



TRABAJO FINAL DE GRADO

Presentado para acceder al título de

Licenciatura en Ciencia de Datos

**Diagnóstico causal y priorización de activos en sistemas de
distribución eléctrica de media tensión : un enfoque basado en
caracterización de firmas eléctricas y aprendizaje no
supervisado**

Autor:

Villarruel, Lucas Andrés – N.º DNI 38812216

Director/a:

Lic./Dr./Dra. Apellido, Nombre

**Lugar:
Rosario**

**Fecha de presentación
27/07/2026**

Trabajo final de Grado:

**Diagnóstico Causal y Priorización de Activos
en Sistemas de Distribución Eléctrica de
Media Tensión: Un Enfoque Basado en
Caracterización de Firmas Eléctricas y
Aprendizaje No Supervisado**

Presentado para lograr el TÍTULO de LICENCIADO EN CIENCIA DE DATOS

Autor: Lucas Andres Villarruel

Institución: Universidad del Gran Rosario - Sede Rosario, Santa Fe

Fecha de presentación: 27 Julio 2026

ÍNDICE

1. RESUMEN

2. INTRODUCCIÓN

3. OBJETIVOS E HIPÓTESIS

- 3.1. Problemática observada
- 3.2. Hipótesis de Investigación
- 3.3. Objetivo General
- 3.4. Objetivos Específicos

4. JUSTIFICACIÓN DEL REENFOQUE METODOLÓGICO

- 4.1. Invalidación del modelo predictivo original

5. MARCO TEÓRICO

- 5.1. Sistemas de Distribución y el Rol del Reconectador Automático
- 5.2. La "Firma Eléctrica" como Vector de Estado
- 5.3. Relación Impedancia-Distancia
- 5.4. Tipologías Físicas de Fallas Eléctricas
- 5.5. Confiabilidad y Frecuencia de Eventos (MTBE)
- 5.6. Aprendizaje Automático No Supervisado (Clustering)

6. DESARROLLO EXPERIMENTAL

- 6.1. Descripción y depuración inicial del conjunto de datos
- 6.2. Extracción de firmas eléctricas (Pipeline ETL y RegEx)
- 6.3. Ingeniería de Características (Agrupamiento de Eventos)
- 6.4. Modelo de Aprendizaje No Supervisado (Clustering)

7. RESULTADOS Y DISCUSIÓN

- 7.1. Análisis de Clústeres: Caracterización Física y Validación Algorítmica
 - 7.1.1. Tipología Dominante: Cluster 1 (Sobrecarga y Desbalance Térmico)
 - 7.1.2. Tipologías Secundarias: Fallas a Tierra y Cortocircuitos Francos
 - 7.1.3. Correlación Causal: Patrones Temporales y Demanda Térmica
- 7.2. Ranking Geoespacial y Priorización de Activos
 - 7.2.1. Análisis Macroscópico: Distribución de Pareto
 - 7.2.2. Auditoría Forense de Reconectores (Análisis Micro)

8. CONCLUSIONES

- 8.1. Validación de la Hipótesis de Investigación y Caracterización de Firmas Eléctricas
- 8.2. Madurez de Datos e Ingeniería de Características: Del Registro SCADA al Evento Único
- 8.3. Diagnóstico Sistémico y Patrones de Estrés Térmico Estival
- 8.4. Impacto Práctico y Prescripción para la Optimización de Activos

9. BIBLIOGRAFÍA

- 9.1. Validación Heurística y Comunicaciones Personales
- 9.2. Ecosistema de Software y Librerías Científicas
- 9.3. Inteligencia Artificial y Asistencia Computacional

10. ANEXOS

- Anexo I: Extractos del Código Fuente (Python)
- Anexo II: Muestra del Conjunto de Datos Limpios (df_limpio)
- Anexo III: Repositorio de Código

1. RESUMEN

La gestión óptima del mantenimiento en redes eléctricas de media tensión requiere métodos analíticos capaces de clasificar y priorizar las fallas basándose en su frecuencia temporal y su tipología eléctrica. El presente trabajo desarrolla una metodología de diagnóstico causal y priorización de activos sobre líneas de distribución de **13.2 kV** y **33 kV** en una red de la región patagónica argentina, utilizando registros históricos de actuación de reconectadores automáticos.

Concebido inicialmente como un modelo predictivo basado en variables climáticas, el enfoque fue reestructurado hacia el aprendizaje automático no supervisado. Esta decisión se fundamentó al comprobar la **ausencia de datos operativos del estado "normal"** de la red y la presencia de severos sesgos temporales y espaciales en los registros meteorológicos. Mediante la implementación de un pipeline de preprocesamiento basado en **expresiones regulares (Regex)**, se extrajo el vector de estado tetradimensional de corrientes $S = [I_A, I_B, I_C, I_N]$ (**firma eléctrica**) capturado en el instante exacto del disparo de las protecciones. Tras aplicar un escalado de datos mediante **StandardScaler** y ejecutar el algoritmo de agrupamiento **K-Means (K=4)**, se logró caracterizar las cuatro tipologías físicas subyacentes de las fallas: cortocircuitos bifásicos, fallas trifásicas, cortocircuitos monofásicos a tierra y eventos de sobrecarga. Asimismo, se utilizó la magnitud del amperaje como indicador indirecto de la localización, reconociendo que los valores de corriente son mayores en las cercanías de la Estación Transformadora (ET) y disminuyen a medida que se aleja.

Los resultados demostraron una incidencia dominante de las sobrecargas operativas, representando el 58.7% de los eventos analizados. Este hallazgo, combinado con el análisis de estacionalidad, permitió estructurar un ranking geoespacial basado en el principio de Pareto, identificando cuáles son los reconectadores con mayor tasa de fallas, en qué momentos del año colapsan y bajo qué firmas eléctricas lo hacen. La metodología propuesta demuestra la viabilidad de transformar registros operativos no estructurados en inteligencia de mantenimiento, facilitando la optimización de los recursos en empresas de distribución de energía.

Palabras clave: *Ciencia de Datos; Aprendizaje No Supervisado; K-Means; Reconectores Automáticos; Firma Eléctrica; Mantenimiento Prescriptivo; SCADA.*

2. INTRODUCCIÓN

Los sistemas de distribución eléctrica de media tensión, particularmente en los niveles de **13.2 kV y 33 kV**, operan bajo dinámicas cambiantes donde confluyen fallas transitorias, cortocircuitos severos y fluctuaciones estacionales de la demanda. En estas redes, los reconectadores automáticos actúan como el elemento principal de protección. Ante una anomalía que supera los umbrales de ajuste, el equipo interrumpe el circuito e inicia ciclos de recierre programados para despejar fallas (ej. descargas atmosféricas o contacto temporal con vegetación).

Sin embargo, los sistemas de adquisición de datos (SCADA) y los softwares de gestión operativa registran estas maniobras como simples incidencias discretas (apertura, recierre, bloqueo). En el flujo de datos principal, a menudo se omite o subutiliza el análisis de las magnitudes de corriente registradas en el instante del disparo. Esta limitación provoca que las empresas distribuidoras tiendan a ejecutar un mantenimiento reactivo basado únicamente en la "tasa de fallas" (conteo bruto de operaciones), sin distinguir si el equipo operó por un cortocircuito bifásico violento, una descarga a tierra, o un evento de sobrecarga por alta demanda estacional.

El presente proyecto de grado introduce un enfoque basado en ciencia de datos para mitigar este sesgo analítico. Se propone utilizar los registros del reconectador no solo como eventos de corte, sino como mediciones de alta velocidad que capturan los valores exactos de corriente en cada fase y el neutro $S = [I_A, I_B, I_C, I_N]$. Además, este enfoque reconoce la relación fundamental de la impedancia del sistema: la magnitud de la corriente de falla permite inferir indirectamente la proximidad a la Estación Transformadora (ET), registrándose altos amperajes en las zonas cercanas y corrientes más bajas en los extremos de la línea.

Mediante técnicas de procesamiento de lenguaje natural (**RegEx**) para la extracción de datos y algoritmos de aprendizaje no supervisado (**K-Means**) para su clasificación, este trabajo busca transformar el registro bruto de corrientes en "firmas eléctricas" diagnósticas. El objetivo central es caracterizar el comportamiento de la red respondiendo a tres interrogantes fundamentales: cuáles son los reconectadores que presentan mayor cantidad de fallas, en qué fechas u horarios se concentran dichos eventos, y a qué tipología eléctrica responden, logrando así un modelo jerárquico que priorice la intervención técnica sobre los activos más críticos.

3. OBJETIVOS E HIPÓTESIS

3.1. Problemática observada

Los sistemas de supervisión y adquisición de datos (SCADA) presentan una limitación estructural en el registro de los eventos eléctricos. Cuando un reconectador ejecuta una secuencia de aperturas y recierres rápidos para despejar una única falla, el software operativo registra cada maniobra como una incidencia independiente. Esta fragmentación de la información distorsiona las estadísticas generales, creando un conteo irreal de fallas que impide a las empresas identificar correctamente el problema real de la red, llevándolas a asignar sus recursos de mantenimiento de forma ineficiente.

3.2. Hipótesis de Investigación

Se postula que el vector tetradimensional de corrientes (las mediciones exactas de cada fase y el neutro en el instante del corte) contiene la firma física propia de la falla. Al procesar estos vectores mediante algoritmos de aprendizaje automático no supervisado, es posible agrupar las firmas eléctricas similares para revelar la tipología de la falla, logrando diferenciar eventos fugaces frente a distintos tipos de sobrecargas y cortocircuitos.

3.3. Objetivo General

Clasificar las tipologías de anomalías en redes de distribución de media tensión mediante técnicas de aprendizaje automático no supervisado sobre registros históricos de corriente, con el fin de diseñar un modelo analítico que priorice la intervención de mantenimiento sobre los activos de mayor recurrencia y gravedad de falla.

3.4. Objetivos Específicos

1. **Desarrollar un pipeline de extracción de datos (ETL):** Implementar expresiones regulares (RegEx) para rescatar y estructurar los valores numéricos de corriente que actualmente se encuentran ocultos en campos de texto libre dentro de los reportes operativos.
2. **Reconstruir la continuidad de los eventos físicos:** Aplicar técnicas de agrupación temporal (ventanas deslizantes de tiempo) para unificar las secuencias de recierres bajo un único identificador de evento, eliminando la duplicación estadística en el conjunto de datos.
3. **Determinar la configuración óptima del modelo de agrupamiento:** Evaluar y definir el hiperparámetro adecuado (valor K) para el algoritmo K-Means, asegurando que los grupos formados representen con precisión las tipologías de falla en la red eléctrica (monofásicas, bifásicas, trifásicas y a tierra).
4. **Elaborar un índice de priorización de activos:** Construir un ranking que ordene los equipos cruzando la tipología de la falla identificada a través de las firmas eléctricas con su recurrencia temporal y estacionalidad.

4. JUSTIFICACIÓN DEL REENFOQUE METODOLÓGICO

4.1. Invalidación del modelo predictivo original

La concepción original de esta investigación proponía un modelo de predicción causal. El objetivo era determinar si variables externas, específicamente mediciones climáticas del Servicio Meteorológico Nacional (SMN) de Argentina, podían predecir o anticipar la ocurrencia de una falla en la red eléctrica.

Para lograr esto, se había desarrollado un script que automatizaba la descarga de datos históricos públicos del SMN correspondientes a las regiones del estudio, fusionándolos con el registro de fallas según la fecha y hora de los eventos. Las variables climáticas incorporadas al análisis incluían **temperatura** (TEMP [C°]), **humedad** (HUM [%]), **dirección del viento** (DD [gr]), **velocidad del viento** (FF [km/hr]) y **presión atmosférica** (PNM [hPa]).

Sin embargo, tras el análisis inicial, este enfoque predictivo se descartó debido a dos limitaciones metodológicas estructurales que invalidaban la viabilidad del modelo:

- **1. Ausencia de muestra negativa (Datos de "Normalidad"):** Para entrenar un modelo de clasificación que reconozca un "patrón de falla" (Clase 1), es imprescindible alimentarlo también con datos de su contraparte, el "patrón normal" o "sano" de la red (Clase 0).
 - Los reconectores, por su diseño operativo, solo generan reportes por excepción; es decir, únicamente registran información cuando ocurre una falla (Clase 1).
 - En consecuencia, no existen registros de corrientes correspondientes al funcionamiento nominal con carga de la red (ej. 150A, 145A, 152A).
 - Sin este grupo de control (estado "sano"), resulta imposible que el modelo aprenda a diferenciar las condiciones que preceden a una falla de las condiciones normales de operación. Intentar predecir anomalías contando exclusivamente con un historial de anomalías es metodológicamente inviable.

- **2. Invalidación de Datos Climáticos (Principio GIGO) y Sesgos Estructurales:** El intento de integrar los datos climáticos del SMN demostró ser inviable. Aunque variables como vientos continuos o temperaturas extremas son relevantes, muchas fallas son desencadenadas por eventos micro-locales y aleatorios (ej. caída de un árbol o ráfagas puntuales) que escapan a la resolución de las estaciones meteorológicas regionales, generando ruido e inexactitudes. Aún más grave, la fusión de los conjuntos de datos reveló sesgos críticos:

Falta de Granularidad y Cobertura en Datos Climáticos: Faltaba granularidad en los datos climáticos y temporales. Si una falla ocurría entre horas, por ejemplo a las 8:39 hs, no se disponía de datos exactos del clima, forzando emparejamientos con reportes horarios estáticos. Adicionalmente, la información dependía de la ubicación geográfica de los centros meteorológicos; al contar únicamente con 2 centros de información en la región, el 97.3% de los registros se concentraba en solo dos ciudades. Filtrar el conjunto de datos bajo estas condiciones lo reducía drásticamente a menos de 500 eventos, una cantidad insuficiente que anulaba la solidez estadística del análisis.

- **Conclusión Metodológica:** Ante la ausencia de una muestra negativa (Clase 0) y los sesgos críticos evidenciados por la falta de granularidad espacio-temporal, el objetivo predictivo causal fue abandonado. Las métricas del SMN se eliminaron por completo para evitar la aplicación del principio GIGO, justificando el reenfoque metodológico hacia el aprendizaje no supervisado, el cual basa el diagnóstico de forma exclusiva en las firmas eléctricas capturadas por los equipos.

5. MARCO TEÓRICO

El desarrollo de esta investigación se sustenta en la intersección de dos disciplinas: la ingeniería eléctrica, específicamente la protección y operación de sistemas de distribución de media tensión, y la ciencia de datos, mediante la aplicación de algoritmos de aprendizaje automático no supervisado. A continuación, se exponen los fundamentos teóricos que permiten traducir registros operativos de la red en diagnósticos accionables.

5.1. Sistemas de Distribución y el Rol del Reconectador Automático

Las redes de distribución de media tensión (típicamente 13.2 kV y 33 kV) operan bajo condiciones dinámicas y conforman la infraestructura crítica que transporta la energía desde las Estaciones Transformadoras (ET) hasta los centros de consumo. Al ser en su mayoría redes aéreas, están expuestas permanentemente a factores ambientales (viento, descargas atmosféricas, vegetación, fauna) y a fluctuaciones en la demanda térmica y operativa.

El **reconectador automático** es el dispositivo de protección principal diseñado para interrumpir corrientes de cortocircuito y sobrecarga. Su característica principal es la capacidad de ejecutar ciclos programados de apertura y recierre rápido. Dado que estadísticamente la mayoría de las fallas en redes aéreas son de naturaleza transitoria (por ejemplo, el contacto temporal de una rama con un conductor), el recierre automático permite despejar la falla y restaurar el servicio sin intervención humana. Si la falla persiste tras agotar los intentos programados, el equipo entra en estado de "bloqueo", resultando en un corte definitivo (Energía No Suministrada - ENS).

En el contexto de este trabajo, el reconectador es conceptualizado no solo como un actuador de protección, sino como un **adquisidor de datos de alta velocidad**. Al momento del disparo, actúa como un osciloscopio, congelando el estado eléctrico de la red y registrando un snapshot de las condiciones exactas.

5.2. La "Firma Eléctrica" como Vector de Estado

Se define como "Firma Eléctrica" al vector tetradimensional capturado por el reconector en el instante t de la actuación:

$$S = [I_A, I_B, I_C, I_N]$$

Donde I_A, I_B, I_C representan las corrientes de las tres fases y I_N representa la corriente residual o de neutro (tierra), medidas en Amperes (A). La interpretación conjunta de este vector encapsula la naturaleza física de la perturbación y se basa en el Principio de Superposición:

- **Fases de Falla:** Aquellas que presentan valores anómalos o de ruptura dieléctrica, registrando la magnitud directa de la anomalía ($I_{falla} + I_{carga}$).
- **El Dato Oculto (Carga Residual):** En fallas asimétricas, las fases no involucradas continúan midiendo exclusivamente la Corriente de Carga Pre-Falla (I_{carga}). Este es un dato vital que actúa como un sensor de contexto operativo, permitiendo inferir si la anomalía ocurrió en un momento de valle o bajo condiciones de alta demanda (estrés del sistema).
 - **Caso A (Falla en Vacío):** Firma [1200, 5, 4, 1205]. Indica una falla masiva en A, pero con B y C casi en cero. La falla ocurrió en un momento de carga mínima (ej. madrugada) o en una línea sin usuarios activos.
 - **Caso B (Falla en Plena Carga):** Firma [1350, 180, 175, 1360]. La misma falla en A, pero B y C muestran ~180A. Esto indica que el evento ocurrió bajo alto estrés operativo.

5.3. Relación Impedancia-Distancia

En ausencia de sistemas avanzados de localización de fallas (como FLISR) o impedanciometros integrados, la magnitud escalar de la corriente de cortocircuito actúa como un estimador indirecto (proxy) de la ubicación geoespacial del evento.

De acuerdo con la Ley de Ohm generalizada:

$$I_{\text{falla}} = \frac{V_{\text{fuente}}}{Z_{\text{sys}} + Z_{\text{línea}} \times d}$$

Dado que el voltaje (V_{fuente}) y la impedancia del sistema (Z_{sys}) son constantes, la corriente es inversamente proporcional a la distancia eléctrica y física desde la subestación.

- **Fallas cercanas a la ET (Zona Roja):** Firmas con valores extremos (ej. >1500A). Indican baja impedancia de línea. Son los eventos más peligrosos para el equipamiento debido al estrés pasante.
- **Fallas lejanas a la ET (Zona de Cola):** Firmas con valores menores (ej. 140A). Indican eventos al final del alimentador o fallas de Alta Impedancia, próximas a los umbrales mínimos de disparo de las protecciones.

Aclaración conceptual: El presente estudio evalúa el comportamiento y la tipología del evento a través de la corriente, sin pretender medir matemáticamente la fatiga, degradación de materiales, o el desgaste térmico estructural de la línea. El enfoque es estrictamente estadístico y eléctrico.

5.4. Tipologías Físicas de Fallas Eléctricas

A partir de la magnitud y la simetría del vector S y el análisis de la secuencia cero (corriente de tierra, I_n), los eventos pueden clasificarse en familias de fenómenos físicos:

1. Familia de Cortocircuitos (Alta Magnitud):

- **Monofásico a Tierra (Componente Homopolar):** Alta correlación entre la fase fallada y la tierra ($Fase \approx Tierra \gg Fases_Sanas$). Involucra una fase y el retorno por tierra. Suele asociarse a aisladores perforados, vegetación o corte de un conductor que cae al suelo.
- **Bifásico:** Corriente elevada en dos fases con corriente de tierra nula o despreciable. La corriente fluye entre fases sin tocar el suelo. Típicamente causado por el choque de dos conductores (galopeo por viento) o la caída de una rama sobre dos fases.
- **Trifásico (Simétrico):** Involucran las tres fases simultáneamente con magnitudes extremas y equilibradas (tierra nula). Son eventos catastróficos o producto de maniobras erróneas (cierres sobre puestas a tierra).

2. Familia de Sobrecargas (Estrés Operativo):

- A diferencia del cortocircuito, la sobrecarga ocurre cuando la impedancia del sistema es normal, pero la demanda de carga supera la capacidad nominal de la infraestructura o la configuración del equipo. Se caracteriza por corrientes moderadas-altas (rango medio-alto), generalizadas o severamente desbalanceadas, con una corriente de tierra moderada o nula en relación a las fases activas. Representan a la red operando al límite por demanda (típicamente estacional) y no por accidentes externos.

5.5. Confiabilidad y Frecuencia de Eventos (MTBE)

Para evaluar la cronicidad de las fallas y auditar la operatividad de los activos, se utilizan dos métricas derivadas:

- **MTBE (Mean Time Between Events):** Distancia temporal que separa incidentes únicos en un mismo reconfigurador. Permite aislar los "accidentes" estadísticos de aquellas fallas sistémicas, recurrentes y crónicas.
- **Índice de Persistencia:** Ratio entre la cantidad de registros brutos (intentos de recierre) y los eventos únicos consolidados. Cuantifica la dificultad de la red para despejar las anomalías, aislando las fallas transitorias (fugaces) de los bloqueos definitivos.

5.6. Aprendizaje Automático No Supervisado (Clustering)

Dado que los sistemas operativos de la red eléctrica solo reportan eventos por excepción (**Clase 1: Falla**), el conjunto de datos carece de una muestra de control del estado nominal (**Clase 0: Operación Normal**). Esta limitación invalida el uso de técnicas de predicción o clasificación supervisada, justificando la adopción de algoritmos no supervisados.

- **K-Means:** Es un algoritmo de particionamiento que recibe los vectores S sin etiquetas predefinidas y los agrupa en K clústeres basándose en su distancia euclidiana. Al estandarizar las magnitudes de corriente, el modelo "redescubre" la estructura oculta de los datos, diferenciando tipologías sin intervención humana. La selección del hiperparámetro K se optimiza heurísticamente mediante el Método del Codo (Elbow Method), evaluando la reducción de la inercia (suma de distancias intra-cluster).
- **Clustering Jerárquico Aglomerativo:** Se utiliza como método de validación estructural. A diferencia de K-Means, este algoritmo construye una jerarquía de agrupamientos (dendrograma) utilizando criterios como el enlace de Ward (minimización de la varianza). El cruce de resultados entre ambos modelos permite certificar matemáticamente la robustez de las tipologías diagnosticadas en la red.

6. DESARROLLO EXPERIMENTAL

El presente capítulo detalla la implementación computacional de la metodología propuesta, estructurada como un pipeline analítico de procesamiento de datos end-to-end. La complejidad de este desarrollo radica en la necesidad de transformar telemetría industrial cruda y no estructurada en un modelo de inteligencia prescriptiva. Se describe paso a paso la transición desde la ingesta del archivo bruto de registros operativos hasta el entrenamiento, optimización y validación algorítmica del modelo de agrupamiento no supervisado. Todo el entorno de desarrollo fue programado utilizando el lenguaje Python, apoyado en el ecosistema de librerías científicas Pandas, Scikit-Learn y NumPy.

6.1. Descripción y depuración inicial del conjunto de datos

El insumo primario de la investigación consistió en una extracción histórica de registros operativos de la flota de reconectadores automáticos. Dado que esta información refleja la vulnerabilidad y el estado operativo de infraestructuras críticas, se estableció un riguroso protocolo de Ingesta Segura para el tratamiento de los datos sensibles. Utilizando criptografía simétrica (mediante la librería *Fernet*) y la derivación de claves bajo el estándar **PBKDF2HMAC**, se aseguró que la información confidencial de la red eléctrica (nombres de subestaciones, coordenadas geográficas precisas y alimentadores) sólo fuera accesible en la memoria temporal volátil del entorno de ejecución. Esto garantizó el anonimato y la seguridad de los activos de la empresa distribuidora frente a posibles filtraciones.

El dataset original, suministrado por el sistema SCADA en formato .tsv (Tab-Separated Values), contenía 3.830 filas que abarcaban el periodo operativo entre los años 2023 y 2025. Estructuralmente, la matriz presentaba 17 columnas que combinaban identificadores de equipamiento, estampas temporales de alta precisión (milisegundos), ubicación geográfica y, de forma problemática, campos de texto libre ingresados manualmente por los operadores del centro de control.

A continuación, se detalla la estructura y el diccionario de datos inicial de este archivo:

1. Variables temporales y de ubicación:

- **Fecha:** Marca temporal exacta del evento (DD/MM/AAAA HH:MM:SS).
- **Cód. Instal.:** Identificador abreviado del sistema o zona.
- **Instalacion:** Nombre descriptivo de la línea de distribución y su nivel de tensión (13.2 kV).
- **Responsable:** Sucursal operativa a cargo de la zona.

2. Identificación del activo:

- **Cód. Equ.:** Código operativo del reconectador.
- **Equipo:** Denominación completa del dispositivo.
- **id_Equipamiento:** Clave numérica única del activo físico.

3. Diagnóstico y telemetría:

- **Sistema:** Módulo informático de origen del log (ej. Norte, Sur).
- **Motivo:** Tipo de evento (constante en el valor "FALLA").
- **Descripción:** Mensaje generado por el sistema SCADA sobre la maniobra ejecutada y el resultado del recierre.
- **Actuaciones:** Detalle técnico de la función de protección disparada.
- **Observaciones:** Campo crítico de texto libre que contiene anotaciones operativas y, fundamentalmente, el vector tetradimensional de corrientes de falla (I_A, I_B, I_C, I_N) bajo formatos no normalizados.
- **CondicionClimatica:** Ingreso textual manual de los operadores con observaciones del clima o del entorno.

4. Variables administrativas y campos nulos:

- **ENS:** Indicador administrativo interno de estado de servicio.
- **Ap:** Variable de control o registro de apertura.
- **valoresFalla:** Columna estructuralmente vacía (sin datos cargados).
- **MotivoFalla:** Columna auxiliar sin datos cargados.

Conociendo esta estructura, y bajo el principio fundamental de la ciencia de datos referente a la calidad de la información (*Garbage In, Garbage Out*), la primera fase del procesamiento consistió en una depuración exhaustiva. El objetivo fue eliminar variables carentes de poder predictivo o explicativo, así como descartar aquellos campos con ruido inmanejable. Para ello, se ejecutaron las siguientes acciones de limpieza estructural:

- **Eliminación por Nulidad Crítica:** Se descartó por completo la columna `valoresFalla`. Al realizar el análisis exploratorio (EDA), se evidenció que esta variable presentaba un índice de valores nulos (NaN) superior al 90%. Al carecer de la densidad de datos mínima requerida, cualquier intento de imputación estadística hubiera introducido sesgos artificiales, resultando irrecuperable para el análisis matricial.
- **Eliminación por Varianza Cero:** Se suprimió la variable categórica `Motivo`. Al explorar su distribución de frecuencias, se constató que contenía un único valor constante ("Falla") para la totalidad de los registros del sistema. En el contexto del aprendizaje automático, una variable sin dispersión estadística carece de capacidad discriminatoria; no aporta información para separar grupos y constituye un gasto computacional innecesario (ruido).
- **Filtrado de Ruido Operativo y Maniobras:** Los reconectores no solo operan frente a cortocircuitos, sino que también son accionados de forma manual o remota para mantenimientos programados, transferencias de carga o pruebas de campo. Para aislar el comportamiento estocástico de las anomalías, se procedió a segregar las fallas reales de estas maniobras administrativas. Para ello, se eliminaron los registros que presentaban valores vacíos en la columna de reporte técnico (`Observaciones`) o en la métrica de impacto directo al usuario (`ENS - Energía No Suministrada`). Tras aplicar esta máscara de filtrado, garantizando que solo se analizaran accionamientos automáticos disparados por perturbaciones de la red, la muestra se redujo a 2681 registros operativos válidos.

6.2. Extracción de firmas eléctricas (Pipeline ETL y RegEx)

La limitación estructural más severa y desafiante del conjunto de datos original residía en la arquitectura de almacenamiento del sistema SCADA heredado. En lugar de archivar las magnitudes de corriente en columnas numéricas tabuladas e independientes (por ejemplo, una base de datos relacional con campos Fase_A, Fase_B, etc.), el software concatenaba estos valores vitales dentro de un único string de texto libre en la columna denominada Observaciones. Este formato es común en registros de alarmas, generando cadenas como: *"Corriente máxima de fase registrada: 145.5; 150; 148; 20 - Bloqueo de equipo"*.

El análisis matemático de vectores requiere dimensiones numéricas puras. Para transformar este texto no estructurado en el vector de estado tetradimensional $S = [I_A, I_B, I_C, I_N]$, fue imperativo desarrollar un algoritmo de Extracción, Transformación y Carga (ETL) fundamentado en técnicas de procesamiento de texto mediante Expresiones Regulares (RegEx).

El motor de procesamiento aplicó una secuencia lógica iterativa sobre todo el conjunto de datos:

1. **Patrón de Búsqueda Flexible:** Se diseñó y aplicó la expresión regular `patron_numerico = r'(\d+\.\d*)'` sobre cada celda individual de la columna Observaciones. Este patrón es altamente resiliente: fue diseñado para escanear, identificar y aislar secuencias de dígitos de longitud variable, admitiendo tanto números enteros como valores de punto flotante (decimales separados por punto), ignorando al mismo tiempo caracteres alfabéticos, signos de puntuación irregulares o errores de tipeo introducidos por el sistema o los operadores.

2. **Estructuración Vectorial:** Una vez identificadas las coincidencias dentro de la cadena de texto, el algoritmo extrajo estrictamente los primeros cuatro valores numéricos encontrados (asumiendo por el protocolo del SCADA que el orden siempre es Fase A, Fase B, Fase C y Neutro/Tierra). Inmediatamente, estos valores fueron casteados (convertidos) de formato string a tipo flotante (`float64`) y mapeados secuencialmente a cuatro nuevas columnas de la matriz: `Corr_A`, `Corr_B`, `Corr_C` y `Corr_Tierra`, reconstruyendo así el vector S
3. **Control de Integridad Física (Bounding Filter):** La extracción de texto es propensa a capturar "falsos positivos" (ej. números de serie colados en el texto). Para blindar la calidad de los datos, se aplicó una máscara condicional basada en las leyes de la física del sistema eléctrico. Se conservaron únicamente las filas donde la totalidad de los cuatro valores se ubicaran en el rango realista de 0A a 5000A. El límite superior de 5000 Amperes obedece a la capacidad de ruptura teórica de los equipos; valores superiores representarían un arco destructivo que fundiría la línea, o bien, un error grave del sensor. Los registros que no cumplieron este criterio termodinámico o que carecían de los cuatro valores exactos fueron descartados definitivamente.

La culminación exitosa de este pipeline ETL arrojó un dataset consolidado y limpio de 2569 registros, dotando por primera vez a los datos de firmas eléctricas numéricamente válidas y listas para el álgebra lineal.

6.3. Ingeniería de Características (Agrupamiento de Eventos)

La naturaleza operativa y mecánica de los reconectores introduce un severo sesgo estadístico de duplicación en las bases de datos de las distribuidoras, conocido en la industria como "multiplicidad de registros por recierre". Ante un único cortocircuito físico en la red (por ejemplo, la caída de una gran rama sobre la línea), el equipo no se bloquea inmediatamente. En cambio, ejecuta iterativamente secuencias automáticas de apertura y cierre (generalmente 3 o 4 intentos en fracciones de segundo) para intentar despejar la falla si esta fuera fugaz.

El problema analítico surge porque el sistema SCADA registra cada uno de estos recierres mecánicos como una fila independiente en la base de datos, con diferencias de milisegundos o segundos en su estampa de tiempo. Analizar estos datos en bruto sin tratamiento previo provocaría el conteo múltiple de una misma anomalía física. Esto inflaría artificialmente las tasas de siniestralidad de ciertos equipos, destruyendo la validez de cualquier ranking de priorización basado en el principio de Pareto, ya que un solo incidente se vería estadísticamente como cuatro fallas distintas.

Para subsanar esta distorsión crítica, se implementó un proceso de ingeniería de características (*Feature Engineering*) creando una nueva variable estructural denominada *Evento_ID*. Esta variable fue diseñada específicamente para unificar la secuencia de recierres bajo el paraguas de un solo evento físico real. La lógica computacional empleó una técnica algorítmica de agrupación por ventana temporal deslizante (*rolling time-window*):

1. **Ordenamiento Espacio-Temporal:** El conjunto de datos se agrupó y ordenó, primariamente, por el identificador del activo físico (`GROUP BY ID_Instalacion`) para evitar mezclar fallas de distintos equipos. Secundariamente, dentro de cada grupo, se ordenaron de forma estrictamente cronológica basándose en la estampa temporal (*Fecha*).
2. **Cálculo del Diferencial Temporal:** Utilizando las funciones vectorizadas de Pandas, específicamente `diff()`, se calculó el tiempo transcurrido exacto (Δt) entre un registro dado y su inmediato anterior dentro del historial del mismo equipo eléctrico.
3. **Definición de la Ventana Heurística (1 Hora):** Se estableció una condición booleana basada en la heurística del mantenimiento eléctrico: si la diferencia temporal entre dos registros consecutivos del mismo equipo era superior a una ventana de 1 hora ($\Delta t > 1 \text{ hora}$), o si el cálculo arrojaba un valor nulo (indicando que era el primer registro histórico de ese equipo), la condición lógica evaluaba como `True`. Esto marcaba inequívocamente el inicio de un nuevo evento físico e independiente. Por el contrario, si $\Delta t \leq 1 \text{ hora}$, evaluaba como `False`, considerando que el accionamiento era simplemente un recierre tardío o un intento manual de restitución perteneciente al mismo evento original.

4. **Asignación del ID Consolidado:** Finalmente, se aplicó una función de suma acumulativa (`cumsum()`) sobre esta matriz de condiciones booleanas. Esto generó un identificador entero, secuencial y único para cada clúster temporal.

El impacto de esta consolidación fue masivo: transformó los 2569 registros brutos originales en 2146 Eventos Únicos Reales. La diferencia matemática de 423 registros no fue descartada, sino que se convirtió en una métrica de inmenso valor: permitió cuantificar el "Índice de Persistencia", revelando el nivel de actividad de los automatismos y la dificultad real de la red para despejar las fallas ambientales a través de los intentos de recierre.

6.4. Modelo de Aprendizaje No Supervisado (Clustering)

Con el espacio vectorial topológicamente definido por eventos únicos y firmas eléctricas multidimensionales tabuladas, se procedió a la fase central de modelado matemático para descubrir, clasificar y diagnosticar las causas ocultas en los registros.

Transformación del Espacio Vectorial: Escalado de Datos (StandardScaler)

Como paso previo, absoluto e indispensable antes de someter los datos al motor de agrupamiento matemático, se aplicó un proceso de estandarización sobre las cuatro columnas de corriente utilizando la clase `StandardScaler`. Esta operación estadística transformó los valores brutos medidos en Amperes en puntuaciones Z (Z -scores), re-escalando los datos de modo que cada variable del vector quedó centrada en una media $\mu = 0$ y con una desviación estándar $\sigma = 1$.

La justificación de este procedimiento radica profundamente en la mecánica matemática del algoritmo K-Means. Este modelo opera minimizando la inercia (la suma de los errores cuadráticos) mediante el cálculo de distancias euclidianas entre los puntos (vectores de corriente) y los centroides del espacio. Los algoritmos basados en distancia son extremadamente vulnerables y sensibles a la escala y magnitud de las dimensiones.

De no haber aplicado el escalado, se habría incurrido en un error analítico grave: una falla trifásica violenta con una magnitud extrema (ej. 2000A en cada fase) generaría un peso de varianza geométricamente tan masivo que dominaría por completo la función de costo del algoritmo. Esto ocultaría y anularía matemáticamente la presencia de una sobrecarga crónica o una falla resistiva de alta impedancia (ej. 200A), agrupándolas incorrectamente o ignorándolas. El escalado anula este sesgo de magnitud bruta (la "fuerza" de la falla dictada por su cercanía a la subestación), forzando al modelo algorítmico a prestar atención primordial a la *forma* de la firma: la geometría vectorial, la asimetría y el desbalance relativo entre las fases. Esto garantiza una agrupación equitativa basada en la "naturaleza" del fenómeno, sin importar cuán cerca o lejos del transformador principal haya ocurrido el cortocircuito.

Configuración, Optimización del Hiperparámetro K y Alineación Teórica

Se seleccionó el algoritmo particional K-Means para ejecutar el motor de diagnóstico principal debido a su eficiencia con matrices de alta densidad y su interpretabilidad de centroides. Sin embargo, la precisión predictiva y descriptiva de este modelo es completamente dependiente de la correcta asignación del hiperparámetro K (el número total de clústeres o agrupaciones a descubrir). Dado que el problema carecía de etiquetas previas, se hizo necesario determinar su valor óptimo empíricamente.

Se evaluó el comportamiento de la Inercia matemática (la suma de las distancias al cuadrado intra-cluster, que mide la compacidad de los grupos) iterando el entrenamiento del modelo en un rango de **K=2** a **K=10**. Al aplicar el **Método del Codo (Elbow Method)** y graficar los resultados, la curva de inercia reveló un comportamiento revelador: evidenció una reducción drástica y acelerada de la inercia desde **K=2** (7109) hasta formar un "codo" pronunciado exactamente en **K=4** (4861). A partir de **K=5**, la curva demostró una rápida estabilización o aplanamiento asintótico. Este aplanamiento indica que agregar más clústeres aporta rendimientos decrecientes y conlleva un altísimo riesgo de sobreajuste (overfitting), resultando en una fragmentación artificial y sin sentido de los datos reales. Este hallazgo matemático fue analizado y corroborado adicionalmente mediante el análisis del coeficiente de **Silueta (Silhouette Score)**, que confirmó que la densidad y separación intra/inter cluster se maximizaba en cuatro particiones.

La selección definitiva de $K=4$ fue un hito en la investigación porque demostró una convergencia física perfecta. La teoría de ingeniería eléctrica de potencia define estrictamente que existen cuatro familias fundamentales de problemas que pueden manifestarse en una red de distribución:

1. Cortocircuito Monofásico a Tierra.
2. Cortocircuito Bifásico (aislado de tierra).
3. Cortocircuito Trifásico (simétrico).
4. Desbalance sostenido por Sobrecarga Térmica.

El hecho de que un algoritmo matemático ciego, trabajando puramente con distancias en un espacio vectorial n -dimensional, haya detectado exactamente la misma cantidad de estructuras que las dictaminadas por las leyes de Kirchhoff, valida profundamente la pertinencia del modelo propuesto.

Validación Cruzada Avanzada mediante Clustering Jerárquico y Dendrogramas

En la ciencia de datos aplicada a infraestructuras críticas, un hallazgo no supervisado requiere pruebas de solidez para asegurar que no se trata de una anomalía del modelo elegido. Para certificar la robustez y la realidad física de la topología descubierta por K-Means, se diseñó un experimento de validación cruzada ejecutando, en paralelo y sobre el mismo espacio muestral, un modelo de Clustering Jerárquico Aglomerativo.

A diferencia de la naturaleza de K-Means (un algoritmo particional iterativo que requiere la predefinición dura de clústeres y agrupa por proximidad espacial a un centroide central), el método jerárquico es de naturaleza ascendente (bottom-up). Asume en su paso inicial que cada uno de los 2146 eventos es su propio micro-clúster individual, y procede a fusionarlos progresivamente paso a paso, evaluando constantemente el incremento de la varianza total mediante la sofisticada métrica del método de enlace de Ward.

El intrincado resultado de este proceso de fusión iterativa se representó visualmente a través de un dendrograma completo. El uso de esta técnica algorítmica cumplió dos funciones analíticas absolutamente críticas para la validez de la tesis:

1. **Confirmación Visual e Independiente del Hiperparámetro K:** La inspección minuciosa del eje de distancias del dendrograma evidenció un "salto" o vacío espacial muy significativo en la distancia de agrupamiento antes de la fusión final. Cortar el árbol en ese umbral de máxima varianza dividía la topología jerárquica exactamente en 4 ramas verticales principales y claramente diferenciadas. Este hallazgo validó de forma independiente, y con una arquitectura matemática distinta, la selección de **K=4** obtenida originalmente mediante el Método del Codo, descartando definitivamente la posibilidad de que dicho valor fuera producto de un mínimo local o un sesgo propio de la inicialización aleatoria de K-Means.
2. **Auditoría de Consenso mediante Matriz de Confusión Algorítmica:** El paso definitivo fue contrastar directamente las asignaciones evento por evento de ambos modelos antagónicos para verificar la solidez de las tipologías encontradas. El foco principal de esta auditoría fue evaluar la tipología de "Sobrecarga Operativa", la cual había emergido sorprendentemente como la anomalía causal dominante en la red. Al cruzar los resultados, se descubrió que el modelo Jerárquico agrupó en la misma rama topológica al 96.3% de los miles de eventos que el modelo K-Means había clasificado como Sobrecarga de forma independiente. Este nivel de consenso algorítmico, casi absoluto entre dos modelos probabilísticos diferentes, demostró estadísticamente más allá de cualquier duda razonable que la tipología de "sobrecarga y fuga" es una estructura de datos real, densa y empíricamente bien definida. Esto confirmó que el diagnóstico principal que señala a la red como térmicamente estresada es un hallazgo causal robusto, fundamentado matemáticamente y no el producto de una fluctuación o anomalía estadística aleatoria.

7 .RESULTADOS Y DISCUSIÓN

7.1. Análisis de Clústeres: Caracterización Física y Validación Algorítmica

Antes de proceder con la caracterización física de las anomalías, fue imperativo determinar la cantidad óptima de particiones (**K**) para el modelo K-Means. Esta selección no se realizó de manera arbitraria, sino mediante una triangulación entre la optimización matemática del error y el conocimiento experto del dominio eléctrico.

En primera instancia, se ejecutó el Método del Codo (Elbow Method), evaluando iterativamente el modelo para un rango de $K \in [2, 10]$. El análisis de la inercia demostró una reducción significativa y continua del error hasta **K=4**. A partir de **K=6**, la curva alcanza un punto de retornos decrecientes pronunciado, indicando que adiciones posteriores de clústeres solo lograrían una fragmentación matemática sin aportar una distinción física real.

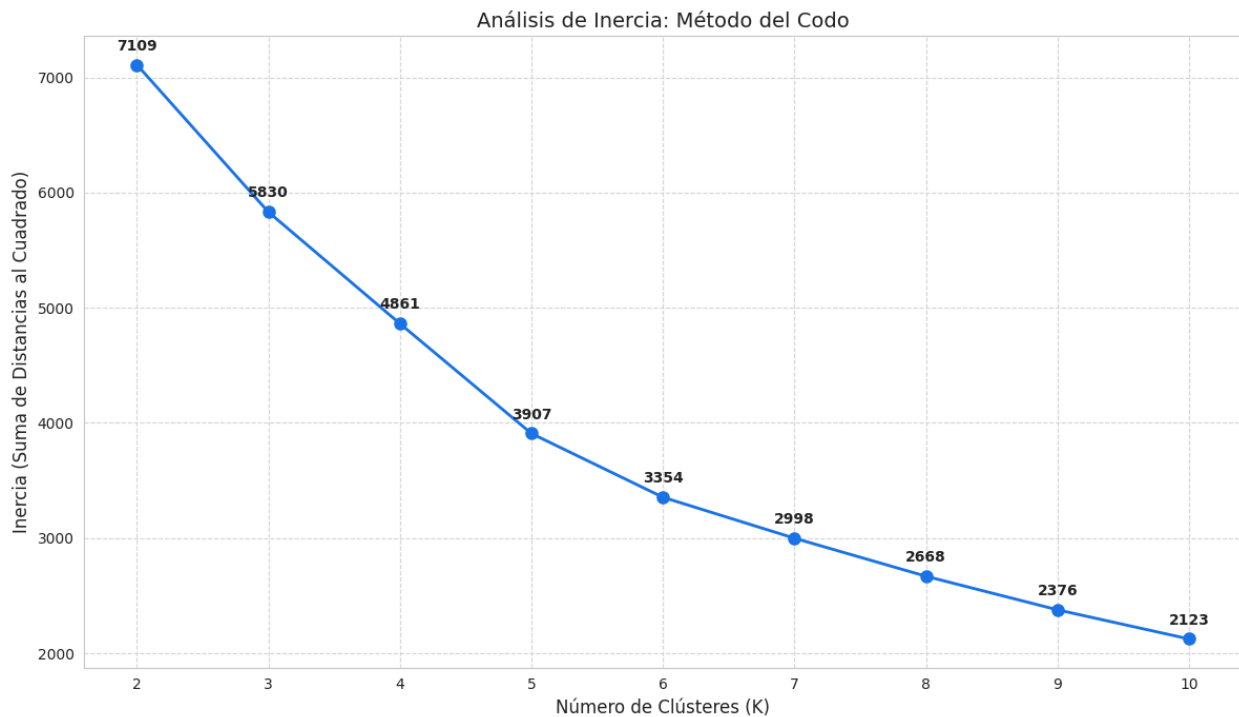


Figura 0. Análisis de Inercia: Método del Codo

Aunque la métrica puramente matemática sugería que **K=5** también ofrecía una partición aceptable, la selección final de **K=4** se fundamentó en el Principio de Parsimonia y en la teoría de protecciones eléctricas. Un sistema de distribución trifásico presenta cuatro tipologías físicas fundamentales de falla: monofásica, bifásica, trifásica y desbalance/sobrecarga. Forzar al modelo a aislar un quinto grupo habría provocado la subdivisión artificial de la categoría de sobrecarga, las cuales operativamente requieren la misma acción correctiva de mantenimiento.

Esta decisión fue posteriormente reconfirmada de manera visual mediante un Dendrograma de Clustering Jerárquico. La estructura genealógica de los datos evidenció un salto aglomerativo limpio, donde un corte horizontal estándar atraviesa exactamente cuatro ramas verticales principales, validando de forma definitiva que el historial de la red se divide naturalmente en cuatro grandes grupos de comportamiento.

7.1.1. Tipología Dominante: Cluster 1 (Sobrecarga y Desbalance Térmico)

Una vez clasificado el dataset, el análisis de las proporciones resultantes revela una distribución fuertemente asimétrica en la siniestralidad de la red. El hallazgo empírico más relevante es la dominancia absoluta del Cluster 1, el cual concentra el **58.7% de la siniestralidad total** del sistema, representando 1.509 eventos de falla documentados.

Cluster ID	Cantidad (N)	Prevalencia (%)	Clasificación Preliminar
Cluster 1	1509	58.7%	Tipología Dominante: Representa casi el 60% de los problemas de la red. Es una falla sistémica y crónica.
Cluster 3	652	25.4%	Tipología Secundaria: Frecuente pero no mayoritaria.
Cluster 2	220	8.6%	Tipología Minoritaria A: Eventos menos frecuentes.
Cluster 0	188	7.3%	Tipología Minoritaria B: Eventos esporádicos.

El análisis de los centroides de este grupo mayoritario revela una configuración física particular: las magnitudes de corriente en las fases sanas operan en rangos bajos a medios (entre 163A y 192A), acopladas a una corriente homopolar (falla a tierra) moderada y persistente de aproximadamente 64.8A. Esta arquitectura eléctrica demuestra empíricamente que la inestabilidad principal de la red no proviene de accidentes mecánicos severos o cortocircuitos catastróficos, sino de un estrés operativo crónico. El patrón se corresponde con desbalances de carga significativos y retornos por neutro, indicando equipos operando sostenidamente al límite de sus curvas térmicas.

Para validar la robustez de esta tipología y asegurar que la concentración de datos no fuera un sesgo del algoritmo K-Means, se ejecutó una prueba de **validación cruzada utilizando Clustering Jerárquico Aglomerativo**. Los resultados evidenciaron un consenso algorítmico casi absoluto: de los 1.509 eventos clasificados como "Sobrecarga" por K-Means, el modelo jerárquico agrupó 1.453 en un mismo clúster equivalente.

Matriz de Contingencia (Validación Cruzada de Algoritmos: K-Means vs. Jerárquico)

Modelo K-Means \ Modelo Jerárquico	Clúster 0	Clúster 1	Clúster 2	Clúster 3	Total K-Means
Clúster 0 (Cortocircuito Bifásico)	110	78	0	0	188
Clúster 1 (Sobrecarga / Térmico)	0	1453	0	56	1509
Clúster 2 (Cortocircuito Trifásico)	98	21	0	101	220
Clúster 3 (Falla Monofásica a Tierra)	0	339	312	1	652
Total Jerárquico	208	1891	312	158	2569

Guía de Análisis:

Filas: *K-Means (Cluster_Firma)*

Columnas: *Jerárquico (Cluster_Jerarquico)*

Diagonalidad: *Si los valores altos están en celdas únicas por fila, significa que ambos modelos ven las mismas estructuras (Alta Robustez).*

Dispersión: *Si una fila de K-Means se reparte en varias columnas Jerárquicas, el modelo Jerárquico está encontrando sub-grupos que K-Means unificó.*

Esta coincidencia del **96.3%** corrobora estadísticamente que la tipología de "Sobrecarga/Fuga" es una estructura de datos densa, real e independiente de la técnica matemática empleada para su extracción.

7.1.2. Tipologías Secundarias: Fallas a Tierra y Cortocircuitos Francos

Si bien la sobrecarga térmica dicta la cronicidad del sistema, el algoritmo identificó de manera robusta otras tres tipologías físicas de falla, las cuales representan eventos de mayor violencia y menor impedancia.

Diagnóstico de Tipologías Físicas (Centroides de K-Means)

Clúster (Tipología Física)	Cantidad de Eventos	Prevalencia (%)	Corriente Fase A (A)	Corriente Fase B (A)	Corriente Fase C (A)	Corriente a Tierra (A)
Clúster 0 (Cortocircuito Bifásico)	188	7.3 %	1010.2	985.2	65.9	15.4
Clúster 1 (Sobrecarga / Desbalance)	1509	58.7 %	163.4	166.7	192.1	64.8
Clúster 2 (Falla Trifásica Masiva)	220	8.6 %	711.3	784.0	1304.8	13.6
Clúster 3 (Falla Monofásica a Tierra)	652	25.4 %	155.3	161.3	170.7	338.6

Guía de Interpretación:

Magnitud > 500A: Indica Falla Franca (Cortocircuito de baja impedancia).

Magnitud < 300A: Indica Sobrecarga o Falla de alta impedancia.

Desbalance: Observe si una fase es mucho mayor a las otras.

Tierra: Si es alto (>50A), hay fuga. Si es ~0A, es falla pura entre fases.

- **Cluster 3 (Falla Monofásica a Tierra):** Constituye el segundo grupo más frecuente, con un **25.4% de los incidentes (652 eventos)**. Físicamente, se caracteriza por una corriente homopolar fuertemente dominante, con un promedio de 338.6A, mientras que las tres fases mantienen niveles de carga residual que oscilan entre los 155A y 170A. Esta asimetría severa diagnostica la típica falla de distribución aérea de media impedancia, originada comúnmente por el contacto de una fase con vegetación, la perforación de un aislador o la rotura de un puente.
- **Cluster 2 (Falla Trifásica Masiva):** Agrupa el **8.6% de los registros (220 eventos)**. Exhibe corrientes generalizadas de altísima energía en todas las fases (entre 711A y 1304A), acopladas a una corriente de tierra nula (13.6A). Esto diagnostica cortocircuitos trifásicos simétricos, indicando fallas catastróficas o errores de maniobra graves, como el cierre de un equipo sobre una puesta a tierra.
- **Cluster 0 (Cortocircuito Bifásico):** Representa la tipología minoritaria con un **7.3% (188 eventos)**. Su firma eléctrica evidencia un evento extremadamente violento, con corrientes que rozan los 1000A en dos fases (1010.2A y 985.2A), mientras que la tercera fase y la corriente de tierra permanecen despreciables (65.9A y 15.4A respectivamente). Esta configuración física delata un choque directo entre conductores (típicamente galopeo por ráfagas de viento) sin derivación al suelo.

El análisis de estos centroides demuestra de forma concluyente que la red de distribución no está siendo atacada primordialmente por eventos violentos de cortocircuito (que apenas suman un 15.9% en conjunto); la infraestructura está cediendo ante el desgaste operativo de desbalances y fugas a tierra menores (84.1%).

7.1.3. Correlación Causal: Patrones Temporales y Demanda Térmica

Para confirmar que el Cluster 1 obedece a un fenómeno de estrés térmico, se proyectaron las firmas diagnósticas sobre el comportamiento de ciclo diario e histórico interanual.

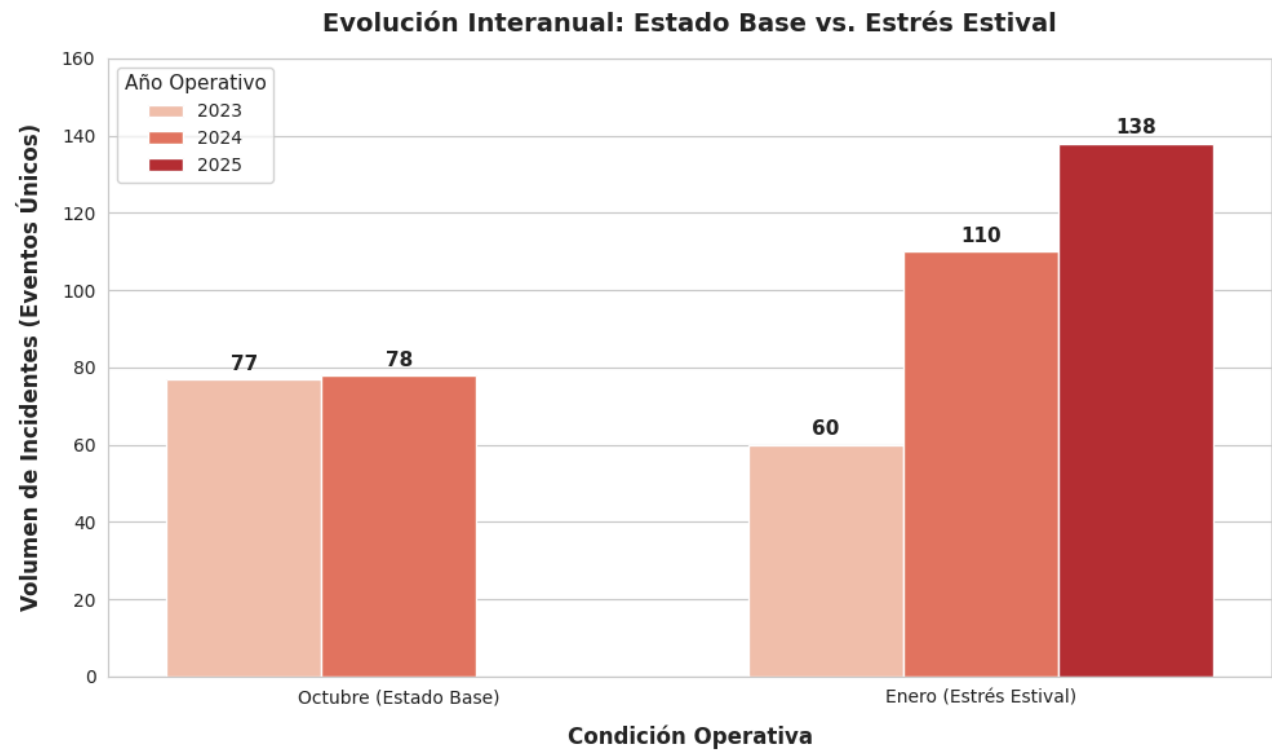


Figura 1. Evolución Interanual: Estado Base vs. Estrés Estival.

Para fundamentar esta correlación, es indispensable definir en primera instancia un **"Estado Base" o línea base de estabilidad estocástica de la red**, es decir, el comportamiento del sistema bajo condiciones climatológicas templadas y sin sobrecarga estacional. Los datos históricos revelan de manera contundente esta regularidad matemática: al analizar el mes de octubre (un período de transición térmica en la región) la siniestralidad se mantiene prácticamente idéntica e imperturbable entre años, registrando **77 eventos en octubre de 2023 y 78 eventos en octubre de 2024**. Esta paridad casi matemática (98.7% de coincidencia interanual) confirma empíricamente la existencia de un ruido basal operativo constante y controlado en el sistema eléctrico.

En contraste con este Estado Base, el cruzamiento de variables reveló que la siniestralidad de

la red se desestabiliza drásticamente al acoplarse a la demanda estival. El análisis estacional interanual (Figura 1) demostró que el inicio del ciclo de altas temperaturas estresa la capacidad instalada: de los 60 eventos totales en enero de 2023, la red escaló de forma crítica a 110 eventos en enero de 2024, alcanzando su máximo histórico absoluto con 138 incidentes únicos en enero de 2025.

Correlación de Patrones (Firma vs. Tiempo)

A. Correlación Estacional (Tipología vs. Mes)

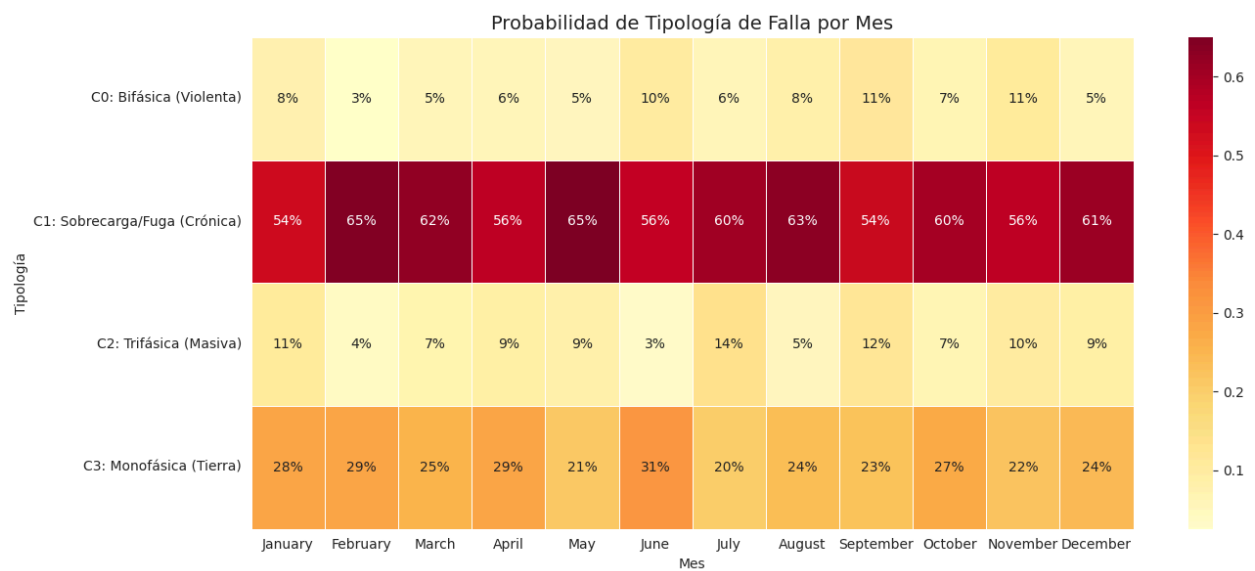


Figura 2. Probabilidad de Tipología de Falla por Mes.

Si bien el volumen absoluto de fallas explota en verano, el análisis de la densidad causal mensual (Figura 2) revela un patrón estructural crítico: **la Sobrecarga (Cluster 1) es una condición crónica del sistema durante todo el año**. La probabilidad de que un corte de suministro pertenezca a esta tipología por estrés térmico nunca desciende del 54% en ningún mes del calendario, alcanzando dominancias relativas del 65% en meses como febrero y mayo. Esto demuestra empíricamente que la red opera permanentemente con desbalances de carga significativos, independientemente de la estación.

Por su parte, las fallas físicas secundarias mantienen su propio comportamiento estacional. La Falla Monofásica a Tierra (Cluster 3) se sostiene como la segunda causa constante, presentando un ligero incremento de su probabilidad relativa hacia los meses de invierno (31% en junio y 29% en abril), posiblemente asociado a factores de humedad, vientos y mayor fricción con la vegetación. Los cortocircuitos violentos (C0 y C2) permanecen estrictamente marginales a lo largo de todo el ciclo anual.

B. Correlación Horaria (Tipología vs. Hora)

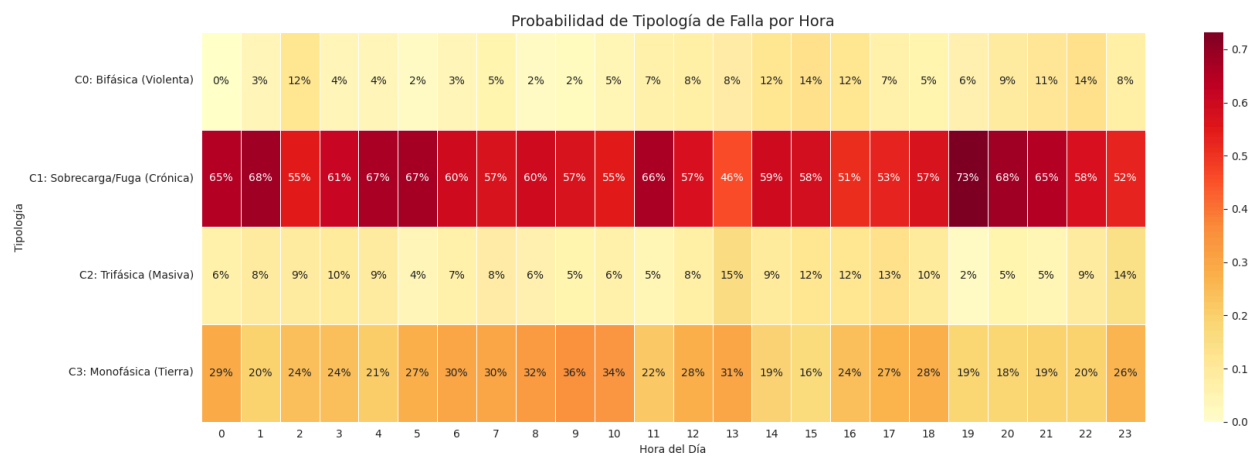


Figura 3. Probabilidad de Tipología de Falla por Hora.

Esta degradación de la red frente a la carga térmica se corrobora de forma intradiaria. El mapa de calor de las 24 horas del día (Figura 3) diferencia claramente la naturaleza de las fallas industriales frente a las residenciales. En el valle del mediodía (13:00 hs), coincidente con el pico de actividad comercial e industrial, las fallas de alta energía (cortocircuitos francos del Cluster 0 y 2) alcanzan sus máximos relativos de participación combinada (23%).

Por el contrario, a medida que el sol descende y la actividad migra a los hogares, la prevalencia de la sobrecarga domina el sistema. En la franja de las 19:00 a las 20:00 hs (momento de máxima demanda residencial por climatización e iluminación) la prevalencia del Cluster 1 se dispara drásticamente, concentrando hasta el 73% de las fallas de la red a las 19:00 hs.

Estos resultados proveen evidencia empírica contundente: el colapso de la red no se debe a inclemencias climáticas imprevistas durante el día, sino a la saturación térmica y el desbalance de fases producto del estrés de carga residencial nocturno, rompiendo diariamente y estacionalmente el estado de equilibrio estocástico demostrado en la línea base del sistema.

7.2. Ranking Geoespacial y Priorización de Activos

Una vez validada la sobrecarga térmica como la causa raíz dominante de la siniestralidad, se procedió a dimensionar su impacto geográfico y operativo. A diferencia de los enfoques tradicionales de mantenimiento basados puramente en el conteo de incidentes, esta etapa implementa una auditoría forense que pondera el volumen absoluto de fallas junto con métricas de cronicidad (Tiempo Medio Entre Eventos o MTBE) y severidad (Índice de Persistencia).

7.2.1. Análisis Macroscópico: Distribución de Pareto

El análisis espacial de la infraestructura revela que la red de media tensión no se degrada de manera uniforme, sino que obedece severamente al Principio de Pareto, exhibiendo una focalización extrema en nodos específicos del sistema.

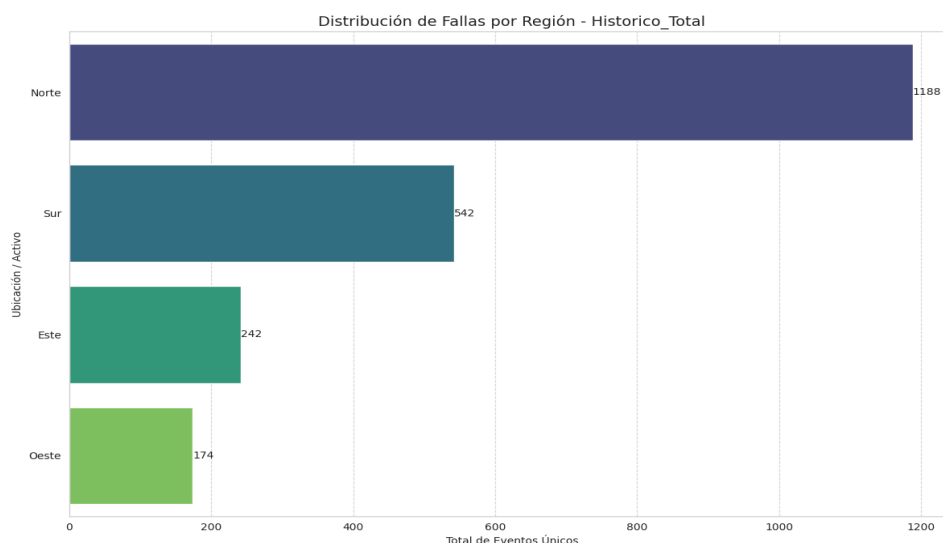


Figura 4. Distribución de Fallas por Región - Histórico Total

A nivel regional, la distribución de los incidentes demuestra una asimetría estructural contundente (Figura 4). El sistema Norte acumula la mayor siniestralidad absoluta de la red con 1.188 eventos únicos, seguido por el sistema Sur con 542 registros, sistema Este (242) y sistema Oeste (174). Sin embargo, al descender en la jerarquía espacial hacia el nivel nodal, los datos confirman que la criticidad se concentra de forma aún más drástica.

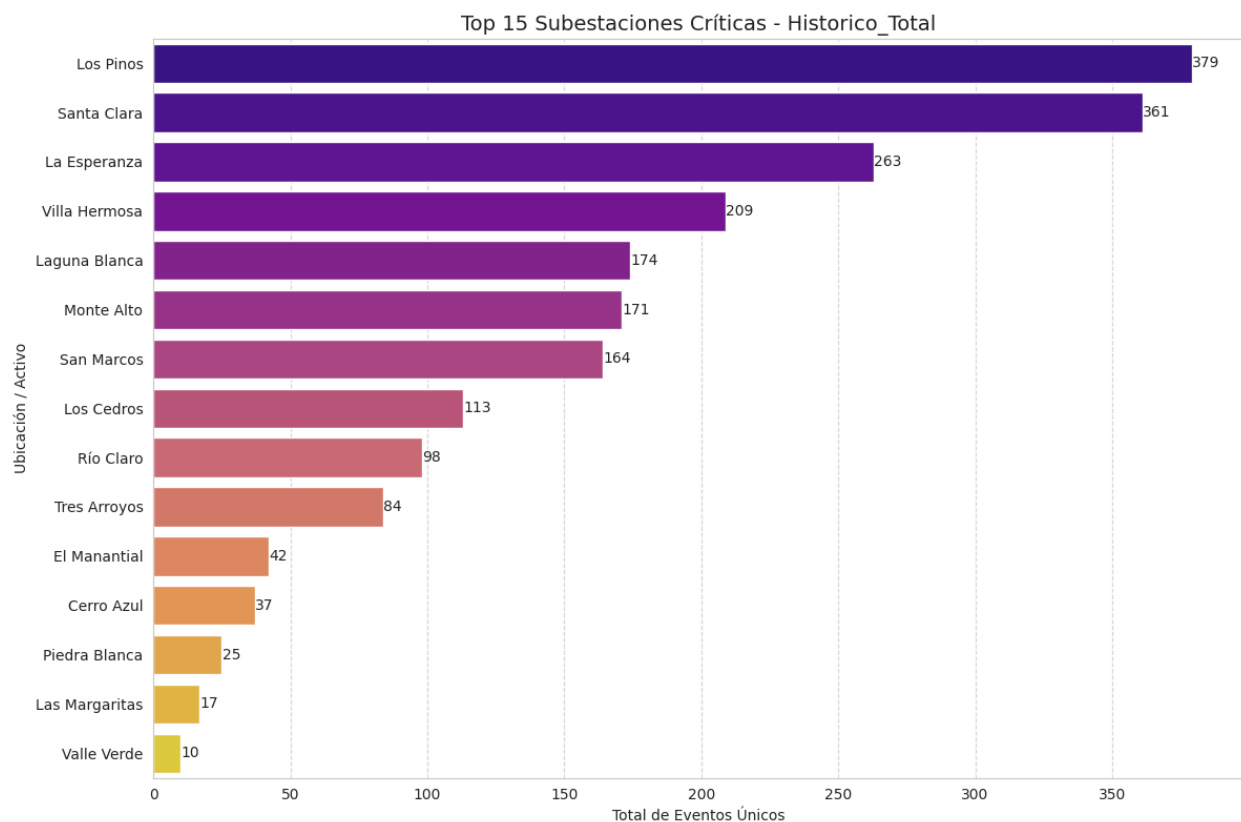


Figura 5. Top 15 Subestaciones Críticas

Como se observa en el ranking de subestaciones (Figura 5), tan solo tres nodos concentran casi la mitad de los problemas de toda la red: Los Pinos (379 eventos), Santa Clara (361 eventos) y La Esperanza (263 eventos). En conjunto, estas tres ubicaciones representan el 46.7% del total de las fallas históricas (1.003 de 2.146 eventos únicos), definiendo geográficamente las "zonas cero" para la focalización del presupuesto de mantenimiento preventivo.

7.2.2. Auditoría Forense de Reconectores (Análisis Micro)

Para traducir el diagnóstico geográfico en acciones de ingeniería concretas, se profundizó la auditoría a nivel de activo individual. Para complementar la visión macro, el análisis del comportamiento histórico de la flota completa permite aislar aquellos reconectores que concentran de manera desproporcionada la inestabilidad operativa.

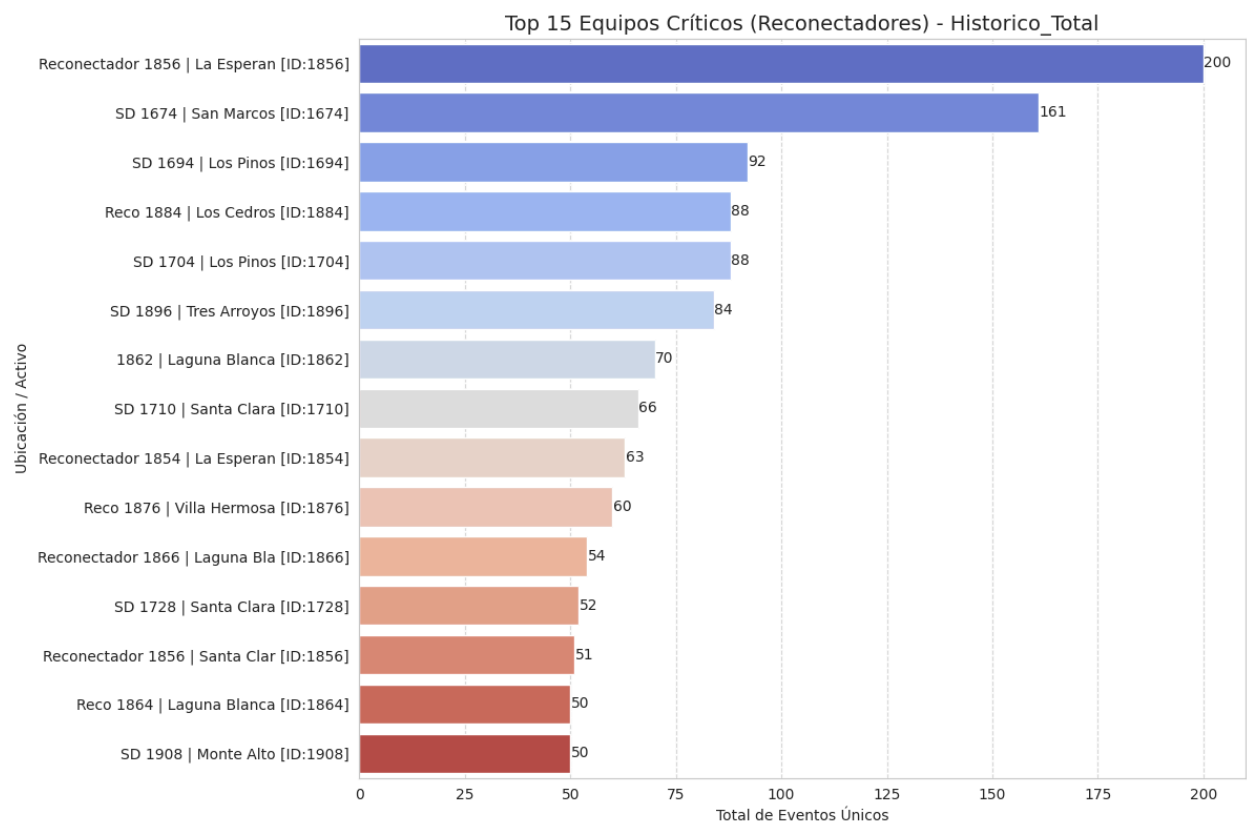


Figura 6. Top 15 Equipos Críticos (Reconectores) - Histórico Total.

El gráfico de los 15 equipos más críticos (Figura 6) expone con claridad una cola larga de decrecimiento en la frecuencia de fallas, donde los primeros cinco activos se desmarcan drásticamente del resto de la flota.

Por razones de confidencialidad y resguardo de la infraestructura crítica, la nomenclatura real de los equipos ha sido ofuscada, utilizando sus identificadores numéricos (ID) en la estructuración del ranking. El cruce de la tipología de falla con el impacto al usuario (Cortes Definitivos) permite identificar aquellos reconectores que han superado su vida útil operativa o que operan bajo regímenes térmicos críticos.

Matriz de Priorización y Auditoría del Top 5 de Activos Críticos.

Ranking	ID	Diagnóstico	Cronicidad (MTBE)	Impacto (Cortes)
#1	1856	Sobrecarga	3.7 días	74
#2	1674	Sobrecarga	6.2 días	24
#3	1704	Sobrecarga	9.9 días	80 (Máx)
#4	1896	Sobrecarga	9.7 días	40
#5	1694	Monofásica	10.9 días	33

El análisis detallado del Top 5 de activos más críticos del sistema revela las siguientes métricas particulares:

1. **Reconector [ID-1856]:** Constituye el activo de mayor criticidad histórica del sistema. Acumula 200 eventos de falla y es responsable directo de 74 cortes de suministro definitivos. Su Tiempo Medio Entre Fallas (MTBE) es de apenas 3.7 días, lo que evidencia que opera de manera crónica como un interruptor de uso regular y no como una protección excepcional. El modelo diagnostica unánimemente sus eventos como **Sobrecarga (Cluster 1)**.
2. **Reconector [ID-1674]:** Registra 161 eventos con un MTBE de 6.2 días. Si bien su volumen de actuación es elevado, genera un menor impacto proporcional en cortes definitivos (24), encontrándose diagnosticado primariamente bajo la tipología de **Sobrecarga (Cluster 1)**.

3. **Reconectador [ID-1704]:** Aunque ocupa el tercer lugar en términos de frecuencia absoluta (88 eventos), representa el activo de **mayor impacto real para la calidad de servicio de los usuarios**. Su alto Índice de Persistencia (1.4) refleja una incapacidad sistémica para despejar transitorios de la red, provocando un máximo de 80 cortes de suministro definitivos. Diagnóstico causal: **Sobrecarga (Cluster 1)**.
4. **Reconectador [ID-1896]:** Presenta 84 eventos con un MTBE de 9.7 días y un impacto de 40 cortes de suministro. Su diagnóstico apunta consistentemente a la **Sobrecarga (Cluster 1)**.
5. **Reconectador [ID-1694]:** Acumula 92 fallas con 33 cortes y un MTBE de 10.9 días. De forma notable, constituye la excepción dentro del Top 5: su diagnóstico principal recae en la tipología de **Falla Monofásica (Cluster 3)**, evidenciando un deterioro focalizado posiblemente en la vegetación circundante o en el aislamiento de una fase específica, diferenciándose del problema sistémico del resto de la flota.

Discusión de los resultados: El cruce final entre el modelado de clústeres no supervisados y el ranking forense de activos expone un hallazgo ineludible para la gestión operativa: **16 de los 20 equipos más críticos de toda la red (80%) fallan predominantemente bajo la Firma 1 (Sobrecarga/Fuga)**.

Esta fuerte correlación demuestra empíricamente que la inestabilidad de la red no responde a un comportamiento estocástico derivado de contingencias climáticas o de agentes externos impredecibles. Los reconectores están operando de forma crónica al límite de sus curvas de disparo térmico como consecuencia directa de una demanda eléctrica que excede la capacidad estructural instalada y de un desbalance sostenido entre las fases del sistema.

8. CONCLUSIONES

8.1. Validación de la Hipótesis de Investigación y Caracterización de Firmas Eléctricas

La culminación de este trabajo de grado permite confirmar de manera contundente la hipótesis de investigación planteada. Se ha demostrado empíricamente que el vector tetradimensional de corrientes ($S = [I_A, I_B, I_C, I_N]$), capturado por los reconfiguradores automáticos en el instante exacto del corte, contiene una firma física intrínseca y suficiente para revelar la tipología subyacente de las fallas en sistemas de distribución de media tensión.

Mediante el uso de aprendizaje automático no supervisado, específicamente a través del algoritmo particional K-Means optimizado con un hiperparámetro $K=4$, el modelo matemático demostró una convergencia física perfecta con las leyes fundamentales de la electrotecnia. La robustez estructural de este espacio vectorial clasificado fue auditada y validada de forma independiente mediante la topología de un modelo de Clustering Jerárquico Aglomerativo bajo el criterio de enlace de Ward. El hecho de que ambas arquitecturas algorítmicas, operando bajo principios matemáticos diferentes, alcanzaran un consenso del 96.3% en la delimitación de los grupos, certifica estadísticamente que las firmas eléctricas descubiertas constituyen estructuras densas, reales y estables de la red, y no el producto de una fluctuación estadística o un sesgo de inicialización aleatoria.

8.2. Madurez de Datos e Ingeniería de Características: Del Registro SCADA al Evento Único

Desde la perspectiva de la Ciencia de Datos aplicada a infraestructuras críticas, esta investigación comprueba la viabilidad de transformar registros transaccionales industriales crudos, degradados y no estructurados en activos analíticos limpios de alto valor prescriptivo. La implementación de un pipeline ETL fundamentado en expresiones regulares (Regex) resolvió con éxito la limitación de los sistemas SCADA heredados de la distribuidora, los cuales confinaban telemetría numérica vital dentro de strings de texto libre o campos de observaciones manuales. Este proceso rescató y estructuró un total de 2569 firmas eléctricas numéricamente válidas y protegidas bajo un rango termodinámico realista.

Asimismo, el desarrollo de una estrategia de ingeniería de características basada en ventanas temporales deslizantes de una hora resolvió el sesgo operativo de la multiplicidad de registros por recierre. Al consolidar las aperturas mecánicas sucesivas pertenecientes a una misma perturbación, se logró transicionar desde los registros brutos hacia un espacio muestral depurado de 2146 eventos únicos reales. Este procedimiento anuló la duplicación estadística que históricamente inflaba de manera artificial las tasas de siniestralidad de ciertos activos, permitiendo además computar y derivar por primera vez métricas críticas de cronicidad y severidad, tales como el Tiempo Medio Entre Eventos (MTBE) y el Índice de Persistencia del sistema.

8.3. Diagnóstico Sistémico y Patrones de Estrés Térmico Estival

El modelado analítico forense derivó en un hallazgo de alta relevancia institucional: el 58.7% de la siniestralidad total de la red de media tensión obedece de forma exclusiva a eventos de sobrecarga y desbalance térmico crónico (Cluster 1). Este resultado cuantitativo refuta empíricamente el sesgo operativo tradicional arraigado en la empresa distribuidora, el cual tendía a atribuir de forma sistemática la mayoría de las interrupciones a cortocircuitos francos y violentos (fenómenos monofásicos, bifásicos o trifásicos que en conjunto apenas representan una fracción menor del sistema) o a factores climáticos aleatorios.

Al cruzar y correlacionar estas tipologías con variables temporales, mensuales e intradiarias a través de mapas de densidad de calor, los datos brutos se convirtieron en inteligencia prescriptiva explícita. El modelo demostró que el colapso recurrente de la red se encuentra fuertemente acoplado y traccionado por la saturación térmica de la infraestructura ante el estrés de la demanda residencial nocturna (con picos de participación relativas del 73% en la franja de 19:00 a 20:00 hs), agudizándose de manera crítica durante la temporada estival hasta alcanzar su máximo histórico absoluto en enero de 2025. Este comportamiento forzado quiebra estacionalmente la línea base o "Estado Base" de estabilidad estocástica que la red registra en periodos templados como octubre, donde la siniestralidad interanual se mantiene matemáticamente idéntica.

8.4. Impacto Práctico y Prescripción para la Optimización de Activos

El impacto directo y el valor transferible de este modelo radica en dotar a la empresa distribuidora de la capacidad analítica para migrar de un paradigma de mantenimiento reactivo e ineficiente (basado en el simple conteo bruto de aperturas de los equipos) hacia una estrategia de priorización jerárquica y prescriptiva de sus activos críticos.

La aplicación del Principio de Pareto a nivel macro y micro identificó con precisión quirúrgica las áreas y equipos donde la inestabilidad operativa se concentra de forma desproporcionada. Al proveer un ranking forense detallado, el modelo prescribe la necesidad de focalizar de manera prioritaria los recursos económicos y técnicos en los **nodos críticos** de Los Pinos y Santa Clara, y de activos específicos de alto impacto y cronicidad extrema como el Reconectador [ID-1856] (cuyo MTBE crítico se determinó en apenas 3.7 días). La reasignación estratégica del presupuesto de inversión hacia estos puntos críticos permitirá mitigar la fatiga térmica acumulada en los conductores y transformadores, optimizando el ciclo de vida de los activos y traducándose en una reducción directa y cuantificable de los índices de Energía No Suministrada (ENS) en favor de la calidad de servicio de los usuarios.

9. BIBLIOGRAFÍA

El desarrollo metodológico de este Trabajo Final de Grado se basó en un enfoque eminentemente heurístico y empírico, priorizando la aplicación práctica sobre la revisión bibliográfica tradicional. La investigación se sustentó en la validación experta de campo, la documentación técnica oficial del ecosistema de desarrollo de software utilizado y la asistencia de modelos de inteligencia artificial para la optimización computacional.

A continuación, se detallan las fuentes, documentaciones y consultas que fundamentaron el trabajo:

9.1. Validación Heurística y Comunicaciones Personales

Dado que el diagnóstico de fallas a través de corrientes de recierre es un conocimiento altamente especializado, el marco teórico y la validación de las firmas eléctricas se construyeron con base empírica en el terreno.

- **Ingenieros eléctricos y técnicos operadores del sistema SCADA (Empresas Distribuidoras).** (2025-2026). *Consultas técnicas, validación heurística de firmas eléctricas, comportamiento de red y análisis de maniobras de reconectores automáticos.* Comunicaciones personales y entrevistas de campo no estructuradas.

9.2. Ecosistema de Software y Librerías Científicas

El pipeline de extracción (RegEx), procesamiento (ETL) y modelado (K-Means, Clustering Jerárquico) se desarrolló utilizando el ecosistema científico de Python, basándose íntegramente en la documentación oficial de sus respectivas librerías:

- **Python Software Foundation.** (2025). *Python Language Reference, version 3.x.* Documentación oficial del lenguaje base. Recuperado de <https://www.python.org/>
- **Pandas Development Team.** (2025). *Pandas: Powerful Python data analysis toolkit.* Documentación utilizada para la estructuración matricial, ventanas temporales y limpieza de datos. Recuperado de <https://pandas.pydata.org/>

- **Scikit-learn developers.** (2025). *Scikit-learn: Machine Learning in Python*. Documentación técnica utilizada para la implementación de StandardScaler, KMeans y AgglomerativeClustering. Recuperado de <https://scikit-learn.org/>
- **Matplotlib Development Team.** (2025). *Matplotlib: Visualization with Python*. Recuperado de <https://matplotlib.org/>
- **Seaborn Development Team.** (2025). *Seaborn: statistical data visualization*. Documentación utilizada para el renderizado de mapas de calor y matrices de densidad temporal. Recuperado de <https://seaborn.pydata.org/>
- **NumPy Developers.** (2025). *NumPy: The fundamental package for scientific computing with Python*. Documentación utilizada para operaciones matriciales y condicionales lógicas. Recuperado de <https://numpy.org/>
- **SciPy Developers.** (2025). *SciPy: Fundamental algorithms for scientific computing in Python*. Documentación utilizada para el motor matemático del agrupamiento jerárquico aglomerativo (enlace de Ward y dendrogramas). Recuperado de <https://scipy.org/>
- **Python Cryptographic Authority (pCA).** (2025). *Cryptography: Python library which exposes cryptographic recipes and primitives*. Documentación utilizada para la ingesta segura, descifrado simétrico AES (Fernet) y derivación de claves. Recuperado de <https://cryptography.io/>

9.3. Inteligencia Artificial y Asistencia Computacional

Se deja constancia del uso de modelos de inteligencia artificial como herramientas de soporte técnico y documentación durante el desarrollo de la investigación:

- **Google.** (2025-2026). *Gemini (Versión Advanced)* [Modelo de lenguaje grande]. Utilizado para asistencia en la estructuración de algoritmos de procesamiento, resolución de dependencias de librerías, generación de Expresiones Regulares (RegEx) y revisión de sintaxis avanzada en Python. Recuperado de <https://gemini.google.com/>

9.4. Redes de Media Tensión y Circuitos Eléctricos

- **CAMMESA. (2026).** *El Sistema Argentino de Interconexión y Distribución*. Compañía Administradora del Mercado Mayorista Eléctrico S.A. Recuperado de <https://portalweb.cammesa.com/>
- **Floyd, T. L. (2007).** *Principios de circuitos eléctricos* (8.^a ed.). Pearson Educación. Recuperado de https://hugochoque.com/wp-content/uploads/2022/05/Principios_de_circuitos_electricos.pdf
- **Schneider Electric. (s.f.).** *Media Tensión - Guía Técnica y Soluciones*. Recuperado de <https://www.se.com/es/es/work/products/product-launch/medium-voltage-technical-guide/>

10. ANEXOS

Este capítulo documenta el soporte técnico y algorítmico desarrollado para la ejecución de la investigación. Se incluyen los extractos de código más relevantes programados en Python (Jupyter Notebooks) y una muestra estructural del conjunto de datos procesado.

Anexo I: Extractos del Código Fuente (Python)

A continuación, se presentan las funciones modulares críticas que conforman el núcleo matemático y de procesamiento de datos del modelo.

1. Pipeline ETL y Extracción de Firmas Eléctricas (RegEx) El siguiente bloque ilustra el uso de expresiones regulares para rescatar los valores de corriente ocultos en los campos de texto libre del SCADA, así como la agrupación temporal de eventos mediante ventanas deslizantes:

```
def ejecutar_pipeline_etl(ruta_archivo_csv: str) -> pd.DataFrame:
    """
    Resumen:
        Ejecuta el ciclo completo de Extracción, Transformación y Carga (ETL).

    Detalle:
        Consolida la ingesta y depuración de datos operativos desde un archivo CSV.
        El flujo incluye el filtrado de ruido (nulos y vacíos), la extracción de
        firmas eléctricas mediante expresiones regulares con validación de rangos
        (0 a 5000), la normalización e ingeniería de variables temporales (desglose
        de fechas y mapeo de días), la estructuración de atributos técnicos del
        equipamiento y la identificación secuencial de eventos bajo ventanas de
        tiempo de una hora por instalación.
    """

    print("--- Iniciando Pipeline ETL: Diagnóstico y Priorización ---")
```

```
print("--- Iniciando Pipeline ETL: Diagnóstico y Priorización ---")

try:
    # 1. CARGA DE DATOS
    print("Fase 1.1: Ingestando archivo fuente...")
    df = pd.read_csv(ruta_archivo_csv, sep=',', on_bad_lines='skip')
    print(f"Ingesta inicial: {len(df)} registros.")

    # 2. FILTRADO DE RUIDO OPERATIVO
    print("Fase 1.2: Filtrando ruido operativo...")
    df = df.dropna(subset=['Observaciones', 'ENS'])

    # Limpieza preliminar de ENS para asegurar que no hay nulos ocultos
    df = df[df['ENS'].astype(str).str.strip() != '']
    print(f"Registros operativos válidos: {len(df)}")

    # 3. EXTRACCIÓN DE FIRMAS ELÉCTRICAS (REGEX)
    print("Fase 1.3: Extracción robusta de firmas eléctricas...")
    patron_numerico = r'(\d+\.\d*)'
    df['_temp_corrientes'] = df['Observaciones'].str.findall(patron_numerico)

    # Filtro de Integridad: Al menos 4 valores
    df = df[df['_temp_corrientes'].apply(len) >= 4].copy()

    # Asignación de columnas
    columnas_corr = ['Corr_A', 'Corr_B', 'Corr_C', 'Corr_Tierra']
    datos_corrientes = df['_temp_corrientes'].apply(lambda x: x[:4]).tolist()
    df_corrientes = pd.DataFrame(datos_corrientes, columns=columnas_corr, index=df.index)
```

```
# Conversión a numérico
for col in columnas_corr:
    df[col] = pd.to_numeric(df_corrientes[col], errors='coerce')

df = df.dropna(subset=columnas_corr)

# Lógica booleana: Primero verificamos el rango en todo el DF, luego .all(axis=1)
condicion_rango = (df[columnas_corr] >= 0) & (df[columnas_corr] <= 5000)
mask_rango = condicion_rango.all(axis=1)

df = df[mask_rango].copy()
print(f"Registros con firmas eléctricas válidas: {len(df)}")

# 4. INGENIERÍA TEMPORAL
print("Fase 1.4: Procesamiento de variables temporales...")

# Intentamos inferir el formato automáticamente para soportar 2023/2024/2025
# dayfirst=True es común en formato latino (dd/mm/yyyy)
df['Fecha'] = pd.to_datetime(df['Fecha'], dayfirst=True, errors='coerce')

# Verificación de pérdida de fechas
n_fechas_invalidas = df['Fecha'].isna().sum()
if n_fechas_invalidas > 0:
    print(f"Advertencia: Se descartarán {n_fechas_invalidas} registros por formato de fecha inválido.")

df = df.dropna(subset=['Fecha'])

# Desglose
df['Año'] = df['Fecha'].dt.year
df['Mes'] = df['Fecha'].dt.month
df['Dia'] = df['Fecha'].dt.day
df['Hora'] = df['Fecha'].dt.hour
df['Minuto'] = df['Fecha'].dt.minute

mapa_dias = {
    'Monday': 'Lunes', 'Tuesday': 'Martes', 'Wednesday': 'Miércoles',
    'Thursday': 'Jueves', 'Friday': 'Viernes', 'Saturday': 'Sábado', 'Sunday': 'Domingo'
}
df['DiaDeLaSemana'] = df['Fecha'].dt.day_name().map(mapa_dias)

# 5. NORMALIZACIÓN
print("Fase 1.5: Normalización de identificadores...")
df['ID_Instalacion'] = pd.to_numeric(df['id_Equipamiento'], errors='coerce').fillna(0)
    .astype(int)
df['SD Ubica.'] = df['Instalacion'].str.extract(r'SD\s(.*)\s*-')
df['Tipo Equ.'] = df['Equipo'].str.extract(r'^(\w+)')
df['Tension Equ. kV'] = df['Equipo'].str.extract(r'(\d+\.\d*)\s*kV').astype(float)
```

```
### CAMBIO 2026
#df['Ubic. Equ.'] = df['Equipo'].str.extract(r'\s*(.*)')
df['Ubic. Equ.'] = df['Equipo'].str.extract(r'\s*(.*)', expand=False) + " | " + df['SD
Ubica.'].

df['ENS'] = df['ENS'].astype(str).str.upper().str.replace('Sí', 'SI').str.strip()

mapa_sistema = {'N': 'Norte', 'S': 'Sur', 'E': 'Este', 'O': 'Oeste'}
df['Sistema'] = df['Sistema'].map(mapa_sistema).fillna(df['Sistema'])

# 6. AGRUPACIÓN DE EVENTOS
print("Fase 1.6: Generación de Evento_ID (Ventana 1h)...")
df = df.sort_values(by=['ID_Instalacion', 'Fecha'])
df['diff_tiempo'] = df.groupby('ID_Instalacion')['Fecha'].diff()
es_nuevo_evento = (df['diff_tiempo'].isna()) | (df['diff_tiempo'] > pd.Timedelta(hours=1
))
df['Evento_ID'] = es_nuevo_evento.cumsum()

# 7. SELECCIÓN FINAL
cols_finales = [
    'ID_Instalacion', 'Evento_ID',
    'Año', 'Mes', 'Día', 'Hora', 'Minuto', 'DiaDeLaSemana',
    'Corr_A', 'Corr_B', 'Corr_C', 'Corr_Tierra',
    'ENS', 'Actuaciones', 'Descripción',
    'Tension Equ. kV', 'Tipo Equ.', 'Ubic. Equ.', 'SD Ubica.', 'Sistema'
]

cols_disponibles = [c for c in cols_finales if c in df.columns]
df_limpio = df[cols_disponibles].copy()

print(f"\n--- ETL Completado Exitosamente ---")
print(f"Dimensiones Finales: {df_limpio.shape}")
print(f"Eventos Únicos Identificados: {df_limpio['Evento_ID'].nunique()}")

if len(df_limpio) == 0:
    print("!!! ERROR CRÍTICO: El DataFrame final está vacío. Revisar lógica de fechas o
    filtros !!!")

return df_limpio

except Exception as e:
    print(f"Error no controlado en Pipeline ETL: {e}")
    traceback.print_exc()
    return None
```

2. Entrenamiento del Modelo de Aprendizaje No Supervisado (K-Means) Este segmento detalla la configuración algorítmica utilizada para la clasificación de las tipologías de falla, asegurando la reproducibilidad mediante una semilla aleatoria fija:

```
def ejecutar_clasificacion_kmeans(df_destino: pd.DataFrame, X_matriz: np.ndarray, k_optimo: int = 4):
    """
    Resumen:
        Entrena el modelo K-Means y asigna las etiquetas de cluster al DataFrame.

    Args:
        df_destino (pd.DataFrame): DataFrame limpio ('df_limpio'). Se modificará in-place o copia.
        X_matriz (np.ndarray): Matriz de datos normalizados (X_scaled) usada para el entrenamiento.
        k_optimo (int): Número de clusters determinado por el Método del Codo (Defecto: 4).

    Returns:
        tuple: (df_etiquetado, modelo_kmeans)
            - df_etiquetado: DataFrame con la nueva columna 'Cluster_Firma'.
            - modelo_kmeans: Objeto del modelo entrenado (útil para análisis de centroides).
    """
    print(f"--- Fase 3.3: Ejecución de Clustering (K={k_optimo}) ---")

    if df_destino is None or X_matriz is None:
        print("Error Crítico: Faltan datos de entrada (DataFrame o Matriz Escalada).")
        return None, None

    try:
        # 1. Configuración y Entrenamiento del Modelo
        print(f"Entrenando algoritmo K-Means con {k_optimo} centroides...")

        # random_state=42 garantiza que los clusters siempre tengan los mismos números
        # en ejecuciones repetidas (Reproducibilidad Académica).
        kmeans_final = KMeans(
            n_clusters=k_optimo,
            init='k-means++',
            n_init=10,
            random_state=42
        )

        # Ajuste (Fit) y Predicción de etiquetas
        etiquetas = kmeans_final.fit_predict(X_matriz)

        # 2. Etiquetado del DataFrame (Integración de Resultados)
        df_etiquetado = df_destino.copy()
        df_etiquetado['Cluster_Firma'] = etiquetas

        print("¡Éxito! Clasificación completada.")
        print(f"Se ha añadido la columna 'Cluster_Firma' al dataset (Valores: 0 a {k_optimo-1}).")

        # Validación de distribución
        conteo = df_etiquetado['Cluster_Firma'].value_counts().sort_index()
        print(f"\nDistribución preliminar de eventos por cluster:\n{conteo}")

        return df_etiquetado, kmeans_final

    except Exception as e:
        print(f"Error durante el etiquetado: {e}")
        traceback.print_exc()
        return None, None
```

Anexo II: Muestra del Conjunto de Datos Limpios (df_limpio)

La siguiente tabla expone las primeras 10 filas del espacio muestral consolidado (compuesto por 2569 registros válidos correspondientes a 2146 eventos únicos). Este esquema evidencia la transición exitosa desde datos no estructurados hacia variables tabulares listas para el análisis matricial.

ID_Instalacion	Evento_ID	Año	Mes	Dia	Hora	Minuto	DiaDeLaSemana	Corr_A	Corr_B	Corr_C	Corr_Tierra	ENS	Actuaciones	Descripción	Tension Equ. kV	Tipo Equ.	Ubic. Equ.	SD Ubica.	Sistema
1674	1	2023	1	16	20	56	Lunes	550.0	700.0	640.0	0.0	NO	Corriente máxima de fase.	Disparo Reconector 13.2 kV - SD 1674 por 13.2	ReconectSD 13.2	ReconectSD 1674	San Marcos	Sur	
1674	2	2023	1	16	22	29	Lunes	55.0	505.0	480.0	1.0	SI	Corriente máxima de fase.	Disparo Reconector 13.2 kV - SD 1674 por 13.2	ReconectSD 13.2	ReconectSD 1674	San Marcos	Sur	
1674	3	2023	1	18	12	41	Miércoles	63.0	283.0	74.0	246.0	NO	Corriente máxima de tierra.	Disparo Reconector 13.2 kV - SD 1674 por 13.2	ReconectSD 13.2	ReconectSD 1674	San Marcos	Sur	
1674	4	2023	1	18	13	46	Miércoles	816.0	766.0	832.0	4.0	NO	Corriente máxima de fase.	Disparo Reconector 13.2 kV - SD 1674 por 13.2	ReconectSD 13.2	ReconectSD 1674	San Marcos	Sur	
1674	5	2023	1	20	2	11	Viernes	745.0	80.0	770.0	1.0	SI	Corriente máxima de fase.	Disparo Reconector 13.2 kV - SD 1674 por 13.2	ReconectSD 13.2	ReconectSD 1674	San Marcos	Sur	
1674	6	2023	1	21	6	31	Sábado	280.0	55.0	25.0	260.0	NO	Corriente máxima de fase y tierra.	Disparo Reconector 13.2 kV - SD 1674 por 13.2	ReconectSD 13.2	ReconectSD 1674	San Marcos	Sur	
1674	7	2023	1	30	18	41	Lunes	580.0	130.0	624.0	1.0	NO	Corriente máxima de fase.	Disparo Reconector 13.2 kV - SD 1674 por 13.2	ReconectSD 13.2	ReconectSD 1674	San Marcos	Sur	
1674	7	2023	1	30	18	54	Lunes	47.0	647.0	622.0	1.0	NO	Corriente máxima de fase.	Disparo Reconector 13.2 kV - SD 1674 por 13.2	ReconectSD 13.2	ReconectSD 1674	San Marcos	Sur	
1674	8	2023	1	31	0	20	Martes	623.0	53.0	642.0	1.0	NO	Corriente máxima de fase.	Disparo Reconector 13.2 kV - SD 1674 por 13.2	ReconectSD 13.2	ReconectSD 1674	San Marcos	Sur	
1674	9	2023	1	31	1	24	Martes	283.0	73.0	64.0	137.0	NO	Corriente máxima de fase y tierra.	Disparo Reconector 13.2 kV - SD 1674 por 13.2	ReconectSD 13.2	ReconectSD 1674	San Marcos	Sur	

(Nota: Los registros con el mismo Evento_ID, como el evento **#7**, demuestran el funcionamiento de la persistencia de fallas y la agrupación por intentos de recierre automatizados).

Anexo III: Repositorio de Código

Por motivos de extensión, el soporte computacional íntegro de la tesis no se transcribe en su totalidad. El ecosistema completo de desarrollo, que incluye los algoritmos de ingesta segura, el procesamiento de las 2569 firmas eléctricas, los scripts de generación automatizada de reportes EDA y las rutinas de validación cruzada jerárquica (Ward Linkage), se encuentra versionado y disponible para auditoría externa.

- **Enlace al repositorio:**

<https://colab.research.google.com/drive/1SljTkbrO7ERa6y6GoiI5oq-bFSFBZmFz?usp=sharing>

Nota sobre disponibilidad: En caso de indisponibilidad del enlace, una copia del código fuente (notebook) puede solicitarse a la universidad o al autor. Por motivos de confidencialidad y restricciones de la empresa proveedora, el conjunto de datos original no está disponible para su distribución ni divulgación, en su lugar se deja el conjunto de datos anonimizado.

Dedicatoria:

Mamá, gracias por apoyarme en cada paso de mi vida, en mis errores y en mis triunfos. Fue un camino difícil y, aún en los momentos donde más dudé, siempre estuviste acompañándome y diciéndome que no bajara los brazos, me cuidaste y me esperaste. Gracias por confiar en mí y ayudarme a demostrarme que las cosas pueden cambiar y mejorar, tanto para uno mismo como para los demás. Fue un trabajo de ambos, un esfuerzo de ambos, y por tanto, este título es de ambos.

Te amo.

- Lucas Villarruel