UNIVERSIDAD TÉCNICA ESTATAL DE QUEVEDO

FACULTAD DE CIENCIAS DE LA INGENIERÍA

CARRERA INGENIERÍA EN ELECTRICIDAD

Proyecto de Investigación previo a la obtención del Título de Ingeniero Eléctrico.

Título del Proyecto de Investigación

"PREDICCIÓN DEL HURTO DE ENERGÍA ELÉCTRICA A TRAVÉS DEL USO DE LA INTELIGENCIA ARTIFICIAL MEDIANTE ALGORITMOS DE MACHINE LEARNING PARA CNEL EP UNIDAD DE NEGOCIOS SANTO DOMINGO."

AUTORES:

MACAO SÁNCHEZ RICHARD ALEX

PUJOTA CUASAPAZ EDISON JAVIER

DIRECTOR:

ING. ORTIZ GONZALEZ YADYRA MONSERRATH MSc.

QUEVEDO – LOS RÍOS – ECUADOR

2022

DECLARACIÓN Y AUTORÍA Y CESIÓN DE DERECHOS

Yo, **Richard Alex Macao Sánchez** declaro desenvueltamente que el trabajo aquí descrito es de mi autoría; que no ha sido previamente presentado para ningún grado o calificación profesional; y que he consultado las referencias bibliográficas que se incluyen en este documento.

La Universidad Técnica Estatal de Quevedo puede hacer uso de los derechos correspondientes a este trabajo, según establecido por la ley de Propiedad Intelectual, por su Reglamento y por la normativa institucional vigente.

Firma: _____

Richard Alex Macao Sánchez

C.C.: 171657529-3

DECLARACIÓN Y AUTORÍA Y CESIÓN DE DERECHOS

Yo, **Edison Javier Pujota Cuasapaz** declaro desenvueltamente que el trabajo aquí descrito es de mi autoría; que no ha sido previamente presentado para ningún grado o calificación profesional; y que he consultado las referencias bibliográficas que se incluyen en este documento.

La Universidad Técnica Estatal de Quevedo puede hacer uso de los derechos correspondientes a este trabajo, según establecido por la ley de Propiedad Intelectual, por su Reglamento y por la normativa institucional vigente.

Firma: _____

Edison Javier Pujota Cuasapaz

C.C.: 230051922-6

CERTIFICACIÓN DEL DIRECTOR DE TESIS

El suscrito, Ortiz González Yadyra Monserrath Docente de la Universidad Técnica Estatal de Quevedo, certifica que el Egresado Sr, Macao Sánchez Richard Alex y Pujota Cuasapaz Edison Javier, quienes realizarón la tesis de grado titulada **"PREDICCIÓN DEL HURTO DE ENERGÍA ELÉCTRICA A TRAVÉS DEL USO DE LA INTELIGENCIA ARTIFICIAL MEDIANTE ALGORITMOS DE MACHINE LEARNING PARA CNEL EP UNIDAD DE NEGOCIOS SANTO DOMINGO."** Previo a la obtención del título de Ingeniero Eléctrico, bajo mi dirección, habiendo cumplido con las disposiciones reglamentarias establecidas para el efecto.

Firmado electrónicamente por:
YADYRA MONSERRATH
ORTIZ GONZALEZ

ING. ORTIZ GONZALEZ YADYRA MONSERRATH MSc.

DIRECTOR DE TESIS

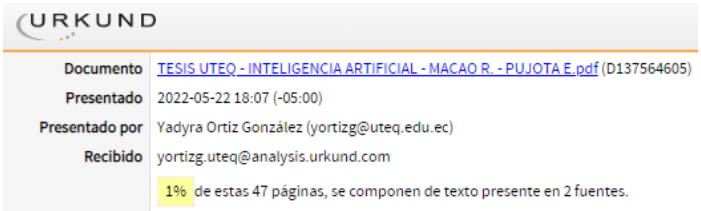
CERTIFICADO DEL REPORTE DE LA HERRAMIENTA DE PRECAUCIÓN DE COINCIDENCIAS Y/O PLAGIO ACADÉMICO

Sr. Ing. Washington Chiriboga Casanova, MSc
DECANO DE LA FACULTAD CIENCIAS DE LA INGENIERIA DE LA UTEQ
En su despacho.
De mi consideración.

En calidad de director del trabajo de investigación titulado: **"PREDICCIÓN DEL HURTO DE ENERGÍA ELÉCTRICA A TRAVÉS DEL USO DE LA INTELIGENCIA ARTIFICIAL MEDIANTE ALGORITMOS DE MACHINE LEARNING PARA CNEL EP UNIDAD DE NEGOCIOS SANTO DOMINGO",** me permito manifestar a usted lo siguiente:

Los señores **RICHARD ALEX MACAO SÁNCHEZ, EDISON JAVIER PUJOTA CUASAPAZ,** estudiantes de la carrera de Ingeniería en Electricidad modalidad presencial del paralelo A, han cumplido con las correcciones pertinentes e ingresado su trabajo de investigación al sistema URKUND, tenga bien certificar la siguiente información sobre el informe del sistema reflejando con un porcentaje favorable del 1%, cumpliendo con el reglamento de graduación de Estudiantes de Pregrado y la Normativa establecida por la Universidad.

Por la aprobación que se sirva de dar a la presente, quedo ante usted muy agradecida.


URKUND

Documento	TESIS UTEQ - INTELIGENCIA ARTIFICIAL - MACAO R. - PUJOTA E.pdf (D137564605)
Presentado	2022-05-22 18:07 (-05:00)
Presentado por	Yadyra Ortiz González (yortizg@uteq.edu.ec)
Recibido	yortizg.uteq@analysis.urkund.com
	1% de estas 47 páginas, se componen de texto presente en 2 fuentes.

Atentamente


Firmado electrónicamente por:
YADYRA MONSERRATH ORTIZ GONZALEZ

Ing. Yadyra Monserrath Ortiz González, MSc.
DIRECTOR DEL PROYECTO DE INVESTIGACIÓN.

DIRECTOR DEL PROYECTO DE INVESTIGACIÓN.

TÍTULO

"PREDICCIÓN DEL HURTO DE ENERGÍA ELÉCTRICA A TRAVÉS DEL USO DE LA INTELIGENCIA ARTIFICIAL MEDIANTE ALGORITMOS DE MACHINE LEARNING PARA CNEL EP UNIDAD DE NEGOCIOS SANTO DOMINGO."

Presentado al Consejo Directivo como requisito previo a la obtención del título de Ingeniero Eléctrico.

Aprobado por:

JORGE PATRICIO MURILLO OVIEDO

Firmado digitalmente por JORGE PATRICIO MURILLO OVIEDO
Nombre de reconocimiento (DN): c=EC, o=BANCO CENTRAL DEL ECUADOR, ou=ENTIDAD DE CERTIFICACION DE INFORMACION-ECIBCE, l=QUITO, serialNumber=0000477687, cn=JORGE PATRICIO MURILLO OVIEDO
Fecha: 2022.07.05 17:18:16 -05'00'

PRESIDENTE DEL TRIBUNAL

Ing. Jorge Murillo Oviedo, PhD.

Firmado electrónicamente por:
MILTON GEOVANNY CUENCA CABRERA

MIEMBRO DEL TRIBUNAL

Ing. Milton Cuenca Cabrera, MSc.

ANDRES ALEXANDER DE LA TORRE MACIAS

Firmado digitalmente por ANDRES ALEXANDER DE LA TORRE MACIAS
DN: cn=ANDRES ALEXANDER DE LA TORRE MACIAS, c=EC, ou=Certificado de Clase 2 de Persona Fisica EC, email=andres.delatorrem@gmail.com
Fecha: 2022.07.05 14:08:13 -05'00'

MIEMBRO DEL TRIBUNAL

Ing. Andrés De la Torre Macías, MSc.

QUEVEDO – LOS RÍOS – ECUADOR

2022

AGRADECIMIENTO

Agradezco a Dios por regalarme la vida y haberme permitido llegar a culminar los estudios superiores.

Este logro se lo debo a mi compañera de toda la vida, a mi esposa a la **Sra. Angelica María Ortega Fuentes**, destacando su paciencia su amor y el deseo de superación en el hogar.

Al **Ing. Romel Analuiza** jefe del departamento de Control de pérdidas de energía **CNEL EP Santo Domingo**, por habernos facilitado los datos para poder realizar el trabajo de investigación.

Autor: Richard Alex Macao Sánchez.

AGRADECIMIENTO

A lo largo de mi carrera eh conocido decenas de personas, pero las que siempre han estado allí son mi familia, en primera instancia están mis padres, **Angel Pujota y Lourdes Cuasapaz**, que en todo momento me han apoyado de manera incondicional en cualquier circunstancia, tanto moralmente como económicamente, al igual que mis hermanas, **Marjorie, Evelyn y Mailin,** que para bien siempre estuvieron presentes en cada paso que daba en mi carrera con su apoyo leal y entusiasta, mis hermanos de igual manera alentándome a que sea el mejor, sin dejar a un lado también a mis sobrinas y a mi cuñado que me han visto dar un paso a la vez en mi carrera.

Gracias a todos ellos y a mi novia que está desde la mitad de mi carrera a mi lado, me ha dado palabras de aliento, ha estado en mis momentos de tristeza y también de alegría, por eso hoy por hoy me encuentro en esta etapa de mi vida, en la que, a pesar de los muchos obstáculos, cada uno logre atravesar y sobresalir para cumplir con mi objetivo, además de ser una alegría para mí, sé que también es para todas las personas que me apoyaron y estuvieron para mí en todo momento.

Autor: Edison Javier Pujota Cuasapaz.

DEDICATORIA

Este trabajo de investigación se lo de dedico a mis hijos **Jhon Macao y Angie Macao**, quienes fueron mi inspiración, mi motivación para no claudicar y seguir siempre adelante, que la vida me ha enseñado que no siempre es necesario tener una ruma de dinero para cumplir los sueños, tan solo basta tener a la persona ideal a su lado que tenga sus mismos anhelos de superación en beneficio de una empresa que se llama familia.

.

Autor: Richard Alex Macao Sánchez.

DEDICATORIA

La dedicación y las ganas de seguir y ser alguien mejor en la vida está dando frutos, por eso dedico este trabajo de investigación a mis padres, **Angel Pujota y Lourdes Cuasapaz,** quienes siempre han confiado en mí de una u otra manera me han brindado su apoyo para que consiga culminar mi carrera universitaria.

De igual manera dedico este trabajo a todos los integrantes que conforman la **Familia Pujota**, para que sea una inspiración para ellos y puedan superarse día a día, con las ganas de cumplir sus objetivos al igual que yo los estoy cumpliendo.

Esto también me lo dedico a mí mismo, porque siempre eh tratado de dar lo mejor de mí, sin decepcionar a nadie y darles una alegría más por mi triunfo a todos los que han estado en mi etapa de estudiante, esperando que lo que eh logrado sea una motivación para todo aquel que quiera superarse en la vida.

Autor: Edison Javier Pujota Cuasapaz.

RESUMEN EJECUTIVO.

Las pérdidas de energía eléctrica son un problema que no se han logrado reducir en su totalidad en los sistemas eléctricos de potencia. Estas pérdidas se pueden presentar durante las etapas de Generación, Transmisión o Distribución de energía eléctrica. El presente proyecto de investigación se desarrolló en el área de distribución eléctrica, con el objetivo de analizar las razones y posibles soluciones que generan pérdidas no técnicas en las unidades de negocios, mediante la implementación de software inteligente.

Las pérdidas técnicas se producen desde la etapa de generación eléctrica y representan la energía que no es aprovechada o que se pierde durante la transmisión, subtransmisión y distribución de energía eléctrica las cuales son ocasionadas debido al efecto Joule, corrientes Foucault o histéresis. No obstante, este tipo de pérdidas no pueden ser eliminadas en su totalidad, debido a los fenómenos fisicoquímicos que se presentan en los núcleos ferromagnéticos, calentamiento en los conductores, que no pueden ser eliminados.

Las pérdidas de tipo no técnicas en el área de distribución se generan por las siguientes razones: 1.- Administrativas: Usuarios sin medición de consumo, lecturas erróneas, sistemas informáticos imprecisos y cultura del “no pago”. 2.- Operación deficiente: Mantenimiento deficiente, descalibración del equipo de medición y desbalance entre fases. 3.- Fraudulentas: Acometidas no autorizadas, conexión de carga antes del medidor, intervención en la base de datos, medidor intervenido o manipulado siendo como tal "hurto de energía", resultando esta última, irrecuperable para la unidad de negocio.

Reducir las pérdidas técnicas, implica una gran inversión económica ya que se debería realizar un redimensionamiento y reemplazo de conductores y transformadores. Las empresas distribuidoras de energía eléctrica han buscado eliminar las pérdidas no técnicas implementando acometidas antihurto, reemplazando los medidores electromecánicos por medidores digitales, sin embargo, ciertos usuarios utilizan diferentes métodos, dispositivos o mecanismos para vulnerar dichas acometidas y aparatos de medición, con el objetivo de disminuir fraudulentamente su consumo eléctrico, y de esta forma evitar el pago de la factura eléctrica total o parcialmente.

En la actualidad la Inteligencia Artificial (IA), se usa en un amplio rango de áreas, una de ellas es el sector eléctrico. Por medio de la IA, se puede analizar gran cantidad de datos obtenidos en las etapas de generación, transmisión, subtransmisión y distribución de energía eléctrica, mediante la IA se puede encontrar una solución óptima a los problemas de una manera lógica y razonable.

El presente proyecto de investigación fue aplicado al área de distribución, para analizar mediante algoritmos de Machine Learning (ML), el comportamiento de las pérdidas no técnicas en los diferentes tipos de usuarios de la unidad de negocios Santo Domingo con el objetivo de encontrar posibles infractores (usuarios que hurtan de energía).

Palabras claves: Pérdidas técnicas y no técnicas, inteligencia artificial, machine learning, hurto de energía, algoritmos, aprendizaje de máquina.

ABSTRACT.

Electrical energy losses are a problem, which have not been completely resolved in electric power systems. These losses may happen during the stages of generation, transmission and distribution of electrical energy. This research project was developed in the area of electrical distribution, with the aim of analyzing the reasons and possible solutions that generate non-technical losses, through the implementation of intelligent software.

Technical losses occur from generation and they represent the energy that is not used or that is lost during transmission, sub-transmission and distribution, caused by the Joule effect, Eddy currents (Foucault's currents) or hysteresis. However, this type of loss cannot be completely eliminated, due to the physicochemical phenomena that occur in ferromagnetic nuclei, heating in the conductors, which cannot be eliminated.

Non-technical losses in the distribution of energy are generated due to the following reasons: 1.- Administrative: Users without metering, erroneous readings, inaccurate computer systems and a culture of "non-payment". 2.- Poor operation: Poor maintenance, poor accuracy of measurement equipment and phase unbalance. 3.- Fraudulent: Unauthorized connections, intervention in the database, meter intervened or targeted, being as such "energy theft".

Reducing technical losses implies a large economic investment since a resizing of conductors and transformers should be carried out. Electricity distribution companies have sought to eliminate non-technical losses by implementing anti-theft connections, replacing electromechanical meters with digital meters, but people use any method, device or mechanism to violate those connections and measurement devices, to seek their personal good that is, not paying in full monthly electricity consumption.

Currently Artificial Intelligence (AI) is used in a wide range of areas, one of them is the electricity sector. Through AI, a large amount of data obtained in the stages of generation, transmission, sub-transmission and distribution of electric power can be analyzed, and with help of AI algorithms an optimal solution to problems can be found in a logical and reasonable way.

This research project was applied in the distribution area, to analyze through Machine Learning (ML) algorithms, the behavior of non-technical losses in the different types of users of the Santo Domingo electrical company with the aim of finding possible offenders (users who steal energy).

Keywords: Technical and non-technical losses, artificial intelligence, machine learning, energy theft, algorithms, machine learning.

Revisión:

MSc. Gardenia Vélez

1709104234

Registro senescyt 1025-12-86029008

TABLA DE CONTENIDO.

ÍNDICE DE TABLAS.

ÍNDICE DE ECUACIONES.

ÍNDICE DE FIGURAS.

ABREVIATURAS

CNEL.- Corporación Nacional De Electricidad.

ARCONEL.- Agencia De Regulación Y Control De Electricidad.

S.I.N.- Sistema Nacional Interconectado

LOSPEE.- Ley Orgánica Del Servicio Público De Energía Eléctrica

ML.- Machine Learning.

AI. Inteligencia Artificial.

MWh. - Megavatios hora

ADS. - Conjunto de datos analíticos

SIG.- Sistema de información geográfica

KDD. - Knowledge Discovery in database

CO.- Comerciales

RD. - Residenciales

BP. - Beneficiario Publico baja tensión

AS.- Asistencia social baja tensión

RH. -Bombeo de agua MT SE

CR.-Culto religioso baja tensión

CH. -Comercial Media tensión con demanda

ID. -Industrial media tensión

TE. - Tercera edad

RP. - Residencial PEC

PEC. - Programa de energía cocción eficiente

DP. - Discapacidad PEC

DS. - Discapacidad

OF. - Entidades Oficiales

MU. - Entidad Municipal

AP.- Alumbrado Publico

CÓDIGO DUBLÍN

Titulo	PREDICCIÓN DEL HURTO DEL HURTO DE ENERGÍA ELÉCTRICA A TRAVÉS DEL USO DE LA INTELIGENCIA ARTIFICIAL MEDIANTE ALGORITMOS DE MACHINE LEARNING PARA CNEL EP UNIDAD DE NEGOCIOS SANTO DOMINGO			
Autores	Richard Alex Macao Sánchez, Edison Javier Pujota Cuasapaz			
Palabras clave	Pérdidas técnicas y no técnicas	Inteligencia Artificial	Hurto de energía eléctrica	Aprendizaje de máquina
Fecha de publicación:	1 de julio del 2022			
Editorial	UTEQ			
Resumen	El presente proyecto de investigación se desarrolló en el área de distribución eléctrica, con el objetivo de analizar las razones y posibles soluciones que generan pérdidas no técnicas en las unidades de negocios, mediante la implementación de software inteligente. En la actualidad la Inteligencia Artificial (AI), se usa en un amplio rango de áreas, una de ellas es el sector eléctrico. Por medio de la AI, se puede analizar gran cantidad de datos obtenidos en las etapas de generación, transmisión, subtransmisión y distribución de energía eléctrica, mediante la AI se puede encontrar una solución óptima a los problemas de una manera lógica y razonable.			
Abstract:	This research project was developed in the area of electrical distribution, with the aim of analyzing the reasons and possible solutions that generate non-technical losses, through the implementation of intelligent software. Currently Artificial Intelligence (AI) is used in a wide range of areas, one of them is the electricity sector. Through AI, a large amount of data obtained in the stages of generation, transmission, sub-transmission and distribution of electric power can be analyzed, and with help of AI algorithms an optimal solution to problems can be found in a logical and reasonable way.			
Descripción:	130 hojas: dimensiones, 29 x 21 cm +CD-ROM 6162			
URL:				

INTRODUCCIÓN.

La energía eléctrica es un pilar fundamental para el desarrollo tecnológico y socioeconómico de un país, ayudando a aumentar la producción industrial y comercial, en la actualidad la mayoría de productos son fabricados por industrias las cuales necesitan como fuente principal la energía eléctrica.

El servicio de energía eléctrica se divide en cuatro etapas: generación, transmisión, distribución y comercialización. La primera etapa se basa en el proceso mediante el cual se utilizan fuentes de energía primaria para convertirla en energía eléctrica, la segunda etapa es la parte del sistema que se encarga en transportar la energía eléctrica a grandes distancias a los puntos de consumo, la distribución por su lado es la cual lleva la energía eléctrica desde las subestaciones de las centrales hasta el consumidor final y la última etapa es la venta de la energía eléctrica a los clientes siendo estos industriales, comerciales y residenciales.

Todos los componentes instalados para el servicio de energía eléctrica sirven para suministrar, transferir y usar dicha energía, desde la etapa de generación hasta la etapa de comercialización, en donde se presentan pérdidas las cuales son; técnicas que se basa en el efecto Joule que es el calor que se produce cuando la corriente pasa a través de las redes eléctricas debido a las largas distancias y las pérdidas no técnicas son causadas por alguna actividad fraudulenta de usuarios que puedan o no estar conectados a la red tomando del sistema sin que el medido de energía registre su consumo, además puede ser por una incorrecta facturación o equipos de medición en mal estado o alterados.

Las pérdidas del tipo no técnicas tendieron al aumento en los últimos 5 años a nivel nacional, (según los reportes del informe estadístico anual y multianual del sector eléctrico ecuatoriano 2021). Este problema representa pérdidas millonarias al país y una de sus principales causantes es la conexión indebida de usuarios al sistema y los equipos de medición en controversia.

El hurto de energía eléctrica se presenta cuando el medidor no registra el consumo del usuario debido a equipos de medición en controversia o de los sistemas de facturación, se propone en este proyecto de investigación detectar a los posibles usuarios que hurtan energía eléctrica por medio del uso de algoritmos basados en inteligencia artificial.

Para la detección de los posibles usuarios que infringen pérdidas comerciales a la empresa distribuidora, se usan datos de los 12 últimos meses de consumo de energía eléctrica, datos que son provistos por el departamento de control de energía de la CNEL EP Santo Domingo, conformado por los cantones: El Carmen, Flavio Alfaro, La Concordia, Pedernales, Santo Domingo y la 14. Además, con la utilización del software ArcGIS, sectorizar a los posibles infractores para evidenciar en que sector existen mayores pérdidas por el hurto de energía.

CAPÍTULO I

CONTEXTUALIZACIÓN DE LA INVESTIGACIÓN

1.1. Problematización.

1.1.1. Diagnóstico.

En la provincia de Santo Domingo de los Tsáchilas, las pérdidas no técnicas tienen relevancia debido al hurto de energía por usuarios que se conectan a la red ilícitamente para disminuir su consumo, presentándose en zonas rurales y urbanas. En el cantón Flavio Alfaro, el cual posee 8261 clientes y una disponibilidad de energía de 22494 MWh, presenta un total de 7062 MWh en pérdidas de energía, lo cual indica que existe una pérdida del 31,4% de energía eléctrica, siendo éste el cantón en presentar mayor perjuicio económico para la unidad de negocios CNEL EP Santo Domingo durante el mes de junio del 2021.

Por ejemplo, en el cantón Santo Domingo, donde existen 148608 clientes con una disponibilidad de 478887 MWh, presenta pérdidas de 48090 MWh, lo cual representa el 10,0% de pérdidas de la energía eléctrica total, comparando con el cantón Flavio Alfaro, se toma como referencia estos dos cantones, en el cual se puede evidenciar que el cantón que tiene menor número de usuarios, presenta un mayor porcentaje de pérdidas de energía con respecto al cantón con mayor número de usuarios.

Considerando, que, si se habla de energía eléctrica en MWh entregada a los usuarios pertenecientes a cada cantón, Santo Domingo presenta 48090 MWh en pérdidas de energía siendo el 46.22%, a comparación con Flavio Alfaro que posee 7062 MWh con 6,79%, siendo así el cantón con mayor porcentaje, lo que implica pérdidas económicas importantes para la empresa distribuidora.

1.1.2. Formulación del problema de investigación.

Las pérdidas en el sector eléctrico desde sus inicios no se han logrado corregir en su totalidad, debido a que siempre existirán pérdidas durante la transmisión, subtransmisión, distribución y comercialización.

Para poder mitigar este tipo de pérdidas se debe tener en cuenta los siguientes parámetros como es: el estado actual del sistema, conocer la proyección de la carga y realizar estudios de flujos de carga para optimizar la operación de líneas y redes, siendo de esta manera que sólo se pueden reducir las pérdidas técnicas a través del mejoramiento de la red.

Las pérdidas no técnicas por su parte se basa en que no toda la energía eléctrica generada se vende y se factura, debido a que los dispositivos de medición no contabilizan como entregado a los usuarios y por lo tanto las empresas suministradoras de electricidad no reciben un pago por la prestación del servicio, además que los usuarios cometen actos fraudulentos manipulando los dispositivos de medición o en los sistemas de facturación, necesitando así realizar una inversión muy alta para minimizar estos actos ilícitos es por ellos que en este proyecto nos enfocamos en las pérdidas no técnicas basándonos en la inteligencia artificial para detectar a los usuarios que hurtan la energía eléctrica.

Las personas que cometen actos fraudulentos a las empresas distribuidoras, buscaran de una u otra manera su bien personal sin conocer las consecuencias legales que serán sancionadas por el delito de hurto a robo, en el artículo 8 de la ley de Régimen del Sector Eléctrico reformado mediante la ley 2006-65 así lo establece, además estarían limitando el crecimiento de la empresa en beneficio de los clientes [40].

1.1.3. Sistematización del problema.

¿Cómo ayudará la implementación de un algoritmo basado en inteligencia artificial en el control de pérdidas por hurto de energía?

¿Qué algoritmo de inteligencia artificial es el óptimo para analizar las pérdidas no técnicas en la CNEL EP de Santo Domingo de los Tsáchilas?

¿Determinar la categoría de los usuarios que incurren en hurtar energía en la CNEL EP de Santo Domingo de los Tsáchilas?

1.2. Objetivos.

1.2.1. Objetivo General.

Desarrollar un algoritmo basado en Inteligencia Artificial que permita detectar a los posibles clientes que hurtan energía eléctrica distribuida por la empresa CNEL EP Santo Domingo.

1.2.2. Objetivos Específicos.

- Desarrollar un algoritmo mediante técnicas de Machine Learning que permita detectar a los posibles infractores que hurtan energía.

- Aplicar el algoritmo de inteligencia artificial mediante el uso del software Python.

- Detectar los sectores donde se produce el hurto de energía mediante el software ArcGIS.

- Clasificar a los posibles infractores de acuerdo a la ubicación geográfica, basándose en datos reales obtenidos por la unidad de control de energía de la CNEL EP Santo Domingo.

1.3. Justificación.

El presente proyecto se realizará con la finalidad de identificar a los usuarios que hurtan energía eléctrica distribuida por la unidad de negocio CNEL EP Santo Domingo, usuarios que vulneran los instrumentos de medición y de esta manera ocasionan pérdidas económicas a la empresa eléctrica.

La inteligencia artificial ha tomado en los últimos años una gran relevancia en el sector eléctrico desarrollando una gran cantidad de aplicaciones como: Mantenimiento predictivo en donde la AI, mejora la eficiencia en el sector de energías renovables, además de reducir los costos de mantenimiento en sus instalaciones. Planificación y ajuste de oferta y demanda, la AI en este caso optimiza la capacidad de monitorización, operación y control de la red. Otra aplicación es la Relación 'Inteligente' con el consumidor, indica que la AI permite aportar nuevos servicios y capacidades Smart, la tecnología y el análisis y monitorización de sus consumos está aportando grandes ahorros en su factura eléctrica.

A través de algoritmos de inteligencia artificial se va detectar a los posibles usuarios infractores que eluden el pago por su consumo de energía eléctrica , para que la CNEL EP, realice una inspección a la instalación de los equipos de medición, en donde usuarios podrían haber realizado manipulaciones, con el objetivo de disminuir su consumo eléctrico fraudulentamente.

En el marco legal el hurto de energía es sancionado con el 300% del valor emitido por la factura real, se pretende llegar a los clientes y concienciar que, al pagar el valor real de su consumo, contribuye económicamente a la empresa distribuidora, permitiéndole desarrollar más proyectos eléctricos, además mejorando la calidad del servicio.

CAPÍTULO II

FUNDAMENTACIÓN TEÓRICA DE LA INVESTIGACIÓN

2.1. Marco Teórico.

2.1.1. ¿Qué es un algoritmo?

Se define un algoritmo como una secuencia de pasos ordenados y finito de operaciones, con una particularidad única puesto que nos permite hallar la solución de un problema, además que siguen una secuencia en base a los datos para cumplir un objetivo [1].

A lo largo del tiempo los algoritmos han ganado un amplio terreno en la humanidad, ya que ahora por medio de ellos, es posible desarrollar software que tengan la capacidad de realizar tareas complejas, por lo general dichas tareas antes eran asignadas a los seres humanos [1].

2.1.1.1. ¿Qué es un algoritmo inteligente?

Un algoritmo es una secuencia de pasos matemáticos que produce un resultado. Se utilizan para la exploración de datos, para descubrir ideas y patrones, para construir modelos, para generar predicciones, recomendaciones e incluso razonamiento y toma de decisiones automatizados. Los algoritmos utilizan técnicas avanzadas como: Big Data, inteligencia artificial, aprendizaje automático y automatización de procesos robóticos [2].

Datos → Información → Información → Inteligencia → Acción

2.1.2. Estadística.

La estadística es la ciencia que se ocupa del desarrollo y estudio de métodos para recopilar, analizar, traducir y brindar una presentación de datos de EMP Factos. La estadística nos permite cuantificar datos para analizarlos y mejorar la comprensión a partir de la información disponible. Al desarrollar métodos y estudiar la teoría que subyace a los métodos, la estadística recurre a una diversa metodología de fácil acceso de herramientas matemáticas y computacionales [3].

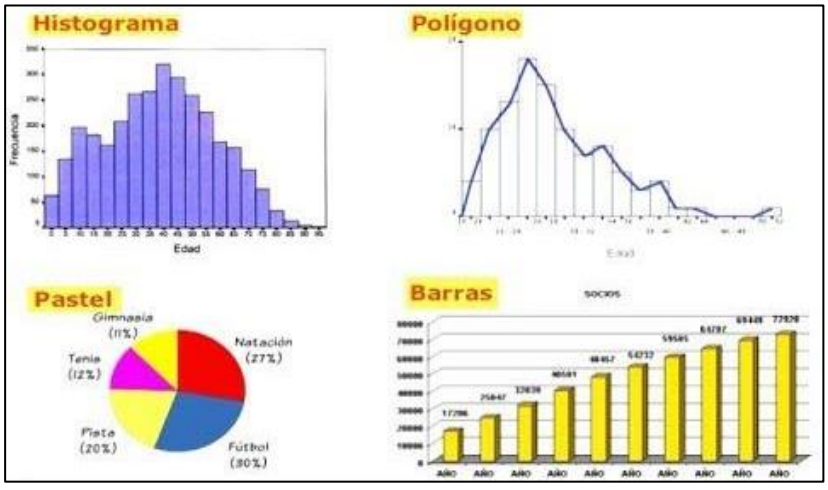
La estadística aplicada se ocupa de la aplicación de la metodología general a problemas específicos. Esto a menudo requiere el uso de técnicas de análisis de datos. Estos son algunos ejemplos de problemas estadísticos [3]:

- Interpretar la evidencia de una conexión entre los factores ambientales y la enfermedad.
- Diseñar experimentos para evaluar la eficacia farmacéutica,
- Extraer datos para encontrar segmentos objetivo de la población,
- Estudios de mercado para estimar la demanda de nuevos productos,
- Significado en política,
- Tamaño estimado de la población animal para ayudar a establecer reglas de conservación,
- Estudios de confiabilidad para determinar garantías,
- Mejorar la calidad de un servicio o producto de producción,
- Pronóstico del tiempo,
- Predicción de los precios del mercado de valores.

2.1.2.1. Estadígrafos.

Los estadísticos analizan e interpretan datos numéricos para permitir una planificación y una toma de decisiones informadas. Recopilan datos, aplican técnicas estadísticas y analíticas a los datos e identifican tendencias en función de los resultados de sus cálculos y pronósticos. Pueden trabajar en organizaciones públicas o privadas, empresas o como consultores independientes [3].

Figura 2. 1. Gráficos utilizados en estadística.



Fuente: (M. Valentina Márquez y Daniela Solano, 2018)

2.1.2.1.1. Media aritmética.

El promedio o media aritmética corresponde a un conjunto de datos de los cuales definimos su valor característico de la serie de datos como la suma de todos los valores u observaciones dividido por el número total de datos [4].

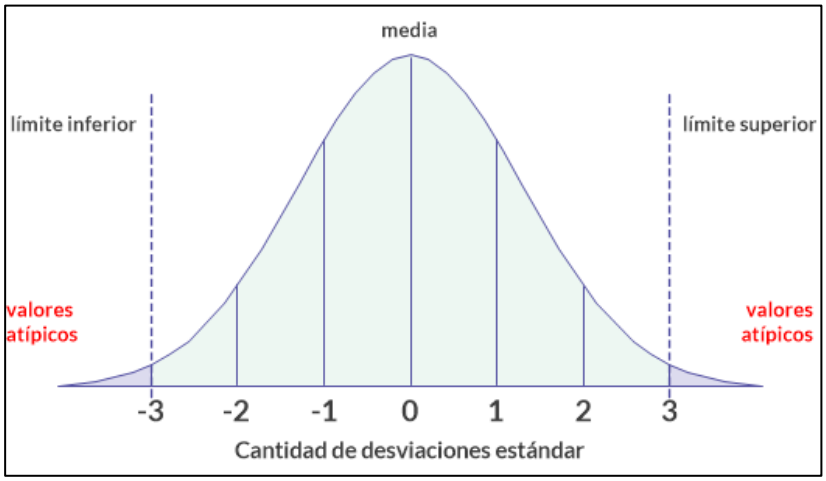
Algunas características de la media o promedio son:

- Considera todas las puntuaciones.
- El numerador de la fórmula es la cantidad de datos u observaciones.
- Cuando hay puntuaciones extremas, no tiene una representación exacta de la muestra.

2.1.2.1.2. Desviación Estándar.

La desviación estándar se usa para reconocer valores atípicos en un conjunto de datos, simplemente observando cuántas "desviaciones estándar"; es un valor del conjunto apartado de la media. Cualquier valor que esté a más de una desviación estándar de la media puede considerarse anormal [5].

Figura 2. 2. Características de la desviación estándar.



Fuente: (Galvaníze, 2019)

2.1.2.1.3. Mediana.

La mediana de un conjunto de números es el número del medio en un conjunto de datos. Sin embargo, los datos deben ordenarse numéricamente (de mayor a menor o de menor a mayor) antes de encontrar este promedio. Si el número del medio está entre dos números, encuentra la media de esos dos (súmalos y divide por 2) [6].

2.1.2.2. Conjunto de datos.

Un conjunto de datos es una lista o tabulación ordenada de datos. Como sabemos, una colección de información obtenida a través de la observación, medición, encuesta o análisis se llama datos. Puede contener información como hechos, cifras, ilustraciones, nombres o incluso descripciones básicas de objetos [7].

Un elemento es un conjunto de números o valores que son específicos de un tema. Los registros generalmente están etiquetados para ayudar a comprender lo que representan los datos. Sin embargo, cuando se trabaja con conjuntos de datos, no siempre se sabe qué significan los datos y no necesariamente se necesita comprender qué representan los datos para resolver el problema [7].

Tenemos diferentes conjuntos de datos para diferentes tipos de información entre los cuales se destacan a continuación:

- Registros numéricos.
- Datos bilaterales.
- Datos con variables múltiples.
- Entradas categóricas o absolutas.
- Consignación de correlación.

Tabla 2. 1. Conjunto de datos con variables múltiples.

TABLA DE CONTENIDO DE USUARIOS DE LA CNEL EP SANTO DOMINGO.

CLICOD	MDENUMFAB	SGCORX	SGCORY	USOCOD	CLIPRVCODP	CLICANCODP
3	1001661413	703451,036	9971964,85	CO	23	1
5	1001193561	703466,052	9971965,92	CO	23	1
7	1001149404	703466,512	9971970,44	CO	23	1
9	1001196309	703468,425	9971984,18	CO	23	1
11	1001190561	703469,271	9971990,18	RD	23	1
12	1001181073	703469,271	9971990,18	CO	23	1
13	1001181074	703470,981	9972000,19	CO	23	1
14	1001196310	703471,674	9972005,3	RD	23	1
15	1001196316	703472,332	9972009,94	CO	23	1
16	1001181056	703472,386	9972010,58	CO	23	1
18	6087807	703473,653	9972023,33	CO	23	1
19	50210701	703457,563	9972025	CO	23	1
20	12729964	703454,09	9972025,27	CO	23	1
21	50211102	703466,668	9972029,19	CO	23	1
22	50211103	703452,421	9972025,32	RD	23	1
26	13093400	703437,314	9972031,4	RD	23	1

Fuente: CNEL EP Unidad de Negocios Santo Domingo.
Elaborado por: Pujota E. – Macao R. (2022).

2.1.2.2.1. Valores fuera de rango.

Son valores que pueden afectar a los algoritmos de Machine Learning, pueden dar predicciones no convincentes, datos que normalmente se encuentran alejados de un conjunto de datos para ello se debe normalizar los datos de la Data set [8].

2.1.2.2.2. Valores Nulos.

Son valores desconocidos, valores que, al momento de ingresar los datos en una hoja de cálculo, algunas de las celdas suelen quedar vacías, dichos valores pueden ocasionar una alteración al momento de realizar una operación o programación [9].

2.1.2.2.3. Normalización de datos.

La normalización de datos es una técnica que se aplica como parte de la preparación de datos para Machine Learning permitiéndole que funcione mejor el algoritmo, con el objetivo de cambiar los valores del conjunto de datos para utilizar una escala común [10].

La normalización es necesaria para que el algoritmo modele los datos correctamente, una mala aplicación o elección errónea del método de normalización puede arruinar los datos y en general el análisis final [10].

La normalización permite evitar estos problemas mediante la creación de nuevos valores que mantienen la distribución general y las relaciones en los datos de origen, manteniendo los valores dentro de una escala que se aplica a todas las columnas numéricas utilizadas en los datos de los modelos [10].

La normalización de datos ofrece varias opciones para transformar los datos numéricos:

- Permite cambiar todos los valores a la escala de 0 a 1, también puede transformar los valores para representarlos como percentiles en lugar de valores absolutos.
- La normalización se puede aplicar a una sola o varias columnas en el mismo conjunto de datos.
- Puede guardar los pasos de transformación de normalización y aplicarla a otros conjuntos de datos que posean el mismo esquema y así poder repetir varias veces como desee en los valores con dicha normalización [10].

2.1.2.2.4. Reducción de la dimensionalidad.

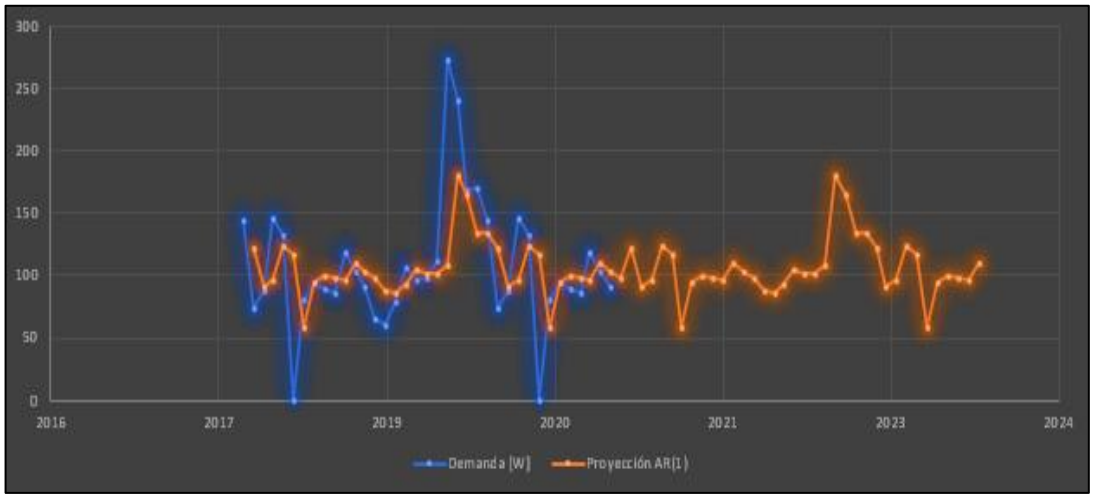
Los problemas de aprendizaje de la máquina se resuelven a través de métodos y dependen de muchos factores, siendo estos llamados características que son las variables de los datos. Cuantas más características proporcione el algoritmo, más difícil será entender los conjuntos de entrenamiento que se le esté dando al algoritmo. La reducción de la dimensionalidad lo que hará es reducir las variables aleatorias obteniendo de esta manera un principio para estas variables, dividiendo así en diferentes características y extracción de selección [11].

2.1.2.2.4. Series Temporales.

Las series temporales son un conjunto de datos etiquetados temporalmente de acuerdo con el orden secuencial en el que se generan, y además analizar su comportamiento a corto mediano y largo plazo para poder hacer una predicción a futuro en función del tiempo [12].

Figura 2. 3. Series temporales en proyección de la demanda.

PROYECCIÓN DE LA DEMANDA EN POTENCIA ACTIVA(EXCEL).



Fuente: (Excel).
Elaborado por: Pujota E. (2021).

2.1.2.3. Gráficas estadísticas.

Los problemas de aprendizaje de la máquina se resuelven a través de métodos y dependen de muchos factores, siendo estos llamados características que son las variables de los datos. Cuantas más características proporcione el algoritmo, más difícil será entender los conjuntos de entrenamiento que se le esté dando al algoritmo [13].

La reducción de la x` lo que hará es reducir las variables aleatorias obteniendo de esta manera un principio para estas variables, dividiendo así en diferentes características y extracción de selección [13].

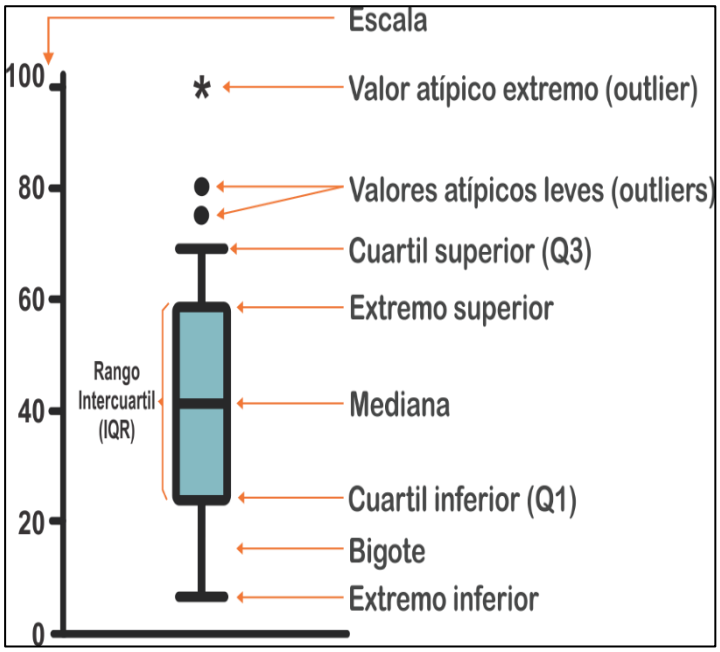
2.1.2.3.1. Diagrama de caja o Boxplot.

El diagrama de caja o caja de bigotes se crea para mostrar un conjunto de dispersión de datos, pueden ayudar a obtener rápidamente una idea del rango de valores mínimos y máximos además de la mediana para especificar de mejor manera los datos. La caja de bigotes tiene las siguientes propiedades que son el: mínimo, primer cuartil, mediana, tercer cuartil y máximo. Las líneas que se extienden y que sobresalen de la caja paralelamente se llaman bigotes [14].

Boxplot es una herramienta muy importante para el análisis de datos en Python al momento de crear diagramas de caja, es así que ayuda a entender los datos, puesto proporciona la información sobre la posición del mínimo el primer cuartil el medio el tercer cuartil además los valores máximos de los datos, es así que en la gráfica se puede observar el resumen de un conjunto de datos [14].

El matplotlib.pyplot módulo de la biblioteca matplotlib proporciona una boxplot()función con la ayuda de la cual podemos crear diagramas de caja [14].

Figura 2. 4. Representación de una caja de bigotes con sus partes.



Fuente: (PArapa, 2019)

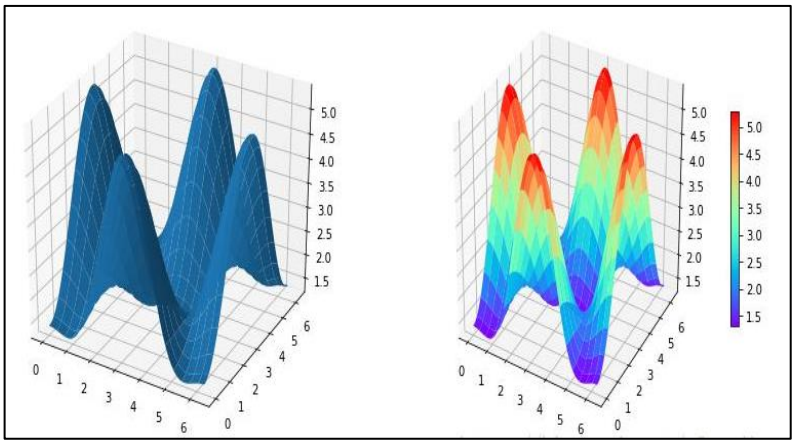
2.1.2.3.2. Histograma de frecuencia.

Permite analizar datos por medio de una gráfica, de una forma muy resumida, para poder interpretar la frecuencia de un conjunto de datos. Se utiliza para saber si los valores de la disposición de una variable se repiten frecuentemente además que por medio de estás gráficas estadísticas podemos determinar si una variable está o no en función del tiempo, permitiendo además agregar especificaciones o características al histograma [15].

2.1.2.3.3. Diagrama de densidad.

La gráfica de Density PLot es un instrumento de visualización de datos, es una variación continua del histograma a partir de los datos, de esta manera también se la conoce como densidad de kernel, la cual consiste en dibujar una curva continua en cada punto de los datos individuales, al final todas las curvas se suman para obtener una única estimación de densidad uniforme la región más alta es donde encontramos los datos máximos, tiene una ventaja que se pueden analizar las gráficas de una manera más clara, puesto que no dependen del tamaño de la columna [16].

Figura 2. 5. Diagrama de densidad utilizado para visualizar datos.



Fuente: (Matplotlib, 2022)

2.1.3. Lenguaje de programación.

Lenguaje de programación informática, cualquiera de los diferentes lenguajes utilizados para expresar un conjunto de instrucciones detalladas para una computadora digital. Dichas instrucciones se pueden implementar directamente cuando están en forma numérica específica de computadora conocida como lenguaje de máquina, después de un proceso de reemplazo simple cuando se expresan en un lenguaje de compilación correspondiente, o después de la traducción de un lenguaje "avanzado". Aunque hay muchos lenguajes de programación, relativamente pocos son ampliamente utilizados [17].

Ejemplos de lenguajes de programación: FORTRAN**,** ALGOL**,** Pascal**,** Logo**,** C Y C++**,** Ada**,** Java**,** Visual Basic**,** Python **,** PROLOG, entre otros [17].

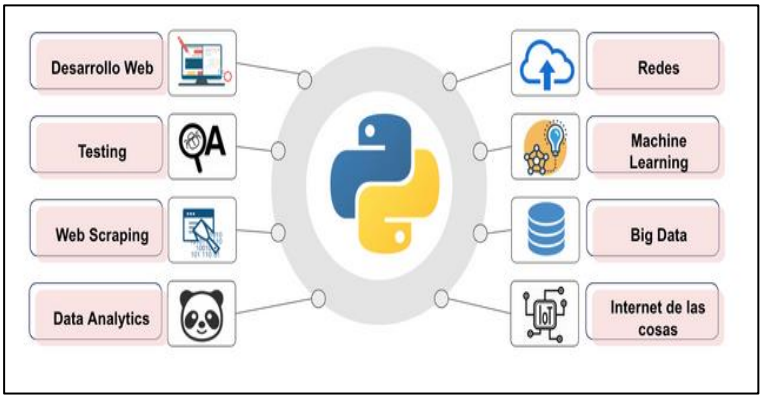
2.1.3.1. Python.

Python es un lenguaje de programación de computadoras que asiduamente se usa para crear sitios web y software, automatizar tareas y realizar análisis de datos. Python es un lenguaje de propósito general, lo que significa que puede usarse para crear muchos programas diferentes y no está especializado para ningún problema en particular [18].

Usos de Python.

Python permite a los ingenieros implementar sistemas SCADA (vigilancia y recopilación de datos) de código abierto que pueden sincronizar proyectos industriales sin problemas. El lenguaje de código puede manejar grandes conjuntos de datos en los que los ingenieros pueden crear algoritmos para cumplir con los objetivos que necesitan para su configuración [18].

Figura 2. 6. Aplicaciones con Python.



Fuente: (Romina Méndez, 2021)

2.1.3.1.1. Listas.

Una lista es un conjunto de estructuras de datos en Python, que permite trabajar con varios elementos a la vez, un conjunto de componentes de orden volátiles o mutuamente organizados. Cada elemento o valor de una lista se denomina elemento. Las oraciones se definen como caracteres en las referencias; una lista se crea colocando elementos entre corchetes [ ] cada uno de los elementos deben ir separados por una coma. Además, una lista puede contener otra lista como elemento para utilizar según se requiera [18].

2.1.3.1.2. Matrices.

Una matriz de Python es una colección de tipos comunes de estructuras de datos que tienen elementos con el mismo tipo de datos los cuales están almacenados en filas y columnas Se utiliza para almacenar colecciones de datos. En la programación de Python, las matrices son manejadas por el módulo "matriz". cuando se crean matrices utilizando el módulo de matriz, los elementos de la matriz deben ser del mismo tipo numérico [19].

2.1.3.1.3. Arreglos.

Un arreglo es una estructura, es decir, es un conjunto de variables que se citan y manejan con el mismo nombre, empleándose para almacenar múltiples valores en una sola variable, y que además permite el uso individual de sus elementos, es decir un arreglo es simplemente un conjunto finito de datos del mismo tipo [20].

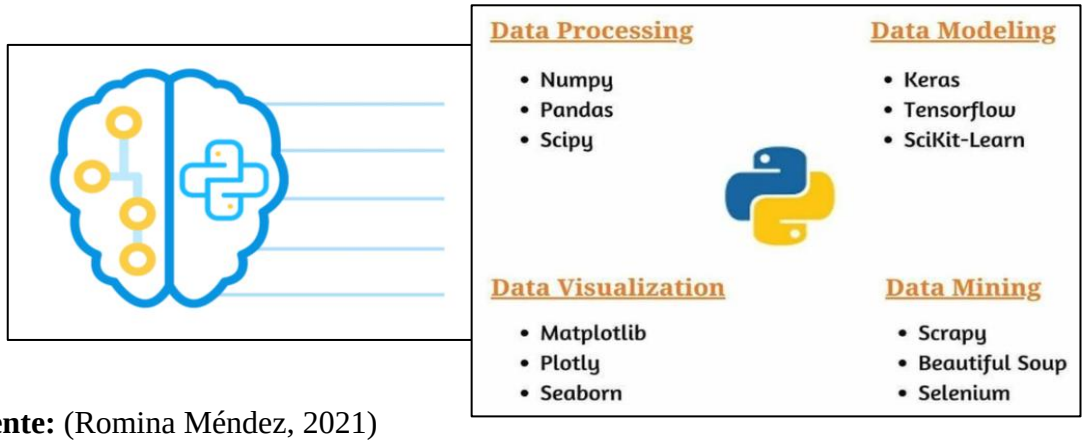
2.1.3.1.4. Diccionarios.

Es la ejecución de Python de una estructura de datos que se conoce generalmente como una matriz asociativa se utilizan para almacenar datos, además los diccionarios en Python están ordenados y no admiten duplicados, los diccionarios se escriben con corchetes y tienen claves y valores, estos tienen una cualidad muy importante ya que son cambiables esto quiere decir que se puede agregar, cambiar o en peor de los casos eliminar elementos una vez que se haya creado el diccionario [20].

2.1.3.1.5. Librerías.

Python utiliza diferentes tipos de librerías que permite codificar, con el objetivo de crear una interfaz tanto para el proceso de los datos como los es; Numpy, Pandas, entre otros, siendo así que para la visualización requiere de otro tipo de librerías como los son; Matplotlib, Seaborn lo cual permite realizar las funciones que tiene el software Python en base a lo que se quiera realizar [21].

Figura 2. 7. Librerías de Python para el aprendizaje automático.



Fuente: (Romina Méndez, 2021)

2.1.3.1.5.1. Pandas.

Pandas se basa en dos bibliotecas principales de Python: matplotlib para visualización de datos y NumPy para operaciones matemáticas permite trabajar con series temporales, analiza la estructura de datos estas estructuras se construyen a partir de arrays la librería NumPy.

Pandas actúa como un envoltorio sobre estas bibliotecas, lo que le permite acceder a muchos de los métodos de matplotlib y NumPy con menos código [21].

2.1.3.1.5.2. Numpy.

NumPy es una librería de Python que se especializa para trabajar con matrices y vectores, en el análisis de datos y en el cálculo numérico, para un gran volumen de datos, Incorpora una nueva clase de objetos llamados arrays que permite representar colecciones de datos de un mismo tipo en varias dimensiones, y funciones muy eficientes para su manipulación [21].

2.1.3.1.5.3. Sklearn.

Es una biblioteca clave para el lenguaje de programación y aprendizaje automático más útil e indispensable en Python, incluyendo algoritmos matemáticos y estadísticos. Proporciona una gama de herramientas efectivas de aprendizaje automático y modelado estadísticos que incluyen clasificación, regresión, agrupación y reducción de dimensiones a través de una interfaz de coherencia en Python. Basándose en la librería SciPy, NumPy y Matplotlib para la utilización en Machine Learning [21].

2.1.4. ArcGIS.

ArcGIS es un sistema información geográfica (SIG) completo para la recopilación, organización, gestión, análisis, de información geográfica. es la plataforma pionera en mundo para construir y usar sistemas de información geográfica (SIG), ArcGIS para aplicar el conocimiento geográfico al gobierno, los negocios, la ciencia, la educación y los medios. ArcGIS permite publicar información geográfica de forma accesible para todos [22].

Figura 2. 8. Ejemplo de una data set en ArcGIS con su ubicación geográfica.

Fuente: (ArcGIS).
Elaborado por: Macao R – Pujota E. (2022).

2.1.4.1. ArcMap.

ArcMap también le permite agregar otros estilos y símbolos definidos o generados por el usuario. ArcMap incluye una gran cantidad de estilos predefinidos para diferentes tipos de geometría, ofreciendo una amplia gama de opciones para mostrar datos [23].

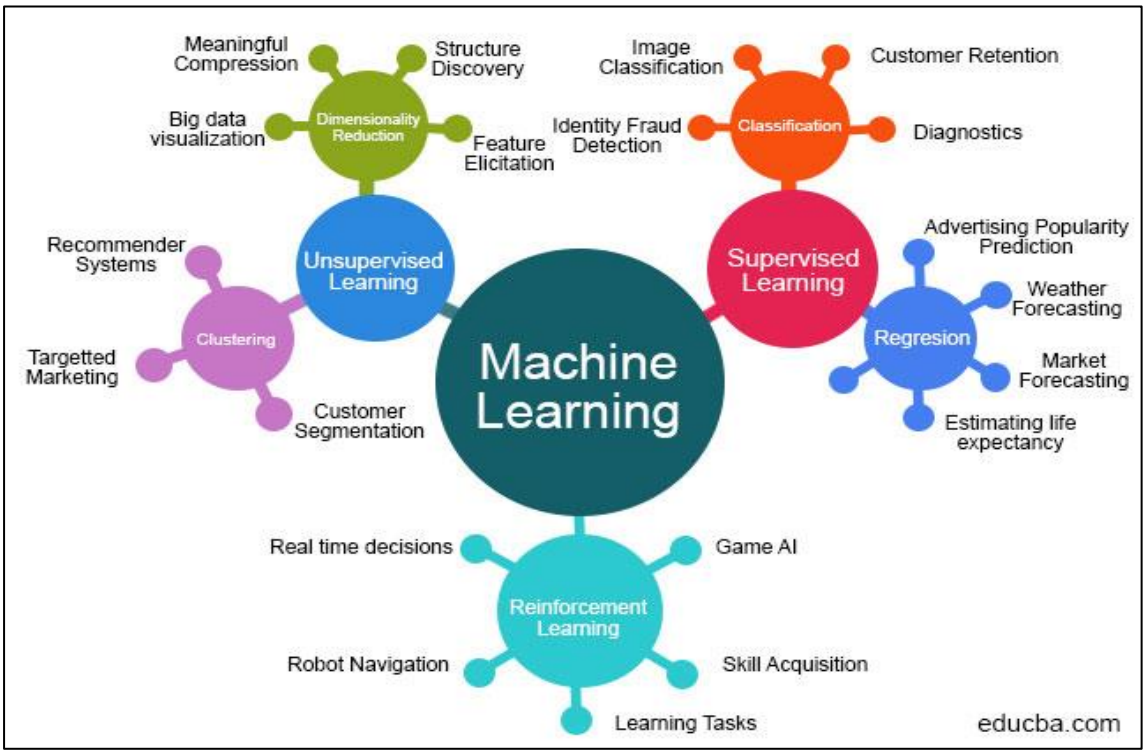
2.1.5. Inteligencia Artificial.

La inteligencia artificial (AI), se basa en la combinación de algoritmos planteados con el propósito de crear una máquina con las capacidades similares a la de un ser humano, permitiendo de esta manera optimizar el trabajo en base al tiempo, siendo más eficiente y que exista menor error en la producción. Realiza tareas frecuentes controladas por computadora de gran volumen de manera confiable y sin fatiga [24].

2.1.5.1. Machine Learning (ML).

El aprendizaje automático (ML) es un tipo de inteligencia artificial (AI) que permite que las aplicaciones de software predigan resultados con mayor precisión sin estar programadas explícitamente para hacerlo. Los algoritmos de aprendizaje automático se clasifican normalmente en supervisados y no supervisados, utilizando como base datos históricos como entrada para encontrar patrones que como resultado predigan nuevos valores de salida [25].

Figura 2. 9. Algoritmos de aprendizaje automático.



Fuente: (Ashish Patel, 2018)

2.1.5.2. Deep Learning.

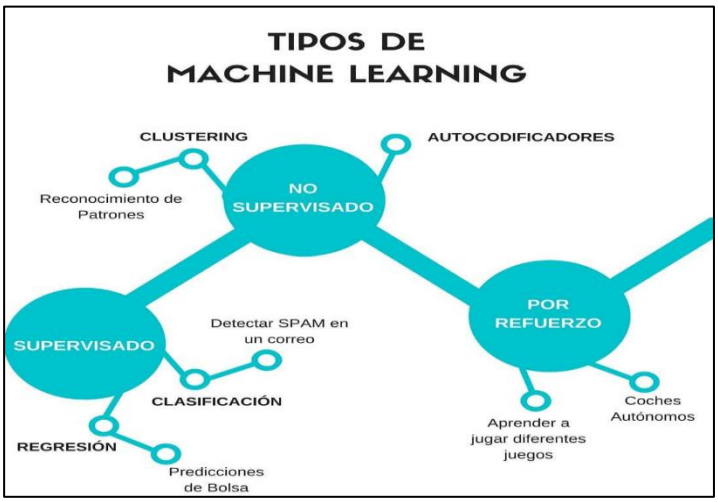
El aprendizaje profundo es un subgrupo del aprendizaje automático, que es esencialmente una red neuronal con tres o más capas. Estas redes neuronales intentan simular el comportamiento del cerebro humano, aunque lejos de igualar sus capacidades, para que pueda "aprender" a partir de grandes cantidades de datos. Si bien una red neuronal de una sola capa aún puede hacer predicciones aproximadas, varias capas ocultas pueden ayudar a ajustar y refinar la precisión. El aprendizaje profundo ejecuta muchas aplicaciones y servicios de inteligencia artificial (AI) que mejoran la automatización al realizar tareas analíticas y físicas sin intervención humana [25].

2.1.5.3. Ingeniería del conocimiento.

Es una técnica que se basa en la teoría de (actor-red), ubicando de manifiesto redes en función de datos históricos y creando nuevas en base a la información emulando la mente de una persona para solucionar problemas en un campo especifico. Además, se debe limitar una parte del total de la base de datos para que el ser humano pueda estudiar y analizar transformando de esta manera los datos en conocimiento con el objeto de captar e interiorizar mentalmente las relaciones que existe entre todos los elementos [26].

2.1.5.4. Tipos de Aprendizaje.

Figura 2. 10. Tipos de Machine Learning.



Fuente: (Cristian Santander, 2020)

2.1.5.4.1. Aprendizaje supervisado.

El aprendizaje supervisado es un conjunto de técnicas que permiten realizar predicciones futuras en base a datos históricos o etiquetados para obtener como resultado una salida que es de tipo numérico en problemas de regresión y de tipo categórico en problemas de clasificación [27].

Regresión lineal.

Es un algoritmo de aprendizaje supervisado que se utiliza en Machine Learning, para predecir el valor de una variable según el valor de la otra encontrando los mejores parámetros como son la variable pendiente y la independiente. La regresión lineal se ajusta a una línea recta midiendo el error con respecto a los puntos de entrada y el valor de salida real, el algoritmo debe minimizar el coste de la función de error cuadrático. La variable que se va a predecir se la denomina como dependiente, mientras que la variable que se está utilizando para predecir el valor de la otra se la conoce como independiente [28].

Clasificación.

La clasificación en Machine Learning tiene la tarea de asignar una clase entre un número limitado de clases, es decir que se utiliza cuando el resultado es una etiqueta discreta, comprendiendo que los datos discretos tienen límites conocidos y definibles en un conjunto finito de resultados, en general predice a que categoría pertenece un conjunto de datos [29].

2.1.5.4.2. Aprendizaje no supervisado.

En esta técnica no es necesario compartir o supervisar los datos etiquetados con el modelo. En su lugar el algoritmo toma sus propias decisiones y comienza a aprender de los datos sin guía. El aprendizaje no supervisado encuentra maneras de clasificar variables y comprobar si se ajustan a lo que se quiere obtener, además con este método se puede obtener nueva información y que no haya sido identificada previamente, funcionan sin ningún entrenamiento adecuado y responden tan pronto como reciben los datos [29].

2.1.5.4.2.1. Tareas que se pueden realizar en el aprendizaje no supervisado.

Data Mining.

El Data Mining nos permite analizar una gran cantidad de datos, de esta manera se puede sacar mayor provecho a la información, de hecho, la minería de datos, sondea, prepara, y explora los datos más relevantes [29].

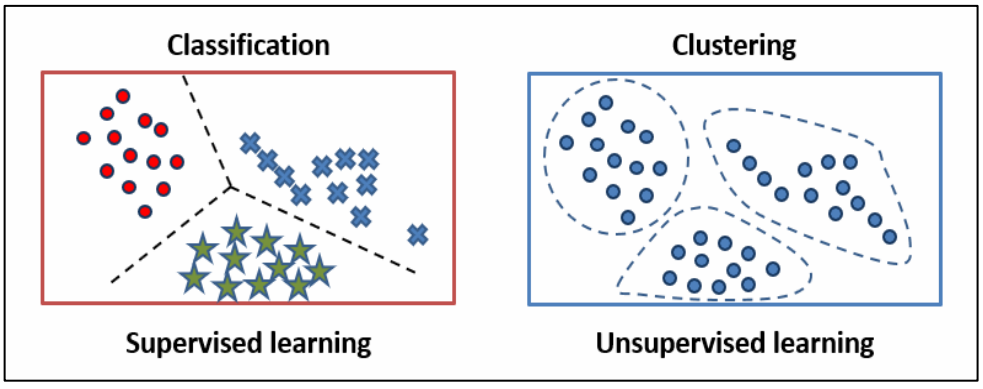
Agrupación.

Es una de las técnicas más utilizadas por el aprendizaje no supervisado, este algoritmo encuentra el patrón y categoriza la recolección de los datos, también identifica los grupos a partir de la información. Esta técnica se divide en diferentes grupos que son: Exclusivo Aglomerado, Superposición y Probabilístico [29].

Clusterización.

Clustering de los datos nos permite agrupar en procesos de machine Learning datos no etiquetados en conjuntos, para así construir subconjuntos de datos, estos datos resultan ser similares entre sí con elementos diferentes de esta manera se puede segmentar datos en grupos de dimensiones similares basándose en las características, para encontrar posibles desaciertos dentro de su funcionamiento, es un proceso fundamental que le permite a los algoritmos de aprendizaje automatizado comprender los datos con el cual ellos van a desarrollar sus actividades [30].

Figura 2. 11. Diferencia entre clasificación y Clusterización.



Fuente: (Héctor Klie, 2017)

2.1.6. Algoritmos de Clusterización.

Hay varios tipos de algoritmos de agrupación que manejan todos los tipos de datos únicos.

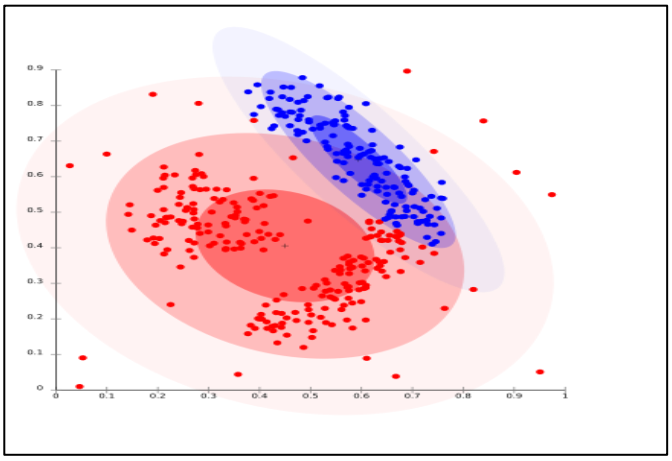
2.1.6.1. Algoritmos de clusterización basados en la densidad.

La agrupación en clústeres basada en densidad se refiere a métodos de aprendizaje automático desatendidos que identifican distintos clústeres en los datos, basados en la idea de que un clúster/grupo es una sala de objetos en particular como ejemplo las computadoras que serían una región contigua con una alta densidad de puntos, separada de otros clústeres de regiones dispersas. Hay muchas familias de algoritmos de agregación de datos, y es posible que esté familiarizado con los más populares: k- [31].

2.1.6.2. Algoritmos basados en distribución.

La agrupación o clusters basada en la distribución, elabora clústeres que parten de modelos matemáticos definidos con precisión subyacentes a la base de datos. En el agrupamiento basado en distribución, todos los puntos de datos se consideran parte de un grupo en función de la probabilidad de que un punto pertenezca a un grupo en particular. La agrupación en el algoritmo de distribución puede ser normal o gaussiana. La distribución gaussiana es más evidente cuando tenemos un número fijo de distribuciones y todos los datos cercanos se ajustan de tal manera que se puede maximizar la distribución de datos [32].

Figura 2. 12. Representación gráfica de un algoritmo basado en distribución.

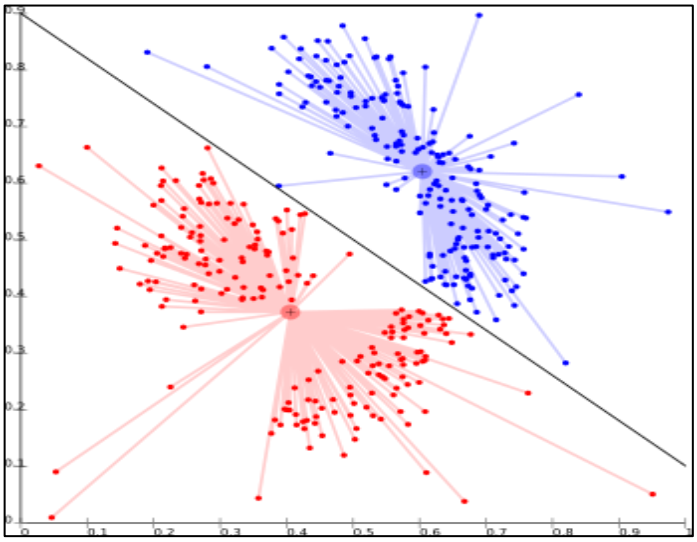


Fuente: (Surya Priy, 2021)

2.1.6.3. Algoritmo basado en Centroides.

Uno de los algoritmos iterativos de agrupación en clústeres en los que los clústeres se forman según la proximidad de los puntos de datos al centroide de los clústeres. Estos tipos de algoritmos son rápidos y eficientes, y separan puntos de datos en función de múltiples centroides en los datos. por lo tanto, cada punto de datos se asigna a un grupo en función de su distancia al cuadrado desde el centroide. Este tipo de algoritmo tiene una desventaja, que es la de agrupar distribuciones basadas en la densidad [33].

Figura 2. 13. Representación gráfica de un algoritmo basado en centroides.



Fuente: (Surya Priy, 2021)

2.1.6.4. Algoritmo basado en Jerarquías.

Esta técnica es muy utilizada en Machine Learning. se utiliza normalmente en datos jerárquicos, la extracción de agrupaciones naturales de objetos de datos similares, la asociación Jerárquica se basa en el uso de estas técnicas de agrupación para encontrar una jerarquía de agrupaciones dependiendo de la dirección en la que el algoritmo ejecute el agrupamiento, donde esta jerarquía se asemeja a una estructura de un árbol es decir agrupa los datos basándose en la distancia entre cada uno y buscando que los datos que están dentro de un clúster sean los más similares entre sí [34].

2.1.6.5. K-Means.

K-Means es un algoritmo iterativo de agrupamiento que intenta dividir el conjunto de datos en subgrupos, además es un método no supervisado. Se utiliza cuando tenemos muchos datos sin etiquetar. El objetivo de este algoritmo iterativo es encontrar grupos "K" (clusters) entre el conjunto de datos sin procesar. Son agrupados en base a la similitud de sus columnas los grupos se van ajustando en posición en cada iteración del proceso hasta que converge el algoritmo [35].

Identificar similitudes: Se basa en identificar a las variables que tengan similitudes, o características que se asemejen a las diferentes variables con el fin de que permita clasificarlas según su similitud [36].

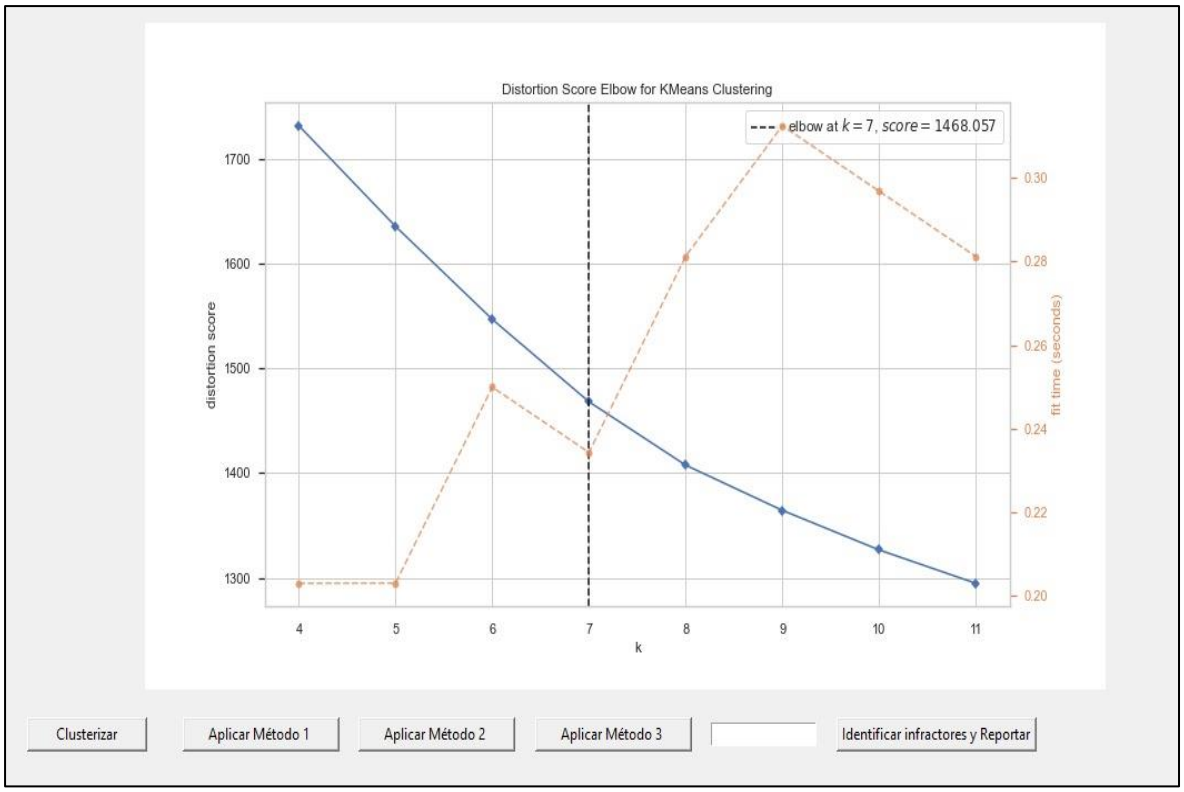
2.1.6.5.1. Método del codo.

Probablemente el método más conocido, el método del codo, en el que calcula y gráfica la suma de cuadrados en cada número de grupos, y allí buscas un cambio en la pendiente de empinada poco profunda, un codo, para determinar el número óptimo de grupos. Este método es inexacto, pero aun potencialmente útil [37].

El método de la curva del codo es útil porque muestra cómo aumentar el número de conglomerados contribuye a separar los conglomerados de manera significativa, no marginal. La curva indica que los grupos que se adicionaron y se encuentran sobre o más allá del tercero tienen poco valor. El método del codo es bastante bueno, pero es una solución ingenua basada en la varianza intracluster. La estadística de brecha es un método más sofisticado para manejar datos que tienen una distribución sin agrupamiento obvio [37].

Este método funciona de la siguiente manera: se calcula la suma de los errores al cuadrado dentro del clúster para diferentes valores de “K” y se elige la “K” para la cual la suma de los errores al cuadrado comienza a disminuir. Esto es visible como un codo [37].

Figura 2. 14. Representación gráfica del método del codo en Machine Learning.



Fuente: CNEL EP Unidad de Negocios Santo Domingo.
Elaborado por: Pujota E. – Macao R. (2022).

2.1.6.5.2. Distancia euclidiana.

La distancia euclidiana son cálculos de distancia que se utilizan tanto en el aprendizaje supervisado como no supervisado, generalmente para calcular la similitud entre los puntos de datos. Un cálculo de distancia efectivo mejora el rendimiento de nuestro modelo de aprendizaje automático, ya sea para ordenar o agrupar tareas [38].

La distancia euclidiana suele ser la distancia "estándar" utilizada, p. B. en K-medias más cercanas (clasificación) o K-medias (agrupación) se utiliza para encontrar "k puntos más cercanos" a un punto de prueba dado. Otro ejemplo notable es el agrupamiento jerárquico, el agrupamiento aglomerativo (unión completa y única) donde se quiere encontrar la distancia entre clusters [38].

2.2. Marco referencial.

2.2.1. El artículo 314 de la constitución de la Republica del Ecuador preceptúa.

El estado será el responsable de la provisión de los servicios públicos de agua y de riego, saneamiento, Energía eléctrica, telecomunicaciones, vialidad, infraestructuras portuarias y aeroportuarias y los de más que determine la ley [39].

El Estado garantizará que los servicios públicos y su provisión respondan a los principios de obligatoriedad, generalidad, uniformidad, eficiencia, responsabilidad, universalidad, accesibilidad, regularidad, continuidad y calidad. El Estado dispondrá que los precios y tarifas de los servicios públicos sean equitativos, y establecerá su control y regulación [39].

2.2.2. Ley Orgánica del servicio público de energía eléctrica (LOSPEE).

El artículo 17 de la LOSPEE faculta al directorio de la ARCONEL entre otros aprobar los pliegos tarifarios para el servicio público de energía eléctrica y para el servicio de alambrado público general [40].

El artículo 71 de la LOSPEE. - Establece la suspensión de servicio, La empresa eléctrica podrá suspender el suministro de energía eléctrica al consumidor o usuario final, por cualquiera de los casos siguientes [40]:

1. Por falta de pago oportuno del consumo de energía eléctrica, al día siguiente de la fecha máxima de pago previamente notificada al consumidor o usuario final;
2. Cuando se detecte consumos de energía eléctrica, a través de instalaciones clandestinas, directas y/o similares, que alteren o impidan el normal funcionamiento del medidor;
3. Cuando la acometida del usuario final no cumpla con las condiciones técnicas establecidas para el efecto;
4. Cuando se compruebe el consumo de energía eléctrica en circunstancias que alteren lo estipulado en el contrato respectivo;

5. Cuando la empresa eléctrica previo aviso, mediante adecuados medios de comunicación, comunique oportunamente al usuario final que por motivos de mantenimiento o reparación se producirá una suspensión de energía eléctrica;
6. Cuando se consuma energía eléctrica sin haberse celebrado el respectivo contrato de suministro de electricidad;
7. Cuando existan conexiones al sistema de la empresa eléctrica sin contar con su autorización;
8. Por causas de fuerza mayor o caso fortuito.

2.2.2.1. El artículo 159 de la LOSPEE.

Análisis y determinación de costos del servicio público de energía eléctrica establece que: Corresponde al ARCONEL, elaborar anualmente el análisis para determinación de los costos del servicio público de energía eléctrica a partir de los costos de las actividades de generación, transmisión, distribución y comercialización de energía eléctrica, y en conformidad con las políticas que para efecto defina el Ministerio de Energía y Recursos Naturales No Renovables [40].

2.2.3. Categorías al consumidor regulado.

2.2.3.1. Categorías tarifarias:

La distribuidora tiene la responsabilidad de aplicar correctamente las tarifas al consumidor regulado para aplicar dicha tarifa la distribuidora debe hacer una evaluación de la carga y el uso de la energía que el usuario le dará a esta, con base a la información otorgada se establece la tarifa y el tipo de consumidor regulado [41].

2.2.3.2. Categoría residencial:

Es la energía que se entrega exclusivamente al servicio doméstico, es decir la vivienda de una familia independiente del tamaño de la carga, es una de las categorías donde se incluye a los consumidores de bajos consumo, de escasos recursos económicos, en cuyos domicilios tienen una minúscula activad económica, como puede ser comercial o artesanal [41].

2.2.3.3. Categoría general:

Esta categoría se diferencia de la categoría residencial, en esta categoría la energía eléctrica abarca al comercio, la industria y la prestación de los servicios públicos y privados, Se considera dentro de esta categoría entre otros a los siguientes [41]:

Locales y establecimientos comerciales públicos o privados:

- Tiendas, almacenes, salas de cine o teatro, restaurantes, hoteles y afines;
- Plantas de radio, televisión y cualquier otro servicio de telecomunicaciones;
- Clínicas y hospitales privados
- Instituciones educativas privadas;
- Vallas publicitarias;
- Organismos internacionales, embajadas, legaciones y consulados;
- Asociaciones civiles y entidades con o sin fines de lucro; y,
- Cámaras de comercio e industria tanto nacionales como extranjeras; entre otros.

2.2.4. Tarifas en el servicio eléctrico.

Subsidio del adulto mayor.

Es un beneficio que está contemplado en la constitución de la Republica el otorga un descuento del 50% del pago del servicio de energía eléctrica hasta 138KWh/mes, para las instituciones sin fines de lucro las cuales desarrollen actividades como, atención a las personas adultas mayores, como: asilos, albergues, comedores e instituciones gerontológicas [41].

Tarifas de la dignidad.

Estos usuarios continuaran pagando 4cUSD por KWh. esta tarifa se aplica a los usuarios residenciales en la región sierra cuyos consumos de energía no sobrepasen los 110 KWh/ mes en la región sierra, y cuyos usuarios de las empresas distribuidoras de las Regiones Costa/oriente y región insular nos sobrepasen los 130KWh/mes [41].

Tarifas en el sector eléctrico.

Las tarifas promedio de consumo en 2021 en Ecuador son de USD 0,1031 para el sector residencial; USD 0,1044 para el comercial; USD 0,799 para el industrial; y USD 0,712 para otros, que constituyen los recursos con los cuales se cubren los costos de todo el sector eléctrico [41].

2.2.5. Reducir las pérdidas de electricidad para garantizar la seguridad energética.

Son inevitables prevenir las pérdidas de energía en cualquier sistema eléctrico, a pesar de esto, 20 de los 26 países de Latinoamérica y el Caribe presentan pérdidas mayores al 10% del total de la energía eléctrica generada. En términos económicos, estas pérdidas de energía se traducen en un costo anual de entre US$11 y US$17 mil millones para las empresas eléctricas a nivel de América Latina. Este costo tiende hacer el 0.3% del PIB (Producto Interno Bruto) de la región, Siendo así comparado con el monto destinado a programas sociales como lo es Oportunidades (hoy Prospera) en México o la Bolsa Familia en Brasil [42].

2.2.5.1. ¿Dónde se pierde esta energía?

Las pérdidas de energía se dan en dos etapas, durante el transporte de energía esto se debe a largas distancias de transmisión, y en las distribuidoras atreves del consumidor final no medido y a la vez no facturado, esto se debe a que los medidores de energía son manipulados por los clientes y de esta manera crean un perjuicio económico a la empresa distribuidora, o a la vez existen personas inescrupulosas que se conectan directamente a la red.

2.2.6. Pérdidas de energía eléctrica.

2.2.6.1 Clasificación de las pérdidas.

En las pérdidas de energía se registran dos categorías de manera general para los sistemas eléctricos siendo estas: las pérdidas técnicas y no técnicas [43].

2.2.6.2 Pérdidas de energía eléctrica no - técnicas:

Se basa en la energía eléctrica entregada a los usuarios y que no son canceladas por el consumo que estos presentan, siendo por motivos como la manipulación de los elementos de medición ocasionando pérdidas financieras directas a la empresa o proveedor [43].

Las pérdidas no técnicas son causadas por factores externos tales como los climatológicos o económicos en los sistemas eléctricos, pero relativos a la gestión de las empresas generadoras y factores institucionales del sector energético [43].

Hurto: Esto se da cuando el usuario no a suscrito un contrato con la empresa distribuidora, y se conecta directamente a la red de baja tensión y así elude el pago de la planilla de energía eléctrica mensual [43].

Fraude: Se da cuando el usuario esta suscrito como cliente de la empresa distribuidora y manipula intencionalmente los aparatos de medición para eludir su pago de la planilla de energía eléctrica [43].

Electricidad no contabilizada: En este caso se toma como referencia los errores de los lectores y errores de medición, el alumbrado público y como ultimo las señaléticas de tránsito [43].

Problemas de gestión: Son los errores que se abarca en la contabilidad y el mantenimiento deficiente de los registros de usuarios [43].

Las Empresas Distribuidoras del servicio de energía eléctrica pierden ingresos económicos por el consumo de electricidad que no pueden identificar por los diferentes factores que se puedan presentar o a la vez que sean pérdidas técnicas. Sin embargo, existen situaciones en las que, al no cobrar por el servicio, este se registra también como pérdida. Es decir, aunque el consumo tenga la medición correcta y se factura correctamente, se le considera como pérdida debido al bajo costo de cobro que tiene establecida la empresa distribuidora [43].

2.2.6.3 Pérdidas de electricidad técnicas:

Se denomina a las pérdidas que se propagan en las líneas de transmisión y distribución. Estas son propias al transporte de la electricidad, y se agrupan de manera significativa a las características que presenta la infraestructura en los sistemas de energía eléctrica. De esta manera se dividen en fijas y variables [43]:

2.2.6.4. Pérdidas de energía técnicas fijas.

Estas pérdidas son ocasionadas solo por el simple hecho de energizar un circuito, además que se presentan en el transformador en el cual se producen por deficiencias físicas, siendo estas como pérdidas en el núcleo, por histéresis y el efecto corona en las líneas de transmisión. Estas pérdidas técnicas fijas, son proporcionales a la tensión e independientes del flujo de electricidad [43].

El voltaje varía relativamente poco respecto de su valor nominal, estas pérdidas de energía son tratadas como una constante que depende sobre todo de la calidad de la línea. Estos tipos de pérdidas representan entre 20% y 40% del total de las pérdidas técnicas en las líneas de transmisión y que no pueden ser eliminadas produciéndose en el sistema, aunque la carga conectada fuera igual a cero [43].

2.2.6.5. Pérdidas técnicas variables.

De facto estas pérdidas se relacionan con el transporte, puesto que la energía se pierde durante la transmisión. y es casi imposible eliminarlas en su totalidad [43].

2.2.6.6. ¿Cuál es la diferencia entre pérdidas técnicas y no técnicas?

Las pérdidas técnicas son ocasionadas debido a la corriente que fluye en los conductores produciendo deficiencias físicas. No se puede evitar, pero sí disminuir mediante una instalación eléctrica adecuada. Las pérdidas no técnicas se basan en la energía eléctrica consumida y que no es facturada o cobrada por las empresas distribuidoras. Además, las perdidas técnicas se pueden calcular [43].

- Pérdidas debidas a la inducción de campos electromagnéticos.
- Pérdidas dieléctricas debido al material de aislamiento entre los conductores.
- Pérdidas por distorsión armónica.
- Pérdidas por mala conexión a tierra.

Las pérdidas no técnicas son más difíciles de reducir.

- Manipulación del medidor.
- Enganchar o evitar el medidor.
- Relaciones de transformador de instrumento programadas incorrectas en el medidor.
- Carga para transformadores de instrumentos demasiado alta.
- Lecturas incorrectas del medidor.
- El medidor está defectuoso o no está en la clase de precisión.
- Facturas de electricidad impagas.

Las perdidas no técnicas no se eliminarán por completo, puesto que las personas siempre idearán un plan de como hurtar energía y así eludir su responsabilidad económica ante las empresas distribuidoras [43].

2.2.7. Categorías tarifarias y niveles de tensión.

Los niveles de tensión en el Ecuador dependen de las características propias del consumidor, puesto este puede ser consumidor residencial, comercial e industrial, de acuerdo al tipo de consumidor se establece el nivel de tensión y estas pueden ser, media, baja y alta tensión [44].

2.2.7.1. Tarifa residencial.

Esta tarifa se aplica a todos los consumidores que están sujetos a la categoría de tarifa residencial, siendo independiente el tamaño de consumo de carga conectada a los usuarios [44].

El consumidor debe pagar:

a) Un cargo por comercialización en USD/consumidor-mes, independiente del consumo de energía que el usuario presente.

b) Cargos incrementales por energía en USD/kWh, en función de la energía consumida por el usuario.

2.2.7.2. Tarifa residencial temporal.

Esta tarifa tiene una particularidad con respecto a las demás tarifas puesto que se aplica a los consumidores residenciales que no tienen su residencia permanente en el área de servicio de la distribuidora y utiliza la energía eléctrica de forma asidua es decir los fines de semana, o cuando la ocasión lo amerite como son las fechas cívicas entre otras [44].

El consumidor debe pagar:

a) Un cargo por comercialización en USD/consumidor-mes, independiente del consumo de energía.

b) Un cargo único por energía en USD/kWh, independiente de la energía consumida.

2.2.8. Facturación.

La facturación del servicio público de energía eléctrica se da en base a la sumatoria de los rubros económicos establecidos por la empresa distribuidora siendo estos: potencia, energía, pérdidas técnicas, comercialización y penalización por el bajo factor de potencia que presente alguna entidad, de acuerdo a las características que debe presentar el consumidor regulado [44].

CAPÍTULO III

METODOLOGÍA DE LA INVESTIGACIÓN

3.1. Localización.

La investigación está orientada a encontrar los posibles usuarios de la CNEL EP unidad de negocios Santo Domingo que hurtan energía. La unidad de negocios Santo Domingo según reportes del 2014, es abastecida por 10 Subestaciones, que a su vez poseen 12 transformadores de potencia que abastecen la carga de 34 alimentadores de Distribución.

Tabla 3. 1. Área de servicio de la CNEL EP Unidad de Negocios Santo Domingo.

Nº	UNIDAD DE NEGOCIO	SUBESTACIÓN	ALIMENTADOR
1	CNEL REGIONAL SANTO DOMINGO	S/E # 1 VIA QUITO	Vía Quito
2			Centro
3			Lorena
4			Las Mercedes
5			Sur
6		S/E # 2 VIA QUEVEDO	Circuito # 1
7			Circuito # 2
8			Circuito # 3
9			Circuito # 4
10			Circuito # 5
11		S/E # 3 LA CONCORDIA	Puerto Quito
12			La Unión
13			Pepe Pan
14			Concordia
15			Monterrey
16		S/E # 4 EL CARMEN	Nuevo Israel
17			Vía Chone
18			Porvenir
19			Maicito
20		S/E # 5 PATRICIA PILAR	Luz de América
21			Patricia Pilar
22			La 14
23		S/E # 6 ALLURIQUIN	Chiguilpe
24			Tandapi - Faisanes
25		S/E # 7 VALLE HERMOSO	Epacem
26			Valle Hermoso
27		S/E # 8 EL CENTENARIO	Juan Eulogio
28			Rosales
29			GENERAL S/E # 8
30		S/E # 9 LA CADENA	Santa Martha
31			Rio Verde
32			Rio Toachi
33		S/E # 10 EL ROCÍO	Puerto Limón
34			Bramahadora

Fuente: CNEL EP Unidad de Negocios Santo Domingo (2014).

3.2. Tipos de investigación.

3.2.1. Investigación exploratoria.

Por medio de este tipo de investigación se puede analizar información específica que no ha sido estudiada previamente, con el objetivo de explorarla y posteriormente hacer una investigación más detallada.

Al ser fenómenos poco estudiados y en algunos casos incluso sin estudios previos, no se apoyan en teorías, sino en la recolección de datos, lo cual permite determinar, explicar o encontrar ciertos patrones para una investigación más profunda y minuciosa. En el presente proyecto de tesis se busca encontrar patrones en los consumos eléctricos de los últimos 12 meses de los usuarios de la CNEL EP Unidad de Negocios Santo Domingo, con el propósito de encontrar a posibles hurtos, y así reducir las pérdidas.

3.2.2. Investigación aplicada.

Su principal objetivo es resolver un determinado problema o tarea, por medio de la búsqueda y consolidación del conocimiento. Los conocimientos adquiridos durante la carrera, fueron aplicados en el estudio de los consumos mensuales de energía, para así solucionar y minimizar el problema de hurto de energía utilizando técnicas de Inteligencia Artificial.

3.3. Métodos.

3.3.1. Método Deductivo.

Extrayendo conclusiones en base a una premisa o a una serie de proposiciones asumidas como verdaderas, se determina que los usuarios que presentan un consumo normal de energía eléctrica durante algunos meses y luego este consumo baja drásticamente, y se mantiene así durante los meses consecuentes, podría estar infringiendo la ley hurtando energía eléctrica.

3.3.2. Método Inductivo.

A partir de la evidencia que se encuentre en este proyecto, los resultados arrojados por el algoritmo de Inteligencia Artificial y la futura recolección y adición de más datos se pretende construir un modelo general de detección de infractores.

3.3.3. Método Comparativo.

La comparación entre consumos de clientes es la piedra fundamental del funcionamiento del algoritmo a desarrollarse, ya que, para encontrar a los posibles infractores, se debe agrupar los datos dependiendo de su similitud o diferencias, además de encontrar posibles patrones que muestren relaciones entre consumos.

3.3.4. Método Analítico.

Por medio de este meta análisis, se trata de descomponer los datos, separándolos entre sí, con el objetivo de determinar la naturaleza de los consumos de los posibles infractores, lo cual podría prevenir que se siga cometiendo ilegalidades y reducir pérdidas.

3.4. Fuentes de recopilación.

La información fue obtenida de fuentes secundarias como: libros, artículos y tesis relacionadas a encontrar patrones en una gran cantidad de datos, especialmente en series temporales. En adición se recopilo información en técnicas y modelos empleados para la óptima limpieza de datos y el descubrimiento de patrones.

Los datos analizados fueron obtenidos de fuentes primarias, siendo provistos por la CNEL EP Unidad de Negocios Santo Domingo.

3.5. Diseño de la investigación.

El objetivo principal de la investigación, radica en el descubrimiento o identificación de posibles hurtos de energía eléctrica por parte de los usuarios, lo cual causa grandes pérdidas económicas a las empresas distribuidoras de energía.

Para la identificación de los posibles usuarios quienes hurtan energía, se ha hecho uso del programa Python, el cual permite manipular, analizar y visualizar datos, además de proveer potentes librerías para el desarrollo e implementación de algoritmos de Inteligencia Artificial, en donde se realizaron las siguientes tareas:

- Carga de datos al programa.
- Limpieza de datos.
- Determinación de los datos para el análisis.
- Análisis de los datos e identificación de posibles usuarios que hurtan energía.

3.5.1. Enunciación del problema de investigación.

La CNEL EP Unidad de Negocios Santo Domingo utiliza una fórmula matemática para la determinación de los posibles usuarios que hurtan energía eléctrica, la cual está basada en el consumo promedio y el último consumo mensual de energía eléctrica.

Dicha fórmula se trata de un cálculo sin base científica, y en la actualidad no existe una forma específica para determinar si un usuario hurta o no energía eléctrica. Por lo tanto, a el presente estudio, se utilizará el método K-Means, para la detección de los usuarios descritos anteriormente.

3.5.2. Datos.

Para el presente proyecto se utilizó un conjunto de datos provisto por la CNEL EP Unidad de Negocios Santo Domingo, el cual incluye 264237 filas y 26 columnas. Cada una de las filas representa un usuario, entre las columnas se encuentran datos como: nombre de usuario, código de cliente, numero de medidor, dirección en coordenadas geográficas y consumo eléctrico durante los últimos 12 meses.

3.5.3. Consideraciones para la identificación de los posibles infractores.

Para el desarrollo del proyecto se tienen en cuenta las siguientes consideraciones:

- **Valores fuera de rango;** el conjunto de datos sobre el cual se va a trabajar, contiene datos de clientes de tipo residencial, comercial e industrial. Por lo tanto, existen mediciones de consumo desde unos cuantos kWh mensuales hasta decenas de miles, que a su vez representan consumos fuera de rango (outliers). Los valores fuera de rango pueden afectar el rendimiento del algoritmo a desarrollarse por lo cual, su detección y eliminación son pasos necesarios a realizar.

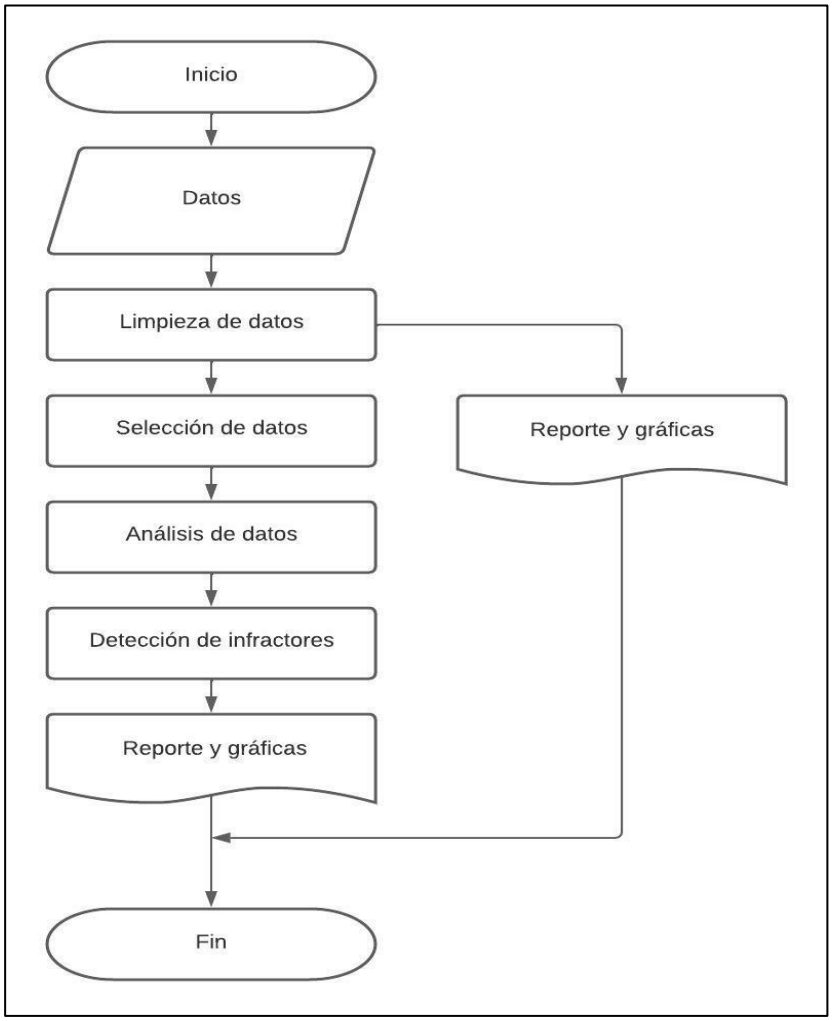
- **Valores nulos y duplicados;** si se encuentran casos en donde no se realizó la medición del consumo mensual de energía eléctrica de los usuarios, o en donde se repiten las filas, no se deben tomar en cuenta dichos usuarios con el objetivo de desarrollar un algoritmo robusto y con alto rendimiento.

3.6. Descripción general del algoritmo.

En el siguiente diagrama de flujo se muestra el proceso llevado a cabo para la identificación de los posibles usuarios que hurtan energía eléctrica.

- Paso 1: Inicio se ejecuta el algoritmo.
- Paso 2: Datos en donde se carga la información o la data set de los usuarios suscritos a la CNEL EP Santo Domingo para ser analizados.
- Paso 3: Limpieza de datos se refiere a separar datos nulos y en cero de los usuarios que reportan consumo en los 12 meses. Además, en el diagrama de flujo, figura 3.1. que parte hacia la derecha se puede generar un reporte total de los usuarios antes mencionados y generar una gráfica individual de cada uno con su código de cliente o en general como caja de bigotes o histogramas.
- Paso 4: Selección de datos no permite aplicar el algoritmo a la data set que ya queda sin los usuarios en cero o nulos.
- Paso 5: Análisis de datos se encarga de aplicar los métodos de inteligencia artificial escogidos para la detección de los posibles infractores.
- Paso 6: Detección de infractores en base al algoritmo de Machine Learning aplicado se detecta a los posibles infractores.
- Paso 7: Reporte y gráficas nos permite obtener un documento en formato Excel de los posibles usuarios infractores además de generar graficas de consumo en kWh individual.

Figura 3. 1. Diagrama de flujo para la identificación de posibles infractores.

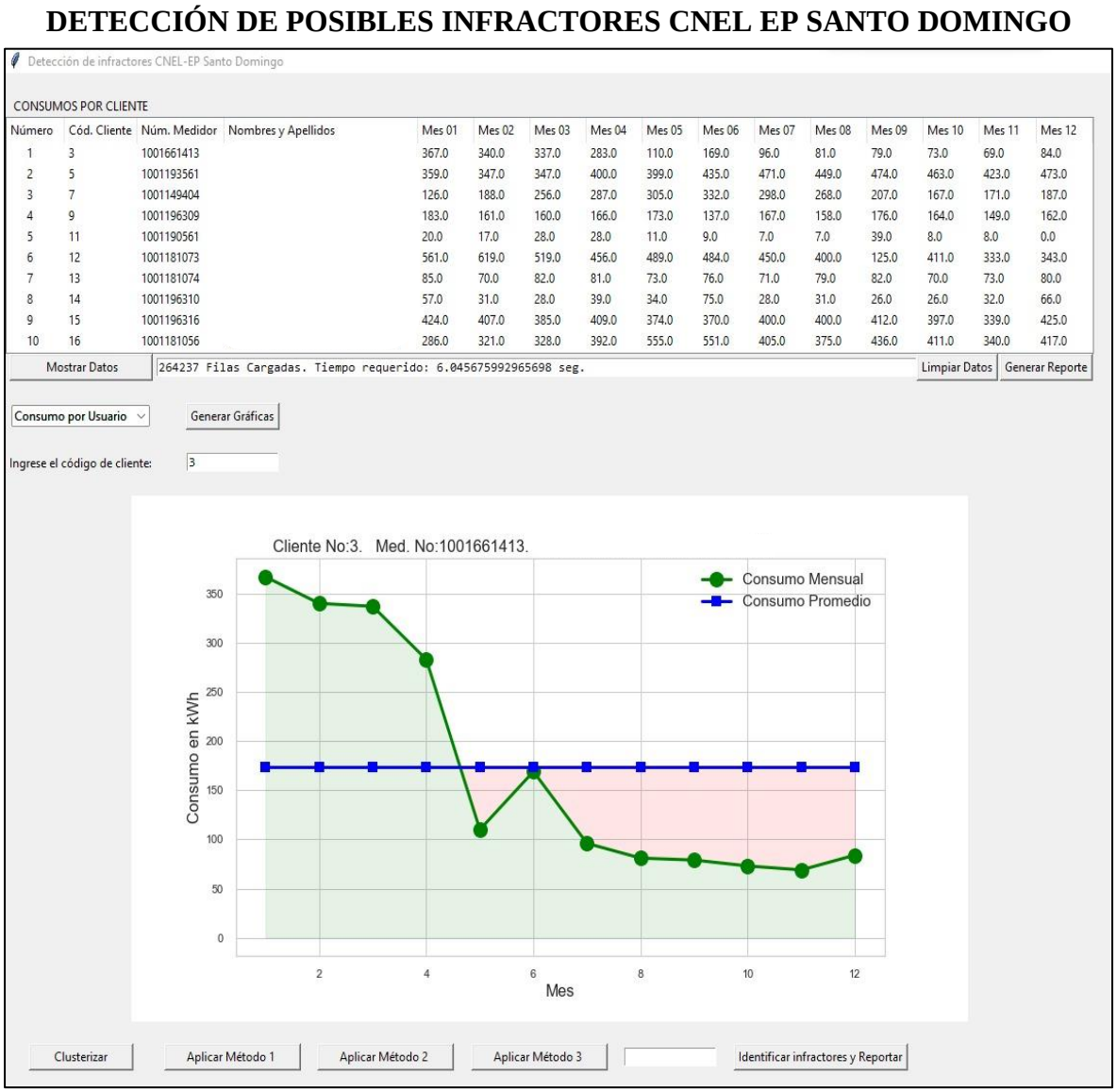


Fuente: (Word).
Elaborado por: Macaco R. – Pujota E. (2022).

3.6.1. Interfaz gráfica.

En orden de mejorar la experiencia y facilidad de uso del algoritmo, se decidió crear una interfaz gráfica que nos permite visualizar de manera grafica los datos generados además de diagramas para una mejor interpretación de los datos, siendo Python un lenguaje sencillo y fácil de aprender, la interfaz se muestra en la figura 3.2.

Figura 3. 2. Interfaz gráfica para visualización de datos e identificación de infractores.



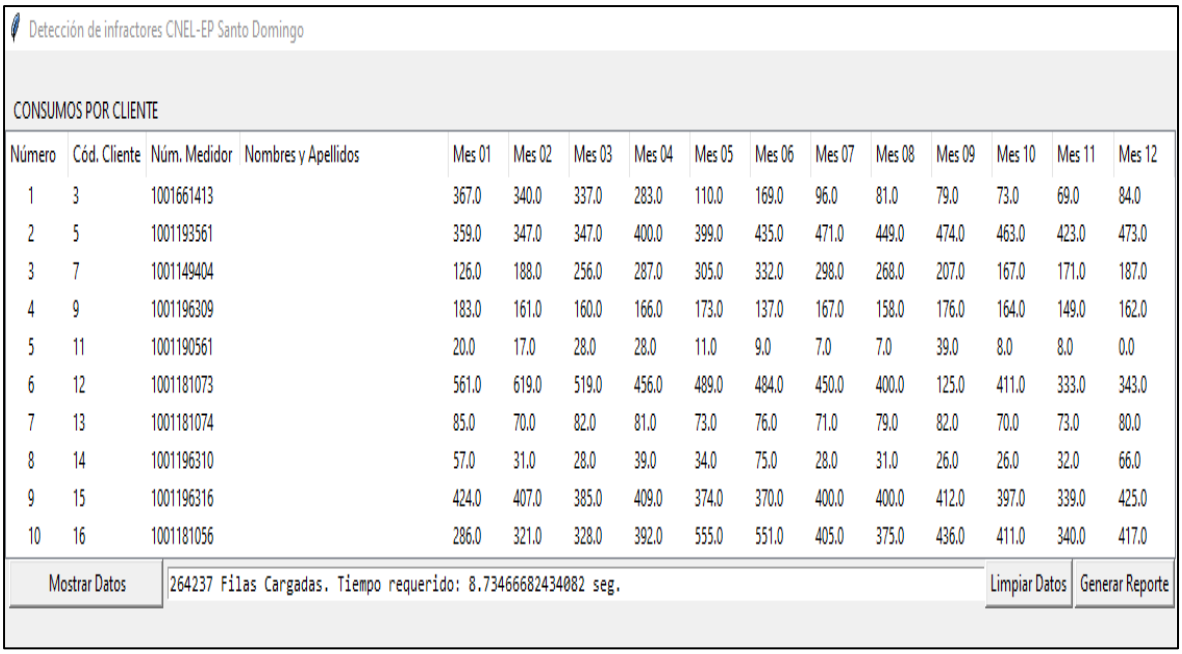
Fuente: (Python).
Elaborado por: Macao R. – Pujota E. (2022).

3.6.2 Importación de datos.

Los datos con un tamaño aproximado de 42 MB fueron importados dentro de PYTHON, una vez importados, los datos son mostrados en una tabla, como se describió anteriormente, el conjunto de datos posee 26 columnas, por lo cual, para su visualización, se escogieron las columnas más representativas: código de cliente, número de medidor, nombres y apellidos y los consumos de energía eléctrica durante los últimos 12 meses.

Tabla 3. 2. Datos cargados en Python y mostrados en forma de tabla.

DATA SET DE USUARIOS PERTENECIENTES A CNEL EP SANTO DOMINGO



Número	Cód. Cliente	Núm. Medidor	Nombres y Apellidos	Mes 01	Mes 02	Mes 03	Mes 04	Mes 05	Mes 06	Mes 07	Mes 08	Mes 09	Mes 10	Mes 11	Mes 12
1	3	1001661413		367.0	340.0	337.0	283.0	110.0	169.0	96.0	81.0	79.0	73.0	69.0	84.0
2	5	1001193561		359.0	347.0	347.0	400.0	399.0	435.0	471.0	449.0	474.0	463.0	423.0	473.0
3	7	1001149404		126.0	188.0	256.0	287.0	305.0	332.0	298.0	268.0	207.0	167.0	171.0	187.0
4	9	1001196309		183.0	161.0	160.0	166.0	173.0	137.0	167.0	158.0	176.0	164.0	149.0	162.0
5	11	1001190561		20.0	17.0	28.0	28.0	11.0	9.0	7.0	7.0	39.0	8.0	8.0	0.0
6	12	1001181073		561.0	619.0	519.0	456.0	489.0	484.0	450.0	400.0	125.0	411.0	333.0	343.0
7	13	1001181074		85.0	70.0	82.0	81.0	73.0	76.0	71.0	79.0	82.0	70.0	73.0	80.0
8	14	1001196310		57.0	31.0	28.0	39.0	34.0	75.0	28.0	31.0	26.0	26.0	32.0	66.0
9	15	1001196316		424.0	407.0	385.0	409.0	374.0	370.0	400.0	400.0	412.0	397.0	339.0	425.0
10	16	1001181056		286.0	321.0	328.0	392.0	555.0	551.0	405.0	375.0	436.0	411.0	340.0	417.0

Fuente: CNEL EP Unidad de Negocios Santo Domingo, (Python).
Elaborado por: Macao R. – Pujota E. (2022).

3.6.3. Limpieza de datos.

Una vez seleccionadas las columnas, es necesario realizar la limpieza de datos, lo cual nos permitirá encontrar datos anómalos, que a su vez podrían afectar el rendimiento del algoritmo desarrollado.

3.6.3.1. Identificación de valores nulos y duplicados.

Para identificar los valores nulos y duplicados, debemos encontrar usuarios cuyas mediciones de consumo de energía eléctrica no han sido realizadas, o no han sido registradas en el conjunto de datos y también nos da la oportunidad de encontrar casos en los que se repitan usuarios.

Tabla 3. 3. Ejemplo de usuarios con datos nulos.

REPRESENTACIÓN DE USUARIOS CON DATOS NULOS

CLICOD	MDENUMFAB	12	11	10	9	8	7	6	5	4	3	2	1
85	1810243028	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
103	139842	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
178	2110710781	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
224	809005996	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
328	1001151983	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
339	1001189545	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
458	1001210014	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
461	7100116412	Nan	Nan	Nan	Nan	Nan	Nan	Nan	Nan	Nan	Nan	Nan	Nan
609	1001216741	Nan	Nan	Nan	Nan	Nan	Nan	Nan	Nan	Nan	Nan	Nan	Nan
612	1001216734	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
663	1001212796	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
701	1001209394	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Fuente: CNEL EP Unidad de Negocios Santo Domingo, (Python).
Elaborado por: Macao R. – Pujota E. (2022).

Como se muestra en la figura 3.4, dentro del conjunto de datos existen valores nulos, por lo cual se procedió a eliminarlos. Se encontraron 22162 datos nulos, datos con consumo en 0 y datos repetidos, por lo cual nuestro conjunto de datos se redujo de 264237 a 242075 usuarios.

3.6.3.2 Consumos en cero.

Existen usuarios que registran un consumo energético mensual de cero, durante los últimos doce meses, estos datos no aportan valor al desarrollo del algoritmo desde el punto de vista del análisis de datos, por lo cual se decidió eliminarlos.

Tabla 3. 4. Representación gráfica de usuarios con datos de consumo en cero.

USUARIO CON DATOS EN CERO

Cód Clienet	Núm Medidor	Mes 01	Mes 02	Mes 03	Mes 04	Mes 05	Mes 06	Mes 07	Mes 08	Mes 09	Mes 10	Mes 11	Mes 12
7590	1001197406	283.0	294.0	292.0	295.0	300.0	287.0	309.0	286.0	301.0	297.0	277.0	274.0
7592	1001229865	14.0	20.0	18.0	16.0	21.0	14.0	16.0	20.0	17.0	18.0	14.0	13.0
7593	1001187089	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
7595	1001166445	462.0	475.0	428.0	449.0	436.0	340.0	383.0	3432.0	285.0	307.0	317.0	325.0
7597	1810251694	111.0	128.0	135.0	162.0	170.0	145.0	173.0	180.0	162.0	159.0	169.0	145.0
7598	1001187445	46.0	65.0	84.0	60.0	77.0	67.0	47.0	41.0	37.0	103.0	141.0	122.0
7599	1001187449	126.0	107.0	126.0	116.0	121.0	111.0	107.0	118.0	117.0	112.0	125.0	135.0

Fuente: CNEL EP Unidad de Negocios Santo Domingo, (Excel).
Elaborado por: Macao R. – Pujota E. (2022).

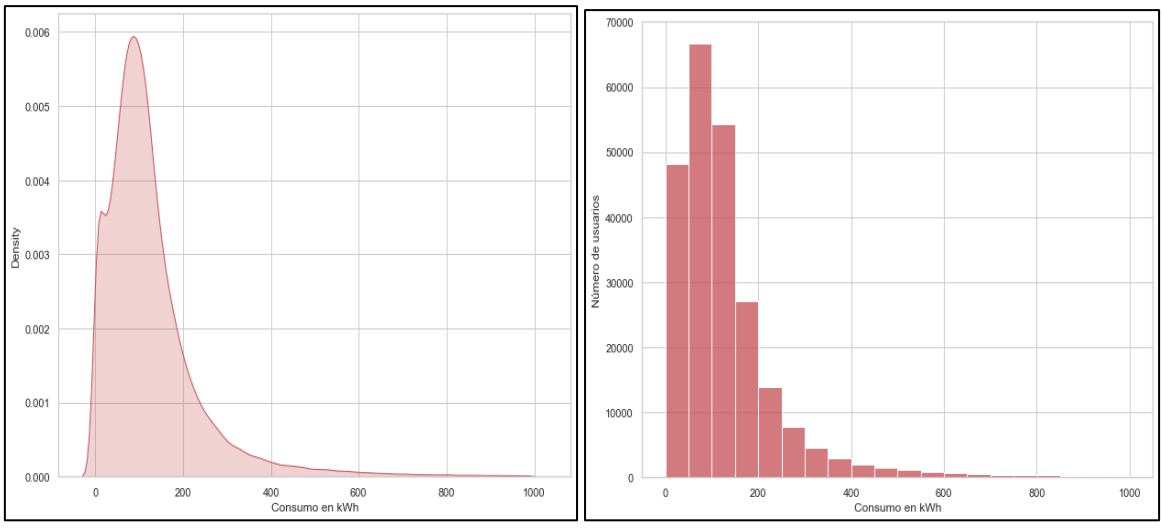
Se encontraron 5230 usuarios, cuyas mediciones de consumo de energía eléctrica son igual a cero durante los últimos 12 meses. Luego de la eliminación de estos datos, por tanto, el conjunto de datos se redujo a un total de 236845 usuarios.

3.6.3.3. Valores fuera de rango.

Para nuestro análisis los clientes con un consumo mensual promedio mayor o igual a 1000 kWh no fueron tomados en cuenta, debido a que solo un grupo minino de 2882 usuarios registra consumos superiores, por lo cual luego de su eliminación contamos con un total de 233963 usuarios.

El diagrama de densidad e histograma de frecuencia presentados en la figura 3.6 (a) y figura 3.6 (b) respectivamente, son un buen indicador de la distribución de datos, por medio de un análisis visual de las gráficas es claro que los datos están mayormente concentrados alrededor de un consumo mensual promedio de 100 kWh, y conociendo que la desviación estándar es igual a 120 kWh, se puede concluir que existen valores fuera de rango.

Figura 3. 3. De izquierda a derecha: a – Diagrama de densidad, b – Histograma de frecuencia.

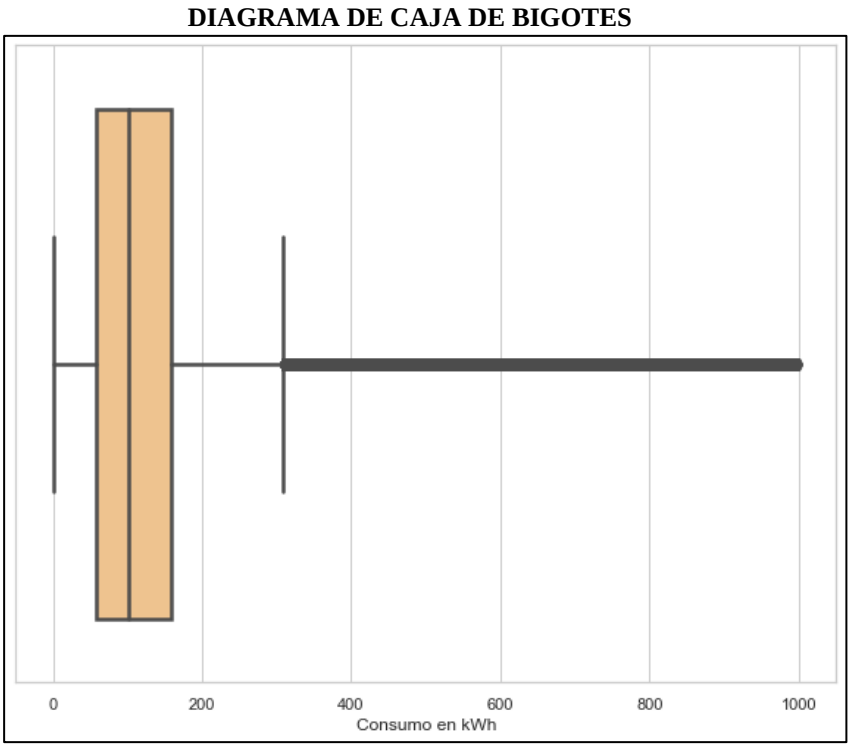


a *b*

Fuente: CNEL EP Unidad de Negocios Santo Domingo, (Python).
Elaborado por: Macao R. – Pujota E. (2022).

Los valores fuera de rango o “outliers”, pueden interferir y reducir el rendimiento y precisión de los algoritmos de Inteligencia Artificial, uno de los métodos más efectivos y de menor complejidad para la detección de valores fuera de rango son los diagramas de caja o “boxplots”. En la figura 3.7 se muestra el diagrama de caja del consumo promedio mensual de los usuarios.

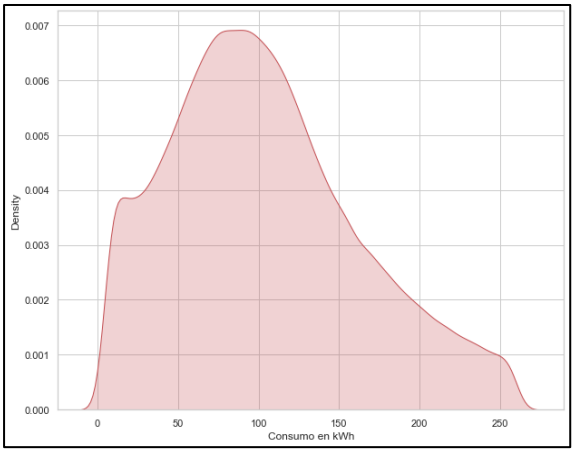
Figura 3. 4. Diagrama de caja del consumo promedio mensual.



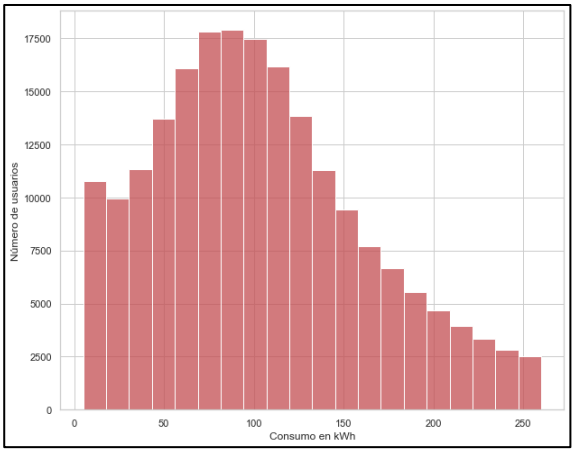
Fuente: CNEL EP Unidad de Negocios Santo Domingo, (Python).
Elaborado por: Macao R. – Pujota E. (2022).

Como se observa en la figura 3.7, existe una gran cantidad de valores fuera de rango, por lo cual se el análisis se enfocó a usuarios con consumos mensuales promedios menores que 260 kWh y mayores que 5 kWh. Debido a ello, los usuarios que no cumplen estos requerimientos fueron eliminados del conjunto de datos, resultando en la remoción de 31246 usuarios, dando como resultado un Dataset con 202717 usuarios. En la figura 3.8 (c) se observa que, al eliminar estos usuarios, ya no existen valores fuera de rango.

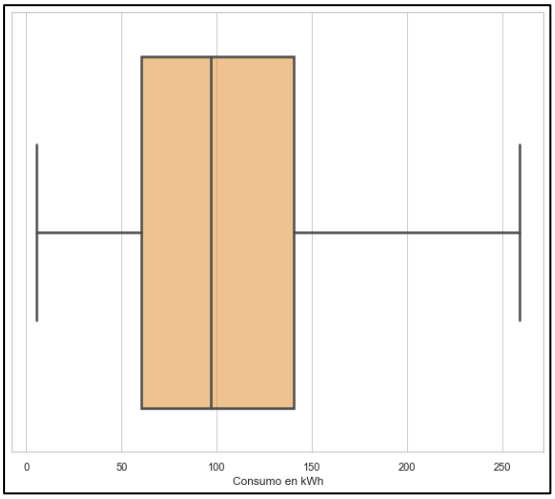
Figura 3. 5. Gráficas nuevas luego de la limpieza de datos; a – Diagrama de densidad, b - Histograma de frecuencia y c – Diagrama de caja.



a



b



c

Fuente: CNEL EP Unidad de Negocios Santo Domingo, (Python).
Elaborado por: Macao R. – Pujota E. (2022).

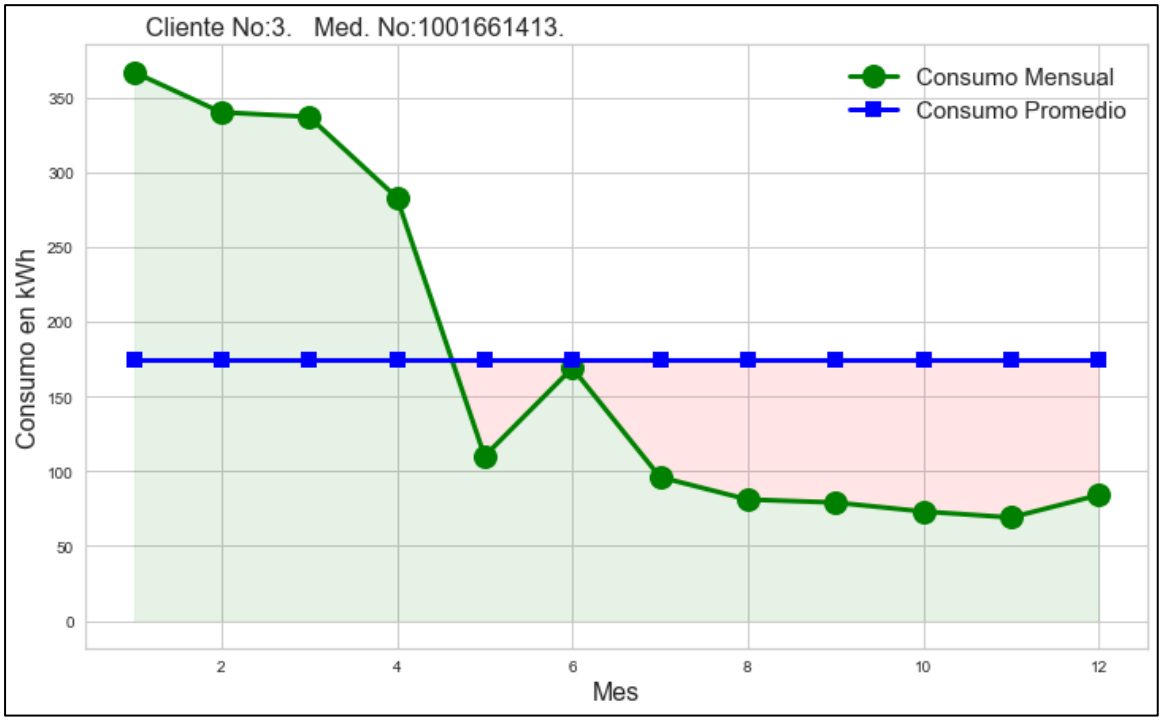
3.6.4. Algoritmo de Inteligencia Artificial para el análisis de datos.

Una vez los datos han sido limpiados y se tiene un conjunto de datos sin valores nulos, valores duplicados, valores en cero y valores fuera de rango, se puede proceder a desarrollar el algoritmo para el análisis de datos y consecuentemente para la detección de infractores.

3.6.4.1. Series temporales.

Debido a que los datos son generados mensualmente, es necesario tratarlos como series temporales, por lo tanto, su estudio es diferente al resto de las variables estadísticas porque el principal interés en este tipo de datos recae sobre la evaluación y análisis de sus cambios en función del tiempo.

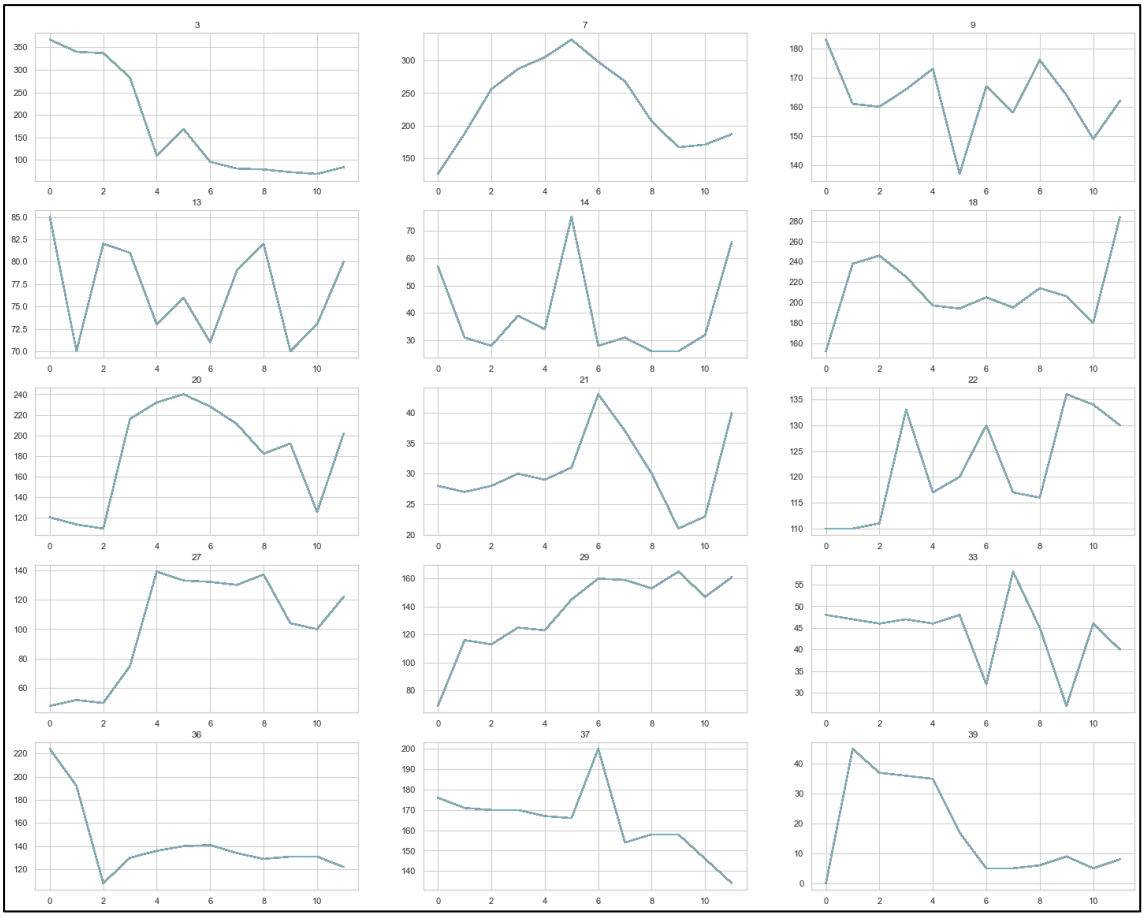
Figura 3. 6. Ejemplo de serie temporal de uno de los usuarios de la CNEL EP Unidad de Negocios Santo Domingo.



Fuente: CNEL EP Unidad de Negocios Santo Domingo, (Python).
Elaborado por: Macao R. – Pujota E. (2022).

En la figura 9, es posible observar (de color verde) la serie temporal generada a partir de los consumos mensuales de energía eléctrica. Este proceso fue realizado para los 202898 usuarios que forman parte del conjunto de datos.

Figura 3. 7. Ejemplo de series temporales de 15 usuarios.



Fuente: CNEL EP Unidad de Negocios Santo Domingo, (Python).
Elaborado por: Macao R. – Pujota E. (2022).

3.6.4.2. Selección del Algoritmo de Inteligencia Artificial.

Debido a que dentro de los datos no se dispone de etiquetas predefinidas (resultados deseados), es decir; dentro del conjunto de datos no está determinado si un usuario es infractor o no, se debe escoger un algoritmo de aprendizaje no supervisado. Los algoritmos de aprendizaje no supervisado nos permiten agrupar los datos, en este caso las series temporales de acuerdo a sus similitudes o patrones.

El objetivo del algoritmo será descubrir patrones desconocidos dentro del conjunto de datos, crear grupos o clústeres con los datos que presentan similitudes, en donde uno de esos grupos será el de los posibles infractores.

Existen varios algoritmos de aprendizaje no supervisado, sin embargo, un factor importante a tomar en cuenta, es que los datos están dados en forma de series temporales, por lo cual se eleva la complejidad del análisis. En este caso se debe seleccionar un algoritmo de "Clustering" o agrupamiento de datos para series temporales con el propósito de maximizar la similaridad de los datos dentro de los clústeres y minimizarla entre clústeres.

Basados en los criterios presentados anteriormente, se decidió seleccionar el algoritmo K-Means, el cual ha mostrado producir los resultados más precisos, por lo tanto, en el consecuente análisis es el algoritmo usado.

3.6.4.3. Normalización de datos.

Una parte primordial de la preparación de datos para Inteligencia Artificial, y en este caso en específico para Machine Learning, es la normalización de datos. El propósito de la normalización de datos es cambiar los valores numéricos del conjunto de datos a una escala común, sin distorsionar las diferencias en los rangos de los valores, y está dado por la siguiente fórmula:

$$z_i = \frac{x_i - \min(x)}{\max(x) - \min(x)}$$

Ecuación 3. 1

Dónde:

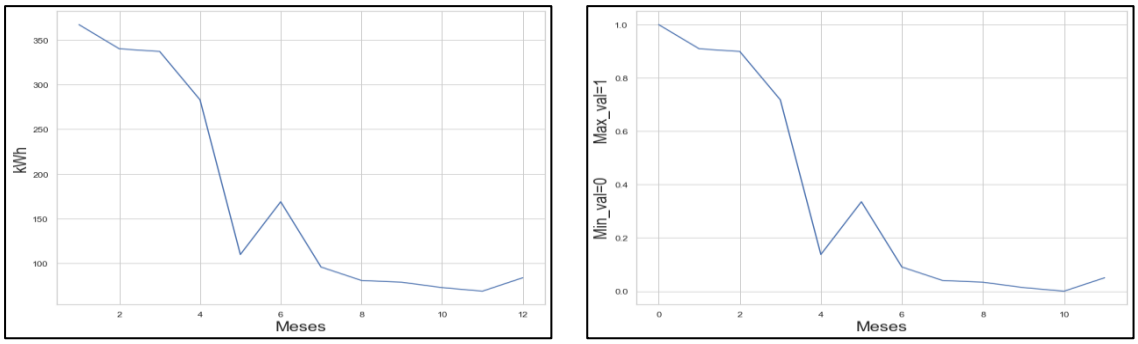
z_i : El $i^{ésimo}$ valor normalizado en el conjunto de datos.

x_i : el $i^{ésimo}$ valor en el conjunto de datos.

min(x) : el valor mínimo en el conjunto de datos.

max(x) : el valor máximo en el conjunto de datos.

Figura 3. 8. Ejemplo de normalización de datos para una serie temporal: a – Serie temporal sin normalizar, b – Serie temporal normalizada con valor mínimo igual a 0 y valor máximo igual a 1.



a *b*

Fuente: CNEL EP Unidad de Negocios Santo Domingo, (Python).
Elaborado por: Macao R. – Pujota E. (2022).

3.6.4.4. Determinación del número de clústeres o grupos.

Uno de los retos en agrupamiento de datos es la determinación del número de clústeres en un conjunto de datos. El algoritmo de agrupamiento K-Means que se determine previamente el número de clústeres, y en realidad es una tarea distinta al proceso de resolución del problema de agrupamiento.

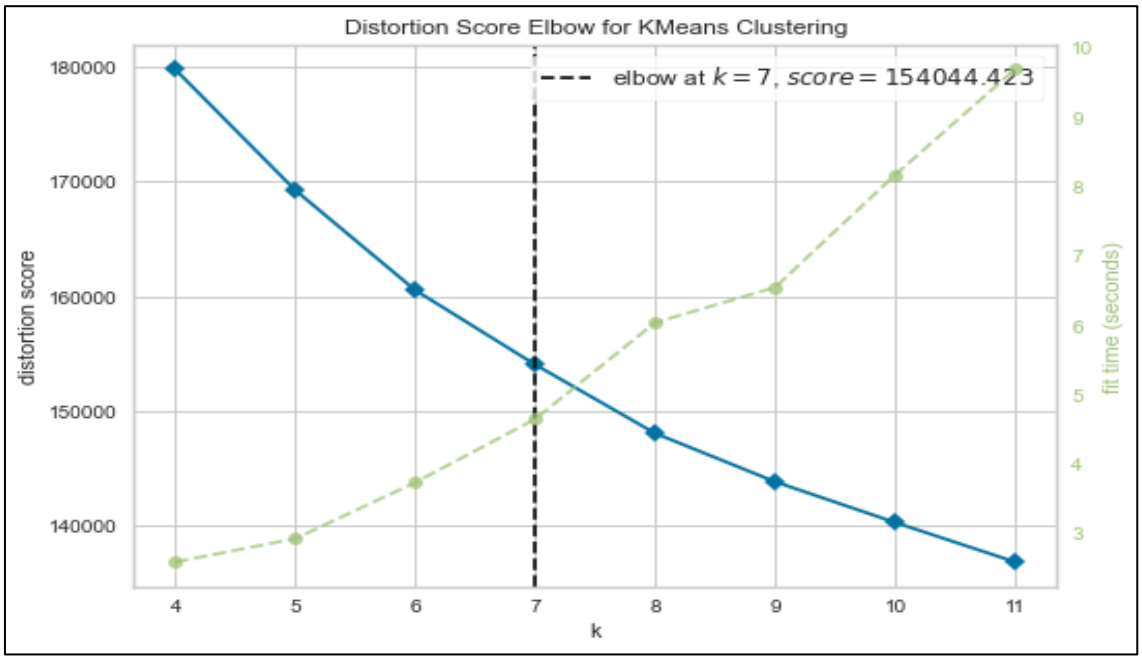
La selección correcta de ***k*** (número de clústeres) es frecuentemente ambiguo, con interpretaciones que dependen de la forma y la escala de la distribución de puntos en un conjunto de datos. Además, incrementando *k* indiscriminadamente siempre reducirá el error en el agrupamiento resultante. Por lo tanto, la selección óptima de *k* deberá lograr el equilibrio entre la máxima compresión de datos usando un solo clúster, y la máxima precisión por medio de la asignación de cada serie temporal a su propio clúster.

3.6.4.4.1. El método del codo.

El método del codo corre el algoritmo K-Means sobre el conjunto de datos para un rango de *k* determinado, y luego para cada uno de los valores *k* calcula la distorsión promedio para todos los clústeres. La distorsión se define como la suma de las distancias cuadradas desde cada punto a su centro asignado.

Cuando todas las métricas mencionadas anteriormente son graficadas, es posible determinar el mejor valor para *k*. Si la gráfica luce como un brazo, entonces el "codo" (el punto de inflexión de la curva) es el mejor valor *k*. El brazo puede estar hacia arriba o hacia abajo, pero si existe un fuerte punto de inflexión, es una buena indicación que el modelo subyacente se ajusta mejor en ese punto.

Figura 3. 9. Método del codo aplicado al conjunto de datos.



Fuente: CNEL EP Unidad de Negocios Santo Domingo, (Python).
Elaborado por: Macao R. – Pujota E. (2022).

Como se puede observar en la figura 3.12, el punto de inflexión estaría en 7, por lo cual nuestro conjunto de datos debe dividirse en 7 clústeres, de los cuales uno de ellos sería el grupo de posibles infractores.

3.6.4.5. K-Means.

Una vez definido el número de clústeres, se procedió a desarrollar el algoritmo de clusterización. El algoritmo K-Means empieza con un grupo de *k* centroides seleccionados aleatoriamente, los cuales son usados como punto de inicio para cada clúster, y luego realiza calculaciones iterativas (repetitivas) con el objetivo de optimizar las posiciones de los centroides.

La optimización de clústeres solo es detenida cuando:

- Los centroides se han estabilizado – no hay cambio en sus valores porque la clusterización se ha realizado exitosamente.
- Se ha alcanzado el número definido de iteraciones.

En el algoritmo K-Means, se mide la distancia o similitud entre los puntos del conjunto de datos y cada centroide inicializado. Basado en los valores encontrados, los puntos son asignados a lo centroides con la distancia mínima. Por lo tanto, el cálculo de esta distancia juego un papel vital en el algoritmo de agrupamiento.

Como es sabido, la distancia entre dos puntos puede ser computada por medio de diferentes técnicas disponibles, por lo tanto, escoger la técnica apropiada es una tarea que representa un gran reto. En este proyecto, se han probado tres técnicas disponibles para la medición de la similitud entre dos series temporales: Distancia Euclidiana, DTW Barycenter Averaging y Soft-DTW.

3.6.4.5.1. Distancia Euclidiana.

La distancia euclidiana computa la raíz de la diferencia de cuadrados entre las coordenadas de par de puntos.

$$Dist_{XY} = \sqrt{\sum_{k=1}^{m} \left(X_{ik} - X_{jk}\right)^2}$$

Ecuación 3. 2.

Una de las ventajas de este método es que la distancia entre cualesquiera dos puntos no se ve afectada por la adición de nuevos puntos al análisis, los cuales podrían ser valores fuera de rango [1].

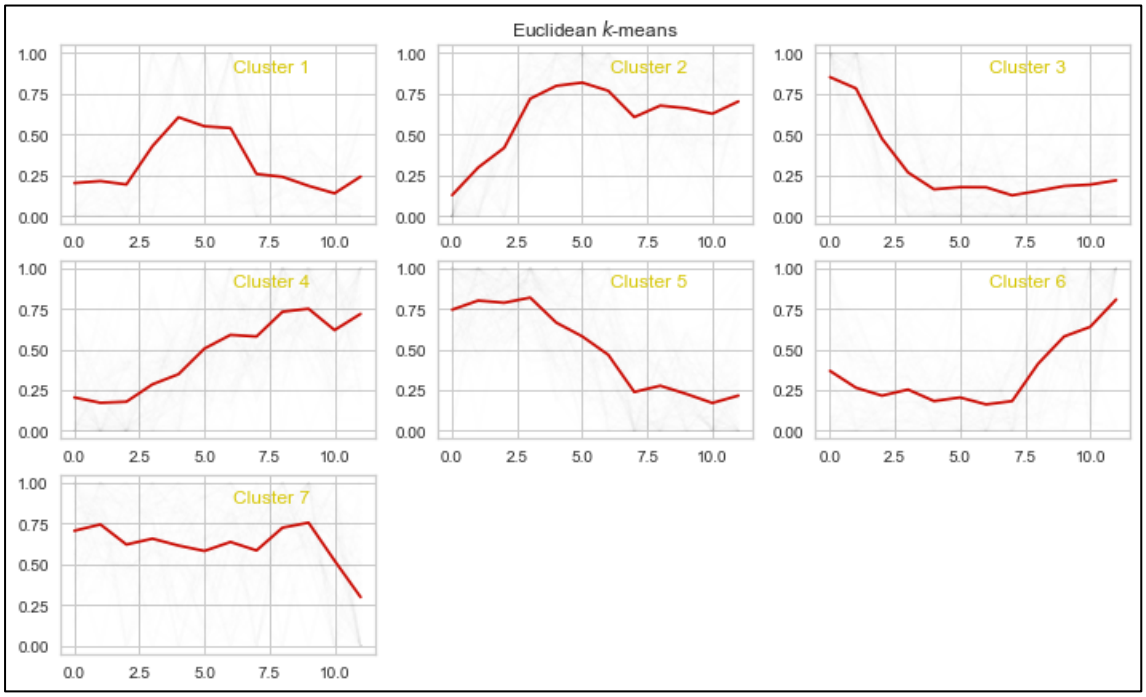
Sin embargo, las distancias pueden verse afectadas enormemente por diferencias en la escala entre las dimensiones de las cuales las distancias son calculadas.

Por ejemplo, si una de las dimensiones denota un consumo de energía eléctrica en kWh, y luego se convierten a Wh, la distancia euclidiana resultante se ve altamente afectada (sesgada por dimensiones que tienen una escala más grande), y consecuentemente, los resultados del análisis clústeres podrían variar.

Una de las mayores desventajas de esta técnica para encontrar similitudes entre series temporales, es que es invariable a los cambios de tiempo, ignorando la dimensión temporal de los datos. Si dos series temporales están altamente correlacionadas, pero una se desfasa incluso un solo paso de tiempo, la distancia euclidiana las mediría erróneamente como más separadas.

Aplicando el algoritmo K-Means al conjunto de datos, y definiendo la técnica de cómputo de similitud entre series temporales como distancia euclidiana se obtienen los resultados mostrados en la figura 3.14.

Figura 3. 10. Resultado del algoritmo de agrupamiento tomando como métrica de similitud la distancia euclidiana.



Fuente: CNEL EP Unidad de Negocios Santo Domingo, (Python).
Elaborado por: Macao R. – Pujota E. (2022).

3.6.4.5.2. DTW Barycenter Averaging (DBA).

En análisis de series temporales, dynamic time warping (DTW), es un algoritmo para medir la similitud entre dos secuencias temporales. Por medio de este algoritmo se encuentra la alineación óptima entre dos secuencias de valores numéricos, y captura las similitudes flexibles alineando las coordenadas dentro de las dos secuencias.

Sean **X** e **Y** dos series temporales con un número de elementos *m* y *n* respectivamente. La distancia DTW desde **X** hasta **Y** es formulada como el siguiente problema de optimización:

$$DTW_{XY} = \min \sqrt{\sum_{(i,j)\in\pi} d(x_i, y_i)^2}$$

Ecuación 3. 3.

Donde $\pi = [\pi_0, \dots, \pi_K]$ es la ruta que satisface las siguientes propiedades:

- Es una lista de pares de índices $\pi_k = (i_k, j_k)$ con $0 \leq i_k \leq m$ y $0 \leq j_k \leq n$
- $\pi_0 = (0, 0)$ y $\pi_K = (m-1, n-1)$
- Para todo $k > 0, \pi_k = (i_k, j_k)$ esta relacionado con $\pi_{k-1} = (i_{k-1}, j_{k-1})$ como se muestra:
    - $i_{k-1} \leq i_k \leq i_{k-1} + 1$
    - $j_{k-1} \leq j_k \leq j_{k-1} + 1$

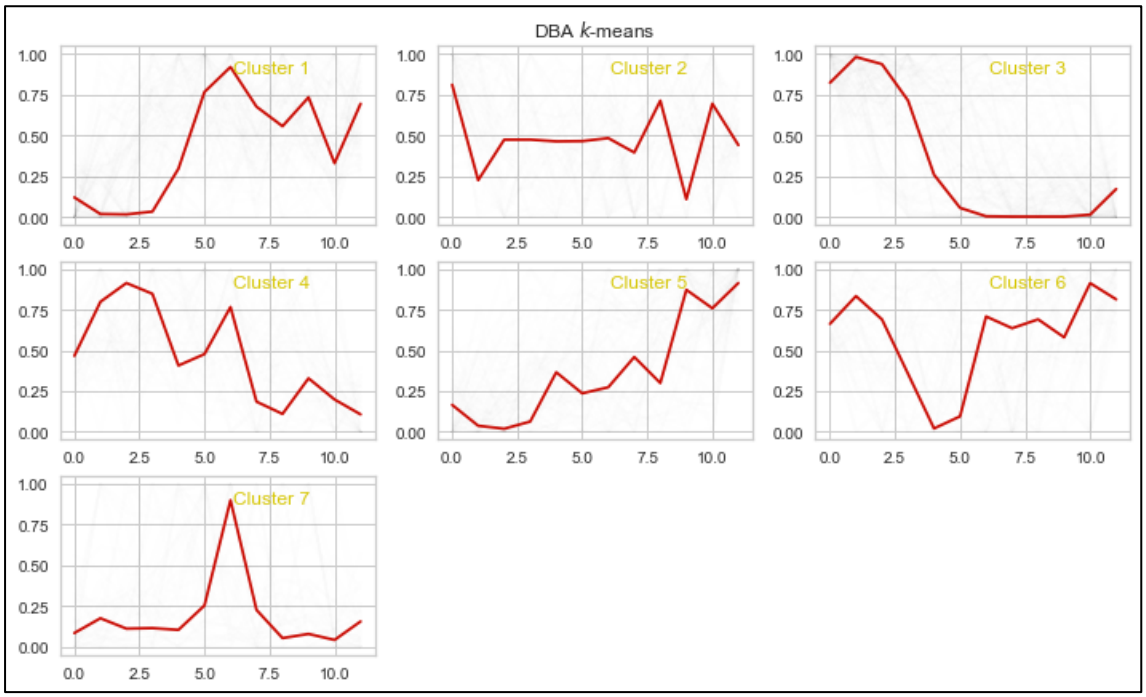
En resumen, DTW es calculado como la raíz cuadrada de la suma de las distancias cuadradas entre cada elemento en **X** y su punto más cercano en **Y**.

Esto crea una ruta deformada entre **X** e **Y** que alinea cada punto en **X** con el punto más cercano en **Y**. La ruta es una alineación temporal de tiempo que minimiza la distancia euclidiana entre series alineadas.

Desafortunadamente, una implementación directa de este algoritmo conlleva un costo muy elevado en tiempo de procesamiento, además de la imposibilidad de determinar si la solución encontrada es la óptima, debido a que se debe aproximar la serie temporal promedio.

Es allí cuando el algoritmo DBA, el cual está diseñado como un método global de aproximación supera este reto. DBA consiste en refinar iterativamente una serie temporal promedio inicial, para de este modo minimizar la suma de distancias DTW cuadradas entre el centroide y la serie temporal promedio. Como resultado, los centroides tienen una forma promedio que se parece a las series temporales miembros del clúster, a pesar que ocurran cambios temporales entre los miembros.

Figura 3. 11. Resultado del algoritmo de agrupamiento tomando como métrica de similitud la distancia DTW Barycenter Averaging.



Fuente: CNEL EP Unidad de Negocios Santo Domingo.
Elaborado por: Macao R. – Pujota E. (2022).

3.6.4.5.3. Soft-DTW.

Una de las fuertes limitaciones de la distancia DTW, es que puede ser computada solo en tiempo cuadrático usando programación dinámica, a pesar que el número de posibles alineaciones es exponencial en la longitud de dos series temporales. Soft-DTW propone reemplazar este mínimo por un mínimo suave. Al igual que el DTW original, el soft-DTW se puede calcular en tiempo cuadrático mediante programación dinámica.

Sin embargo, la principal ventaja de soft-DTW radica en el hecho de que es diferenciable en todas partes y que su gradiente también se puede calcular en tiempo cuadrático. Esto permite usar soft-DTW para promediar series temporales o como una función de pérdida, entre una serie de tiempo real y una serie de tiempo predicha por una red neuronal, entrenada de extremo a extremo usando retropropagación. La siguiente es la definición formal de Soft-DTW:

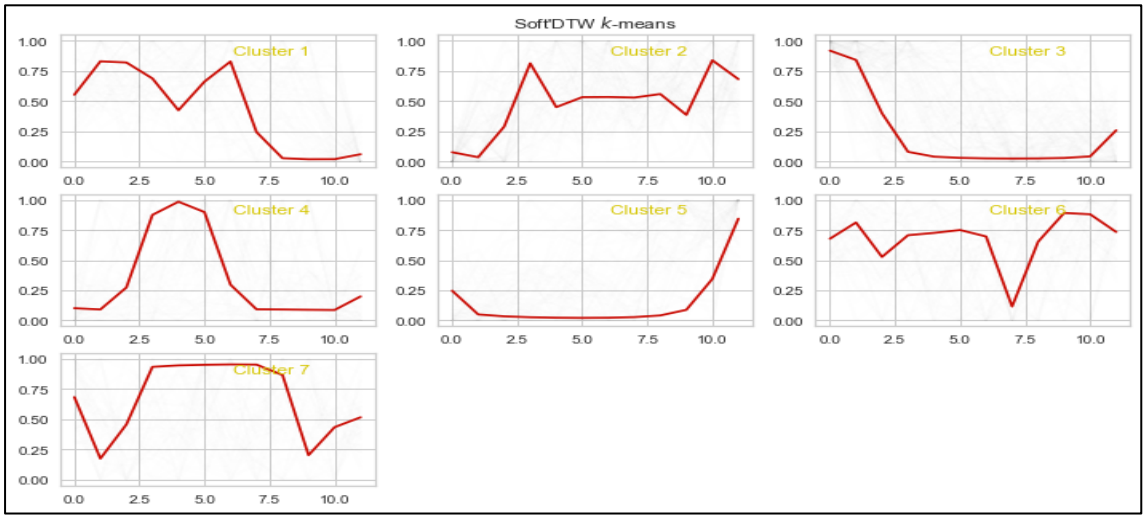
$$Soft - DTW^{\gamma}{}_{XY} = min^{\gamma}_{\pi \in A(X,Y)} \sum_{(i,j)\in\pi} d(X_i, Y_j)^2$$

Ecuación 3. 4.

Dónde min^{γ} es el operador soft-min parametrizado por el factor de suavidad γ.

Soft-DTW no es invariable a los cambios de tiempo, como lo es DTW. Teniendo que x es una serie de tiempo que es constante excepto por un motivo que ocurre en algún punto de la serie, y denotamos por x_{+k} una copia de x en la que el motivo se desplaza temporalmente por k marcas de tiempo. Entonces la cantidad $|soft - DTW^{\gamma}(x, x_{+k}) - soft - DTW^{\gamma}(x, x)|$ crece linealmente con γk^2. La razón detrás de esta sensibilidad a los cambios de tiempo es que soft-DTW proporciona una calificación de similitud promedio ponderado en todas las rutas de alineación (donde se asignan pesos más fuertes a mejores rutas), en lugar de centrarse en la mejor alineación única como se hace en DTW.

Figura 3. 12. Resultado del algoritmo de agrupamiento tomando como métrica de similitud la distancia Soft-DTW.



Fuente: CNEL EP Unidad de Negocios Santo Domingo, (Python).
Elaborado por: Macao R. – Pujota E. (2022).

3.6.5. Presentación de resultados.

Una vez que el algoritmo detectó los posibles infractores usando tres métricas diferentes, solo se toman en cuenta los usuarios que hayan sido detectados por los tres métodos, y se genera un conjunto de datos. A su vez el conjunto de datos se guarda como un archivo .xlsx (archivo de Excel), cada hoja de Excel tiene como nombre la subestación a la cual están conectados los posibles infractores y el porcentaje de posibles infractores (en relación a todos los usuarios conectados a esa subestación).

Tabla 3. 5. Ejemplo de presentación de resultados en forma de archivo Excel.

	CLICOD	MDENUMFAB	SGCORX	SGCORY	USOCOD	CLIPRVCODP	CLIRLSCOD
0	74	1001620277	704027,8998	9972183,289	RD	23	153
7	475	1810244882	703159,6771	9972529,868	CO	23	136
10	522	1001230630	703288,6356	9972687,84	RD	23	136
11	530	1001205941	703297,175	9972752,839	RD	23	136
12	531	1001205934	703297,445	9972755,728	CO	23	136
13	560	1001622146	703304,0234	9972623,175	RD	23	144
14	590	1001173508	703345,1944	9972720,629	TP	23	144
15	600	1710091659	703376,7931	9972821,334	RD	23	144
16	641	1001210894	703383,2429	9972558,657	RD	23	144
66	2340	50276949	702930,6303	9971782,926	CO	23	151
80	2891	1001623174	704158,9886	9973102,382	CO	23	165

SUBESTACIONES	PÉRDIDAS
EL CENTENARIO	**0.047%**
NO ASIGNADOS	**----**
PAMBILES	**0.135%**
QUEVEDO	**0.0208%**

Fuente: CNEL EP Unidad de Negocios Santo Domingo, (Excel).
Elaborado por: Macao R. – Pujota E. (2022).

3.7. Instrumentos de investigación.

Como instrumentos se utilizó un conjunto de datos provisto por la CNEL EP Unidad de Negocios Santo Domingo, al no existir investigaciones previas basadas en el enfoque aplicado en el presente proyecto, para el desarrollo del algoritmo solamente se usó la documentación provista por los desarrolladores del lenguaje de programación seleccionado.

3.8. Tratamiento de datos.

EL procesamiento de datos y el desarrollo del algoritmo fue realizado completamente en el lenguaje de programación Python, debido a su gran cantidad de librerías para del manejo de datos y desarrollo de aplicaciones de Inteligencia Artificial.

3.9. Recursos humanos y materiales.

Tabla 3. 6. Recursos humanos y materiales.

Humano				
Ítem	**Detalle**	**Descripción**	**Costo Unitario**	**Total**
1	480H	Macao Richard	$10	$4800
2	480H	Pujota Edison	$10	$4800
3	90H	Ing. Yadyra Ortiz	$10	$900
			Total $	$10500
Material				
Ítem	**Detalle**	**Descripción**	**Costo Unitario**	**Total**
1	1 U	Resma de hojas	$5	$5
2	1 U	Memoria USB 32GB	$15	$15
3	3 U	Computador personal	$1200	$3600
			Total $	$3620

Fuente: (Word).
Elaborado por: Macao R. – Pujota E. (2022).

CAPÍTULO IV

RESULTADOS Y DISCUSIÓN

4.1. Resultados.

Se presentan los resultados obtenidos al aplicar el algoritmo de inteligencia artificial para la detección de posibles infractores que hurtan energía eléctrica en la unidad de negocios CNEL EP Santo Domingo, permitiéndonos además generar un reporte en formato Excel y graficarlos geográficamente en ArcGIS.

4.1.1. Reporte de contribución de pérdidas de energía proporcionado por el departamento de control de energía CNEL EP Santo Domingo.

En base a los datos obtenidos de las de pérdidas totales de energía en porcentajes de la Unidad de Negocio CNEL EP Santo Domingo, conformado por 6 cantones de los cuales en la siguiente tabla 4.1. Se puede evidenciar que el cantón Flavio Alfaro presenta el 31,4%, de pérdidas, debido a que tiene una disponibilidad inferior que los otros cantones y por ende tiende a presentar mayores pérdidas de energía. Esto se debe a que es el cantón que posee un menor número clientes suscritos a la empresa distribuidora de energía eléctrica y ocupa el penúltimo lugar en la contribución económica para la empresa eléctrica.

Tabla 4. 1. Contribución de pérdidas de energía CNEL EP Santo Domingo.

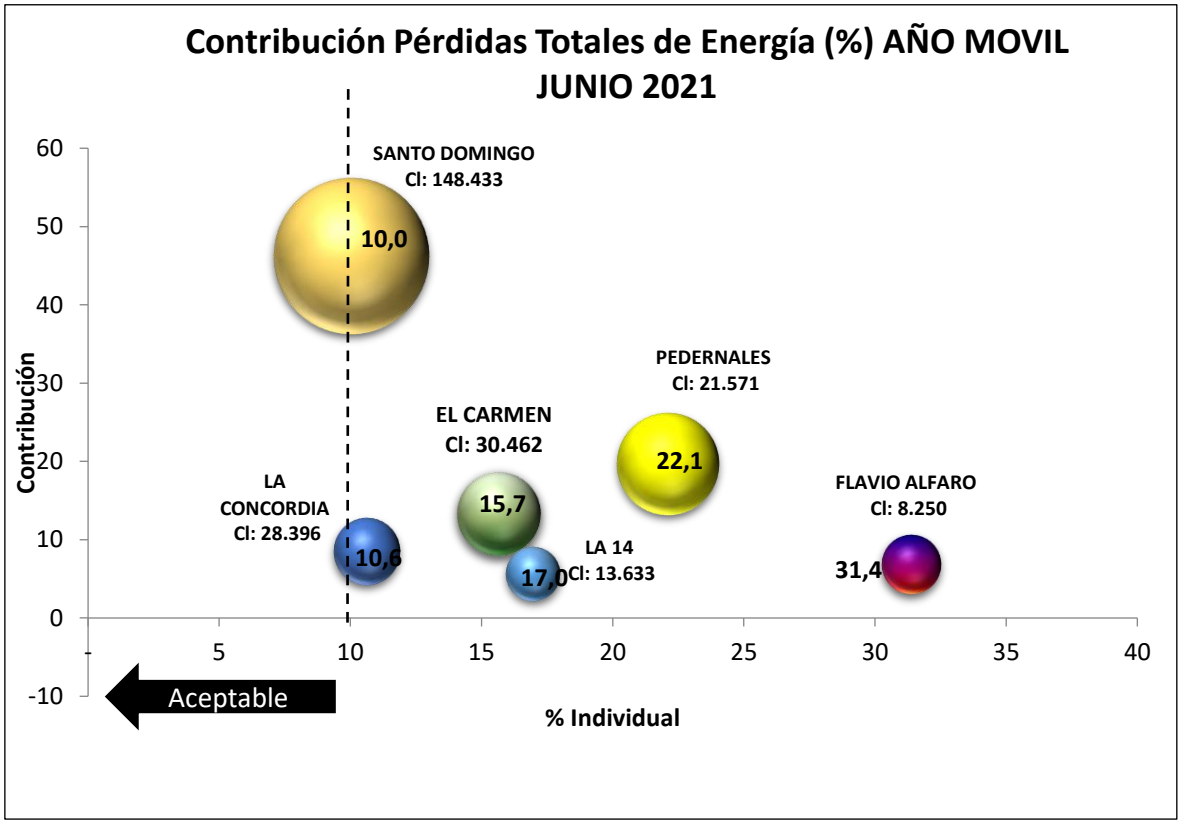
Contribución Pérdidas de Energía (%) AÑO MOVIL JUNIO 2021					
CANTÓN	**CLIENTES**	**DISPONIBILIDAD MWh**	**PÉRDIDAS MWh**	**% PÉRDIDAS**	**% CONTRIBUCIÓN**
EL CARMEN	30430	88110	13813	15,7	13,3
FLAVIO ALFARO	8261	22494	7062	31,4	6,8
LA CONCORDIA	28456	82616	8784	10,6	8,4
PEDERNALES	21512	92386	20433	22,1	19,6
SANTO DOMINGO	148608	478887	48090	10,0	46,2
LA 14	13617	34491	5857	17,0	5,6
TOTAL	**250884**	**798984**	**104038**	**13,0**	**100,0**

Fuente: CNEL EP Unidad de Negocios Santo Domingo, (Word).
Elaborado por: Macao R. – Pujota E. (2022).

4.1.1.1. Diagrama de esferas en base a la contribución de pérdidas de energía eléctrica.

En la siguiente grafica denominada gráfica de bolitas proporcionada por la CNEL EP Santo domingo, se puede visualizar de forma gráfica las pérdidas de energía de los 6 cantones, que se describen en la tabla 4.1. Se observa que el cantón Santo Domingo posee el mayor número de usuarios con respecto a los demás cantones, de tal manera que su disponibilidad también es mayor y por ende existen menores pérdidas comerciales para la empresa distribuidora. En la figura 4.1. Se visualiza mediante el diagrama de esferas de una forma más didáctica la contribución individual de cada cantón.

Figura 4. 1. Representación gráfica de esferas de la contribución de pérdidas totales de energía (%).

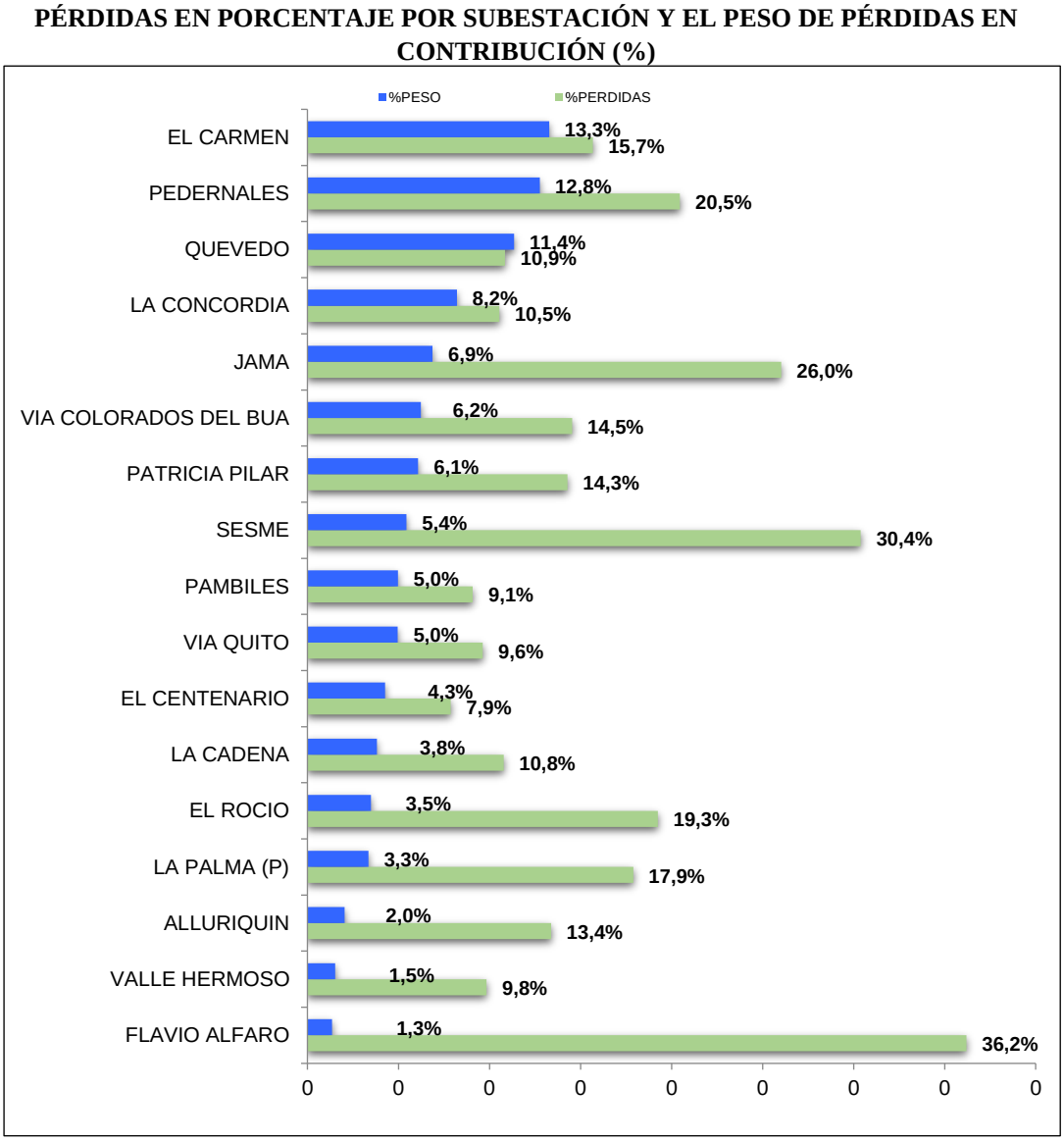


Fuente: CNEL EP Unidad de Negocios Santo Domingo, (Excel).
Elaborado por: Macao R. – Pujota E. (2022).

4.1.2. Reporte de pérdidas por subestación de la distribuidora CNEL EP Santo Domingo.

En la figura 4.2. Se detalla las pérdidas por subestaciones, la subestación Flavio Alfaro presenta mayor porcentaje de perdidas, en comparación con la subestación el centenario quien es el que tiene menor pérdidas de energía con un 7,9 % de pérdidas para la empresa distribuidora.

Figura 4. 2. Representación gráfica de perdidas por subestación de la CNEL EP Santo Domingo.



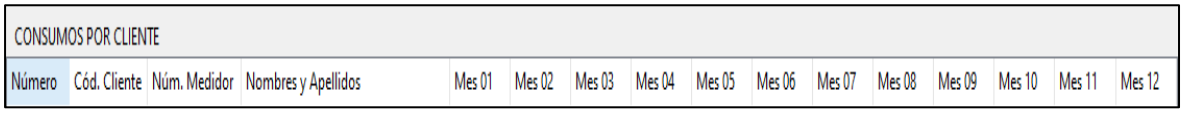
Fuente: CNEL EP Unidad de Negocios Santo Domingo, (Excel).
Elaborado por: Macao R. – Pujota E. (2022).

4.1.3. Interfaz gráfica del algoritmo para la obtención de reportes y gráficas de los posibles infractores.

A continuación, se presenta la interfaz gráfica para visualizar los diferentes parámetros que son:

- Número de cliente.
- Código de cliente.
- Nombre del cliente.
- Número de medidor.
- Consumo individual de 12 meses.

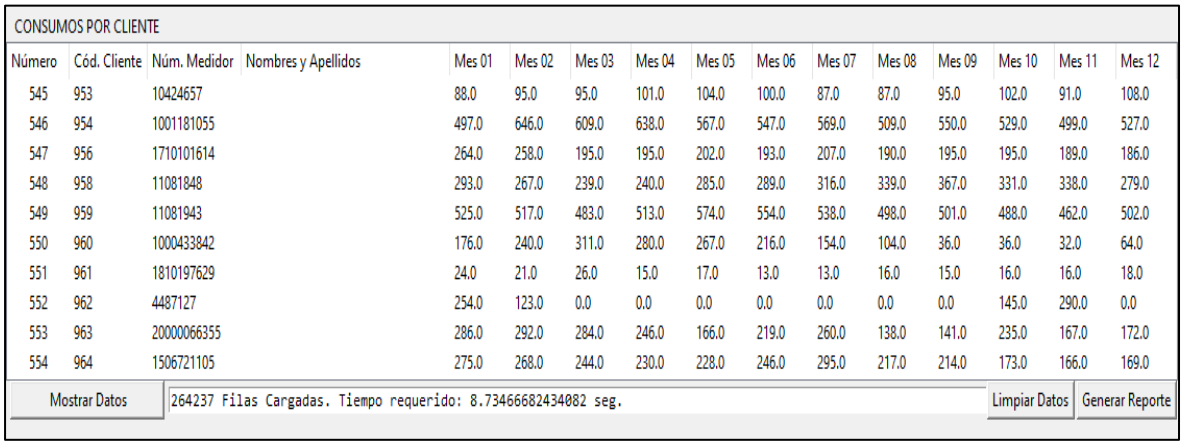
Figura 4. 3. Interfaz gráfica de los parámetros de usuarios suscritos a la CNEL EP Santo Domingo.



Fuente: (Python).
Elaborado por: Macao R. – Pujota E. (2022).

Al pulsar el botón “mostrar datos” el algoritmo carga nuestro archivo de Excel que contiene la información de cada uno de los usuarios para luego clasificar los datos en dos grupos siendo estos: clientes que poseen consumo mayor a 0 de 12 meses y clientes que tienen los consumos en 0, o que no registran consumo en uno o varios meses.

Tabla 4. 2. Representación en Python de los usuarios con el consumo de 12 meses.



Fuente: (Python).
Elaborado por: Macao R. – Pujota E. (2022).

Además, en la figura 4.4. Se visualiza los clientes suscritos a la CNEL EP Santo Domingo, conformado por 264237 clientes, y el tiempo que demora el algoritmo en cargar esta base de datos es alrededor de 8.734 segundos y este tiempo depende de la capacidad de la máquina en la cual se esté ejecutando el algoritmo.

Se puede obtener mediante el botón "generar reporte" un archivo en formato (.csv), que contiene a los usuarios que presentan el consumo mensual "0" en KWh o clientes que no registran consumo los cuales posiblemente sean usuarios prepagos y por ende tienen dicho consumo.

Tabla 4. 3. Representación de datos nulos y datos cero.

CLICOD	MDENUMFAB	12	11	10	9	8	7	6	5	4	3	2	1
85	1810243028	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
103	139842	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
178	2110710781	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
224	809005996	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
328	1001151983	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
339	1001189545	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
458	1001210014	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
461	7100116412												
609	1001216741	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

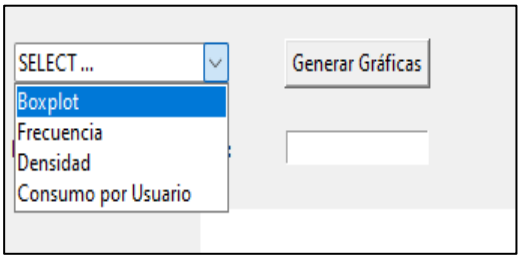
Fuente: (Python).
Elaborado por: Macao R. – Pujota E. (2022).

El reporte de datos nulos o en cero, contiene un total de 27392 clientes, los cuales no son tomados en el algoritmo para comprobar si son usuarios infractores o no, debido a que, para ser más eficiente en los resultados, solo tomamos los clientes que tienen los 12 meses de consumo.

4.1.4. Resultados mediante gráficas de los usuarios analizados por el algoritmo de inteligencia artificial.

En la figura 4.6. Se observa una lista de las diferentes gráficas que el algoritmo permite mostrar para un mejor entendimiento el comportamiento de consumo individual de cada uno de los usuarios, permitiéndonos de tal manera generar en la gráfica el consumo en kWh de 12 meses por medio del "código de cliente".

Figura 4. 4. Lista de gráficas utilizadas en Python.

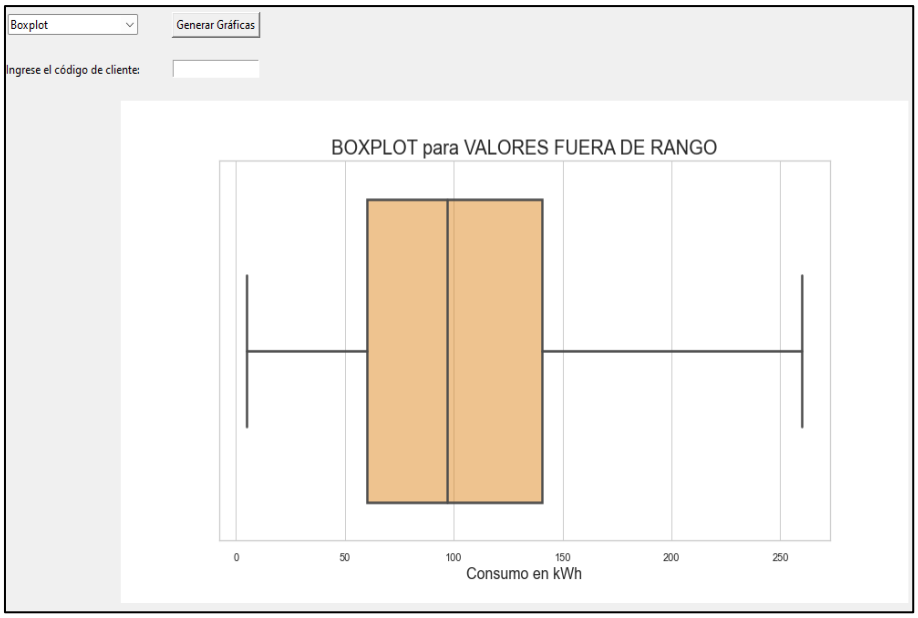


Fuente: (Python).
Elaborado por: Macao R. – Pujota E. (2022).

4.1.4.1. Resultado de los usuarios mediante el diagrama BoxPlot.

En la figura 4.7. Se puede visualizar el diagrama caja de bigotes en el cual se indican los valores atípicos que no serán tomados para la detección de los posibles infractores, de tal manera que se tomaran a los usuarios con consumo mayor a 5 kWh y menores a 260 kWh, que son clientes de categoría residencial y comercial, tomando en cuenta que son los más propensos a generar conexiones clandestinas para su bien común.

Figura 4. 5. Diagrama Boxplot para valores fuera de rango.

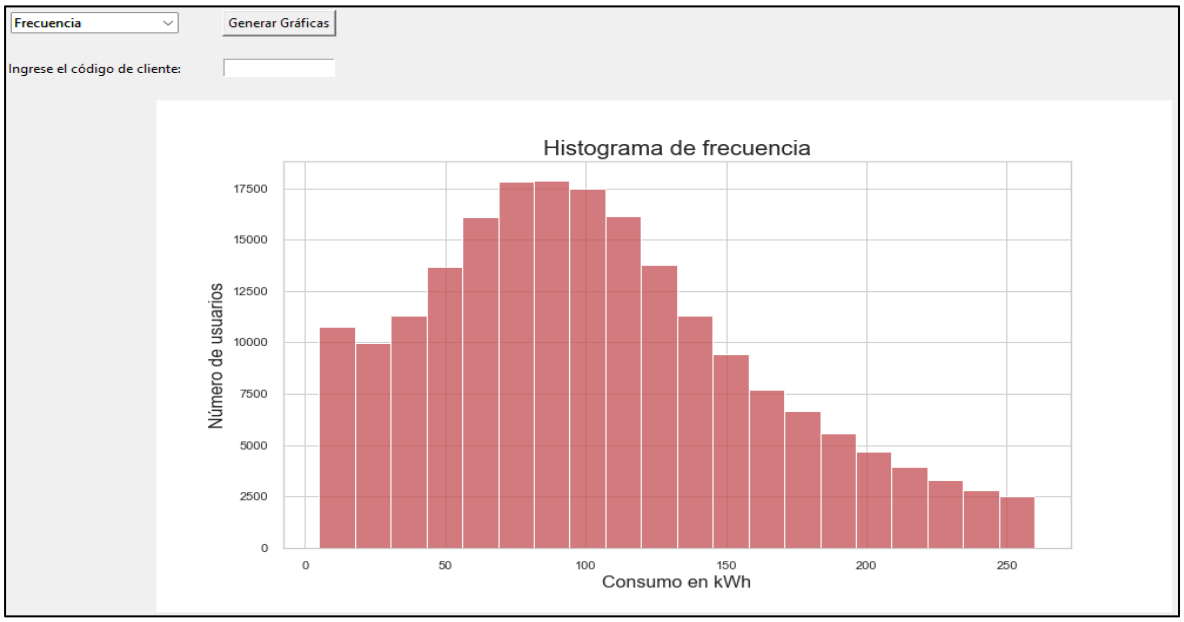


Fuente: (Python).
Elaborado por: Macao R. – Pujota E. (2022).

4.1.4.2. Resultado de los usuarios mediante el histograma de frecuencia.

Se visualiza en la figura 4.8. Un gráfico de barras el cual nos permite comprender como se distribuyen los usuarios en función de su consumo en kWh, en donde el punto máximo de usuarios es de 17600 con consumo promedio de 90 kWh a 95 kWh. Aproximadamente 2500 son los usuarios que consumen alrededor de 260 kWh, que es el consumo mínimo, además estos datos nos van a permitir la detección de posibles infractores en la unidad de negocios CNEL Santo Domingo.

Figura 4. 6. Representación de un histograma de frecuencia en base al consumo de usuarios de la CNEL EP Santo Domingo.

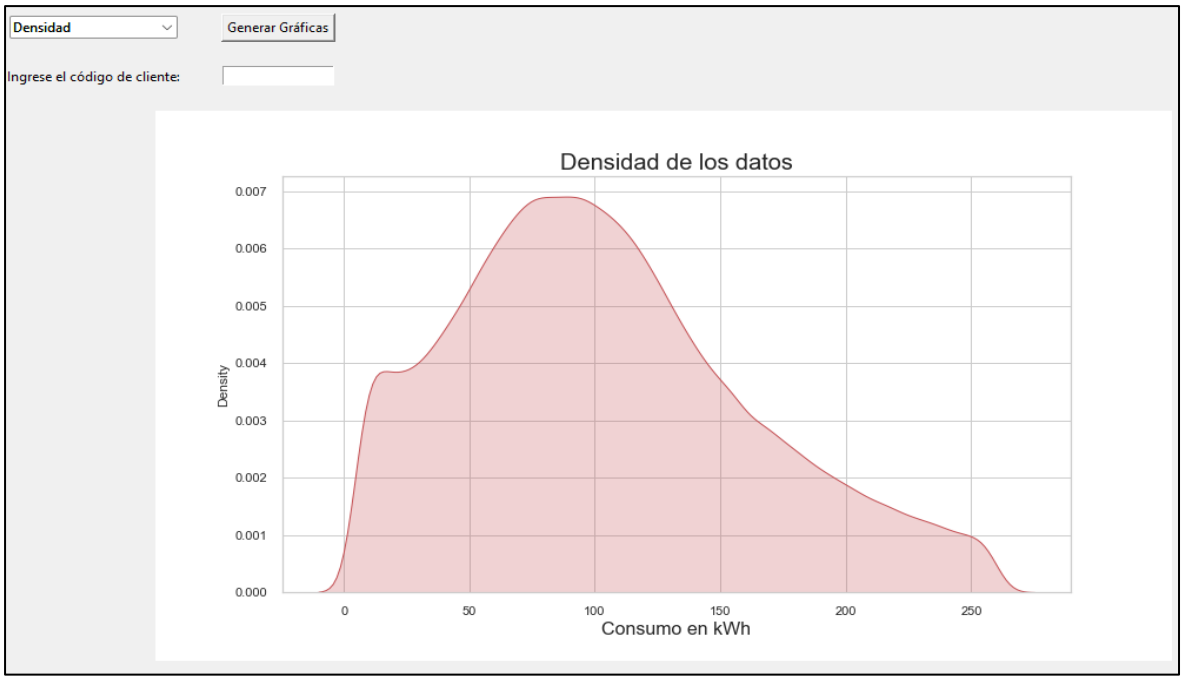


Fuente: (Python).
Elaborado por: Macao R. – Pujota E. (2022).

4.1.4.3. Resultado de los usuarios mediante el diagrama de densidad.

El diagrama de densidad nos permite visualizar a los clientes en un intervalo de tiempo la distribución mediante el consumo mensual que han presentado, en donde el límite de densidad de los usuarios es alrededor del 0.007 en función del consumo en KWh. como se muestra en la siguiente figura 4.9.

Figura 4. 7. Diagrama de densidad de los datos para la detección de posibles infractores.



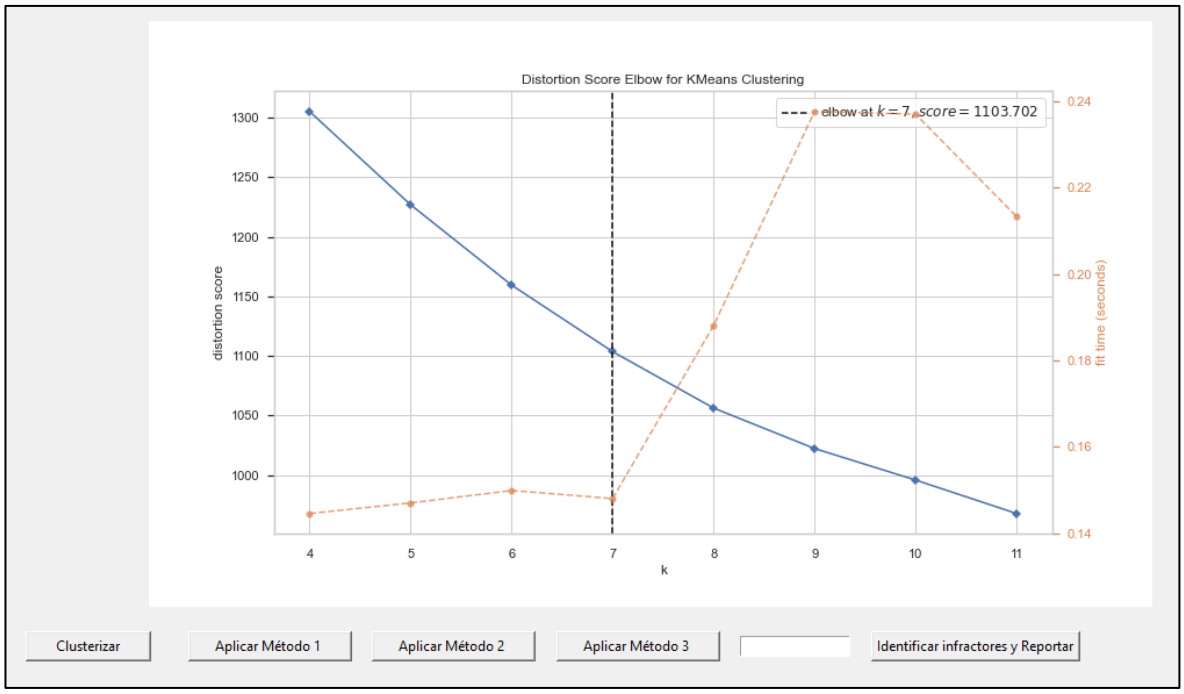
Fuente: (Python).
Elaborado por: Macao R. – Pujota E. (2022).

4.1.5. Detección de infractores como base de muestra previo a la ejecución total de usuarios.

Se toma como muestra la cantidad de 1000 usuarios para la detección de posibles infractores desde la base de datos que consta con cada parámetro individual de los clientes suscritos en la unidad de negocios CNEL EP Santo Domingo.

La ejecución del algoritmo se lo realiza mediante el botón “Clusterizar”, el cual permite determinar en cuantos grupos la data set va hacer distribuida, para posterior aplicar 3 métodos basados en la clusterización para filtrar cada uno de los usuarios y de esta manera realizar la medición entre dos series temporales en base a su similitud.

Figura 4. 8. Representación gráfica del método del codo para conocer el número de clusters.



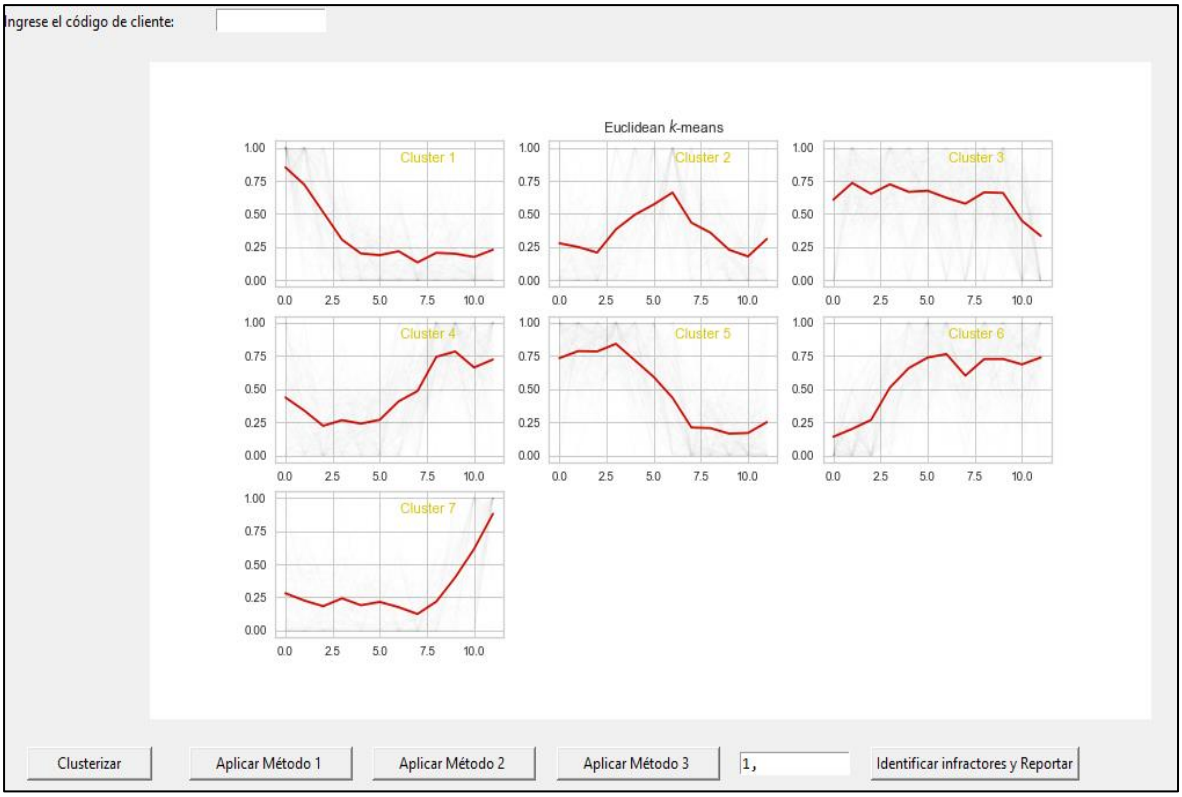
Fuente: (Python).
Elaborado por: Macao R. – Pujota E. (2022).

En la figura 4.8. Se observa que al momento de aplicar el algoritmo de inteligencia artificial de Machine Learning por clusterización, se utiliza K-Means, debido a que ha mostrado resultados más precisos, arrojando en este caso de muestra K=7, que comprende a que existen 7 grupos en los cuales los 1000 usuarios fueron distribuidos según su similitud.

4.1.5.1. Resultado obtenido al aplicar el método 1 "Distancia Euclidiana" en el algoritmo de K-Means de muestra.

Se procede a ejecutar el primer método con una muestra de 1000 usuarios, se escoge el clúster # 1, los usuarios de este grupo tienen un consumo promedio aproximado de 90 kWh y en el tercer mes del año su consumo decae paulatinamente manteniéndose casi constate hasta el mes #12.

Figura 4. 9. Resultado obtenido al aplicar el método 1 “Distancia Euclidiana”.

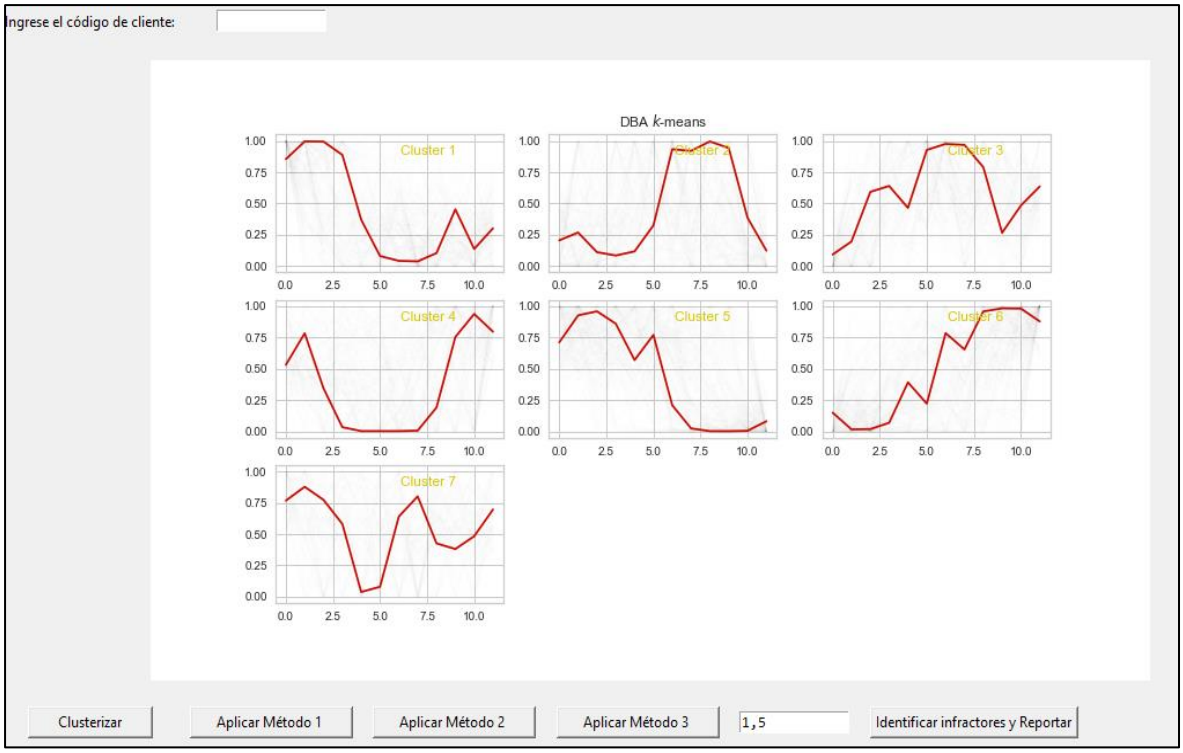


Fuente: (Python).
Elaborado por: Macao R. – Pujota E. (2022).

4.1.5.2. Resultado obtenido al aplicar el método 2 “Barycenter Averaging (DBA)” en el algoritmo de K-Means de muestra.

Se procede a ejecutar el segundo método con una muestra de 1000 usuarios, se escoge el clúster # 5, los usuarios de este grupo tienen un consumo promedio aproximado de 80 kWh, en la gráfica se observa que el consumo en los primeros tres meses es elevado para luego decaer y volver a subir nuevamente, al sexto mes el consumo decae de una forma abrupta y se mantiene en un bajo consumo constante, al final del último mes la curva tiende a subir.

Figura 4. 10. Resultado obtenido al aplicar el método 2 “Barycenter Averaging (DBA)”.

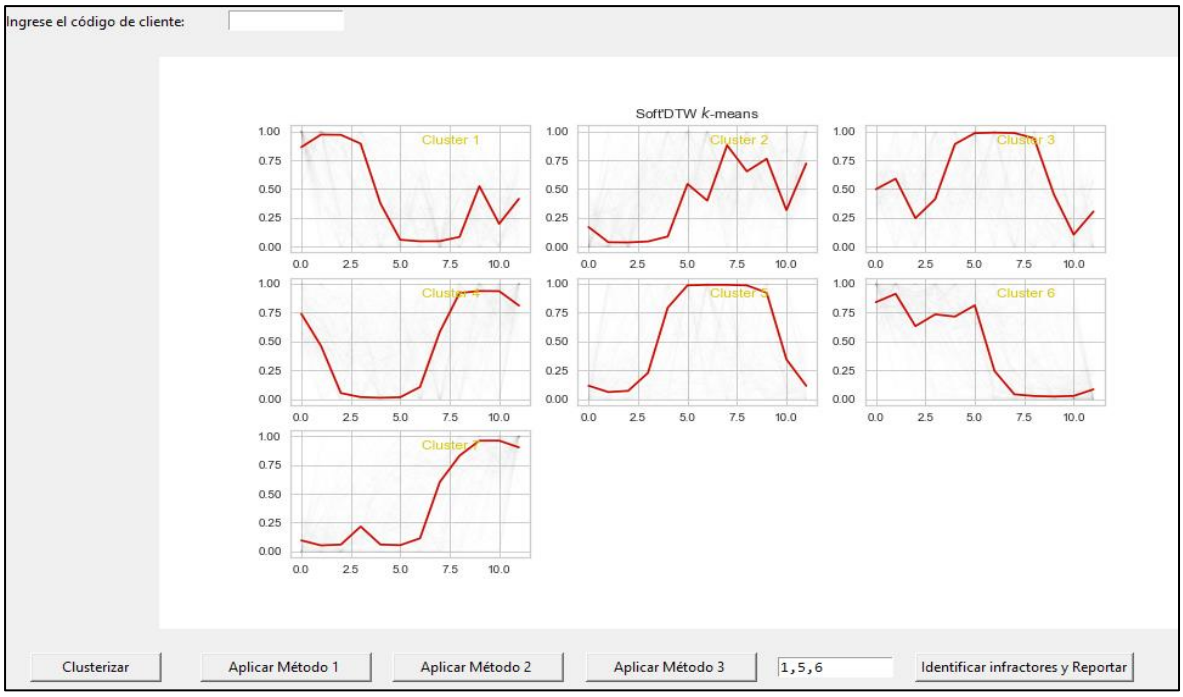


Fuente: (Python).
Elaborado por: Macao R. – Pujota E. (2022).

4.1.5.3. Resultado obtenido al aplicar el método 3 “Soft-DTW” en el algoritmo de K-Means de muestra.

Al ejecutar el tercer método con una muestra de 1000 usuarios, se escoger el clúster # 6, los usuarios de este grupo tienen un consumo promedio aproximado de 80 KWh en la gráfica se observa que el consumo en el primer mes es elevado para luego decaer y volver a subir levemente, al sexto mes el consumo comienza a decaer paulatinamente hasta llegar a un bajo consumo constante, para al final del últimos la curva tiende a subir de una forma muy leve.

Figura 4. 11. Resultado obtenido al aplicar el método 3 “Soft-DTW”.



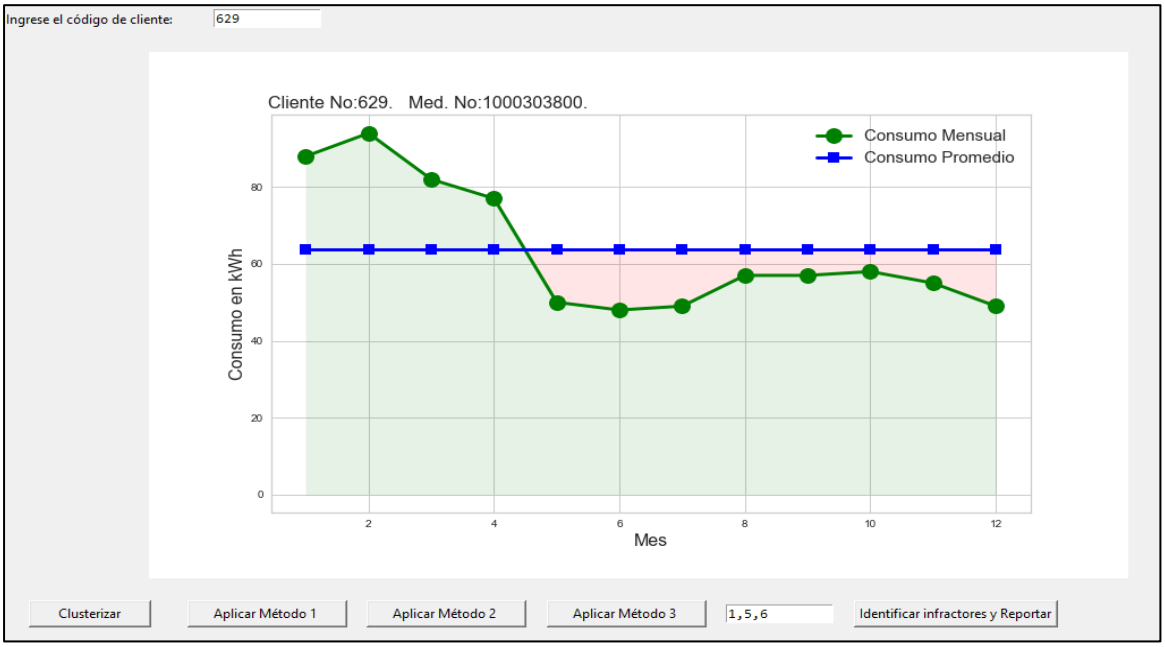
Fuente: (Python).
Elaborado por: Macao R. – Pujota E. (2022).

4.1.5.4. Resultados de los tres métodos de clusterización.

El algoritmo toma en cuenta a los usuarios que tienen la misma similitud en los tres métodos de clusterización, si un usuario tiene similitud en uno o dos métodos el algoritmo lo descarta. El resultado del análisis de los 1000 usuarios nos da por subestaciones en porcentaje en un archivo de Excel.

El algoritmo da como resultado tres subestaciones que posiblemente tengan perdidas comerciales. La subestación el Centenario tiene un 0,025% de perdidas no técnicas, y existen 6 usuarios que quizás estén hurtando energía eléctrica, se analiza por código de cliente para poder visualizar y analizar durante los 12 meses de consumo, se procede a escoger el código 629. el color azul corresponde al consumo promedio y el color verde al consumo mensual.

Figura 4. 12. Representación gráfica al ingresar el código de un cliente para analizar su consumo de 12 meses.



Fuente: (Python).
Elaborado por: Macao R. – Pujota E. (2022).

Como se observa en la gráfica 4.15. El cliente tiene un consumo promedio 65 KWh, del primero al segundo mes existe un consumo que tiende a subir, del segundo al tercer mes su consumo comienza a decaer hasta llegar al cuarto mes, del cuarto al quinto mes su consumo baja más de su consumo promedio manteniéndose así hasta el mes 12, esto no quiere decir que el cliente este hurtando energía, existen muchos factores para que el consumo de este haya descendido, entre ellos esta que era una vivienda que se utilizaban varias familias, que había un negocio “tienda, bazar, taller”.

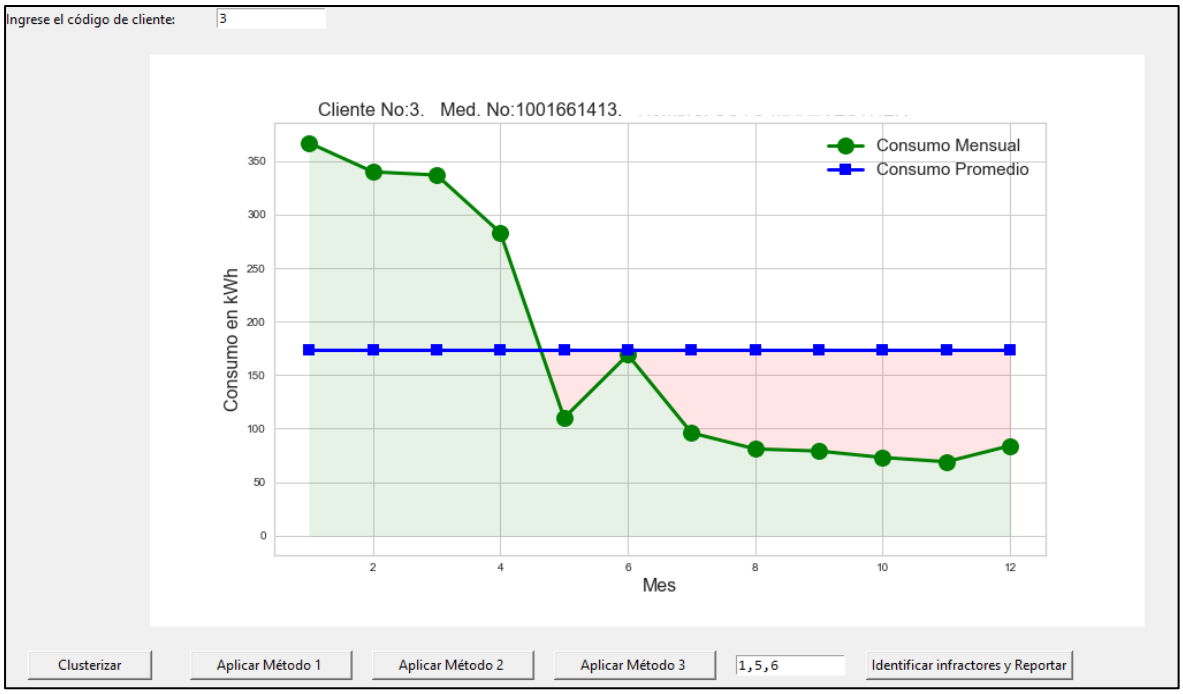
La subestación los Pambiles tiene un porcentaje de 0,079% de pérdidas no técnicas y 17 clientes que posiblemente estén hurtando energía.

Tabla 4. 4. Datos de posibles usuarios que presenten pérdidas comerciales a la subestación los Pambiles.

	CLICOD	MDENUMFAB	SGCOR X	SGCOR Y	USOCOD	CLIPRVCODP
0	3	1001661413	7034510361	99719648454	CO	23
20	36	4262982	7034539404	99720348668	CO	23
94	141	1001199966	7034356508	99722020997	CO	23
113	171	1001171481	7033855692	99723150518	RD	23
123	193	1001197418	7033961605	99722247322	RD	23
168	286	1001171680	7032601708	99721553836	RD	23
220	372	1001593794	7033217666	99724341166	RD	23
253	433	13085	7032188726	99721840148	CO	23
261	444	1502728158	7032313119	99722534946	CO	23
285	486	1001624907	7032263969	99724684865	RD	23
368	629	1000303800	7033655075	99725671431	RD	23
446	777	1001182434	70314256	99723492201	RD	23
448	779	1001182426	7031413859	99723430699	RD	23
451	785	1001200577	7031000634	99723153706	RD	23
SUBESTACIÓN PAMBILES 0.079%						

Fuente: (Python).
Elaborado por: Macao R. – Pujota E. (2022).

Figura 4. 13. Cliente código 3 de la subestación los Pambiles posibles pérdidas comerciales.



Fuente: (Python).
Elaborado por: Macao R. – Pujota E. (2022).

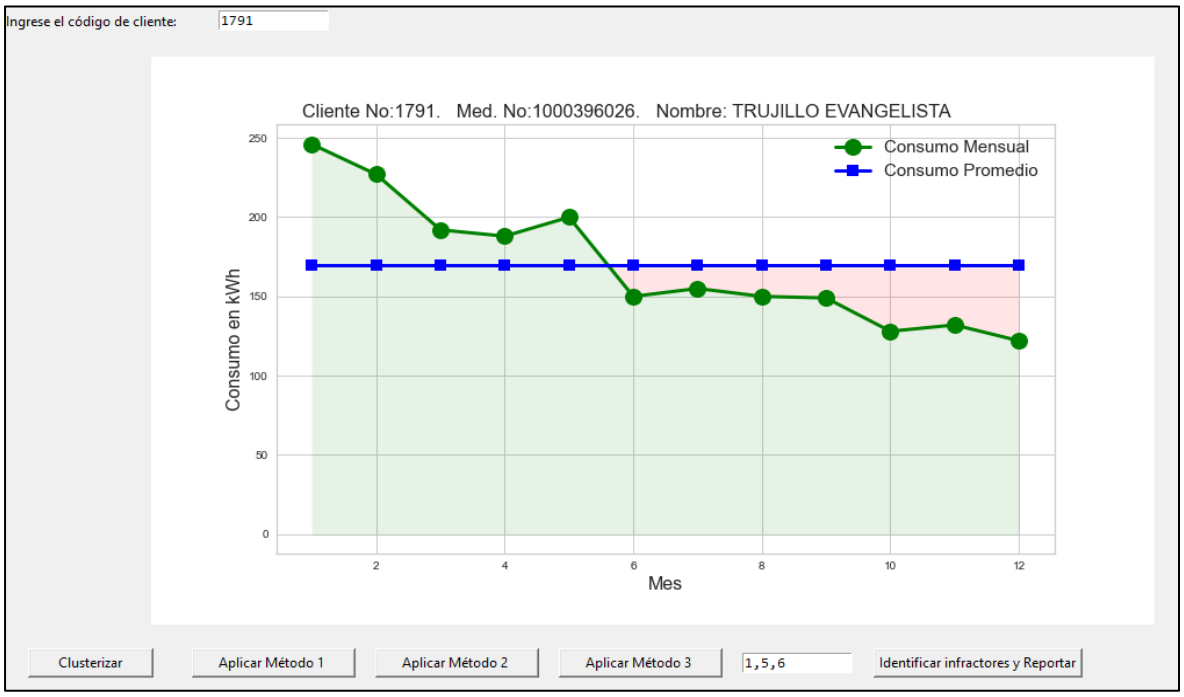
La subestación de Quevedo tiene un porcentaje de 0,016% de pérdidas no técnicas y 6 clientes que posiblemente estén hurtando energía.

Tabla 4. 5. Datos de usuarios que pueden ser infractores de la subestación Quevedo.

CLICOD	MDENUMFAB	SGCORX	SGCORY	USOCOD	CLIPRVCODP	CLIRLSCOD	CLISECINM
1413	1E+09	701983	9971122	RD	23	143	18107
1566	1E+09	701212	9969732	RD	23	165	12200
1676	1E+09	701810	9971098	RD	23	134	11700
1791	1E+09	701901	9971292	RD	23	135	14300
2063	1,5E+09	702097	9971388	RD	23	133	15500
2133	1,00E+09	702233	9971537	RD	23	132	14002
SUBESTACIÓN QUEVEDO	**0,016%**						

Fuente: (Python).
Elaborado por: Macao R. – Pujota E. (2022).

Figura 4. 14. Cliente código 1791 de la subestación de Quevedo.



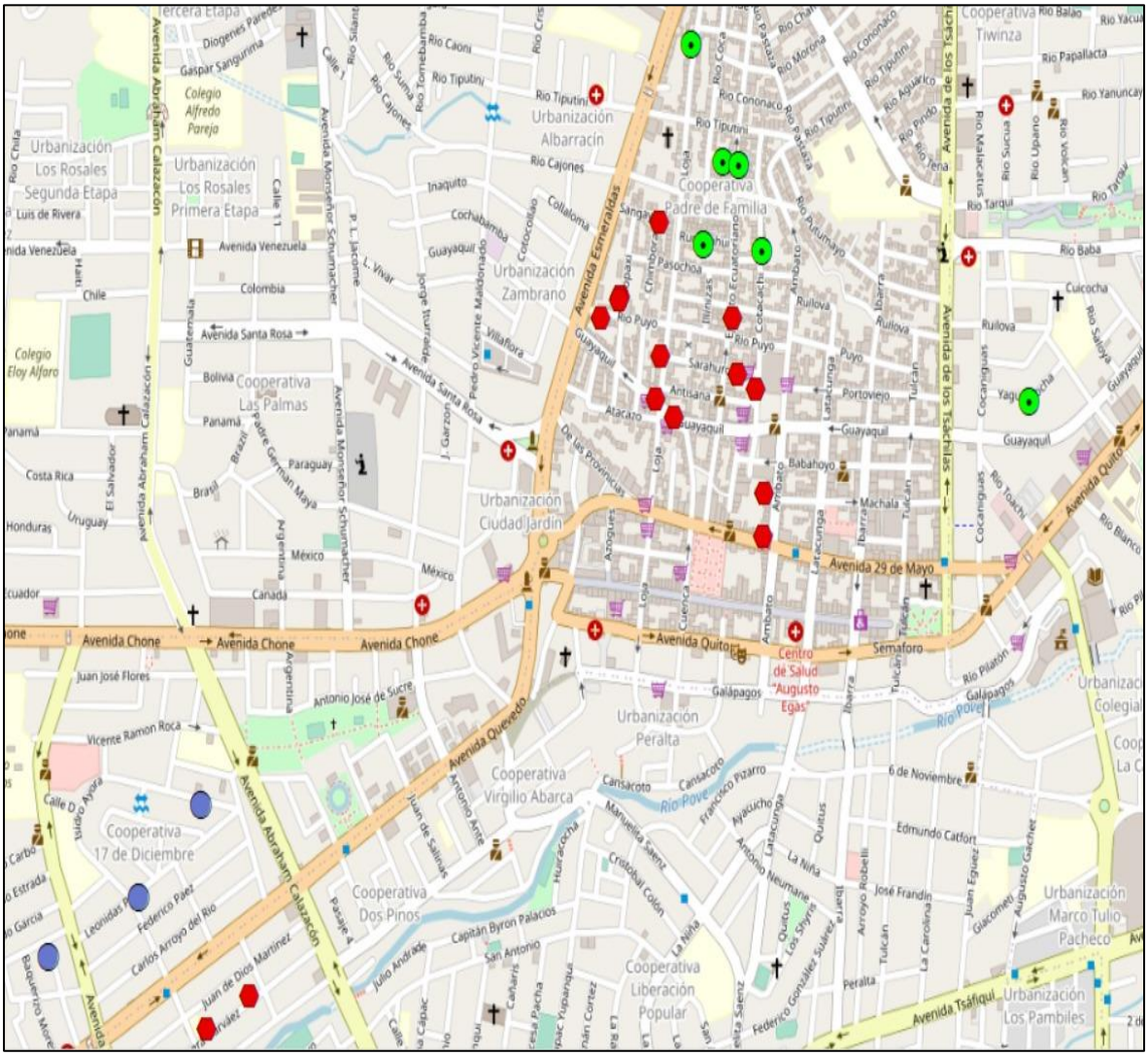
Fuente: (Python).
Elaborado por: Macao R. – Pujota E. (2022).

4.1.5.5 Representación geográfica de los posibles infractores como base de muestra previo a la ejecución total de usuarios utilizando el software ArcGIS.

- El color amarillo corresponde a la subestación el Centenario.
- El color rojo corresponde a la subestación Quevedo.
- El color verde corresponde a la subestación Pambiles.

Se puede ubicar a los usuarios de acuerdo a la subestación y su ubicación geográfica.

Figura 4. 15. Representación geográfica de los posibles infractores como base de muestra previo a la ejecución total de usuarios utilizando el software ArcGIS.

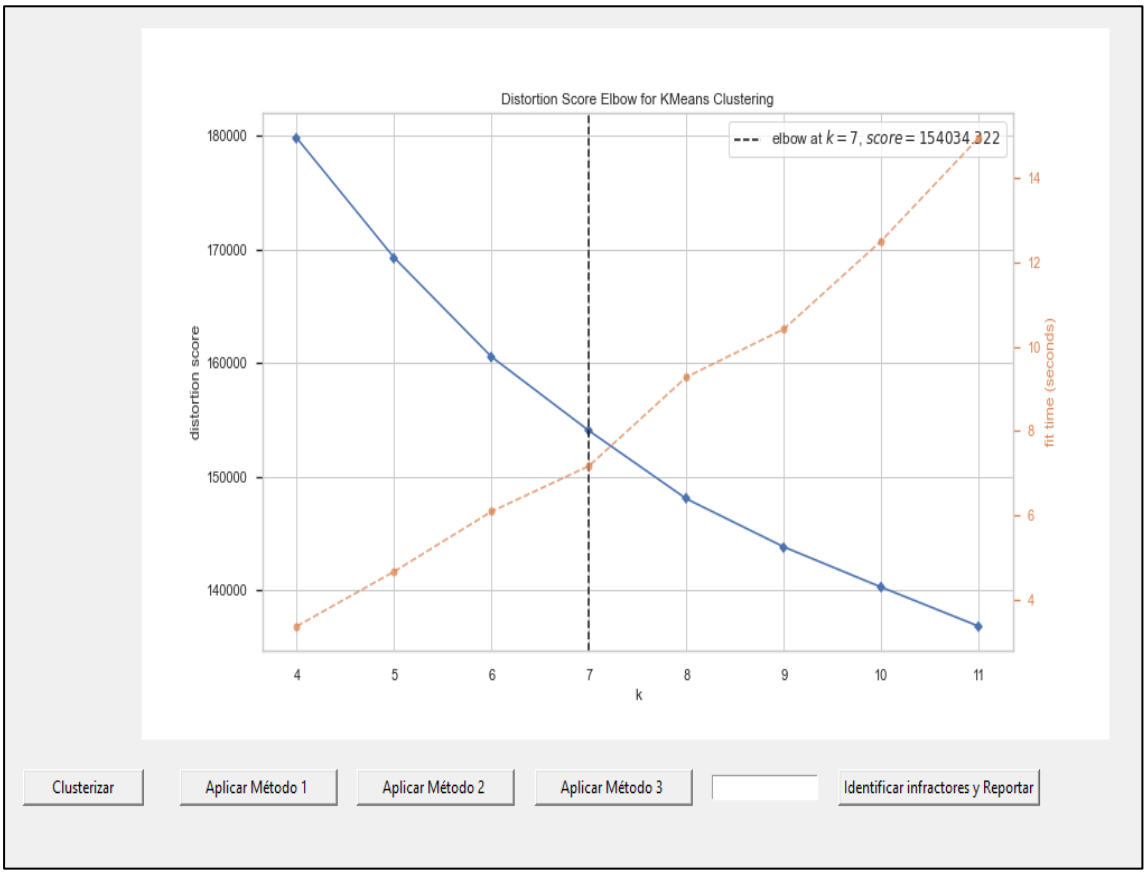


Fuente: (Python).
Elaborado por: Macao R. – Pujota E. (2022).

4.1.6. Detección de posibles usuarios infractores que pertenecen a la unidad de negocios CNEL EP Santo Domingo.

Se ejecuta el algoritmo con los 202717 usuarios, mediante el botón "Clusterizar", el cual nos indica que la Data set será distribuida en 7 clústeres, para posterior aplicar 3 métodos **"**Distancia Euclidiana, Barycenter Averaging (DBA), Soft-DTW**"** que se ha utilizado para la medición entre las series temporales en base a su similitud y de esta manera obtener un reporte de los posibles infractores en pérdidas comerciales para la unidad de negocios CNEL EP Santo Domingo.

Figura 4. 16. Representación gráfica de los números de clúster para aplicar los métodos de Machine Learning.



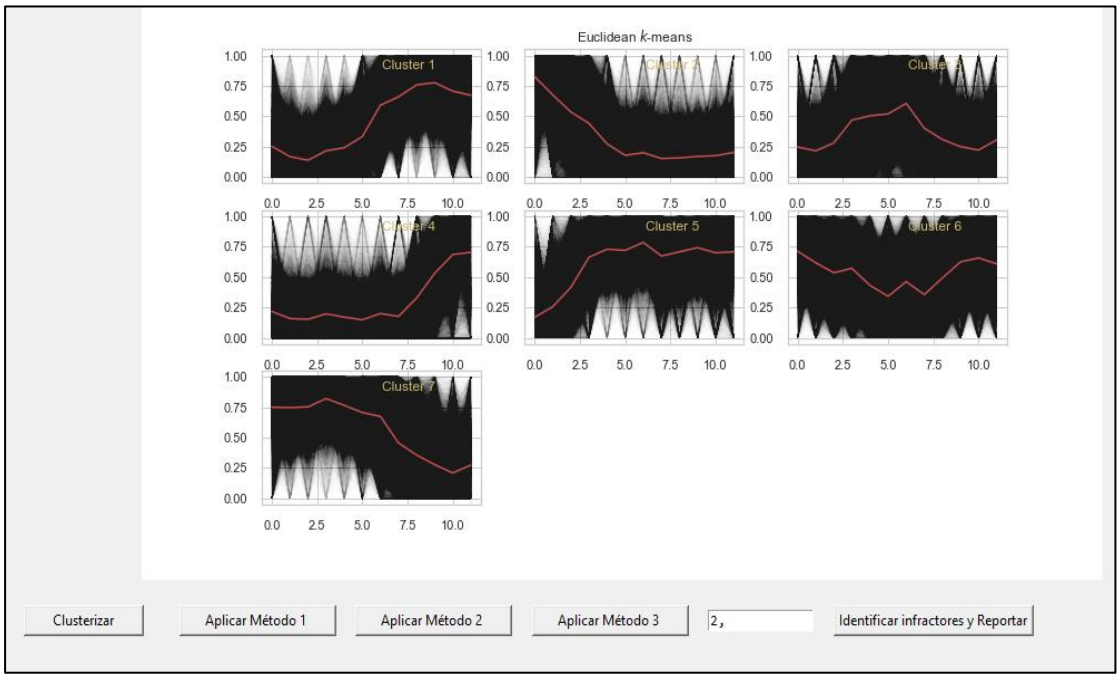
Fuente: (Python).
Elaborado por: Macao R. – Pujota E. (2022).

4.1.6.1. Resultado obtenido al aplicar el método 1 “Distancia Euclidiana” en el algoritmo de K-Means.

Se ejecuta el primer método cuya función es encontrar la distancia más corta entre dos puntos, para el análisis se escoge la segunda gráfica, como se puede observar los usuarios de este grupo, la curva comienzan descender y se mantienen por debajo del consumo promedio hasta que la curva llega a su fin.

En la figura 4.22. Se observa 7 clústeres, los cuales indica el consumo de los usuarios, los cuales están distribuidos acorde a su similitud en cada grupo, eligiendo el clúster 2 debido a que el consumo desciende y se mantiene por debajo del consumo promedio.

Figura 4. 17. Representación gráfica al aplicar el método 1 “Distancia Euclidiana” en el algoritmo de K-Means.

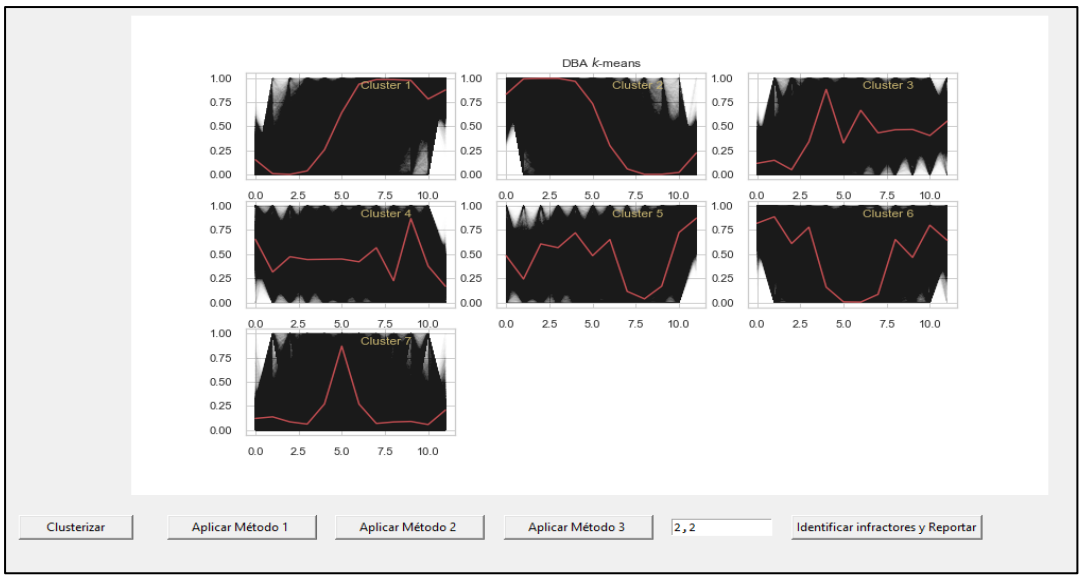


Fuente: (Python).
Elaborado por: Macao R. – Pujota E. (2022).

4.1.6.2. Resultado obtenido al aplicar el método 2 “Barycenter Averaging (DBA)” en el algoritmo de K-Means.

El segundo método el cual ayuda a refinar interactivamente una secuencia promedio inicial, como se puede observar las 7 gráficas, la segunda de estas marcas la diferencia con relación a las otras, la cual indica el consumo de este grupo desciende muy por debajo de su consumo promedio y se mantiene así casi hasta el final donde comienza a ascender lentamente.

Figura 4. 18. Representación gráfica al aplicar el método 2 “Barycenter Averaging (DBA)” en el algoritmo de K-Means.

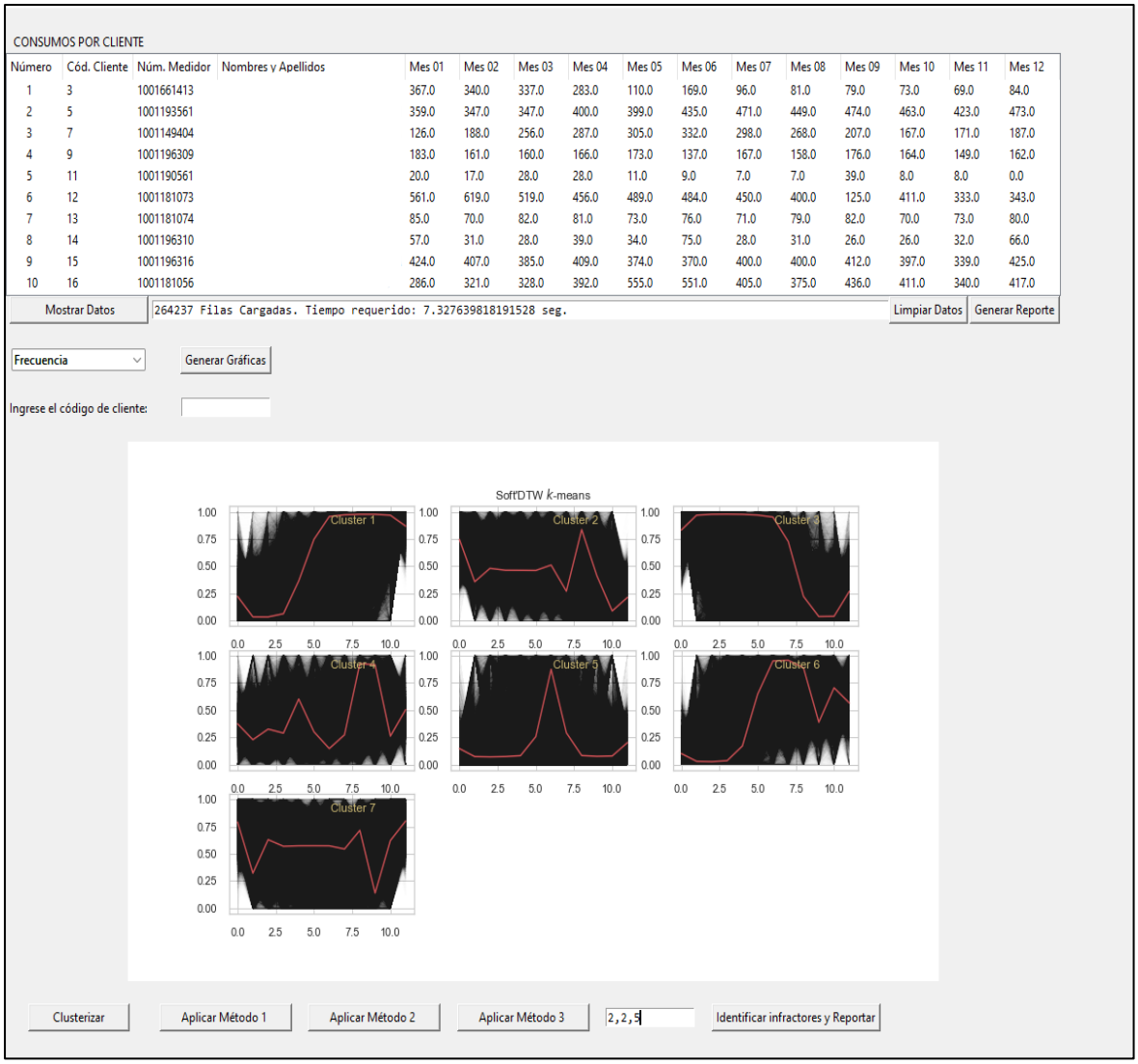


Fuente: (Python).
Elaborado por: Macao R. – Pujota E. (2022).

4.1.6.3. Resultado obtenido al aplicar el método 3 “Soft-DTW” en el algoritmo de K-Means.

El tercer método del algoritmo es un método agrupamiento tomando como métrica de similitud, en este método se escoge la tercera gráfica 4.24. Puesto que es muy divergente con relación a las otras gráficas los usuarios de este grupo tienen un consumo normal, pero a la mitad de la curva decrece de una forma muy abrupta casi hasta llegar a cero y al final de ella comienza a ascender nuevamente.

Figura 4. 19. Representación gráfica al aplicar el método 3 “Soft-DTW” en el algoritmo de K-Means.



Fuente: (Python).
Elaborado por: Macao R. – Pujota E. (2022).

4.1.6.4. Resultados de los tres métodos de clusterización de posibles usuarios infractores que pertenecen a la unidad de negocios CNEL EP Santo Domingo.

Luego de aplicar los tres métodos de clusterización, el algoritmo genera un archivo en Excel donde da como resultado que 18400 usuarios posiblemente estarían causando pérdidas comerciales a la CNEL EP Santo Domingo, en el mismo archivo de Excel se indican las pérdidas por subestaciones en porcentaje.

Tabla 4. 6. Resultados de los tres métodos de clusterización de posibles usuarios infractores.

CLICOD	MDENUMFAB	USOCOD	12	11	10	9	8	7	6	5	4	3	2	1	SUBESTACIÓN
19233	1710100852	RD	0	0	0	0	0	0	0	0	0	0	48	39	ALLURIQUIN
19296	1810380291	RD	86	77	81	67	65	51	49	51	51	117	635	761	ALLURIQUIN
19341	1001214209	RD	204	197	183	194	177	214	214	181	174	234	241	246	ALLURIQUIN
19353	1506718593	RD	151	184	101	107	145	204	168	137	179	235	276	282	ALLURIQUIN
19369	1506718595	RD	151	139	156	152	162	159	143	141	123	139	155	225	ALLURIQUIN
19406	1506718866	RD	197	240	234	195	186	166	166	168	184	215	310	270	ALLURIQUIN
19412	1710163024	TE	139	127	148	134	139	142	148	191	214	210	232	201	ALLURIQUIN
19422	1011000574	RD	77	71	73	74	92	82	73	73	75	69	92	109	ALLURIQUIN
19441	1710091521	RD	25	0	0	0	2	0	0	68	81	136	305	282	ALLURIQUIN
19480	50375240	RD	35	46	33	36	37	30	37	36	50	46	60	42	ALLURIQUIN
19489	1001217871	BP	17	23	23	19	16	30	0	0	0	0	0	90	ALLURIQUIN
19539	1001218092	RD	10	11	9	14	13	11	4	1	62	47	60	50	ALLURIQUIN
19579	1001197385	RD	122	134	128	110	118	135	136	131	157	147	215	190	ALLURIQUIN
19597	1001193386	RD	47	53	41	41	46	53	63	72	73	67	87	65	ALLURIQUIN
19674	1810245213	RD	94	113	104	91	92	93	92	87	111	114	140	120	ALLURIQUIN

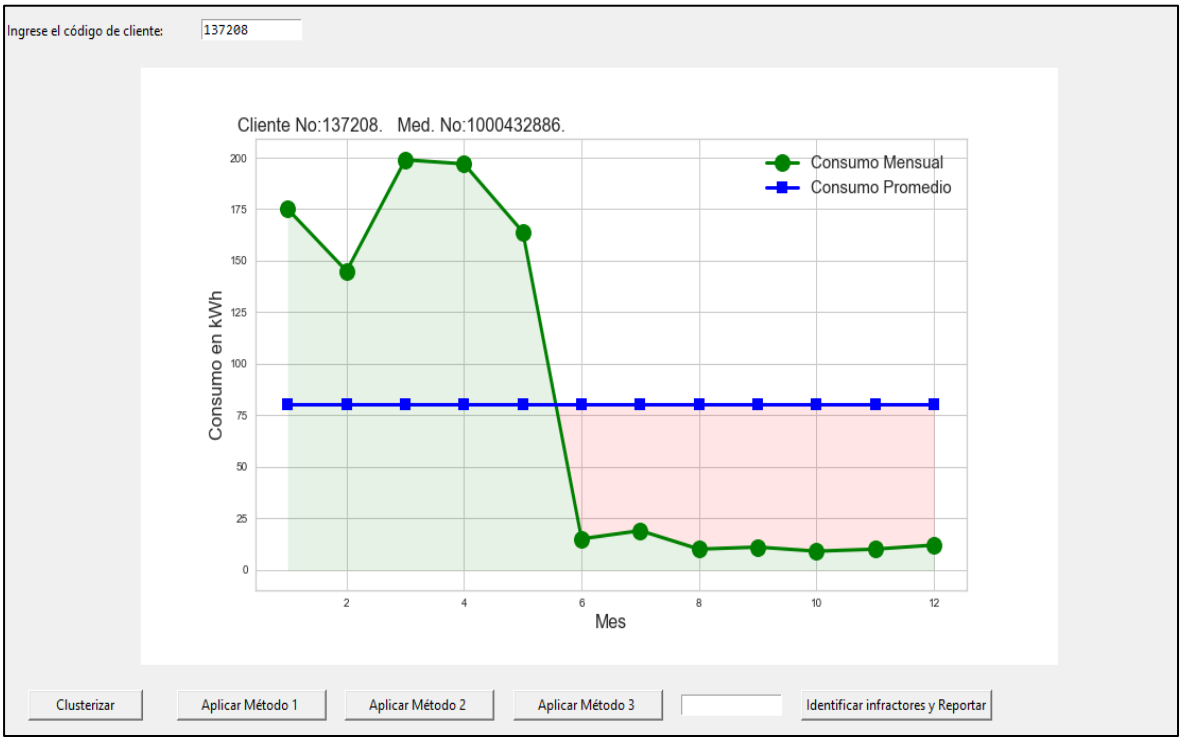
Fuente: (Excel).
Elaborado por: Macao R. – Pujota E. (2022).

Ejemplo de un posible cliente infractor detectado por el algoritmo con código 137208.

Esta gráfica indica el comportamiento del consumo durante 12 meses, es un usuario RP, residencial PEC "Programa energía cocción eficiente", perteneciente a la subestación Flavio Alfaro, como se puede observar sus 5 meses de consumo fueron normales, pero al mes 6 descendió drásticamente, hasta llegar a consumir energía eléctrica menor a su consumo promedio, manteniéndose así hasta el mes número 12.

Existen varios factores que indican el comportamiento de consumo energético de este usuario, puede ser que la vivienda estaba alquilada, dejo de utilizar la cocina de inducción o la vez era una vivienda vacacional, le corresponde a la CNEL EP verificar al usuario en mención para descartar manipulación en el medidor de energía eléctrica.

Figura 4. 20. Ejemplo de un posible cliente infractor detectado por el algoritmo.

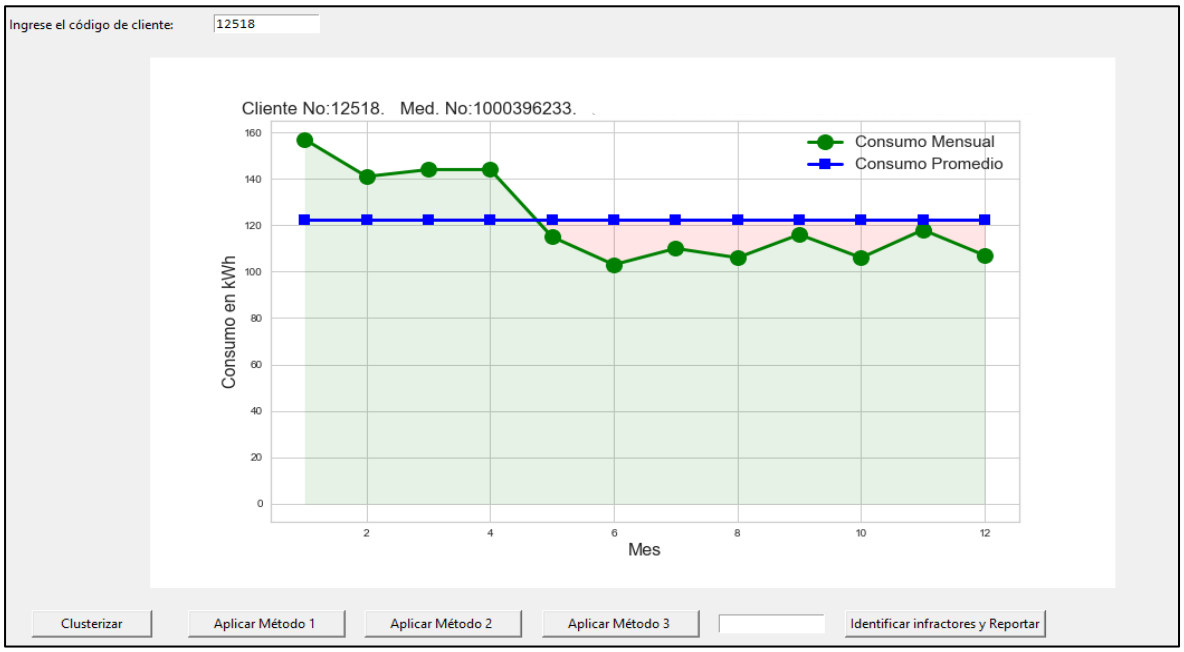


Fuente: (Excel).
Elaborado por: Macao R. – Pujota E. (2022).

Ejemplo de un posible cliente infractor detectado por el algoritmo con código 12518.

Usuario residencial "RD" perteneciente a la subestación Concordia los primeros cuatro meses mantiene un consumo regular, al quinto mes su consumo desciende por debajo de su consumo promedio, este comportamiento lo mantiene hasta el mes doce. hay varios factores que indiquen el comportamiento de consumo energético de este usuario, puede ser que la vivienda estaba alquilada, era una vivienda vacacional, o talvez ahí existía una actividad comercial le corresponde a la CNEL EP verificar al usuario en mención para descartar cualquier irregularidad en su medidor de energía eléctrica.

Figura 4. 21. Ejemplo de un posible cliente infractor detectado por el algoritmo.



Fuente: (Excel).
Elaborado por: Macao R. – Pujota E. (2022).

4.1.7. Representación geográfica del reporte generado por el algoritmo de los posibles infractores en ArcGIS.

El análisis realizado en esta investigación, difiere con los porcentajes de pérdidas de la proporcionados por la CNEL EP Santo Domingo, en la presente investigación se analizó a los clientes comerciales y residenciales, además se analizó solo pérdidas no técnicas de los usuarios antes mencionados. en los datos que fueron proporcionados se encuentran pérdidas técnicas y no técnicas, pérdidas por alimentadores. Los porcentajes que se encuentran detallados por subestaciones en la tabla 4.2. Se puede visualizar que la subestación Alluriquín ‘color verde claro’ tiene un 8.102% de pérdidas, presentando 517 usuarios que posiblemente estén ocasionando pérdidas comerciales.

La subestación Flavio Alfaro ‘color rojo’ es la que ocupa el segundo lugar en pérdidas no técnicas para la CNEL EP, debido que tiene una población muy baja suscrita a la CNEL EP. y representa un 7,98% un total de 487 usuarios que posiblemente estén causando pérdidas comerciales a la unidad de negocios.

La Subestación Pedernales ‘color melón’ tiene el menor número de pérdidas con un porcentaje de 5,343% de pérdidas presentando 615 usuarios que sus medidores tendrán que ser revisados por personal de la CNEL EP, para descartar posibles manipulaciones en sus medidores.

Todas las subestaciones presentan pérdidas no técnicas como se puede visualizar en la tabla 4.2. CNEL EP Santo Domingo, se tendrá que analizar y corroborar de manera práctica a cada uno de los usuarios que el algoritmo reporto como posibles infractores.

Tabla 4. 7. Pérdidas por subestaciones CNEL EP Santo Domingo.

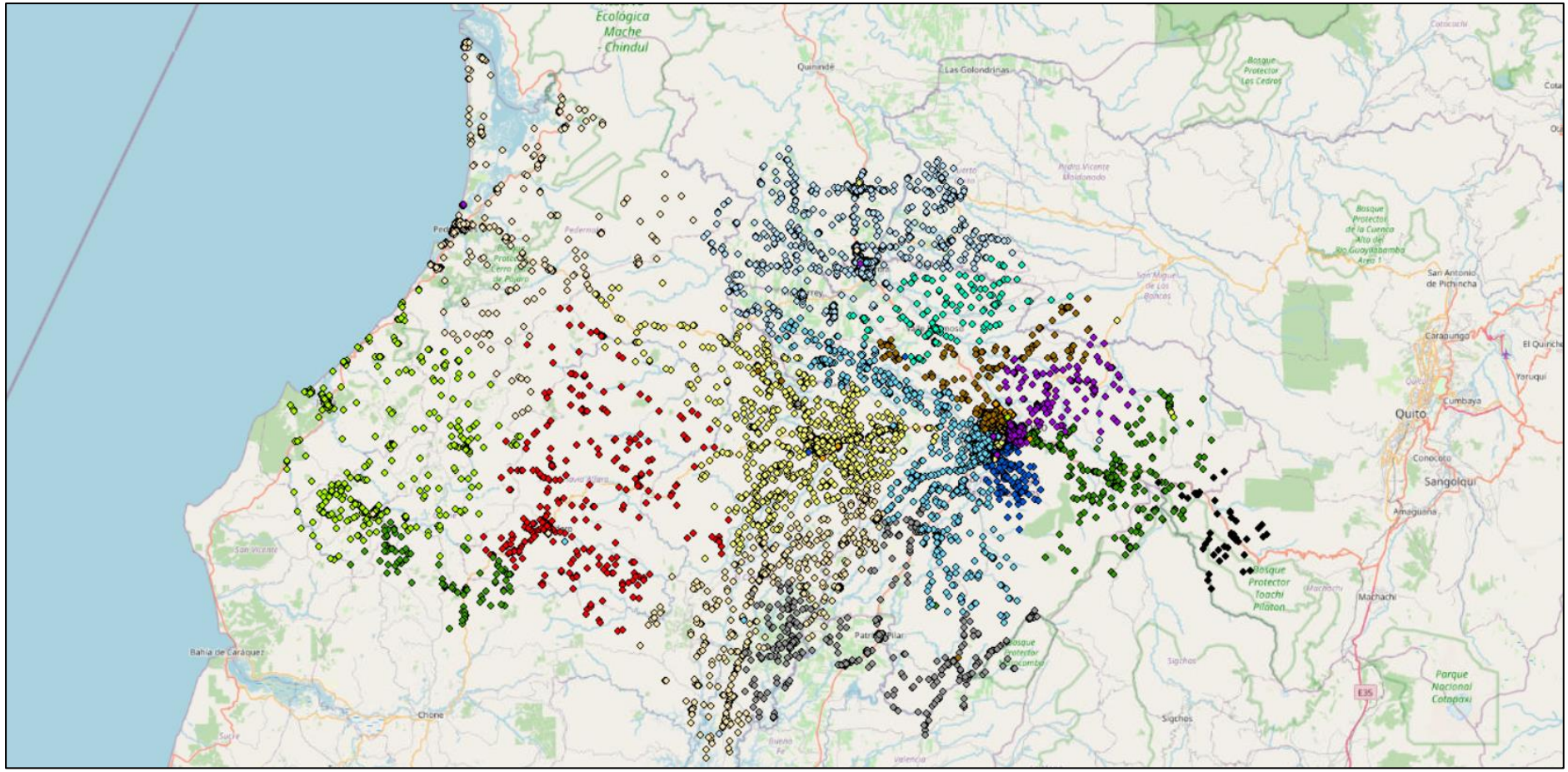
SUBESTACIONES	PÉRDIDAS EN %	COLOR	ARGIS
ALLURIQUÍN	8,102	VERDE CLARO	
EL CARMEN	7,196	AMARILLO	
EL CENTENARIO	7,702	NARANJA	
EL ROCÍO	7,585	ORO	
FLAVIO ALFARO	7,98	ROJO	
JAMA	6,838	VERDE E1	
LA CADENA	7,479	AZUL	
LA CONCORDIA	7,391	CELESTE E1	
LA PALMA (P)	7,763	NEGRO E2	
NO ASIGNADOS	No Asignados	AZUL OSCURO	
PAMBILES	7,418	ORO E4	
PATRICIA PILAR	6,785	GRIS E2	
PEDERNALES	5,343	MELON	
QUEVEDO	7,624	CELESTE	
SESME	7,574	VERDE OSCURO	
VALLE HERMOSO	6,25	TURQUESA	
VÍA COLORADOS DEL BÚA	7,888	CAFÉ	
VÍA QUITO	6,945	MORADO	

Fuente: (Excel).
Elaborado por: Macao R. – Pujota E. (2022).

4.1.7.1. Representación geográfica de los posibles infractores en ArcGIS.

En la figura 4.28. Se visualizan los posibles usuarios infractores en base al reporte obtenido del algoritmo de inteligencia artificial los cuales son 184000 usuarios ubicados geográficamente e identificados por subestaciones y a la vez por colores, cada color representa una subestación como se muestra en la tabla 4.2.

Figura 4. 22. Ubicación geográfica de los posibles infractores distribuidos por subestaciones.



Fuente: (ArcGIS).
Elaborado por: Macao R. – Pujota E. (2022).

CAPÍTULO V

CONCLUSIONES Y RECOMENDACIONES

5.1. Conclusiones.

Python es un software versátil y libre que permite la creación de aplicaciones, siendo uno de los más utilizados en machine Learning, debido a que su lenguaje es sencillo y rápido de aprender, además permite realizar algoritmos en menos líneas de código, realizando la programación con o sin conectividad a internet, facilitando su interpretación por la interfaz que utiliza para diferentes tipos de diagramas.

Para desarrollar algoritmos de inteligencia artificial se tiene que tomar en cuenta la información que se dispone y que se quiere obtener a partir de dicha información, en base a los datos proporcionados se debe analizar el tipo de aprendizaje que se empleara para implementar un algoritmo, este puede ser aprendizaje supervisado o aprendizaje no supervisado.

Para obtener datos de salida se debe utilizar el aprendizaje no supervisado, ya que solo se tiene datos de entrada que son los 12 meses de consumo individual de cada usuario, para luego poder aplicar los diferentes métodos de clusterización el cual permite agrupar a los datos en grupos de acuerdo a su similitud.

Se debe limpiar los datos nulos, datos duplicados y datos en cero, para evitar errores al momento de ejecutar el algoritmo, además los clientes que tienen un consumo mayor a 1000 kWh deben ser excluidos, estos corresponden a grandes clientes. Se crea una restricción a los usuarios que poseen un consumo de 5 kWh hasta 260 kWh, los cuales son tomados en cuenta, debido a que si se considera a todos los usuarios de la Data set se crearían más grupos de clusterización que al momento de la ejecución del algoritmo los resultados se distorsionarían y en base a ese problema, es que se crea una restricción para que el análisis en base al algoritmo sea más eficiente.

Los beneficios de utilizar las técnicas de Machine Learning es que permite analizar una gran cantidad de información en un menor tiempo posible. El algoritmo de Clustering K-Means es el más utilizado para poder encontrar etiquetas en un gran conjunto cuando no se tiene datos de salida, facilita al ser humano la agrupación de los grupos de acuerdo a su similitud ya que de manera manual no lo podría distinguir. de esta manera se evita el error humano y se puede encontrar y diferenciar las características únicas de los K grupos de clusterización.

Los resultados obtenidos al aplicar el algoritmo para detectar a los posibles infractores de la unidad de negocios CNEL EP Santo Domingo, determina que la subestación Alluriquín presenta el mayor porcentaje de pérdidas de energía, no obstante la subestación Flavio Alfaro es la segunda en presentar un alto índice de pérdidas no técnicas a pesar que es quien representa menor número de usuarios conectados a la empresa distribuidora, se debe a la idiosincrasia de las personas de obtener posiblemente un beneficio propio en base a cometer un acto ilícito.

Para evidenciar la ubicación geográfica de los posibles infractores se utiliza el software ArcGIS, el cual permite detectar de manera georreferenciada en que sectores se encuentra la mayor parte de pérdidas no técnicas. de esta manera la empresa distribuida podrá realizar la respectiva revisión a los clientes para descartar o sancionar a los usuarios, siendo así una manera más eficiente de reducir las posibles pérdidas comerciales.

Para realizar una comparación entre las pérdidas no técnicas de la CNEL EP con el resultado obtenido en base a algoritmos de inteligencia artificial, se debe contar con un historial de pérdidas basados en un algoritmo, no obstante debido a que es la primera vez que se realiza este tipo de estudio no podemos obtener un porcentaje para comparar dichas pérdidas de energía eléctrica, por lo tanto los resultados del algoritmo deben ser comprobados realizando un trabajo de campo para la revisión de cada uno de los posibles infractores que el algoritmo detecto.

5.2. Recomendaciones.

El reporte generado de los clientes que tienen consumos en cero se genera en un archivo Csv, al momento de importarlo a Excel las coordenadas geográficas de los clientes tanto en X como en Y, se pierde la coma y este genera un problema al importar las coordenadas al software ArcGIS, para evitar este inconveniente se recomienda convertir el archivo en un formato bloc de notas, este formato si es compatible con ArcGIS.

Para analizar algoritmos de Inteligencia Artificial es necesario tener un equipo de alta gama, con esto se reduciría el tiempo de la ejecución del algoritmo, este proyecto fue ejecutado en una laptop MSI GF63 8RD i7+8th Gen, el tiempo de ejecución en su totalidad fue de 23 horas, la cantidad total de datos analizados por la maquina fue 3’977.963, si al momento de la ejecución del algoritmo se disponía de otro equipo con características superiores, el tiempo empleado por el computador para terminar el análisis se hubiese reducido considerablemente.

Para crear algoritmos de inteligencia artificial basados en series temporales es necesario tener la mayor cantidad de información de lo que se va analizar en función del tiempo, para poder identificar patrones de acuerdo a su similitud por medio de la clusterización y así los resultados en cuanto a la predicción o pronóstico serán más exactos.

El presente trabajo investigativo queda a disposición para futuras investigaciones, una de ellas es la implementación del algoritmo para un análisis de los grandes clientes debido a que su consumo es muy alto y agruparlo con valores de consumo inferiores va a distorsionar las curvas para el análisis de posibles infractores y usuarios prepagos que tienen su consumo en cero, por lo tanto en este estudio se analizaron clientes comerciales y residenciales, para la detección de posibles pérdidas comerciales en la empresa distribuidora.

Con una base de datos que contenga varios parámetros se puede lograr mediante los algoritmos inteligentes, mejorar la eficiencia, optimizar costos, generar nuevas fuentes de ingresos y administrar el riesgo y el fraude. Además de que se lo puede implementar como aplicación para teléfonos celulares y que sea más accesible para que el operador analice el comportamiento de consumo de energía eléctrica a cualquier tipo de cliente.

CAPÍTULO VI

BIBLIOGRAFÍA

[1] E. L. G. Wiechers, «Definición de algoritmo,» de Análisis Diseño e Implantación de Algoritmos Apunte electrónico , Ciudad de México, UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO, 2017, pp. 11-12.

[2] J. C. L. GARCÍA, «Algoritmo Inteligente en Inteligencia Artificial,» de ALGORITMOS Y PROGRAMACIÓN , Colombia, Eduteka , 2018, pp. 22-29.

[3] T. A. Williams, «Britannica,» Britannica, vol. l, nº 10, pp. 10-12, 2018.

[4] Y.-Y. Y. H.-W. Z. &. Y.-M. C. Wei-Mao Qian, «Optimal two-parameter geometric and arithmetic mean bounds for the Sándor–Yang mean,» Journal of Inequalities and Applications, nº 287, p. 25, 2019.

[5] Microsoft, «support.microsoft,» 8 Abril 2020. [En línea]. Available: https://support.microsoft.com/es-es/office/desvesta-funci%C3%B3n-desvesta-5ff38888-7ea5-48de-9a6d-11ed73b29e9d. [Último acceso: 11 Marzo 2022].

[6] J. Smith, «Medidas de tendencia central,» 02 Agosto 2018. [En línea]. Available: http://www.cca.org.mx/cca/cursos/estadistica/html/m9/promedio_ponderado.htm. [Último acceso: 12 Marzo 2022].

[7] BYJU'S , «Datasets,» BYJU'S , nº 202, p. 18, 2020.

[8] Na8, «Detección de outliers en Python,» Aprende Machine Learning, nº 69, p. 52, 2020.

[9] Home, «Data Carpentry.org,» Home, 2018. [En línea]. Available: https://datacarpentry.org/python-ecology-lesson-es/04-data-types-and-format/. [Último acceso: 12 Marzo 2022].

[10] P. Clu, «Microsoft Build,» Microsoft, 26 Abril 2021. [En línea]. Available: https://docs.microsoft.com/es-es/azure/machine-learning/component-reference/normalize-data. [Último acceso: 12 Abril 2022].

[11] M. Raymer, W. Punch, E. Goodman, L. Kuhn y A. Jain, «Journals & Magazines,» 18 Noviembre 2019. [En línea]. Available: https://ieeexplore.ieee.org/abstract/document/850656. [Último acceso: 12 Marzo 2022].

[12] A. A. Fierro, «SEDICI,» 3 Mazo 2021. [En línea]. Available: http://sedici.unlp.edu.ar/handle/10915/114857. [Último acceso: 12 Marzo 2022].

[13] Elizabeth-H., J. García-García y C. López, «Universidad de los Andes,» 20 Octubre 2019. [En línea]. Available: http://funes.uniandes.edu.co/23493/. [Último acceso: 12 Marzo 2022].

[14] L. Jenkins, «Linuxteaching,» Linuxteaching, 26 Julio 2020. [En línea]. Available: https://www.linuxteaching.com/article/how_to_use_boxplot_in_python. [Último acceso: 12 Marzo 2022].

[15] L. Updated, «IBM,» 7 Diciembre 2021. [En línea]. Available: https://www.ibm.com/docs/es/spss-statistics/beta?topic=types-histogram-charts. [Último acceso: 15 Marzo 2022].

[16] Acervo Lima, «ACERVO LIMA,» 15 Mayo 2020. [En línea]. Available: https://es.acervolima.com/graficos-de-densidad-con-pandas-en-python/. [Último acceso: 10 Marzo 2022].

[17] D. Hemmendinger, «Lenguaje de programación,» de Programación, Mexico , ProgramWorld, 2019, pp. 25-29.

[18] C. R. Severance, Explorando la información con Python 3, EE.UU: Elliott Hauser, Sue Blumenberg, 2018.

[19] D. Jacob, «Opportunistic acceleration of array-centric Python,» University of Glasgow, vol. l, nº 126, p. 96, 2020.

[20] C. Manzino, «Programaci´on usando Python,» 16 Febrero 2018. [En línea]. Available: chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://www.fceia.unr.edu.ar/~alpi/laboratorio/python2.pdf. [Último acceso: 15 Marzo 2022].

[21] S. Gupta, «Springboard,» Springboard, 7 Marzo 2022. [En línea]. Available: https://www.springboard.com/blog/data-science/python-libraries-for-machine-learning/. [Último acceso: 18 Marzo 2022].

[22] C. Harder, ArcGIS Book, Houston: ESRI, Incorporated, 2017.

[23] D. University, «Duke University,» Duke, 29 Abril 2020. [En línea]. Available: https://guides.library.duke.edu/c.php?g=289313&p=1929408. [Último acceso: 25 Marzo 2022].

[24] S. Dick, «Artificial Intelligence,» HDSR, vol. l, nº 12, p. 23, 2019.

[25] IBM Cloud Education, «IBM,» IBM Cloud Education, 15 Julio 2020. [En línea]. Available: https://www.ibm.com/cloud/learn/machine-learning. [Último acceso: 24 Marzo 2022].

[26] Z. Shi, «Science Direct,» trademark of Elsevier B.V., 17 Agosto 2021. [En línea]. Available: https://www.sciencedirect.com/topics/computer-science/knowledge-engineering. [Último acceso: 23 Marzo 2022].

[27] J. E. v. E. &. H. H. Hoos, «A survey on semi-supervised learning,» SpringerLink, vol. l, nº 109, p. 373, 2019.

[28] Á. Gonzalo, «Regresión lineal en Python,» Machine Learning , España, 2018.

[29] S. Salayhin, Aprendizaje machine-learning, Mexico : riptutorial, 2020.

[30] S. Priy, «Clustering in Machine Learning,» GeeksforGeeks, nº 2, p. 3, 2021.

[31] M. Nandi, «Domino,» 2 Diciembre 2020. [En línea]. Available: https://blog.dominodatalab.com/topology-and-density-based-clustering. [Último acceso: 25 Marzo 2022].

[32] A. K. J. y. R. C. Dubes, Algorithms for Clustering Data, New Jersey : Barbara Marttine, 2018.

[33] H. JIMENEZ SALAZAR, «Selección de atributos mediante separación de centroides,» Universidad Autónoma Metropolitana, vol. l, nº 5, p. 10, 2019.

[34] Y. Z. a. G. Karypis, «Researchgate,» 26 Octubre 2018. [En línea]. Available: https://www.researchgate.net/publication/220451867_Hierarchical_Clustering_Algorithms_for_Document_Datasets. [Último acceso: 18 Marzo 2022].

[35] K. P. Sinaga y M.-S. Yang, «Unsupervised K-Means Clustering Algorithm,» IEEE, vol. Vlll, nº 1958, pp. 16-27, 2020.

[36] M. A. A. Carmona, «AlvarezCMA,» 14 Septiembre 2018. [En línea]. Available: chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://inaoe.repositorioinstitucional.mx/jspui/bitstream/1009/168/1/AlvarezCMA.pdf. [Último acceso: 18 Marzo 2022].

[37] L. Gonzalez, «aprendeIA,» 12 Mayo 2020. [En línea]. Available: https://aprendeia.com/seleccionar-el-numero-adecuado-de-clusteres/. [Último acceso: 22 Marzo 2022].

[38] J. Brownlee, «Machine Learning Mastery,» 25 Marzo 2020. [En línea]. Available: https://machinelearningmastery.com/distance-measures-for-machine-learning/. [Último acceso: 26 Marzo 2022].

[39] A. N. D. ECUADOR, «ariae,» 17 Enero 2015. [En línea]. Available: https://www.ariae.org/servicio-documental/ley-organica-del-servicio-publico-de-energia-electrica. [Último acceso: 25 Marzo 2022].

[40] A. d. R. y. C. d. Electricidad, «ARCONEL,» 17 Mayo 2017. [En línea]. Available: https://www.regulacionelectrica.gob.ec/wp-content/uploads/downloads/2017/05/2017_05_17_regulaci%C3%B3n_modelo_contrato_suministro_difusion_externa.pdf. [Último acceso: 25 Marzo 2022].

[41] E. D. D. L. A. D. R. Y. C. D. E. -. ARCONEL, «Resolución Nro. ARCONEL-043/18,» 22 Octubre 2018. [En línea]. Available: https://www.regulacionelectrica.gob.ec/wp-content/uploads/downloads/2019/03/043-18.pdf. [Último acceso: 25 Marzo 2022].

[42] T. S. J. M. Raúl Jiménez, «IDB Inter-American,» 26 Julio 2017. [En línea]. Available: https://publications.iadb.org/publications/spanish/document/Electricidad-perdida-Dimensionando-las-p%C3%A9rdidas-de-electricidad-en-los-sistemas-de-transmisi%C3%B3n-y-distribuci%C3%B3n-en-Am%C3%A9rica-Latina-y-el-Caribe.pdf. [Último acceso: 24 Marzo 2022].

[43] L. Arias, «Programa de rehabilitación de redes eléctricas,» 13 Septiembre 2017. [En línea]. Available: https://redeselectricasrd.cdeee.gob.do/que-y-como-son-las-perdidas-electricas-se-esta-haciendo-para-reducirlas/. [Último acceso: 25 Marzo 2022].

[44] P. T. P. L. E. E. D. DISTRIBUCIÓN, «SERVICIO PÚBLICO DE ENERGÍA ELÉCTRICA,» 30 Diciembre 2018. [En línea]. Available: https://www.regulacionelectrica.gob.ec/wp-content/uploads/downloads/2018/01/2018-01-11-Pliego-y-Cargos-Tarifarios-del-SPEE-20182.pdf. [Último acceso: 25 Marzo 2022].

CAPÍTULO VII

ANEXOS

Anexo 1.1. Programación del algoritmo de inteligencia artificial para la detección de posibles infractores suscritos a la CNEL EP Sto Dmgo

In [5]:

```
## Libraries
from tkinter import *
from PIL import ImageTk, Image
#import mysql.connector as mysql
#from mysql.connector import errorcode
import pandas as pd
import numpy as np
import tkinter as tk
from tkinter import ttk
import matplotlib.pyplot as plt
import statsmodels.api as sm
from datetime import date
import seaborn as sns
from matplotlib.backends.backend_tkagg import FigureCanvasTkAgg
import pylab
from sklearn.model_selection import train_test_split
import scipy.stats as stats
from sklearn.metrics import mean_squared_error,r2_score, mean_absolute_error
from sklearn.datasets import make_blobs
from sklearn.preprocessing import StandardScaler
#from mpl_toolkits.basemap import Basemap
from sklearn.ensemble import RandomForestRegressor
import time
import itertools
from tqdm import tqdm
from sklearn.preprocessing import MinMaxScaler
from sklearn.cluster import KMeans
import matplotlib.patches as mpatches

from yellowbrick.cluster import KElbowVisualizer

from tslearn.clustering import TimeSeriesKMeans
from tslearn.datasets import CachedDatasets
from tslearn.preprocessing import TimeSeriesScalerMeanVariance, \
    TimeSeriesResampler
```

In [2]:

```
# INSERTAR LA DIRECCION DEL ARCHIVO A ANALIZAR
df = pd.read_excel(r"C:\Users\richa\Desktop\TESIS FINAL UTEQ\DATOS USUARIOS CNEL- MEDIDORES
```

In [4]:

```
root = Tk()
root.title("Detección de infractores CNEL-EP Santo Domingo")
root.geometry("1124x980")

label01 = Label(root,text=" ")
label01.grid(row=0,column=0)

label02 = Label(root,text=" ")
label02.grid(row=0,column=1)

label03 = Label(root,text=" ")
label03.grid(row=0,column=2)

label04 = Label(root,text=" ")
label04.grid(row=0,column=3)

label05 = Label(root,text=" ")
label05.grid(row=0,column=4)

label06 = Label(root,text=" ")
label06.grid(row=0,column=5)

label07 = Label(root,text=" ")
label07.grid(row=0,column=6)

label08 = Label(root,text=" ")
label08.grid(row=0,column=7)


# creacion de la etiqueta sobre la tabla
label1 = Label(root,text="CONSUMOS POR CLIENTE")
label1.grid(row=1,column=0, columnspan=1)

# creacion de la tabla para la visualizacion
my_tree = ttk.Treeview(root)
my_tree["columns"]=('Column1','Column2','Column3','Column4','Column5','Column6','Column7','
my_tree.column('#0',width=60,minwidth=25)
my_tree.column('Column1',width=75,minwidth=25)
my_tree.column('Column2',width=90,minwidth=25)
my_tree.column('Column3',width=200,minwidth=25)
my_tree.column('Column4',width=58,minwidth=25)
my_tree.column('Column5',width=58,minwidth=25)
my_tree.column('Column6',width=58,minwidth=25)
my_tree.column('Column7',width=58,minwidth=25)
my_tree.column('Column8',width=58,minwidth=25)
my_tree.column('Column9',width=58,minwidth=25)
my_tree.column('Column10',width=58,minwidth=25)
my_tree.column('Column11',width=58,minwidth=25)
my_tree.column('Column12',width=58,minwidth=25)
my_tree.column('Column13',width=58,minwidth=25)
my_tree.column('Column14',width=58,minwidth=25)
my_tree.column('Column15',width=58,minwidth=25)
my_tree.heading('#0',text="Número",anchor=W)
my_tree.heading("Column1",text="Cód. Cliente",anchor=W)
my_tree.heading("Column2",text="Núm. Medidor",anchor=W)
my_tree.heading("Column3",text="Nombres y Apellidos",anchor=W)
my_tree.heading("Column4",text="Mes 01",anchor=W)
my_tree.heading("Column5",text="Mes 02",anchor=W)
my_tree.heading("Column6",text="Mes 03",anchor=W)
```

```
my_tree.heading("Column7",text="Mes 04",anchor=W)
my_tree.heading("Column8",text="Mes 05",anchor=W)
my_tree.heading("Column9",text="Mes 06",anchor=W)
my_tree.heading("Column10",text="Mes 07",anchor=W)
my_tree.heading("Column11",text="Mes 08",anchor=W)
my_tree.heading("Column12",text="Mes 09",anchor=W)
my_tree.heading("Column13",text="Mes 10",anchor=W)
my_tree.heading("Column14",text="Mes 11",anchor=W)
my_tree.heading("Column15",text="Mes 12",anchor=W)
my_tree.grid(row=2,column=0, columnspan=8)

def search():
    start = time.time()

    #seleccion de las columnas que se usaran para la aplicacion
    df1 = df.iloc[:,[0,1,2,14,15,16,17,18,19,20,21,22,23,24,25]]
    _cols = list(df1.columns)

    #Eliminamos la palabra mes de cada columna del archivo para tener solo valores numerico
    df1.columns = pd.Index(_cols[:3] + [int(c.replace("MES_","")) for c in _cols[3:]])
    del _cols
    cols = df1.columns.tolist()

    #Cambiamos el orden de los meses para que vayan desde mas antiguo a mas nuevo
    cols=['CLICOD', 'MDENUMFAB', 'IDCNOMIMP',1 ,2,3,4,5,6,7,8,9,10,11,12]
    df1 = df1[cols]
    df_col = df1.values.tolist()

    #Insertamos los valores en la tabla
    for i in range(len(df1)):
         my_tree.insert(parent='',index='end',text=str(i+1),values=(df_col[i]))
    end = time.time()

    #Calculamos el tiempo de carga de los datos en la tabla
    time_elapsed=(end - start)

    #Mostramos los datos que se han cargado, el tiempo que se ha demorado
    my_text1.config(state="normal")
    my_text1.delete(1.0, END)
    my_text1.insert(tk.INSERT,"{} Filas Cargadas. Tiempo requerido: {} seg.\n".format(len(d
    my_text1.config(state="disabled")


def limpiar():

    start = time.time()

    df1 = df.iloc[:,[0,1,2,14,15,16,17,18,19,20,21,22,23,24,25]]
    df2=df1.dropna() # Eliminar las filas que tienen valores nulos, o sin datos
    null_rows=len(df1)-len(df2) # Contar el numero de datos nulos

    df3=df2.loc[(df2.iloc[:, 3:15]!=0).any(1)] #Seleccionar las columnas de consumo y elima
    zero_rows=len(df2)-len(df3) # Contar las columnas que son ceros

    duplicated_rows=df3.duplicated().sum() # Eliminar y contar los clientes repetidos

    end = time.time()

    time_elapsed=(end - start)

    #Mostramos los datos que se han eliminado, el tiempo que se ha demorado
```

```
    my_text1.config(state="normal")
    my_text1.insert(tk.INSERT,"De {} Usuarios. Filas en cero: {}. Datos nulos {}. Datos dup
    my_text1.config(state="disabled")


def clean(df1):
    df2=df1.dropna()
    null_rows=len(df1)-len(df2)
    df3=df2.loc[(df2.iloc[:, 3:15]!=0).any(1)]
    zero_rows=len(df2)-len(df3)
    duplicated_rows=df3.duplicated().sum()
    return df3


def report():
    start = time.time()
    df1 = df.iloc[:,[0,1,2,14,15,16,17,18,19,20,21,22,23,24,25]]
    df4=pd.concat([df1,clean(df1)]).drop_duplicates(keep=False) # Tomar todos los valores n
    df4.to_csv(r"C:\Users\richa\Desktop\null_zero_data.csv") # Guardarlos en un archivo csv
    path = "C:\richa\Desktop\null_zero_data.xlsx" # Imprimir en donde se guarda xxxxxxxxxx
    end = time.time()
    time_elapsed=(end - start)
    my_text1.config(state="normal")
    my_text1.insert(tk.INSERT,"Dataset creado en {}. Número de filas: {}. Tiempo: {:.3f} se
    my_text1.config(state="disabled")


def graph():
    df1 = df.iloc[:,[0,1,2,14,15,16,17,18,19,20,21,22,23,24,25]]
    figure = plt.figure(num=None, figsize=(12, 7))
    bar2 = FigureCanvasTkAgg(figure, root)
    bar2.get_tk_widget().grid(row=9,column=0,columnspan=8)
    df_mean = clean(df1)
    fh=df_mean
    fh['mean'] = df_mean.iloc[:,3:15].mean(axis=1)
    fd=fh.loc[(fh['mean']<260)]
    fa=fd.loc[(fd['mean']>5)]

    if drop1.get()=="Boxplot":
        sns.set(style="whitegrid")
        ax = sns.boxplot(x=fa['mean'], color='darkorange',linewidth=2.5)
        ax.set_title('BOXPLOT para VALORES FUERA DE RANGO', fontsize=20)
        ax.set_xlabel('Consumo en kWh', fontsize=16)
        for patch in ax.artists:
            r, g, b, a = patch.get_facecolor()
            patch.set_facecolor((r, g, b, .5))
        plt.show()

    if drop1.get()=="Frecuencia":
        sns.set(style="whitegrid")
        ax = sns.histplot(fa['mean'], bins=20, color="r")
        ax.set_title('Histograma de frecuencia', fontsize=20)
        ax.set_ylabel('Número de usuarios', fontsize=16)
        ax.set_xlabel('Consumo en kWh', fontsize=16)
        plt.show()

    if drop1.get()=="Densidad":
        sns.set(style="whitegrid")
        ax = sns.kdeplot(fa['mean'], shade=True, color="r")
        ax.set_title('Densidad de los datos', fontsize=20)
        ax.set_xlabel('Consumo en kWh', fontsize=16)
```

```
        plt.show()

    if drop1.get()=="Consumo por Usuario":

        df1 = df.iloc[:,[0,1,2,14,15,16,17,18,19,20,21,22,23,24,25]]
        _cols = list(df1.columns)

        #Eliminamos la palabra mes de cada columna del archivo para tener solo valores nume
        df1.columns = pd.Index(_cols[:3] + [int(c.replace("MES_","")) for c in _cols[3:]])
        del _cols
        cols = df1.columns.tolist()

        #Cambiamos el orden de los meses para que vayan desde mas antiguo a mas nuevo
        cols=['CLICOD', 'MDENUMFAB', 'IDCNOMIMP',1 ,2,3,4,5,6,7,8,9,10,11,12]
        df1 = df1[cols]

        def series_from_id(_CLICOD:str) -> pd.DataFrame:
            return (df1.loc[df1.CLICOD == _CLICOD].iloc[:,3:].T, df1.loc[df1.CLICOD == _CLI
        CLIENTE=my_text2.get("1.0","end-1c")
        y=series_from_id(int(CLIENTE))[0].values

        x= [1,2,3,4,5,6,7,8,9,10,11,12]
        y = list(itertools.chain(*y))
        z=(series_from_id(int(CLIENTE))[1].values)[0][1]
        medidor=(series_from_id(int(CLIENTE))[1].values)[0][0]
        y1=[np.mean(y)]*12

        plt.grid()
        plt.plot(x,y,color='green',marker='o',linewidth=3, markersize=15,label='Consumo Men
        plt.plot(x,y1,color='blue',marker='s',linewidth=3, markersize=10, label='Consumo Pr
        plt.fill_between(x,y, color='green', alpha=0.1)
        plt.fill_between(x,y,y1[0],where=y <= y1[0],color='red', alpha=0.1, interpolate=Tru
        plt.title('Cliente No:{}.   Med. No:{}.   Nombre: {}'.format(CLIENTE,medidor,z),fon
        plt.legend(loc="best", prop={'size': 15})
        plt.xlabel("Mes",fontsize=16)
        plt.ylabel("Consumo en kWh",fontsize=16)
        plt.grid()
        plt.show()

def cluster1():

    df1 = df.iloc[:,[0,1,2,14,15,16,17,18,19,20,21,22,23,24,25]]

    dfclean = clean(df1)
    _cols = list(dfclean.columns)

    #Eliminamos la palabra mes de cada columna del archivo para tener solo valores numerico
    dfclean.columns = pd.Index(_cols[:3] + [int(c.replace("MES_","")) for c in _cols[3:]])
    del _cols
    cols = dfclean.columns.tolist()

    #Cambiamos el orden de los meses para que vayan desde mas antiguo a mas nuevo
    cols=['CLICOD', 'MDENUMFAB', 'IDCNOMIMP',1,2,3,4,5,6,7,8,9,10,11,12]
    dfclean = dfclean[cols]

    fh=dfclean
    fh['mean'] = dfclean.iloc[:,3:15].mean(axis=1)
    fd=fh.loc[(fh['mean']<260)]
    fa=fd.loc[(fd['mean']>5)]

    fa=fa.drop(columns=['mean','MDENUMFAB', 'IDCNOMIMP'])
```

```
    def series_from_id(_CLICOD:str) -> pd.DataFrame:
        """
        Get a daily time series for a single id
        """
        return fa.loc[fa.CLICOD == _CLICOD].iloc[:,1:].T

    index_list = fa['CLICOD'].tolist()
    numerodeseries=202717
    #numerodeseries=len(index_list)
    series=[]
    for i in tqdm(range(numerodeseries)): #
        series.append(series_from_id(index_list[i]))

    numero_series= numerodeseries
    series1=[]
    global names
    names=[]
    for i in range(numero_series):
        series1.append(series[i].values)
        names.append(index_list[i])


    scaledSeries=series1[i]
    for i in tqdm(range(numero_series)):
        scaler = MinMaxScaler()
        series1[i] = MinMaxScaler().fit_transform(series1[i])
        series1[i] = series1[i].reshape(len(series1[i]))

    # Instantiate the clustering model and visualizer

    global series2
    series2 = np.array(series1)
    series2.shape

    figure = plt.figure(num=None, figsize=(12, 7))
    bar2 = FigureCanvasTkAgg(figure, root)
    bar2.get_tk_widget().grid(row=9,column=0,columnspan=8)

    model = KMeans()

    visualizer = KElbowVisualizer(model, k=(4,12))

    visualizer.fit(series2)        # Fit the data to the visualizer
    visualizer.show()

    plt.show()


def graph1():

    seed = 0
    np.random.seed(seed)
    km = TimeSeriesKMeans(n_clusters=7, verbose=True, random_state=seed)
    y_pred = km.fit_predict(series2)

    global labels
    labels = km.labels_
```

```
    figure = plt.figure(num=None, figsize=(12, 7))
    bar2 = FigureCanvasTkAgg(figure, root)
    bar2.get_tk_widget().grid(row=9,column=0,columnspan=8)

    for yi in range(7):

        plt.subplot(3,3, yi + 1)
        for xx in series2[y_pred == yi]:
            plt.plot(xx.ravel(), "k-", alpha=.002)
        plt.plot(km.cluster_centers_[yi].ravel(), "r-")
        #plt.xlim(0, sz)
        #plt.ylim(-4, 4)
        plt.text(0.55, 0.85,'Cluster %d' % (yi + 1),
                 transform=plt.gca().transAxes,color='y')
        if yi == 1:
            plt.title("Euclidean $k$-means")
    plt.show()

def graph2():
    seed = 0
    np.random.seed(seed)
    dba_km = TimeSeriesKMeans(n_clusters=7,
                              n_init=2,
                              metric="dtw",
                              verbose=True,
                              max_iter_barycenter=10,
                              n_jobs=-1,
                              random_state=seed)
    y_pred = dba_km.fit_predict(series2)

    global labels1
    labels1 = dba_km.labels_

    figure = plt.figure(num=None, figsize=(12, 7))
    bar2 = FigureCanvasTkAgg(figure, root)
    bar2.get_tk_widget().grid(row=9,column=0,columnspan=8)

    for yi in range(7):

        plt.subplot(3, 3, yi + 1)
        for xx in series2[y_pred == yi]:
            plt.plot(xx.ravel(), "k-", alpha=.002)
        plt.plot(dba_km.cluster_centers_[yi].ravel(), "r-")
        plt.text(0.55, 0.85,'Cluster %d' % (yi + 1),
                 transform=plt.gca().transAxes,color='y')
        if yi == 1:
            plt.title("DBA $k$-means")

def graph3():
    seed = 0
    np.random.seed(seed)
    sdtw_km = TimeSeriesKMeans(n_clusters=7,
                               metric="softdtw",
                               metric_params={"gamma": .01},
                               verbose=True,
                               random_state=seed)
    y_pred = sdtw_km.fit_predict(series2)

    global labels2
    labels2 = sdtw_km.labels_
```

```
    figure = plt.figure(num=None, figsize=(12, 7))
    bar2 = FigureCanvasTkAgg(figure, root)
    bar2.get_tk_widget().grid(row=9,column=0,columnspan=8)

    for yi in range(7):

        plt.subplot(3, 3, yi + 1)
        for xx in series2[y_pred == yi]:
            plt.plot(xx.ravel(), "k-", alpha=.002)
        plt.plot(sdtw_km.cluster_centers_[yi].ravel(), "r-")
        #plt.xlim(0, sz)
        #plt.ylim(-4, 4)
        plt.text(0.55, 0.85,'Cluster %d' % (yi + 1),
                 transform=plt.gca().transAxes, color='y')
        if yi == 1:
            plt.title("Soft'DTW $k$-means")
    plt.show()

def identificar():
    val = my_text3.get("1.0","end-1c")


    primer_cluster = pd.DataFrame(
    {'Usuario': names,
     'Cluster': labels,
    })
    primer_cluster1 = primer_cluster[primer_cluster["Cluster"] == (int(val[0])-1)]


    segundo_cluster = pd.DataFrame(
    {'Usuario': names,
     'Cluster': labels1,
    })
    segundo_cluster1 = segundo_cluster[segundo_cluster["Cluster"] == (int(val[2])-1)]


    tercer_cluster = pd.DataFrame(
    {'Usuario': names,
     'Cluster': labels2,
    })
    tercer_cluster1 = tercer_cluster[tercer_cluster["Cluster"] == (int(val[4])-1)]


    infractores=list(set(primer_cluster1.Usuario) & set(segundo_cluster1.Usuario) & set(ter

    x=df[df['CLICOD'].isin(infractores)]

    df2 = pd.read_excel(r"C:\Users\richa\Desktop\TESIS FINAL UTEQ\DATOS USUARIOS CNEL- MEDI
    mdf = pd.concat(df2, axis=0, ignore_index=True)
    df3=mdf[['CODIGOCLIENTE', 'SUBESTACION']]
    df3=df3.replace(['S/E FLAVIO ALFARO'], ['FLAVIO ALFARO'])
    df3.rename(columns={'CODIGOCLIENTE': 'CLICOD'}, inplace=True)
    int_df = x.merge(df3, how='left', on='CLICOD')
    int_df.fillna('NO_ASIGNADO', inplace=True)
    res = list(tuple(int_df.groupby('SUBESTACION')))
    tot=df3.groupby('SUBESTACION')
    writer = pd.ExcelWriter(r'C:\Users\richa\Desktop\Reporte_Infractores.xlsx', engine='xls
    barras=[]
    nombres=[]
    resid=[]
    for i in range(len(res)):
```

```
        h=(res[i][0])
        if h==('NO_ASIGNADO'):
            a=(res[i][1])
            a.to_excel(writer, sheet_name='{}%'.format(res[i][0]))
            datos
            barras.append(len(a))
            nombres.append(res[i][0])
            resid.append(res[i][1].USOCOD.value_counts().RD)
        else:
            a=(res[i][1])
            porcentaje=(len(a)/len(tot.get_group(res[i][0])))*100
            porcentaje = round(porcentaje, 3)
            a.to_excel(writer, sheet_name='{}-{}%'.format(res[i][0],porcentaje))
            barras.append(len(a))
            nombres.append(res[i][0])
            resid.append(res[i][1].USOCOD.value_counts().RD)

    writer.save()

    porcentaje=[]
    for i in range(len(resid)):
        valor=100-((barras[i]-resid[i])/barras[i])*100
        porcentaje.append(int(valor))


    figure = plt.figure(num=None, figsize=(12, 7))
    bar2 = FigureCanvasTkAgg(figure, root)
    bar2.get_tk_widget().grid(row=9,column=0,columnspan=8)


    def add_value_label(x_list,y_list):
        for i in range(len(x_list)):
            plt.text(i,y_list[i],y_list[i], ha="center")


    plt.bar(nombres, barras, color = 'g', alpha=0.5 )
    plt.bar(nombres, resid, color='darkorange', linewidth=2.5)

    red_patch = mpatches.Patch(color='g', alpha=0.5, label='Otros')
    blue_patch = mpatches.Patch(color='darkorange', label='Residenciales')

    add_value_label(nombres,porcentaje)

    plt.legend(handles=[red_patch, blue_patch],  loc='best')

    plt.ylabel('Número de Infractores', fontsize=14)

    plt.xticks(rotation=90)
    plt.grid(True)
    plt.show()




# Boton para cargar los datos
button1=Button(root,text="          Mostrar Datos          ", command=search)
button1.grid(row=3,column=0)
```