas de energía.
- **Etapa 2: Pre – procesamiento de datos:** Se realiza una limpieza de los datos, es decir, se elimina cualquier dato que pueda alterar el proceso de los algoritmos de agrupamiento o clasificación. Se considera un dato anómalo aquel que es muy distinto a los demás, pudiendo ser este un dato inexistente o un dato atípico.
- **Etapa 3: Procesamiento de datos:** Con los datos una vez "limpios", se procede a aplicar alguna de las técnicas ya revisadas en el capítulo 1, con la finalidad de obtener listados de sistemas de medición que tengan patrones con indicativos de posible fraude o daño, para posteriormente proceder con las revisiones en sitio.

  Para ello se aplican dos técnicas no supervisadas: 1) el agrupamiento k-medias; y 2) la detección de sistemas de medición anómalos mediante el coeficiente de Pearson. Además, se aplican tres técnicas supervisadas: 1) k-vecinos, 2) árbol de decisión y 3) red neuronal. Estos algoritmos son aplicados en MATLAB®.
- **Etapa 4: Evaluación e interpretación de resultados:** Esta etapa evalúa las técnicas aplicadas, la cual será explicada y analizada en detalle en el capítulo 3.

3. CAPÍTULO 3 – EVALUACIÓN E INTERPRETACIÓN DE RESULTADOS

Este capítulo evalúa las técnicas de minería de datos expuestas en el capítulo anterior mediante "métricas"; los resultados de esta evaluación, permiten proponer una metodología para la determinación de listados de sistemas de medición, que contienen una mayor probabilidad de tener un daño y/o sufrir algún tipo de manipulación que afecte la cuantificación del consumo de energía eléctrica, y que por lo tanto requieran una verificación de su funcionamiento en campo (o lugar en donde se encuentren instalados). Se generaron aleatoriamente listados de sistemas de medición para ser revisados en sitio y al final realizar un análisis técnico-económico, lo que permitirá evaluar la efectividad de los listados despachados, así como también obtener las conclusiones y las recomendaciones que conlleven mejorar la efectividad de la metodología de control, en el futuro.

3.1. EVALUACIÓN DE LAS TÉCNICAS DE MINERÍA DE DATOS MEDIANTE MÉTRICAS

Las métricas que se utilizan para la evaluación son: Exactitud, razón de verdaderos positivos (TPR), precisión o confianza y razón de falsos positivos (FPR) (ecuaciones (1, (2, (3 y (4 respectivamente).

La evaluación se realiza con una matriz de datos con 400 ejemplos de sistemas de medición comprobados de fraude y no fraude; entre ellos: 200 etiquetados con 1 (fraude) y 200 etiquetados con 0 (no fraude).

3.1.1. Evaluación de la técnica no supervisada K-Medias

Para la evaluación de la técnica no supervisada K-Medias, se ejecuta el algoritmo agrupando los ejemplos de sistemas de medición en diferentes números de grupos K, es decir, formaciones de K igual a 2, 3, 5, 7 y 9 grupos.

En la Figura 3.1, se aprecia los sistemas de medición formados, en:

(a) K = 2;
(b) K = 3;
(c) K = 5;
(d) K = 7;
(e) K = 9;

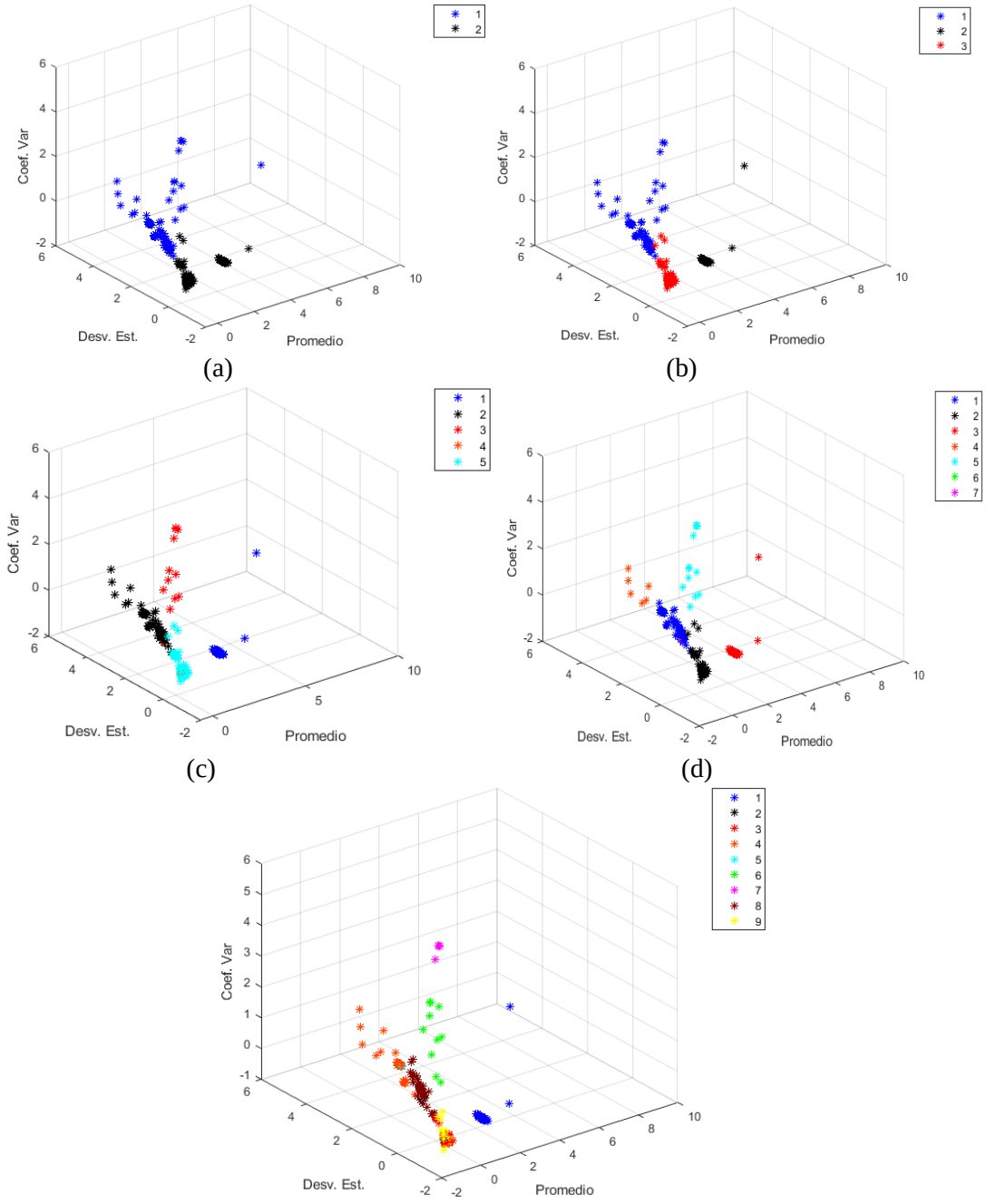


Figura 3.1, Resultado de agrupamiento K-Medias. (a) K=2; (b) K=3; (c) K=5; (d) K=7; (e) K=9

Este algoritmo si bien no permite identificar que grupo o grupos se consideran como fraudulentos y como no fraudulentos, partiendo de la información ingresada, posibilita detectar las pérdidas no técnicas de energía.

Las técnicas de agrupamiento de datos posibilitan tener como fraudulentos a aquellos datos que se observan alejados de la media, porque se presentan como "atípicos" o que de acuerdo a los parámetros ingresados para el agrupamiento son distintos a los demás, haciendo necesario revisar que está sucediendo en esos sistemas de medición.

Los grupos electos como fraudulentos de la Figura 3.1 son:

(a) Grupo 2;
(b) Grupos 2 y 3;
(c) Grupos 1, 2 y 4;
(d) Grupos 3, 4, 5 y 6;
(e) Grupos 1, 2, 5, 6 y 8.

La Tabla 3.1 muestra los resultados obtenidos y las métricas calculadas; se obtuvieron buenos resultados al formar 2, 3, 5 y 7 grupos, consiguiendo números altos de verdaderos positivos y verdaderos negativos y números bajos de falsos positivos y falsos negativos; con esto, altos porcentajes de exactitud, TPR y precisión y, bajo porcentaje en FPR; en tanto que con 9 grupos, el resultado fue medio.

Pese a que los resultados fueron buenos, podría suceder que en otros casos con menos o más grupos se logre buenos o malos resultados; es decir, no hay un método que ayude a determinar el número correcto de grupos a desarrollarse y cuál de ellos elegir como fraudulentos; dependiendo entonces de la cantidad de datos con que se cuente, el número de grupos se tendrá que elegir de manera empírica.

Tabla 3.1, Evaluación del algoritmo K-Medias

K-Medias	2 Grupos	3 Grupos	5 Grupos	7 Grupos	9 Grupos
Verdaderos positivos (TP)	160	159	158	158	98
Verdaderos negativos (TN)	167	152	152	152	152
Falsos positivos (FP)	33	48	48	48	48
Falsos negativos (FN)	40	41	42	42	102
Total	**400**	**400**	**400**	**400**	**400**
MÉTRICAS					
Exactitud	82%	78%	78%	78%	63%
TPR	80%	80%	79%	79%	49%
Precisión	83%	77%	77%	77%	67%
FPR	17%	24%	24%	24%	24%

3.1.2. Evaluación de las técnicas supervisadas

La Tabla 3.2 presenta los resultados de evaluar la técnica K-Vecinos con diferentes valores de K, esto con el objetivo de identificar el valor K que de mejores resultados para la aplicación de esta técnica, pues como se indicó en el capítulo 2 no existe un método adecuado para obtener este dato.

Como se observa de los valores con los que se probó la técnica, el que dio mejores resultados fue K=10, de aquí en adelante para el desarrollo de este algoritmo K tendrá este valor.

Tabla 3.2, Evaluación con métricas de la técnica supervisada K-Vecinos

K-Vecinos	K=2	K=3	K=5	K=10	K=20
Verdaderos positivos (TP)	25	32	48	66	49
Verdaderos negativos (TN)	76	76	88	95	36
Falsos positivos (FP)	124	124	112	105	164
Falsos negativos (FN)	175	168	152	134	151
Total	**400**	**400**	**400**	**400**	**400**
MÉTRICAS					
Exactitud	25%	27%	34%	40%	21%
TPR	13%	16%	24%	33%	25%
Precisión	17%	21%	30%	39%	23%
FPR	62%	62%	56%	53%	82%

La Tabla 3.3 presenta los resultados de la evaluación con métricas de las técnicas supervisadas. En esta, podemos observar que, de los tres métodos la técnica que presentó mejores resultados fue la red neuronal.

La red neuronal obtuvo porcentajes considerables de exactitud, TPR y precisión (59%, 60% y 59% respectivamente), sin embargo, presentó valores altos de FPR (43%), indicando que existe altos números de falsos positivos.

Tabla 3.3, Evaluación con métricas de las técnicas supervisadas

Técnicas Supervisadas	Red Neuronal	Árbol de decisión	K-Vecinos
Verdaderos positivos (TP)	120	80	66
Verdaderos negativos (TN)	115	74	95
Falsos positivos (FP)	85	126	105
Falsos negativos (FN)	80	120	134
Total	**400**	**400**	**400**
MÉTRICAS			
Exactitud	59%	39%	40%
TPR	60%	40%	33%
Precisión	59%	39%	39%
FPR	43%	63%	53%

Comparando los resultados de las técnicas de minería de datos; el agrupamiento K-Medias, es el que dio mejores resultados; sin embargo, se debe tener en cuenta que el entrenamiento de las técnicas supervisadas, al no tener una base de datos con suficientes ejemplos, impidió una evaluación acertada.

3.1.3. Resultados de combinar una técnica no supervisada con una supervisada

Se realizó una evaluación, aplicando juntas una técnica no supervisada (K-Medias) con una técnica supervisada. Esto es que sistemas de medición considerados como fraudulentos obtenidos del K-Medias, son clasificados mediante una técnica supervisada. El resultado fue el siguiente:

Tabla 3.4, Resultados de juntar una técnica no supervisada con una supervisada

	K-Medias + Red Neuronal	**K-Medias + Árbol de Decisión**	**K-Medias + K-Vecinos**
Verdaderos positivos (TP)	174	109	106
Verdaderos negativos (TN)	169	123	133
Falsos positivos (FP)	31	77	67
Falsos negativos (FN)	26	91	94
Total	400	400	400
MÉTRICAS			
Exactitud	86%	59%	60%
TPR	87%	55%	53%
Precisión	85%	59%	61%
FPR	16%	39%	34%

Para la evaluación se utilizó dos grupos para K-Medias y como se muestra en la Tabla 3.4, Al combinar las técnicas se obtuvo un mejor resultado, en donde los porcentajes de exactitud, TPR y precisión, mediante combinaciones se incrementó; aunque FPR disminuyó relativamente.

De las combinaciones realizadas, K-Medias con Red Neuronal resultó ser la más eficiente, presentó los porcentajes más altos de exactitud, TPR y precisión, y el porcentaje más bajo de FPR.

Cabe recalcar que la técnica de detección de patrones de consumo mediante el coeficiente de Pearson no fue evaluada con métricas, porque esta técnica solo detecta casos específicos, omitiendo casos que pueden dar lugar a las excepciones de las pérdidas no técnicas.

3.2.METODOLOGÍA

El análisis previo permite, proponer una metodología que conduzca a mejorar y optimizar los procesos operativos y administrativos de la empresa distribuidora en lo relacionado a las pérdidas no técnicas.

El método está conformado por un total de 5 fases: planificación, pre-procesamiento de datos, aplicación de técnicas de análisis y minería de datos, filtrado de resultados y despacho de órdenes de trabajo a campo; a continuación, se detalla cada fase:

1. **Fase 1: Planificación**

   Paso 1: Criterio de revisión

   Una vez obtenida la base de datos (con variables técnicas, económicas y sociales) del sistema de comercialización de la empresa distribuidora de energía, cuyas variables de análisis son estrictamente necesarias para la aplicación de uno o varios criterios de investigación de pérdidas, estos criterios, también denominados "criterio experto" responderán a la experiencia y conocimiento sobre la temática de control y mitigación de pérdidas que aplique el personal experto de la Empresa. Ejemplos de "criterio experto" pueden ser: selección por ubicación geográfica, tarifaria, por estrato de consumo, o características del sistema de medición, también por análisis de variación de consumo como reducción, incremento o variación atípica, etc. Como se puede inferir, existe una amplia variedad y combinaciones de criterios que él experto con su experiencia y características puede aportar al sistema bajo análisis.

2. **Fase 2: Pre-procesamiento de datos**

   Paso 2: Eliminación de datos inconsistentes y registros erróneos

   En este paso, a través de técnicas de filtrado y selección de registros, se identificarán y eliminarán de la base de datos aquellos registros de sistemas de medición que presenten valores en blanco, inconsistencias o errores de la información, ejemplo de este pre-procesamiento es la exclusión de la base de datos los registros de consumo negativo pues estos datos corresponden a re-facturaciones de consumo en devolución a favor del cliente situación que puede presentarse debido a errores de lectura y/o de aplicación tarifaria, esta situación introduce "ruido" al análisis posterior pues corresponden a casos muy específicos y especiales en los cuales la operatividad de análisis no arrojará resultados válidos.

3. **Fase 3: Aplicación de técnicas de minería de datos**

   En esta fase se procede con la aplicación de una técnica de agrupamiento de datos, en el caso particular del presente estudio se han definido dos posibles técnicas a aplicar: Coeficiente de Pearson y Agrupamiento K-Medias. Una vez seleccionada técnica se procede con su aplicación.

Paso 3: Aplicación de la técnica "Coeficiente de Pearson"

Como se indicó anteriormente en el capítulo 2, esta técnica agrupa únicamente casos con características específicas, para el presente estudio serán aquellos sistemas de medición que presentan los siguientes patrones de consumo:

- Caída drástica con posterior estabilización.
- Caída progresiva con posterior estabilización.

Al ser una técnica que agrupa casos específicos, se obtendrán listas pequeñas de sistemas de medición que registraron solamente dichos patrones de consumo en un período preestablecido de tiempo.

Paso 4: Aplicación de la técnica de "Agrupamiento K-Medias"

En caso de no haber aplicado la técnica anterior, es posible aplicar el agrupamiento K-Medias, esta es una opción más robusta pues agrupa los datos dependiendo de los patrones de entrada, sin embargo, es de suma importancia la correcta elección de los grupos a ser revisados, los cuales generalmente son aquellos que tienen pocos elementos y se encuentran alejados de los demás grupos. Con el objeto de aumentar la eficiencia de revisiones, las listas obtenidas por medio de esta técnica pasan por una nueva clasificación mediante cualquiera de las técnicas supervisadas las cuales podrían ser: K-Vecinos, árbol de decisión o Red Neuronal.

4. **Fase 4: Filtrado de resultados**

   Paso 5: Eliminación de sistemas de medición con revisiones previas en un lapso de 12 meses previos

   Una vez concluida la ejecución de cualquier técnica de minería de datos seleccionada, el listado resultante debe ser verificado con la finalidad de que las revisiones no incluyan casos que previamente hayan sido inspeccionados y/o revisados en campo, pues de lo contrario se tendría un impacto negativo sobre la eficiencia operativa y se revisarían casos en los cuales ya se determinó una novedad o el funcionamiento correcto del sistema de medición.

5. **Fase 5: Despacho de órdenes de trabajo a campo**

   Finalmente se procederá con una revisión muestral de los datos de los listados de sistemas de medición obtenidos en el paso anterior, esta revisión deberá ser realizada por el analista de pérdidas, con miras a determinar aspectos generales que podrían incidir negativamente en la eficiencia operativa de los grupos de trabajo, como por ejemplo: ubicación geográfica, sistema de medición, historial de consumo, etc. Una vez efectuada la revisión, se despacharán únicamente aquellas revisiones que tengan mayor probabilidad de encontrar novedades en el registro de consumo y que posibiliten un mayor grado de éxito.

La Figura 3.2 presenta la metodología explicada.

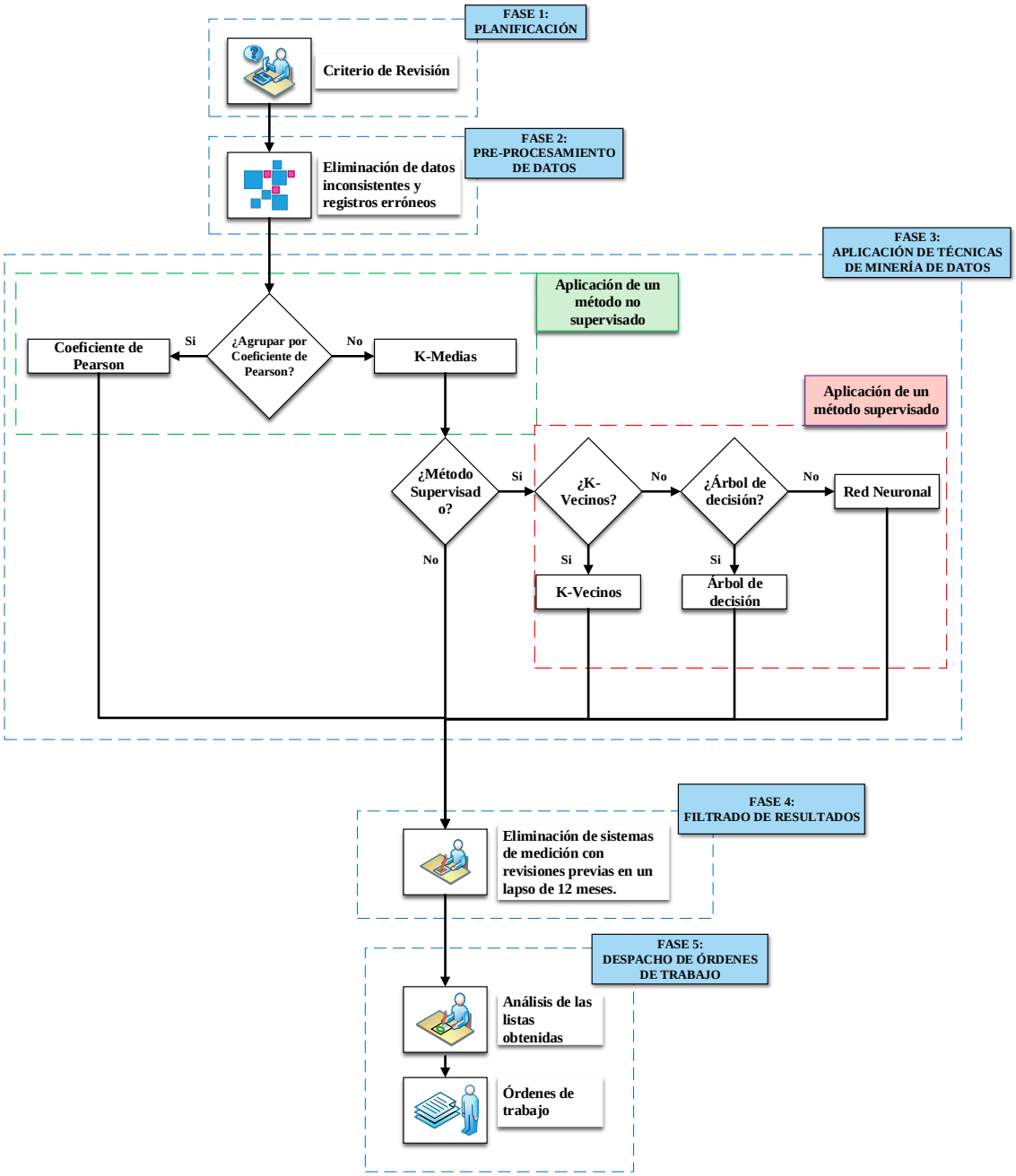


Figura 3.2, Metodología para el control de pérdidas no técnicas

3.3.RESULTADOS

Para la obtención de los listados de sistemas de medición que deben ser inspeccionados en campo, se generaron aleatoriamente siete listas aplicando la metodología planteada previamente (Figura 3.2), con diferentes criterios de revisión o "criterios experto".

En la Tabla 3.5, se enumeran los criterios considerados para la inspección y la técnica de minería de datos utilizada para el agrupamiento y posterior clasificación.

Tabla 3.5, Criterios de revisión y técnica aplicada para minería de datos

# Lista	Criterio de revisión	Técnica de Agrupamiento y Clasificación
L1	• Provincia: Azuay. • Cantón: Cuenca. • Parroquias: Checa, Chiquintad y Octavio Cordero.	K-Medias + Red Neuronal.
L2	• Provincia: Azuay. • Cantón: Cuenca. • Parroquias: San Sebastián y El Valle. • Grupo de consumo: Residencial y Comercial.	K-Medias + Red Neuronal.
L3	• Provincia: Azuay. • Cantón: Cuenca. • Grupo de consumo: Industrial.	K-Medias + Red Neuronal.
L4	• Provincia: Azuay. • Cantón: Cuenca. • Tipo de medición: monofásico 1F3H indirecto, Trifásico 3F3H indirecto, Trifásico 3F4H indirecto	K-Medias + Red Neuronal.
L5	• Provincia: Azuay. • Cantón: Cuenca. • Tarifa: Media Tensión. • Demanda Cero.	K-Medias + Red Neuronal.
L6	• Provincia: Azuay. • Cantón: Cuenca. • Parroquias: Monay, Sayausí y Totoracocha.	K-Medias + Red Neuronal.
L7	• Provincia: Azuay. • Cantón: Cuenca. • Meses adeudados: >3 meses. • Deuda: >60 USD.	K-Medias + Red Neuronal.

Se aplica el criterio para la obtención de L1 (listado 1) y el resultado fue el siguiente:

Al iniciar con el proceso se tiene la base de datos integrada por todos los sistemas de medición de la empresa distribuidora, que son en total 393.960.

Aplicando los criterios de revisión (ver Tabla 3.5), el caso de L1, se reduce a sistemas de medición ubicados en la provincia del Azuay, dentro del cantón Cuenca y específicamente en las parroquias Checa, Chiquintad y Octavio Cordero.

Luego, se obtiene una lista de 5.615 sistemas de medición. En ella se encontraron 266 casos con presencia de datos inexistentes (NaN) y 86 casos de datos erróneos; dando como resultado una lista de 5.263, a la que se denomina "matriz base".

Con la "matriz base" se ejecuta el algoritmo de agrupamiento K-Medias; este punto es crítico, pues, de la selección de los grupos dependerá las revisiones en sitio. Para L1 el resultado del agrupamiento K-Medias se presenta en la Figura 3.3.

Los grupos elegidos para este caso fueron:

- Residenciales: Grupo 5.
- Comerciales: Grupos 2 y 4.
- Industriales: Grupos 2 y 3.

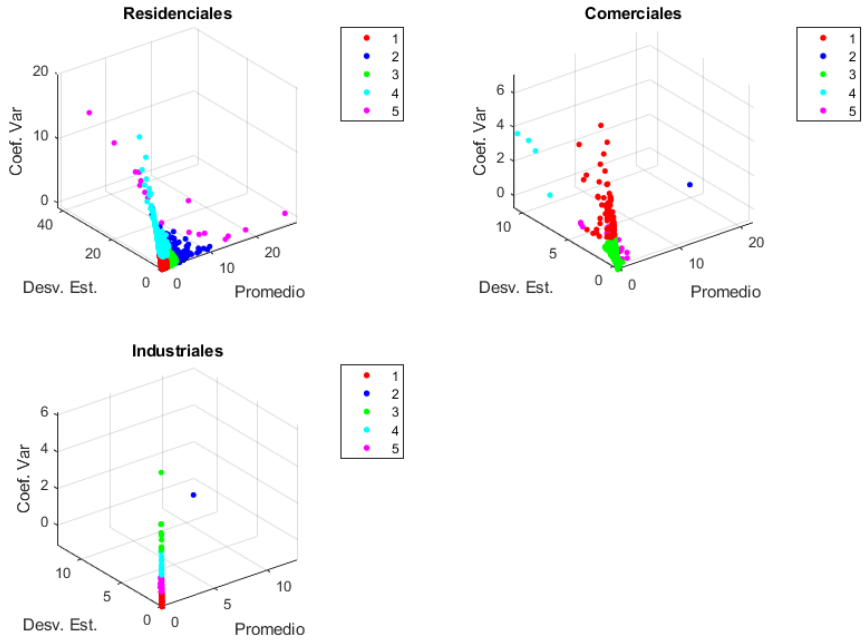


Figura 3.3, Resultado del agrupamiento K-Medias. Parroquias: Checa, Chiquintad y Octavio Cordero.

De los grupos seleccionados, se obtuvieron 39 sistemas de medición. Estos a su vez son clasificados mediante una red neuronal, determinando que 32 de ellos se clasifican para una posible revisión.

De la lista de 32 sistemas de medición, se eliminan aquellos que ya fueron revisados en el período de un año, en el presente caso, se encontró que 6 de ellos ya fueron revisados, reduciendo la cantidad a 26 sistemas de medición.

Finalmente, es revisado el patrón de consumo de los 26 sistemas de medición y se determina que todos los sistemas de medición resultado del algoritmo deben ser inspeccionados en sitio.

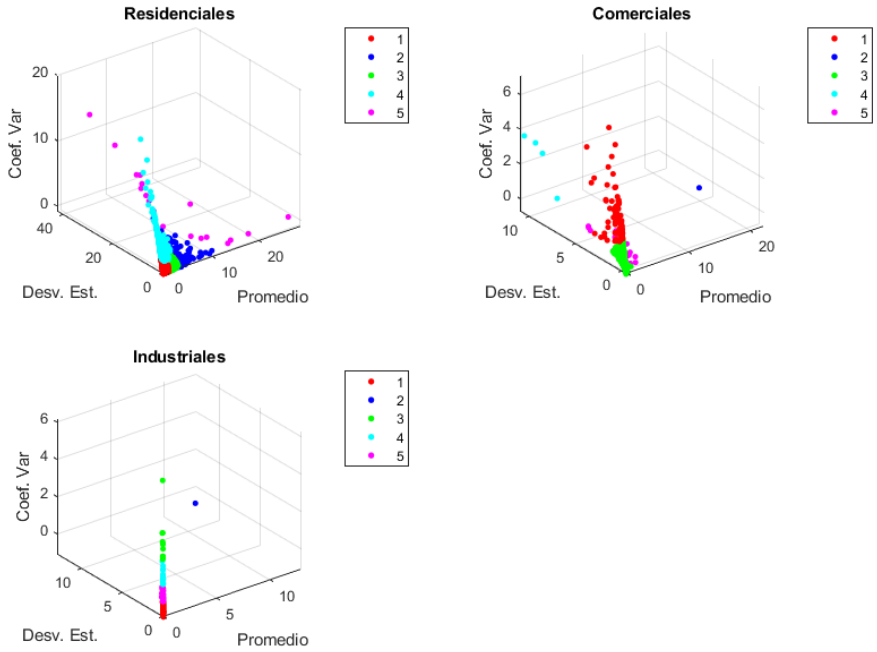
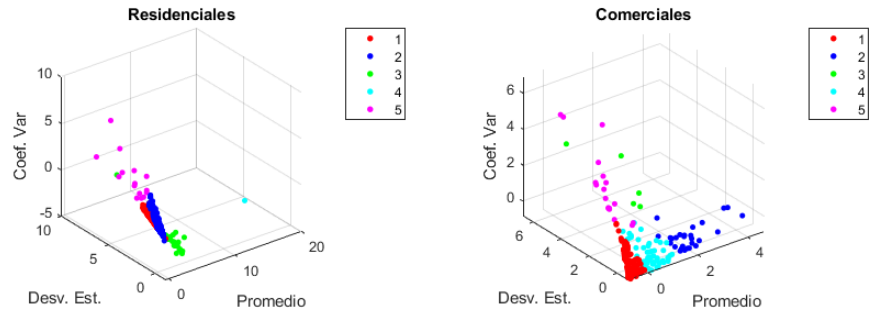
En la siguiente tabla, se presenta un resumen de los resultados de aplicar la metodología.

Tabla 3.6, Resultados de la minería de datos

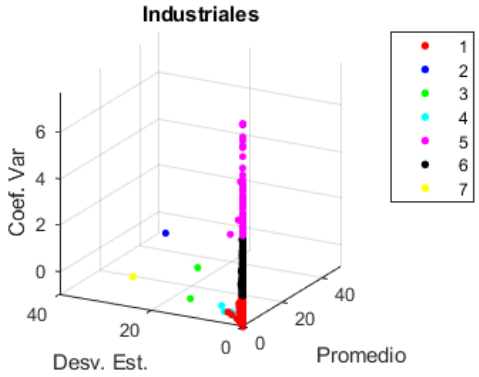
#	Total, Sist. Medi.	CRITERIO	Pre-Procesamiento		Matriz Base	K-Medias	Red Neuronal	Revisados últimos 12 meses	Supresión Por Análisis	Revisión
			NaN	Datos Erróneos.						
L1	393960	5.615	266	86	5263	39	32	6	0	26
L2	393960	24533	1615	2080	20802	37	32	9	0	28
L3	393960	3685	130	254	3301	63	58	11	24	23
L4	393960	103	0	2	101	22	5	0	0	5
L5	393960	3218	166	127	2925	28	28	12	0	16
L6	393960	21453	952	354	20147	128	125	27	10	88
L7	393960	1809	118	81	1610	46	46	22	5	19
								Total Revisiones		**205**

La Tabla 3.7 muestra la gráfica de los grupos K-Medias y cuáles de ellos fueron seleccionados para cada lista.

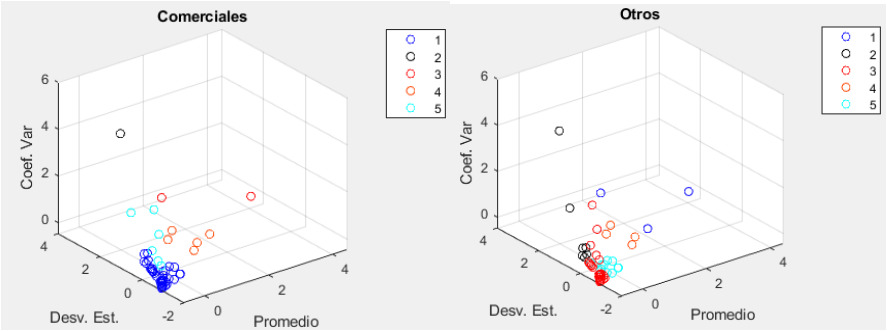
Tabla 3.7, K-Medias y Selección de grupos

Lista	Grupos Seleccionados
L1	Residenciales: Grupo 5. Comerciales: Grupos 2 y 4 Industriales: Grupos 2 y 3
	
L2	Residenciales: Grupos 4 y 5 Comerciales: Grupos 3 y 5
	

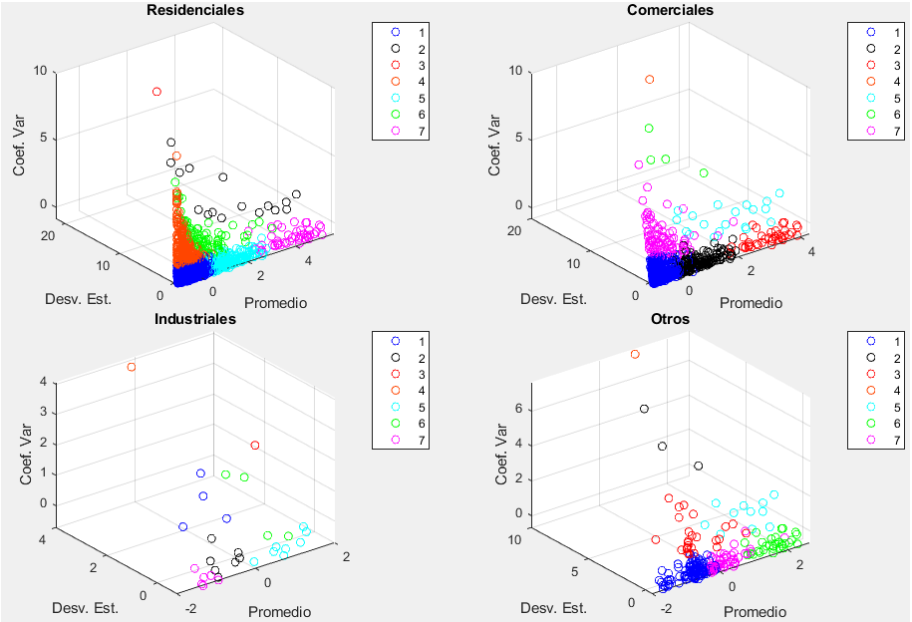
L3	Industriales: Grupos 2, 3, 5 y 7



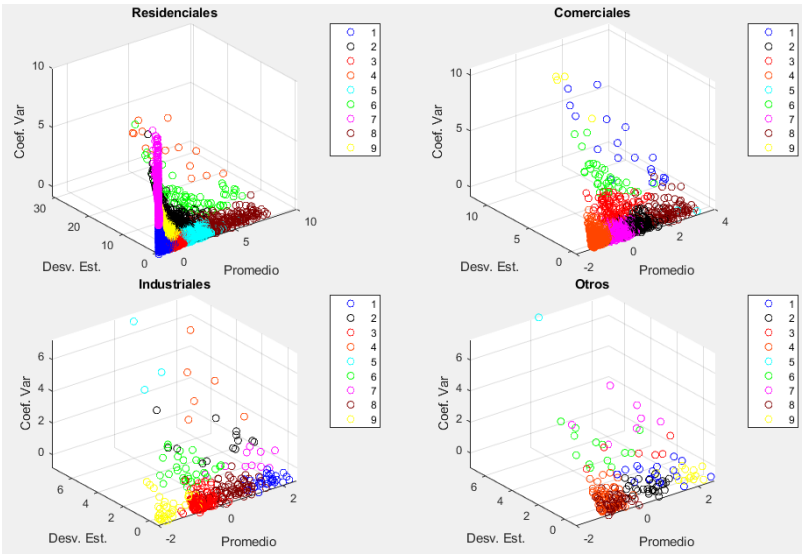
L4	Comercial: Grupos 2, 3 y 4 Otros: Grupos 1, 2 y 4



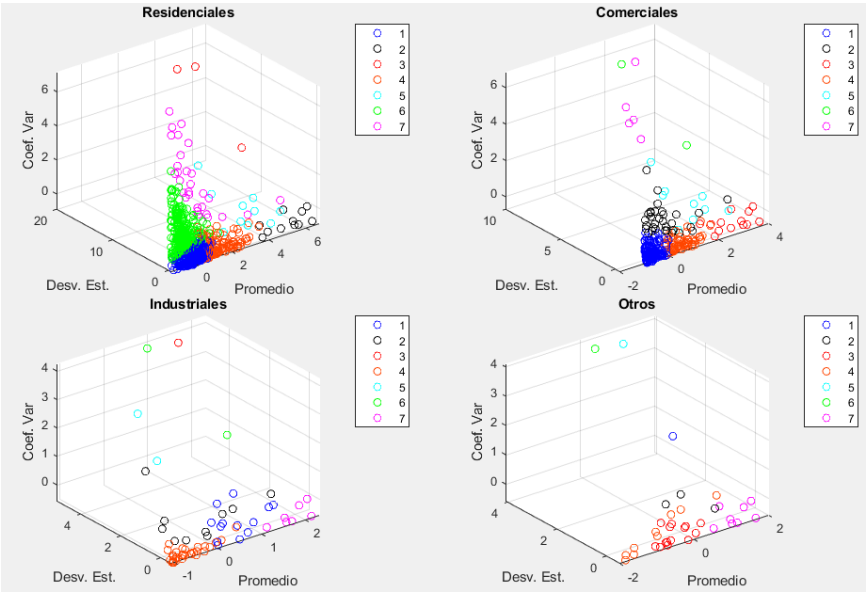
L5	Residenciales: Grupos 2 y 3. Comerciales: Grupos 4 y 6. Industriales: Grupos 3 y 4. Otros: Grupos 2 y 4.



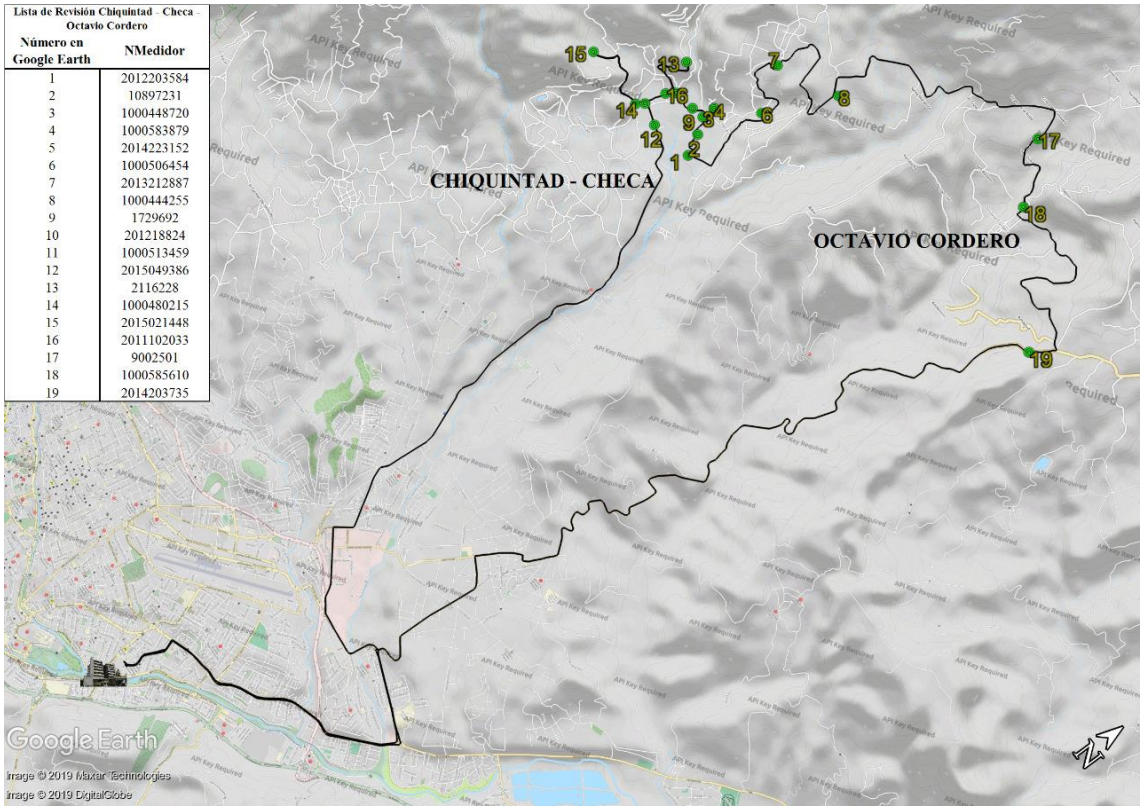
L6	Residenciales: Grupos 4 y 6. Comerciales: Grupos 1 y 9. Industriales: Grupos 4 y 5. Otros: Grupos 5 y 7.



L7	Residenciales: Grupos 3, 6 y 7 Comerciales: Grupos 6 y 7. Industriales: Grupos 3, 5 y 6 Otros: Grupos 5 y 6.

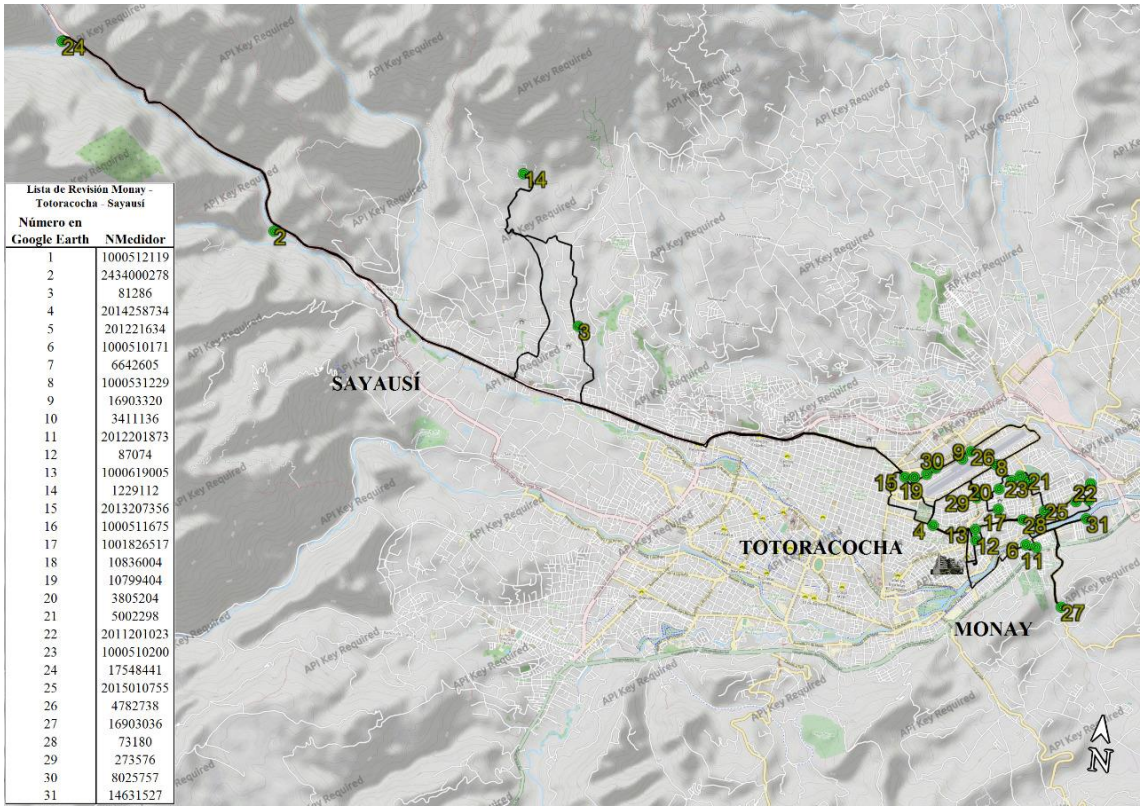


Previo a cumplir con las inspecciones en campo de los listados obtenidos, se diseñó una ruta para optimizar el tiempo de las revisiones, maximizando la eficiencia en el traslado operativo de los grupos encargados de la revisión de los sistemas de medición. Por ejemplo, en la Figura 3.4 se muestra la ruta planificada para las revisiones del listado 1 (L1) y en la Figura 3.5 para el listado 6 (L6). Las rutas planificadas para los demás listados se pueden observar en ANEXO A6.



Lista de Revisión Chiquintad - Checa - Octavio Cordero	
Número en Google Earth	NMedidor
1	2012203584
2	10897231
3	1000448720
4	1000583879
5	2014223152
6	1000506454
7	2013212887
8	1000444255
9	1729692
10	201218824
11	1000513459
12	2015049386
13	2116228
14	1000480215
15	2015021448
16	2011102033
17	9002501
18	1000585610
19	2014203735

Figura 3.4, Ruta de revisión para L1
Fuente: Google Earth



Lista de Revisión Monay - Totoracocha - Sayausí	
Número en Google Earth	NMedidor
1	1000512119
2	2434000278
3	81286
4	2014258734
5	201221634
6	1000510171
7	6642605
8	1000531229
9	16903320
10	3411136
11	2012201873
12	87074
13	1000619005
14	1229112
15	2013207356
16	1000511675
17	1001826517
18	10836004
19	10799404
20	3805204
21	5002298
22	2011201023
23	1000510200
24	17548441
25	2015010755
26	4782738
27	16903036
28	73180
29	273576
30	8025757
31	14631527

Figura 3.5, Ruta de revisión para L6
Fuente: Google Earth

3.4.ANÁLISIS Y EVALUACIÓN DE RESULTADOS

Se realizaron 91 revisiones en sitio, de un total de 205 sistemas de medición obtenidos con la aplicación de la metodología, esto en razón de aspectos operativos y/o administrativos del Departamento de Control de la Medición, que limitaron el cumplimiento del 100% de casos. Entre las principales novedades detectadas en las revisiones, están:

- Tarifas mal aplicadas;
- Fugas eléctricas;
- Un medidor oxidado por ingreso de agua en el tablero;
- Medidores dañados;
- Un reflector conectado directamente a la acometida de la empresa;
- Una conexión directa.

Todas estas novedades representaron la ejecución posterior de actividades de mejora a las condiciones técnicas y comerciales de prestación del servicio, pero, no en todos los casos representaron una recuperación de energía perdida para la Empresa; sin embargo, uno de los objetivos institucionales del departamento de Control de la Medición es el de permitir que se mantenga el correcto funcionamiento de los sistemas de medición en toda el área de concesión de la E. E. CENTROSUR.

Tabla 3.8, Análisis Técnico-Económico

N°	N° Revisiones	N° Realizadas	Costo Revisiones [USD]	Revisiones con Novedad	Revisiones - Recuperación	Efectividad [%]	Recuperación [kWh]	Recuperación [USD]
L1	26	19	594,13	2	0	10,53%	0	0
L2	28	23	719,21	4	1	17,39%	1169,58	322,10
L3	23	18	562,86	2	0	11,11%	0	0
L4	5	0	0,00	0	0	0,00%	0	0
L5	16	0	0,00	0	0	0,00%	0	0
L6	88	31	1.125,72	7	0	22,58%	0	0
L7	19	0	0,00	0	0	0,00%	0	0
Total	**205**	**91**	**3.001,92**	**15**	**1**	**16,48%**	**1169,58**	**322,10**

El costo por cada revisión representó para la empresa CENTROSUR 31,27 USD, dando un total por las revisiones realizadas de 3.001,92 USD. Por cuanto se encontró una conexión directa, esto representó una recuperación de energía de 1169,58 kWh y económica de 322,10 USD.

Se calcula la efectividad de la metodología, dividiendo el número de revisiones con novedad para el número total de revisiones realizadas. (Ecuación (9)

Lo que nos permite establecer que la efectividad de la metodología es de 16.48%.

$$efectividad\ [\%] = \frac{\#\ de\ revisiones\ con\ novedad}{\#\ total\ de\ revisiones} x100\% \qquad (9)$$

La E. E. CENTROSUR, tiene uno de los porcentajes más bajos de perdida de energía no técnica en el país (1,27% en el año 2018). Siendo importante destacar que la Empresa cuenta con personal calificado, con un excelente criterio de honestidad y responsabilidad, apoyado con el uso de tecnología de avanzada.

La posibilidad siempre latente de que se den casos de fraude y/o daños en los equipos de medición, crea la necesidad de que se mantengan controles ante todo de carácter disuasivo.

Esto evidencia plenamente el que la Empresa, mantenga una metodología orientada a minimizar el fraude o casos de daño en los equipos de medición. Metodología que no debe escatimarse bajo pretexto de que su costo no se justifica.

3.5.PROPUESTA DE PLANIFICACIÓN

La integración de una metodología y que ésta sea apropiada, ayudará para que la empresa distribuidora y comercializadora de energía, haga el control de las pérdidas no técnicas de manera eficaz, a través de determinar y planificar el proceso de revisiones, buscando que se mantenga en el mínimo ese porcentaje de pérdidas, y así la eficacia en su control.

Entonces, esa metodología se proyecta con los siguientes objetivos:

3.5.1. Objetivos

- Optimizar el tiempo de trabajo. - Realizar las revisiones en campo en el menor tiempo posible y así lograr un aumento de revisiones diarias.
- Optimizar grupos de trabajo. - Cumplir con todas las revisiones planificadas con el menor número de grupos y personal de la empresa.
- Optimizar revisiones en sitio. -Que las revisiones planificadas sean efectivas, aumentando los verdaderos positivos y reduciendo los falsos negativos, con esto la reducción de gastos de recursos de la empresa.

Estos objetivos deberán reducir el costo medio unitario de las revisiones en campo.

Los objetivos planteados están sujetos a restricciones tanto operativas como técnicas, estas son:

3.5.2. Restricciones

- Grupos de revisión. – Como se indicó anteriormente, la CENTROSUR históricamente ha mantenido 3 grupos de trabajo, dos de los cuales revisan sistemas de medición masivos y uno revisa sistemas de medición especiales; adicionalmente los grupos atienden reclamos relacionados con los sistemas de medición de energía.
- Número de revisiones por día/semana/mes. - Se estima que el número de revisiones diarias sería de 12 para los grupos de sistemas de medición masivos y 8 para el grupo de sistemas de medición especiales; teniendo un total de 32 revisiones por día, 160 por semana y 640 por mes. Esto se presenta en la Tabla 3.9.

Tabla 3.9, Número de revisiones por grupo

		N° de Revisiones				
Grupo	**Eq. Medición**	**Día**	**Mes**	**3 Meses**	**6 Meses**	**Año**
Grupo1	Masivos	12	240	720	1440	2880
Grupo 2	Masivos	12	240	720	1440	2880
Grupo 3	Especiales	8	160	480	960	1920
	Total	**32**	**640**	**1920**	**3840**	**7680**

3.5.3. Plan de trabajo

Teniendo en consideración las restricciones que implica la ejecución de la metodología, se propone la planificación de revisiones, por tiempo, por ubicación geográfica y por categoría tarifaria.

3.5.3.1. Planificación por tiempo:

- A Corto Plazo: Período de un mes calendario.
- A Mediano Plazo: Tres meses calendario.
- A Largo plazo: Seis meses, hasta un año calendario.

En la planificación por plazo, dentro del tiempo propuesto se deberá revisar los sistemas de medición que se obtuvieron a través de la metodología propuesta; es decir, se plantea un criterio de revisión y se genera un listado de sistemas de medición que deben ser inspeccionados en campo. Por ejemplo, para el Corto Plazo, se consideró que se pueden realizar 640 inspecciones; lo que permitirá determinar que en caso de que no se esté obteniendo buenos resultados, se cambie el criterio de revisión aplicado y volver a generar nuevos listados con la metodología para nuevas revisiones en campo. En los de Mediano Plazo y Largo Plazo, entraran entonces aquellos que no se examinaron, porque se consideró cambiar el criterio de revisión, dependiendo por cierto de la efectividad de las inspecciones realizadas.

En los lapsos de tiempo se recomienda haber cumplido con todos los listados, intentando identificar cual o cuales criterios de revisión han sido los que dieron mejores resultados. Esto servirá para aplicarlos en futuros análisis en la búsqueda de fraude o daño en los sistemas de medición.

3.5.3.2. Planificación por zona geográfica

Permite optimizar el tiempo en las inspecciones, porque los sistemas de medición que deben ser revisados están cercanos entre sí. Se parte de realizar un análisis social en las zonas urbanas y rurales, para identificar qué sectores son las que más pérdidas de energía producen, permitiendo generar listas de sistemas de medición.

El propósito es todas las zonas identificadas y pertenecientes a la empresa distribuidora, deben ser revisadas por lo menos una vez cada dos años.

3.5.3.3. Planificación por categoría tarifaria

Se obtienen listados de sistemas de medición a través de la metodología propuesta, mediante un criterio de revisión por tarifas (residencial, comercial e industrial).

Al haber una gran cantidad de sistemas de medición residenciales y comerciales, la metodología, permitirá detectar las variaciones que deben ser inspeccionados, lo que se recomienda hacerlo con una periodicidad de seis meses, es decir, que cada seis meses se puede correr el algoritmo de la metodología y obtener los listados de sistemas de medición residenciales y comerciales.

No se debe perder de vista, que hay ciertos sistemas de medición que son comerciales o industriales "especiales" por su gran consumo mensual. Un daño en estos sistemas de medición puede generar importantes pérdidas para la empresa distribuidora, por lo que dichos sistemas deben ser inspeccionados constantemente. Sistemas de medición deben ser filtrados por la metodología por lo menos cada tres meses.

Es importante resaltar, que no se puede prescindir del "criterio del experto" para la planificación, el cual es utilizado para definir las reglas, para la optimización al momento de aplicar el algoritmo que se está utilizando en la metodología.

CONCLUSIONES

Las pérdidas no técnicas anualmente generan un millonario menoscabo económico, de ahí la necesidad de un control constante. Del estudio realizado con los datos de la empresa distribuidora, mediante técnicas de minería de datos, y la metodología propuesta, se puede concluir:

- La E. E. Regional CENTROSUR, está dentro de las empresas más eficientes a nivel país, en el control de pérdidas no técnicas; teniendo el 1,27%. Pese a que el porcentaje de pérdidas no técnicas es bajo, económicamente son representativas para la empresa, por esta razón, el control debe ser constante, con tendencia a evitar un potencial aumento.
- La "matriz base", está integrada por la información obtenida de los reportes de la base comercial de la Empresa Distribuidora CENTROSUR, cuyo corte fue hasta el 14 de marzo del 2019.
- La formación de la "matriz base" se realizó en Excel. El cruce de datos entre los diferentes reportes, se realizó en su mayoría mediante las variables "NMedidor" y "CuentaContrato", sin embargo, el proceso de obtener los reportes y realizar los cruces requiere un tiempo considerable. En un futuro se deberá desarrollar los reportes con las variables presentadas en este estudio, permitiendo optimizar tiempos.
- La limpieza de datos inexistentes (NaN) y datos erróneos se realiza en MATLAB® por su facilidad, rapidez en la lectura y computo de los datos. Pese a estas ventajas, existen softwares especializados en el manejo de datos como WEKA, STATGRAPH y SPSS.
- Los datos con los que se trabaja deben aportar información fiable para conseguir un óptimo funcionamiento de los algoritmos (o técnicas de minería de datos). Por esta razón, previo la ejecución de los algoritmos, los datos tienen que estar "limpios" y normalizados.
- En las bases de datos se encontraron una elevada cantidad de datos inexistentes y datos erróneos. En todas las listas ejecutadas se experimentó este tipo de problema, que condujo a que se eliminen una considerable cantidad de sistemas de medición, que podrían ser causantes de pérdidas no técnicas.
- De las técnicas de minería de datos: Supervisadas y las no supervisadas, que se aplicaron para el control de pérdidas no técnicas, con el estudio realizado a los datos de la empresa distribuidora, se definió ejecutar varias de esas técnicas en MATLAB®, entre ellas: La detección de patrones de consumo mediante el coeficiente de Pearson; agrupamiento en equipos de medición con K-Medias; y, clasificación de datos con red neuronal, árbol de decisión y vecinos más cercanos.

- Se determina las siguientes ventajas y desventajas de las técnicas:

Técnica	Ventajas	Desventajas
Detección de patrones de consumo mediante el coeficiente de Pearson.	▪ No necesita de ejemplos para la agrupación de datos. ▪ Generación de listas potenciales para revisión en sitio.	▪ Detecta solamente casos específicos, eludiendo gran parte de los posibles causantes de pérdidas no técnicas.
Agrupamiento K-Medias	▪ No necesita de ejemplos para la agrupación de datos. ▪ Generación de listas potenciales para revisión en sitio.	▪ No conocer el número adecuado de grupos a desarrollarse. ▪ Tener un conocimiento considerable para la correcta elección de los grupos que deben ser inspeccionados en sitio.
Red Neuronal	▪ El proceso de entrenamiento, validación y prueba de la red neuronal en MATLAB no requiere de mucho tiempo. ▪ Generación de listas potenciales para revisión en sitio.	▪ La dependencia de una base de datos con ejemplos comprobados de fraude y no fraude. ▪ El parámetro del número de neuronas en la capa oculta puede necesitar ajuste.
Árbol de decisión	▪ Generación de listas potenciales para revisión en sitio.	▪ La dependencia de una base de datos con ejemplos comprobados de fraude y no fraude. ▪ La necesidad de trabajar en conjunto con otra técnica de minería de datos para un mejor rendimiento.
K-Vecinos	▪ Generación de listas potenciales para revisión en sitio. ▪ Simplicidad en la ejecución del algoritmo.	▪ La dependencia de una base de datos con ejemplos comprobados de fraude y no fraude. ▪ La necesidad de saber el número de "K" vecinos cercanos a considerar.

- El algoritmo de detección de patrones de consumo mediante el coeficiente de Pearson, demuestra ser un excelente agrupador de sistemas de medición, que permite detectar casos específicos de patrones de consumo; pero, no todos los casos detectados con este algoritmo son causas de pérdidas no técnicas, esto se debe a varias razones, siendo una de ellas que las ventanas de tiempo para la ejecución del mismo fueron muy cortas; los creadores de este algoritmo recomiendan que para el análisis es necesario contar con el consumo de mínimo dos años y en este caso solamente se utilizaron ventanas de tiempo de un año. Otra razón es que una caída drástica en el consumo del cliente no siempre se debe a que el medidor se haya dañado, o haya hurto de energía, la mayoría de las veces se origina en reducciones reales de consumo por inmuebles deshabitados o desocupados.
- De las técnicas ejecutadas, el agrupamiento K-Medias resultó ser el método que más se adecúa a los datos de la empresa distribuidora. Al ser una técnica no supervisada, no requirió una base de datos con ejemplos comprobados de fraude y no fraude, para el entrenamiento. Hay que ser minucioso al momento de elegir los grupos de equipos de medición a ser revisados, pues una mala elección del grupo provocará que se revisen equipos que no deben ser inspeccionados.
- Al no contar con suficientes ejemplos para el entrenamiento de las técnicas supervisadas, estas no rindieron de manera óptima; sin embargo, al juntar una técnica no supervisada (K-Medias) con una supervisada (Red neuronal, árbol de decisión o K-Vecinos), se permite eliminar clientes obtenidos en la técnica no supervisada, que por algún motivo entraron en el

grupo de los que deben ser revisados, obteniendo una mejora de rendimiento al momento de obtener las listas para revisión.

- Se realizaron 91 inspecciones en sitio, de ellas 15 se encontraron con alguna novedad, obteniendo una efectividad de la metodología propuesta de 16,48%. En la realidad es difícil realizar un análisis de efectividad, pues, para que una metodología alcance porcentajes altos de certeza, implicaría que el sistema de distribución tenga porcentajes altos de pérdidas no técnicas.
- El control de pérdidas no técnicas, no contempla un análisis de costo beneficio; es decir, no se puede medir el rendimiento de la metodología por la cantidad de ingreso recuperado. Queda claro que el departamento del Control de la Medición, no se encarga de generar ingresos, su función es la de controlar, mantener y mitigar las pérdidas de energía como medida constante dentro del balance energético de la CENTROSUR, orientada precisamente al mínimo posible de pérdidas no técnicas.
- El beneficio que obtiene la empresa distribuidora con éste estudio, está en la aplicación de una metodología esquematizada, que permite al funcionario que esté a cargo del Departamento del Control de la Medición, poder aplicarlo.

RECOMENDACIONES

- Analizar y corregir los reportes de datos, puesto que al realizar el estudio de los mismos se detectó redundancia en la información, debido a variables con diferente nombre, pero con el mismo contenido, además de esto se encontraron numerosos datos inexistentes y datos erróneos.
- Ampliar a mínimo dos años un reporte de datos de consumo y de demanda, aplicados a la matriz base propuesta.
- Realizar un alistamiento de los sistemas de medición en los que se detectaron hurto de energía o errores en los equipos de medición, con el informe del mes en que se realiza la re-facturación. Información que permitirá en futuros análisis contar con una base sólida de ejemplos para la ejecución de los métodos supervisados.
- Cuando se cuente con una base sólida con perfiles de clientes que fueron comprobados con fraude, realizar un reajuste en el entrenamiento de la red neuronal ya que éste es de los mejores métodos informáticos para la clasificación de datos.
- Existen otras técnicas de minería de datos como: las reglas de inducción; máquina de soporte vectorial (SVM) o mapa de auto organización (SOM) para la clasificación y el agrupamiento. Son técnicas altamente usadas en la literatura; no se aplicaron por cuestiones de tiempo, confiando que en futuros estudios se pueda aplicar estas técnicas y comprobar su eficacia con las que ya fueron aplicadas en este proyecto de grado.

REFERENCIAS

[1] W. G. on L. R. CIRED, "Reduction of Technical and Non-Technical Losses in Distribution Networks," *Int. Conf. Electr. Distrib.*, p. 114, 2017.

[2] J. L. Viegas, P. R. Esteves, R. Melício, V. M. F. Mendes, and S. M. Vieira, "Solutions for detection of non-technical losses in the electricity grid : A review," *Renew. Sustain. Energy Rev.*, vol. 80, no. August 2016, pp. 1256–1268, 2017.

[3] A. Tama, "Las pérdidas de energía eléctrica," *Criell*, pp. 12–17, 2014.

[4] G. M. Messinis and N. D. Hatziargyriou, "Review of non-technical loss detection methods," *Electr. Power Syst. Res.*, vol. 158, pp. 250–266, 2018.

[5] M. D. Monteiro and R. S. Maciel, "Detection of commercial losses in electric power distribution systems using data mining techniques," *SBSE 2018 - 7th Brazilian Electr. Syst. Symp.*, pp. 1–6, 2018.

[6] CENTROSUR, "ÁREA DE CONCESIÓN," 2019. [Online]. Available: http://www.centrosur.gob.ec/sites/default/files/AREA DE CONCESION CENTROSUR.pdf.

[7] E. Ramírez, "METODOLOGÍA PARA REDUCIR LAS PÉRDIDAS TÉCNICAS EN LAS REDES DE DISTRIBUCIÓN DE MEDIA TENSIÓN CON APLICACIÓN EN EL CIRCUITO INDUSTRIAL NORTE," Corporación Universitaria de la Costa. CUC, 2005.

[8] J. Mercado, R. Jiménez, and T. Serebrisky, "Dimensionando las pérdidas de electricidad en los sistemas de transmisión y distribución en América Latina y el Caribe," *Banco Interam. Desarro.*, p. 42, 2014.

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

[10] Asamblea Nacional de la República del Ecuador, "Ley Orgánica del Servicio Público de Energía Eléctrica (LOSPEE)," pp. 1–28, 2015.

[11] ARCONEL, "Regulación 'Modelo de contrato de suministro de energía eléctrica,'" 2017.

[12] ARCONEL, "Módulos de Estadística Anual y Multianual del Sector Eléctrico Ecuatoriano desde el año 2007 hasta 2018." [Online]. Available: https://www.regulacionelectrica.gob.ec/boletines-estadisticos/.

[13] MEER, "Plan Maestro de ELECTRICIDAD 2016-2025." p. 433, 2016.

[14] C. Papadimitriou, G. Messinis, D. Vranis, S. Politopoulou, and N. Hatziargyriou, "Non-technical losses: detection methods and regulatory aspects overview," *CIRED - Open Access Proc. J.*, vol. 2017, no. 1, pp. 2830–2832, 2017.

[15] G. M. Messinis and N. D. Hatziargyriou, "Unsupervised classification for non-technical loss detection," *20th Power Syst. Comput. Conf. PSCC 2018*, vol. 2018, pp. 1–7, 2018.

[16] S. A. S. B. Practices and K. Nevala, "The MACHINE LEARNING Primer," *SAS Inst. Inc.*, p. 52, 2017.

[17] J. Pulz, R. B. Muller, F. Romero, A. Meffe, Á. F. Garcez Neto, and A. S. Jesus, "Fraud detection in low-voltage electricity consumers using socio-economic indicators and billing profile in smart grids," *CIRED - Open Access Proc. J.*, vol. 2017, no. 1, pp. 2300–2303, 2017.

[18] J. Hernandez, M. J. Ramirez, and C. Ferri, *Introducción a la Minería de Datos*, PEARSON ED. Madrid, 2004.

[19] C. Pérez, *Minería de datos: técnicas y herramientas*, Reimpresa. 2007.

[20] O. Maimon and R. Lior, *Data Mining With Decision Trees*, 2nd Editio. 2014.

[21] P. Daniel, *Analisis de Datos Multivariantes*, no. January. Madrid, 2002.

[22] C. Cuadras, *Nuevos Métodos de Análisis Multivariante*, no. Septiembre. Barcelona, 2018.

[23] J. Gironés, J. Casas, J. Minguillón, and R. Caihuelas, *Minería de datos. Modelos y algoritmos*, First Edit. 2017.

[24] C. Guevara, "Reconocimiento de patrones para identificación de usuarios en accesos informáticos," Universidad Complutense de Madrid, 2012.

[25] I. Monedero, F. Biscarri, C. León, J. I. Guerrero, J. Biscarri, and R. Millán, "Detection of frauds and other non-technical losses in a power utility using Pearson coefficient, Bayesian networks and decision trees.," *Int. J. Electr. Power Energy Syst.*, vol. 34, pp. 90–98, 2012.

[26] P. Bholowalia and A. Kumar, "EBK-Means : A Clustering Technique based on Elbow Method and K-Means in WSN," *Int. J. Comput. Appl.*, vol. 105, no. 9, pp. 17–24, 2014.

[27] G. Berástegui, "Implementación del algoritmo de los k vecinos más cercanos (k-NN) y estimación del mejor valor local de k para su cálculo," Universidad Pública de Navarra, 2018.

[28] The MathWorks Inc., "MATLAB," 2019. [Online]. Available: https://es.mathworks.com/.

[29] R. Barros, G. E. S. A, and P. Dee, "Use of ANN for Identification of Consumers with Irregular Electrical Installations," *2018 Simp. Bras. Sist. Eletr.*, pp. 1–6, 2018.

A1. ANEXO 1 – MATRIZ DE CORRELACIONES

	Atributo	V27	V28	V29	V30	V31	V32	V33	V34	V35	V36	V37	V38	V39	V40	V41	V42	V43	V44	V45	V46	V47	V48	V49	V50	V51	V52	V53	V54	V55	V56	V57	V58	V59	V60	V61	V62	V63	V64	V65	V66	V67	V68
V27	TerceraEdad	1,0																																								
V28	Bdh	0,0	1,0																																							
V29	TipoConsumo	0,0	0,0	1,0																																						
V30	Tension	0,0	0,0	0,0	1,0																																					
V31	FabricanteMedidor	0,0	0,1	0,0	0,0	1,0																																			
V32	TipMedicion	0,0	0,1	0,0	0,1	-0,3	1,0																																		
V33	GrupoDeConsumo	0,0	0,0	0,7	0,2	0,0	0,0	1,0																																	
V34	Fases	0,0	0,1	0,0	0,1	-0,3	1,0	0,0	1,0																																
V35	MesesAdeudados	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	1,0																															
V36	Deuda	0,0	0,0	0,1	0,0	0,0	0,0	0,1	0,0	0,5	1,0																														
V37	PromedioFact_6Ult_Meses	0,0	0,0	0,1	0,3	0,0	0,1	0,2	0,1	0,0	0,5	1,0																													
V38	ValorUltimaFactura	0,0	0,0	0,1	0,4	0,0	0,1	0,2	0,1	0,0	0,6	1,0	1,0																												
V39	ConsumoKWhActual	0,1	0,0	0,1	0,4	0,0	0,1	0,2	0,1	0,0	0,4	0,9	0,9	1,0																											
V40	ConsumoKWh1MesAntes	0,1	0,0	0,0	0,4	0,0	0,1	0,1	0,1	0,0	0,4	0,8	0,8	0,9	1,0																										
V41	ConsumoKWh2MesesAntes	0,1	0,0	0,2	0,4	0,0	0,1	0,3	0,1	0,0	0,5	0,9	0,9	1,0	0,8	1,0																									
V42	ConsumoKWh3MesesAntes	0,1	0,0	0,2	0,2	0,0	0,1	0,3	0,1	0,0	0,5	0,9	0,9	0,9	0,7	0,9	1,0																								
V43	ConsumoKWh4MesesAntes	0,1	0,0	0,2	0,3	0,0	0,1	0,2	0,1	0,0	0,4	0,9	0,9	0,9	0,8	0,9	0,9	1,0																							
V44	ConsumoKWh5MesesAntes	0,1	0,0	0,2	0,3	0,0	0,1	0,2	0,1	0,0	0,4	0,9	0,9	0,9	0,8	0,9	0,9	1,0	1,0																						
V45	ConsumoKWh6MesesAntes	0,1	0,0	0,2	0,3	0,0	0,1	0,2	0,1	0,0	0,5	0,9	0,9	0,9	0,8	0,9	0,9	0,9	1,0	1,0																					
V46	ConsumoKWh7MesesAntes	0,1	0,0	0,2	0,3	0,0	0,1	0,2	0,1	0,0	0,4	0,9	0,9	0,9	0,8	0,9	0,9	0,9	1,0	1,0	1,0																				
V47	ConsumoKWh8MesesAntes	0,1	0,0	0,1	0,3	0,0	0,1	0,2	0,1	0,0	0,4	0,9	0,8	0,9	0,8	0,9	0,8	0,9	0,9	0,9	0,9	1,0																			
V48	ConsumoKWh9MesesAntes	0,0	0,0	0,1	0,3	0,0	0,1	0,2	0,1	0,0	0,4	0,8	0,8	0,8	0,7	0,8	0,8	0,8	0,8	0,8	0,9	0,9	1,0																		
V49	ConsumoKWh10MesesAntes	0,1	0,0	0,2	0,3	0,0	0,1	0,2	0,1	0,0	0,4	0,8	0,8	0,8	0,7	0,8	0,8	0,8	0,9	0,9	0,9	0,9	0,9	1,0																	
V50	ConsumoKWh11MesesAntes	0,1	0,0	0,2	0,3	0,0	0,1	0,2	0,1	0,0	0,4	0,9	0,9	0,9	0,8	0,9	0,9	0,9	0,9	1,0	1,0	0,9	0,9	0,9	1,0																
V51	ConsumoKWh12MesesAntes	0,1	0,0	0,2	0,3	0,0	0,1	0,2	0,1	0,0	0,4	0,9	0,9	0,9	0,7	0,9	0,9	0,9	0,9	0,9	0,9	0,9	0,8	0,9	1,0	1,0															
V52	ConsumoPro_	0,1	0,0	0,2	0,3	0,0	0,1	0,2	0,1	0,0	0,5	1,0	0,9	0,9	0,8	1,0	0,9	1,0	1,0	1,0	1,0	1,0	0,9	0,9	1,0	1,0	1,0														
V53	DemandaActualKW	0,0	0,0	0,0	0,8	0,0	0,1	0,2	0,1	0,0	0,0	0,5	0,5	0,6	0,6	0,5	0,3	0,5	0,5	0,4	0,5	0,4	0,5	0,4	0,5	0,4	0,5	1,0													
V54	DemandaKW1MesAntes	0,0	0,0	0,0	0,9	0,0	0,1	0,2	0,1	0,0	0,0	0,4	0,5	0,5	0,5	0,5	0,3	0,5	0,4	0,4	0,4	0,4	0,5	0,4	0,4	0,4	0,4	1,0	1,0												
V55	DemandaKW2MesesAntes	0,0	0,0	0,0	0,9	0,0	0,1	0,2	0,1	0,0	0,0	0,4	0,5	0,5	0,5	0,5	0,3	0,5	0,4	0,4	0,4	0,4	0,5	0,4	0,4	0,4	0,4	1,0	1,0	1,0											
V56	DemandaKW3MesesAntes	0,0	0,0	0,0	0,9	0,0	0,1	0,2	0,1	0,0	0,0	0,4	0,4	0,5	0,5	0,4	0,2	0,4	0,4	0,4	0,4	0,4	0,4	0,3	0,4	0,3	0,4	1,0	1,0	1,0	1,0										
V57	DemandaKW4MesesAntes	0,0	0,0	0,0	0,8	0,0	0,1	0,2	0,1	0,0	0,0	0,5	0,5	0,6	0,6	0,5	0,3	0,5	0,5	0,5	0,5	0,4	0,5	0,4	0,5	0,4	0,5	1,0	1,0	1,0	0,9	1,0									
V58	DemandaKW5MesesAntes	0,0	0,0	0,0	0,8	0,0	0,1	0,2	0,1	0,0	0,0	0,5	0,5	0,6	0,6	0,5	0,3	0,5	0,5	0,5	0,5	0,5	0,5	0,4	0,5	0,4	0,5	1,0	1,0	1,0	0,9	1,0	1,0								
V59	DemandaKW6MesesAntes	0,0	0,0	0,0	0,8	0,0	0,1	0,2	0,1	0,0	0,0	0,5	0,5	0,6	0,6	0,5	0,3	0,5	0,5	0,5	0,5	0,5	0,5	0,4	0,5	0,4	0,5	1,0	1,0	1,0	0,9	1,0	1,0	1,0							
V60	DemandaKW7MesesAntes	0,0	0,0	0,0	0,8	0,0	0,1	0,2	0,1	0,0	0,0	0,5	0,5	0,6	0,6	0,5	0,3	0,5	0,5	0,5	0,5	0,5	0,5	0,4	0,5	0,4	0,5	1,0	1,0	1,0	0,9	1,0	1,0	1,0	1,0						
V61	DemandaKW8MesesAntes	0,0	0,0	0,0	0,8	0,0	0,1	0,1	0,1	0,0	0,0	0,5	0,5	0,6	0,6	0,5	0,3	0,5	0,5	0,5	0,5	0,5	0,5	0,4	0,5	0,4	0,5	1,0	0,9	0,9	0,9	1,0	1,0	1,0	1,0	1,0					
V62	DemandaKW9MesesAntes	0,0	0,0	0,0	0,8	0,0	0,1	0,1	0,1	0,0	0,0	0,5	0,5	0,6	0,6	0,5	0,3	0,5	0,5	0,5	0,5	0,5	0,5	0,4	0,5	0,4	0,5	1,0	0,9	0,9	0,9	1,0	1,0	1,0	1,0	1,0	1,0				
V63	DemandaKW10MesesAntes	0,0	0,0	0,0	0,8	0,0	0,1	0,2	0,1	0,0	0,0	0,5	0,5	0,6	0,6	0,5	0,3	0,5	0,5	0,5	0,5	0,5	0,5	0,4	0,5	0,4	0,5	1,0	1,0	1,0	0,9	1,0	1,0	1,0	1,0	1,0	1,0	1,0			
V64	DemandaKW11MesesAntes	0,0	0,0	0,0	0,6	0,0	0,1	0,1	0,1	0,0	0,0	0,5	0,5	0,6	0,6	0,5	0,3	0,5	0,6	0,5	0,5	0,5	0,5	0,4	0,6	0,5	0,5	0,9	0,8	0,8	0,7	0,9	0,9	0,9	0,9	1,0	1,0	0,9	1,0		
V65	DemandaKW12MesesAntes	0,0	0,0	0,0	0,6	0,0	0,1	0,1	0,1	0,0	0,0	0,5	0,5	0,6	0,6	0,5	0,3	0,5	0,6	0,5	0,5	0,5	0,6	0,4	0,6	0,5	0,5	0,9	0,8	0,8	0,7	0,9	0,9	0,9	0,9	1,0	1,0	0,9	1,0	1,0	
V66	PecPorCliente	0,1	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,1	0,1	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,1	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	1,0	
V67	EstratoGeografico	-0,1	0,0	-0,1	0,0	-0,1	-0,1	-0,1	-0,1	0,1	-0,2	-0,4	-0,4	-0,4	-0,5	-0,5	-0,5	-0,5	-0,4	-0,5	-0,4	-0,4	-0,4	-0,4	-0,4	-0,4	-0,5	-0,1	-0,1	-0,1	-0,1	-0,1	-0,1	-0,1	-0,1	-0,1	-0,1	-0,1	0,0	0,0	0,0	1,0	
V68	A_oFabricacion	0,0	0,0	0,0	-0,2	0,0	0,0	0,0	0,1	0,0	0,0	-0,1	-0,2	-0,2	-0,2	-0,1	-0,1	-0,2	-0,2	-0,1	-0,1	-0,1	-0,2	-0,1	-0,2	-0,1	-0,1	-0,3	-0,3	-0,3	-0,2	-0,3	-0,3	-0,3	-0,3	-0,3	-0,3	-0,3	-0,3	-0,3	0,0	0,0	1,0

A2. ANEXO 2 - CÓDIGO DESARROLLADO EN MATLAB®

A2.1. Limpieza de datos NaN

```
%% ----------------------LIMPIEZA DE DATOS NaN-----------------------------
clear all; close all; clc %Limpieza de pantalla y comandos
%% Lectura de datos
datos=readtable('Matriz Base.xlsx');
%% Guardado de variables - Variables "técnicas"
nombreVariables=datos.Properties.VariableNames; %Nombre de las variables
TerceraEdad=datos.TerceraEdad;[f1]=find(isnan(TerceraEdad));TerceraEdad(f1)=0;%1
Bdh=datos.Bdh;[f2]=find(isnan(Bdh));Bdh(f2)=0;%2
TipoConsumo=datos.TipoConsumo;%3
Tension=datos.Tension;%4
FM=datos.FabricanteMedidor;%5
TipoMedicion=datos.TipMedicion;%6
GrupoConsumo=datos.GrupoDeConsumo;%7
Fases=datos.Fases; %8
MesesAdeudados=datos.MesesAdeudados;%9
Deuda=datos.Deuda;%10
PromedioFact_6Ult_Meses=datos.PromedioFact_6Ult_Meses;%11
ValorUltimaFactura=datos.ValorUltimaFactura;%12
ConsumoKWhActual=datos.ConsumoKWhActual;%13
ConsumoKWh1MesAntes=datos.ConsumoKWh1MesAntes;%14
ConsumoKWh2MesesAntes=datos.ConsumoKWh2MesesAntes;%15
ConsumoKWh3MesesAntes=datos.ConsumoKWh3MesesAntes;%16
ConsumoKWh4MesesAntes=datos.ConsumoKWh4MesesAntes;%17
ConsumoKWh5MesesAntes=datos.ConsumoKWh5MesesAntes;%18
ConsumoKWh6MesesAntes=datos.ConsumoKWh6MesesAntes;%19
ConsumoKWh7MesesAntes=datos.ConsumoKWh7MesesAntes;%20
ConsumoKWh8MesesAntes=datos.ConsumoKWh8MesesAntes;%21
ConsumoKWh9MesesAntes=datos.ConsumoKWh9MesesAntes;%22
ConsumoKWh10MesesAntes=datos.ConsumoKWh10MesesAntes;%23
ConsumoKWh11MesesAntes=datos.ConsumoKWh11MesesAntes;%24
ConsumoKWh12MesesAntes=datos.ConsumoKWh12MesesAntes;%25
ConsumoPro=datos.ConsumoPro_;%26
DemandaActualKW=datos.DemandaActualKW;%27
DemandaKW1MesAntes=datos.DemandaKW1MesAntes;%28
DemandaKW2MesesAntes=datos.DemandaKW2MesesAntes;%29
DemandaKW3MesesAntes=datos.DemandaKW3MesesAntes;%30
DemandaKW4MesesAntes=datos.DemandaKW4MesesAntes;%31
DemandaKW5MesesAntes=datos.DemandaKW5MesesAntes;%32
DemandaKW6MesesAntes=datos.DemandaKW6MesesAntes;%33
DemandaKW7MesesAntes=datos.DemandaKW7MesesAntes;%34
DemandaKW8MesesAntes=datos.DemandaKW8MesesAntes;%35
DemandaKW9MesesAntes=datos.DemandaKW9MesesAntes;%36
DemandaKW10MesesAntes=datos.DemandaKW10MesesAntes;%37
DemandaKW11MesesAntes=datos.DemandaKW11MesesAntes;%38
DemandaKW12MesesAntes=datos.DemandaKW12MesesAntes;%39
PecCli=datos.PecPorCliente;[f3]=find(isnan(PecCli));PecCli(f3)=0;%40
Estrato=datos.EstratoGeografico;%41
AFabricacion=datos.A_oFabricacion;[f4]=find(isnan(AFabricacion));AFabricacion(f2)=0;
%% Guardado en una matriz general
matrizBase=[TerceraEdad,Bdh,TipoConsumo,Tension,...
            FM,TipoMedicion,GrupoConsumo,Fases,...
            MesesAdeudados,Deuda,PromedioFact_6Ult_Meses,...
            ValorUltimaFactura,ConsumoKWhActual,...
            ConsumoKWh1MesAntes,ConsumoKWh2MesesAntes,...
            ConsumoKWh3MesesAntes,ConsumoKWh4MesesAntes,...
            ConsumoKWh5MesesAntes,ConsumoKWh6MesesAntes,...
            ConsumoKWh7MesesAntes,ConsumoKWh8MesesAntes,...
            ConsumoKWh9MesesAntes,ConsumoKWh10MesesAntes,...
            ConsumoKWh11MesesAntes,ConsumoKWh12MesesAntes,...
            ConsumoPro,DemandaActualKW,DemandaKW1MesAntes,...
            DemandaKW2MesesAntes,DemandaKW3MesesAntes,...
            DemandaKW4MesesAntes,DemandaKW5MesesAntes,...
            DemandaKW6MesesAntes,DemandaKW7MesesAntes,...
            DemandaKW8MesesAntes,DemandaKW9MesesAntes,...
            DemandaKW10MesesAntes,DemandaKW11MesesAntes,...
            DemandaKW12MesesAntes,PecCli,Estrato,AFabricacion];
%% Limpieza de datos NaN
[filNaN colNaN]=find(isnan(matrizBase));%Busca datos NaN
filNaN=unique(sort(filNaN)); %Filas en donde existe datos NaN
% Eliminación de datos Nulos
datos(filNaN,:)=[ ]; %Elimina los clientes con NaN de la matriz general
writetable(datos,'Rep_MatrizDatosSinNaN.xlsx','Sheet',1);
```

A2.2. Limpieza de datos Atípicos

```
%% --------------------LIMPIEZA DE DATOS ATÍPICOS-------------------------
clear all; close all; clc%Limpieza de pantalla y comandos
datos=readtable('Rep_MatrizDatosSinNaN.xlsx'); %Lectura de datos
%% Guardado de variables - Variables "técnicas"
nombreVariables=datos.Properties.VariableNames; %Nombre de las variables
TerceraEdad=datos.TerceraEdad;%1
Bdh=datos.Bdh;%2
TipoConsumo=datos.TipoConsumo;%3
Tension=datos.Tension;%4
FM=datos.FabricanteMedidor;%5
TipoMedicion=datos.TipMedicion;%6
GrupoConsumo=datos.GrupoDeConsumo;%7
Fases=datos.Fases; %8
MesesAdeudados=datos.MesesAdeudados;%9
Deuda=datos.Deuda;%10
PromedioFact_6Ult_Meses=datos.PromedioFact_6Ult_Meses;%11
ValorUltimaFactura=datos.ValorUltimaFactura;%12
ConsumoKWhActual=datos.ConsumoKWhActual;%13
ConsumoKWh1MesAntes=datos.ConsumoKWh1MesAntes;%14
ConsumoKWh2MesesAntes=datos.ConsumoKWh2MesesAntes;%15
ConsumoKWh3MesesAntes=datos.ConsumoKWh3MesesAntes;%16
ConsumoKWh4MesesAntes=datos.ConsumoKWh4MesesAntes;%17
ConsumoKWh5MesesAntes=datos.ConsumoKWh5MesesAntes;%18
ConsumoKWh6MesesAntes=datos.ConsumoKWh6MesesAntes;%19
ConsumoKWh7MesesAntes=datos.ConsumoKWh7MesesAntes;%20
ConsumoKWh8MesesAntes=datos.ConsumoKWh8MesesAntes;%21
ConsumoKWh9MesesAntes=datos.ConsumoKWh9MesesAntes;%22
ConsumoKWh10MesesAntes=datos.ConsumoKWh10MesesAntes;%23
ConsumoKWh11MesesAntes=datos.ConsumoKWh11MesesAntes;%24
ConsumoKWh12MesesAntes=datos.ConsumoKWh12MesesAntes;%25
ConsumoPro=datos.ConsumoPro_;%26
DemandaActualKW=datos.DemandaActualKW;%27
DemandaKW1MesAntes=datos.DemandaKW1MesAntes;%28
DemandaKW2MesesAntes=datos.DemandaKW2MesesAntes;%29
DemandaKW3MesesAntes=datos.DemandaKW3MesesAntes;%30
DemandaKW4MesesAntes=datos.DemandaKW4MesesAntes;%31
DemandaKW5MesesAntes=datos.DemandaKW5MesesAntes;%32
DemandaKW6MesesAntes=datos.DemandaKW6MesesAntes;%33
DemandaKW7MesesAntes=datos.DemandaKW7MesesAntes;%34
DemandaKW8MesesAntes=datos.DemandaKW8MesesAntes;%35
DemandaKW9MesesAntes=datos.DemandaKW9MesesAntes;%36
DemandaKW10MesesAntes=datos.DemandaKW10MesesAntes;%37
DemandaKW11MesesAntes=datos.DemandaKW11MesesAntes;%38
DemandaKW12MesesAntes=datos.DemandaKW12MesesAntes;%39
PecCli=datos.PecPorCliente;%40
Estrato=datos.EstratoGeografico;%41
AFabricacion=datos.A_oFabricacion;%42
%% Análisis de datos atípicos
fil1=find(TerceraEdad~=0&TerceraEdad(:,1)~=1);%3° Edad
fil2=find(Bdh~=0&Bdh~=1);%BDH
fil3=find(TipoConsumo~=1&TipoConsumo~=2);%Tipo Consumo
fil4=find(Tension~=1&Tension~=2&Tension~=3);%Tension
fil5=find(FM~=0&FM~=1&FM~=2&FM~=3&FM~=4&FM~=5&FM~=6&FM~=7&FM~=8&FM~=9&...
          FM~=10&FM~=11&FM~=12&FM~=13&FM~=14&FM~=15&FM~=16&FM~=17&...
          FM~=18&FM~=19&FM~=12&FM~=20&FM~=21&FM~=22&FM~=23&FM~=24&...
          FM~=25&FM~=26&FM~=27&FM~=28&FM~=29&FM~=30&FM~=31&FM~=32&...
          FM~=33&FM~=34&FM~=35&FM~=36&FM~=37&FM~=38&FM~=37&FM~=38&...
          FM~=39&FM~=40&FM~=41&FM~=42&FM~=43);%Fabricante Medidor
fil6=find(TipoMedicion~=0&TipoMedicion~=1&TipoMedicion~=2&...
          TipoMedicion~=3&TipoMedicion~=4&TipoMedicion~=5&...
          TipoMedicion~=6&TipoMedicion~=7&TipoMedicion~=8&...
          TipoMedicion~=9&TipoMedicion~=10);%Tipo Medición
fil7=find(GrupoConsumo~=1&GrupoConsumo~=2&GrupoConsumo~=3&...
          GrupoConsumo~=4&GrupoConsumo~=5);%Grupo Consumo
fil8=find(Fases~=1&Fases~=2&Fases~=3);%Fases
fil9=find(MesesAdeudados<0);%Se elimina Individuos con valores negativos
fil10=find(Deuda<0);%Se elimina individuos con valores negativos
fil11=find(PromedioFact_6Ult_Meses<0);%Se elimina individuos con negativos
fil12=find(ValorUltimaFactura<0);%Se elimina individuos con negativos
% Variables de consumo, se elimina únicamente valores negativos.
matrizConsumo=[ConsumoKWhActual,...
               ConsumoKWh1MesAntes,ConsumoKWh2MesesAntes,...
               ConsumoKWh3MesesAntes,ConsumoKWh4MesesAntes,...
               ConsumoKWh5MesesAntes,ConsumoKWh6MesesAntes,...
```

```
               ConsumoKWh7MesesAntes,ConsumoKWh8MesesAntes,...
               ConsumoKWh9MesesAntes,ConsumoKWh10MesesAntes,...
               ConsumoKWh11MesesAntes,ConsumoKWh12MesesAntes];
[fil13 col13]=find(matrizConsumo<0);% Se elimina solamente valores negativos.
matrizDem =[DemandaKW1MesAntes,...
            DemandaKW2MesesAntes,DemandaKW3MesesAntes,...
            DemandaKW4MesesAntes,DemandaKW5MesesAntes,...
            DemandaKW6MesesAntes,DemandaKW7MesesAntes,...
            DemandaKW8MesesAntes,DemandaKW9MesesAntes,...
            DemandaKW10MesesAntes,DemandaKW11MesesAntes,...
            DemandaKW12MesesAntes];
[fil14 col14]=find(matrizDem<0);

%Unimos las filas en donde se encontraron atípicos
atiUni=vertcat(fil1,fil2,fil3,fil4,fil5,fil6,fil7,fil8,fil9,fil10,...
               fil11,fil12,fil13,fil14);%Concatena los atípicos encontrados
                                        %verticalmente
InfoCli=unique(sort(atiUni));%En caso de tener filas repetidas se elimina
                             %y se ordena ascendentemente
% Filtrado de datos atípicos
datos(InfoCli,:)=[]; %Se elimina los datos atípicos
% Se crea un archivo Excel que presenta la matriz base limpia (sin datos nulos y sin
atípicos)
writetable(datos,'Matriz Base_Limpia.xlsx','Sheet',1);
```

A2.3. Aplicación de técnicas: K-Medias

```
%% --------------------Aplicación de técnicas: K-Medias---------------
% Este algoritmo realiza el agrupamiento de clientes "similares".
% Mediante la técnica k-medias se busca un agrupamiento con sistemas de medición
% "sospechosos".
%% Limpieza de pantalla y comandos
clear all
close all
clc
%% Lectura de datos
datos=readtable('MatrizBase.xlsx');
%% Guardado de Variables
GrupoConsumo=datos.GrupoDeConsumo;%8
PromedioFact=datos.PromedioFact_6Ult_Meses;%12
ValUltFac=datos.ValorUltimaFactura;%13
ConsumoKWhActual=datos.ConsumoKWhActual;%14
ConsumoKWh1MesAntes=datos.ConsumoKWh1MesAntes;%15
ConsumoKWh2MesesAntes=datos.ConsumoKWh2MesesAntes;%16
ConsumoKWh3MesesAntes=datos.ConsumoKWh3MesesAntes;%17
ConsumoKWh4MesesAntes=datos.ConsumoKWh4MesesAntes;%18
ConsumoKWh5MesesAntes=datos.ConsumoKWh5MesesAntes;%19
ConsumoKWh6MesesAntes=datos.ConsumoKWh6MesesAntes;%20
ConsumoKWh7MesesAntes=datos.ConsumoKWh7MesesAntes;%21
ConsumoKWh8MesesAntes=datos.ConsumoKWh8MesesAntes;%22
ConsumoKWh9MesesAntes=datos.ConsumoKWh9MesesAntes;%23
ConsumoKWh10MesesAntes=datos.ConsumoKWh10MesesAntes;%24
ConsumoKWh11MesesAntes=datos.ConsumoKWh11MesesAntes;%25
ConsumoKWh12MesesAntes=datos.ConsumoKWh12MesesAntes;%26

%% Guardado de datos en una matriz general
matrizConsumo=[ConsumoKWh12MesesAntes,ConsumoKWh11MesesAntes,...
               ConsumoKWh10MesesAntes,ConsumoKWh9MesesAntes,...
               ConsumoKWh8MesesAntes,ConsumoKWh7MesesAntes,...
               ConsumoKWh6MesesAntes,ConsumoKWh5MesesAntes,...
               ConsumoKWh4MesesAntes,ConsumoKWh3MesesAntes,...
               ConsumoKWh2MesesAntes,ConsumoKWh1MesAntes,...
               ConsumoKWhActual];
%% Cálculo de los atributos
% Se calcula los atributos para realizar el agrupamiento en base al consumo
% del cliente.
promedio=mean(matrizConsumo,2); %Promedio de consumo
maxCon= max(matrizConsumo,[],2); %Consumo máximo
minCon=min(matrizConsumo,[],2); %Consumo mínimo
rango=maxCon-minCon; %Diferencia máxima absoluta: módulo de la diferencia entre los
valores máximo y mínimo de consumo
desvEst=std(matrizConsumo,0,2); %Desviación Estándar
coefVar=desvEst./promedio; %Coeficiente de variación

% Eliminamos los datos NaN encontrados en la variable coeficiente de
```

```
% variación
[NanCoefVar]=find(isnan(coefVar));
coefVar(NanCoefVar,:)=0;

%Matriz
Atributos=[promedio,desvEst,coefVar,maxCon,minCon,rango,...
           PromedioFact,ValUltFac];

%% División de clientes por grupo de consumo
% Aquí se divide los clientes por grupo de consumo
% 1 = Residencial
% 2 = Comercial
% 3 = Industrial
% 4 = Otros y Alumbrado Público
filRes=find(GrupoConsumo==1);
filCom=find(GrupoConsumo==2);
filIndu=find(GrupoConsumo==3);
filOtros=find(GrupoConsumo==4&& GrupoConsumo==5);

AtributosR=Atributos(filRes,:);
AtributosC=Atributos(filCom,:);
AtributosI=Atributos(filIndu,:);
AtributosO=Atributos(filOtros,:);

%% Normalización de las matrices
XNorR=zscore(AtributosR);
XNorC=zscore(AtributosC);
XNorI=zscore(AtributosI);
XNorO=zscore(AtributosO);
%% Técnica k-medias
dist_K='sqeuclidean'; %Distancia Euclidiana
GR=5; GC=5; GI=5; GO=5;%Grupos
% idx: Índices de agrupamiento, devueltos como un vector.
% C:   Ubicaciones de centro de agrupamiento.
% sum: Sumas dentro del grupo de distancias de punto a centroide.
% D:   Distancias desde cada punto hasta cada centroide.
[idxR,CR,sumR,DR]=kmeans(XNorR,GR,'Distance',dist_K);
[idxC,CC,sumC,DC]=kmeans(XNorC,GC,'Distance',dist_K);
[idxI,CI,sumI,DI]=kmeans(XNorI,GI,'Distance',dist_K);
[idxO,CO,sumO,DO]=kmeans(XNorO,GO,'Distance',dist_K);
%% Generación de Reportes
% Grupos Residenciales
writetable(datos(filRes,:),'Rep_KMedias.xlsx','Sheet',1);
writetable(table(idxR),'Rep_KMedias.xlsx','Sheet',1,'Range','BQ1');
% Grupos Comerciales
writetable(datos(filCom,:),'Rep_KMedias.xlsx','Sheet',2);
writetable(table(idxC),'Rep_KMedias.xlsx','Sheet',2,'Range','BQ1');
% Grupos Industriales
writetable(datos(filIndu,:),'Rep_KMedias.xlsx','Sheet',3);
writetable(table(idxI),'Rep_KMedias.xlsx','Sheet',3,'Range','BQ1');
% Grupos Otros
writetable(datos(filOtros,:),'Rep_KMedias.xlsx','Sheet',4);
writetable(table(idxO),'Rep_KMedias.xlsx','Sheet',4,'Range','BQ1');
```

A2.4. Aplicación de técnicas: K-Vecinos

```
%% -------------------Aplicación de técnicas: K-Vecinos--------------------
% ---------------------------K-Nearest Neighbors----------------------------
% En este programa, se utiliza el toolbox de MATLAB para el entrenamiento
% y la clasificación de nuevos valores.
% El comando utilizado para el entrenamiento es "fitcknn".
% El comando utilizado para la clasificación es "predict"
close all; clear all; clc %Limpieza de pantalla
%% Lectura de datos - Matriz de Entrenamiento
datos=readtable('Entrenamiento.xlsx');
 % Guardado de Variables - Matriz de Entrenamiento
ConsumoKWhActual=datos.ConsumoKWhActual;%12
ConsumoKWh1MesAntes=datos.ConsumoKWh1MesAntes;%13
ConsumoKWh2MesesAntes=datos.ConsumoKWh2MesesAntes;%14
ConsumoKWh3MesesAntes=datos.ConsumoKWh3MesesAntes;%15
ConsumoKWh4MesesAntes=datos.ConsumoKWh4MesesAntes;%16
ConsumoKWh5MesesAntes=datos.ConsumoKWh5MesesAntes;%17
ConsumoKWh6MesesAntes=datos.ConsumoKWh6MesesAntes;%18
ConsumoKWh7MesesAntes=datos.ConsumoKWh7MesesAntes;%19
ConsumoKWh8MesesAntes=datos.ConsumoKWh8MesesAntes;%20
```

```
ConsumoKWh9MesesAntes=datos.ConsumoKWh9MesesAntes;%21
ConsumoKWh10MesesAntes=datos.ConsumoKWh10MesesAntes;%22
ConsumoKWh11MesesAntes=datos.ConsumoKWh11MesesAntes;%23
ConsumoKWh12MesesAntes=datos.ConsumoKWh12MesesAntes;%24
Clase=datos.Etiqueta;
% Guardado de datos en una matriz general
matrizConsumo=[ConsumoKWh12MesesAntes,ConsumoKWh11MesesAntes,...
               ConsumoKWh10MesesAntes,ConsumoKWh9MesesAntes,...
               ConsumoKWh8MesesAntes,ConsumoKWh7MesesAntes,...
               ConsumoKWh6MesesAntes,ConsumoKWh5MesesAntes,...
               ConsumoKWh4MesesAntes,ConsumoKWh3MesesAntes,...
               ConsumoKWh2MesesAntes,ConsumoKWh1MesAntes,...
               ConsumoKWhActual];

%% Lectura de datos - matriz de prueba
datos=readtable('MatrizBase.xlsx');
% Guardado de variables
ConsumoKWhActualP=datos.ConsumoKWhActual;%12
ConsumoKWh1MesAntesP=datos.ConsumoKWh1MesAntes;%13
ConsumoKWh2MesesAntesP=datos.ConsumoKWh2MesesAntes;%14
ConsumoKWh3MesesAntesP=datos.ConsumoKWh3MesesAntes;%15
ConsumoKWh4MesesAntesP=datos.ConsumoKWh4MesesAntes;%16
ConsumoKWh5MesesAntesP=datos.ConsumoKWh5MesesAntes;%17
ConsumoKWh6MesesAntesP=datos.ConsumoKWh6MesesAntes;%18
ConsumoKWh7MesesAntesP=datos.ConsumoKWh7MesesAntes;%19
ConsumoKWh8MesesAntesP=datos.ConsumoKWh8MesesAntes;%20
ConsumoKWh9MesesAntesP=datos.ConsumoKWh9MesesAntes;%21
ConsumoKWh10MesesAntesP=datos.ConsumoKWh10MesesAntes;%22
ConsumoKWh11MesesAntesP=datos.ConsumoKWh11MesesAntes;%23
ConsumoKWh12MesesAntesP=datos.ConsumoKWh12MesesAntes;%24
% Matriz General de prueba
matrizPrueba=[ConsumoKWh12MesesAntesP,ConsumoKWh11MesesAntesP,...
              ConsumoKWh10MesesAntesP,ConsumoKWh9MesesAntesP,...
              ConsumoKWh8MesesAntesP,ConsumoKWh7MesesAntesP,...
              ConsumoKWh6MesesAntesP,ConsumoKWh5MesesAntesP,...
              ConsumoKWh4MesesAntesP,ConsumoKWh3MesesAntesP,...
              ConsumoKWh2MesesAntesP,ConsumoKWh1MesAntesP,...
              ConsumoKWhActualP];

%% Obtención de atributos estadísticos - Matriz de entrenamiento
promedio=mean(matrizConsumo,2); %Promedio de consumo
desvEst=std(matrizConsumo,0,2); %Desviación Estándar
coefVar=desvEst./promedio;%Coeficiente de variación
[FilNaN]=find(isnan(coefVar));
coefVar(FilNaN,:)=0;%Reemplazamos los datos NaN por 0
minE=min(matrizConsumo,[],2); %Valor máximo
maxE=max(matrizConsumo,[],2); %Valor mínimo
RangoE=maxE-minE;

% Matriz de Entrenamiento
Entrenamiento=[promedio,desvEst,coefVar,minE,maxE,RangoE];

%% Atributos estadísticos - Matriz Prueba
promedioP=mean(matrizPrueba,2); %Promedio de consumo
desvEstP=std(matrizPrueba,0,2); %Desviación Estándar
coefVarP=desvEstP./promedioP; %Coeficiente de variación
[FilNaNP]=find(isnan(coefVarP));
coefVarP(FilNaNP,:)=0;%Reemplazamos los datos NaN por 0
minP=min(matrizPrueba,[],2);
maxP=max(matrizPrueba,[],2);
RangoP=maxP-minP;

% Matriz de Prueba
Prueba=[promedioP,desvEstP,coefVarP,minP,maxP,RangoP];

%% Normalización de los datos
%Normalización de la matriz de entrenamiento
ZEntrenamiento=(Entrenamiento-min(Entrenamiento(:))./...
                (max(Entrenamiento(:))-min(Entrenamiento(:)));

%Normalización de la matriz de prueba con max-min
ZPrueba=(Prueba-min(Prueba(:))./...
        (max(Prueba(:))-min(Prueba(:)));

%% K-Vecinos
K=10;
% Entrenamiento
```

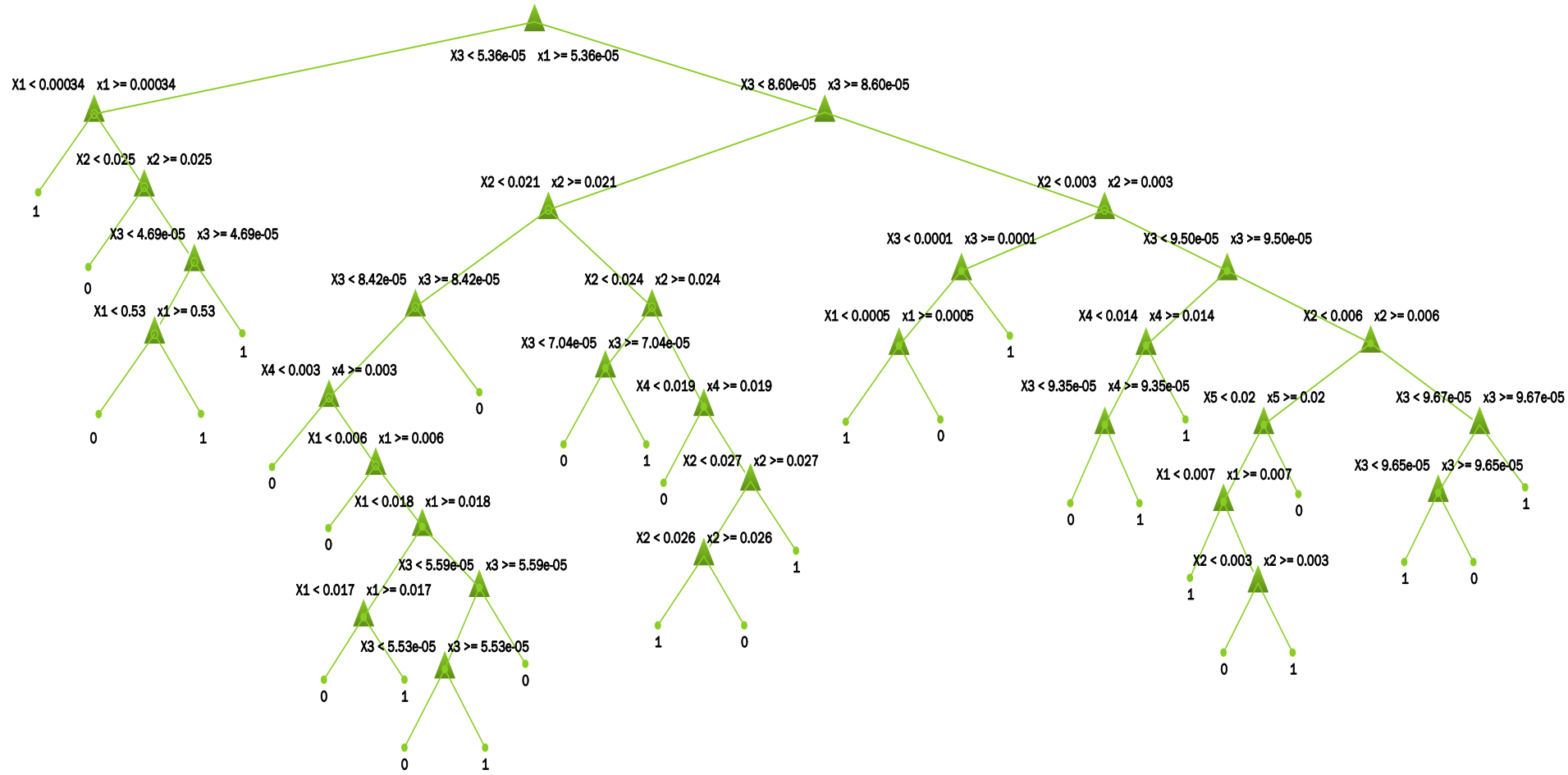
```
entrenar=fitcknn(ZEntrenamiento,Clase,'NumNeighbors', K)
% Clasificación
[etiqueta, puntaje, costo] = predict(entrenar, ZPrueba);
%% Generación de reportes
[fil]=find(etiqueta==1); %Busca filas marcadas por 1="sospechoso"
writetable(datos(fil,:),'SistemasDeMedicion_Sospechosos.xlsx'); % Genera reporte Excel
```

A2.5. Aplicación de técnicas: Árbol de decisión

```
%% ------------Aplicación de técnicas: Árbol de decisión--------------
% Este programa realiza la clasificación mediante el algoritmo de árbol de decisión
con % el toolbox de MATLAB.
clear all; close all; clc % Limpieza de pantalla
%% Lectura de datos
datos=readtable('Entrenamiento.xlsx');
%% Guardado de Variables
ConsumoKWhActual=datos.ConsumoKWhActual;%12
ConsumoKWh1MesAntes=datos.ConsumoKWh1MesAntes;%13
ConsumoKWh2MesesAntes=datos.ConsumoKWh2MesesAntes;%14
ConsumoKWh3MesesAntes=datos.ConsumoKWh3MesesAntes;%15
ConsumoKWh4MesesAntes=datos.ConsumoKWh4MesesAntes;%16
ConsumoKWh5MesesAntes=datos.ConsumoKWh5MesesAntes;%17
ConsumoKWh6MesesAntes=datos.ConsumoKWh6MesesAntes;%18
ConsumoKWh7MesesAntes=datos.ConsumoKWh7MesesAntes;%19
ConsumoKWh8MesesAntes=datos.ConsumoKWh8MesesAntes;%20
ConsumoKWh9MesesAntes=datos.ConsumoKWh9MesesAntes;%21
ConsumoKWh10MesesAntes=datos.ConsumoKWh10MesesAntes;%22
ConsumoKWh11MesesAntes=datos.ConsumoKWh11MesesAntes;%23
ConsumoKWh12MesesAntes=datos.ConsumoKWh12MesesAntes;%24
Clase=datos.Etiqueta;
% Guardado de datos en una matriz general
matrizConsumo=[ConsumoKWh12MesesAntes,ConsumoKWh11MesesAntes,...
               ConsumoKWh10MesesAntes,ConsumoKWh9MesesAntes,...
               ConsumoKWh8MesesAntes,ConsumoKWh7MesesAntes,...
               ConsumoKWh6MesesAntes,ConsumoKWh5MesesAntes,...
               ConsumoKWh4MesesAntes,ConsumoKWh3MesesAntes,...
               ConsumoKWh2MesesAntes,ConsumoKWh1MesAntes,...
               ConsumoKWhActual];
%% Carga de datos de prueba
datos=readtable('MatrizBase.xlsx');
% Guardado de variables
ConsumoKWhActualP=datos.ConsumoKWhActual;%12
ConsumoKWh1MesAntesP=datos.ConsumoKWh1MesAntes;%13
ConsumoKWh2MesesAntesP=datos.ConsumoKWh2MesesAntes;%14
ConsumoKWh3MesesAntesP=datos.ConsumoKWh3MesesAntes;%15
ConsumoKWh4MesesAntesP=datos.ConsumoKWh4MesesAntes;%16
ConsumoKWh5MesesAntesP=datos.ConsumoKWh5MesesAntes;%17
ConsumoKWh6MesesAntesP=datos.ConsumoKWh6MesesAntes;%18
ConsumoKWh7MesesAntesP=datos.ConsumoKWh7MesesAntes;%19
ConsumoKWh8MesesAntesP=datos.ConsumoKWh8MesesAntes;%20
ConsumoKWh9MesesAntesP=datos.ConsumoKWh9MesesAntes;%21
ConsumoKWh10MesesAntesP=datos.ConsumoKWh10MesesAntes;%22
ConsumoKWh11MesesAntesP=datos.ConsumoKWh11MesesAntes;%23
ConsumoKWh12MesesAntesP=datos.ConsumoKWh12MesesAntes;%24
% Matriz General de prueba
matrizPrueba=[ConsumoKWh12MesesAntesP,ConsumoKWh11MesesAntesP,...
              ConsumoKWh10MesesAntesP,ConsumoKWh9MesesAntesP,...
              ConsumoKWh8MesesAntesP,ConsumoKWh7MesesAntesP,...
              ConsumoKWh6MesesAntesP,ConsumoKWh5MesesAntesP,...
              ConsumoKWh4MesesAntesP,ConsumoKWh3MesesAntesP,...
              ConsumoKWh2MesesAntesP,ConsumoKWh1MesAntesP,...
              ConsumoKWhActualP];
%% Obtención de atributos estadísticos de matriz de entrenamiento
promedio=mean(matrizConsumo,2); %Promedio de consumo
desvEst=std(matrizConsumo,0,2); %Desviación Estándar
coefVar=desvEst./promedio; %Coeficiente de variación
[FilNaN]=find(isnan(coefVar));
coefVar(FilNaN,:)=0;%Reemplazamos los datos NaN por 0
minE=min(matrizConsumo,[],2); %Valor máximo
maxE=max(matrizConsumo,[],2); %Valor mínimo
RangoE=maxE-minE;
% Matriz de Entrenamiento
Entrenamiento=[promedio,desvEst,coefVar,minE,maxE,RangoE];
% Normalización de los datos max-min
ZEntrenamiento=(Entrenamiento-min(Entrenamiento(:))./...
```

```
                (max(Entrenamiento(:))-min(Entrenamiento(:)));
%% Obtención de Atributos estadísticos de matriz prueba
promedioP=mean(matrizPrueba,2); %Promedio de consumo
desvEstP=std(matrizPrueba,0,2); %Desviación Estándar
coefVarP=desvEstP./promedioP; %Coeficiente de variación
[FilNaNP]=find(isnan(coefVarP));
coefVarP(FilNaNP,:)=0;%Reemplazamos los datos NaN por 0
minP=min(matrizPrueba,[],2);
maxP=max(matrizPrueba,[],2);
RangoP=maxP-minP;

% Matriz de Prueba
Prueba=[promedioP,desvEstP,coefVarP,minP,maxP,RangoP];

%Normalización de la matriz prueba
ZPrueba=(Prueba-min(Prueba(:)))./...
        (max(Prueba(:))-min(Prueba(:)));

%% Creación del árbol de decisión
arbol = fitctree(ZEntrenamiento, Clase) %Crea el árbol de decisión

%% Muestra del árbol de decisión
view(arbol,'Mode','graph')
view(arbol)

%% Clasificación de nuevos datos
clasificacion = predict(arbol,ZPrueba)

%% Generación de reporte
[fil]=find(clasificacion==1); %Busca filas marcadas por 1="sospechoso"
writetable(datos(fil,:),'Sospechosos_Arbol.xlsx'); % Genera reporte Excel
```

A3. ANEXO 3 – ÁRBOL DE DECISIÓN



A4. ANEXO 4 – ENTRENAMIENTO DE RED NEURONAL

El entrenamiento de la red neuronal que ayudará a clasificar los sistemas de medición, se realiza mediante el Toolbox de MATLAB®. El proceso es el siguiente:

El comando "*nftool*" inicia el proceso de entrenamiento de la red, con él, aparece una pantalla como se muestra en la Figura A4. 1.

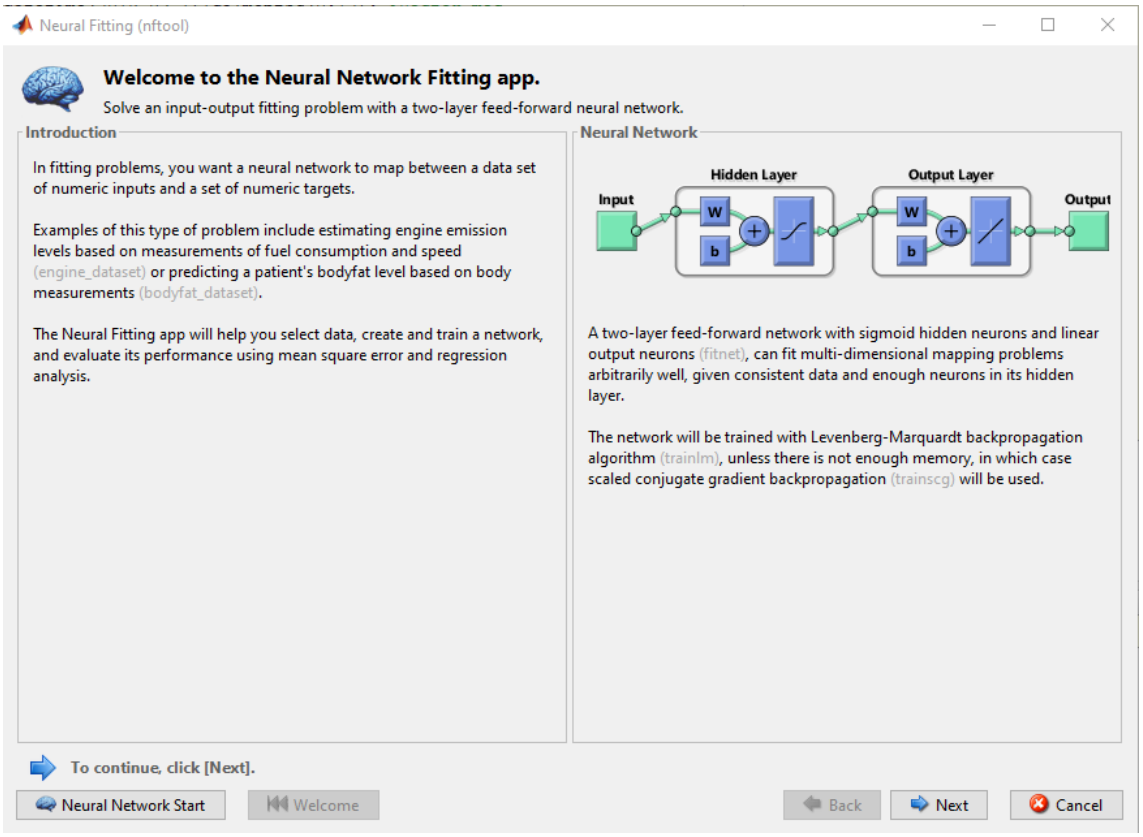


Figura A4. 1, Pantalla de "nftool"

Como ya se explicó, la red neuronal es una técnica supervisada, por lo que, se entrena con ejemplos etiquetados de fraude y no fraude y estos tienen que estar normalizados.

En este caso "Input" son los ejemplos de entrenamiento (Figura A4. 2) y "Target" es la clase, etiqueta u objetivo de los ejemplos (Figura A4. 3).

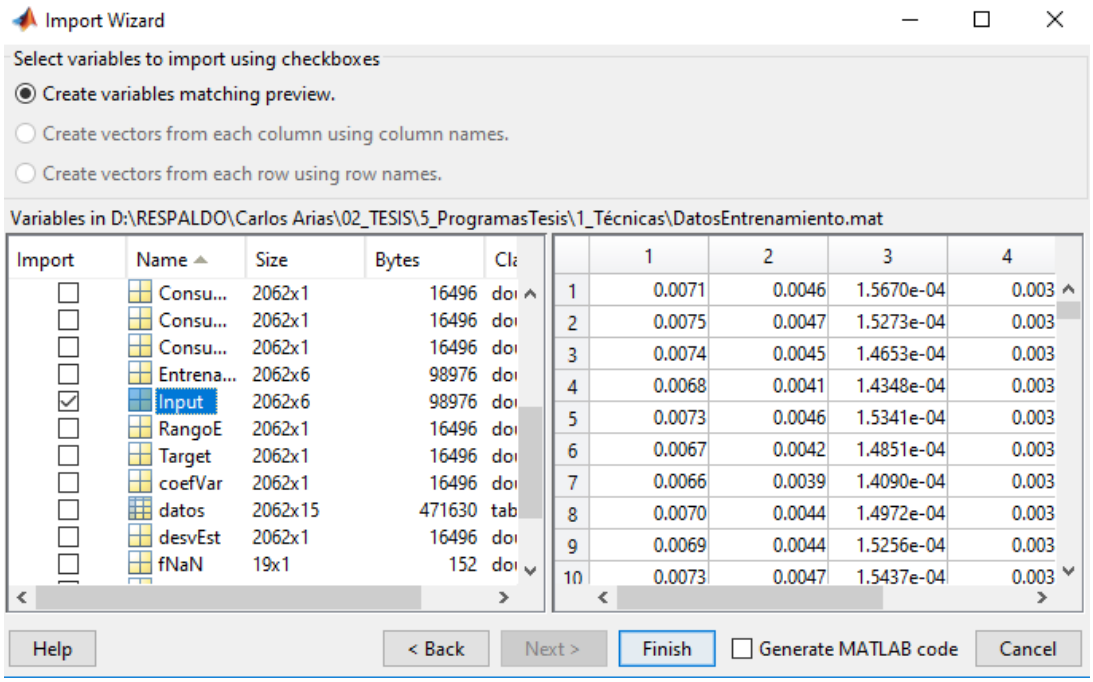


Figura A4. 2, Ejemplo de entrenamiento

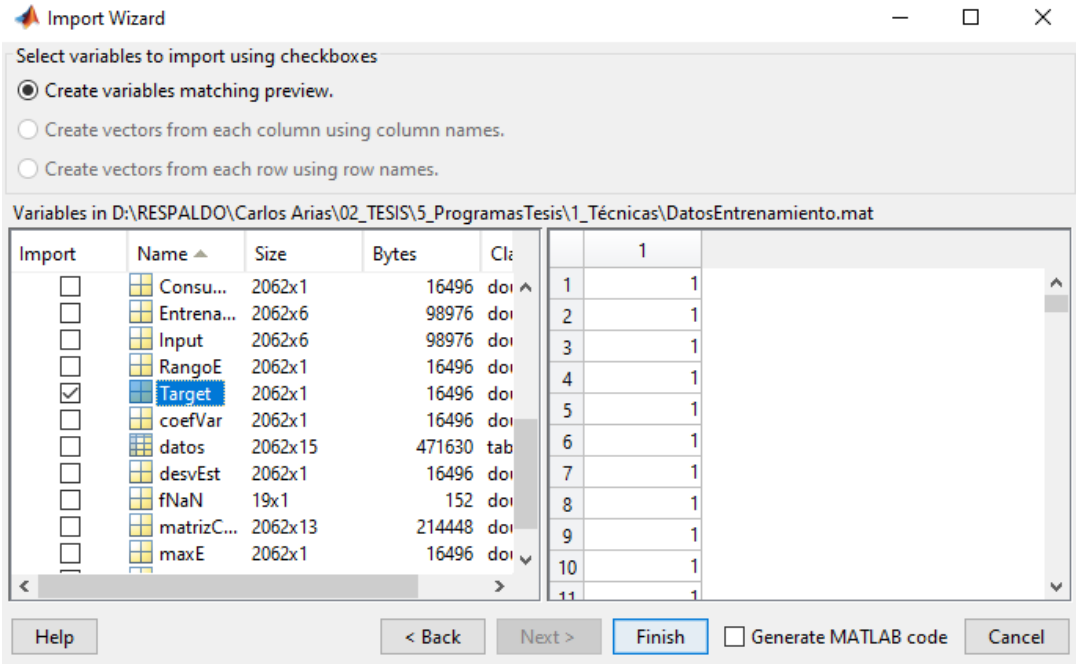


Figura A4. 3, Clase u Objetivo

La base de datos de entrenamiento consta de 2062 ejemplos de 6 variables (promedio, desviación estándar, coeficiente de variación, máximo, mínimo y rango). Esto se puede constatar en la Figura A4. 4.

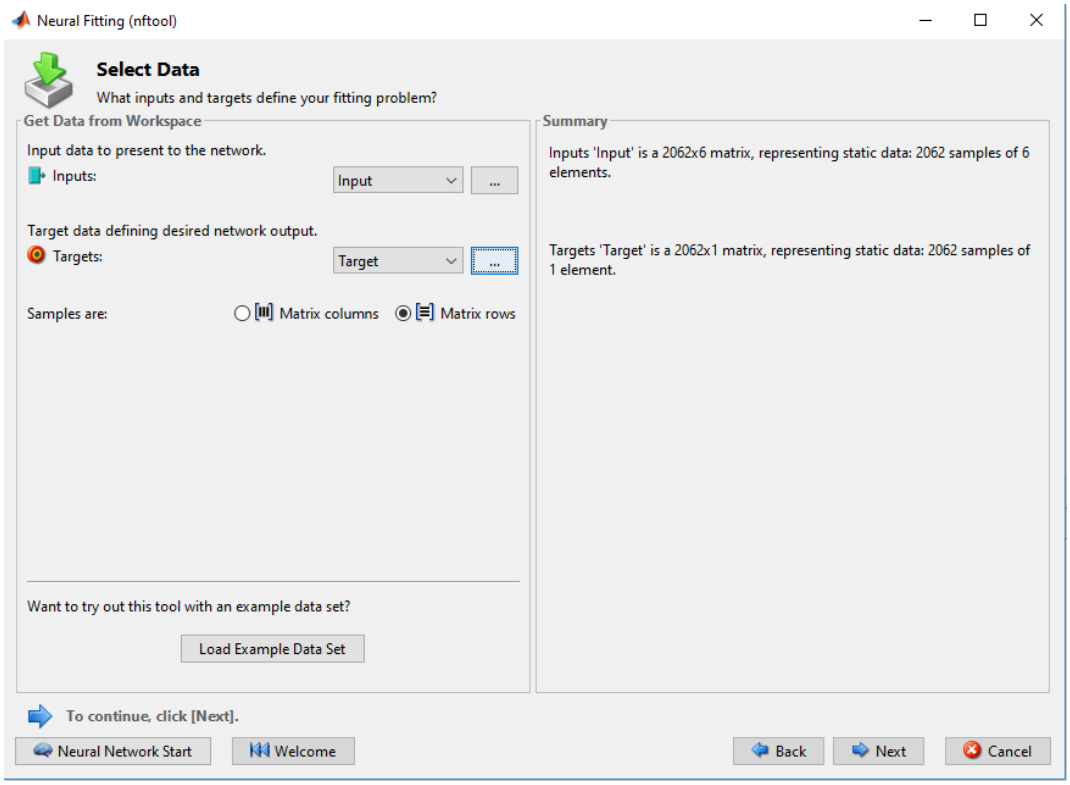


Figura A4. 4, Selección de datos de entrenamiento

Como se aprecia en la Figura A4. 5, se selecciona aleatoriamente el 70% de datos para el entrenamiento, 15% para la validación y 15% para la prueba.

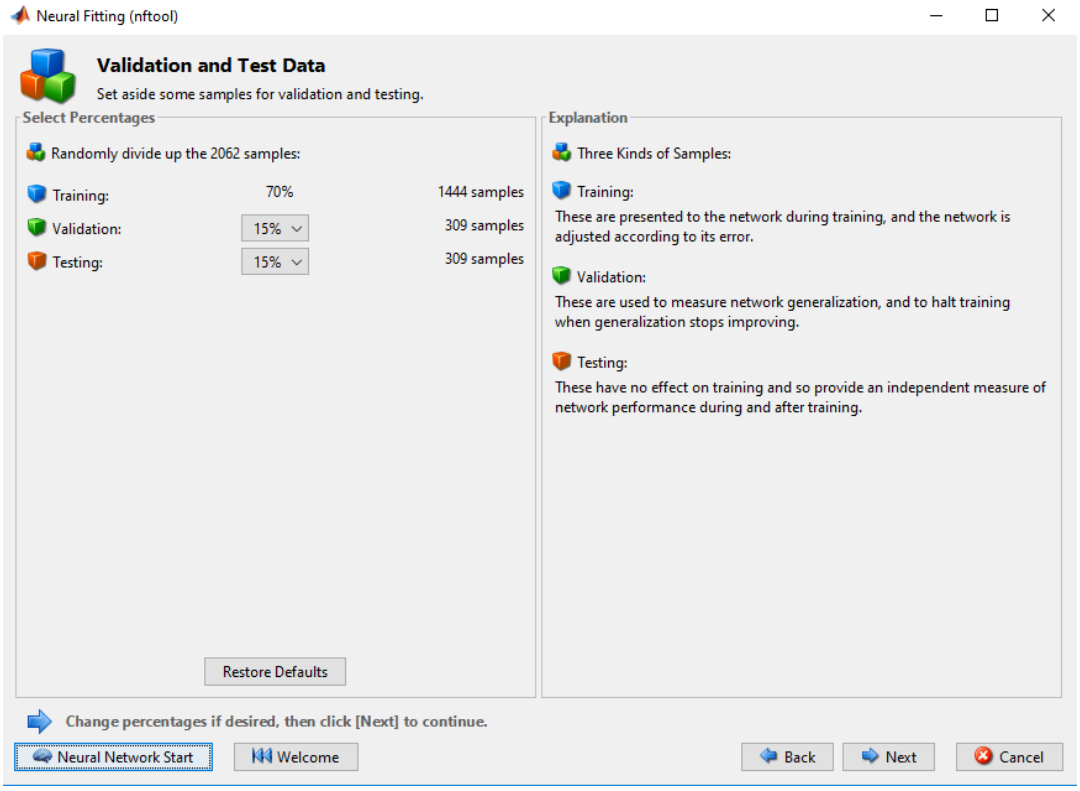


Figura A4. 5, División de datos: Entrenamiento, Validación y Prueba

En la Figura A4. 6, se selecciona la estructura de la red neuronal, para esto se selecciona 10 neuronas en la capa oculta y una neurona en la capa de salida que será para la clasificación.

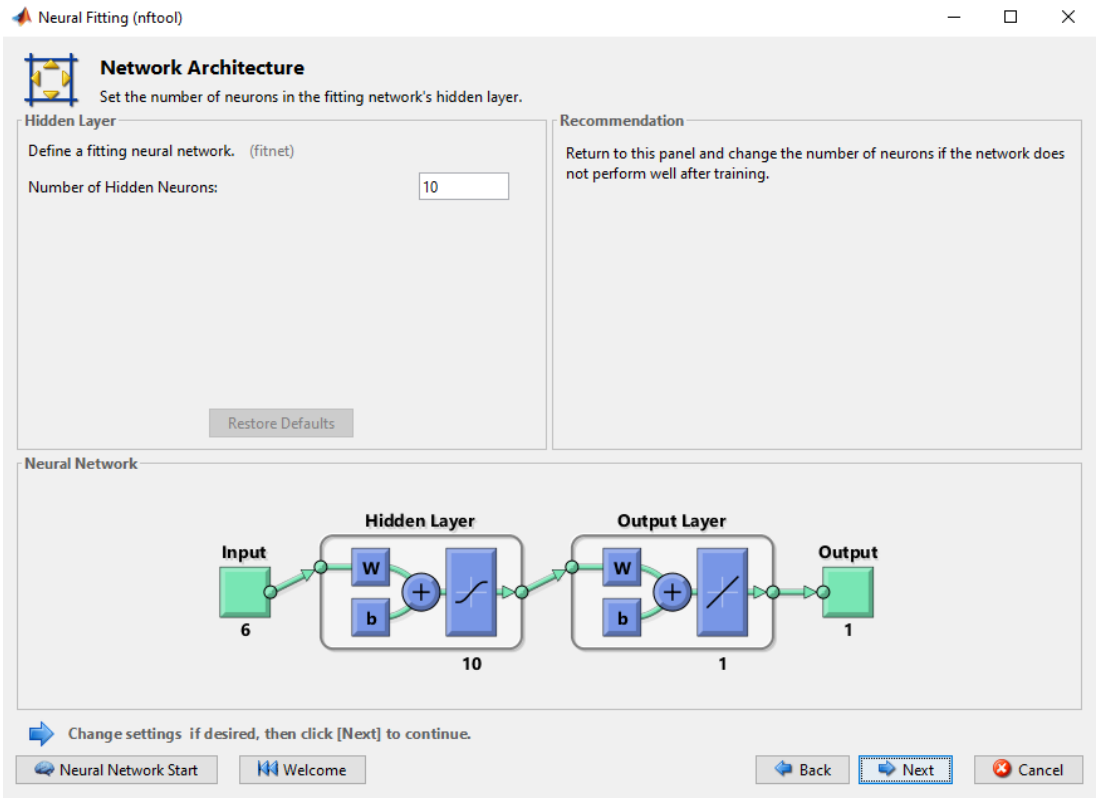


Figura A4. 6, Estructuración de la Red Neuronal

La Figura A4. 7, muestra la estructura de la Red Neuronal.

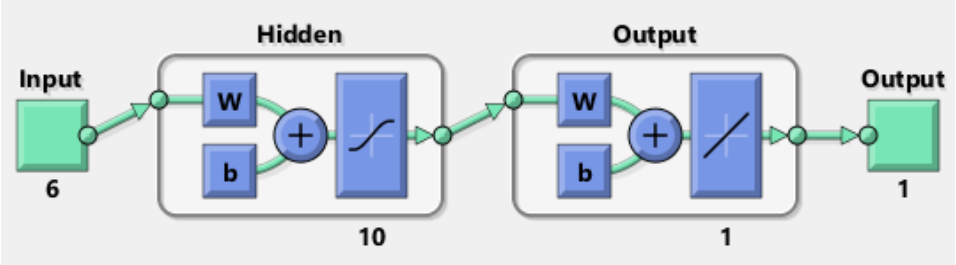


Figura A4. 7, Estructura de la Red Neuronal

El siguiente paso es seleccionar el algoritmo de entrenamiento. El algoritmo de entrenamiento seleccionado es el "Levenberg-Marquardt Backpropagation" como se muestra en la Figura A4. 8.

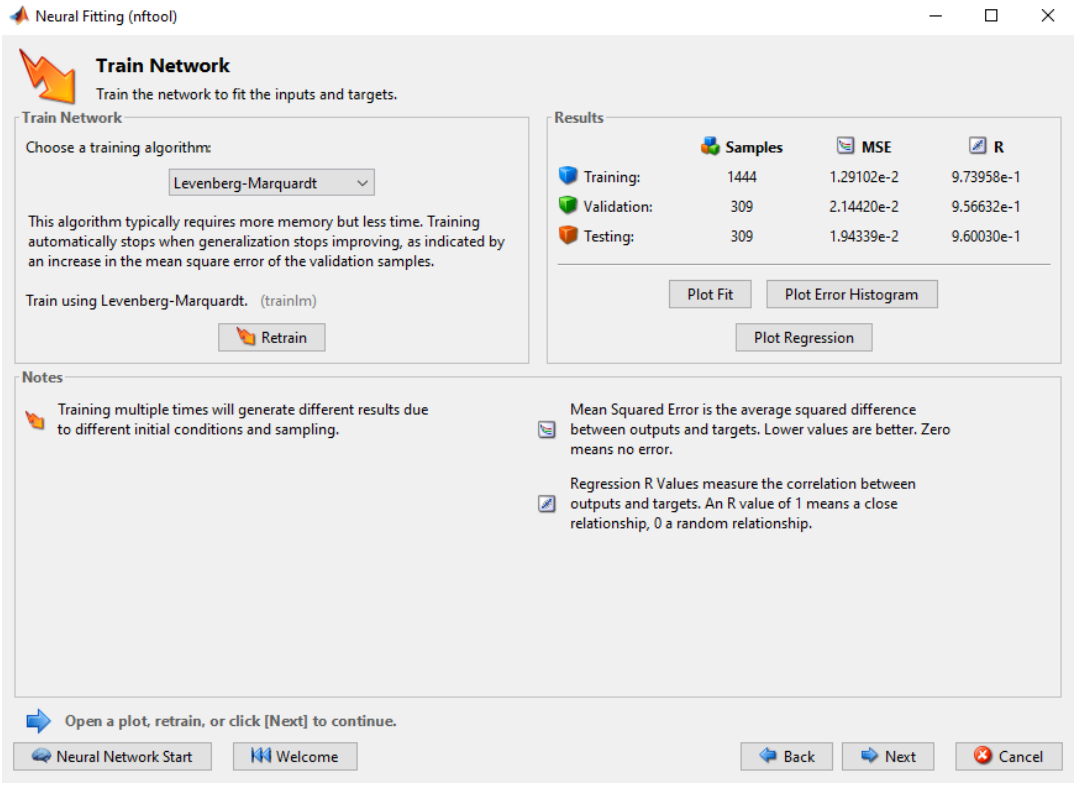


Figura A4. 8, Entrenamiento de la Red

A continuación, se muestra los resultados del entrenamiento:

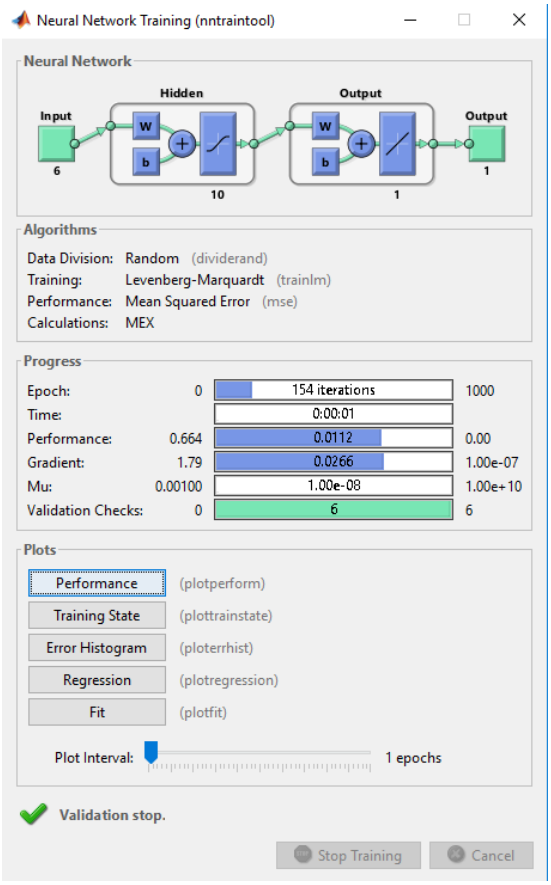


Figura A4. 9, Resultados del entrenamiento

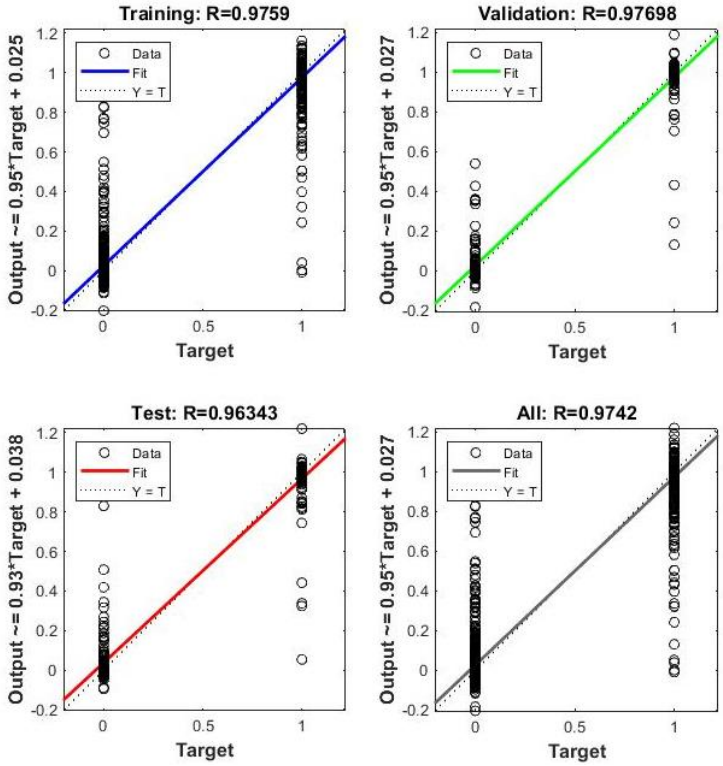


Figura A4. 10, Resultados de entrenamiento - Regresión.

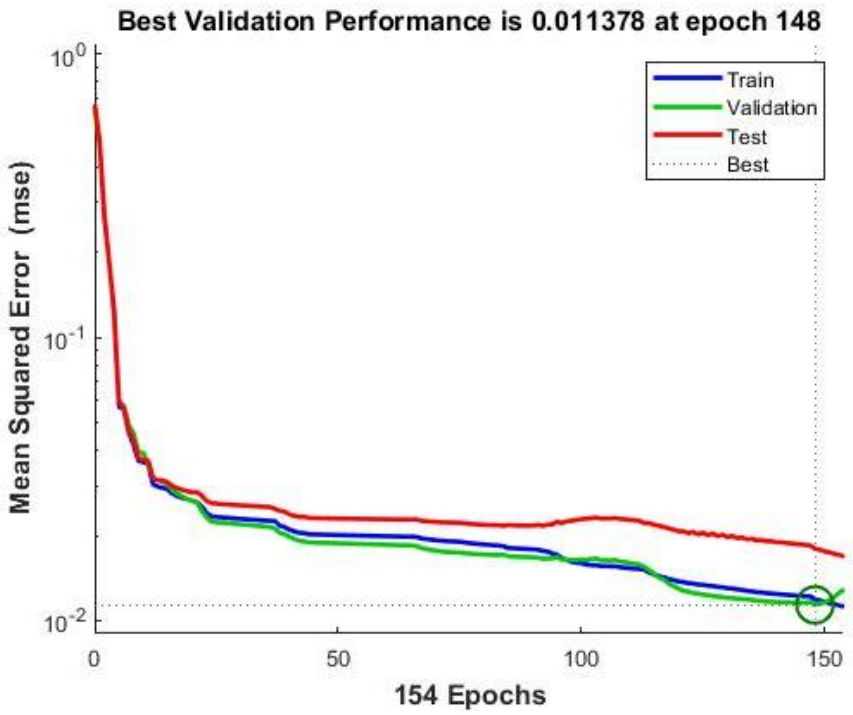


Figura A4. 11, Resultados de entrenamiento - Validación de rendimiento

A5.ANEXO 5 – MANUAL DE USUARIO DE INTERFAZ GRÁFICA

Para el desarrollo de listas de sistemas de medición "sospechosos" de mal funcionamiento o posible hurto de energía y para agilizar el proceso, se efectuó una interfaz gráfica en MATLAB – GUIDE denominado "CONTROL DE PÉRDIDAS NO TÉCNICAS". A continuación, se describe el manual de usuario:

A5.1. Menú principal

La Figura A5. 1muestra la pantalla del menú principal y como se observa, tiene varios botones que representa el proceso de minería de datos a desarrollarse para el control de pérdidas no técnicas.

El proceso inicia con la selección de datos, seguido del pre-procesamiento de datos, Procesamiento y finalmente Resultados.

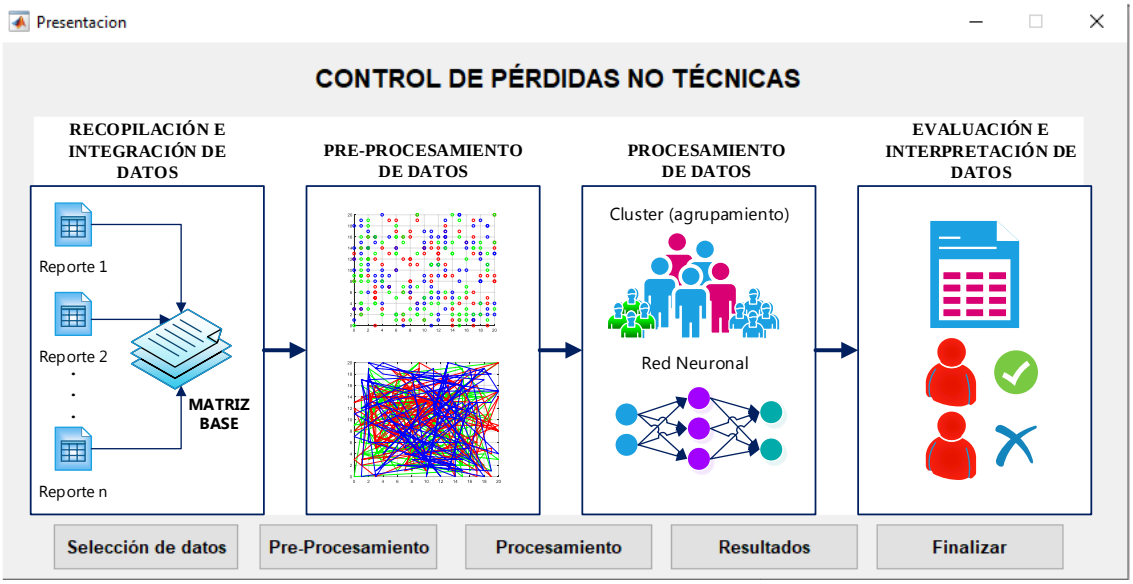


Figura A5. 1, Menú principal

A5.2. Selección de datos

Se inicia el proceso dando clic en "Selección de datos" y se abre la pantalla que se observa en la Figura A5. 2.

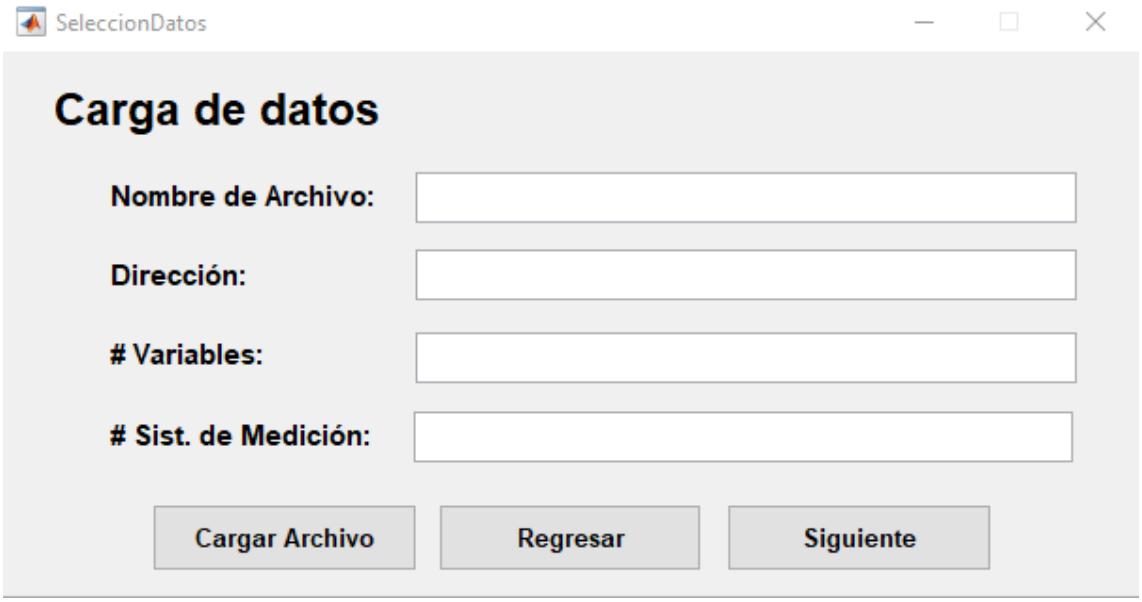


Figura A5. 2, Selección de datos

Para cargar el archivo que contendrá los datos a analizarse, clic en "Cargar Archivo" y se abre la pantalla que está en la Figura A5. 3. Seleccionar el archivo y dar clic en "Abrir"

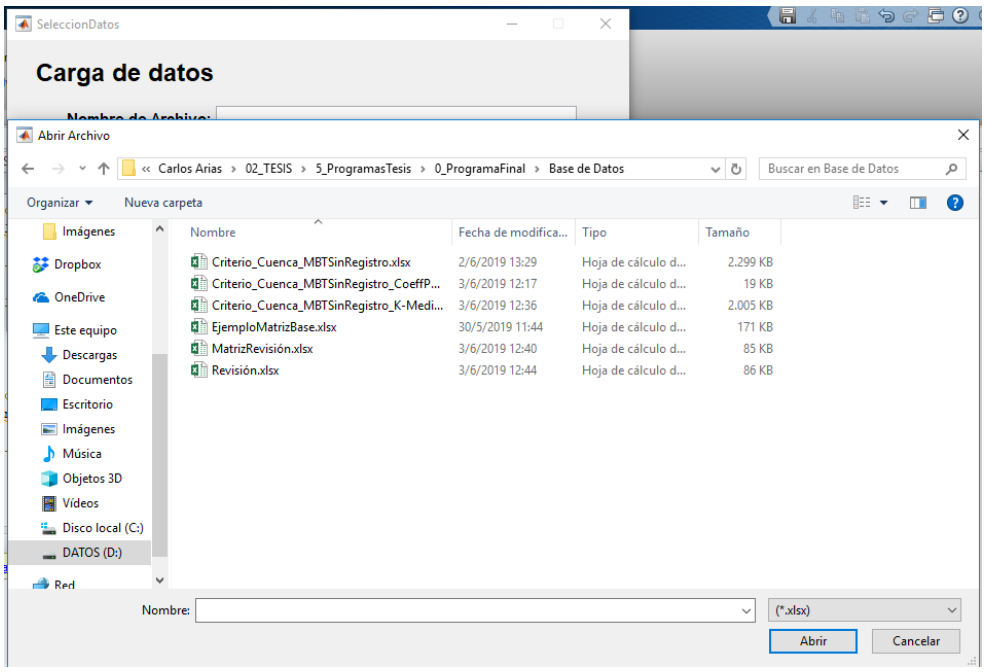


Figura A5. 3, Cargar Datos

Empezará a cargar el archivo y mostrará un aviso como se muestra en la Figura A5. 4. Esperar unos instantes; este tiempo de carga dependerá del tamaño del archivo seleccionado. Una vez terminada la carga aparecerá otro aviso como se observa en la Figura A5. 5. Clic en "OK" y se puede continuar con el proceso.

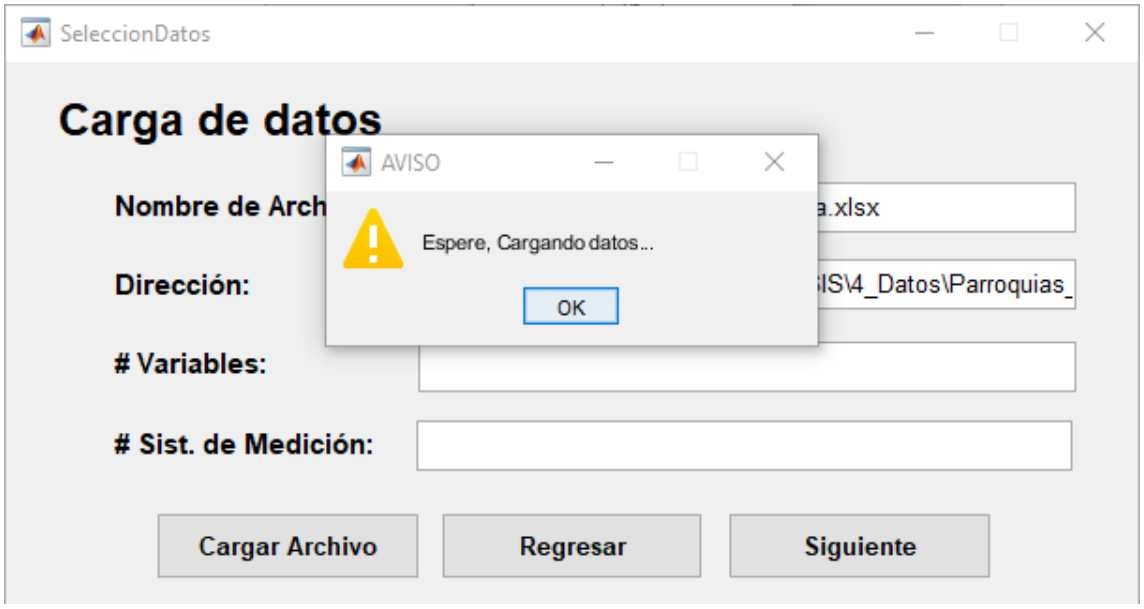


Figura A5. 4, Aviso de "cargando datos"

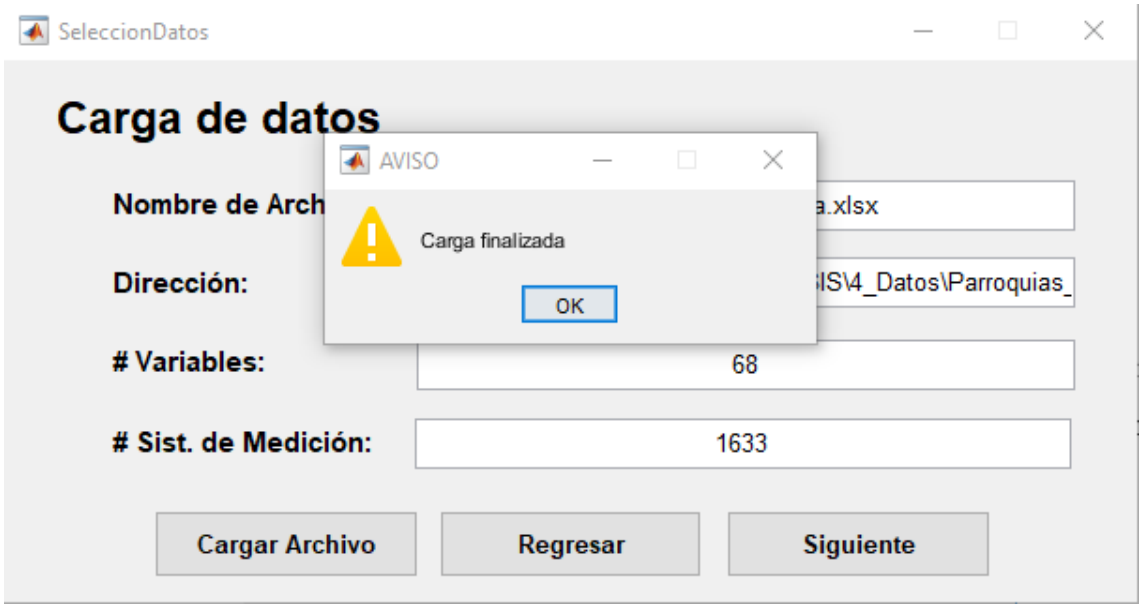


Figura A5. 5, Aviso de "carga finalizada"

Cargado el archivo, se muestra los parámetros de los datos, en donde, "# Variables" tendrán que ser las 68 para el análisis y en "# Sist. de Medición" la cantidad de sistemas de medición cargados como se aprecia en la Figura A5. 6.

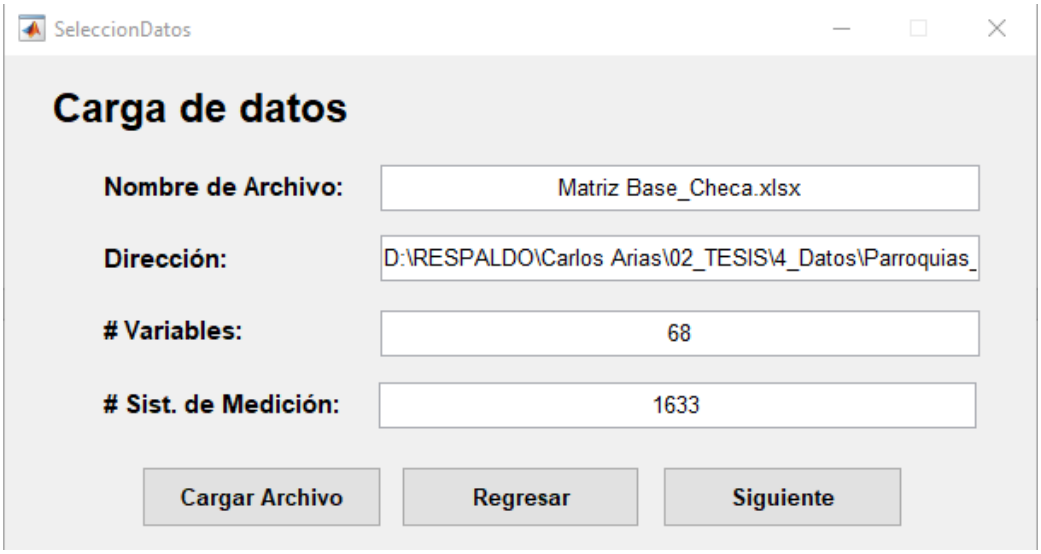


Figura A5. 6, Carga completada

A5.3. Pre-Procesamiento de datos

Con la carga lista, se puede proceder dando clic en "Siguiente" y aparecerá una pantalla como está en Figura A5. 7.

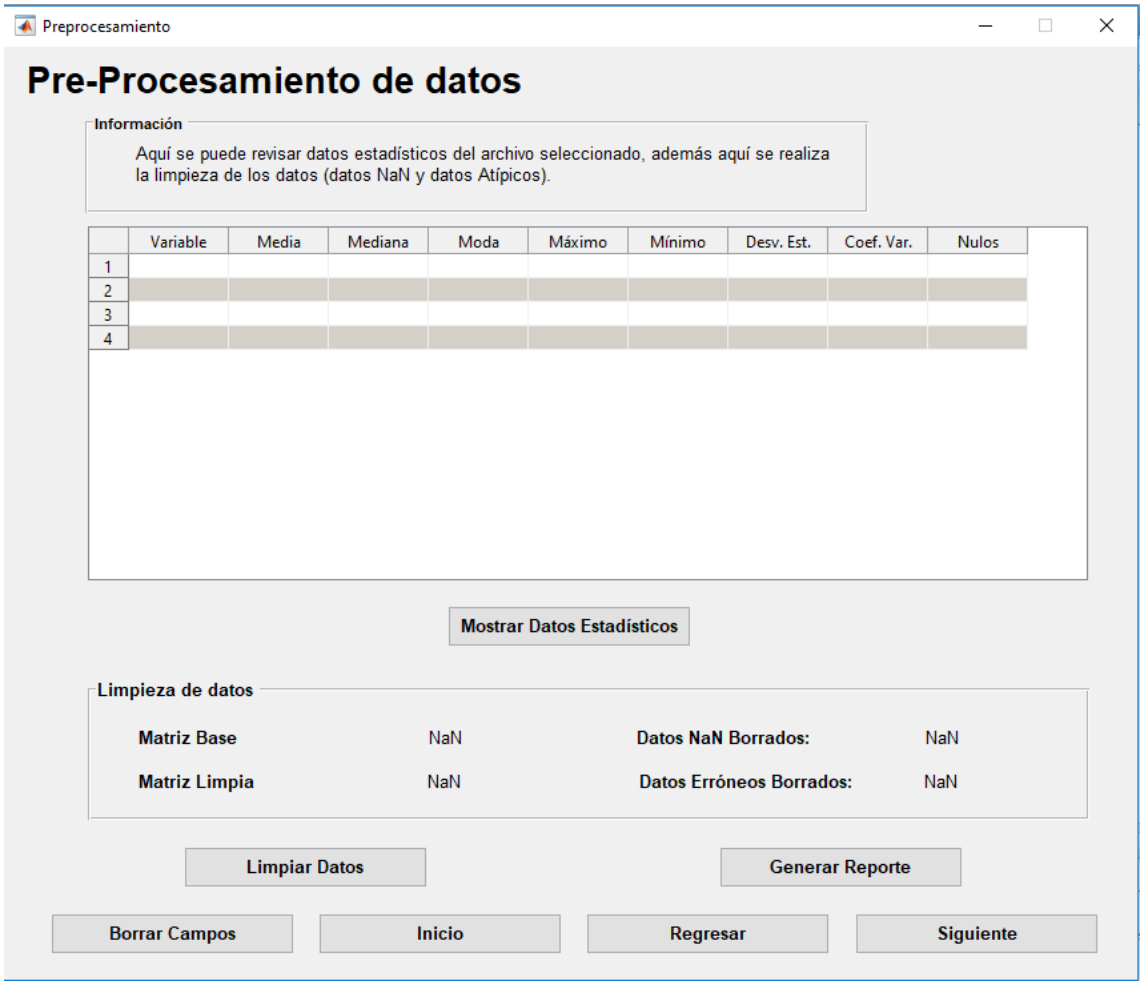


Figura A5. 7, Pre-Procesamiento de datos

En esta etapa se realiza la limpieza de los datos. Dando clic en “Mostrar Datos Estadísticos”, se muestra una tabla con datos estadísticos de las variables cargadas; además, al dar clic en “Limpiar Datos” la matriz cargada se “Limpiara”, es decir, eliminará datos NaN y datos erróneos e indicará cuantos datos corresponden a estos y el número de sistemas de medición que quedará en la matriz “limpia”. Esto se puede observar en la Figura A5. 8.

	Variable	Media	Mediana	Moda	Máximo	Mínimo	Desv. Est.	Coef. Var.	Nulos
1	TerceraEdad	0.0502	0	0	1	0	0.2185	435.0428	0
2	Bdh	0.0116	0	0	1	0	0.1073	921.9513	0
3	TipoConsumo	1.0098	1	1	2	1	0.0985	9.7572	0
4	Tension	1.0024	1	1	2	1	0.0494	4.9326	0
5	FabricanteM...	10.9173	8	8	26	1	6.5938	60.3973	0
6	TipMedicion	NaN	NaN	5	10	1	NaN	NaN	122
7	GrupoDeCon...	1.1035	1	1	5	1	0.5532	50.1353	0
8	Fases	1.6895	2	2	3	1	0.4668	27.6280	0
9	MesesAdeu...	1.1310	1	1	30	0	1.9719	174.3443	0
10	Deuda	11.3158	3.8800	0	581.8300	-5.9400	28.5696	252.4759	0
11	PromedioFac...	10.5755	6.3900	3.5600	321.9600	0	17.5900	166.3276	0
12	ValorUltimaF...	10.4714	6.0100	3.7300	337.6400	1.6700	17.5430	167.5324	0
13	ConsumoKW...	NaN	NaN	0	2.6663e+03	-88	NaN	NaN	56

Figura A5. 8, Limpieza de datos

Al dar clic en “Generar Reporte”, aparece una pantalla como está en la Figura A5. 9. Como se observa, se tiene dos opciones:

- Al dar clic en “Datos “Anómalos”” se generará un archivo Excel que contendrá en la “Hoja1” aquellos sistemas de medición que tuvieron datos inexistentes y en la “Hoja2” sistemas de medición que presentaron datos erróneos.
- “Datos Limpios” genera un archivo Excel con los datos de sistemas de medición que no presentaron datos anómalos.

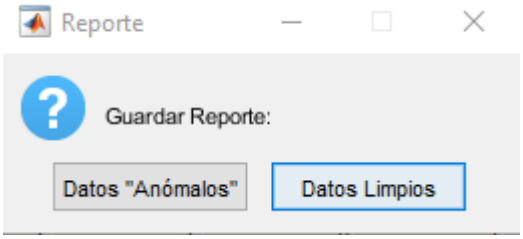


Figura A5. 9, Generar Reporte de Datos

A5.4. Procesamiento de datos

Con la matriz “Limpia”, se puede continuar dando clic en “Siguiente” (Figura A5. 8) y se muestra una pantalla como está en la Figura A5. 10, la cual pide elegir una técnica para la agrupación de datos. Como se observa estas técnicas son K-Medias y Coeficiente de Pearson.

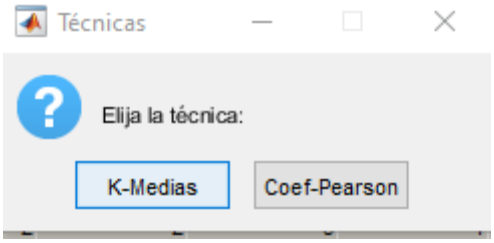


Figura A5. 10, Elección de técnica

Si se elige la opción “Coef-Pearson”, el programa dirige a la pantalla que se observa en la Figura A5. 11. De otra manera, si se elige la opción “K-Medias” se abre la pantalla que se observa en la Figura A5. 13.

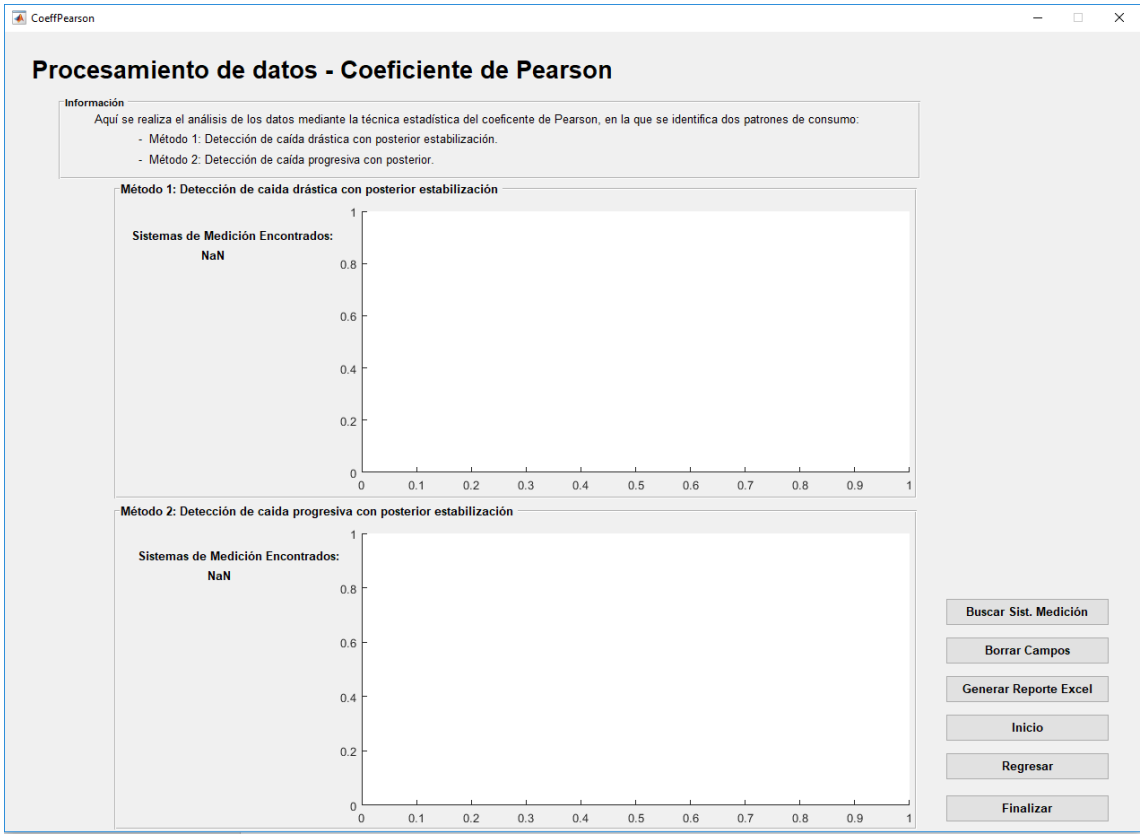


Figura A5. 11, Pantalla de Coeficiente de Pearson

Al dar clic en "Buscar Sist. Medición" (Figura A5. 11), el programa buscará sistemas de medición con patrones de consumo explicados en el capítulo 2. Como se aprecia en la Figura A5. 12, de la matriz cargada, encontró 14 sistemas de medición con el método 1 y 12 con el método 2. Los consumos registrados por estos sistemas se grafican.

El programa permite generar un reporte en Excel, en el que la "Hoja1" del archivo generado contendrá los sistemas de medición encontrados con el método 1 y, en la "Hoja2" los sistemas de medición encontrados con el método 2.

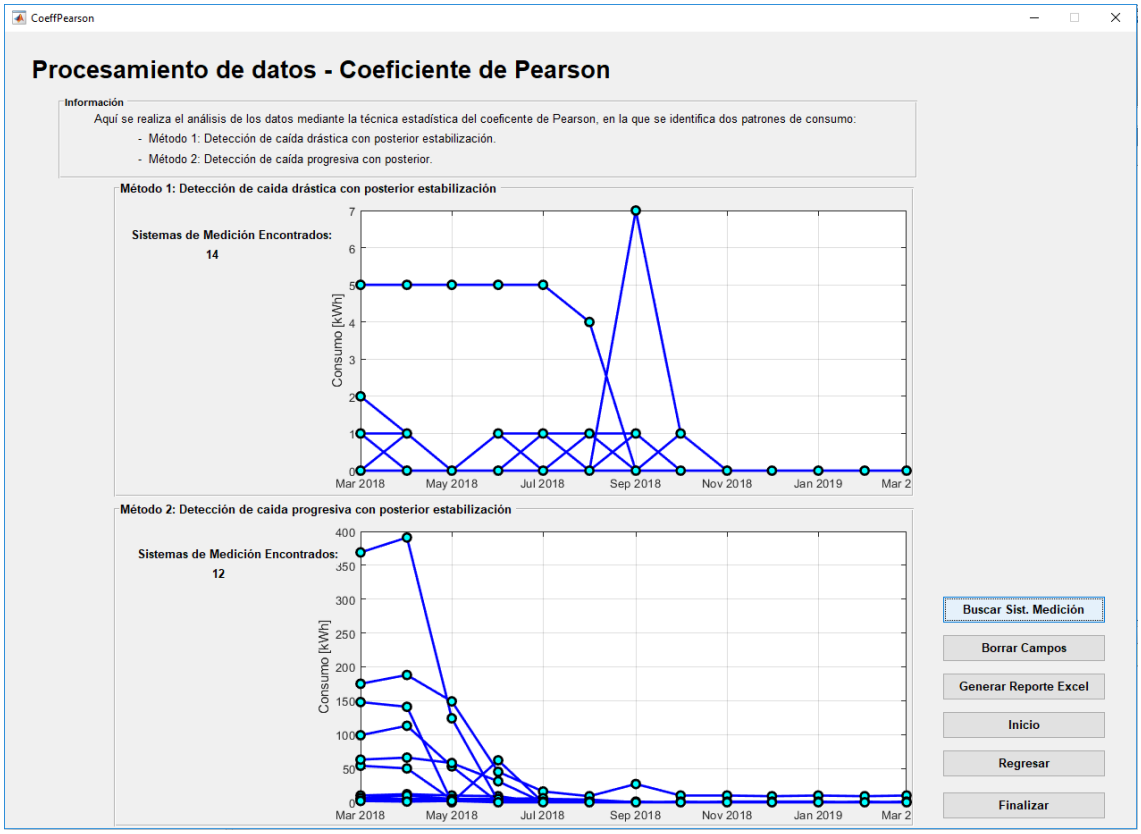


Figura A5. 12, Resultados de agrupación por el coeficiente de Pearson

El programa de "Agrupamiento K-Medias" (Figura A5. 13), permite realizar el agrupamiento de los sistemas de medición, para esto será necesario ingresar el número requerido de grupos K.

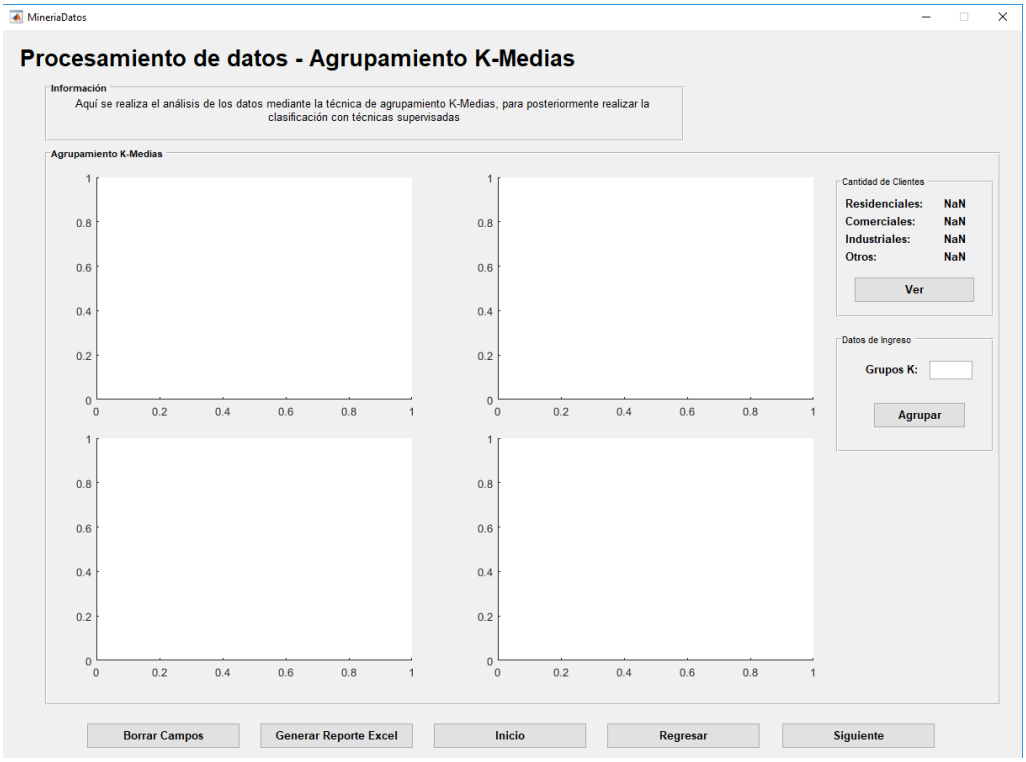


Figura A5. 13, Pantalla de agrupamiento K-Medias

Para agrupar los clientes, dar clic en “Agrupar” y el programa agrupará a los sistemas de medición. Al dar clic en “Ver” mostrara la cantidad de sistemas de medición pertenecientes a cada grupo de consumo, esto en la Figura A5. 14.

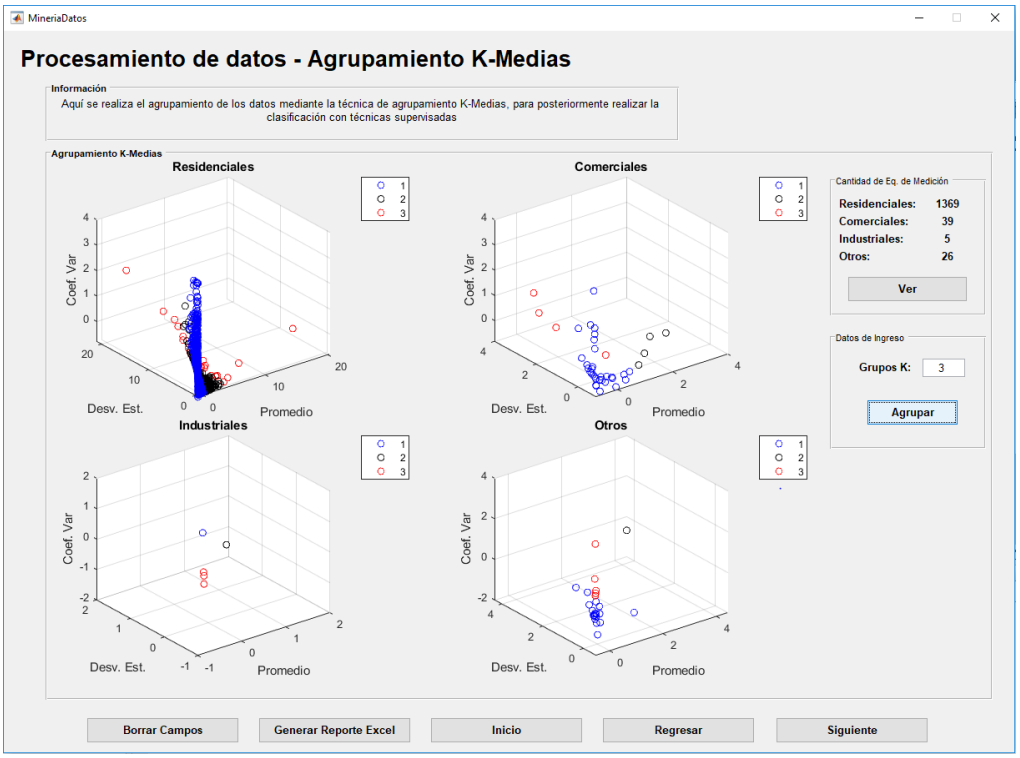


Figura A5. 14, Resultados de agrupamiento por K-Medias

El programa permite generar un reporte en archivo Excel (Figura A5. 15); en la Hoja 1 del archivo Excel estarán los clientes residenciales, en la Hoja 2 los comerciales, en la Hoja 3 los Industriales y en la Hoja 4 “Otros”.

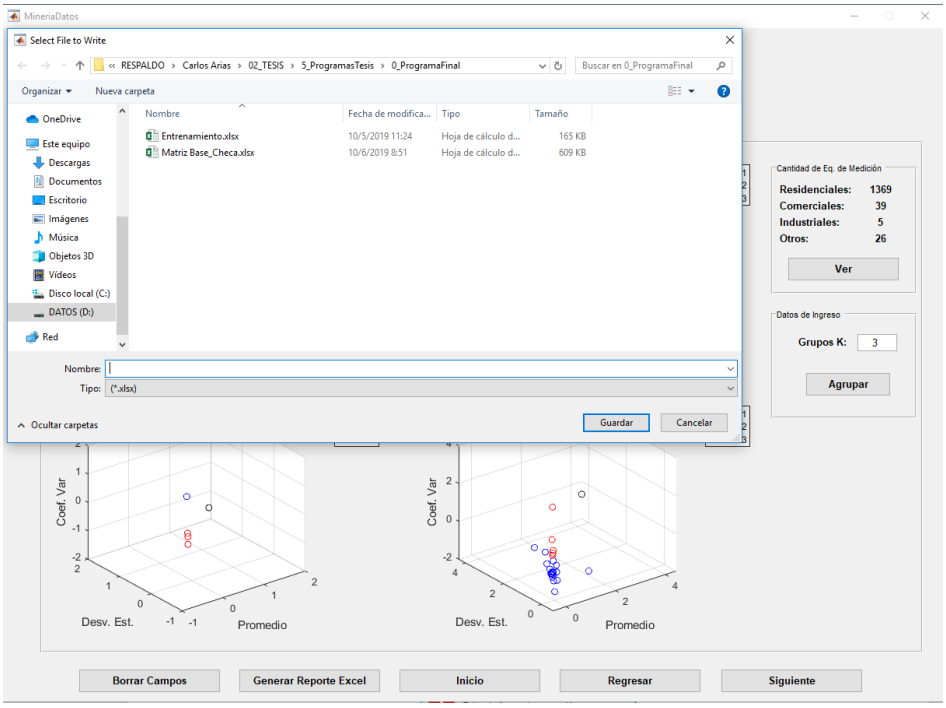


Figura A5. 15, Generación de reporte - K-Medias

En caso de dar clic en "Agrupar" sin ingresar el valor K, aparecerá un mensaje de error como está en la Figura A5. 16.

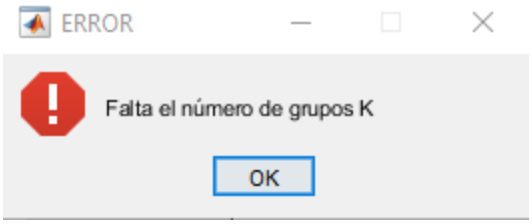


Figura A5. 16, Pantalla de Error

Si se ingresa un valor K mayor a 9, aparecerá un mensaje de advertencia como está en la Figura A5. 17.

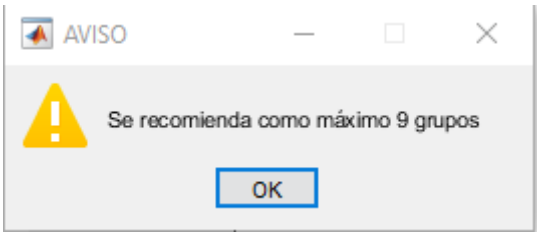


Figura A5. 17, Pantalla de advertencia

Ya realizado el análisis de los grupos K-Medias y luego de la elección de los grupos de aquellos sistemas de medición que posiblemente deben ser revisados en sitio, se puede realizar una clasificación de dichos sistemas mediante cualquiera de las técnicas supervisadas que se ve en la Figura A5. 18.

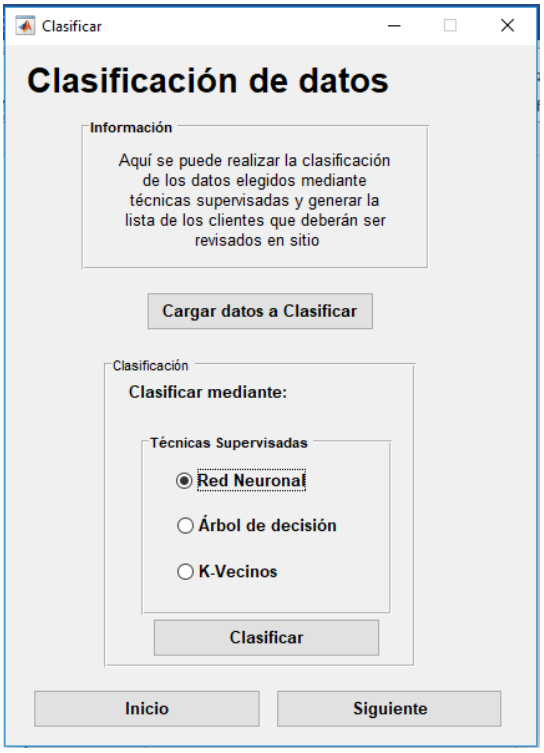


Figura A5. 18, Clasificación de datos

Antes de realizar la clasificación se tiene que cargar el archivo Excel que contiene los sistemas de medición de los grupos obtenidos por el agrupamiento K-Medias. Para esto, dar clic en "Cargar datos a Clasificar" y se abrirá la una pantalla como se muestra en la Figura A5. 19.

Figura A5. 19, Cargar datos a clasificar

Con los datos cargados, dar clic en "Clasificar" y el programa clasificará los sistemas de medición con la técnica elegida. Esto se puede observar en la Figura A5. 20.

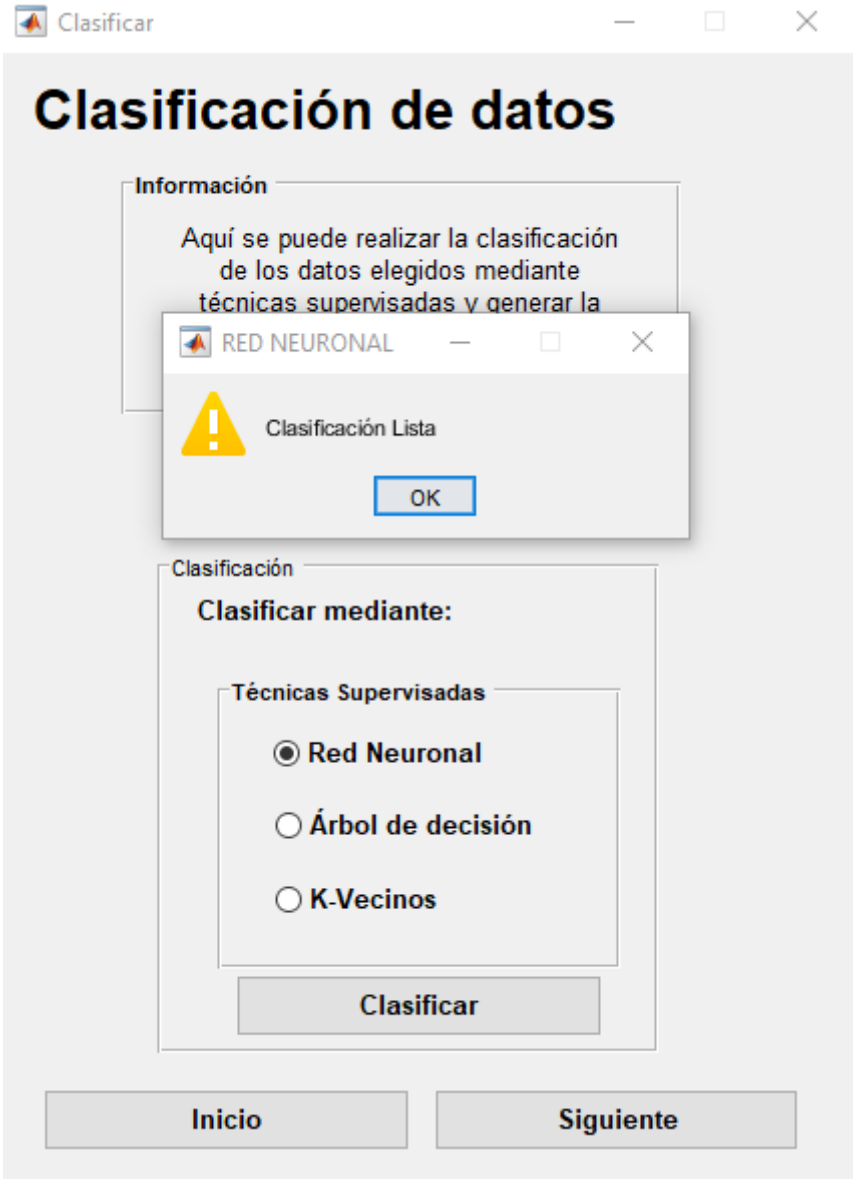


Figura A5. 20, Clasificación mediante Red Neuronal

Este programa también permitirá generar un reporte en Excel (Figura A5. 21); en este reporte estarán los sistemas de medición considerados para revisión en sitio.

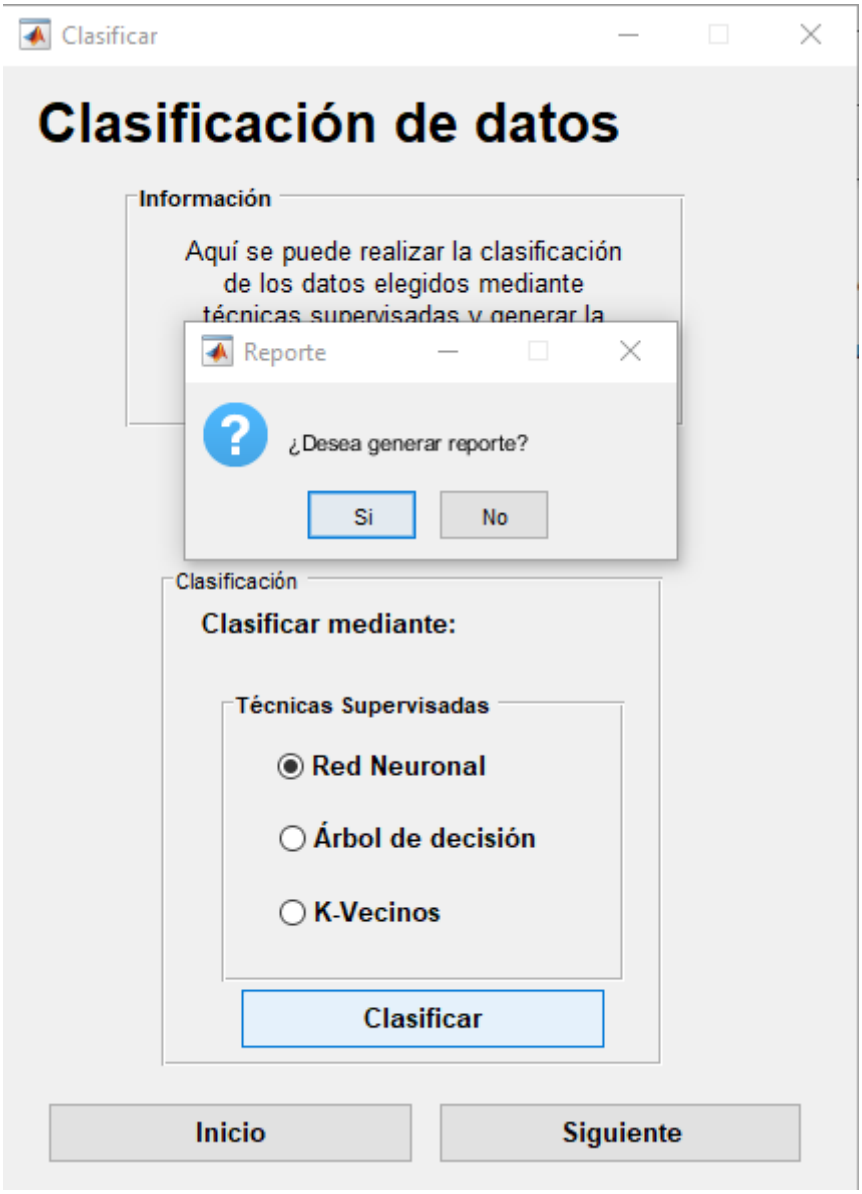


Figura A5. 21, Generar reporte de sistemas de medición "sospechosos"

A5.5. Análisis de resultados

En análisis de resultados (Figura A5. 22), se puede observar los resultados obtenidos de la minería de datos. También se puede graficar el patrón de consumo de cualquiera de los sistemas de medición que el algoritmo consideró como "sospechoso".

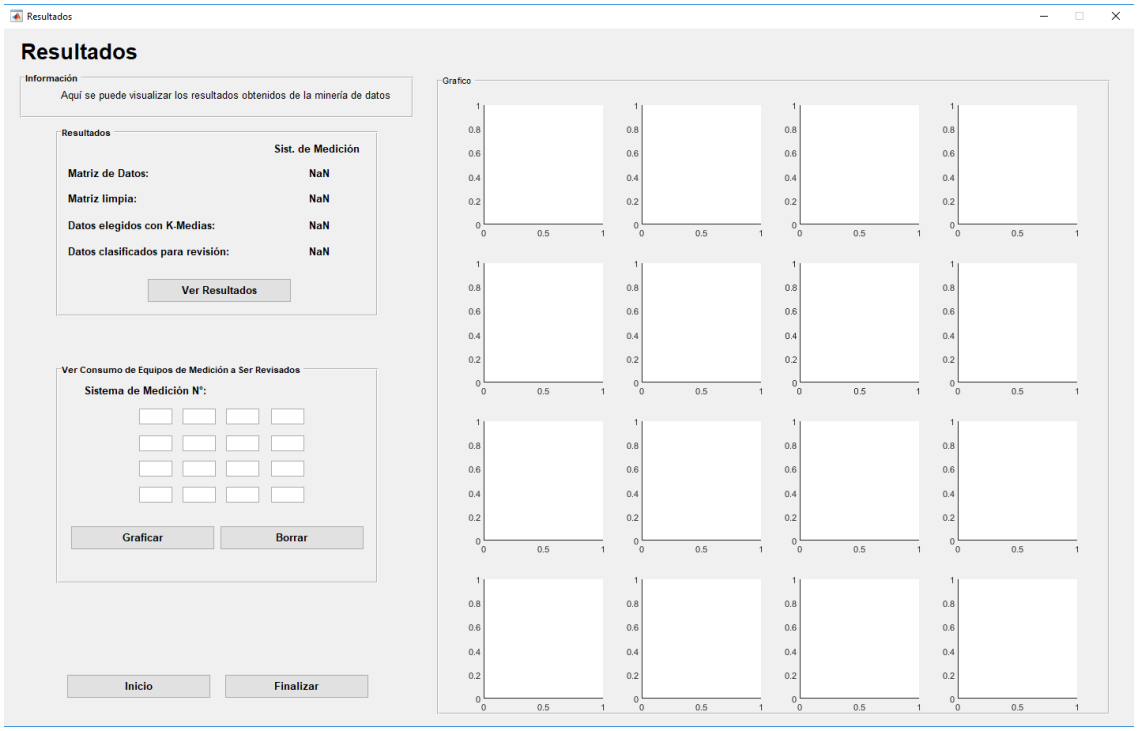


Figura A5. 22, Pantalla para Análisis de resultados

En "Ver Resultados", aparecerá los datos que se observa en la Figura A5. 23. Como análisis del ejemplo, se aprecia que, al iniciar el proceso se cargó un archivo con 1633 sistemas de medición; una vez limpia la base, redujo este dato a 1439.

Con la agrupación K-Medias y eligiendo los grupos adecuados de sistemas de medición, se obtuvo un total de 23. Después de esto se clasifica mediante la red neuronal, la cual clasifico a 18 para revisión en sitio.

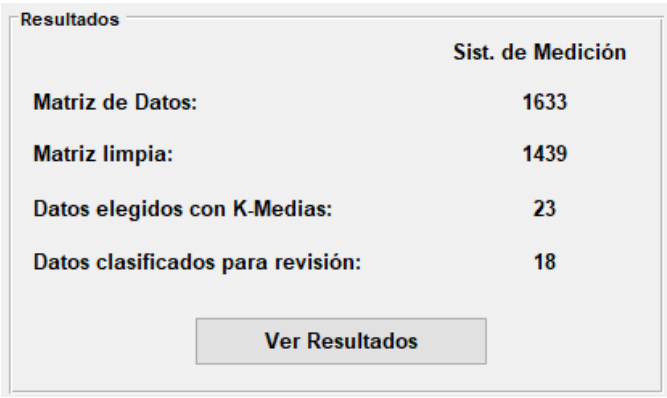


Figura A5. 23, Resultados

Como se presenta en la Figura A5. 24, se puede realizar el gráfico del consumo de cualquiera de los sistemas de medición clasificados como “sospechosos”.

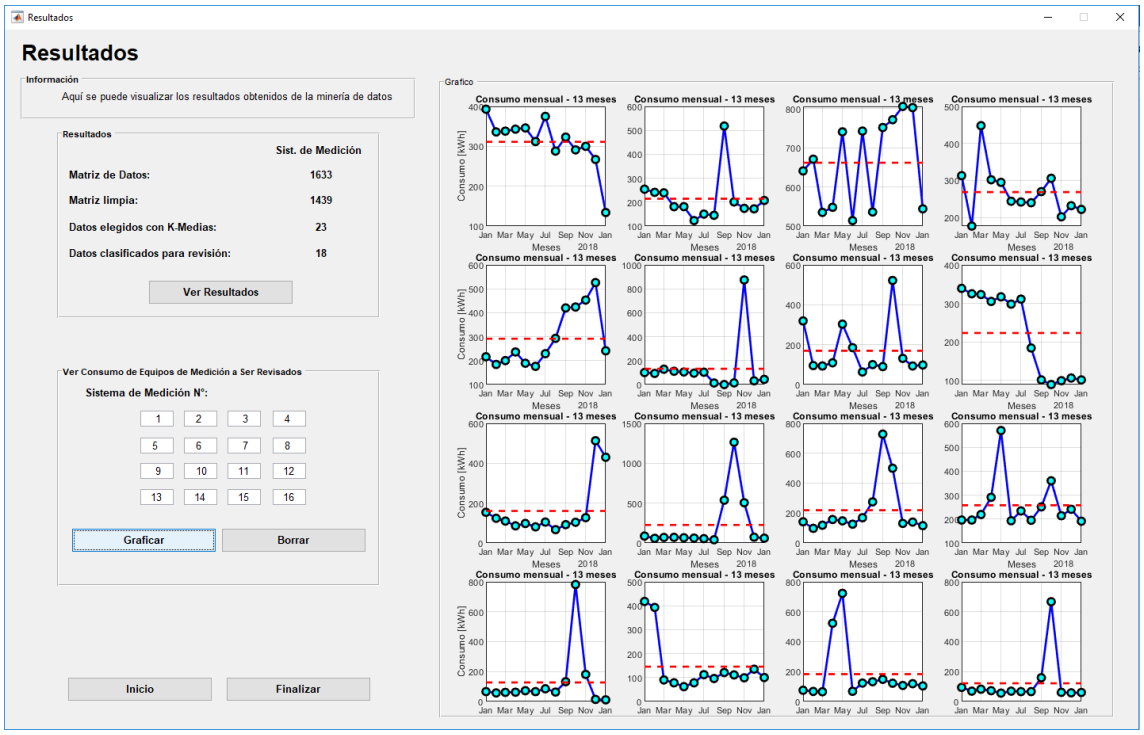
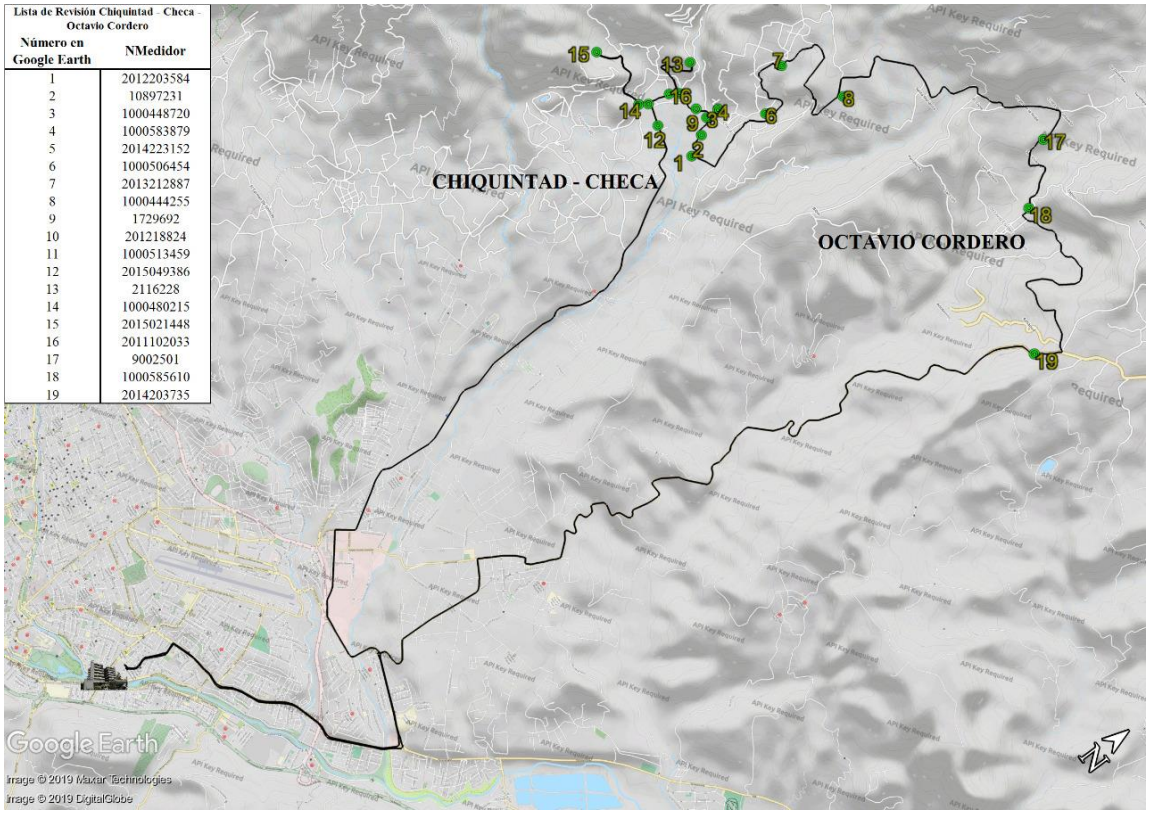


Figura A5. 24, Consumo de sistemas de medición “sospechosos”

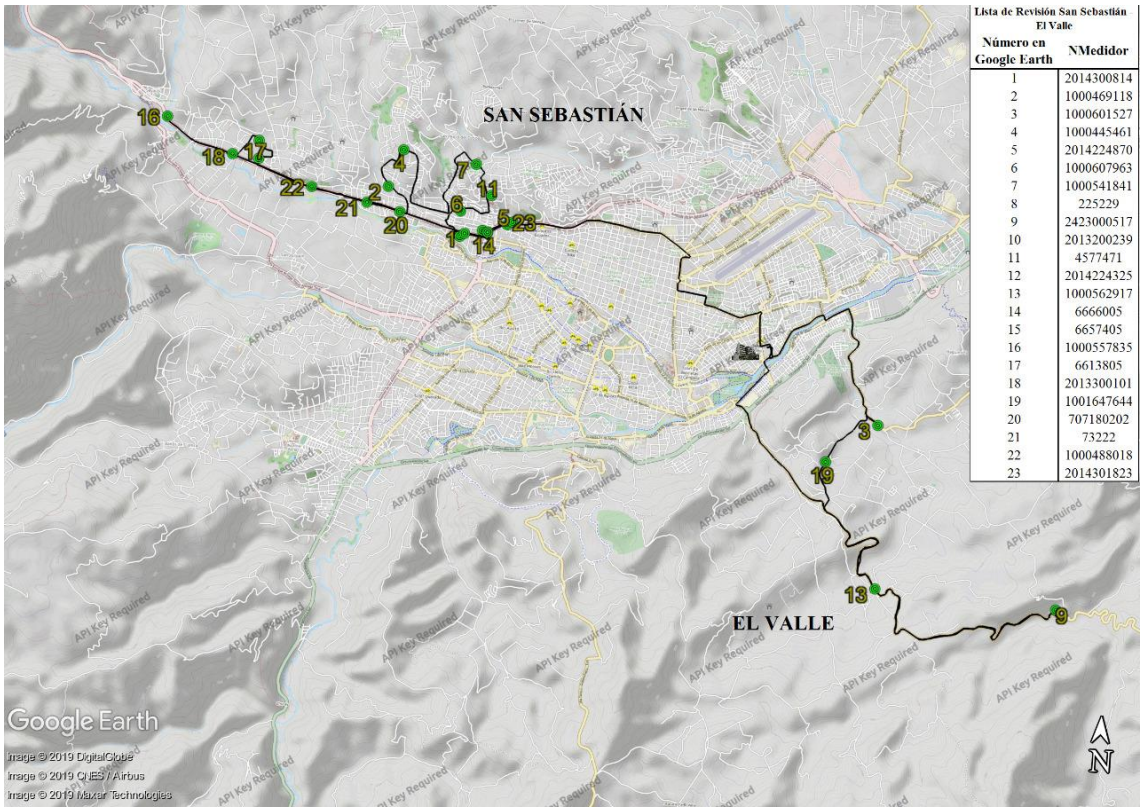
A6. ANEXO 6 – RUTAS PARA REVISIONES EN CAMPO

A continuación, se presenta las rutas planificadas de las revisiones que se hicieron en sitio:



Lista de Revisión Chiquintad – Checa – Octavio Cordero	
Número en Google Earth	NMedidor
1	2012203584
2	10897231
3	1000448720
4	1000583879
5	2014223152
6	1000506454
7	2013212887
8	1000444255
9	1729692
10	201218824
11	1000513459
12	2015049386
13	2116228
14	1000480215
15	2015021448
16	2011102033
17	9002501
18	1000585610
19	2014203735

Figura A6. 1, Ruta para L1.
Fuente: Google Earth



Lista de Revisión San Sebastián – El Valle	
Número en Google Earth	NMedidor
1	2014300814
2	1000469118
3	1000601527
4	1000445461
5	2014224870
6	1000607963
7	1000541841
8	225229
9	2423000517
10	2013200239
11	4577471
12	2014224325
13	1000562917
14	6666005
15	6657405
16	1000557835
17	6613805
18	2013300101
19	1001647644
20	707180202
21	73222
22	1000488018
23	2014301823

Figura A6. 2, Ruta para L2
Fuente: Google Earth

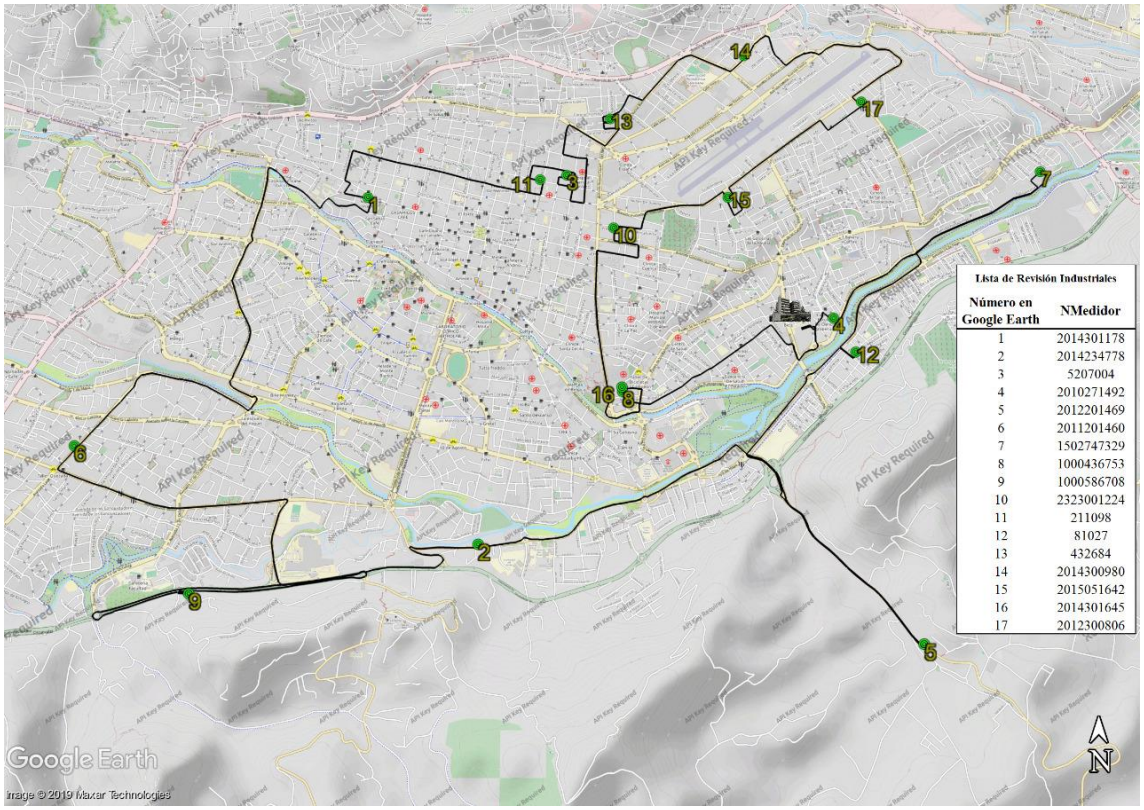


Figura A6. 3, Ruta para L3
Fuente: Google Earth

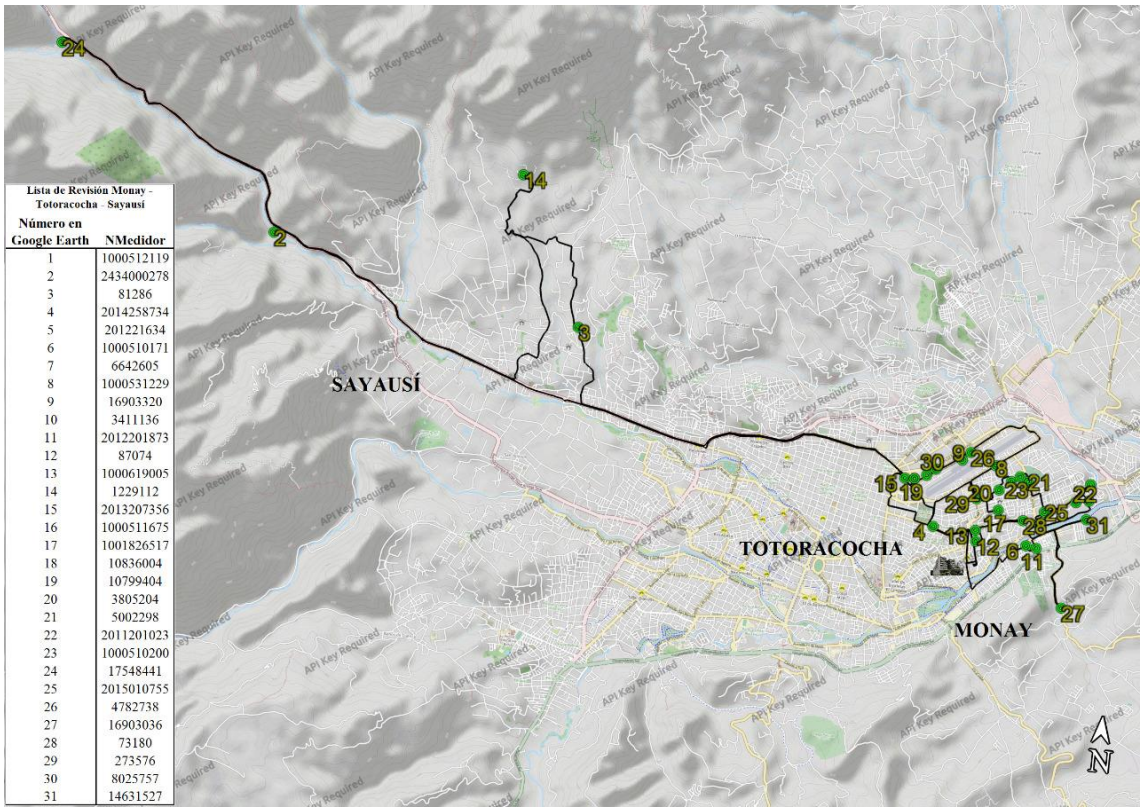


Figura A6. 4, Ruta para L6
Fuente: Google Earth

A7. ANEXO 7 – FOTOGRAFÍAS DE REVISIONES REALIZADAS EN CAMPO

Se presenta fotografías de algunas novedades que se encontraron en las revisiones que se realizaron en sitio:

Conexión directa desde barras:

Se encontró que la línea estaba conectada directo a las barras. Se sanciona al cliente con una refacturación, además el medidor se encontraba dentro de la propiedad y se pide modificar la ubicación del mismo.



Medidor con batería baja:

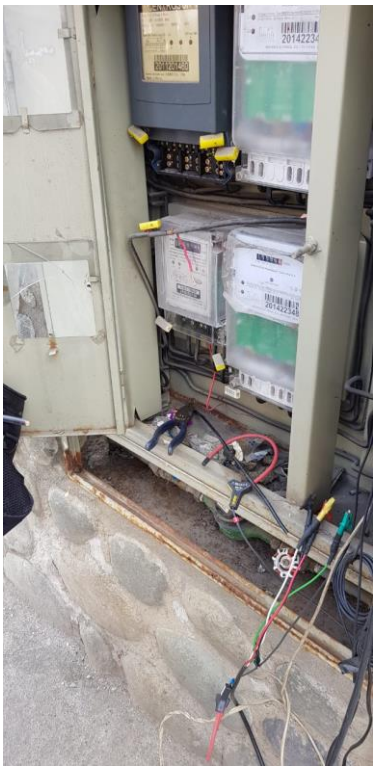
El medidor necesitaba un cambio de batería. La batería tiene que ser cambiada en el Laboratorio de medidores de la Empresa Distribuidora.

Medidor dañado:

El disco del medidor no giraba, por ende, no registraba energía.

Medidor mal instalado:

Las fases estaban mal instaladas en el medidor. En este caso se realiza la corrección de la instalación.



Reflector conectado directamente a la red de la empresa

Se encontró un reflector de 200W con conexión directa a la red de la Empresa Distribuidora.