UNIVERSIDAD DE EL SALVADOR

FACULTAD DE INGENIERÍA Y ARQUITECTURA

ESCUELA DE INGENIERÍA ELÉCTRICA

APLICACIÓN DE HERRAMIENTAS DE MACHINE LEARNING PARA LA PREDICCIÓN DE INTERRUPCIONES EN SISTEMAS DE DISTRIBUCIÓN ELÉCTRICA EN EL SALVADOR

PRESENTADO POR:

CESAR ALBERTO MORALES ORELLANA

EDUARDO ENRIQUE MUNTO UCEDA

PARA OPTAR AL TITULO DE:

INGENIERO ELECTRICISTA

CIUDAD UNIVERSITARIA, NOVIEMBRE DE 2025

UNIVERSIDAD DE EL SALVADOR

RECTOR:

MSC. JUAN ROSA QUINTANILLA

SECRETARIO GENERAL:

LIC. PEDRO ROSALÍO ESCOBAR CASTANEDA

FACULTAD DE INGENIERÍA Y ARQUITECTURA

DECANO:

MSC. LUIS SALVADOR BARRERA MANCÍA

SECRETARIO:

ARQ. RAÚL ALEXANDER FABIÁN ORELLANA

ESCUELA DE INGENIERIA ELECTRICA

DIRECTOR:

ING. WERNER DAVID MELÉNDEZ VALLE

UNIVERSIDAD DE EL SALVADOR

FACULTAD DE INGENIERÍA Y ARQUITECTURA

ESCUELA DE INGENIERÍA ELÉCTRICA

Trabajo de Graduación previo a la opción al grado de:

INGENIERO ELECTRICISTA

Titulo:

APLICACIÓN DE HERRAMIENTAS DE MACHINE LEARNING PARA LA PREDICCIÓN DE INTERRUPCIONES EN SISTEMAS DE DISTRIBUCIÓN ELÉCTRICA EN EL SALVADOR

Presentado por:

CESAR ALBERTO MORALES ORELLANA

EDUARDO ENRIQUE MUNTO UCEDA

Trabajo de Graduación Aprobado por

Docente Asesor:

ING. WERNER DAVID MELENDEZ VALLE

SAN SALVADOR, NOVIEMBRE DE 2025

Trabajo de Graduación Aprobado por

Docente Asesor:

ING. WERNER DAVID MELENDEZ VALLE

NOTA Y DEFENSA FINAL

En esta fecha, lunes 6 de octubre de 2025, en la Sala de Lectura de la Escuela de Ingeniería Eléctrica, a las 2:00 p.m. horas, en presencia de las siguientes autoridades de la Escuela de Ingeniería Eléctrica de la Universidad de El Salvador:

1. Ing. Werner David Meléndez Valle
   Director

   Firma

2. MSc. José Wilber Calderón Urrutia
   Secretario

   Firma

Y, con el Honorable Jurado de Evaluación integrado por las personas siguientes:

- ING. WERNER DAVID MELÉNDEZ VALLE
  (Docente Asesor)

  Firma

- ING. LUIS ERNESTO ESCOBAR BRIZUELA

  Firma

- DR. CARLOS OSMIN POCASANGRE JIMÉNEZ

  Firma

Se efectuó la defensa final reglamentaria del Trabajo de Graduación:

APLICACIÓN DE HERRAMIENTAS DE MACHINE LEARNING PARA LA PREDICCIÓN DE INTERRUPCIONES EN SISTEMAS DE DISTRIBUCIÓN ELÉCTRICA EN EL SALVADOR

A cargo de los Bachilleres:

- MUNTO UCEDA EDUARDO ENRIQUE
- MORALES ORELLANA CESAR ALBERTO

Habiendo obtenido en el presente Trabajo una nota promedio de la defensa final: 8.9
(OCHO PUNTO NUEVE)

AGRADECIMIENTOS

Antes que todo, quiero agradecer profundamente a **Dios**, por ser mi guía en cada paso de este camino. Gracias por darme la fortaleza en los momentos difíciles, la sabiduría para continuar cuando las cosas se complicaban y por permitirme llegar hasta este momento tan especial. Sin Su presencia y Su gracia, nada de esto habría sido posible.

A mi **familia**, mi mayor fuente de amor y motivación. A mi **mamá, Roxana de Morales**, gracias por estar siempre a mi lado, por acompañarme en cada decisión, idea y meta que me he propuesto en la vida. Tu amor incondicional, tu fe en mí y tu apoyo constante han sido la base sobre la cual he podido construir mis sueños. Gracias por tus palabras, tus consejos y por nunca dejarme rendir.

A mi **papá, César Morales**, gracias por tus enseñanzas, por tus consejos llenos de sabiduría y por ser ese apoyo firme en la recta final de mi carrera. Tu confianza en mis capacidades y tu ayuda, tanto personal como profesional, me impulsaron a seguir adelante y a esforzarme cada día por ser un mejor profesional y persona.

A mi **hermana, Akemi Morales**, gracias por tu apoyo en cada momento de este proceso, por tu compañía, tu paciencia y por estar siempre ahí cuando lo necesité. Tu cariño y tus palabras de aliento fueron un gran motor para mí, especialmente en los días en que más los necesitaba.

A mis **abuelas**, gracias por sus oraciones, por su amor inmenso y por estar siempre pendientes de mí. Sus palabras y su fe me dieron paz y fortaleza, recordándome siempre que todo esfuerzo tiene su recompensa y que los sueños se cumplen con perseverancia.

A mis **amigos**, aquellos que se volvieron parte fundamental de este viaje, gracias por compartir conmigo los días de estudio, las desveladas, las risas y los momentos de estrés. Cada uno de ustedes aportó algo valioso en mi formación, y me alegra haber recorrido este camino acompañado de personas tan auténticas y solidarias.

También quiero expresar mi agradecimiento a mis **mentores**: **Miguel Carrillo, Luis Flores, Josué Palacios, Miguel Portillo y Armando Solórzano**, quienes con su guía, paciencia y consejos me ayudaron a crecer y a creer más en mis capacidades. Gracias por compartir su conocimiento y su tiempo para ayudarme a mejorar.

Finalmente, mi agradecimiento especial a nuestro **asesor, Werner Meléndez**, por su compromiso, por su apoyo constante y por acompañarnos en todo este proceso con dedicación y paciencia. Gracias por orientarnos, por sus valiosas observaciones y por motivarnos a dar siempre lo mejor de nosotros.

A todos ustedes, gracias de corazón. Cada palabra, cada gesto y cada momento compartido formaron parte de este logro. Este triunfo no es solo mío, sino también de todos los que creyeron en mí y me acompañaron a lo largo de este camino.

Cesar Morales

AGRADECIMIENTOS

Estoy agradecido con Dios, con todo lo que esta vida me ha otorgado, con mi familia y en especial con mi madre, mi padre, mis hermanos, quienes son mis seres queridos, a quienes aprecio con el corazón en cada paso que doy.

Agradezco profundamente a mi madre, **Rosa Estrada**, quien siempre me apoyó incondicionalmente a lo largo de mi vida. Ella se encargó de que nunca me faltase nada y nunca dudó en darlo todo por el bienestar y desarrollo de nuestra familia, actuando como madre y padre, demostrándome con su ejemplo el significado del esfuerzo y el sacrificio.

A mi padre, **Nelson Munto**, por su responsabilidad y entrega constante; a mi hermano, **Alex Munto**, quien me brindó lo necesario para cumplir con las exigencias de la universidad y por despertar en mí el interés por la ciencia y la ingeniería.

A mis tres hermanos, **Alex, Princela y Amanda**, quienes han estado a mi lado en cada desafío, siendo una referencia de constancia, dedicación y esfuerzo, dejando una huella que también inspiró mi crecimiento académico.

A **mis tíos,** que a través de su ejemplo me enseñaron el valor de la responsabilidad, el respeto, el trabajo duro y la fe.

A mí mismo, por mantener la constancia, perseverancia y dar mi mejor esfuerzo, aun cuando todo estuviera en contra para no dejar de avanzar.

Agradezco también a quienes me brindaron la oportunidad de desarrollarme profesionalmente e instruyeron en el ámbito eléctrico: **Miguel Carrillo, Luis Flores, Palacios y Miguel Portillo**, por su guía, conocimientos, consejos y por acompañarme con amistad y apoyo, siendo parte importante de mi desarrollo personal y profesional.

Además, expreso mi agradecimiento al **Ing. Armando Solórzano**, por su liderazgo, confianza y acompañamiento. Su calidad humana, orientación y apoyo constante fueron fundamentales en mi formación y crecimiento.

A su vez, agradezco al **Ing. Numa Jiménez**, por su apoyo y confianza, al compartir con nosotros sus conocimientos y experiencia.

Finalmente, deseo expresar mi profundo agradecimiento a nuestro asesor **Ing. Wener Meléndez**, por su valiosa orientación, paciencia y acompañamiento constante durante cada etapa de la elaboración de esta tesis. Sus aportes y su guía fueron determinantes para la finalización de este trabajo.

A todos ustedes, quienes han contribuido de manera invaluable a mi crecimiento personal y profesional, les expreso mi más profundo agradecimiento. Este logro no es únicamente mío, sino también resultado del apoyo, la confianza y las enseñanzas que he recibido de cada uno de ustedes a lo largo de este camino. Finalizando así una etapa significativa en mi vida, con la satisfacción de haber llegado hasta aquí gracias a su acompañamiento, y con la convicción y esperanza de afrontar los nuevos desafíos que vienen con la misma dedicación y gratitud que me han inspirado.

Eduardo Munto

Índice

Índice de Tablas

Índice de Figuras

Introducción

Los sistemas de distribución eléctrica constituyen el eslabón final del suministro de energía y, por tanto, el punto donde convergen las exigencias técnicas de continuidad, calidad del producto y seguridad para el usuario final. En El Salvador, estas exigencias están normadas por la Superintendencia General de Electricidad y Telecomunicaciones (SIGET), que desde 2005 estableció índices, tolerancias y formatos estandarizados de reporte que las empresas distribuidoras remiten mensualmente. La disponibilidad de esta información estructurada ha abierto la posibilidad de abordar, con rigor cuantitativo, problemas clásicos de la operación de las redes, como la predicción de fallas de transformadores de distribución y la priorización de mantenimiento.

El contexto operativo de la zona central del país con niveles de tensión de 4.16 kV y 23 kV, alta humedad, vegetación abundante y una de las mayores actividades de descargas atmosféricas en Centroamérica condiciona fuertemente el comportamiento de los activos. A ello se suman prácticas de operación y crecimiento de la demanda que, en conjunto, definen perfiles de carga heterogéneos a lo largo del tiempo. El análisis histórico 2008–2024 evidencia un incremento en las fallas asociadas a degradación prematura y fin de vida, mientras que se observa una disminución relativa de eventos por descargas atmosféricas y por cortocircuitos en baja tensión. Estas tendencias sugieren que los factores internos del activo (condición térmica, esfuerzos dieléctricos, estado de conexiones, fugas) han ganado peso relativo frente a causas exógenas.

La disponibilidad de datos estandarizados (SIGET), complementada con historiales de activación de protecciones y perfiles de consumo por tarifa (factores de responsabilidad horaria actualizados quinquenalmente), permite construir un conjunto de variables explicativas con valor operativo: cargabilidad máxima (MAX), composición de clientes por tarifa, número de usuarios conectados, potencia nominal efectiva, urbanidad/densidad de carga (U/R) y activaciones

históricas de protecciones. Sobre esta base, se conformó un universo de 19,344 transformadores en operación y un conjunto de entrenamiento supervisado 2013–2023 con 257 casos positivos (transformadores que fallaron por degradación o fin de vida) y casos negativos.

Frente a la naturaleza estocástica del fenómeno de falla donde coexisten relaciones lineales y no lineales, y donde el desbalance entre clases es significativo, esta tesis adopta un enfoque de aprendizaje automático que combina modelos supervisados (Regresión Logística, Árbol de Decisión y Bosque Aleatorio, Redes Neuronales) con un componente no supervisado (DBSCAN) para detección de contextos atípicos. Los modelos supervisados permiten estimar la probabilidad de falla y analizar la contribución de las variables, mientras que el análisis por densidad identifica subconjuntos de transformadores cuyo comportamiento operativo se aparta del patrón global y que, con frecuencia, se vinculan a eventos de falla posteriores.

Los resultados experimentales muestran que, aunque todos los modelos alcanzan altas especificidades (buena clasificación de transformadores sin falla), la sensibilidad (detección de fallas) es más desafiante debido al desbalance de clases. El Bosque Aleatorio y la Red Neuronal (MLP) proporcionan el mejor compromiso desempeño–robustez; la Regresión Logística y el Árbol de Decisión aportan interpretabilidad para auditoría y explicación de riesgos; y DBSCAN resulta útil como herramienta de screening para anomalías sin etiquetas. Sobre estas salidas se construye un Índice Compuesto de Riesgo (ICR) que integra tres dimensiones ponderadas: probabilidad de falla, condición de carga y consecuencia económica. Este índice permite clasificar los activos en cuatro estados (Bueno, Regular, Malo y Crítico) y sustentar una estrategia de mantenimiento escalable y con la criticidad del servicio.

Objetivos

Objetivo General

Proponer una herramienta analítica basada en machine learning para la predicción de interrupciones en sistemas de distribución eléctrica de la zona central de El Salvador.

Objetivos Específicos

Recopilar datos históricos asociados a las interrupciones en el servicio de los sistemas de distribución eléctrica en algunas zonas de San Salvador, incluyendo el punto de falla y usuarios afectados.

Identificar factores que influyen en las interrupciones en el servicio, sus causas y posibles alternativas de acciones preventivas.

Desarrollar algoritmos de machine learning que permitan predecir las interrupciones en los sistemas de distribución eléctrica en la zona central de El Salvador.

Evaluar la precisión y efectividad de los algoritmos desarrollados, usando datos históricos y simulaciones.

Proponer estrategias de prevención de interrupciones, utilizando el algoritmo previamente desarrollado.

Planteamiento del problema

Las interrupciones en los sistemas de distribución eléctrica causan importantes inconvenientes económicos y sociales. Los métodos tradicionales de mantenimiento, basados en cronogramas fijos y reacciones a fallos, resultan insuficientes para garantizar una alta continuidad del servicio. Existe la necesidad de una herramienta que, mediante el uso de técnicas avanzadas de machine learning, pueda predecir interrupciones y permitir la planificación proactiva del mantenimiento.

Los transformadores de distribución son fundamentales en la red eléctrica, ya que convierten la energía de alta tensión en niveles utilizables por los consumidores. Sin embargo, estos equipos son propensos a fallos debido a la degradación natural de sus componentes y, especialmente, a la sobrecarga prolongada que excede la capacidad del transformador, acelera su deterioro y aumenta la frecuencia de fallos. Estos fallos no solo interrumpen el suministro eléctrico, sino que también requieren un tiempo considerable para la reparación o reemplazo del transformador afectado, lo cual puede prolongar la interrupción del servicio, causa pérdidas económicas significativas y afecta a la calidad del suministro eléctrico.

Alcances

- **Cobertura:** La herramienta estará diseñada para sistemas de distribución eléctrica de mediana y gran escala.
- **Funcionalidad:** La herramienta ofrecerá predicciones sobre posibles interrupciones y sugerencias de mantenimiento preventivo.
- **Integración:** Será compatible con los sistemas de gestión de distribución eléctrica existentes.

Justificación

El fomento del uso de tecnologías avanzadas en la gestión de sistemas de distribución eléctrica es fundamental para incrementar la eficiencia operativa y garantizar un servicio de calidad. La adopción de tecnologías nuevas contribuye significativamente a la reducción de la frecuencia y duración de las interrupciones del servicio, lo que se traduce en una mayor continuidad del suministro y satisfacción. Asimismo, la mejora en la planificación y ejecución del mantenimiento preventivo permite prevenir fallas inesperadas, asegurando un funcionamiento más confiable y optimizando los recursos disponibles.

Antecedentes

En los últimos años, la fiabilidad en los sistemas de distribución eléctrica ha cobrado gran importancia debido al aumento en la demanda de energía y la creciente dependencia de la electricidad en todas las actividades económicas y sociales. Las interrupciones en el suministro eléctrico no solo provocan pérdidas económicas significativas, sino que también pueden afectar la seguridad y el bienestar de la población. Tradicionalmente, las compañías eléctricas han dependido de métodos reactivos para abordar las interrupciones, reparando los fallos después de que ocurren.

Con el avance de las tecnologías de la información y el auge del machine learning, se ha abierto una nueva perspectiva para la gestión proactiva de los sistemas eléctricos. Estas técnicas permiten analizar grandes volúmenes de datos históricos para identificar patrones y predecir posibles fallos en la red antes de que ocurran, lo que representa un cambio significativo respecto a las metodologías convencionales.

Capitulo I: Marco Teórico y Fundamentos

1.1 Sistemas de Distribución Eléctrica

1.1.1 Estructura y Características de los Sistemas de Distribución Eléctrica

Los sistemas de distribución eléctrica representan la etapa final del suministro de energía, con la función de transportar la electricidad desde las subestaciones hasta los usuarios finales, que incluyen hogares, comercios, industrias e infraestructura pública. Estos sistemas transforman la energía de alta tensión, recibida de las redes de transmisión, en niveles de media tensión (generalmente entre 4,000 y 35,000 voltios) y baja tensión (120/240 voltios para residencias, hasta 480 voltios para industrias), adecuados para el consumo. La estructura de un sistema de distribución integra componentes físicos, sistemas de monitoreo y estrategias operativas diseñadas para garantizar un suministro continuo, seguro y de alta calidad, manteniendo estabilidad en parámetros como el voltaje y la frecuencia (Glover et al., 2016).

1.1.1.1 Componentes Físicos.

La infraestructura física de un sistema de distribución comprende diversos elementos interconectados, cada uno con roles específicos para transportar, transformar y proteger la energía eléctrica. Las líneas de distribución pueden ser aéreas, instaladas en postes con conductores de aluminio o cobre, o subterráneas, utilizando cables aislados para mayor seguridad y resistencia a condiciones ambientales. Las líneas aéreas son más económicas y fáciles de inspeccionar o reparar, pero están expuestas a eventos climáticos, como vientos fuertes o tormentas. Las líneas subterráneas, frecuentes en áreas urbanas densas, son más costosas, pero ofrecen mayor confiabilidad y estética, al reducir la exposición a daños externos como caídas de árboles o relámpagos (Kersting, 2017).

Los transformadores de distribución convierten el voltaje de niveles altos a niveles más bajos, permitiendo que la electricidad sea utilizable por los consumidores. Por ejemplo, un transformador puede reducir la media tensión de 13,800 voltios a 120/240 voltios para hogares.

Los transformadores están diseñados para operar con alta eficiencia, minimizando pérdidas de energía por calor, y pueden ser monofásicos, para cargas pequeñas como residencias, o trifásicos, para industrias o comercios grandes. Incorporan sistemas de protección, como fusibles o relés, para evitar daños por sobrecargas o fallos eléctricos (Gonen, 2015).

Los equipos de maniobra y protección incluyen interruptores, fusibles, recierres automáticos y relés. Los interruptores permiten conectar o desconectar secciones de la red manual o automáticamente. Los fusibles protegen contra sobrecorrientes al interrumpir el flujo eléctrico en caso de fallos graves, como cortocircuitos. Los recierres automáticos detectan fallos temporales, como el contacto momentáneo de una rama con una línea, y reintentan el suministro en fracciones de segundo, evitando interrupciones prolongadas. Los relés monitorean continuamente parámetros como la corriente o el voltaje, activando protecciones rápidamente para prevenir daños a los equipos (Momoh, 2008).

Los medidores registran el consumo de energía de los usuarios, expresado en kilovatios-hora. Los medidores tradicionales son mecánicos o electrónicos, pero los medidores inteligentes, parte de la Infraestructura de Medición Avanzada (AMI), recopilan datos en tiempo real sobre consumo, voltaje, corriente y calidad de la energía, como la estabilidad del voltaje. Los medidores inteligentes se comunican con centros de control mediante redes inalámbricas, fibra óptica o señales transmitidas por las propias líneas eléctricas, conocidas como comunicación por corriente portadora (Weedy et al., 2012).

En redes aéreas, los postes, fabricados en madera, concreto o metal, sostienen los conductores y equipos, diseñados para soportar cargas mecánicas, como el peso de cables, y climáticas, como vientos fuertes. Los sistemas de puesta a tierra, conectados a los postes o equipos, protegen contra descargas eléctricas y relámpagos, dirigiendo corrientes no deseadas al suelo (Glover et al., 2016).

Los capacitores compensan las variaciones en la demanda de energía reactiva, mejorando la eficiencia y estabilidad del sistema al reducir pérdidas. Los reguladores de voltaje

ajustan dinámicamente el voltaje para mantenerlo dentro de rangos aceptables, especialmente en redes extensas donde las caídas de voltaje son comunes debido a la longitud de las líneas o la alta demanda (Kundur, 1994).

1.1.1.2 Topologías de Red.

Los sistemas de distribución se organizan en diferentes configuraciones o topologías, que determinan cómo fluye la electricidad y cómo se gestiona la continuidad del suministro ante fallos. La topología radial es la más simple y común, donde la electricidad fluye desde un único punto de alimentación, como una subestación, hacia los usuarios en una configuración lineal o ramificada. Es económica y fácil de diseñar, ya que requiere menos infraestructura, pero un fallo en el punto de origen o en una línea principal interrumpe el suministro a todos los usuarios conectados aguas abajo, lo que reduce la confiabilidad en áreas críticas como hospitales o centros comerciales (Kersting, 2017).

La topología en anillo conecta puntos de alimentación en un circuito cerrado, permitiendo que la electricidad fluya por múltiples rutas. Si ocurre un fallo en una sección, los interruptores automáticos o manuales reconfiguran la red para redirigir la energía por una ruta alternativa, manteniendo el suministro. Es más confiable que la radial, pero requiere más equipos, como interruptores adicionales, y un diseño más complejo, lo que incrementa los costos. Es común en áreas urbanas donde la continuidad del servicio es prioritaria (Gonen, 2015).

La topología mallada ofrece el mayor nivel de redundancia, con múltiples conexiones entre nodos de la red, permitiendo que la electricidad fluya por varias rutas simultáneamente. Esto asegura alta confiabilidad, ya que un fallo en una línea no interrumpe el suministro, pero su diseño y operación son extremadamente complejos y costosos, reservándose para sistemas críticos, como redes industriales, hospitales o grandes centros urbanos (Glover et al., 2016).

La elección de la topología depende de factores como el presupuesto disponible, la densidad de usuarios, la importancia de las cargas (por ejemplo, hospitales frente a zonas rurales) y las normativas locales. En la práctica, muchas redes combinan topologías híbridas, utilizando secciones radiales para áreas de baja densidad y anillos o mallas parciales en zonas urbanas, optimizando el equilibrio entre costo, confiabilidad y complejidad operativa (Grainger & Stevenson, 1994).

1.1.1.3 Operación y Control.

La operación de un sistema de distribución busca garantizar que la electricidad se entregue de manera estable, eficiente y segura, manteniendo parámetros clave como el voltaje, la frecuencia y la calidad de la energía dentro de rangos establecidos por estándares técnicos. Por ejemplo, el voltaje debe permanecer dentro de un margen del ±5% de su valor nominal, como 120 voltios ±6 voltios, y la frecuencia debe ser estable, como 60 Hz ±0.1 Hz en sistemas de 60 Hz (Bergen & Vittal, 1999).

El balanceo de cargas implica distribuir la demanda de energía entre transformadores, líneas y fases para evitar sobrecargas, que pueden causar caídas de voltaje, sobrecalentamiento de equipos o fallos. Por ejemplo, en sistemas trifásicos, las cargas deben distribuirse equitativamente entre las tres fases para evitar desequilibrios que reduzcan la eficiencia y dañen los equipos (Saadat, 2010).

La gestión de la calidad de la energía mitiga problemas como fluctuaciones de voltaje, como caídas breves o picos, armónicos, que son distorsiones causadas por equipos electrónicos, y transitorios, que son picos momentáneos por relámpagos o conmutaciones. Estos problemas afectan la operación de equipos sensibles, como computadoras o maquinaria industrial, y requieren dispositivos como filtros o reguladores (Weedy et al., 2012).

La respuesta a eventos implica coordinar acciones rápidas ante interrupciones, como reconfigurar la red para aislar secciones dañadas, despachar equipos de reparación o comunicar el estado del servicio a los usuarios. Esto requiere sistemas de monitoreo avanzados, procedimientos establecidos y personal capacitado para actuar con rapidez (Momoh, 2008).

Los sistemas de monitoreo y control son esenciales para estas tareas. Los sistemas SCADA (Supervisory Control and Data Acquisition) recopilan datos en tiempo real de sensores instalados en subestaciones, líneas, transformadores y otros equipos. SCADA permite a los operadores visualizar el estado de la red, incluyendo niveles de carga, alarmas de fallos y configuraciones de interruptores, y ejecutar comandos remotos, como abrir o cerrar circuitos para reconfigurar el flujo de energía (International Electrotechnical Commission, 2017).

Otros sistemas, como la Infraestructura de Medición Avanzada (AMI), recopilan datos detallados de medidores inteligentes, mientras que los Sistemas de Gestión de Interrupciones (OMS) y los Sistemas de Gestión de Distribución (DMS) integran información para optimizar la operación, detectar anomalías, planificar mantenimiento preventivo y coordinar respuestas ante eventos críticos, como tormentas o fallos masivos (Momoh, 2008).

1.1.1.4 Normativas y Estándares.

Los sistemas de distribución eléctrica están regulados por estándares internacionales que aseguran su diseño, operación y seguridad. Por ejemplo, el Instituto de Ingenieros Eléctricos y Electrónicos (IEEE) publica guías como la IEEE Std 1366, que define índices de confiabilidad, como la frecuencia y duración promedio de las interrupciones. La Comisión Electrotécnica Internacional (IEC) establece normas como la IEC 61850, que especifica protocolos de comunicación para sistemas de automatización en redes eléctricas, asegurando que los equipos de diferentes fabricantes puedan interoperar. Estos estándares garantizan protección contra riesgos eléctricos, como descargas o incendios, y cumplimiento de requisitos de calidad de la

energía, como límites para fluctuaciones de voltaje o armónicos, promoviendo sistemas seguros y eficientes (IEEE Standards Association, 2018).

1.1.2 Tipos y Causas de Interrupciones

Las interrupciones en los sistemas de distribución eléctrica son eventos que afectan la continuidad del suministro, impidiendo que la electricidad llegue a los usuarios. Las interrupciones transitorias son fallos breves, generalmente resueltos en segundos o minutos por dispositivos automáticos, como recierres o fusibles. Estos fallos suelen ser causados por eventos temporales, como el contacto momentáneo de una rama con una línea eléctrica o una descarga atmosférica que no causa daños permanentes (Short, 2014).

Las interrupciones prolongadas son fallos que requieren intervención humana, como reparaciones en sitio, reemplazo de equipos dañados o reconfiguración manual de la red. Estas interrupciones pueden durar desde unas pocas horas hasta varios días, dependiendo de la gravedad del daño, la disponibilidad de recursos y la accesibilidad de la infraestructura afectada (Momoh, 2008).

Las causas de las interrupciones son variadas. Los factores climáticos, como lluvias intensas, vientos fuertes, relámpagos, tormentas tropicales o inundaciones, son una de las principales causas, ya que pueden dañar líneas aéreas, postes, transformadores o subestaciones. Por ejemplo, los vientos pueden derribar líneas o postes, mientras que los relámpagos pueden provocar cortocircuitos o dañar equipos electrónicos sensibles, como relés o medidores (Glover et al., 2016).

El envejecimiento de la infraestructura aumenta la probabilidad de fallos, ya que componentes como transformadores, cables, aisladores y postes tienen una vida útil limitada, que puede variar entre 20 y 40 años según el material y las condiciones de operación. El desgaste

por uso prolongado, la corrosión, especialmente en ambientes húmedos o salinos, o la falta de mantenimiento preventivo generan fallos mecánicos, eléctricos o térmicos (Kersting, 2017).

Las conexiones defectuosas, como empalmes, terminales o conectores mal instalados, corroídos o de baja calidad, pueden generar arcos eléctricos, sobrecalentamiento o cortocircuitos, que interrumpen el suministro. Estas fallas son comunes en redes antiguas o mal mantenidas, donde las conexiones no se inspeccionan regularmente (Short, 2014).

El contacto de ramas, árboles u otra vegetación con líneas aéreas es una causa frecuente de interrupciones, especialmente durante tormentas o vientos fuertes. La vegetación puede provocar cortocircuitos al conectar conductores o al interrumpir el aislamiento, afectando una o varias líneas, lo que requiere poda regular para prevenir estos incidentes (Momoh, 2008).

Las interferencias externas incluyen el contacto de animales, como aves, roedores o serpientes, con conductores o transformadores, lo que puede causar cortocircuitos o activación de protecciones. Los errores operativos, como configuraciones incorrectas de relés o interruptores, y los actos vandálicos, como el robo de cables de cobre o daños intencionales, también contribuyen (Glover et al., 2016).

La creciente digitalización de los sistemas de distribución, con la adopción de redes inteligentes, medidores inteligentes y sistemas de control como SCADA, introduce vulnerabilidades a ciberataques. Las intrusiones en sistemas de control, la manipulación de datos o el sabotaje remoto pueden interrumpir el suministro, alterar configuraciones críticas o comprometer la seguridad de la red, afectando a miles de usuarios (Weedy et al., 2012).

Comprender estas causas es fundamental para diseñar estrategias de prevención, como el mantenimiento predictivo, la poda regular de vegetación, el reforzamiento de la infraestructura con materiales resistentes y la implementación de medidas de ciberseguridad, como sistemas de detección de intrusos (Momoh, 2008).

1.1.3 Tecnologías de Monitoreo y Gestión

La gestión eficiente de los sistemas de distribución eléctrica depende de tecnologías avanzadas que permiten monitorear el estado de la red, optimizar su operación y responder a eventos críticos de manera oportuna. Estas tecnologías integran hardware, como sensores y medidores, software, como algoritmos de análisis, y redes de comunicación, como inalámbricas o fibra óptica, para recopilar, procesar y analizar datos en tiempo real, mejorando la confiabilidad y la eficiencia del suministro (Momoh, 2008).

La Infraestructura de Medición Avanzada (AMI) utiliza medidores inteligentes instalados en los puntos de consumo, como hogares, comercios o industrias, para recopilar datos detallados sobre el consumo de energía, el voltaje, la corriente, la potencia y la calidad de la energía, como la estabilidad del voltaje o la presencia de distorsiones. Estos medidores se comunican con centros de control mediante redes inalámbricas, como Zigbee o 4G/5G, fibra óptica o señales transmitidas por las líneas eléctricas, conocidas como comunicación por corriente portadora. Los datos de AMI permiten facturación precisa, detección de anomalías, como consumo inusual que indique fallos o fraudes, y planificación de la red (Weedy et al., 2012).

Los Sistemas de Gestión de Interrupciones (OMS) integran datos de múltiples fuentes, incluyendo sensores en la red, alarmas de equipos, reportes de usuarios y datos externos como pronósticos meteorológicos, para localizar fallos rápidamente, estimar el número de usuarios afectados, priorizar acciones de reparación y coordinar la respuesta de los equipos técnicos. Por ejemplo, un OMS puede identificar que un fallo en una línea afecta a 500 hogares y sugerir una reconfiguración para restaurar el servicio parcialmente mientras se repara (Short, 2014).

Los Sistemas de Gestión de Distribución (DMS) optimizan la operación de la red mediante herramientas analíticas que balancean la distribución de cargas entre transformadores y líneas, reconfiguran rutas de suministro para evitar sobrecargas, previenen fallos por congestión y gestionan la integración de fuentes de energía renovables, como paneles solares o turbinas eólicas. Los DMS son esenciales en redes modernas con alta penetración de generación distribuida (Gonen, 2015).

Los sistemas SCADA monitorean y controlan la red en tiempo real, recopilando datos de sensores instalados en subestaciones, líneas, transformadores y otros equipos. Los operadores utilizan interfaces gráficas para visualizar el estado de la red, incluyendo niveles de carga, alarmas de fallos, posiciones de interruptores y flujos de energía. SCADA permite ejecutar comandos remotos, como abrir o cerrar interruptores para aislar secciones dañadas o redirigir el suministro (International Electrotechnical Commission, 2017).

Las tecnologías emergentes incluyen sistemas de detección de fallos basados en inteligencia artificial, que analizan datos históricos y en tiempo real para predecir interrupciones, identificando patrones asociados con fallos inminentes, como fluctuaciones anómalas en el voltaje. Los drones equipados con cámaras y sensores realizan inspecciones aéreas de líneas y postes, detectando daños o vegetación. Los sensores del Internet de las Cosas (IoT) monitorean parámetros como la temperatura de los transformadores, y los sistemas de almacenamiento de energía proporcionan respaldo durante interrupciones (Momoh, 2008).

La implementación de estas tecnologías enfrenta desafíos significativos, como el alto costo de instalación y mantenimiento, la necesidad de personal capacitado para operar sistemas complejos y la integración con infraestructura heredada. Además, la dependencia de redes de comunicación introduce riesgos de ciberseguridad, como ataques que comprometan el control de la red o la privacidad de los datos. Para abordar estos desafíos, las empresas distribuidoras invierten en capacitación, estándares de interoperabilidad y medidas de protección (International Electrotechnical Commission, 2017).

1.2 Fundamentos del Machine Learning

1.2.1 Introducción al Machine Learning

El *machine learning* (ML), o aprendizaje automático, es una disciplina dentro de la inteligencia artificial que se centra en el desarrollo de algoritmos capaces de aprender patrones,

identificar relaciones y tomar decisiones a partir de datos, sin necesidad de instrucciones explícitas programadas para cada tarea. En lugar de seguir reglas predefinidas, los algoritmos de ML ajustan sus parámetros internos basándose en la información proporcionada, mejorando su rendimiento con la experiencia, un proceso conocido como aprendizaje inductivo (Alpaydin, 2020).

El ML combina principios de estadística, matemáticas, informática y optimización para modelar datos complejos, generar predicciones precisas o tomar decisiones informadas. Su versatilidad lo hace aplicable a una amplia gama de problemas, desde clasificar información hasta predecir valores numéricos, agrupar datos similares, detectar anomalías o generar nuevos datos, siendo un pilar fundamental en disciplinas como la ingeniería, la ciencia de datos, la medicina y las finanzas (Bishop, 2006).

El proceso de desarrollo de un modelo de ML consta de varias etapas clave, cada una esencial para lograr resultados confiables. La recopilación de datos implica obtener un conjunto de datos representativo que incluya las características relevantes del problema, como medidas de consumo eléctrico, tipo de usuario o series de tiempo. La calidad y cantidad de los datos son críticas para el éxito del modelo, ya que datos incompletos o sesgados pueden llevar a predicciones erróneas (James et al., 2013).

El preprocesamiento prepara los datos para el análisis, incluyendo limpiar datos incorrectos o incompletos, como eliminar valores atípicos o completar datos faltantes, normalizar escalas para que todas las variables tengan rangos comparables, codificar variables categóricas en formatos numéricos, como convertir categorías en valores binarios, y seleccionar las características más relevantes para reducir la complejidad y mejorar la eficiencia del modelo (Tan et al., 2019).

El entrenamiento ajusta el modelo a los datos mediante un algoritmo que optimiza su capacidad para predecir o clasificar correctamente, ajustando parámetros internos para capturar patrones en los datos. Este proceso implica comparar las predicciones del modelo con los valores

reales y realizar ajustes iterativos para mejorar la precisión, utilizando técnicas de optimización que buscan el mejor conjunto de parámetros (Hastie et al., 2009).

La evaluación mide el rendimiento del modelo utilizando métricas específicas, como el porcentaje de predicciones correctas, conocido como precisión, para tareas de clasificación, o la diferencia promedio entre valores predichos y reales, conocido como error, para regresión. La evaluación se realiza en un conjunto de datos separado del usado para el entrenamiento, asegurando que el modelo pueda generalizar a datos nuevos (Murphy, 2022).

El despliegue implementa el modelo en un entorno operativo, integrándolo con sistemas de datos para generar predicciones en tiempo real o por lotes. Esto incluye monitorear el rendimiento del modelo para detectar posibles degradaciones, como cambios en los patrones de los datos que requieran reentrenamiento, asegurando que las predicciones sigan siendo precisas (Russell & Norvig, 2020).

1.2.2 Tipos de Aprendizaje en Machine Learning

1.2.2.1 Aprendizaje Supervisado.

El aprendizaje supervisado utiliza datos etiquetados, donde cada entrada, representada por un conjunto de características, está asociada a una salida conocida, llamada etiqueta. El objetivo es construir un modelo que aprenda a asociar las entradas con las etiquetas correspondientes, de modo que pueda predecir etiquetas para datos nuevos no vistos previamente (Russell & Norvig, 2020).

El aprendizaje supervisado se divide en clasificación, que predice categorías discretas, por ejemplo, determinar si un usuario es residencial o comercial, y regresión, que predice valores numéricos continuos, por ejemplo, estimar el consumo de energía de un usuario en kilovatios-hora. Las métricas de evaluación incluyen precisión para clasificación y error promedio para regresión, pero el aprendizaje supervisado es sensible a problemas como el desbalance de

clases o el sobreajuste, donde el modelo memoriza los datos de entrenamiento (James et al., 2013).

1.2.2.2 Aprendizaje No Supervisado.

El aprendizaje no supervisado trabaja con datos no etiquetados, donde no se proporcionan categorías o valores de salida predefinidos. El objetivo es descubrir patrones, estructuras o relaciones inherentes en los datos, sin guía externa (Tan et al., 2019).

Se divide en clustering, que agrupa los datos en conjuntos basándose en la similitud de sus características, por ejemplo, agrupar usuarios según patrones de consumo similares, y reducción de dimensionalidad, que simplifica los datos al reducir el número de características, preservando la mayor cantidad posible de información relevante. Es útil para explorar datos desconocidos, pero requiere interpretar los resultados manualmente y es sensible a datos ruidosos (Aggarwal, 2015).

1.2.2.3 Aprendizaje por Refuerzo.

El aprendizaje por refuerzo entrena a un agente para tomar decisiones secuenciales en un entorno dinámico, aprendiendo a maximizar una recompensa acumulada a largo plazo. El agente interactúa con el entorno, observando su estado actual, tomando acciones y recibiendo recompensas o penalizaciones según los resultados. Una política define cómo el agente selecciona acciones, y el objetivo es ajustarla para elegir las acciones más beneficiosas, aunque requiere simulaciones precisas y es computacionalmente intensivo (Russell & Norvig, 2020).

1.2.2.4 Aprendizaje Semi-Supervisado.

El aprendizaje semi-supervisado combina datos etiquetados y no etiquetados, aprovechando la estructura de los datos no etiquetados para mejorar la precisión del modelo. Es útil cuando etiquetar datos es costoso, ya que asume que los datos no etiquetados contienen información valiosa sobre las relaciones entre las entradas. Por ejemplo, un modelo podría usar un pequeño conjunto de datos etiquetados para inicializar el aprendizaje y propagar etiquetas a datos no etiquetados basándose en su similitud (Goodfellow et al., 2016).

1.2.3 Principales Algoritmos de Machine Learning

1.2.3.1 Árboles de Decisión.

Los árboles de decisión son modelos que dividen los datos en regiones basadas en una serie de preguntas o condiciones sobre las características, organizadas en una estructura similar a un árbol. Cada nodo representa una pregunta, por ejemplo, si el valor de una característica es mayor que un cierto umbral, cada rama indica una respuesta posible, y las hojas proporcionan la predicción, que puede ser una categoría para clasificación o un valor numérico para regresión (Quinlan, 1986).

Las variantes incluyen árboles de clasificación, que predicen categorías discretas, árboles de regresión, que predicen valores numéricos, y CART, que utiliza divisiones binarias para ambos tipos de tareas. Estas variantes adaptan el modelo a diferentes problemas, facilitando su implementación (Alpaydin, 2020).

El entrenamiento construye el árbol dividiendo los datos repetidamente, eligiendo la característica y el umbral que mejor separen los datos en grupos homogéneos. La evaluación mide el rendimiento usando métricas como el porcentaje de predicciones correctas para clasificación o la diferencia promedio para regresión. Técnicas como la poda evitan el sobreajuste, asegurando generalización (James et al., 2013).

Los desafíos incluyen el sobreajuste, ya que los árboles profundos memorizan los datos, la sensibilidad a pequeños cambios, que altera la estructura del árbol, la dificultad para capturar relaciones no lineales, y el sesgo hacia características con muchos valores, lo que requiere ajustes en el algoritmo (Hastie et al., 2009).

1.2.3.2 Random Forest.

Random Forest combina múltiples árboles de decisión para mejorar la precisión y la estabilidad de las predicciones. Cada árbol se construye con un subconjunto aleatorio de los datos y considera un subconjunto aleatorio de características en cada división, reduciendo la similitud entre los árboles. Las predicciones se combinan seleccionando la categoría más común para clasificación o el promedio para regresión (Breiman, 2001).

Las variantes incluyen bagging estándar, que usa subconjuntos de datos con reemplazo, Extremely Randomized Trees, que eligen umbrales al azar, y Random Forest ponderado, que prioriza árboles más precisos. Estas variantes optimizan el rendimiento en diferentes contextos (Alpaydin, 2020).

El entrenamiento crea árboles independientes, ajustando parámetros como el número de árboles, la profundidad y el número de características por división. La evaluación usa métricas como precisión para clasificación o error promedio para regresión, a menudo con datos fuera de bolsa para estimar el rendimiento sin un conjunto de prueba adicional (James et al., 2013).

Los desafíos incluyen el alto costo computacional, la falta de interpretabilidad, el rendimiento limitado en datos con características irrelevantes, y el uso intensivo de memoria, lo que puede limitar su aplicación en sistemas con recursos restringidos (Hastie et al., 2009).

1.2.3.3 Máquinas de Soporte Vectorial (SVM).

Las máquinas de soporte vectorial buscan la mejor manera de separar datos en categorías para clasificación o predecir valores numéricos para regresión. En clasificación, el modelo identifica un límite que separa las categorías con la mayor distancia posible entre los puntos más cercanos al límite, maximizando la claridad de la separación. Para datos no separables linealmente, transforma los datos a un espacio donde la separación es más sencilla, usando funciones llamadas kernels (Cortes & Vapnik, 1995).

Las variantes incluyen C-SVM, que controla el equilibrio entre un límite amplio y la tolerancia a errores, v-SVM, que ajusta los vectores de soporte, SVR, para regresión, y One-Class SVM, para detectar anomalías. Estas variantes adaptan el modelo a diferentes tareas (Bishop, 2006).

El entrenamiento ajusta el modelo para encontrar el límite óptimo, considerando la separación y la tolerancia a errores, resolviendo un problema de optimización. La evaluación mide el rendimiento con métricas como precisión para clasificación o error promedio para regresión, usando datos separados para asegurar generalización (James et al., 2013).

Los desafíos incluyen la escalabilidad limitada, el ajuste complejo de parámetros, la necesidad de normalización para evitar distorsiones, y la dificultad para interpretar los límites generados, especialmente con transformaciones no lineales (Hastie et al., 2009).

1.2.3.4 Regresión Lineal y Logística.

La regresión lineal predice valores numéricos asumiendo que las características tienen una relación directa y constante con el valor a predecir. Por ejemplo, el modelo podría asumir que el consumo de energía aumenta proporcionalmente con el número de electrodomésticos, ajustando una línea recta para representar esta relación (Hastie et al., 2009).

La regresión logística predice la probabilidad de que un dato pertenezca a una categoría, por ejemplo, determinar si un usuario es de alto o bajo consumo, transformando los datos para

producir una probabilidad entre 0 y 1. Las variantes incluyen regresión lineal simple, múltiple, regularizada (Ridge, Lasso) y logística multinomial (James et al., 2013).

El entrenamiento ajusta el modelo para minimizar las diferencias entre predicciones y valores reales en regresión lineal, o maximizar la probabilidad de clasificaciones correctas en regresión logística. La evaluación usa métricas como error promedio para regresión lineal o precisión para regresión logística (Bishop, 2006).

Los desafíos incluyen la suposición de relaciones directas, la sensibilidad a datos extremos, los problemas con variables correlacionadas, y la necesidad de preprocesamiento para datos categóricos o con escalas diferentes (Alpaydin, 2020).

1.2.3.5 Naive Bayes.

Naive Bayes es un método de clasificación que predice categorías basándose en la probabilidad de que un dato pertenezga a cada categoría, utilizando un principio estadístico que calcula cómo las características contribuyen a esa probabilidad. Asume que las características son independientes entre sí, simplificando los cálculos y permitiendo clasificar rápidamente (Domingos & Pazzani, 1997).

Las variantes incluyen Gaussian Naive Bayes, para datos continuos, Multinomial Naive Bayes, para datos discretos, y Bernoulli Naive Bayes, para datos binarios. Estas variantes adaptan el modelo a diferentes tipos de datos (Bishop, 2006).

El entrenamiento estima las probabilidades de las categorías y las contribuciones de cada característica, un proceso eficiente incluso con datasets pequeños. La evaluación mide el rendimiento con métricas como precisión o capacidad de distinguir categorías, comparando con datos no utilizados en el entrenamiento (James et al., 2013).

Los desafíos incluyen la suposición de independencia, que rara vez se cumple, los datos desbalanceados, que generan predicciones sesgadas, la limitación a clasificación, y la necesidad de suposiciones específicas para datos continuos (Alpaydin, 2020).

1.2.3.6 K-Nearest Neighbors (KNN).

K-Nearest Neighbors clasifica o predice basándose en los datos más similares, o "vecinos", en el conjunto de entrenamiento. Para un nuevo dato, identifica los puntos más cercanos según una medida de similitud, como la distancia directa. En clasificación, selecciona la categoría más común entre los vecinos; en regresión, calcula el promedio de sus valores (Cover & Hart, 1967).

Las variantes incluyen KNN ponderado, que da más importancia a vecinos cercanos, y KNN adaptativo, que ajusta el número de vecinos según la densidad de los datos. Estas variantes mejoran la sensibilidad a patrones locales (Alpaydin, 2020).

No hay entrenamiento explícito, ya que el modelo almacena los datos de entrenamiento. Durante la predicción, calcula la similitud con todos los puntos almacenados, seleccionando los más cercanos. La evaluación mide el rendimiento con métricas como precisión o error promedio, ajustando el número de vecinos (James et al., 2013).

Los desafíos incluyen el alto costo computacional, la sensibilidad a escalas, el impacto del ruido, y la selección del número de vecinos, que afecta el rendimiento si es demasiado pequeño o grande (Hastie et al., 2009).

1.2.3.7 K-means.

K-means organiza los datos en un número predefinido de grupos, buscando que los puntos dentro de cada grupo sean lo más similares posible entre sí y lo más diferentes posible de los puntos en otros grupos. El algoritmo asigna cada punto al grupo cuyo centro

representativo, o centroide, esté más cerca, ajustando los centroides iterativamente hasta que los puntos dejan de cambiar de grupo (MacQueen, 1967).

Las variantes incluyen K-means++, que mejora la selección inicial de centroides, Mini-batch K-means, que usa subconjuntos para mayor rapidez, y K-medoids, que selecciona puntos reales como centros para mayor robustez frente a datos atípicos (Jain & Dubes, 1988).

La calidad de los grupos se mide con métricas que evalúan la cohesión dentro de cada grupo y la separación entre grupos. La elección del número de grupos se basa en métodos que comparan la mejora en la cohesión al aumentar los grupos (Tan et al., 2019).

Los desafíos incluyen la selección del número de grupos, la asunción de grupos uniformes, la dependencia de la inicialización, y la sensibilidad a datos atípicos, que distorsionan los centroides (Aggarwal, 2015).

1.2.3.8 Clustering Jerárquico.

El clustering jerárquico organiza los datos en una estructura de árbol, agrupando puntos en niveles progresivos de similitud. Puede ser aglomerativo, empezando con cada punto como un grupo individual y combinándolos gradualmente, o divisivo, comenzando con todos los puntos en un solo grupo y dividiéndolos en grupos más pequeños (Jain & Dubes, 1988).

Las variantes incluyen enlace simple, que combina grupos basándose en los puntos más cercanos, enlace completo, que usa los puntos más alejados, enlace promedio, que considera la distancia promedio, y el método de Ward, que minimiza la dispersión interna (Tan et al., 2019).

La calidad de los grupos se evalúa con métricas de cohesión y separación, o analizando la estructura del árbol generado, conocido como dendrograma, para identificar niveles de agrupamiento significativos (Aggarwal, 2015).

Los desafíos incluyen el alto costo computacional, la sensibilidad a ruido, las decisiones irreversibles en las combinaciones, y la elección del criterio de combinación, que produce resultados distintos (Jain & Dubes, 1988).

1.2.3.9 DBSCAN.

DBSCAN agrupa puntos basándose en la densidad de datos, identificando regiones donde los puntos están muy juntos y marcando como ruido los puntos aislados. Define un grupo como un conjunto de puntos donde cada uno tiene suficientes vecinos cercanos dentro de una distancia específica, conectando puntos densos para formar grupos de formas variadas (Ester et al., 1996).

Las variantes incluyen HDBSCAN, que maneja densidades variables, y OPTICS, que genera una estructura para identificar grupos a diferentes niveles de densidad, ofreciendo más flexibilidad (Jain & Dubes, 1988).

La calidad de los grupos se mide con métricas de cohesión y separación. DBSCAN no requiere especificar el número de grupos, pero necesita ajustar parámetros como la distancia máxima y el número mínimo de vecinos (Tan et al., 2019).

Los desafíos incluyen el ajuste de parámetros, la dificultad con densidades variables, el costo computacional, y el rendimiento en datos de alta dimensionalidad, donde las medidas de distancia pierden significado (Aggarwal, 2015).

1.2.3.10 Redes Neuronales.

Las redes neuronales son modelos inspirados en el cerebro humano, compuestos por unidades interconectadas, conocidas como neuronas, organizadas en capas. Cada neurona procesa información combinando las entradas recibidas, ajustándolas según su importancia, y

aplicando una transformación para producir una salida, permitiendo modelar relaciones complejas (Goodfellow et al., 2016).

Las variantes incluyen el Red Neuronal Multicapa, para datos tabulares, redes convolucionales (CNN), para datos espaciales, redes recurrentes (RNN), para datos secuenciales, LSTM/GRU, para relaciones a largo plazo, transformers, para secuencias largas, autoencoders, para reducir dimensionalidad, y GAN/VAE, para generar datos (Haykin, 2008).

El entrenamiento ajusta la importancia de las conexiones entre neuronas para minimizar las diferencias entre predicciones y valores reales, propagando los errores hacia atrás. La evaluación mide el rendimiento con métricas como precisión o error promedio, usando datos no utilizados en el entrenamiento (Nielsen, 2015).

Los desafíos incluyen la necesidad de grandes cantidades de datos y poder computacional, el ajuste complejo de parámetros, el riesgo de sobreajuste, y la falta de transparencia en las predicciones (Aggarwal, 2018).

1.2.3.11 Gradient Boosting.

Gradient Boosting combina múltiples árboles de decisión, construyéndolos uno tras otro, donde cada árbol corrige los errores de los anteriores. El modelo mejora progresivamente, enfocándose en los datos predichos incorrectamente, para producir predicciones más precisas (Friedman, 2001).

Las variantes incluyen XGBoost, que mejora velocidad y precisión, LightGBM, que optimiza el uso de memoria, y CatBoost, que maneja datos categóricos sin transformaciones previas, facilitando su uso en problemas con variables mixtas (Chen & Guestrin, 2016).

El entrenamiento construye árboles secuencialmente, ajustando parámetros como el número de árboles, su profundidad y la velocidad de aprendizaje. La evaluación mide el

rendimiento con métricas como precisión o error promedio, usando datos separados para verificar generalización (James et al., 2013).

Los desafíos incluyen el lento entrenamiento, el riesgo de sobreajuste, la necesidad de optimizar múltiples parámetros, y la interpretación limitada debido a la combinación de muchos árboles (Hastie et al., 2009).

Capitulo II: Metodología y Procedimiento

2.1 Normas de Calidad del Servicio de los Sistemas de Distribución

Normas de Calidad del Servicio reguladas por la Súper Intendencia General de Electricidad y Telecomunicaciones (SIGET), tienen por objeto regular los índices e indicadores de referencia para calificar la calidad con que las empresas distribuidoras de energía eléctrica, suministran los servicios de energía eléctrica a los usuarios de la Red de Distribución, tolerancias permisibles, métodos de control y compensaciones respecto de a la calidad del servicio técnico, calidad del producto técnico y calidad del producto comercial. Iniciando la etapa de régimen en 2005 para garantizar la adecuación de los sistemas de distribución a las normas establecidas.

La normativa utilizada, establece formatos estandarizados de información que permiten a la SIGET evaluar los índices de calidad del servicio y son presentados mensualmente por las empresas distribuidoras a la SIGET, en cumplimiento de los requerimientos normativos. Esto permite a las distribuidoras establecer procesos de análisis de información generalizados, con información relevante sobre los transformadores de distribución permitiendo crear un histórico fiable del comportamiento de la red eléctrica.

El estudio se enfocó en la zona central de El Salvador, con un histórico de datos completo desde el 2008 al 2024, para poder verificar las relaciones de las variables relevantes para el sistema y el comportamiento de las fallas de los transformadores de distribución.

2.2 Metodología de Aplicación

Representar un entorno de distribución eléctrica de forma precisa, presenta dificultades inherentes debido a la naturaleza aleatoria y no lineal de las variables de operación. Ante esta condición, la metodología desarrollada se basa en la integración de múltiples fuentes de información historial de fallas, niveles de cargabilidad, caracterización de la carga, estudios históricos de fallas y condiciones de operación con el fin de representar el comportamiento del sistema. Esta información se analiza mediante diversos métodos de aprendizaje automático, lo

que permite evaluar y comparar su desempeño en la identificación de comportamientos anómalos o potenciales fallas, que cada uno de esto métodos.

A continuación, se presentan etapas de la metodología aplicada, en el estudio realizado para la detección de interrupciones en los sistemas de distribución, enfocado en las fallas de transformadores de distribución, en la zona central de El Salvador.

2.2.1 Contexto operacional de los equipos

Los transformadores analizados corresponden a activos de distribución conectados a la red del operador a niveles de tensión de 13.2 kV y 23 kV, en la zona central de El Salvador. En esta región se presentan condiciones climáticas que incluyen alta humedad y presencia significativa de vegetación, lo cual favorece la ocurrencia de fallas por contactos accidentales con ramas. Adicionalmente, se debe considerar que El Salvador es uno de los países con mayor actividad eléctrica atmosférica en Centroamérica, lo que incrementa la probabilidad de daños por descargas atmosféricas, especialmente durante la temporada lluviosa (mayo a octubre).

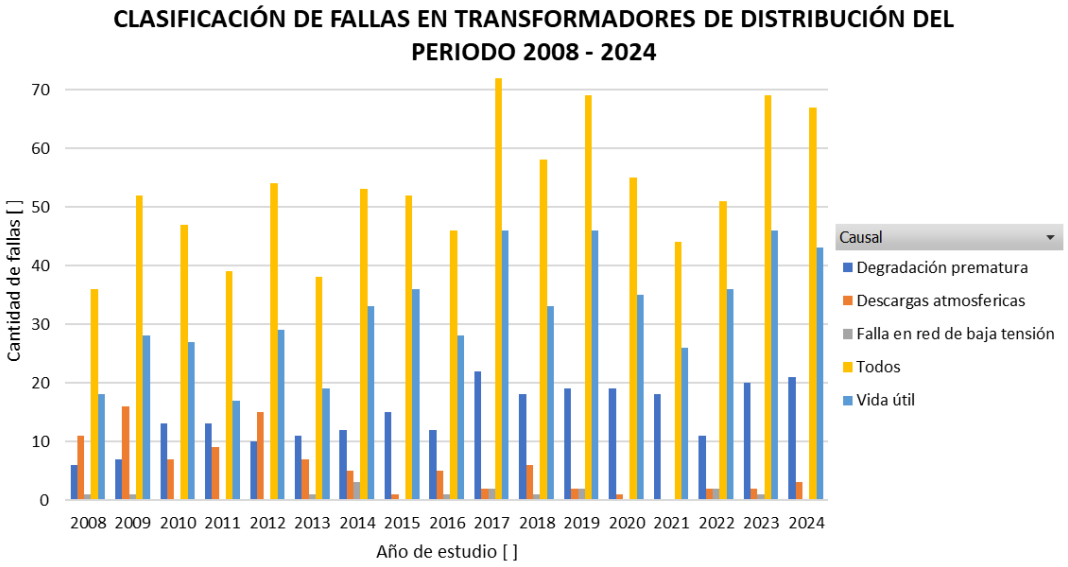
2.2.2 Identificación del conjunto de transformadores de distribución bajo estudio

El estudio se centró en transformadores de distribución conectados a niveles de tensión de hasta 23 kV, ubicados en la zona central de El Salvador, la cual concentra la mayor infraestructura eléctrica del país en términos de demanda energética y crecimiento poblacional. Esta región incluye tanto áreas urbanas densamente desarrolladas como localidades remotas de carácter rural, lo que permite un análisis representativo de diversas condiciones operativas. Se excluyeron del análisis los transformadores de propiedad privada, ya que su configuración, operación y mantenimiento son gestionados directamente por los clientes, bajo criterios distintos a los establecidos por la distribuidora. Esta última realiza únicamente mantenimiento correctivo sobre los equipos que presentan fallas, considerando en el presente estudio un universo total de

19,344 transformadores en operación dentro del área de análisis. Como parte de la evaluación, se identificaron las principales causas asociadas a la reducción de la vida útil y a la degradación prematura de estos equipos, tanto por factores técnicos como ambientales. A partir de esta base, se presenta a continuación un análisis detallado del histórico de fallas en transformadores de distribución entre los años 2008 y 2024.

Figura 1.

Clasificación de fallas en transformadores de distribución.



Los resultados obtenidos, del estudio realizado en los trasformadores de distribución, reflejan un incremento general en la cantidad de fallas, destacando las fallas asociadas a degradaciones prematuras de los transformadores y fallos por vida útil. En contraste, se aprecia una disminución en la cantidad de fallos por descargas atmosféricas y cortocircuitos presentes en la red de baja tensión.

Las principales causas de la degradación de componentes que causaron fallas reparables en los transformadores fueron desconexiones de las borneras BT y fugas de aceites, siendo necesaria una sustitución del transformador para realizar el posterior diagnóstico y reparación, acumulando compensaciones por energía no servida y costos de reparación.

2.3 Identificación de las variables relevantes para el sistema

2.3.1Tipo de servicio

Variable de clasificación binaria que indica el tipo de servicio entregado en el transformador, según su densidad de carga, definido en las “Normas de Calidad del Servicio de los Sistemas de Distribución”.

Área de densidad de carga: Es el área geográfica comprendida dentro de un cuadrado de un kilómetro por lado, de forma que para una empresa distribuidora las áreas de densidad de carga no se traslapen entre sí, debiendo contener en su conjunto a la totalidad de la red eléctrica y usuarios de la distribuidora.

Área de densidad de carga alta: Es aquella área de densidad de carga que contiene al menos mil habitantes o en donde la demanda de energía eléctrica de los usuarios es al menos 250 kilowatts, y que además se encuentre en una región que aglomere al menos 10 áreas contiguas que bajo los dos parámetros antes indicados puedan ser clasificadas como áreas de densidad de carga alta. Por otra parte, independientemente al resultado de la aplicación de los criterios antes señalados, en las metodologías de control de las presentes normas se podrán definir excepciones para que áreas adicionales también sean consideradas de alta densidad de carga.

Área de densidad de carga baja: Es aquella área de densidad de carga que no ha sido definida como un área de densidad de carga alta.

Así clasificado para densidad de carga alta como “U” y para densidad de carga baja “R”, en la tabla ‘DATOS_CENTROS’, remitida mensualmente por las empresas distribuidoras a la SIGET.

2.3.2 Potencia nominal

NumTrafo: Definido como “Número de Trafos en el Centro MT/BT” (Entero), indicando la cantidad de transformadores instalados en la estructura con el identificador reportado.

kVAinst: Definido como, “kVA instalado total en el Centro MT/BT”, reportando la totalidad de carga en kVA que puede entregar la instalación del centro de transformación.

Para el estudio se utiliza la potencia nominal como la relación de la potencia instalada y la cantidad de transformadores instalados, clasificando los comportamientos de los transformadores según su capacidad individual.

2.3.3 Tipo de clientes suministrados.

Clasificación de cada transformador según la tarifa de los clientes conectados y la agrupación de la demanda y nivel de tensión suministrado.

Tabla 1.

Clasificación de Usuarios por su tarifa de consumo.

Tarifa	Clasificación	Descripción
109	1	Tarifa simple para Usuarios conectados en baja tensión, pequeña demanda residencial con consumo menor o igual a 99 kWh.
110	2	Tarifa simple para Usuarios conectados en baja tensión, pequeña demanda residencial con consumo mayor que 99 kWh y menor o igual a 200 kWh.
111	3	Tarifa simple para Usuarios conectados en baja tensión, pequeña demanda residencial con consumo mayor de 200 kWh.
112	3	Tarifa simple para Usuarios conectados en baja tensión, pequeña demanda uso general
113	3	Tarifa de Alumbrado Público
121	4	Tarifa simple para Usuarios conectados en baja tensión, mediana demanda sin medición de potencia
122	4	Tarifa simple para Usuarios conectados en baja tensión, mediana demanda con medición de potencia
123	4	Tarifa simple para Usuarios conectados en baja tensión, mediana demanda con medidor horario

131	5	Tarifa simple para Usuarios conectados en baja tensión, grandes demandas con medidor horario
132	5	Tarifa simple para Usuarios conectados en baja tensión, grandes demandas con medidor electromecánico
211	6	Tarifa para Usuarios conectados en medianas demandas en MT, sin medidor de potencia
212	6	Tarifa para Usuarios conectados en medianas demandas en MT, con medidor de potencia
213	6	Tarifa para Usuarios conectados en medianas demandas en MT, con medidor horario
221	7	Tarifa para Usuarios conectados en grandes demandas en MT, con medidor horario
222	7	Tarifa para Usuarios conectados en grandes demandas en MT, con medidor electromecánico

Nota. Clasificación de usuarios propia, basado en la definición de tarifas de las "Normas de Calidad del Servicio de los Sistemas de Distribución".

2.3.4 Número de usuarios.

Cantidad de usuarios reportados anualmente en la tabla "Datos_Usuarios", tabla que posee la información de los clientes agrupados por transformador.

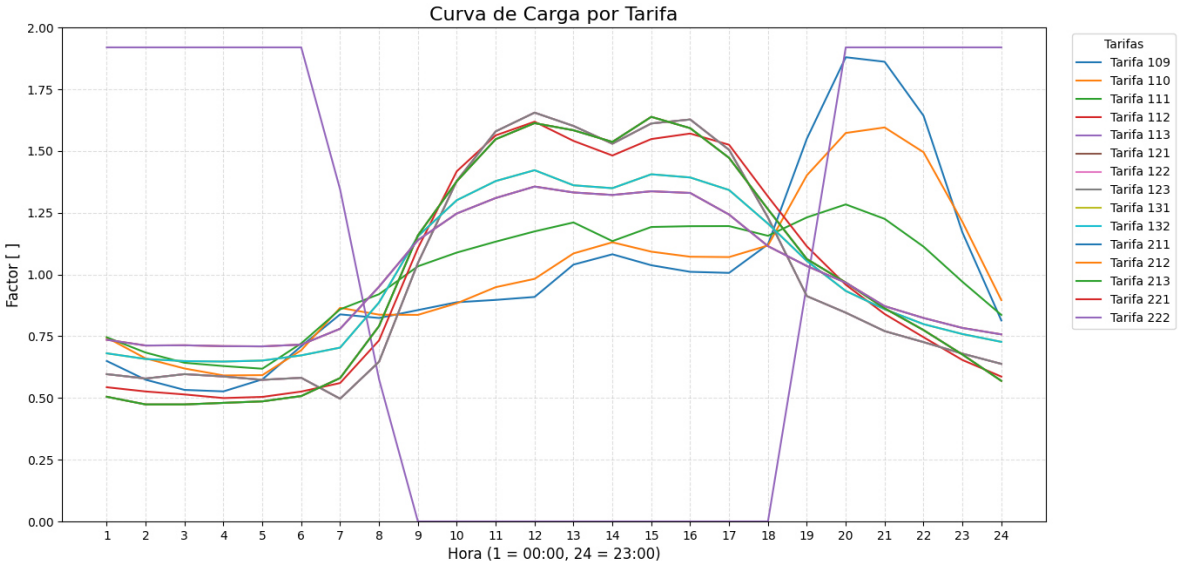
2.3.5 Cargabilidad máxima

Se llevó a cabo un análisis de la cargabilidad histórica, utilizando como referencia el valor máximo de demanda registrado por hora durante el mes en que se produjo la falla de los transformadores del estudio.

Para la estimación de la demanda, se consideró el consumo eléctrico segmentado por tipo de tarifa, el cual fue distribuido conforme a los perfiles de consumo establecidos en el estudio de demanda típica elaborado por la SIGET, actualizado de forma quinquenal y categorizado según la tarifa asignada a los usuarios.

Figura 2.

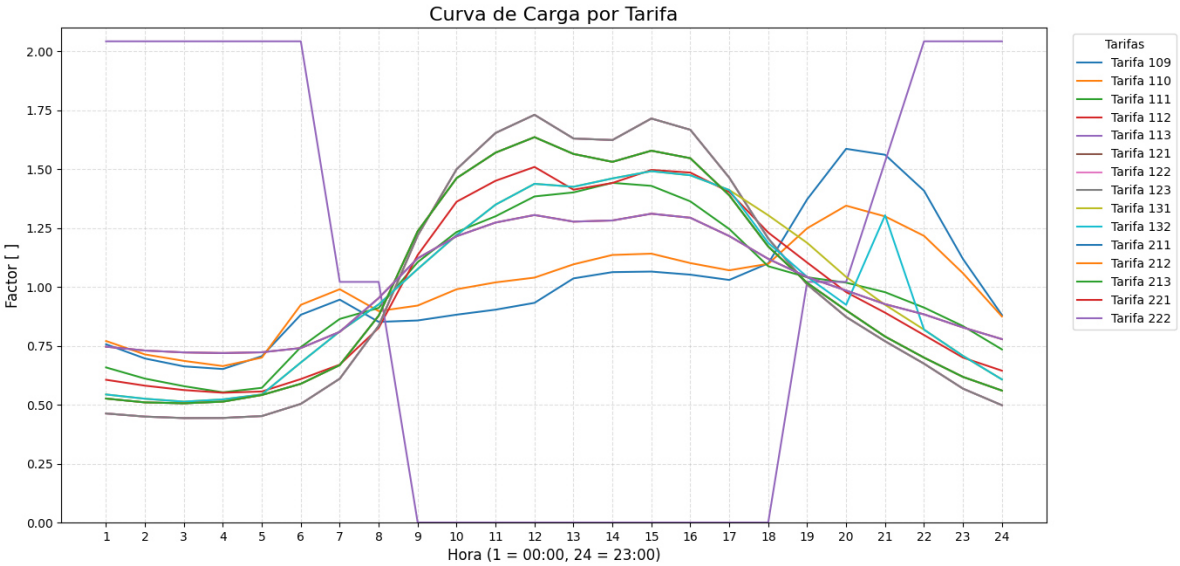
Factores de responsabilidad horaria, definidos por SIGET, periodo 2013-2018.



Nota. En la figura se muestran los factores de responsabilidad horaria para cada tarifa definida por la SIGET. Fuente: SIGET.

Figura 3.

Factores de responsabilidad horaria, definidos por SIGET, periodo 2019-2024



Nota. En la figura se muestran los factores de responsabilidad horaria para cada tarifa definida por la SIGET. Fuente: SIGET.

Como resultado del estudio de cargabilidad, se llevó a cabo un proceso de consulta individual por transformador, con el fin de evaluar su comportamiento histórico. Esta revisión permitió identificar variaciones en los consumos mensuales de los usuarios, así como redistribuciones de carga realizadas sobre los transformadores analizados. Para complementar este análisis, se implementó un script en Python que produjo representaciones gráficas a partir de los datos procesados, permitiendo una visualización estructurada de los resultados y optimizando los procesos de análisis e interpretación de la información recopilada.

Figura 4

Sección del código que posee la definición de datos a utilizar.

```
def Cargabilidad(Fecha, Empresa):
        #Estructura de datos
                mes = Fecha.month
                año = Fecha.year
                Datos_Usuario_1 = 'Datos_Usuario/' + Empresa + 'C' + str(año-2000) + '1' + '
                        _DATOS_USUARIOS_ACUMULADA.txt'
                Facturacion_1 = 'Facturacion/'+'/' + Empresa + '/' + Empresa + 'T' +  str(año) + str(mes).
                        zfill(2) + '_FACTURACION.txt'#XT201301_FACTURACION
                mes_L = mes
                if mes_L == 10:
                        mes_L = 'O'
                if mes_L == 11:
                        mes_L = 'N'
                if mes_L == 12:
                        mes_L = 'D'
                Datos_Centro_D = 'Datos_Centro/' + Empresa + 'T' +  str(año-2000) + str(mes_L) + '
                        _DATOS_CENTROS.txt'#XT172_DATOS_CENTROS
                Factores_Resp = 'Factores_Resp/Factores_'+ Empresa+ '_2013-2018.txt'
        #Seleccion de Factores de Responsabilidad
                Q1 = '2018-02-1'
                Q1 =datetime.strptime(Q1, '%Y-%m-%d')
                if(Fecha <= Q1):
                        Factores_Resp = 'Factores_Resp/Factores_' + Empresa + '_2013-2018.txt'
                else:
                        Factores_Resp = 'Factores_Resp/Factores_' + Empresa + '_2018-2024.txt'
        #Lecuturas de Archivos
                print(Facturacion_1)
                print(Datos_Usuario_1)
                print(Factores_Resp)
                print(Datos_Centro_D)
                print(Factores_ambientales)

                Facturacion = pd.read_csv(Facturacion_1,sep="|",low_memory=False)
                Facturacion =Facturacion.rename(columns={'IDUsuario': 'IDUSUARIO'})
                Facturacion =Facturacion.rename(columns={'Energia': 'ENERGIA'})
                F_Ambientales = pd.read_csv(Factores_ambientales,sep="  ",low_memory=False)
                Datos_Usuario = pd.read_csv(Datos_Usuario_1,sep="|",low_memory=False)
                Factores = pd.read_csv(Factores_Resp,sep="|",low_memory=False)
                Datos_Centro_2 = pd.read_csv(str(Datos_Centro_D),sep="|",low_memory=False, encoding='latin-1
                        ')
```

Figura 5

Sección del código que realiza la relación de tablas utilizadas y cálculo de la cargabilidad.

```
#Enlace con DU_FACTURACION
        Datos_Usuario = Datos_Usuario.rename(columns={'IDUsuario': 'IDUSUARIO'})
        Consumo = pd.merge(Datos_Usuario, Facturacion, on='IDUSUARIO', how='inner')
#Agrupacion de energía por tarifa
        Consumo['Tarifa'] = Consumo['Tarifa'].apply(lambda x: (int(x)) if pd.notnull(x) else x)
        conteo = Consumo.groupby(['CenMTBT', 'Tarifa']).agg({'ENERGIA': 'sum',  'IDUSUARIO': 'count'}).
                rename(columns={'IDUSUARIO': 'Conteo_IDUSUARIO'}).reset_index()
#Enlace con Factores
        conteo = pd.merge(conteo, Factores, on='Tarifa', how='left')
        a=['1','2','3','4','5','6','7','8','9','10','11','12','13','14','15','16','17','18','19','20','
                21','22','23','24']
#Distribución de energía consumida por hora
        conteo[a] = conteo[a].apply(lambda x: (x * conteo['ENERGIA']/24/30))
        a1=['1','2','3','4','5','6','7','8','9','10','11','12','13','14','15','16','17','18','19','20','
                21','22','23','24','Conteo_IDUSUARIO', 'ENERGIA']
        conteo2 = conteo[['CenMTBT', 'Tarifa','Conteo_IDUSUARIO']]
        conteo = conteo.groupby(['CenMTBT'])[a1].sum().reset_index()
#Agraga el conteo de usuarios por tarifa usando una trasposicon (pivot), maneniendo la distribución de
energía
        conteo_tarifas = conteo2.pivot(index='CenMTBT', columns='Tarifa', values='Conteo_IDUSUARIO').
                fillna(0).reset_index()
        conteo_tarifas.columns.name = None  # Limpiar el nombre del índice de columnas
        conteo = pd.merge(conteo, conteo_tarifas, on='CenMTBT', how='left')
#Enlace con Datos_Centro
        try:
                Datos_Centro_2 = Datos_Centro_2[['CenMTBT', 'NumTrafo', 'kVAinst', 'Dirección']]
        except:
                Datos_Centro_2 = Datos_Centro_2[['CenMTBT', 'NumTrafo', 'kVAinst', 'Direccion']]
        conteo = pd.merge(conteo, Datos_Centro_2, on='CenMTBT', how='left')
#Normalización de la carga
        conteo[a] = conteo[a].apply(lambda x: round(x / conteo['kVAinst'], 4))
        conteo['MAX'] = conteo[a].max(axis=1)
        conteo['MED'] = conteo[a].mean(axis=1)
#Añadiendo factores ambientales diarios
        F_Ambientales
        F_Ambientales['date2'] = F_Ambientales['date'].apply(lambda x: datetime.strptime(x, '%Y-%m-%d'))
        F_Ambientales = F_Ambientales[(F_Ambientales['date2'].dt.year == año) & (F_Ambientales['date2'].
                dt.month == mes)]
# Calcular el promedio excluyendo valores faltantes de las temperaturas promedio del mes
        promedio_tavg = F_Ambientales['tavg'].mean()
        promedio_tmax = F_Ambientales['tavg'].max()
        conteo['T_Promedio'] = promedio_tavg
        conteo['T_Max'] = promedio_tmax
```

Este código se ejecutó de forma mensual, para los periodos enero 2013 a diciembre 2023, y conglomero en una base de datos con la estructura de presentada en la tabla 2.

Tabla 2.

Conglomerado de cargabilidad histórica.

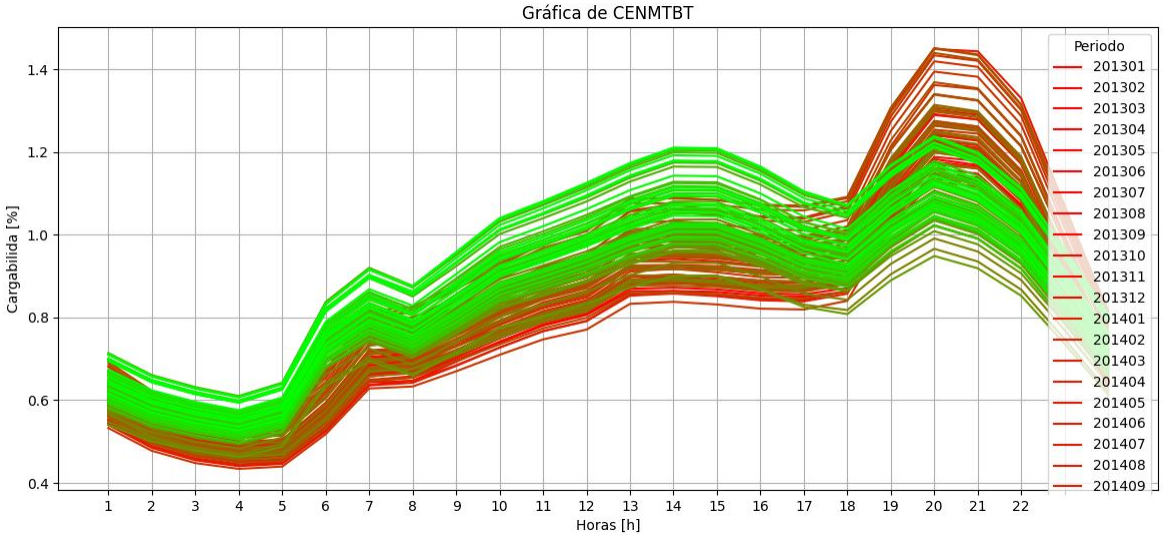
Campo	Descripción
CenMTBT	Placa del transformador
1 - 24	Cargabilidad por cada hora del día.
Conteo_IDUSUARIO	Cantidad de Usuarios conectados al transformador
ENERGIA	Cantidad de energía consumida en el transformador
109 - 221	Conteo de Usuarios por Tarifa
NumTrafo	Numero de transformadores en el centro MT/BT
kVAinst	Potencia Nominal instalada por transformador
Dirección	Dirección del centro de transformación
MAX	Cargabilidad máxima en el transformador

MED	Cargabilidad promedio en el transformador
T_Promedio	Temperatura promedio del transformador
T_Max	Temperatura máxima del transformador
Periodo	Corresponde al mes y año de referencia del registro

Nota. Elaboración propia, de base de datos histórica de factores ambientales, cargabilidad y usuarios asociados al transformador.

Figura 6

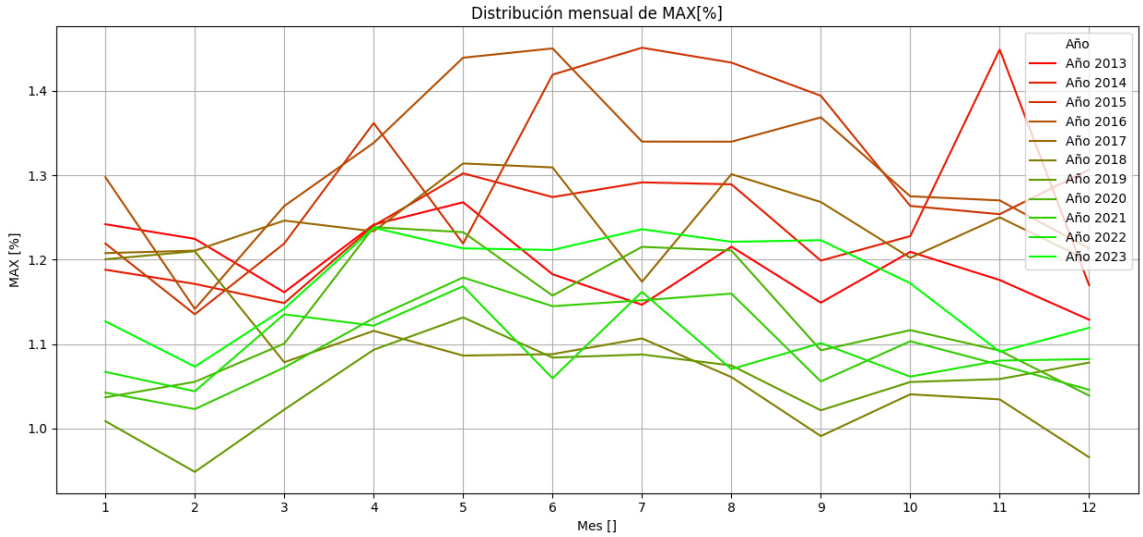
Curva histórica de carga horaria, mensual.



Nota. Ejemplo de curva de carga horaria mensual histórica, para el periodo 2013-2023. Figura de elaboración propia construida a partir del registro de demanda horaria mensual.

Figura 7

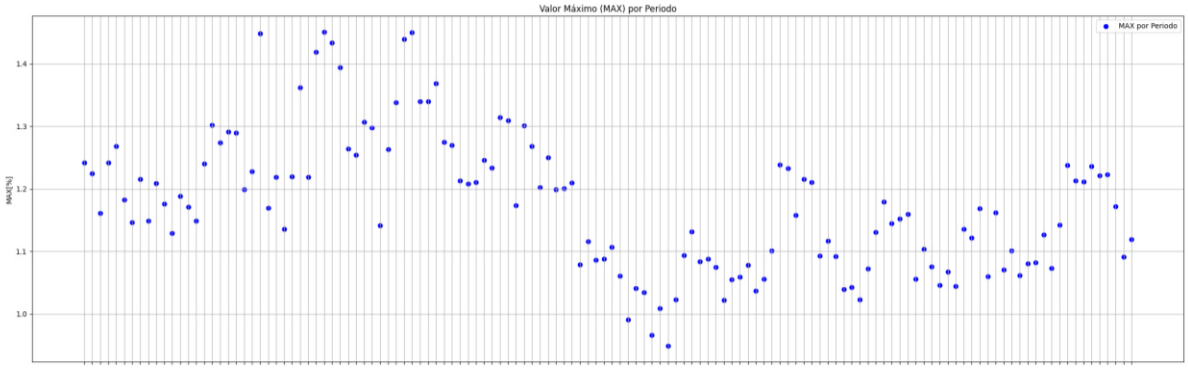
Máximo de cargabilidad mensual por año.



Nota. Ejemplo de curva de carga horaria anual histórica, para el periodo 2013-2023. Figura de elaboración propia construida a partir del registro de demanda horaria anual.

Figura 8

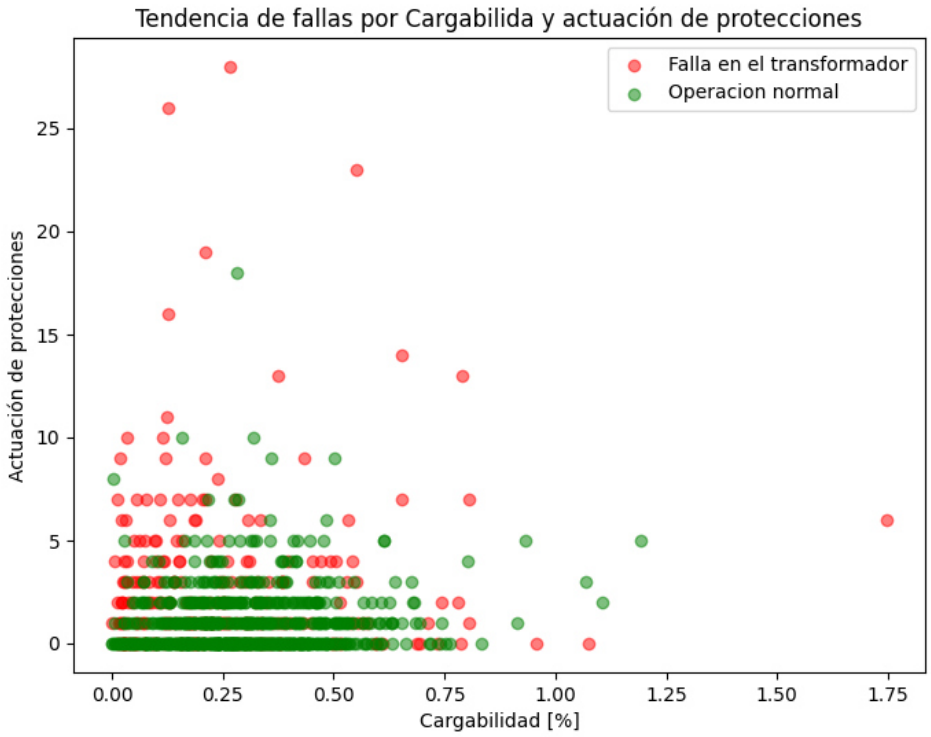
Histórico de máximos de cargabilidad por periodo mensual.



Nota. Se aprecia en la figura los puntos de mayor cargabilidad registrados por mes, elaboración propia a partir de la base generada de cargabilidad.

Figura 9

Gráfico que muestra la tendencia de fallas por Cargabilidad y actuación de protecciones



Nota. Figura que muestra las fallas de los transformadores y la activación de sus protecciones, creación propia.

2.3.6 Activaciones históricas de protecciones de los transformadores

Este conjunto de datos contiene registros de las activaciones de los dispositivos de protección asociados a cada transformador. Para cada evento, se registra el identificador único del transformador y la fecha exacta en la que se produjo la activación. Esta información es fundamental para analizar patrones históricos de fallas o sobrecargas, así como para evaluar el desempeño y confiabilidad del sistema de protección.

Tabla 3.

Formato de base de datos histórico de activación de protecciones

Campo	Descripción
CenMTBT	Placa del transformador afectado
Fecha	Fecha y hora en la que se presentó la interrupción debido a la actuación de protecciones del transformador

Nota. Base de datos organizada a partir de la información remitida por la empresa distribuidora y utilizada para el presente estudio.

2.4 Correlación de variables

Para mejora la interpretación de las relaciones entre las variables operativas y la ocurrencia de fallas, se realizó un análisis de correlación sobre una muestra compuesta por los 181 transformadores históricos que presentaron fallas relacionadas a degradaciones de los transformadore y un grupo de transformadores en operación normal seleccionados con criterios de representatividad por tipo de cliente y nivel de urbanidad. Este procedimiento permitió identificar asociaciones lineales entre variables técnicas, de carga y de entorno, así como descartar aquellas cuyo aporte explicativo resultó limitado o inconsistente.

La Tabla 4 presenta la matriz de correlación de variables, segmentada según la clasificación de urbanidad (urbano/rural) y el tipo de cliente predominante. Este enfoque segmentado busca capturar diferencias estructurales en el comportamiento operativo de los transformadores, ya que las condiciones de carga, densidad de usuarios y frecuencia de disparos difieren sustancialmente entre entornos urbanos y rurales.

Tabla 4.

Matriz de correlación de variables por clasificación de transformadores según su urbanidad y consumo de clientes.

Urbanidad	Tipo Cliente	Activación de protecciones	Conteo Usuarios	Energía	Max	KVA Nominal	Vida Útil	Registros
0	1	0.427	0.243	0.128	0.335	-0.116	0.058	468
0	2	0.467	0.817	0.556	0.697	-0.133	0.019	12
0	3	0.191	0.239	0.480	0.670	-0.066	-0.131	28

1	1	0.486	0.145	0.228	0.300	-0.152	-0.084	302
1	2	0.461	0.161	0.167	0.369	-0.315	-0.150	64
1	3	0.437	0.134	0.278	0.365	-0.198	-0.051	135

Nota. Matriz de correlación de elaboración propia. La muestra incluye transformadores en operación normal y fallidos, con foco en clientes residenciales.

Para mejorar la interpretabilidad de los resultados, los coeficientes de correlación se consolidaron mediante un promedio ponderado considerando el número de registros de cada grupo. Este índice ponderado representa la relación media entre las variables más relevantes para el análisis de falla y operación.

Tabla 5.

Índice ponderado de correlación de variables.

Activación de protecciones	Conteo Usuarios	Energía	MAX	KVA Nominal	Vida útil
0.442	0.201	0.195	0.344	-0.150	-0.018

Nota. índice ponderado correlación de variables con muestra de datos de transformadores fallidos históricos y en operación normal.

Los resultados obtenidos muestran que las variables “Activación de protecciones”, “Cargabilidad máxima (MAX)” y “Conteo de usuarios” presentan los coeficientes de correlación más elevados con el estado operativo de los transformadores. Esto evidencia que los eventos de activación de protecciones, cargabilidad y la densidad de carga tienen una influencia significativa en la degradación del aislamiento y, en consecuencia, en la probabilidad de ocurrencia de fallas.

En contraste, la variable “KVA nominal” exhibe una correlación negativa moderada, lo que sugiere que los transformadores de mayor capacidad tienden a operar con un margen de carga más holgado y, por tanto, muestran menor susceptibilidad a fallas bajo condiciones similares de operación y entorno.

Por su parte, la variable “Vida útil” registró un coeficiente de correlación bajo (-0.018), resultado que se asocia a la presencia de valores atípicos y registros concentrados en dos fechas específicas de instalación. Esta condición refleja inconsistencias en la trazabilidad de los datos históricos y debido a que no fue posible recuperar la esta información, se excluirá de los datos de entrenamiento. Sin embargo, desde un punto de vista operativo, se reconoce que la vida útil constituye un factor determinante en la estimación de la probabilidad de falla, quedando abierta a mejoras en el entrenamiento de los modelos y exactitud de los modelos de machine learning.

En conjunto, el análisis de correlación permitió validar la coherencia estadística de las variables seleccionadas y priorizar aquellas con mayor influencia sobre el comportamiento operativo de los transformadores. Estos resultados constituyen una base cuantitativa sólida para la fase de entrenamiento de los modelos supervisados y para la posterior construcción del Índice Compuesto de Riesgo (ICR), instrumento central en la clasificación de criticidad y planificación de mantenimiento predictivo.

2.5 Base de datos de entrenamiento

Para el desarrollo del modelo predictivo, se construyó un conjunto de datos estructurado a partir de la información remitida por las distribuidoras a la SIGET y los historiales de fallas correspondientes al período 2013–2023.

El enfoque principal se centró en aquellos transformadores que presentaron fallas asociadas a degradación prematura o al fin de su vida útil.

La muestra de transformadores fallados se constituyó por un total de 181 unidades, consideradas como casos positivos y clasificados con “1” en la variable de salida binaria correspondiente a degradaciones prematuras y finalización de la vida útil de los transformadores. Estos casos fueron cuidadosamente seleccionados tras una depuración y validación de los reportes de mantenimiento, historiales de activación de protecciones y análisis técnico post-falla,

garantizando que efectivamente correspondieran a eventos atribuibles al deterioro progresivo del equipo y no a eventos externos o aleatorios.

Con el fin de construir un modelo supervisado equilibrado y robusto, se incorporaron adicionalmente registros de transformadores que no presentaron fallas en el mismo período analizado. Estos transformadores operaron bajo condiciones normales o controladas y no registraron incidentes de desconexión definitiva ni eventos relevantes asociados a pérdida de funcionalidad. Estos casos fueron catalogados como negativos o con un valor de “0” en la variable objetivo.

La inclusión de estos dos tipos de registros permitió estructurar una variable de salida binaria, codificada como 1 para transformadores fallados y 0 para transformadores sin falla, lo que viabilizó la aplicación de técnicas de clasificación supervisada. Esta configuración del conjunto de datos permite no solo entrenar modelos de aprendizaje automático orientados a la predicción de fallas, sino también analizar patrones diferenciadores entre transformadores en condiciones críticas y aquellos en condiciones operativas estables.

Tabla 6.

Formato de base de entrenamiento

Campo	Descripción	Tipo de dato
CenMTBT	Placa del transformador	Texto
Fecha	Fecha y hora en la que se presentó la interrupción debido a la actuación de protecciones del transformador	Fecha
Salida	Valor lógico de falla del transformador	Binario
Activación de protecciones	Conteo histórico de activaciones de protecciones	Entero
Conteo Usuarios	Cantidad de Usuarios conectados al transformador	Entero
Energía	Cantidad de energía consumida en el transformador	Decimal
Max	Cargabilidad máxima en el transformador	Decimal
KVANominal	Potencia instalada por transformador	Decimal
Tipo Cliente	Clasificación de clientes suministrados en el transformador	Entero

Urbanidad	Área de densidad de carga	Binario

Nota. Base creada a partir la información proporcionada por una distribuidora.

2.6 Caracterización de los datos.

La información presentada en la tabla corresponde a un conjunto de registros asociados a elementos identificados con un código único (por ejemplo, T00001, T00002, etc.), los cuales contienen variables tanto numéricas como categóricas que describen aspectos técnicos, operativos y de ubicación de cada equipo. En este sentido, la base de datos refleja la naturaleza mixta de la información, integrando variables cuantitativas, cualitativas y temporales que permiten realizar un análisis integral del comportamiento de los elementos observados.

Tabla 7.

Caracterización de datos utilizados en los modelos de predicción.

Elemento Repos	Fecha	Salida	Activación de protecciones	Conteo Usuarios	Energía	Max	KVA Nominal	Tipo Cliente	Urbanidad
T00001	23/6/2022	0	2	7	417	0.0612	15	1	0
T00002	12/5/2020	0	2	6	1181	0.0592	38	1	1
T00003	9/12/2018	1	7	18	740	0.1087	15	1	0
T00004	19/7/2014	0	1	41	3861	0.1765	50	1	1
T00005	9/5/2017	0	0	1	118	0.0174	15	2	0
T00006	24/6/2016	0	7	8	336	0.0343	25	1	0
T00007	17/6/2013	0	0	34	3239	0.2794	25	1	1
T00008	9/11/2019	0	0	2	34586.72	0.5072	75	3	0
T00009	30/5/2017	0	0	55	5279	0.2484	50	1	1
T00010	9/8/2014	0	4	5	267	0.0465	15	1	0

Nota. Datos extraídos directamente de las bases compartidas por la distribuidora.

Dentro de las variables cuantitativas se encuentran aquellas que expresan medidas o conteos, tales como Activación de protecciones, Conteo Usuarios Energía, Max, KVANominal. Estas variables representan, respectivamente, los conteos de la cantidad de disparos de protecciones y reposiciones realizadas, el conteo de usuarios vinculados al transformador, la

energía registrada durante el mes, el valor máximo de cargabilidad registrado en el mes de la falla y la capacidad nominal en kilovoltamperios (kVA). Los valores observados presentan una amplia variabilidad, con rangos que van desde valores mínimos muy bajos hasta máximos considerablemente altos. Esta dispersión indica diferencias significativas en el uso, capacidad y desgaste operativo de los equipos, evidenciando posibles casos atípicos o de operación intensiva, como ocurre con el registro que alcanza un conteo de energía con un valor superior a 34,000.

Por otra parte, las variables categóricas y temporales Tipo Cliente, Urbanidad y Fecha permiten clasificar los elementos según su naturaleza técnica, su asociación con un cliente, el entorno geográfico en el que operan (urbano o rural) y la fecha de inicio o instalación. Estas variables aportan información relevante para el análisis contextual, ya que permiten relacionar la ubicación o el tipo de cliente con el comportamiento operativo.

En términos generales, la caracterización de los datos muestra un conjunto heterogéneo, con variaciones amplias en las magnitudes registradas y una diversidad en la clasificación de los equipos. Este tipo de información resulta esencial para los procesos de análisis y modelado, ya que la presencia de datos con diferentes escalas, unidades y dispersiones puede influir en la interpretación de los resultados y en la identificación de patrones de comportamiento. En consecuencia, la comprensión de la naturaleza de los datos constituye un paso fundamental previo a cualquier análisis estadístico o predictivo, asegurando una interpretación coherente y técnicamente sustentada del fenómeno estudiado.

2.7 Análisis Estadístico Descriptivo de las Variables del Modelo

El análisis estadístico descriptivo constituye una etapa fundamental en la investigación cuantitativa, ya que permite caracterizar las propiedades inherentes de los datos, identificar patrones preliminares y evaluar la adecuación de las variables para su posterior modelado. En el

contexto de esta tesis, se emplea un enfoque descriptivo para examinar las variables involucradas en un modelo de aprendizaje automático, orientado a predecir una variable de salida binaria ("Salida"). Las medidas calculadas incluyen indicadores de tendencia central (media y mediana), dispersión (desviación estándar, mínimo y máximo), forma de la distribución (asimetría y curtosis) y completitud de los datos (valores faltantes).

Tabla 8.

Estadísticos descriptivos de las variables numéricas.

Variable	Media	Mediana	Desviación estándar	Mínimo	Máximo	Asimetría
Salida	0.18	0	0.39	0	1	1.65
Activación de protecciones	1.1	0	2.37	0	28	5.2
Conteo Usuarios	32.06	23	30.75	1	169	1.26
Energía	4177.6	2481	5303.05	1	55633.8	3.25
Max	0.2	0.16	0.17	0	1.75	1.84
KVANominal	35.16	25	22.29	10	225	1.9
Tipo Cliente	1.5	1	1.02	1	7	2.57
Urbanidad	0.5	0	0.5	0	1	0.01
Fecha	2017.93	2018	3.09	2013	2023	-0.02

Nota. La tabla presenta las medidas de tendencia central y dispersión de las variables numéricas registradas en el conjunto de datos. Los valores de asimetría permiten identificar distribuciones no simétricas, lo cual es relevante para el ajuste de los modelos predictivos.

La Tabla 8 resume las estadísticas descriptivas de las variables numéricas, revelando una diversidad en sus distribuciones que influye directamente en la selección de técnicas de preprocesamiento y modelado. La variable dependiente "Salida", de naturaleza binaria (0 o 1), presenta una media de 0.18, lo que indica que aproximadamente el 18% de las observaciones corresponden a la categoría positiva. Su mediana de 0.00 y desviación estándar de 0.39 sugieren una distribución sesgada hacia valores bajos, con un rango limitado entre 0.00 y 1.00. La asimetría positiva de 1.65 confirma esta tendencia, implicando un desbalance de clases que

podría sesgar el rendimiento de algoritmos de clasificación sensibles a la prevalencia de categorías, como la regresión logística o las redes neuronales.

Tabla 9.

Estadísticos descriptivos de las variables utilizadas en el modelo de aprendizaje automático.

Variable	Media	Mediana	Desviación estándar	Mínimo	Máximo	Curtosis
Activación de protecciones	1.1	0	2.37	0	28	41.46
Conteo Usuarios	32.06	23	30.75	1	169	1.28
Energía	4177.6	2481	5303.05	1	55633.8	18.05
Max	0.2	0.16	0.17	0	1.75	8.2
KVA Nominal	35.16	25	22.29	10	225	8.15
Tipo Cliente	1.5	1	1.02	1	7	8.11
Urbanidad	0.5	0	0.5	0	1	-2

Nota. Se presentan las variables empleadas como características de entrada en los modelos de regresión logística, árboles de decisión, bosques aleatorios, redes neuronales y DBSCAN.

La Tabla 9 complementa el análisis al incorporar la curtosis, que mide el grado de apuntamiento o achatamiento de la distribución relativa a una normal. "Activación de protecciones" presenta una curtosis extremadamente alta (41.46), indicando una distribución con picos pronunciados y colas pesadas, lo que podría generar inestabilidad en modelos paramétricos. De manera análoga, "Energía" (curtosis: 18.05) y "Max" (8.20) muestran concentraciones elevadas en torno a la media, sugiriendo la presencia de subclústeres en los datos que DBSCAN podría explotar para detección de anomalías.

Variables como "KVANominal" (8.15) y "Tipo Cliente" (8.11) también exhiben curtosis positiva, interpretada como evidencia de heterogeneidad categórica subyacente, especialmente

en "Tipo Cliente" (rango: 1-7), que podría beneficiarse de codificación one-hot para evitar suposiciones de ordinalidad en modelos de árboles de decisión. "Urbanidad" presenta una curtosis negativa (-2.00), consistente con su naturaleza binaria y distribución uniforme.

Tabla 10*.*

Distribución de frecuencias de la variable de salida (Salida).

Categoría	Frecuencia	Porcentaje (%)
0	845	81.8
1	188	18.2

Nota. La variable de salida representa las categorías de respuesta del modelo. La distribución porcentual permite identificar posibles desbalances de clases que pueden afectar el desempeño predictivo.

La Tabla 10 detalla la distribución frecuencial de "Salida", con 845 observaciones (81.80%) en la categoría 0 y 188 (18.20%) en 1. Este desbalance pronunciado es un hallazgo crítico, ya que puede inducir a modelos a sobre predecir la clase mayoritaria, resultando en métricas infladas como la precisión, pero deficientes para la clase. En un contexto académico, esto justifica estrategias de remuestreo o ponderación de clases en algoritmos como regresión logística y redes neuronales, para mejorar la generalización.

2.8 Herramientas de machine learning utilizadas en el modelo de estudio

Con base en la información histórica de fallas registrada entre los años 2008 y 2024, se desarrollaron diferentes modelos de aprendizaje automático con el objetivo de representar de forma precisa el comportamiento del sistema de distribución eléctrica en la zona central de El Salvador. El propósito principal de estos modelos es identificar patrones en los datos que permitan predecir la probabilidad de falla de los transformadores de distribución y, de esta forma, facilitar decisiones de mantenimiento preventivo y reemplazo estratégico.

Dado que la naturaleza del sistema es estocástica y no permite establecer relaciones determinísticas simples entre variables técnicas, ambientales y operativas, se optó por una aproximación multivariada basada en aprendizaje automático.

Para abordar la problemática de la predicción de fallas en transformadores de distribución, se seleccionaron y aplicaron distintos algoritmos de aprendizaje automático que permiten capturar tanto relaciones lineales como no lineales en los datos. La regresión logística fue utilizada como un modelo base debido a su capacidad para estimar la probabilidad de falla a partir de variables explicativas, proporcionando interpretabilidad directa de los coeficientes asociados.

Por su parte, los árboles de decisión permiten representar reglas de decisión en forma jerárquica, facilitando la comprensión de los factores críticos que conducen a una falla. Para mejorar la capacidad de generalización y mitigar el sobreajuste inherente a los árboles de decisión, se implementó un bosque aleatorio (Random Forest), que combina múltiples árboles mediante técnicas de agregación, mejorando la precisión en escenarios con alta variabilidad en los datos. Asimismo, se exploraron redes neuronales artificiales, dada su habilidad para modelar relaciones complejas y no lineales entre múltiples variables de entrada, lo cual resulta adecuado en contextos donde los patrones de falla no siguen estructuras fácilmente separables. Complementariamente, se utilizó el algoritmo de clustering DBSCAN con el objetivo de identificar estructuras anómalas o agrupamientos atípicos dentro del conjunto de datos, particularmente útil para detectar subconjuntos de transformadores con comportamientos operativos que difieren significativamente del resto, sin requerir una variable de salida.

Todos estos modelos fueron implementados y evaluados utilizando la biblioteca Scikit-learn en Python, mediante el desarrollo de un código modular que permite ejecutar los algoritmos, realizar la validación cruzada, generar métricas de desempeño, y visualizar los resultados de forma sistemática.

Figura 10

Sección del código Python que evalúa los modelos de aprendizaje automático para el desarrollo de la tesis.

```python
modelos = [
    ("Regresión Lógistica", LogisticRegression()),
    ("Arbol de decisión", DecisionTreeClassifier()),
    ("Bosque Aleatorio", RandomForestClassifier(n_estimators=100)),
    ("Redes Neuronales", MLPClassifier(hidden_layer_sizes=(100, 100), activation='relu', solver='adam', max_iter
        =1000, random_state=seed))
]
for nombre, modelo in modelos:
    modelo.fit(X_Entrena, y_entrena)
    puntaje = modelo.score(X_Prueba, y_prueba)
    print(f'{nombre}: {puntaje:.7f}')
    predicciones = modelo.predict(X_Prueba)
    total = len(y_prueba) #Total
    aciertos = np.sum(predicciones == y_prueba)# Aciertos totales
    pct_aciertos = aciertos / total * 100
    pct_fallas_bien = np.mean(predicciones[y_prueba == 1] == 1) * 100 # Fallas bien clasificadas (1 -> 1)
    pct_normales_bien = np.mean(predicciones[y_prueba == 0] == 0) * 100# Normales bien clasificadas (0 -> 0)
    print(f"\n🔷 {nombre}")
    print(f"✔️ Acierto en resultados positivos: {pct_fallas_bien:.2f}%")
    print(f"✔️ Acierto en resultados negativos: {pct_normales_bien:.2f}%")
    print(f"🎯 Aciertos totales (global): {pct_aciertos:.2f}%")
    df_resultado = X_Prueba.copy()
    df_resultado["Prediccion"] = predicciones
    df_resultado["Real"] = y_prueba.values
    df_resultado.to_csv("Evaluaciones Globales/Entrenamiento_Salidas_" + nombre + ".csv", index=False, sep=";")
    if nombre == "Arbol de decisión":
        plt.figure(figsize=(20, 10))
        plot_tree(modelo,
                  feature_names=X_Entrena.columns,)  # nombres de las variables
        plt.title("Árbol de Decisión")
        plt.show()
    probabilidades = modelo.predict_proba(X_Prueba)
    probabilidades
    prob_Falla = probabilidades[:,1]
    df_Eval = pd.DataFrame({
    "Evaluacion (Prob_Falla)": prob_Falla,
    "Salida Real (y_prueba)": y_prueba.tolist()})
    print(df_Eval)
    print(X_Prueba)
```

2.8.1 Configuración de regresión logística

Se implementó en el modelo de regresión logística (LogisticRegression), el solucionador 'lbfgs' como algoritmo de optimización, utilizado para ajustar los coeficientes en modelos, el cual es apropiado para problemas de clasificación de tamaño moderado, y se aplicó una regularización L2 que ayuda a evitar el sobre ajuste durante el entrenamiento. La clasificación se llevó a cabo en un contexto binario, considerando el umbral de decisión predeterminado del 50%. Esta configuración ofrece un balance adecuado entre capacidad predictiva y simplicidad

interpretativa, siendo especialmente eficaz cuando existe una relación aproximadamente lineal entre las variables independientes y la clase objetivo.

2.8.2 Configuración de árboles de decisión

Se implementaron dos enfoques basados en árboles de decisión para abordar el problema de clasificación. El primero consistió en un modelo de árbol de decisión individual, utilizando “DecisionTreeClassifier”. El árbol se construyó utilizando el criterio de impureza de Gini, sin restricciones en la profundidad máxima ni en el número mínimo de muestras por división, lo que permitió un ajuste completo a los datos de entrenamiento.

El segundo enfoque fue un modelo de bosque aleatorio (RandomForestClassifier) configurado con 100 árboles de decisión (n_estimators=100). Este método combina múltiples árboles entrenados de manera independiente sobre subconjuntos aleatorios de los datos y las características, lo cual mejora la capacidad de generalización del modelo y reduce significativamente el riesgo de sobreajuste típico de los árboles individuales.

Ambos modelos se incluyeron para comparar su rendimiento, destacando las ventajas del enfoque individual en términos de interpretabilidad, y del enfoque ensemble en cuanto a robustez y estabilidad predictiva.

2.8.3 Configuración del DBSCAN

Se implementó un modelo de agrupamiento utilizando el algoritmo DBSCAN, configurado con un radio de vecindad (eps) de 0.09 y una cantidad mínima de puntos (min_samples) de 20. Esta configuración fue seleccionada para identificar patrones de densidad en los datos normalizados, permitiendo clasificar observaciones en distintos clústeres sin necesidad de definir previamente su número.

Una característica clave de DBSCAN es su capacidad para detectar puntos atípicos, a los cuales asigna la etiqueta -1. En este análisis, dichos puntos fueron interpretados como datos que se desvían del comportamiento esperado o de la tendencia dominante. Estos registros atípicos fueron posteriormente comparados con los historiales de fallas reales, observándose una coincidencia relevante: muchos de ellos correspondían a transformadores que habían experimentado fallas operativas. Esto sugiere que la salida de estos puntos del patrón general podría estar relacionada con condiciones anómalas previas a una falla, lo cual valida el enfoque del modelo como herramienta para la detección temprana de riesgos.

2.8.4 Configuración de Redes Neuronales

Se utilizó un clasificador de redes neuronales con una arquitectura de dos capas ocultas de entrada salida y la configuración por defecto de 100 neuronas siendo el valor por defecto. La función de activación seleccionada fue ReLU (Rectified Linear Unit), que es el valor por defecto en este tipo de modelos debido a su buen desempeño en tareas no lineales. Para la optimización de los pesos durante el entrenamiento, se empleó el algoritmo Adam, un optimizador eficiente basado en descenso de gradiente estocástico. Además, se estableció un máximo de 1000 iteraciones para asegurar la convergencia del modelo.

Capitulo III: Resultados.

3.1 Resultados Obtenidos

Con el objetivo de evaluar la efectividad de los modelos de aprendizaje automático aplicados a la predicción de fallas en transformadores de distribución, se realizó una validación cruzada de cada algoritmo implementado. Para medir el rendimiento de los modelos, se utilizaron tres métricas clave: precisión general, exactitud en la clasificación de casos positivos (transformadores que presentaron fallas) y exactitud en la clasificación de casos negativos (transformadores que no presentaron fallas). Estas métricas se derivaron de la matriz de confusión y permiten una evaluación integral del comportamiento de cada modelo.

Tabla 11.

Resumen de los resultados obtenidos durante el entrenamiento de los modelos aplicados.

Modelo	Precisión general	Exactitud de casos positivos	Exactitud de casos negativos
Regresión Logística	83.65%	15%	100%
Árbol de decisión	81.73%	65%	85.71%
Bosque Aleatorio	84.62%	50%	92.86%
Clúster DBSCAN	66.21%	27.07%	86.36%
Redes Neuronales	83.65%	40%	94.05%

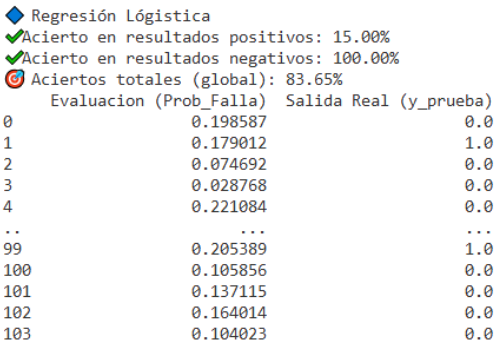
Nota. Datos obtenidos a partir del modelo diseñado en Python de autoría propia.

3.2 Desempeño de los Modelos

Regresión Logística. El modelo logró una **precisión general del 83.65%**, con una exactitud del **15.00% en la detección de fallas** y **100.00% en la clasificación correcta de transformadores sin fallas**. Este rendimiento revela una fuerte tendencia a clasificar las evaluaciones como negativas, aunque su sensibilidad al detectar fallas es limitada. Su utilidad radica en la interpretabilidad de los coeficientes, aunque puede presentar restricciones cuando se enfrentan patrones no lineales.

Figura 11

Resultado visto desde la consola de Visual Studio para el modelo de Regresión Logística.



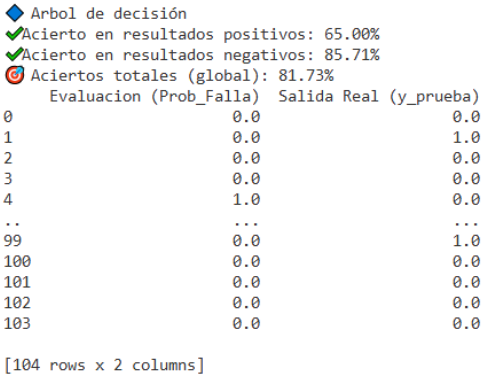
```
🔷 Regresión Lógistica
✔️Acierto en resultados positivos: 15.00%
✔️Acierto en resultados negativos: 100.00%
🎯 Aciertos totales (global): 83.65%
     Evaluacion (Prob_Falla)  Salida Real (y_prueba)
0                   0.198587                     0.0
1                   0.179012                     1.0
2                   0.074692                     0.0
3                   0.028768                     0.0
4                   0.221084                     0.0
..                       ...                     ...
99                  0.205389                     1.0
100                 0.105856                     0.0
101                 0.137115                     0.0
102                 0.164014                     0.0
103                 0.104023                     0.0
```

Nota. Resultados de la evaluación del modelo de Regresión Logística. Generado mediante un script en Python de autoría propia para visualizar el desempeño del modelo.

Árbol de Decisión. Alcanzó una precisión general del 81.73% una exactitud del 65% en los casos positivos, y 85.71% en los casos negativos. Su estructura jerárquica facilita la comprensión de los factores críticos que conducen a una falla, aunque su capacidad para generalizar puede verse afectada por la variabilidad de los datos.

Figura 12

Resultado visto desde la consola de Visual Studio para el modelo de Árbol de Decisión.



```
🔷 Arbol de decisión
✔️Acierto en resultados positivos: 65.00%
✔️Acierto en resultados negativos: 85.71%
🎯 Aciertos totales (global): 81.73%
     Evaluacion (Prob_Falla)  Salida Real (y_prueba)
0                       0.0                     0.0
1                       0.0                     1.0
2                       0.0                     0.0
3                       0.0                     0.0
4                       1.0                     0.0
..                      ...                     ...
99                      0.0                     1.0
100                     0.0                     0.0
101                     0.0                     0.0
102                     0.0                     0.0
103                     0.0                     0.0

[104 rows x 2 columns]
```

Nota. Resultados de la evaluación del modelo de Árbol de decisión. Generado mediante un script en Python de autoría propia para visualizar el desempeño del modelo.

Bosque Aleatorio. Mostró una precisión general del 84.62%, con una exactitud del 50% para fallas y 92.86% para operaciones normales. Este modelo mejora la robustez frente al sobreajuste mediante la combinación de múltiples árboles y proporciona mejores niveles de generalización, aunque aún presenta limitaciones en la sensibilidad hacia los casos positivos.

Figura 13

Resultado visto desde la consola de Visual Studio para el modelo de Bosque Aleatorio.

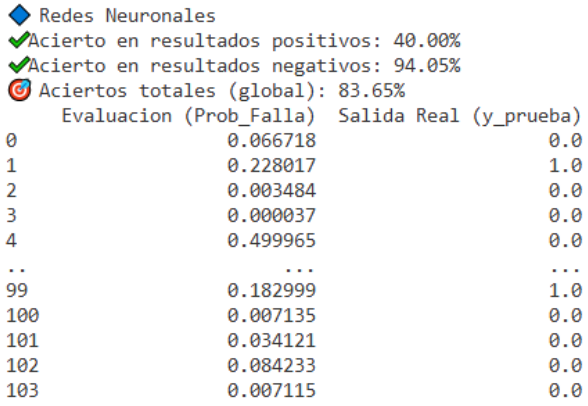
```
Bosque Aleatorio: 0.8461538

🔷 Bosque Aleatorio
✔️Acierto en resultados positivos: 50.00%
✔️Acierto en resultados negativos: 92.86%
🎯 Aciertos totales (global): 84.62%
     Evaluacion (Prob_Falla)  Salida Real (y_prueba)
0                       0.17                     0.0
1                       0.12                     1.0
2                       0.02                     0.0
3                       0.00                     0.0
4                       0.80                     0.0
..                       ...                     ...
99                      0.04                     1.0
100                     0.01                     0.0
101                     0.04                     0.0
102                     0.00                     0.0
103                     0.00                     0.0
```

Nota. Resultados de la evaluación del modelo de Bosque Aleatorio. Generado mediante un script en Python de autoría propia para visualizar el desempeño del modelo.

Redes Neuronales. Presento una precisión general de 83.65%, con una exactitud del 40% en la predicción de fallas y 95.05% para los casos negativos. Aunque su desempeño en la detección de fallas no fue el más alto, el modelo demostró buena capacidad de clasificación general y destaca por su habilidad para capturar relaciones no lineales complejas.

Figura 14

Resultado visto desde la consola de Visual Studio para el modelo de Redes Neuronales.

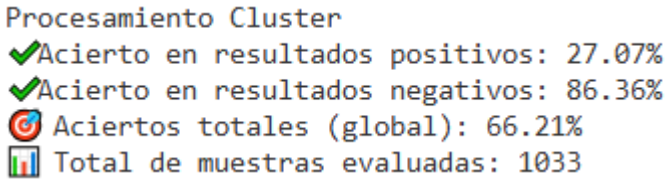


Nota. Resultados de la evaluación del modelo de Redes Neuronales. Generado mediante un script en Python de autoría propia para visualizar el desempeño del modelo.

Clustering DBSCAN. Si bien no es un modelo supervisado, se evaluó su capacidad para identificar patrones anómalos. Obtuvo una precisión general del 66.21%, con una exactitud del 27.07% en la detección de fallas y 86.36% en casos negativos. Su utilidad principal radica en la segmentación no lineal del espacio de datos, especialmente útil para detectar agrupamientos atípicos sin requerir etiquetas predefinidas.

Figura 15

Resultado visto desde la consola de Visual Studio para el modelo de Clúster DBSCAN.



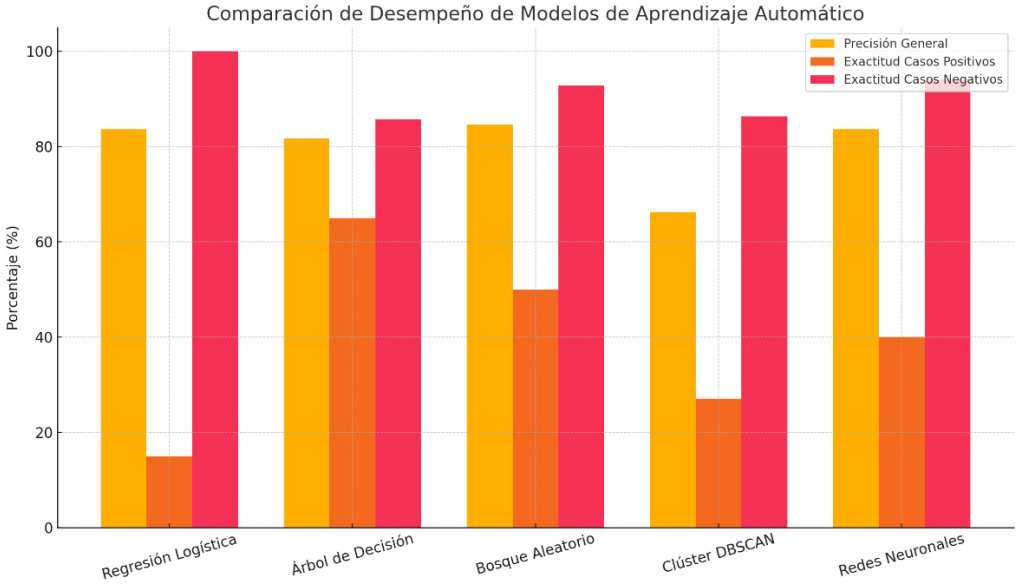
Nota. Resultados de la evaluación del modelo de Clustering DBSCAN. Generado mediante un script en Python de autoría propia para visualizar el desempeño del modelo.

3.2.1 Comparación Gráfica

La Figura 1 muestra una visualización comparativa del desempeño de los cinco modelos evaluados, lo que permite identificar fácilmente cuáles presentan mejor balance entre sensibilidad y especificidad.

Figura 16

Comparación de desempeño de modelos de aprendizaje automático aplicados a la predicción de fallas en transformadores de distribución.



Nota. Datos obtenidos a partir del script de Python que evalúa los modelos evaluados en este proyecto de creación propia.

3.2.2 Síntesis de Resultados

Los resultados reflejan una tendencia generalizada de los modelos a presentar mejor desempeño en la clasificación de transformadores sin fallas que en la detección de fallas reales, lo cual es un fenómeno común en escenarios de clases desbalanceadas. Las redes neuronales y el bosque aleatorio se destacan como las mejores opciones por su capacidad de adaptación a patrones no lineales, aunque es necesario considerar técnicas adicionales de balanceo o reentrenamiento para mejorar la sensibilidad hacia los casos positivos. Por otro lado, la regresión logística y el árbol de decisión ofrecen interpretabilidad directa, lo que las hace útiles en análisis explicativos.

3.3 Propuesta de Estrategia de Mantenimiento Basada en Índice Compuesto de Riesgo

1.Construcción del Índice Compuesto

Para priorizar la gestión de los transformadores de distribución se desarrolló un **índice compuesto de riesgo** que integra tres factores clave:

1. **Probabilidad de falla**: se refiere a la estimación estadística de la posibilidad de que el transformador sufra una avería. Se construyó a partir del historial de fallas menores, condiciones operativas.
2. **Cargabilidad**: corresponde al grado de utilización del transformador respecto a su capacidad nominal. Los equipos que operan de forma sostenida por encima del 80–90% y de picos máximos de 120% de su capacidad presentan mayor deterioro térmico y reducción de vida útil, lo cual incrementa significativamente el riesgo operativo.
3. **Costo de interrupción y reparación**: considera el impacto económico asociado a la indisponibilidad del transformador, incluyendo las pérdidas por interrupción del suministro eléctrico y los costos directos de reparación o reposición. Se ponderó con menor peso que los factores técnicos, pero se mantuvo dentro del modelo para reflejar la criticidad económica de la falla.

El índice final se calculó como una combinación lineal ponderada de los tres factores:

IRC = α Probabilidad + β Consecuencia + γ Condición de carga

α = 0.4

β = 0.2

γ = 0.4

2. Clasificación del Estado de los Transformadores

Con base en el índice compuesto, los 18,147 transformadores analizados fueron clasificados en cuatro estados:

Tabla 12.

Clasificación de estado de transformadores según su índice compuesto.

Estado	Descripción	Cantidad	Porcentaje
0 – Bueno	Bajo riesgo, condiciones normales de operación.	16,776	92.45%
1 – Regular	Riesgo moderado, necesidad de seguimiento.	1,251	6.89%
2 – Malo	Alto riesgo, degradación evidente.	92	0.51%
3 – Crítico	Riesgo critico de falla, con degradado preexistente existente.	28	0.15%

3. Estrategia de Mantenimiento Escalable

Dada la magnitud de la cantidad de transformadores evaluados, se propone una estrategia escalonada que permita optimizar recursos y priorizar los equipos con mayor impacto en la continuidad y calidad del servicio.

Estado 0 – Bueno

- Acciones: Mantenimiento preventivo básico (inspección visual anual, limpieza de aisladores, revisión de bornes).
- Pruebas técnicas: Ensayos dieléctricos por muestreo estadístico (5–10% anual).

- Gestión de carga: Monitoreo continuo mediante SCADA o medidores inteligentes, con alertas >80% de capacidad.
- Objetivo: Mantener la confiabilidad con un costo mínimo.

Estado 1 – Regular

- Acciones: Mantenimiento preventivo reforzado (ensayos de aceite, cromatografía de gases disueltos, termografía).
- Inspección: Revisión semestral en campo (fugas, ruidos, temperatura).
- Gestión de carga: Redistribución de carga en caso de sobreutilización.
- Objetivo: Evitar la transición hacia estados de mayor criticidad.

Estado 2 – Malo

- Acciones: Intervención correctiva programada (evaluación dieléctrica completa, descargas parciales, resistencia de aislamiento).
- Decisión económica: Reemplazo si el costo de reparación supera el 50% del costo de un transformador nuevo.
- Medida temporal: Transferencia parcial de carga mientras se ejecuta la intervención.
- Objetivo: Reducir el riesgo de falla inesperada y planificar la sustitución progresiva.

Estado 3 – Crítico

- Acciones: Intervención inmediata para prevenir falla en servicio.
- Medidas urgentes: Sustitución programada en el corto plazo (<12 meses), transferencia de carga preventiva.
- Documentación: Registro detallado de condiciones al momento de retiro, para retroalimentar el modelo predictivo.
- Objetivo: Evitar interrupciones masivas y asegurar continuidad en zonas críticas.

Discusión.

Las variables utilizadas en el modelo permitieron caracterizar aspectos fundamentales del comportamiento de los transformadores, destacando especialmente la cargabilidad máxima histórica, el tipo de cliente suministrado y el número de usuarios conectados, las cuales ofrecieron una representación significativa de la demanda y condiciones operativas a las que estuvo expuesto cada equipo.

No fue posible incorporar el análisis de la **vida útil de los transformadores** como variable explicativa, debido a la **falta de una base sólida sobre las fechas de instalación**. Esta limitación redujo la capacidad del modelo para evaluar el impacto del envejecimiento en las fallas, dejando abierta la posibilidad de mejorar el enfoque en estudios futuros con una base de datos más completa.

El modelo de redes neuronales presentó el mejor desempeño general entre los algoritmos evaluados, alcanzando una precisión global del 83.65% y una alta exactitud en la detección de casos negativos 94.05%, indicando una gran capacidad de predicción de los transformadores que no presentarán fallas, los cuales predominan en el universo de estudio.

El bosque aleatorio también mostró resultados consistentes, con una precisión general del 81.73% y un buen equilibrio entre la detección de casos positivos y negativos, posicionándose

como una opción robusta frente a problemas de sobreajuste comunes en modelos basados en árboles individuales.

Aunque la regresión logística obtuvo una alta exactitud en los casos negativos (100%), su desempeño en la identificación de fallas fue más limitado (15%), lo que reduce su efectividad en escenarios donde la prioridad es la detección temprana de eventos críticos.

A pesar de los resultados alcanzados, no se obtuvo un modelo que logre una clasificación completamente satisfactoria de los eventos de falla, especialmente considerando la sensibilidad requerida en contextos operativos. En este sentido, se deja abierto el camino para futuras investigaciones que incluyan la aplicación de modelos de aprendizaje automático más avanzados y especializados, tales como técnicas de aprendizaje profundo, algoritmos de clasificación no lineales más complejos, o enfoques híbridos que combinen conocimiento experto con técnicas supervisadas y no supervisadas.

La naturaleza del desbalance entre clases (casos fallidos frente a no fallidos) representó un desafío importante durante el entrenamiento de los modelos, especialmente en algoritmos sensibles al equilibrio de clases como la regresión logística. Aunque se aplicaron técnicas para mitigar este efecto, la baja proporción de transformadores con fallas en la muestra utilizada correspondiente a una región geográfica específica limitó la capacidad del modelo para generalizar con precisión en la detección de eventos críticos. En futuras investigaciones, se recomienda ampliar el análisis a una zona más extensa del sistema de distribución, incorporando una mayor cantidad de registros de fallas que permita fortalecer la representación de la clase minoritaria y mejorar la robustez de los modelos.

La interpretación práctica de los resultados del modelo y su vinculación con procesos de mantenimiento predictivo constituye una línea de aplicación clave para mejorar la gestión operativa de transformadores. Si bien las redes neuronales demostraron un alto rendimiento en la clasificación, su naturaleza de "caja negra" puede dificultar la interpretación directa de sus decisiones. Por ello, se propone complementar su uso con un análisis detallado de variables

clave, como la cargabilidad máxima histórica, la urbanidad del entorno y la tipología de clientes atendidos, que han mostrado ser consistentes en la caracterización del riesgo. Esta integración permitiría identificar con mayor claridad los factores relevantes que influyen en la ocurrencia de fallas, facilitando una aplicación más efectiva y fundamentada de estrategias de mantenimiento predictivo en campo. Esta línea representa una oportunidad valiosa para profundizar en estudios futuros.

Bibliografía.

Aggarwal, C. C. (2015). *Data Mining: The Textbook*. Springer.

Aggarwal, C. C. (2018). *Neural Networks and Deep Learning: A Textbook*. Springer.

Alpaydin, E. (2020). *Introduction to Machine Learning* (4th ed.). MIT Press.

Bergen, A. R., & Vittal, V. (1999). *Power Systems Analysis* (2nd ed.). Prentice Hall.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5-32.

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785-794.

Cho, K., van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using RNN encoder-decoder. *arXiv preprint arXiv:1406.1078*.

Comaniciu, D., & Meer, P. (2002). Mean shift: A robust approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *24*(5), 603-619.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273-297.

Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, *13*(1), 21-27.

Domingos, P., & Pazzani, M. (1997). On the optimality of the simple Bayesian classifier. *Machine Learning*, *29*(2), 103-130.

Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters. *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, *96*(34), 226-231.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, *29*(5), 1189-1232.

Glover, J. D., Sarma, M. S., & Overbye, T. J. (2016). *Power Systems Analysis and Design* (6th ed.). Cengage Learning.

Gonen, T. (2015). *Electric Power Distribution Engineering* (3rd ed.). CRC Press.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial nets. *Advances in Neural Information Processing Systems*, *27*.

Grainger, J. J., & Stevenson, W. D. (1994). *Power System Analysis*. McGraw-Hill.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning* (2nd ed.). Springer.

Haykin, S. (2008). *Neural Networks and Learning Machines* (3rd ed.). Pearson.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735-1780.

IEEE Standards Association. (2018). *IEEE Guide for Electric Power Distribution Reliability Indices* (No. Std 1366-2012). IEEE.

International Electrotechnical Commission. (2017). *IEC 61850: Communication Networks and Systems*. IEC.

Jain, A. K., & Dubes, R. C. (1988). *Algorithms for Clustering Data*. Prentice Hall.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning*. Springer.

Kersting, W. H. (2017). *Distribution System Modeling and Analysis* (4th ed.). CRC Press.

Kingma, D. P., & Welling, M. (2013). Auto-encoding variational Bayes. *arXiv preprint arXiv:1312.6114*.

Kundur, P. (1994). *Power System Stability and Control*. McGraw-Hill.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436-444.

MacQueen, J. (1967). Some methods for classification and analysis. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, *1*(14), 281-297.

Momoh, J. A. (2008). *Electric Power Distribution, Automation, Protection, and Control*. CRC Press.

Murphy, K. P. (2022). *Probabilistic Machine Learning: An Introduction*. MIT Press.

Nielsen, M. A. (2015). *Neural Networks and Deep Learning*. Determination Press.

Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning*, *1*(1), 81-106.

Russell, S., & Norvig, P. (2020). *Artificial Intelligence: A Modern Approach* (4th ed.). Pearson.

Saadat, H. (2010). *Power System Analysis* (3rd ed.). McGraw-Hill.

Short, T. A. (2014). *Electric Power Distribution Handbook* (2nd ed.). CRC Press.

Tan, P.-N., Steinbach, M., & Kumar, V. (2019). *Introduction to Data Mining* (2nd ed.). Pearson.

Vaswani, A., Shazeer, N., Parmar, N., Uszoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *30*.

Weedy, B. M., Cory, B. J., Jenkins, N., Ekanayake, J. B., & Strbac, G. (2012). *Electric Power Systems* (5th ed.). Wiley.

Witten, I. H., Frank, E., & Hall, M. A. (2016). *Data Mining: Practical Machine Learning Tools and Techniques* (4th ed.). Morgan Kaufmann.