ESCUELA SUPERIOR POLITÉCNICA DEL LITORAL
Facultad de Ingeniería en Electricidad y Computación

Desarrollo de modelos de pronósticos mediante el uso de aprendizaje automático para la caracterización de desconexiones eléctricas que ocurren en la ciudad de estudio

PROYECTO DE TITULACIÓN

Previo a la obtención del título de:
Magíster en Ciencias de Datos

Presentado por:
Doménica Nicole Apolo Loaiza
Edwars Geovanny Sabando Muñiz

GUAYAQUIL - ECUADOR
Año: 2025

Dedicatoria

A Dios, a mi familia, a mis amigos, a la comarca, y a quien sostuvo mi ánimo cuando la jornada fue extensa.
Doménica Nicole Apolo Loaiza

A Dios, mi familia, y amigos, por su apoyo y amor incondicional.
Edwars Geovanny Sabando Muñiz

Agradecimientos

Queremos expresar nuestro más sincero agradecimiento a todas las personas e instituciones que hicieron posible la realización de este proyecto.

A nuestros padres, parejas y amigos por su constante apoyo, paciencia y motivación a lo largo de este proceso.

De igual manera, extendemos nuestro agradecimiento a la Escuela Superior Politécnica del Litoral (ESPOL), por las oportunidades de formación y los recursos brindados durante nuestra etapa de estudios.

Finalmente, deseamos reconocer a la empresa distribuidora de energía eléctrica que nos proporcionó los datos necesarios para el análisis y modelado, cuya colaboración fue esencial para concretar este trabajo.

Declaración Expresa

Nosotros Doménica Nicole Apolo Loaiza y Edwars Geovanny Sabando Muñiz acordamos y reconocemos que: La titularidad de los derechos patrimoniales de autor (derechos de autor) del proyecto de graduación corresponderá al autor o autores, sin perjuicio de lo cual la ESPOL recibe en este acto una licencia gratuita de plazo indefinido para el uso no comercial y comercial de la obra con facultad de sublicenciar, incluyendo la autorización para su divulgación, así como para la creación y uso de obras derivadas. En el caso de usos comerciales se respetará el porcentaje de participación en beneficios que corresponda a favor del autor o autores. El o los estudiantes deberán procurar en cualquier caso de cesión de sus derechos patrimoniales, incluir una cláusula en la cesión que proteja la vigencia de la licencia aquí concedida a la ESPOL. La titularidad total y exclusiva sobre los derechos patrimoniales de patente de invención, modelo de utilidad, diseño industrial, secreto industrial, secreto empresarial, derechos patrimoniales de autor sobre software o información no divulgada que corresponda o pueda corresponder respecto de cualquier investigación, desarrollo tecnológico o invención realizada por nosotros durante el desarrollo del proyecto de graduación, pertenecerán de forma total, exclusiva e indivisible a la ESPOL, sin perjuicio del porcentaje que nos corresponda de los beneficios económicos que la ESPOL reciba por la explotación de nuestra innovación, de ser el caso. En los casos donde la Oficina de Transferencia de Resultados de Investigación (OTRI) de la ESPOL comunique a los autores que existe una innovación potencialmente patentable sobre los resultados del proyecto de graduación, no se realizará publicación o divulgación alguna, sin la autorización expresa y previa de la ESPOL. Guayaquil, 14 de noviembre del 2025.

Ing. Doménica Nicole Apolo Loaiza

Ing. Edwars Geovanny Sabando Muñiz

Comite Evaluador

Ph.D. Miguel Torres
PROFESOR TUTOR

Ph.D. José Córdova
PROFESOR EVALUADOR

Resumen

Se presenta un estudio para caracterizar el origen de las desconexiones eléctricas en una zona urbana del litoral del Ecuador mediante aprendizaje automático. El objetivo es clasificar eventos en interna programada, interna no programada, externa programada y externa no programada e identificar patrones espaciales y factores asociados; se plantea como hipótesis que la combinación de variables operativas, de red, temporales, climáticas y de contexto socioeconómico permite discriminar dichas clases con precisión, justificándose por su utilidad para la gestión del sistema de distribución. Se compiló el histórico de interrupciones de una distribuidora de energía eléctrica en los años 2023 y 2024 y se integraron 12 variables predictoras. Se entrenaron clasificadores supervisados, Random Forest y Perceptrón Multicapa, con validación estratificada y métricas de precisión, recall y F1; la georreferenciación se efectuó sobre la ciudad de estudio. Random Forest alcanzó una precisión global cercana al 99 % y la red MLP alrededor del 99.7 %, con mejor desempeño en clases de mayor soporte. Los resultados confirman que la taxonomía del origen de las interrupciones es separable con las 12 variables consideradas y que la visualización geoespacial y la priorización por subestación y sector proveen insumos útiles para planificación y mitigación.

Palabras Clave: Aprendizaje Automático, Sistema de Distribución, Interrupciones Eléctricas, Georreferenciación, Litoral del Ecuador.

Abstract

This study presents a machine-learning approach to characterize the origin of electrical outages in an urban area on Ecuador's coastal region. The objective is to classify events into internally scheduled, internally unscheduled, externally scheduled, and externally unscheduled categories and to identify spatial patterns and associated factors; the hypothesis is that combining operational, network, temporal, climatic, and socio-economic variables enables accurate discrimination, justified by its usefulness for distribution-system management. Historical outage records from a power-distribution utility for 2023 and 2024 were compiled, and twelve predictive variables were integrated. Supervised classifiers, Random Forest and a multilayer perceptron, were trained with stratified validation and precision, recall, and F1 metrics; georeferencing was performed for the study city. Random Forest achieved an overall accuracy close to 99 %, and the MLP around 99.7 %, with higher performance for classes with greater support. The results confirm that the taxonomy of outage origin is separable using the twelve variables considered, and that geospatial visualization and prioritization by substation and sector provide useful inputs for planning and mitigation.

Keywords: Machine Learning; Distribution System; Electrical Outages; Georeferencing; Coastal Ecuador.

Abreviaturas

API Application Programming Interface

EDA Exploratory Data Analysis

ESPOL Escuela Superior Politécnica del Litoral

FMIk Frecuencia Media de Interrupción por kilómetro

K-Means Algoritmo de agrupamiento K-medias

ML Machine Learning

MLP Multi-Layer Perceptron

NLP Natural Language Processing

RF Random Forest

SCADA Supervisory Control and Data Acquisition

SVM Support Vector Machines

TF-IDF Term Frequency–Inverse Document Frequency

TTIk Tiempo Total de Interrupción por kilómetro

kVA Kilovoltio-amperio

MW Megavatio

MWh Megavatio-hora

Índice general

Índice de figuras

Índice de tablas

Capítulo 1

Introducción

1.1. Descripción del problema

La ciudad de estudio, al igual que otras grandes urbes, depende profundamente de un suministro eléctrico continuo y confiable para el desarrollo de sus actividades diarias y el bienestar de sus habitantes. La red de distribución eléctrica desempeña un papel vital en la sociedad moderna; cuando ocurre una interrupción en el servicio, una restauración rápida y adecuada es crucial para mantener la calidad del servicio y la satisfacción de los usuarios [1].

En la actualidad, las ciudades demandan un suministro eléctrico ininterrumpido, por lo que es crítico identificar y resolver las fallas eléctricas tan pronto como sea posible. Cada desconexión del servicio no planificada (apagón), puede afectar semáforos, hospitales, comercios y hogares, generando impactos sociales y económicos significativos en una urbe densamente poblada como la ciudad de estudio [2].

Las desconexiones eléctricas en sistemas de distribución pueden ser provocadas por una variedad de factores, lo que hace complejo su manejo. Entre las causas más comunes se encuentran: fenómenos climáticos adversos como tormentas eléctricas, lluvias intensas e inundaciones; contacto de la infraestructura con vegetación, como caída de árboles o ramas sobre las líneas; interferencia de animales, como aves o roedores que ocasionan cortocircuitos; fallas técnicas o de equipamiento, como averías por sobrecarga, desgaste o defectos; y errores humanos o accidentes, como colisiones vehiculares contra postes, o intervenciones de mantenimiento mal planificadas [3].

Esta multiplicidad de causas, muchas de ellas aleatorias o externas, vuelve impredecible la ocurrencia de cortes de energía con los enfoques tradicionales. Además, los eventos pueden ocurrir en cualquier parte del extenso entramado de circuitos de distribución que abastecen a la ciudad de estudio, caracterizado por una distribución geográfica amplia y condiciones urbanas heterogéneas. En consecuencia, las empresas eléctricas enfrentan el reto de diagnosticar

rápidamente la causa y ubicación de la falla para restaurar el servicio, en un contexto donde cada minuto de interrupción cuenta. Abordar este problema con los métodos actuales resulta particularmente complejo. A diferencia de las redes de transmisión de alta tensión, donde existen métodos bien establecidos para localizar fallas a lo largo de líneas relativamente lineales y con abundantes sensores, en las redes de distribución urbana la localización y diagnóstico de fallas es mucho más difícil. Esta dificultad se explica por diversas razones técnicas, entre las que destacan: la topología típicamente radial de las redes de distribución, la gran cantidad de ramificaciones y cargas conectadas, y la limitada instrumentación disponible en campo dificultan saber con precisión dónde y por qué ocurrió un corte.

Estudios previos señalan que los métodos de localización de averías desarrollados para líneas de transmisión no pueden aplicarse directamente en la distribución debido a diferencias estructurales entre ambos tipos de redes. Por ejemplo, la variabilidad de las impedancias de falla y las múltiples derivaciones en distribución complican el uso de fórmulas tradicionales de distancia de falla [4, 5].

En la práctica, la detección y atención de un apagón en distribución suele depender en gran medida de la experiencia de operadores expertos que, apoyados en alarmas SCADA o llamadas de usuarios, infiere la posible causa y despachan cuadrillas al sitio, lo cual, al tratarse de un procedimiento mayormente manual, puede implicar demoras significativas. Durante ese lapso, la población afectada permanece sin luz y el sistema eléctrico opera en estado de contingencia. Todo lo anterior evidencia que las desconexiones eléctricas en la ciudad de estudio representan un problema complejo, donde confluyen desafíos técnicos, como la identificación, localización y diagnóstico de fallas, y retos contextuales, como garantizar una respuesta oportuna para mitigar afectaciones sociales y económicas. Esta situación plantea la necesidad urgente de explorar enfoques alternativos que fortalezcan la resiliencia del sistema de distribución eléctrica frente a eventos no planificados.

1.2. Justificación del problema

Frente a la problemática descrita, existe una necesidad apremiante de desarrollar enfoques innovadores que mejoren la gestión de fallas en la red de distribución eléctrica de la ciudad de estudio. La literatura técnica reciente sugiere que la automatización en la detección, clasificación y pronóstico de interrupciones puede reducir significativamente los tiempos de respuesta y recuperación del servicio. Sin embargo, los trabajos de investigación centrados en redes urbanas de distribución que logren caracterizar y predecir desconexiones con alta precisión siguen siendo escasos. La mayoría de las soluciones tradicionales son reactivas, es decir, actúan después de ocurrido el corte, y se centran en aislar la zona afectada o

calcular la ubicación de la falla, mas no en prever dónde o cuándo ocurrirá un evento de este tipo. Incluso los sistemas de monitoreo avanzados actuales carecen de la incorporación de variables exógenas, tales como datos ambientales, geoespaciales o socioeconómicos, que podrían influir en la ocurrencia de fallas. Este vacío metodológico representa una oportunidad para aplicar enfoques de ciencia de datos, los cuales permiten extraer y descubrir patrones no evidentes mediante el análisis de históricos, de interrupciones e integración de fuentes de datos heterogéneas.

De hecho, los métodos de reconocimiento de patrones y los algoritmos de Machine Learning se han convertido en componentes importantes para la predicción de fallas en sistemas eléctricos. Aunque inicialmente estos modelos se centraban en variables eléctricas tradicionales, como corriente y voltaje, estudios recientes muestran que, al incorporar información contextual, como datos socioeconómicos, demográficos, meteorológicos e infraestructurales, se mejora significativamente la precisión de los pronósticos de desconexiones a nivel urbano [6].

Lo anterior sugiere que existe un margen amplio para innovar, integrando variables como características del barrio (densidad poblacional, ingreso socioeconómico, antigüedad de las instalaciones), condiciones del clima, e incluso descripciones textuales de reportes de averías, dentro de modelos predictivos más sofisticados. En particular, algoritmos de clasificación basados en redes neuronales han logrado identificar automáticamente causas de cortes, como caídas de árboles, descargas atmosféricas o intervención animal, a partir de las formas de onda de corriente y voltaje, descubriendo características sutiles que difícilmente serían detectados por operadores humanos [7].

El uso de estas técnicas acelera la identificación de la causa raíz de una falla, lo que a su vez permite una restauración más rápida del servicio y mejora la confiabilidad del sistema eléctrico. Además, los modelos basados en datos pueden actualizarse continuamente con nueva información, lo que les permite adaptarse a eventos inéditos, como cortes causados por tormentas atípicas o cambios en la demanda eléctrica, incrementando su precisión con el tiempo. Estos enfoques también permiten combinar información estructurada y no estructurada. Por ejemplo, variables ambientales como temperatura, lluvia o velocidad del viento han demostrado correlacionarse con determinados tipos de fallas. Asimismo, reportes textuales de cuadrillas o avisos ciudadanos pueden analizarse mediante técnicas de procesamiento de lenguaje natural (NLP), proporcionando pistas adicionales sobre la naturaleza y evolución de la avería [8].

En conjunto, estas capacidades superan las de los métodos tradicionales basados únicamente en reglas fijas o en la intuición humana, permitiendo una respuesta más inteligente y proactiva ante las desconexiones. Los beneficios esperados de la solución propuesta abarcan tanto el ámbito técnico-operativo como el socioeconómico.

1.3. Objetivos

1.3.1. Objetivo General

Desarrollar modelos de clasificación usando técnicas de aprendizaje automático, orientados a la caracterización del origen de las desconexiones eléctricas en una zona urbana del litoral del Ecuador, a la identificación de patrones espaciales y factores socioeconómicos asociados, proporcionando información útil para la mejora de la gestión del sistema eléctrico.

1.3.2. Objetivos Específicos

1. Estudiar estadísticamente los datos históricos de interrupciones eléctricas de una de las ciudades principales de Ecuador mediante el preprocesamiento de la base de datos de una empresa distribuidora de energía eléctrica.

2. Evaluar los modelos de aprendizaje supervisado para la clasificación y pronóstico de las desconexiones eléctricas, valorando las medidas de desempeño.

3. Desarrollar una herramienta de georreferenciación para la distribución espacial de las desconexiones eléctricas, su caracterización y clasificación.

1.4. Metodología

El desarrollo del proyecto se basó en una metodología estructurada que integra las etapas de ingeniería de datos, análisis exploratorio y modelado predictivo, complementadas con técnicas de aprendizaje automático y georreferenciación. La base del análisis estuvo conformada por los informes de desconexiones eléctricas, proporcionados por la empresa de distribución de energía eléctrica de la ciudad de estudio, correspondientes a los años 2023 y 2024, los cuales registran de manera detallada los eventos de interrupciones eléctricas, sus causas y características operativas.

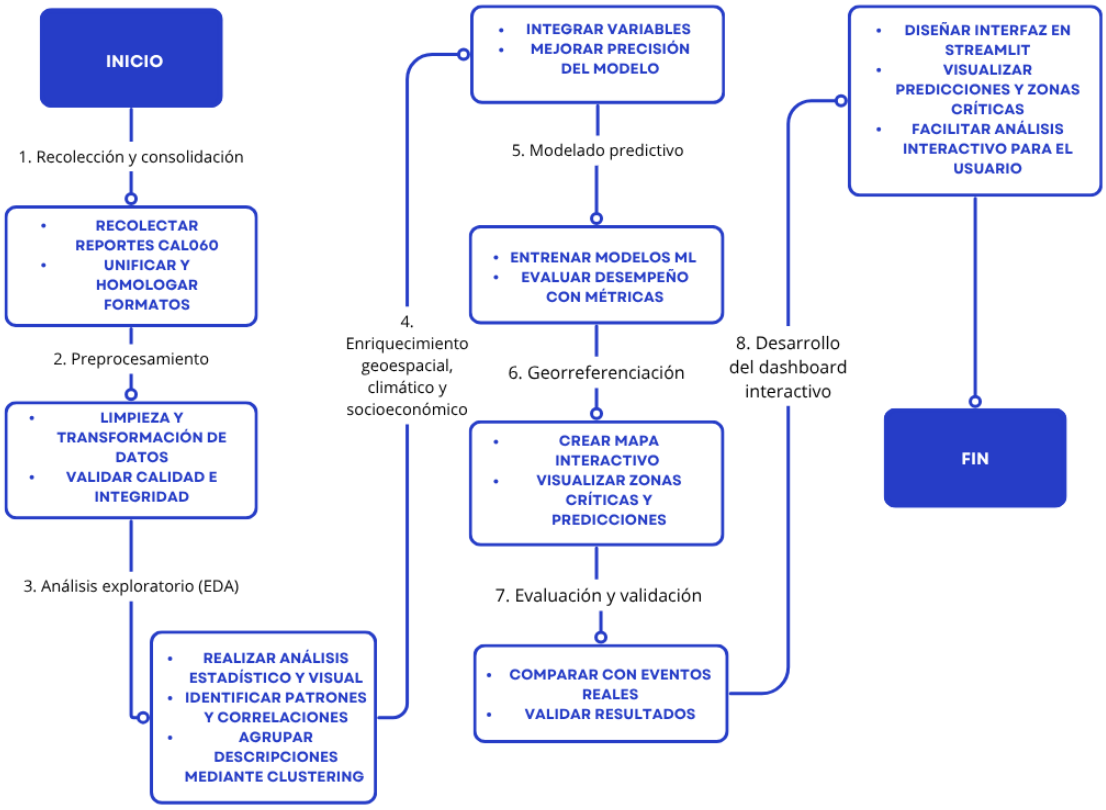

Figura 1.1: Diagrama de flujo de la metodología

1.4.1. Recolección y consolidación de datos

Se recopilan reportes de desconexiones eléctricas de la empresa distribuidora de electricidad, los cuales contienen los registros mensuales de interrupciones eléctricas, y, además, fueron consolidados en una base unificada. Esta etapa implica la homologación de los formatos recibidos desde distintas unidades de negocio, así como la integración de metadatos asociados, como, por ejemplo, la ubicación geográfica de los eventos, las características de las instalaciones, y las condiciones climáticas y demográficas del sector afectado.

1.4.2. Preprocesamiento de datos

Se aplican técnicas de limpieza y transformación de datos usando herramientas como Python (pandas, NumPy) para corregir inconsistencias, eliminar duplicados, imputar valores faltantes y estandarizar variables. Se valida, también, la integridad y calidad de los datos para garantizar la confiabilidad del análisis posterior.

1.4.3. Análisis exploratorio de datos (EDA)

Se realiza un análisis estadístico y visual para identificar distribuciones, tendencias temporales, patrones de frecuencia, zonas geográficas recurrentes y correlaciones entre variables. Este análisis permite reconocer posibles factores determinantes de las desconexiones mediante herramientas como gráficos estadísticos, mapas de calor y técnicas de reducción de dimensionalidad. Complementariamente, se efectuó un agrupamiento no supervisado (clustering) sobre las descripciones textuales de las interrupciones con el fin de detectar patrones semánticos en los reportes de fallas. Para ello, se aplicó una representación vectorial de texto basada en Term Frequency–Inverse Document Frequency (TF-IDF) y en algoritmos de agrupamiento (K-Means), permitiendo agrupar descripciones con vocabulario y contexto operativo similar. Este procedimiento permitió identificar categorías generales de fallas y complementar el análisis exploratorio con una dimensión lingüística, que posteriormente facilitó la interpretación de los eventos y la coherencia del conjunto de datos previo al modelado predictivo.

1.4.4. Enriquecimiento geoespacial, climático y socioeconómico

El conjunto de datos fue complementado con información ambiental, temporal y socioeconómica para ampliar el contexto de las interrupciones eléctricas. Se integraron variables climáticas obtenidas de la plataforma Visual Crossing Weather Data [9]. para la ciudad de estudio (2023–2024), incorporando el estado general del tiempo y descripciones detalladas del mismo, con el fin de relacionar los eventos eléctricos con condiciones meteorológicas como nubosidad o lluvias. Además, se añadieron variables temporales derivadas del calendario, como el día de la semana, la condición de feriado nacional y un indicador de fin de semana, generadas mediante la librería holidays de Python. Estas variables permitieron identificar patrones operativos y estacionales que influyen en la ocurrencia de desconexiones. En el componente socioeconómico, se incorporó un índice de peligrosidad, basado en los resultados presentados en [10] sobre la peligrosidad por zonas en la ciudad de estudio. Este índice se asignó de forma cualitativa: valor 10 para zonas de alta peligrosidad, 6 para zonas de menor riesgo y 7 cuando no existía una referencia específica.

1.4.5. Modelado predictivo

El modelado predictivo se desarrolló bajo un enfoque de aprendizaje supervisado, tomando como variable objetivo el origen de la interrupción eléctrica (interna o externa, programada o no programada). Se implementaron modelos de clasificación multiclase, específicamente Random Forest (RF) y una Red Neuronal Multicapa (MLP), empleando como variables de

entrada los atributos técnicos de los eventos, junto con las variables contextuales obtenidas del proceso de enriquecimiento geoespacial, climático y socioeconómico. El procedimiento incluyó la preparación de datos, con la codificación de variables categóricas, la normalización de las variables numéricas y la división del conjunto de datos en entrenamiento y prueba. Ambos modelos fueron configurados para ajustarse a la naturaleza del problema, garantizando la consistencia entre las fases de preprocesamiento, entrenamiento y validación.

1.4.6. Herramienta de georreferenciación

Se desarrolló una herramienta de visualización geográfica interactiva que permite representar espacialmente los eventos de desconexión eléctrica en la ciudad de estudio. Para ello, se realizó un proceso de geocodificación automática mediante la interfaz de programación de aplicaciones (API) Nominatim de OpenStreetMap, a partir de las direcciones, sectores y cantones reportados en el dataset, obteniendo coordenadas de latitud y longitud asociadas a cada registro. La herramienta fue implementada en Python utilizando la librería Folium, lo que permitió generar un mapa dinámico que muestra la distribución espacial de las interrupciones, diferenciadas por tipo de origen (interna o externa, programada o no programada) y complementadas con las predicciones del modelo de red neuronal perceptrón multicapa (MLP). Cada punto del mapa contiene información contextual del evento, como la subestación, la fecha, la descripción de la interrupción y la coincidencia entre la clasificación real y la predicha. Esta visualización georreferenciada facilita la identificación de zonas críticas y patrones espaciales de fallas eléctricas, permitiendo analizar recurrencias por sectores y contrastar las áreas de mayor peligrosidad o vulnerabilidad. Asimismo, constituye una herramienta útil para el diagnóstico y toma de decisiones operativas por parte de empresas distribuidoras o entidades de planificación eléctrica.

1.4.7. Evaluación y validación del modelo

La evaluación y validación de los modelos se llevó a cabo mediante un enfoque cuantitativo y cualitativo. En el primer caso, se aplicaron métricas de desempeño como la precisión (accuracy), la exhaustividad (recall) y el F1-score, junto con el análisis de la matriz de confusión para examinar el comportamiento del modelo en cada clase. Este proceso permitió verificar la estabilidad del modelo y su capacidad de generalización frente a nuevos datos. En el ámbito cualitativo, las predicciones se contrastaron con eventos reales registrados en los informes de la empresa de distribución de energía eléctrica y se contó con la retroalimentación técnica de especialistas, con el fin de validar la coherencia de las clasificaciones generadas. Asimismo, se documentaron los procedimientos utilizados para comparar modelos, identificando las

configuraciones más adecuadas para la caracterización de interrupciones eléctricas y su aplicación potencial en la planificación operativa.

1.4.8. Aplicación interactiva de visualización de resultados

Como fase final del proyecto, se desarrolló una aplicación web interactiva tipo dashboard, diseñada para facilitar la consulta, visualización y análisis dinámico de los resultados obtenidos en las etapas anteriores. Esta herramienta integra las funciones de los modelos predictivos, la visualización geoespacial y la exploración de los datos enriquecidos, permitiendo al usuario acceder a una interfaz centralizada y amigable. La aplicación fue implementada en Python, utilizando las librerías Streamlit, Plotly, Pandas y Folium, lo que permitió combinar visualizaciones interactivas con gráficos estadísticos y mapas dinámicos. El entorno ofrece diferentes secciones para la exploración de datos históricos, la visualización de predicciones georreferenciadas y la comparación de resultados reales y estimados, de acuerdo con filtros definidos por el usuario, como tipo de interrupción, sector geográfico o período de tiempo. Además, el dashboard permite generar indicadores resumen sobre la frecuencia y duración de las interrupciones, los tipos de fallas más comunes y la distribución de eventos por zona. Esta herramienta constituye un módulo de apoyo a la toma de decisiones operativas, al facilitar el análisis visual del comportamiento de la red eléctrica y de los patrones espaciales asociados a los eventos de desconexión, sirviendo como base para estrategias de mantenimiento preventivo y priorización de recursos.

1.4.9. Resultados Esperados

Al finalizar el proyecto, se espera disponer de una aplicación con un modelo de clasificación validado que permita identificar el origen de las desconexiones eléctricas registradas en la ciudad de estudio, diferenciando entre causas internas y externas, así como entre eventos programados y no programados. Estos modelos deberán presentar un desempeño consistente y verificable, evaluado mediante métricas estándar como accuracy, recall y F1-score. Asimismo, se prevé el desarrollo de una visualización geoespacial interactiva que represente la distribución y frecuencia de las interrupciones, permitiendo identificar zonas críticas o de alta vulnerabilidad dentro de la red de distribución eléctrica. Esta visualización facilitará el análisis de patrones espaciales y temporales, asociados a factores técnicos, climáticos y socioeconómicos.

Como resultado complementario, se espera contar con un conjunto de datos unificado y enriquecido, que integre fuentes geográficas, socioeconómicas y ambientales en un formato estructurado y optimizado para su análisis con técnicas de ciencia de datos. Esta base consolidada servirá como insumo para futuras investigaciones y para la mejora continua

del sistema propuesto. Finalmente, se contempla la implementación de una aplicación web tipo dashboard, desarrollada en Python (Streamlit), que integre los resultados del modelo predictivo, las visualizaciones geoespaciales y los indicadores de desempeño. Esta herramienta permitirá a los usuarios explorar los datos de manera dinámica, consultar predicciones en tiempo real y apoyar la toma de decisiones operativas, orientadas a optimizar el mantenimiento preventivo, priorizar inversiones y mejorar la gestión técnica de la red eléctrica en la ciudad de estudio.

1.4.10. Dataset

El dataset, o conjunto de datos, utilizado corresponde a los reportes de desconexiones eléctricas de la empresa distribuidora de energía eléctrica de la ciudad de estudio, generados mensualmente por la empresa eléctrica para la ciudad de estudio. Estos reportes cuentan con información detallada sobre las interrupciones del servicio eléctrico, incluyendo datos técnicos, operativos y contextuales. Para el presente estudio, se recopilaron los reportes comprendidos entre enero de 2023 y diciembre de 2024, abarcando un período de análisis de dos años completos.

Cada archivo mensual contiene múltiples registros que describen eventos de desconexión eléctrica, con variables como: fecha y hora de ocurrencia, alimentador afectado, ubicación (zona/parroquia), duración del evento, causa o código de interrupción, tipo de interrupción (programada/no programada, interna/externa), entre otros. El conjunto de datos también incluye una columna clave denominada “Origen de Interrupción”, la cual fue utilizada como variable objetivo en los modelos de clasificación supervisada desarrollados en este estudio. Durante el desarrollo, se obtuvo un total de aproximadamente, 6282 registros limpios tras aplicar las etapas de depuración inicial, lo cual representó el conjunto de datos base para el modelado. Sin embargo, este número pudo variar dependiendo del criterio de limpieza adoptado y del tratamiento aplicado a los datos faltantes o inconsistentes. El formato original de los archivos es Excel (.xlsx) y fue transformado a un formato estructurado tipo DataFrame para su procesamiento en Python, utilizando bibliotecas como pandas y NumPy. Adicionalmente, el conjunto de datos fue enriquecido con variables externas de tipo geográfico, climático y socioeconómico, que se integraron para mejorar el poder explicativo de los modelos desarrollados.

En resumen, se trata de un conjunto de datos heterogéneo, que combina variables estructuradas y no estructuradas, con alto potencial para tareas de clasificación y análisis espacial. Su naturaleza compleja requirió una cuidadosa etapa de preprocesamiento antes de ser empleado en los modelos de aprendizaje automático.

Tabla 1.1: Tabla de Detalle de Columnas en Conjunto de Datos

Categoría	**Nombre de la Variable**	**Tipo de Dato**
1. Identificación	Código de Interrupción	string
	Indicador de Mantenimiento o Falla	string
	Etapa funcional en la que se presentó la falla	string
2. Ubicación de la Falla	Instalación / Equipo donde se presentó la falla	string
	Provincia (Falla)	string
	Cantón (Falla)	string
	Sector (Falla)	string
	Ubicación Estimada de la Falla	string
	Propiedad de la Instalación / Equipo donde se presentó la falla	string
	Protección que Operó	string
	Tipo de protección que actuó	string
3. Ubicación de la Interrupción	Etapa funcional en la que se presentó la interrupción de servicio	string
	Instalación / Equipo donde se presentó la interrupción	string
	Provincia (Interrupción)	string
	Cantón (Interrupción)	string
	Sector (Interrupción)	string
4. Detalle del Alimentador Afectado	Línea de Subtransmisión	string
	Subestación	string
	Alimentador primario	string
	Tipo de Alimentador primario	string
	Nivel de Tensión (kV)	float

Continúa en la siguiente página

Categoría	**Nombre de la Variable**	**Tipo de Dato**
	Nivel de afectación de la interrupción a la Red	string
5. Profundidad de la Interrupción	Origen de Interrupción	string
	Causa de Interrupción	string
	Catálogo de Interrupciones	string
	Descripción de Interrupción	string
	Potencia Nominal Instalada del Alimentador (kVA)	float
	Potencia Nominal Fuera de Servicio (kVA)	float
	Potencia Nominal Fuera de Servicio (MW)	float
	Carga Fuera de Servicio (kVA)	float
	Energía No Suministrada (MWh)	float
	Fecha Inicio de Interrupción (dd:mm:aa)	string
	Hora Inicio de Interrupción (hh:mm)	string
	Fecha Fin de Interrupción (dd:mm:aa)	string
	Hora Fin de Interrupción (hh:mm)	string
	Duración de Interrupción (Horas:minutos:segundos)	string
	Duración de Interrupción (Horas)	float
6. Índices	FMIk	float
	TTIk	float

Capítulo 2

Estado del arte

2.1. Conceptos clave sobre fallas en sistemas eléctricos de distribución

Los sistemas eléctricos de distribución son fundamentales para transportar energía desde las subestaciones hasta los usuarios finales. Su topología radial, la gran cantidad de componentes y su exposición a condiciones ambientales variables incrementan significativamente la probabilidad de fallas [1].

En el contexto de los sistemas eléctricos de distribución, las fallas pueden clasificarse ampliamente en **fallas técnicas** y **fallas no técnicas**. Las fallas técnicas incluyen eventos como cortocircuitos, fallas de aislamiento o sobrecargas, y son generalmente originadas por condiciones operativas anómalas o degradación de equipos. Por otro lado, las fallas no técnicas comprenden eventos como el contacto con vegetación, interferencia de animales, accidentes vehiculares o condiciones climáticas extremas. Esta distinción es clave porque las técnicas de diagnóstico requieren abordar ambas categorías, las cuales presentan señales y patrones de comportamiento diferentes en los datos operativos del sistema [2].

Entre las causas técnicas más comunes de fallas en sistemas de distribución se incluyen cortocircuitos originados por deterioro del aislamiento, fallas en interruptores y transformadores, sobrecargas prolongadas y defectos de conexión. Las causas no técnicas más recurrentes están asociadas a eventos externos como la caída de ramas, contacto de animales con los conductores, impactos de vehículos y efectos del viento o lluvia intensa. Estas fallas suelen presentarse con mayor frecuencia en zonas rurales y en redes aéreas debido a su mayor exposición ambiental y menor densidad de sensores de protección [1].

La identificación oportuna y precisa de la causa de una falla es fundamental para mejorar la confiabilidad operativa, reducir el tiempo de respuesta y orientar estrategias de mantenimiento

preventivo. Mientras que las fallas técnicas pueden localizarse mediante mediciones eléctricas, las fallas no técnicas exigen herramientas más sofisticadas que analicen múltiples fuentes de datos y permitan establecer relaciones entre variables ambientales, geográficas y sociales. La clasificación de fallas, según su causa, no solo apoya la restauración eficiente del servicio, sino que también permite segmentar la red en zonas críticas y priorizar inversiones en infraestructura [6].

2.2. Aplicaciones de ciencia de datos y machine learning en sistemas eléctricos

El avance en la digitalización de los sistemas eléctricos ha permitido recolectar grandes volúmenes de datos operacionales en tiempo real, lo que ha favorecido la adopción de técnicas de aprendizaje automático para mejorar la operación, mantenimiento y diagnóstico en redes de distribución. Los algoritmos de *machine learning* permiten construir modelos que capturan relaciones complejas en los datos, facilitando tareas como la predicción de fallas, la clasificación de sus causas y la detección de patrones anómalos. Estas técnicas han sido aplicadas tanto a datos históricos como a flujos en línea, y su eficacia ha sido validada en numerosos estudios recientes [3].

Una de las aplicaciones más destacadas ha sido la detección de pérdidas no técnicas, las cuales representan un problema económico crítico para las empresas distribuidoras. La utilización de modelos supervisados como *Random Forest* y *Gradient Boosting* ha demostrado un desempeño notable al detectar este tipo de eventos a partir de datos de medidores inteligentes, alcanzando métricas de precisión elevadas incluso ante bases de datos desbalanceadas [11].

La predicción de interrupciones eléctricas causadas por factores meteorológicos y estructurales también ha sido abordada mediante redes neuronales profundas, integrando datos climáticos, características de la infraestructura y variables socioeconómicas. Estas técnicas permiten anticipar eventos en zonas específicas, adaptándose a las vulnerabilidades particulares de cada comunidad y mejorando la gestión preventiva de la red [6].

Por otra parte, se han aplicado técnicas de clasificación combinadas con minería de datos para identificar causas específicas de fallas, como contacto con vegetación o interferencia animal. Este enfoque, basado en reglas de asociación y árboles de decisión, permite analizar registros históricos para extraer patrones frecuentes y mejorar las estrategias de mantenimiento predictivo [7].

2.3. Técnicas de clasificación y predicción relevantes

Durante años, en áreas importantes como la salud, los especialistas han tratado con una gran cantidad de datos de los pacientes, los cuales han permitido tomar decisiones relevantes que impactaron en el tratamiento de dichos pacientes. Sin embargo, el análisis realizado en estos datos ha sido breve, debido a las limitaciones humanas para considerar toda la información recopilada. Para solventar estas limitaciones y mejorar los análisis, se ha implementado el aprendizaje de máquina como apoyo para la detección y predicción de posibles problemas basado en los datos clínicos.

El aprendizaje de máquina, o Machine Learning, se divide en paradigmas, cuya diferencia principal es el modo de búsqueda de patrones. Dos de los paradigmas más importantes han sido: el aprendizaje supervisado y el aprendizaje no supervisado. El aprendizaje supervisado se basa en la búsqueda de patrones mediante características conocidas y una variable objetivo existente; sus tareas más comunes son la regresión, para predecir valores continuos, y la clasificación, para la predicción de categorías. En cambio, el aprendizaje no supervisado se basa en la búsqueda de patrones sin tener una variable objetivo previamente definida; sus tareas más frecuentes son el clustering, y la detección de anomalías [12].

En la clasificación, existen varias técnicas que se pueden utilizar, siendo las más usuales la regresión logística, las SVM (Support Vector Machines), y los métodos de conjunto (ensemble methods). Otras técnicas de clasificación son las redes neuronales y el aprendizaje profundo, las cuales han tenido apogeo los últimos años debido a su capacidad de aprendizaje usando representaciones complejas de datos y utilizando sistemas computacionales con una gran capacidad de procesamiento [13].

Si bien la efectividad de cada técnica es dependiente de la precisión del modelo final para el problema que se desea resolver, se ha demostrado que, en áreas como la medicina, las redes neuronales y las máquinas de vectores de soporte (SVM) mostraron ser las mejores en comparación a las demás a pesar de su complejidad [14]. Los modelos de SVM se basan en la clasificación binaria, pues su funcionamiento consiste en la separación de elementos en dos grupos según las propiedades de cada uno, además de tener la capacidad de trabajar con datos de dimensionalidad alta. Por este motivo, se considera las SVM adecuadas para modelos de clasificación que operan con enormes conjuntos de datos.

Las redes neuronales, como indican su nombre, simulan un funcionamiento parecido al de los seres humanos, pues se trabajan con elementos conocidos como “neuronas”, los cuales procesan los datos obtenidos (texto, audio, imagen, etc.) por medio de varias “capas”, hasta dar con un resultado. Las redes neuronales, adicionalmente, son ampliamente utilizadas para el procesamiento de lenguaje natural, reconocimiento facial, y reconocimiento de voz [15].

2.4. Herramientas, software y librerías utilizadas

En las últimas décadas, el aprendizaje de máquina ha sido implementado de diferentes maneras, pues depende en gran medida de los recursos a disposición, así como también del modo de solucionar un problema existente.

La forma en la que varían estas implementaciones incluye los lenguajes de programación, el software empleado, los algoritmos seleccionados para el aprendizaje de máquina, las plataformas o herramientas en la que se realizará el aprendizaje, y las librerías de programación utilizadas.

En estudios relacionados con la arquitectura, por ejemplo, se han usado lenguajes de programación como Java, Python, Matlab, y C#, de entre ellos, siendo Python el más utilizado. Además, en varios de estos estudios, se trabajaron con librerías tales como Scikit-Learn y Pytorch, las cuales son utilizadas en el aprendizaje de máquina y el aprendizaje profundo (Deep learning) [16].

La librería Scikit-Learn contiene varios métodos que ofrecen apoyo durante el preprocesamiento de datos, división de datos en conjunto de entrenamiento y de prueba, implementación de algoritmos comunes de clasificación y regresión, selección y evaluación de modelos predictivos, etcétera [17].

La librería Pytorch, gracias a su facilidad de uso, permite la experimentación, despliegue, y pruebas de soluciones enfocadas en el aprendizaje profundo. Entre sus métodos más conocidos, se encuentran: la creación y operaciones entre tensores (matrices) de cualquier dimensionalidad, la definición de estructura y entrenamiento de redes neuronales, la definición de función de pérdida y optimizador, entre otros [18].

2.5. Análisis de trabajos previos

En la investigación [5], se utilizaron redes neuronales con perceptrones multicapa (MLP) para resolver un problema de detección de tipo de fallas en sistemas distribuidos; lo cual, si bien se puede considerar un caso de estudio relativamente antiguo, puede sentar las bases para resolver, de manera efectiva, problemas relacionados a la clasificación en el área de electricidad.

En adición, en el artículo [4], se ha demostrado que la predicción de fallas eléctricas con ayuda de la inteligencia artificial, frente a otros métodos evaluados en dicho estudio, una fiabilidad y precisión altas. Señalaron, además, como sus desventajas principales, la necesidad de un conjunto de datos constante y en actualización, una implementación compleja y costosa, y una velocidad baja.

En otro estudio, [19], se propuso un algoritmo combinado a partir de otros algoritmos de clasificación para solucionar problemas de identificación de fallas en líneas de transmisión eléctrica. Uno de los puntos más importantes de la investigación fue el proceso de selección de características, o features, los cuales se encuentran se relacionan fuertemente con el objetivo de la predicción (la falla ocurrida) y estas características no se encuentran relacionadas entre sí, lo que implica que este proceso fue realizado de forma correcta.

Además, en el artículo [20], de la misma manera que en el estudio anterior, se sugirió un algoritmo de clasificación por medio de realizar una combinación, esta vez usando redes neuronales convolucionales y redes neuronales artificiales, con el fin de identificar patrones que permitieran determinar la causa de cortes de luz utilizando datos en forma de ondas. Dicho algoritmo combinado demostró un mejor desempeño en la clasificación que otros utilizados en el estudio, y aún más al utilizar la técnica de data augmentation.

Por otro lado, en la investigación [21], se tomó el concepto de la clasificación de texto, el cual implica agrupar datos en formato de texto. Se abordaron los métodos más utilizados para ello, incluyendo también el uso de redes neuronales, y una parte importante de dicha investigación fue el descubrimiento de desafíos a los que cada método se enfrentaba, con lo que es posible determinar que los métodos de clasificación no son perfectos, mas es posible enfrentar dichos desafíos para que los modelos resulten siendo lo más eficientes posibles.

Finalmente, y con relación a la clasificación de texto, se consideró el estudio [22], el cual no solo incluyó redes neuronales como uno de los clasificadores base en su investigación, sino que también lo agregó en comparación junto a otras técnicas de clasificación de aprendizaje de máquina, en especial un método que combinaba redes neuronales con árboles de decisión. Al final, se determinó que el método combinado tuvo mejores resultados en precisión, pero con margen de mejora en la velocidad de procesamiento, y, además, se concluyó que los árboles de decisión, en adición a su popularidad en la creación de modelos de clasificación, resultan eficaces para acelerar el entrenamiento en modelos basados en redes neuronales.

Capítulo 3

Diseño e implementación del modelo

3.1. Recolección y preprocesamiento de datos

La construcción del conjunto de datos se realizó a partir de veinticuatro archivos en formato Excel, correspondientes a registros mensuales de interrupciones eléctricas. La información se extrajo de la segunda hoja de cada archivo, en la que se encontraba el detalle de los eventos. Antes de la integración, se procedió a la unificación y estandarización de los encabezados, fusionando las filas superiores que contenían títulos fragmentados y asignando nombres de columna consistentes. Una vez integrados todos los archivos, el conjunto inicial contenía 866 601 filas y 39 columnas, incluyendo tanto registros completos como incompletos. A este se le aplicaron los siguientes procesos de limpieza y transformación:

- **Eliminación de registros incompletos:** Realizado mediante el descarte de filas con valores nulos en columnas críticas como: código de interrupción, detalle de la línea de subtransmisión, origen y causa de la interrupción.
- **Imputación de valores faltantes:** Realizado en la columna "Observación" (relleno con cadena vacía) y en variables numéricas como potencias y energía no suministrada (relleno con cero).
- **Depuración semántica:** Mediante la eliminación de registros cuya descripción de la interrupción contenía términos fuera del alcance del análisis, como "racionamiento".
- **Eliminación de columnas redundantes y reordenamiento:** Realizado en algunas variables para priorizar la información técnico-operativa.
- **Creación de variables derivadas:** "Tipo de Falla" (clasificación en técnicas y no técnicas) y "Bloque Horario" (segmentación temporal del evento).

Tras este proceso, el conjunto depurado (df_final) quedó compuesto por 6282 filas y 36 columnas, lo que representa una reducción sustancial del volumen inicial y concentra únicamente la información relevante y completa para el análisis y modelado posteriores.

3.2. Análisis exploratorio de datos (EDA)

El análisis exploratorio de datos se realizó con el objetivo de comprender la estructura del conjunto, identificar patrones y posibles anomalías, y evaluar la distribución de la variable objetivo. En primer lugar, se examinó la variable Tipo de Falla, obtenida a partir de la clasificación del "Origen de Interrupción" en técnicas y no técnicas. Los resultados mostraron que, del total de 6282 registros, 4728 correspondían a interrupciones técnicas (75.3 %) y 1554 a interrupciones no técnicas (24.7 %), evidenciando un desbalance moderado en las clases. Se efectuó, además, un análisis univariado de las principales variables numéricas:

- **Duración de Interrupción (horas):** Se observó una distribución sesgada a la derecha, con la mayoría de los eventos concentrados en duraciones cortas y la presencia de valores atípicos que superan las 100 horas.
- **Energía No Suministrada (MWh):** Esta variable mostró una concentración de valores bajos y casos aislados de gran magnitud, lo que indica eventos de alto impacto energético, pero poco frecuentes.

En cuanto a las variables categóricas, se analizaron las distribuciones de Causa de Interrupción, Tipo de Protección que Actuó, Nivel de Tensión y Tipo de Alimentador Primario, encontrando que ciertas causas y configuraciones de red están más asociadas a interrupciones no técnicas. Para conocer la relación entre estas variables, se realizó un análisis bivariado, el cual incluyó:

- **Relación entre Causa de Interrupción y Tipo de Falla:** Identificando que causas como manipulación indebida o hurto de energía presentan mayor proporción de fallas no técnicas.
- **Relación entre Nivel de Tensión y Duración de la Interrupción:** Niveles de tensión más bajos presentaron una mayor dispersión en la duración de eventos.
- **Comparación del promedio de duración por Origen y Causa de Interrupción:** Se realizó mediante mapas de calor, lo que permitió detectar combinaciones críticas de origen–causa que generan interrupciones prolongadas.

Adicionalmente, se realizó un análisis de agrupamiento no supervisado (clustering) sobre las descripciones textuales de las interrupciones, con el fin de detectar patrones lingüísticos en los reportes operativos. Para ello, se utilizó la representación vectorial TF-IDF (Term Frequency–Inverse Document Frequency) junto con el algoritmo K-Means, lo que permitió agrupar registros con vocabulario y contexto similar. Este proceso facilitó la identificación de grupos semánticos recurrentes, tales como maniobras operativas, desconexiones solicitadas, causas climáticas o eventos por terceros, aportando una dimensión cualitativa al análisis general del conjunto de datos. Finalmente, las visualizaciones empleadas incluyeron histogramas para la distribución de variables numéricas, diagramas de barras para frecuencias de categorías, boxplots para la detección de valores atípicos y mapas de calor para el análisis de relaciones promedio. Estas herramientas permitieron identificar comportamientos relevantes y servir de base para la etapa de modelado predictivo.

3.3. Integración de variables geográficas, socioeconómicas y climáticas

Con el propósito de capturar heterogeneidades espaciales, contextuales y ambientales, se integraron variables geográficas, socioeconómicas y climáticas al conjunto de datos base. Estas dimensiones adicionales permitieron contextualizar los eventos eléctricos dentro de un entorno físico y social más representativo de la realidad operativa de la red.

3.3.1. Variables geográficas

- **Geocodificación y columnas de coordenadas:** A partir de los campos de localización del evento se construyó una clave de geocodificación y se obtuvieron coordenadas con Nominatim (vía geopy), manteniendo un caché para reproducibilidad. Las coordenadas se incorporaron como columnas Lat y Lon y se validaron con un bounding box de la ciudad de estudio para filtrar errores.

- **Control de calidad:** Se verificaron rangos válidos (lat $\in$ [-90, 90], lon $\in$ [-180, 180]) y se documentaron las imputaciones conservadoras en los casos sin coordenadas.

3.3.2. Variable socioeconómica

El contexto territorial se incorporó mediante un índice heurístico (6-10) asignado exclusivamente en función de la toponimia de SectorFalla, con el objetivo de reflejar la exposición a violencia y conflictividad reportada públicamente para la ciudad de estudio en los últimos

años. Como regla de asignación, se normalizaron los nombres de sector (mayúsculas, sin tildes, espacios unificados) y se aplicaron reglas por patrones:

- **10 (riesgo máximo):** Sectores de muy alta peligrosidad, localizados predominantemente en el cuadrante sur de la ciudad de estudio.
- **9 (muy alto):** Sectores peligrosos con densidad de eventos, aunque sin cumplir todos los criterios del nivel 10.
- **8 (alto):** Conjunto de sectores que exhiben densidad de eventos elevada.
- **7 (estándar):** no existen datos en SectorFalla (independiente de que exista o no CantonFalla, conforme a la regla acordada).
- **6 (base):** existen datos en SectorFalla, pero no coincide con los patrones anteriores.

3.3.3. Variable climática

Para capturar la influencia de las condiciones meteorológicas sobre las interrupciones eléctricas, se integraron variables climáticas obtenidas de la plataforma Visual Crossing Weather Data, correspondiente al período enero 2023 - diciembre 2024.

- **Fuentes y procesamiento:** Los datos fueron consultados mediante la interfaz de Visual Crossing, vinculando cada evento de interrupción con la fecha correspondiente. Se extrajeron las variables conditions (estado general del tiempo) y description (descripción detallada).
- **Integración y validación:** Estas variables fueron fusionadas con el conjunto de datos principal mediante una unión temporal (timestamp join), garantizando la coherencia cronológica entre el día del evento y las condiciones registradas. Posteriormente, se realizó una verificación de integridad estructural, confirmando la correcta correspondencia entre registros y la presencia completa de datos en las columnas agregadas.

3.3.4. Variable objetivo y features seleccionados para modelado

Para la construcción de los modelos de clasificación se consideraron variables operativas, de red, temporales y contextuales. La variable objetivo fue el Origen de Interrupción (interna/externa y programada/no programada), mientras que las variables de entrenamiento o features fueron las siguientes columnas:

- **Texto libre**

- DescripcionInterrupcion.

- **Numéricas**
  - PeligrosidadZona_1_10 (índice 6–10).
  - FMIk (indicador operativo).
  - NivelTension_kV.
  - PotenciaNominalFueraServicio_MW.
  - EsFeriado (0/1).
  - EsFinDeSemana (0/1).

- **Categóricas**
  - SectorFalla.
  - Subestacion.
  - CondicionClimatica.
  - DiaSemana.
  - BloqueHorario (derivada de la hora de inicio).

3.3.5. Definición de clases de la variable objetivo

A efectos del modelo, cada evento se clasifica en una de las siguientes cuatro clases:

- **Interna Programada:** Intervención planificada y notificada realizada por el operador de distribución dentro de su sistema (p. ej., mantenimiento preventivo, ampliaciones de red, maniobras programadas).

- **Interna No Programada:** Evento no planificado cuyo origen está dentro del sistema de distribución del operador (p. ej., fallas súbitas en alimentadores, transformadores, protecciones o empalmes).

- **Externa Programada:** Interrupción planificada y notificada cuyo origen es externo al sistema de distribución (p. ej., trabajos coordinados con transmisión u obras de terceros que requieren desenergizar parte de la red).

- **Externa No Programada:** Evento imprevisto con origen externo al sistema de distribución (p. ej., clima adverso, contacto de vegetación, colisiones con infraestructura, incidentes en transmisión).

3.4. Selección y entrenamiento de modelos de clasificación

Con el objetivo de elegir el modelo más adecuado, se consideró el desarrollo de dos modelos de aprendizaje supervisado en la tarea de clasificación: Random Forest, y Red Neuronal de Perceptrón Multi Capa (MLP). El primero, basado netamente en el aprendizaje de máquina (Machine Learning), se usó como base para comprobar sus ventajas y desventajas frente a soluciones más complejas. El segundo, basado en el aprendizaje profundo (Deep Learning), se usó como estrategia escalable, es decir, una alternativa en el caso se necesite trabajar con un mayor volumen de datos.

3.4.1. División de conjunto de datos

Para poder realizar un correcto entrenamiento en ambos modelos y evaluar su desempeño, fue necesario dividir el conjunto de datos. Se decidió realizar la división de la siguiente manera: 80 % de los registros fueron usados para el entrenamiento, y el 20 % restante fue destinado para pruebas. Adicionalmente, en el caso del modelo MLP, fue necesario tener un conjunto de validación para monitorizar la ausencia de sobreajuste, por lo que se repartió el conjunto de entrenamiento en 80 % para entrenamiento neto, y 20 % para validación. En total, el conjunto fue dividido en 64 % para entrenamiento, 16 % para validación, y 20 % para pruebas.

3.4.2. Modelo Random Forest

Debido a que el modelo Random Forest opera bajo un esquema de nodos, no se realizó una normalización a las variables predictoras. Por lo tanto, solo fue necesario realizar una configuración de hiperparámetros previo al entrenamiento. Dichos hiperparámetros fueron definidos mediante pruebas preliminares enfocadas en lograr un balance entre precisión, interpretabilidad y eficiencia computacional. Se empleó un número de 100 árboles (n_estimators) y se configuró el criterio de división como entropía (criterion), lo que le permite al modelo seleccionar las divisiones en los nodos de los árboles con base en la ganancia de información. Este enfoque permite capturar patrones informativos más ricos en comparación con el criterio de Gini en ciertos contextos.

3.4.3. Modelo Red Neuronal MLP

En el caso de este modelo, sus variables predictoras fueron normalizadas mediante la técnica de escalado estándar (StandardScaler), lo cual es esencial para asegurar un aprendizaje

eficiente en redes neuronales. Además, las etiquetas de clase, correspondientes al tipo de falla, fueron codificadas utilizando LabelEncoder y transformadas a formato categórico por medio de one-hot encoding. Su arquitectura es secuencial y consta de 3 capas. La primera capa contiene 128 neuronas y función de activación ReLU. La segunda capa contiene 64 neuronas y también activación ReLU. Finalmente, la tercera capa, siendo la salida, tiene activación softmax, pues va acorde a la naturaleza multiclase del modelo. Para mitigar el riesgo de sobreajuste, se incluyeron capas Dropout del 15 % después de cada capa, con excepción de la salida. El modelo fue compilado con el optimizador Adam, debido a su uso universal para este tipo de red neuronal, y la función de pérdida categorical_crossentropy, la cual es efectiva para modelos cuya salida sea una clasificación multiclase. Además, se definieron 50 épocas con un tamaño de lote (batch size) de 32 muestras con el fin de mitigar la posibilidad de que el modelo no tenga problemas de sobreajuste.

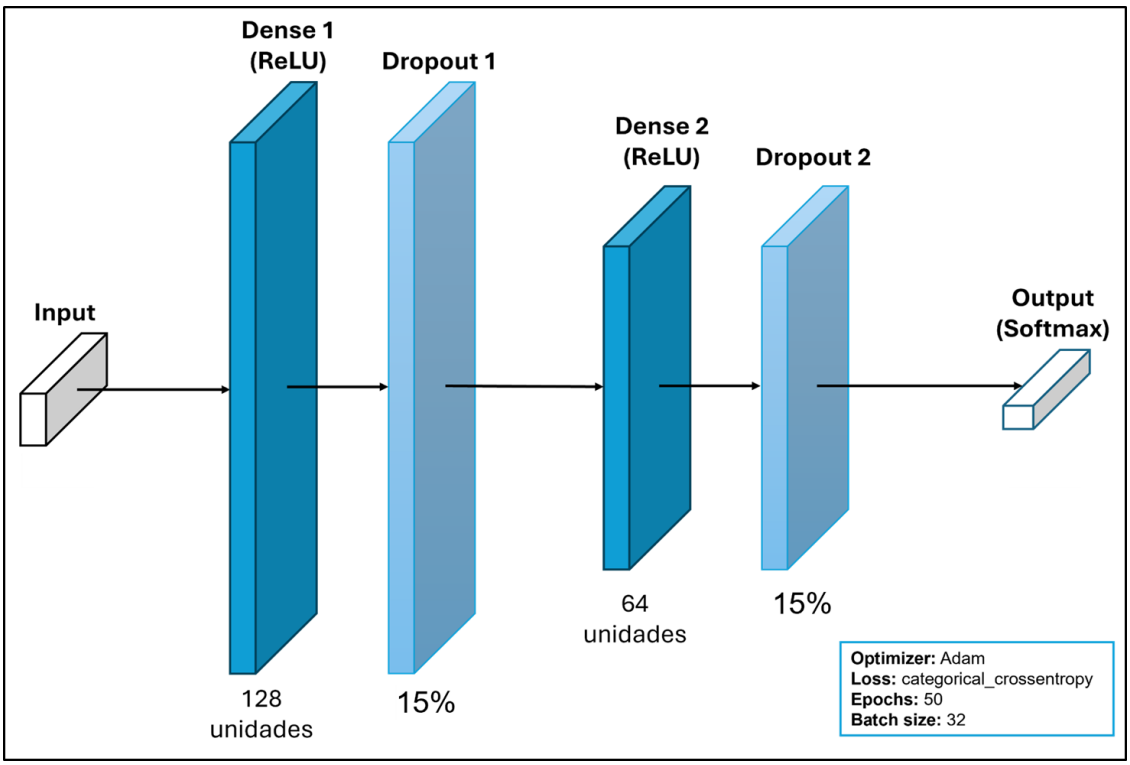

Figura 3.1: Esquema del modelo

3.5. Evaluación de desempeño de los modelos

Para evaluar el desempeño de ambos modelos, se utilizaron métricas estándar de clasificación multiclase: precisión (accuracy), precisión por clase (precision), recuperación (recall) y puntaje F1 (f1-score). Estas métricas pueden obtenerse mediante el uso del reporte de

clasificación generado por la librería Scikit-Learn en Python. Además, se realizó una comparación final en el que se incluyó la curva de aprendizaje con el fin de determinar si existió un sobreajuste en los modelos y determinar cuál es el más adecuado.

3.5.1. Modelo Random Forest

Tabla 3.1: Reporte de Clasificación del Modelo Random Forest

Clase / Medida	**Precisión**	**Recall**	**F1-Score**	**Support**
Externa no programada	1.000	0.983	0.992	361
Externa programada	1.000	1.000	1.000	330
Interna no programada	0.976	1.000	0.988	439
Interna programada	1.000	0.583	0.737	12
Accuracy			0.990	1142
Macro Avg	0.994	0.892	0.929	1142
Weighted Avg	0.991	0.990	0.990	1142

El modelo de clasificación basado en Random Forest mostró un desempeño global notablemente alto en el conjunto de prueba, alcanzando una precisión general de 99.0 % en su accuracy. Sin embargo, al clasificar registros de la clase INTERNA PROGRAMADA, el modelo obtuvo métricas considerablemente inferiores en comparación a las demás clases. Esta caída en el rendimiento puede atribuirse tanto al bajo número de muestras para dicha clase, que afecta la capacidad del modelo para aprender patrones representativos, y la posible similitud semántica o contextual con otras clases. En síntesis, a primera vista, el modelo Random Forest presentó un excelente rendimiento general, pero existieron limitaciones en la clasificación de clases con menos registros, lo cual puede ser relevante para decisiones críticas basadas en detección de fallas específicas.

3.5.2. Modelo Red Neuronal MLP

Tabla 3.2: Reporte de Clasificación del Modelo Red Neuronal MLP

Clase / Medida	**Precisión**	**Recall**	**F1-Score**	**Support**
Externa no programada	1.000	1.000	1.000	361
Externa programada	1.000	1.000	1.000	330
Interna no programada	1.000	1.000	1.000	439
Interna programada	0.910	0.830	0.870	12
Accuracy			0.997	1142
Macro Avg	0.980	0.960	0.970	1142
Weighted Avg	1.000	1.000	1.000	1142

El modelo de red neuronal MLP alcanzó una precisión global del 99.73 % en el conjunto de prueba, lo que representa un rendimiento sólido en el contexto del problema de clasificación multiclase. La arquitectura profunda del modelo le permitió captar relaciones complejas en los datos, lo cual se ve reflejado en su buen rendimiento general, aunque sigue presentando dificultades en la clase de menor cantidad de registros, INTERNA PROGRAMADA, por los motivos descritos durante la evaluación del modelo anterior. En resumen, este modelo ha demostrado ser una solución prometedora a pesar de las dificultades, por lo que fue considerado para la comparativa final.

3.5.3. Comparación final

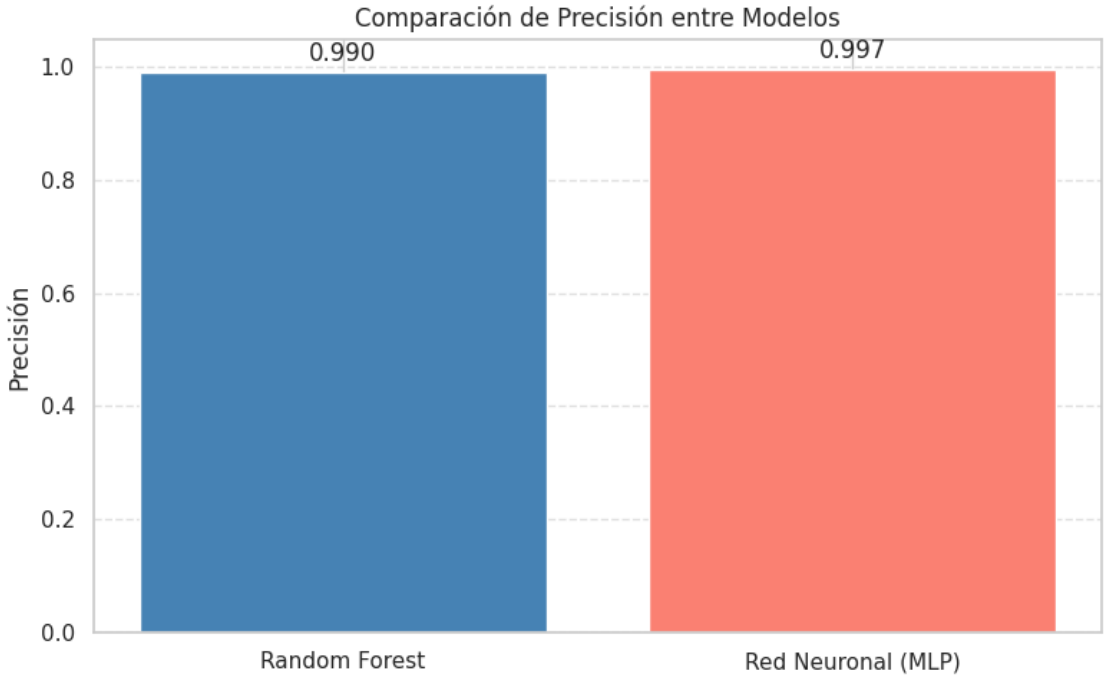

Figura 3.2: Comparación de precisión entre modelos

A nivel de precisión global, ambos modelos alcanzaron un rendimiento sobresaliente en la tarea de clasificación del origen de las interrupciones eléctricas. El modelo Random Forest obtuvo un accuracy del 99.0 %, mientras que la Red Neuronal Multicapa (MLP) alcanzó un 99.7 % de precisión global en el conjunto de prueba. Sin embargo, estas métricas, por si solas, no garantizan la eficiencia de un modelo frente a otro. Debido a esto, también se analizó la curva de aprendizaje de cada uno; es decir, una gráfica que indica qué tan bien se desenvolvieron frente a un conjunto de prueba o validación.

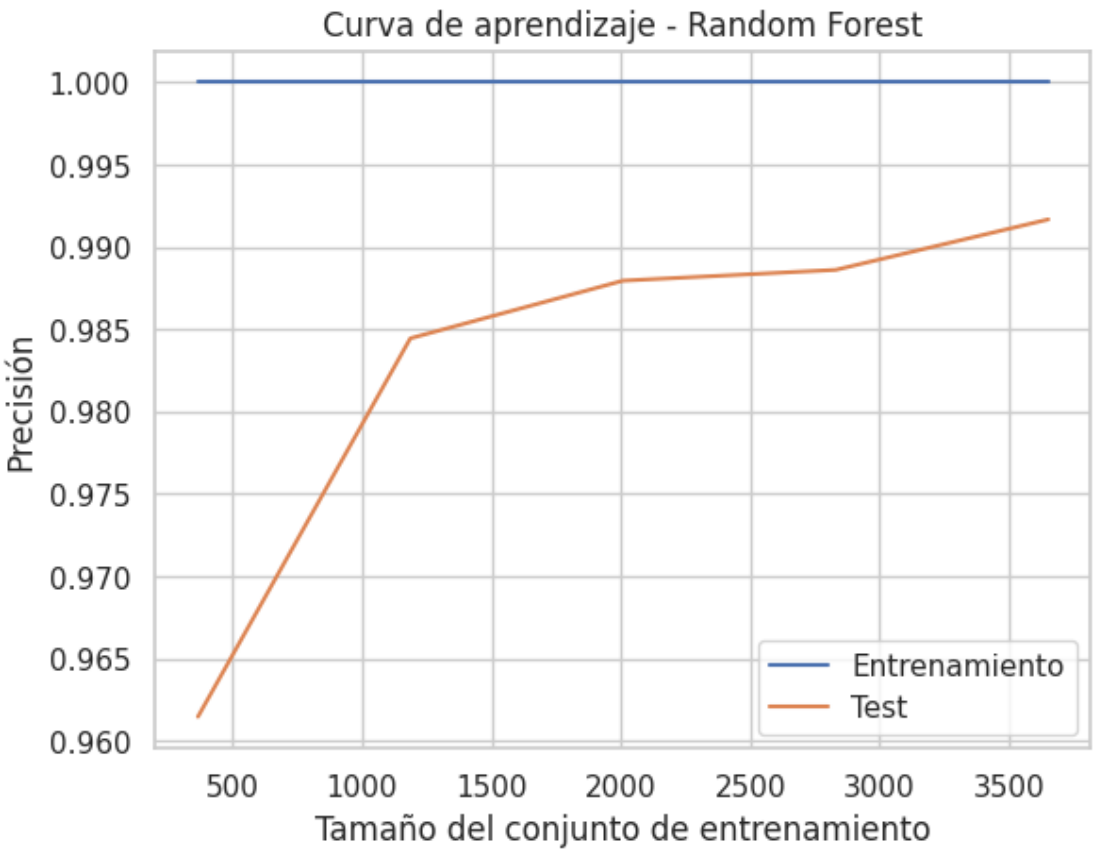

Figura 3.3: Curva de aprendizaje – Modelo Random Forest

El modelo Random Forest muestra un ajuste casi perfecto al conjunto de entrenamiento, mientras que la curva de validación se mantiene también alta, pero con una brecha constante. Esta diferencia entre entrenamiento y validación, si bien es normal en este tipo de modelo, también podría indicar un alto riesgo de sobreajuste (overfitting), es decir, que el modelo habría aprendido demasiado bien los patrones del conjunto de entrenamiento, limitando su capacidad de generalización a nuevos datos.

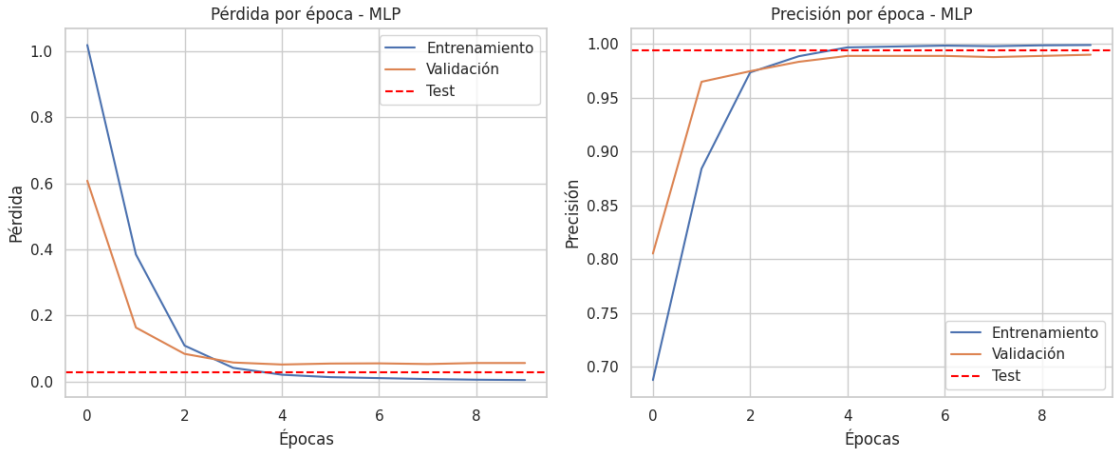

Figura 3.4: Curva de aprendizaje – Modelo Red Neuronal MLP

Por su parte, el modelo MLP (Multilayer Perceptron) muestra un comportamiento más balanceado y progresivo. En la gráfica de pérdida, tanto las curvas de entrenamiento como de validación descienden de forma simultánea hasta estabilizarse cerca de cero, indicando una convergencia adecuada sin indicios de sobreajuste. En la gráfica de precisión, las curvas de entrenamiento y validación evolucionan de manera casi paralela, alcanzando valores cercanos al 100 %, en coherencia con los resultados del conjunto de prueba. Este comportamiento refleja un modelo más estable y con mayor capacidad de generalización, capaz de mantener un rendimiento uniforme incluso con variaciones en los datos de entrada. Aunque el Random Forest alcanzó un excelente rendimiento cuantitativo, la Red Neuronal MLP mostró un entrenamiento más controlado, menor propensión al sobreajuste y una mejor escalabilidad para escenarios con mayor volumen de datos o mayor complejidad de variables. Por su robustez, adaptabilidad y consistencia en las métricas obtenidas, el modelo MLP fue seleccionado como modelo final para la implementación en la aplicación interactiva desarrollada en Streamlit, utilizada para la visualización y análisis geoespacial de los resultados.

3.6. Visualización geoespacial de los resultados

Con el objetivo de facilitar la interpretación de los resultados del modelo y detectar patrones territoriales asociados a la ocurrencia de fallas eléctricas, se implementó una visualización georreferenciada de las predicciones del modelo seleccionado previamente sobre un mapa de la ciudad de estudio. Para esto, se utilizó el campo “UbicaciónFalla”, complementado con los campos “SectorFalla” y “CantonFalla” del conjunto de datos para construir una dirección geográfica aproximada. A falta de coordenadas en el dataset original, se recurrió a un proceso de geocodificación mediante la API de Nominatim (OpenStreetMap), utilizando como clave

combinada la dirección y el cantón. Para optimizar el rendimiento y evitar solicitudes repetidas a la API, las coordenadas se almacenaron en un diccionario persistente y se reutilizaron en ejecuciones posteriores. En los casos en los que el proceso de geocodificación no devolvió una coordenada válida, sea esto ocasionado por ambigüedad en la dirección, o por falta de coincidencia, se asignaron coordenadas por defecto correspondientes al centro urbano de la ciudad de estudio, lo cual permitió mantener la consistencia del mapa y evitar valores nulos.

Una vez obtenido el conjunto completo de predicciones con coordenadas válidas, se generó un mapa utilizando la librería Folium, en el cual cada punto representa una clasificación individual del modelo. Los puntos fueron codificados por color de acuerdo con la clase predicha, permitiendo así una rápida identificación de zonas con alta concentración de ciertos tipos de interrupciones. Esta visualización permitió observar con claridad la distribución espacial de los eventos según su tipología, destacando patrones de concentración en determinados sectores urbanos. De este modo, el mapa georreferencial no solo enriquece el análisis de los resultados, sino que constituye una herramienta clave para la toma de decisiones operativas y estratégicas por parte de la empresa eléctrica, al permitir focalizar esfuerzos de mantenimiento preventivo en zonas con alta incidencia de fallas.

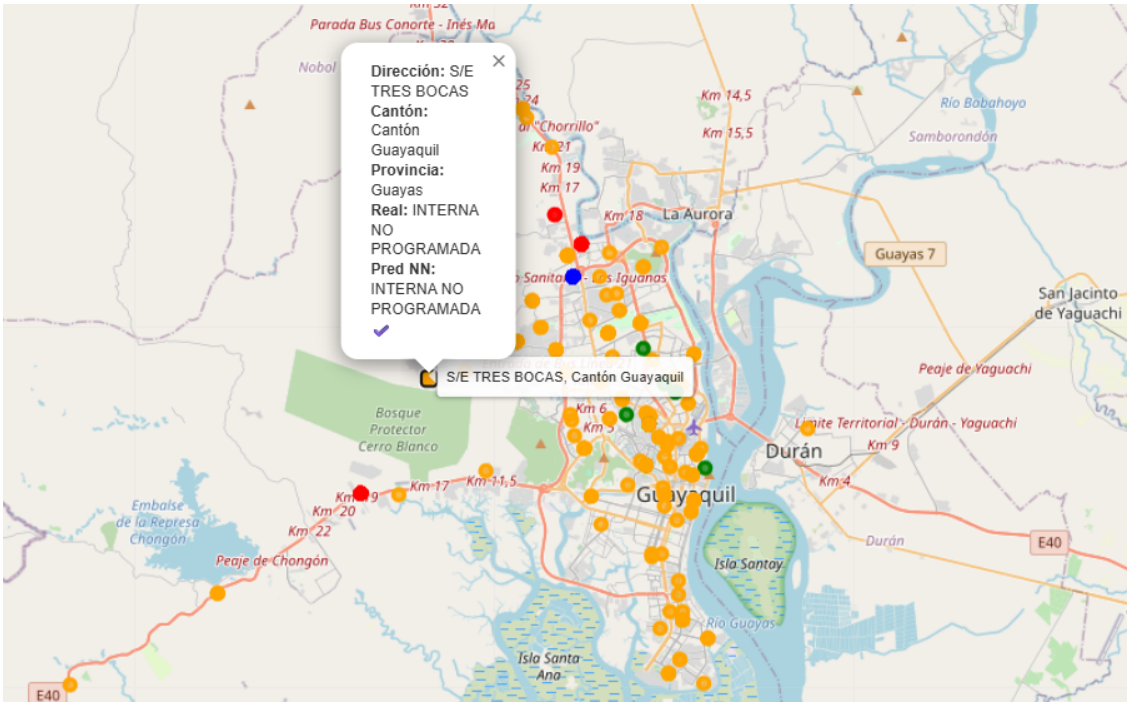

Figura 3.5: Vista preliminar del mapa generado

3.7. Aplicación interactiva

Como etapa final del desarrollo, se implementó una aplicación web interactiva tipo dashboard, desarrollada en Streamlit, con el propósito de integrar los resultados del modelo predictivo y facilitar la exploración dinámica de los datos. Esta aplicación representa la fase de implementación operativa del sistema propuesto, al consolidar los procesos de modelado, análisis y visualización en una única interfaz accesible para el usuario final. El dashboard fue diseñado bajo un enfoque modular, integrando tres componentes principales.

El primero corresponde al módulo de exploración de datos históricos, que permite filtrar los registros por tipo de evento, sector o rango de fechas. En este apartado se presentan estadísticas resumidas, gráficos dinámicos de barras y diagramas de pastel que ilustran la distribución y frecuencia de las interrupciones. El segundo módulo, denominado predicciones del modelo MLP, muestra los resultados clasificados del modelo final seleccionado. Incluye métricas globales de desempeño —como accuracy, recall y F1-score— junto con el detalle individual por clase, lo que permite analizar la efectividad del modelo en la identificación del origen de las interrupciones eléctricas. Finalmente, el módulo de mapa interactivo integra la visualización geográfica desarrollada con Folium, donde cada punto representa un evento de interrupción clasificado por el modelo. Los registros se visualizan sobre el mapa de la ciudad de estudio con codificación de colores según el tipo de evento, y contienen información contextual adicional, como el sector, la subestación y la tipología de la falla.

Esta herramienta constituye una plataforma de apoyo para la toma de decisiones técnicas y de planificación, ya que permite visualizar de forma integral el comportamiento histórico, las tendencias espaciales y la efectividad del modelo de clasificación. Asimismo, fortalece el vínculo entre la analítica predictiva y la gestión operativa de la red eléctrica, al ofrecer una interfaz práctica, escalable y orientada a la mejora continua del sistema eléctrico.

Capítulo 4

Análisis de resultados

4.1. Validación cuantitativa y cualitativa

La validación cuantitativa de los modelos desarrollados se realizó mediante métricas estándar de clasificación: accuracy, precision, recall y F1-score. El modelo Random Forest alcanzó un accuracy global de 99.0 %, destacando por su robustez en la mayoría de las clases, aunque presentó limitaciones en la categoría INTERNA PROGRAMADA, con un F1-score de 0.737, debido al bajo número de muestras disponibles para esa clase. Por su parte, la Red Neuronal Multicapa (MLP) obtuvo un accuracy del 99.73 %, mostrando un desempeño perfecto en tres de las cuatro clases y una notable mejora en la generalización de patrones. Este resultado refleja la capacidad del modelo para aprender relaciones no lineales complejas entre las variables técnicas, geoespaciales y socioeconómicas.

Desde la perspectiva cualitativa, los resultados fueron contrastados con registros reales, verificando la coherencia de las predicciones en distintos contextos urbanos. Se observó que el modelo reflejó adecuadamente la naturaleza de las fallas en sectores de alta densidad poblacional, donde la ocurrencia de interrupciones coincide con los registros históricos de campo. En contraste, en zonas con menor instrumentación o con registros incompletos, el modelo tendió a presentar leves desviaciones, principalmente atribuibles a la falta de información consistente sobre eventos menores o fallas de baja frecuencia. En conjunto, la validación mixta demostró que, aunque el modelo Random Forest ofrece una precisión numérica sobresaliente, el modelo MLP presenta un equilibrio superior entre desempeño, estabilidad y escalabilidad, resultando más adecuado para escenarios futuros de mayor complejidad y volumen de datos.

4.2. Discusión de hallazgos relevantes

Los resultados obtenidos permiten identificar tendencias significativas sobre la naturaleza y el comportamiento de las interrupciones eléctricas en la ciudad de estudio. Una primera observación es la estacionalidad de las fallas, con una mayor frecuencia durante la temporada lluviosa, comprendida entre febrero y abril de cada año, presentando picos notables en marzo de 2023 y abril de 2024. Esta tendencia confirma el carácter estacional del fenómeno, coincidiendo con los meses de mayor precipitación y humedad, donde las condiciones climáticas adversas, como la nubosidad, la lluvia y las descargas atmosféricas, inciden directamente en la estabilidad del sistema eléctrico.

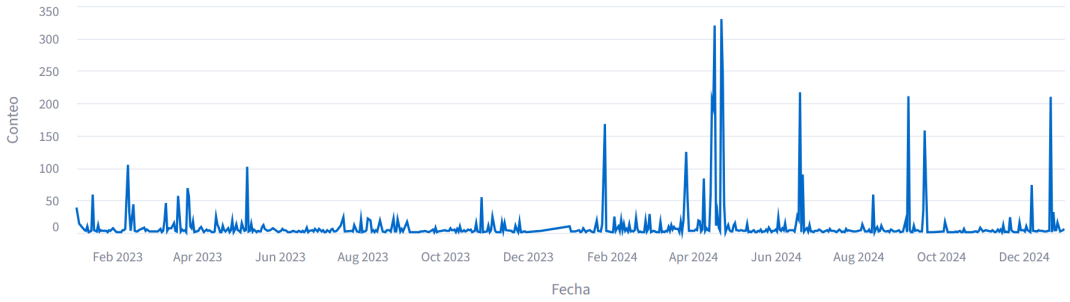

Figura 4.1: Evolución temporal de interrupciones

Fuera de estos periodos, la actividad se reduce significativamente, lo que evidencia que los factores climáticos tienen un efecto directo sobre la ocurrencia de desconexiones. Este comportamiento respalda la relación entre el clima y la confiabilidad en redes aéreas de distribución, en concordancia con estudios técnicos que señalan la vulnerabilidad de este tipo de infraestructura ante lluvias intensas, humedad elevada y descargas eléctricas.

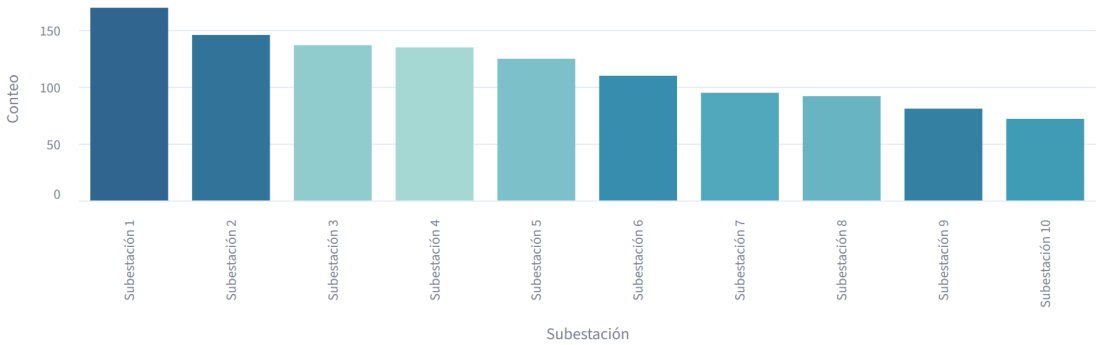

Figura 4.2: Distribución por subestación

El análisis por subestación muestra que los mayores conteos de interrupciones se concentran en las subestaciones 1, 2, 3 y 4, con más de 130 incidencias cada una. Estas subestaciones

atienden sectores con alta densidad de carga y demanda residencial-comercial, lo que se asocia con una mayor frecuencia de eventos. En contraste, subestaciones como la 9 y la 10 registran niveles inferiores de interrupciones, posiblemente por cubrir zonas de menor extensión o con infraestructura más reciente y confiable.

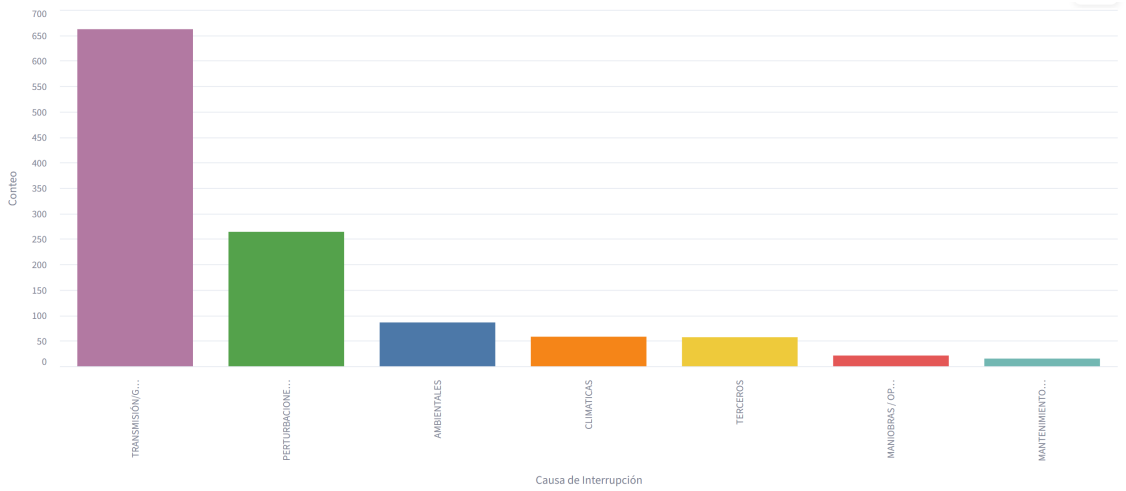

Figura 4.3: Causa de interrupción

Respecto a las causas, los resultados muestran que las perturbaciones del sistema de transmisión y las causas ambientales son las más recurrentes, seguidas en menor medida por eventos climáticos y afectaciones por terceros. En cambio, las categorías de maniobras operativas y mantenimiento planificado representan una fracción muy reducida, lo que confirma que las fallas técnicas o accidentales predominan sobre las interrupciones programadas, evidenciando una brecha en la planificación preventiva del mantenimiento. El resumen por clase muestra 2414 interrupciones internas no programadas, 2143 externas no programadas, 1648 externas programadas y 77 internas programadas. Esta marcada diferencia entre interrupciones externas e internas refleja un sesgo operativo hacia causas externas, tales como perturbaciones en líneas troncales o mantenimientos de red, lo cual coincide con las métricas obtenidas durante la etapa de entrenamiento, donde las clases externas presentaron mayor soporte y precisión.

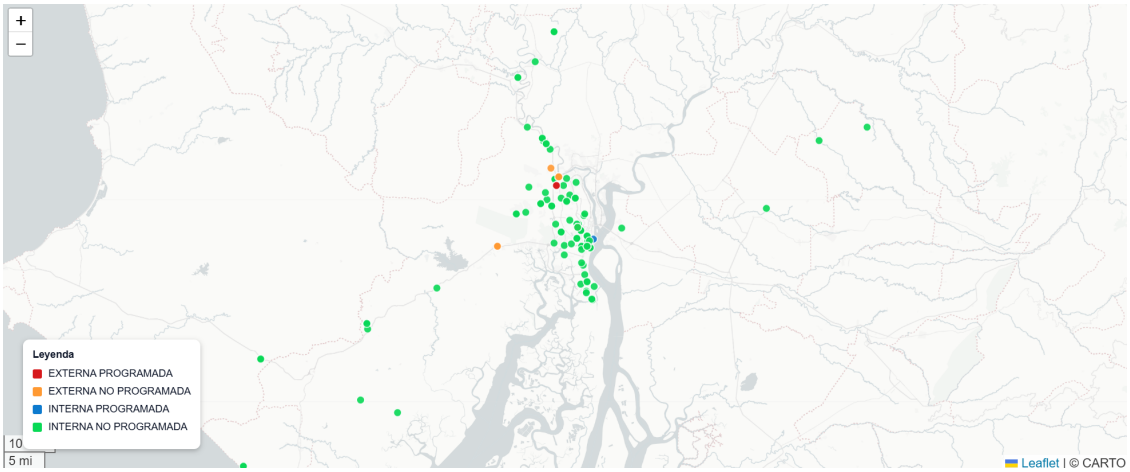

Figura 4.4: Mapa de distribución geográfica

La visualización geográfica evidencia una concentración principal de eventos en el núcleo urbano de la ciudad de estudio, con dispersión menor hacia los bordes de esta. Predominan claramente los marcadores verdes (INTERNA NO PROGRAMADA) y azules (INTERNA PROGRAMADA), lo que indica que las interrupciones de origen interno son las más frecuentes en el área analizada. En contraste, los eventos EXTERNOS aparecen en proporción reducida y con distribución más dispersa: los puntos naranjas (EXTERNA NO PROGRAMADA) y rojos (EXTERNA PROGRAMADA) son puntuales y tienden a ubicarse en sectores periféricos o a lo largo de corredores viales, sin formar concentraciones equivalentes a las internas. Este patrón respalda la heterogeneidad operativa del sistema de distribución: el mayor peso recae en incidencias locales dentro del tejido urbano (internas), mientras que las externas, sean programadas o no, tienen una presencia menor y más aislada. En conjunto, los resultados demuestran que el sistema de clasificación y visualización desarrollado permite identificar patrones reales y coherentes en la operación eléctrica de la ciudad de estudio. El modelo no solo reproduce fielmente los registros históricos, sino que además facilita la interpretación temporal, operativa y espacial de los eventos, constituyéndose en una herramienta eficaz para priorizar mantenimiento, identificar zonas críticas y planificar acciones correctivas con base en evidencia empírica.

4.3. Análisis de patrones en descripciones de fallas

Con el fin de identificar tendencias lingüísticas y operativas dentro de los registros históricos, se realizó un análisis de agrupamiento no supervisado (clustering) sobre las descripciones textuales de las interrupciones. Este procedimiento complementa el análisis exploratorio al permitir detectar similitudes en el vocabulario utilizado en los reportes técnicos y reconocer

grupos representativos de causas recurrentes. Para el procesamiento del texto se utilizó la técnica TF-IDF (Term Frequency–Inverse Document Frequency), que transforma las descripciones en vectores numéricos basados en la frecuencia y relevancia de los términos. Posteriormente, se aplicó el algoritmo K-Means, seleccionando un número óptimo de 10 clústeres de acuerdo con el comportamiento de la métrica de silueta.

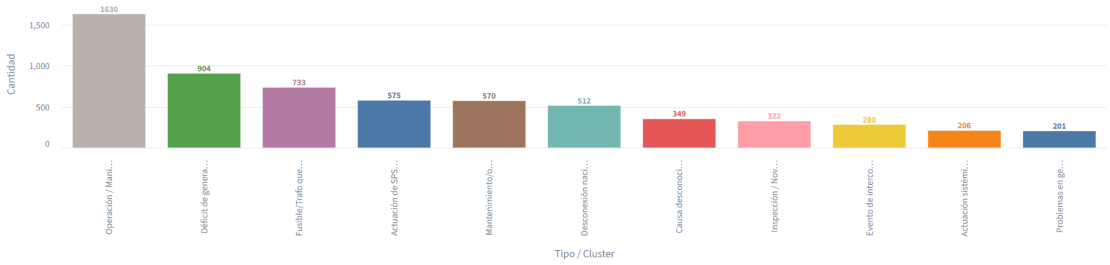

Figura 4.5: Distribución de clústeres en descripciones de fallas

El resultado del agrupamiento permitió identificar diez grupos temáticos diferenciados, entre ellos destacan:

- **Operación / Maniobras:** Agrupa la mayoría de los registros (1 630 casos), asociados a desconexiones operativas o maniobras en subestaciones.
- **Déficit de generación:** Segundo grupo más numeroso (904 casos), relacionado con eventos de falta de generación o restricciones nacionales.
- **Fusible / Trafo que actuó, Actuación de SPS, Mantenimiento o revisión:** Categorías intermedias que representan incidencias técnicas comunes en los sistemas de protección y mantenimiento.
- **Causa desconocida o Desconexión nacional:** Grupos con menor frecuencia, pero relevantes por su potencial impacto sistémico.
- **Problemas en equipo, inspección o intervención, evento de terceros y actuación de sistema:** Categorías menores, pero útiles para la clasificación automática de registros futuros.

El agrupamiento permitió observar que más del 25 % de las descripciones se relacionan con maniobras operativas, mientras que una fracción menor corresponde a causas no determinadas. Este resultado respalda los hallazgos generales del modelo clasificatorio, en los que predominan las interrupciones técnicas sobre las no técnicas, y demuestra la consistencia semántica entre los registros textuales y los patrones de clasificación. En conjunto, este análisis de clúster

aporta una dimensión adicional al estudio, ya que permite estructurar la información textual de los reportes operativos, reducir redundancias y mejorar la coherencia terminológica de los registros históricos, fortaleciendo así la calidad de los datos para análisis futuros y aplicaciones de minería de texto.

4.4. Identificación de zonas críticas y patrones espaciales

La representación georreferenciada de las interrupciones eléctricas permitió identificar zonas de mayor recurrencia de eventos dentro de la ciudad de estudio. A partir de la integración de coordenadas geográficas obtenidas mediante la API de Nominatim y el posterior mapeo en la aplicación interactiva, se observaron agrupaciones espaciales de eventos con distinta intensidad y distribución. En la zona centro y norte de la ciudad se concentra la mayor cantidad de interrupciones, principalmente en sectores urbanos con alta densidad de carga y presencia de subestaciones que abastecen áreas residenciales y comerciales de gran demanda. En estas zonas predominan las interrupciones internas, tanto programadas como no programadas, reflejando una alta actividad operativa y mantenimiento frecuente en los alimentadores principales. Por otro lado, la zona norte muestra una dispersión menor de eventos, con presencia de interrupciones externas no programadas vinculadas a causas operativas o ambientales. Estos sectores registran incidentes localizados, generalmente asociados a fallas reactivas o deficiencias en la infraestructura secundaria de distribución.

El análisis espacial confirma la existencia de corredores eléctricos críticos donde se combinan alta frecuencia de interrupciones y tiempos prolongados de restablecimiento, lo que sugiere la necesidad de priorizar estas áreas en los planes de mantenimiento preventivo. Además, se evidencia una correspondencia entre la distribución geográfica de las fallas y los patrones detectados en los clústeres semánticos, donde las descripciones de tipo maniobras operativas y déficit de generación se concentran en los mismos sectores con alta densidad de eventos. La herramienta de visualización desarrollada permite filtrar los eventos por tipo de interrupción, causa o período temporal, lo que facilita la detección dinámica de patrones espaciales y su relación con variables contextuales como peligrosidad territorial o condiciones climáticas. Este componente constituye un módulo de apoyo a la gestión operativa, al ofrecer una visión geoespacial integral de los puntos críticos de la red y sus comportamientos a lo largo del tiempo. En resumen, los resultados confirman que la integración entre el modelo clasificatorio, el análisis de texto y la visualización geográfica ofrece una perspectiva completa del fenómeno, permitiendo identificar, clasificar y localizar las interrupciones eléctricas con

precisión y coherencia analítica.

4.5. Consideraciones sobre utilidad práctica del sistema

La integración del modelo de red neuronal con la representación georreferencial ofrece varias ventajas. En primer lugar, permite identificar zonas críticas y priorizar trabajos de mantenimiento en áreas urbanas con mayor recurrencia de fallas, por lo que es muy útil para la planificación preventiva. Además, los mapas interactivos facilitan la comunicación de hallazgos a directivos, técnicos y reguladores, mostrando de manera visual dónde se concentran los problemas, lo que implica un apoyo enorme en la toma de decisiones. Por lo tanto, al focalizar los sectores más afectados, los equipos de mantenimiento pueden ser asignados de forma más eficiente, reduciendo costos y tiempos de respuesta.

Por otro lado, el modelo también presenta limitaciones. Entre ellas, la más notoria es que el desempeño depende de la calidad y completitud del dataset, pues si la información de incidentes es parcial o contiene errores, las predicciones pierden confiabilidad. En adición, la representación espacial actual muestra la localización de fallas, pero no cuantifica otras métricas importantes, tales como la cantidad de habitantes afectados por zona, lo que restringe análisis en términos de impacto. Finalmente, se requiere infraestructura tecnológica adecuada para mantener actualizado el sistema y garantizar que pueda integrarse con los sistemas de gestión de la empresa eléctrica, lo que supondría un incremento en los costos iniciales planteados.

Capítulo 5

Conclusiones y recomendaciones

5.1. Conclusiones

El desarrollo e implementación de modelos de clasificación basados en aprendizaje automático permitió caracterizar de manera precisa el origen y tipo de las interrupciones eléctricas registradas en la ciudad de estudio, integrando en el análisis factores técnicos, espaciales y socioeconómicos. El sistema desarrollado combina técnicas de procesamiento de datos, modelado predictivo y visualización geoespacial, lo que proporciona una visión integral del comportamiento operativo de la red eléctrica. Esta integración demostró que el uso de herramientas de inteligencia artificial, complementadas con la georreferenciación y el análisis exploratorio, mejora la comprensión del fenómeno de las desconexiones, facilita la identificación de zonas críticas y apoya la toma de decisiones estratégicas orientadas a fortalecer la confiabilidad y resiliencia del sistema de distribución eléctrica.

El análisis estadístico y exploratorio de los datos históricos de interrupciones eléctricas permitió identificar que la mayoría de los eventos corresponden a fallas técnicas no programadas, con una marcada estacionalidad durante los meses de mayor precipitación (febrero a abril). Las variables Duración de Interrupción y Energía No Suministrada evidenciaron distribuciones asimétricas con presencia de eventos extremos, confirmando la existencia de cortes de alto impacto operativo. El análisis de clustering aplicado sobre las descripciones textuales complementó este estudio, agrupando los registros en diez clústeres semánticos representativos, donde predominan categorías como maniobras operativas y déficit de generación. Este resultado respalda la consistencia entre la información textual de los reportes y la clasificación técnica obtenida, aportando valor adicional al análisis descriptivo tradicional.

La evaluación comparativa de los modelos de clasificación demostró un alto nivel de desempeño para ambos enfoques. El modelo Random Forest alcanzó una precisión global del 99.0 %, mientras que la Red Neuronal Multicapa (MLP) obtuvo una exactitud del 99.7 %,

con un mejor equilibrio entre clases y mayor capacidad de generalización frente a escenarios complejos. Estos resultados evidencian la viabilidad de aplicar modelos de aprendizaje supervisado en la clasificación del origen de las desconexiones eléctricas, confirmando que las redes neuronales pueden adaptarse con eficacia a variables técnicas, geográficas y socioeconómicas, constituyendo una herramienta confiable para el diagnóstico y pronóstico de interrupciones.

La herramienta de georreferenciación e interfaz interactiva desarrollada integró de forma efectiva los resultados del modelado con la dimensión espacial, permitiendo visualizar patrones de recurrencia y zonas críticas dentro de la red de distribución eléctrica de la ciudad de estudio. Los mapas generados evidenciaron concentraciones de eventos en sectores norte y centro de la ciudad, mientras que las interrupciones externas se observaron de forma más dispersa en la periferia urbana. Este componente visual facilita la planificación de mantenimientos preventivos, la priorización de recursos técnicos y el monitoreo operativo en tiempo real, consolidando la utilidad práctica del sistema como una herramienta de apoyo para la gestión técnica y estratégica del servicio eléctrico.

5.2. Recomendaciones

Debido a que el modelo demostró mayor precisión en áreas urbanas densas de la ciudad de estudio, pero presenta limitaciones para cuantificar el impacto real de interrupciones, por lo que es deseable el implementar estudios con datos enriquecidos, tales como información sobre el número de usuarios afectados y zonas con alta densidad de población, lo que mejoraría la eficiencia de los equipos de mantenimiento. Además, si bien la metodología de implementación del modelo ha resultado efectiva, aún tiene un enorme margen de mejora si se integrasen técnicas avanzadas de procesamiento de lenguaje natural, lo que aumentaría su capacidad de clasificación y, por ende, optimizar su funcionamiento.

Cabe destacar que el sistema actualmente opera con datos históricos, lo que implica poca utilidad en tiempo real, por lo que se puede recomendar una integración con plataformas operativas para clasificar las desconexiones eléctricas con datos dinámicos, lo que mejora enormemente su apoyo en la toma de decisiones. En futuros trabajos, se espera no solo implementar las recomendaciones descritas previamente, sino también experimentar con modelos adicionales de inteligencia artificial, que permitan una comparación más efectiva con el modelo elegido para esta investigación, e incluso, se podría validar con datos de otras regiones del país y, en el mejor de los casos, con datos internacionales.

Bibliografía

[1] Y. Zhou, L. Chen, W. Zhang, and X. Xiao, "Fault cause identification of distribution system based on association rule," in *2023 3rd International Conference on Energy Engineering and Power Systems (EEPS)*, pp. 781–786, IEEE, 2023.

[2] M. Azeroual, Y. Boujoudar, K. Bhagat, L. El Iysaouy, A. Aljarbouh, A. Knyazkov, M. Fayaz, M. S. Qureshi, F. Rabbi, and H. E. Markhi, "Fault location and detection techniques in power distribution systems with distributed generation: Kenitra city (morocco) as a case study," *Electric Power Systems Research*, vol. 209, p. 108026, 2022.

[3] M. S. Bashkari, A. Sami, and M. Rastegar, "Outage cause detection in power distribution systems based on data mining," *IEEE Transactions on Industrial Informatics*, vol. 17, no. 1, pp. 640–649, 2020.

[4] R. Dashti, M. Daisy, H. Mirshekali, H. R. Shaker, and M. H. Aliabadi, "A survey of fault prediction and location methods in electrical energy distribution networks," *Measurement*, vol. 184, p. 109947, 2021.

[5] S. Javadian, A. Nasrabadi, M.-R. Haghifam, and J. Rezvantalab, "Determining fault's type and accurate location in distribution systems with dg using mlp neural networks," in *2009 International conference on clean electrical power*, pp. 284–289, IEEE, 2009.

[6] X. Wang, N. Fatehi, C. Wang, and M. H. Nazari, "Deep learning-based weather-related power outage prediction with socio-economic and power infrastructure data," in *2024 IEEE Power & Energy Society General Meeting (PESGM)*, pp. 1–5, IEEE, 2024.

[7] L. Xu and M.-Y. Chow, "A classification approach for power distribution systems fault cause identification," *IEEE Transactions on power systems*, vol. 21, no. 1, pp. 53–60, 2006.

[8] A. Jaech, B. Zhang, M. Ostendorf, and D. S. Kirschen, "Real-time prediction of the duration of distribution system outages," *IEEE Transactions on Power Systems*, vol. 34, no. 1, pp. 773–781, 2018.

[9] V. C. W. Data, "Visual crossing weather query builder." Available at: https://www.visualcrossing.com/weather-query-builder/, 2025. Accessed: September 2025.

[10] C. P. Rodríguez Méndez, S. M. Ortiz Santamaria, H. Roa, *et al.*, *Predicción del riesgo de robo: enfoque bayesiano de modelización espacio-temporal y caso de estudio.* PhD thesis, ESPOL. FCNM, 2022.

[11] M. M. Buzau, J. Tejedor-Aguilera, P. Cruz-Romero, and A. Gómez-Expósito, "Detection of non-technical losses using smart meter data and supervised learning," *IEEE Transactions on Smart Grid*, vol. 10, no. 3, pp. 2661–2670, 2018.

[12] W. A. Zgallai, *Biomedical signal processing and artificial intelligence in healthcare*. Academic Press, 2020.

[13] N. JAKSE, "Classification techniques in machine learning," *Machine Learning in Geomechanics 1: Overview of Machine Learning, Unervised Learning, Regression, Classification and Artificial Neural Networks*, p. 117, 2024.

[14] B. ElOuassif, A. Idri, M. Hosni, and A. Abran, "Classification techniques in breast cancer diagnosis: a systematic literature review," *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, vol. 9, no. 1, pp. 50–77, 2021.

[15] J. Kufel, K. Bargieł-Łączek, S. Kocot, M. Koźlik, W. Bartnikowska, M. Janik, Ł. Czogalik, P. Dudek, M. Magiera, A. Lis, *et al.*, "What is machine learning, artificial neural networks and deep learning?—examples of practical applications in medicine," *Diagnostics*, vol. 13, no. 15, p. 2582, 2023.

[16] B. Topuz and N. Ç. Alp, "Machine learning in architecture," *Automation in Construction*, vol. 154, p. 105012, 2023.

[17] O. Kramer, "Scikit-learn," in *Machine learning for evolution strategies*, pp. 45–53, Springer, 2016.

[18] N. Ketkar, J. Moolayil, N. Ketkar, and J. Moolayil, *Deep learning with Python: learn best practices of deep learning models with PyTorch*. Springer, 2021.

[19] S. He, A. Chang, H. Zhang, Y. Song, J. Shang, G. Zhao, and W. Li, "Fault causes identification of transmission lines based on weighted naive bayes classification algorithm combined with complex algorithm," in *2021 International Conference on Power System Technology (POWERCON)*, pp. 2312–2316, IEEE, 2021.

[20] H. Sun, F. Li, C. Sticht, and S. Mukherjee, "Outage cause classification of power distribution systems with machine learning and real-world data," in *2022 IEEE Power & Energy Society General Meeting (PESGM)*, pp. 1–5, IEEE, 2022.

[21] L. Yang, "A brief introduction of the text classification methods," in *2022 IEEE International Conference on Electrical Engineering, Big Data and Algorithms (EEBDA)*, pp. 495–498, IEEE, 2022.

[22] P. Shiguihara and L. Berton, "Exploring deep neural networks and decision tree for spanish text classification," in *2022 IEEE XXIX International Conference on Electronics, Electrical Engineering and Computing (INTERCON)*, pp. 1–4, IEEE, 2022.