Universidad Técnica Federico Santa María

Departamento de Ingeniería Eléctrica

UNIVERSIDAD TECNICA
FEDERICO SANTA MARIA

Departamento de Ingeniería Eléctrica

Tesis para optar al grado de Magíster en Ciencias de la Ingeniería Eléctrica

Desarrollo de un Sistema Predictivo de descargas eléctricas atmosféricas Nube-Tierra Utilizando Técnicas de Inteligencia Artificial

Candidato:

Sr. Sergio Zumarán Rivera

Director de Tesis:

Dr. Johny Montaña Chaparro

Co-Director de Tesis:

Dr. Carlos Valle

Valparaíso, Chile
24 de noviembre de 2025

CONSTANCIA DE VALIDACIÓN Y CONFIDENCIALIDAD DE MONOGRAFÍA A REPOSITORIO ACADÉMICO

1.- IDENTIFICACIÓN DEL TRABAJO ACADÉMICO

Tipo de monografía (marcar una opción): □ Memoria o trabajo de título X Tesis de Postgrado

Título del trabajo: Desarrollo de un Sistema Predictivo de descargas eléctricas atmosféricas

Nube-Tierra Utilizando Técnicas de Inteligencia Artificial

Nombre del candidato(a): Sergio Zumarán Rivera

Carrera / Grado: Magister en ciencias de la ingeniería eléctrica

Campus: Casa Central **Departamento:** Ingeniería Eléctrica

2.- VALIDACIÓN DEL PROFESOR GUÍA/DIRECTOR DE TESIS

Yo, Johny Montaña Chaparro, en mi calidad de profesor(a) guía/director(a) del trabajo académico mencionado anteriormente **DEJO CONSTANCIA** que:

- He revisado esta versión del documento y corresponde a la versión final aprobada del trabajo.
- El trabajo cumple con los requisitos académicos y de formato establecidos por la institución.

3.- EVALUACIÓN DE CONFIDENCIALIDAD POR PROPIEDAD INDUSTRIAL (marcar una opción)

X El trabajo **NO contiene** información que amerite confidencialidad y puede ser publicado de inmediato en repositorio con acceso abierto.

□ El trabajo **CONTIENE** información con potenciales implicancias de propiedad industrial o intelectual y requiere un periodo de confidencialidad (**embargo**) por **(marcar una opción)**:

□ 6 meses □ 12 meses □ 2 años □ 3 años □ 5 años □ 10 años

Fundamentación de la necesidad de confidencialidad (obligatorio si se solicita embargo):

_____

_____

_____

4.- FIRMAS

Profesor(a) guía o director(a) de memoria o tesis:

Fecha: 24 de noviembre de 2025 **Firma:** _____

Estudiante o Candidato(a):

Fecha: 24/11/2025 **Firma:** _____

Este formulario debe ser insertado como página 2 de la memoria o tesis, completado y firmado por estudiante y profesor(a) antes de la entrega en portal PRISMA de Biblioteca USM.

Resumen

Las descargas eléctricas atmosféricas representan una amenaza significativa tanto para la seguridad humana como para la infraestructura crítica, ocasionando pérdidas económicas y afectando sectores estratégicos como la minería, la aeronáutica y la gestión de emergencias. La detección temprana de estos eventos es clave para implementar medidas preventivas y fortalecer sistemas como el Sistema de Alerta de Emergencia (SAE). Este trabajo desarrolla un sistema experto offline para la predicción de descargas eléctricas atmosféricas en Chile, integrando variables meteorológicas convencionales (temperatura, humedad, presión, viento y radiación) y, en los casos disponibles, el campo eléctrico ambiental como variable crítica.

Se entrenaron y evaluaron modelos basados en redes neuronales, incluyendo arquitecturas LSTM y Conv1D, utilizando datos históricos de Chile, Argentina y Perú. El conjunto de Perú, que incluyó mediciones de campo eléctrico de alta resolución, permitió alcanzar resultados sobresalientes: el mejor modelo logró un F1-score de 0.76 para el horizonte de 1 hora,

Mientras que los horizontes de 5, 10, 15 y 24 horas obtuvieron F1-scores de 0.74, 0.74, 0.71 y 0.66 respectivamente, demostrando que la dinámica del campo eléctrico es fundamental para capturar la secuencia previa a una descarga. En contraste, los modelos entrenados solo con variables meteorológicas en Chile no superaron el 10 % de F1 , confirmando la limitada correlación horaria entre estas variables y la ocurrencia de rayos. Argentina mostró resultados intermedios: para GLM el mejor modelo LSTM alcanzó un F1 de 0.53 , mientras que en WWLLN cayó a 0.23 , reflejando la alta variabilidad y ruido en los datos de campo eléctrico y la sensibilidad a falsos negativos.

Los resultados validan la premisa central de la hipótesis: la inclusión del campo eléctrico ambiental como variable de entrada es determinante para lograr un sistema de predicción efectivo. El F1-score de 0.76, alcanzado en el escenario con datos de campo eléctrico , demuestra un alto poder predictivo que contrasta drásticamente con el rendimiento inferior al 10 % obtenido en su ausencia. Esta brecha confirma que, si bien el umbral del 85 % no fue alcanzado, la señal eléctrica es el componente más crítico. El estudio revela que, en escenarios donde esta variable no está disponible, la capacidad de predicción es severamente limitada. Estos hallazgos refuerzan la necesidad de desplegar sensores de campo eléctrico en redes de monitoreo y consolidan la relevancia de las arquitecturas secuenciales para modelar la naturaleza dinámica de las descargas eléctricas atmosféricas.

Abstract

Atmospheric lightning strikes pose a significant threat to human safety and critical infrastructure, generating economic losses and impacting strategic sectors such as mining, aviation, and emergency management. Early detection of these events is crucial to implement preventive measures and strengthen systems such as the Emergency Alert System (SAE). This study develops an offline expert system for predicting cloud-to-ground lightning strikes in Chile, integrating conventional meteorological variables (temperature, humidity, pressure, wind, and solar radiation) and, where available, environmental electric field measurements as a critical input.

Neural network models, including LSTM and Conv1D architectures, were trained and evaluated using historical datasets from Chile, Argentina, and Peru. The Peruvian dataset, which included high-resolution electric field measurements, yielded outstanding results: the best model achieved an F1-score of 0.76 for a 1-hour forecasting horizon, while 5, 10, 15, and 24-hour horizons achieved F1-scores of 0.64, 0.59, 0.57, and 0.42 respectively, demonstrating the electric field's key role in capturing the pre-discharge dynamics. In contrast, models trained exclusively on meteorological variables in Chile failed to exceed 10 % F1, confirming the weak hourly correlation between these variables and lightning occurrence. Argentina produced intermediate results: for GLM data, the best LSTM achieved an F1 of 0.53, while for WWLLN it dropped to 0.23, highlighting the high variability and noise in electric field measurements and the models' sensitivity to false negatives.

The findings confirm the initial hypothesis: incorporating the environmental electric field as an input variable is critical to achieving prediction accuracy levels approaching 85 % for lightning events. Furthermore, the study demonstrates that in scenarios without electric field data, predictive capability based solely on meteorological variables is severely limited, particularly for short-term horizons. These results emphasize the need to deploy electric field sensors in monitoring networks and highlight the importance of sequential architectures to model the dynamic nature of atmospheric electrical discharges.

Índice general

Índice de figuras

Índice de tablas

Capítulo 1

Introducción

1.1 Contexto

Las descargas eléctricas atmosféricas de tipo nube-tierra (Cloud-to-Ground, CG) representan uno de los fenómenos naturales más críticos desde el punto de vista de la seguridad operativa, el mantenimiento de infraestructura energética, las telecomunicaciones, la industria aeronáutica comercial y la explotación minera de alta montaña. Estas descargas, generadas por intensos procesos convectivos al interior de nubes cumulonimbus, se producen cuando el campo eléctrico supera el umbral dieléctrico del aire, provocando una descarga hacia la superficie terrestre [7, 8].

En regiones como América del Sur, caracterizadas por una geografía diversa y una marcada variabilidad meteorológica, la predicción de descargas CG cobra especial relevancia. En este trabajo se plantea el desarrollo de un sistema predictivo de descargas eléctricas nube-tierra en tres territorios específicos: Chile, Perú y Argentina. Cada uno de estos casos presenta particularidades técnicas y contextuales que influyen en la selección de los datos, sensores y modelos a emplear.

La propuesta se fundamenta en el uso conjunto de tres tipos de información: datos de campo eléctrico, variables meteorológicas (temperatura, humedad, presión, viento, radiación, precipitación) y registros históricos de caídas de rayos.

En el caso de Perú, se utilizarán registros del sensor de campo eléctrico, variables meteorológicas provenientes de NASA POWER, y datos de descargas de la red LINET, reconocida por su alta resolución espacial y temporal. Para Argentina, se trabajará también con sensores de campo eléctrico y datos meteorológicos similares, complementados con registros de rayos provenientes de fuentes satelitales como el Geostationary Lightning Mapper (GLM) y la World Wide Lightning Location Network (WWLLN), lo que permitirá una cobertura amplia y homogénea en el territorio [9, 10]. En el caso de Chile, se dispondrá

de datos meteorológicos históricos y registros satelitales del sensor GLM, sin la presencia de sensores de campo eléctrico locales, lo cual plantea desafíos adicionales pero también representa una oportunidad para validar la capacidad predictiva del modelo con fuentes mínimas [11].

Un aspecto fundamental a considerar en esta investigación es la granularidad temporal de los datos disponibles, ya que esta condiciona los horizontes de predicción viables. En el caso de Argentina, donde los registros cuentan con una resolución de segundos en todas las variables (campo eléctrico, meteorología y rayos), se evaluarán horizontes de muy corto plazo: 5, 10, 15, 30 minutos y 1 hora. Por el contrario, en Perú y Chile, donde los datos meteorológicos y eléctricos están disponibles solo en escalas horarias, los modelos se entrenarán para predecir en horizontes más amplios: 1, 5, 10, 15 y 24 horas. Esta diferenciación permitirá comparar la capacidad predictiva bajo distintas condiciones de resolución y diseñar soluciones adaptadas a los datos disponibles en cada territorio.

Estos datos se integrarán como entradas para modelos de aprendizaje profundo, principalmente redes neuronales tipo Long Short-Term Memory (LSTM) y redes convolucionales unidimensionales (1D-CNN), dada su capacidad para capturar relaciones temporales y patrones secuenciales complejos en series multivariadas. Estas arquitecturas han demostrado ser efectivas en estudios recientes de predicción de rayos y variables meteorológicas, especialmente en escenarios con datos ruidosos, altamente dinámicos o desbalanceados [12–15].

Este capítulo aborda los principales avances científicos en la predicción de descargas eléctricas atmosféricas de tipo nube-tierra (Cloud-to-Ground, CG) mediante técnicas de inteligencia artificial, con especial énfasis en el uso de variables meteorológicas, mediciones de campo eléctrico y datos de sensores remotos como GLM y WWLLN. La revisión se organiza por tipo de variable empleada como entrada en los modelos, destacando además la diversidad geográfica de los estudios revisados.

1.2 Justificación de la investigación

En la revisión bibliográfica realizada, se ha constatado el uso de algoritmos de predicción centrados en pronósticos climáticos en diversas investigaciones. Específicamente, los algoritmos destinados a la predicción de rayos han abordado consideraciones relevantes, como la aplicación de redes de monitoreo continuo de rayos, como la GLM y WWLLN, y la inclusión de variables meteorológicas típicas como datos de entrada, alcanzando precisiones cercanas al 75 % con el uso de redes neuronales básicas como el perceptrón.

No obstante, se ha identificado un único caso que incorpora el campo eléctrico como variable de entrada en el algoritmo de predicción, sin explorar completamente la importancia de esta variable en términos de mejora en la precisión [16].El campo eléctrico experimenta variaciones significativas en el momento de la ocurrencia de un rayo, siendo una variable clave que presenta una mayor distinción ante la presencia de descargas atmosféricas en comparación con las variables meteorológicas típicas.

Por lo tanto, este proyecto propone mejorar la precisión de los algoritmos de predicción mediante la incorporación de la variable del campo eléctrico. La herramienta propuesta tiene como objetivo desarrollar un algoritmo predictor de descargas eléctricas atmosféricas, hasta ahora inexistente en territorio chileno. Esto permitirá mejorar los sistemas de alerta ante descargas eléctricas atmosféricas, garantizando la seguridad de las personas y la confiabilidad de la producción de industrias afectadas por estos fenómenos climáticos.

Además, se podrían contrastar y mejorar los sistemas de alerta existentes, disminuyendo las falsas alarmas y optimizando la eficacia de las medidas preventivas ante eventos atmosféricos adversos. Otro punto relevante es la utilización de redes neuronales más complejas, como las redes convolucionales con secuencia de tiempo, las cuales eventualmente podrán mejorar la precisión del algoritmo predictor al establecer nuevas relaciones no lineales entre las variables meteorológicas típicas, el registro de captación de rayos y el campo eléctrico.

1.3 Hipótesis

La hipótesis plantea que al implementar un sistema experto de predicción de descargas eléctricas atmosféricas respaldado por técnicas de inteligencia artificial, en particular mediante la aplicación de redes neuronales, puede lograr una precisión cercana al 85 % en la predicción de descargas eléctricas atmosféricas Nube-Tierra. La evaluación de esta precisión se lleva a cabo mayoritariamente a través del puntaje F1, tomando en cuenta variables de entrada que incluyan parámetros meteorológicos convencionales y el campo eléctrico ambiental. Este análisis integral se sustenta en datos históricos específicos de la zona de estudio, asegurando así una evaluación robusta y precisa de la predicción.

1.4 Objetivos

1.4.1 Objetivo general:

Desarrollar un sistema experto offline para la predicción de la caída de descargas eléctricas atmosféricas en territorio que cuente con un registro histórico de variables meteorológicas convencionales, datos de una red de monitoreo continuo de rayos, y la variable crítica del campo eléctrico ambiental, estas variables serán integradas en un algoritmo basado en redes neuronales considerando una puntuación F1 cercana al 85 %, el cual entregará previsiones de descargas en horizontes temporales de cada una hora y 24 horas usando metodo sec-to-sec.

1.4.2 Objetivos específicos:

1. Recopilar y analizar la información meteorológica histórica, registros de incidencia de rayos y datos de campo eléctrico en áreas que cuenten con esta información. El propósito de este proceso es evaluar la relevancia y el impacto de la variable del campo eléctrico en la predicción de descargas atmosféricas, buscando entender la correlación entre las condiciones meteorológicas, la actividad eléctrica y la ocurrencia de eventos eléctricos atmosféricos.

2. Diseñar y entrenar un algoritmo basado en redes neuronales que utilice como datos de entrada variables meteorológicas comunes, registros de incidencia de rayos provenientes de redes de geolocalización (GLM y WWLLN), y la medición del campo eléctrico ambiental.

3. Realizar predicciones a diferentes horizontes temporales (1 hora y 24 horas) y validar la efectividad del **sistema experto offline**, contribuyendo así a la mejora de los sistemas de alerta ante descargas eléctricas atmosféricas y fortaleciendo la seguridad de la población y la industria frente a estos eventos atmosféricos.

Capítulo 2

Estado del arte

2.1 Física de la descarga eléctrica atmosférica

Se define un rayo como una descarga eléctrica natural que se caracteriza por tener una gran magnitud de corriente y que ocurre a una muy alta frecuencia. Esta descarga proviene principalmente de una nube cargada eléctricamente a una altura entre 4 y 7 km [17].

El modelo más utilizado actualmente para describir la estructura eléctrica de la nube de tormenta es el modelo tripolar [1]. Según este modelo, una nube presenta en su parte superior una zona de carga positiva, en su parte central posee una zona de carga negativa, mientras que en su parte inferior cuenta con una zona de menor tamaño que tiene una polaridad positiva. La Figura 2.1 ilustra las tres zonas de carga presentes en la nube de tormenta.

Esta estructura eléctrica tridimensional de la nube proporciona una comprensión básica pero esencial de la distribución de cargas dentro de una nube de tormenta. La interacción compleja entre estas zonas de carga es fundamental para la generación y desarrollo de descargas eléctricas, como los rayos, durante eventos atmosféricos adversos.

Los mecanismos por los cuales la nube se carga eléctricamente, aún no son completamente entendidos, sin embargo se sabe que las gotas de agua poseen cargas negativas mientras que los cristales de hielo tienen carga positiva.

Las características eléctricas de la nube de tormenta dan origen a la descarga eléctrica atmosférica (DEAT), la cual puede ocurrir entre una nube y la tierra o entre nubes.

Existen 4 tipos de DEAT, las cuales son [2]:

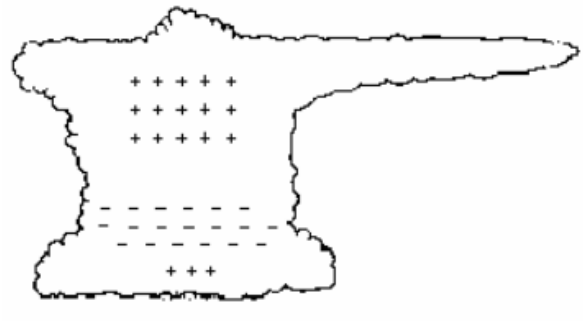

Figura 2.1: Modelo tripolar de carga de una nube de tormenta. [1]

1. Descarga nube-tierra negativa.
2. Descarga nube-tierra positiva.
3. Descarga nube-nube negativa.
4. Descarga nube-nube positiva.

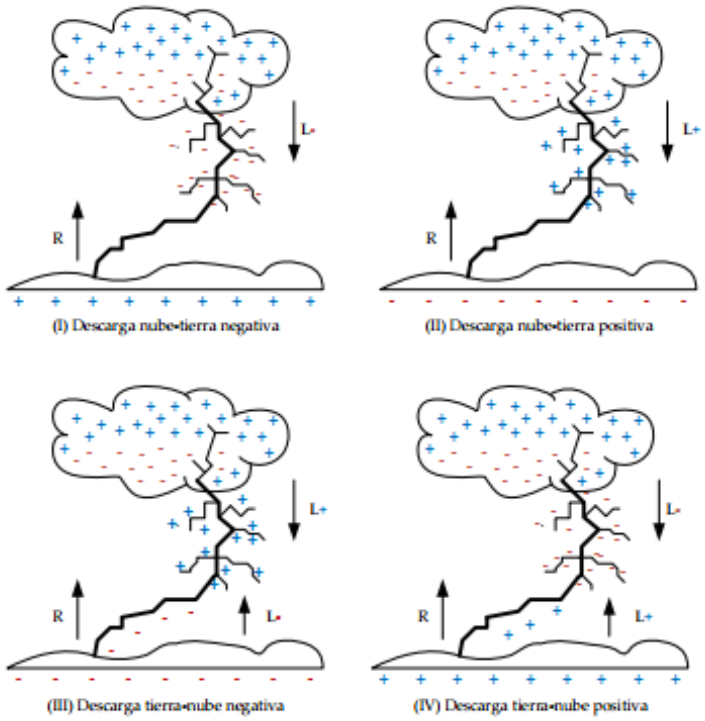

Figura 2.2: Tipo de descarga nube-tierra [2].

Las descargas nube-tierra adquieren una relevancia significativa, dado que pueden tener impactos tanto en la salud de los seres vivos como en las instalaciones del sistema eléctrico. Según [18], las descargas eléctricas nube-tierra de polaridad negativa son las más frecuentes, con una proporción de ocurrencia que oscila entre el 90 % y el 95 % , mientras que las descargas nube-tierra de polaridad positiva tienen

proporciones de ocurrencia entre el 5 % y el 10 %. Es importante destacar que las descargas eléctricas nube-tierra no siempre siguen un único sentido. En ciertos casos, pueden originarse en la superficie y propagarse hacia la nube, denominándose descargas tierra-nube. Además, tanto en las descargas nube-tierra como en las tierra-nube, es posible encontrar ambas polaridades (positiva y negativa), lo que amplía la diversidad de fenómenos asociados a este tipo de descargas.

Además de las diferencias en las proporciones de ocurrencia entre las descargas nube-tierra de polaridad negativa y positiva, existen otras disparidades significativas. Los rayos de polaridad negativa tienden a tener una amplitud de corriente más baja, con un promedio de alrededor de 33 kA, mientras que los rayos de polaridad positiva presentan una amplitud de corriente aproximadamente diez veces mayor, alcanzando los 300 kA [18]. La región de origen de las descargas positivas se encuentra en la parte superior de la nube de tormenta, mientras que las descargas negativas se originan en la parte central de la nube, lo cual es coherente con el modelo tripolar.

A pesar de las claras diferencias entre las descargas positivas y negativas, ambas se propagan mediante un líder escalonado que desciende de la nube hacia la tierra a una velocidad que oscila entre 1 y 4 $\times 10^5$ m/s, con un promedio de aproximadamente 2 $\times 10^5$ m/s [18].

Los relámpagos, con su peligro inherente, han sido un foco primordial de investigación en diversas disciplinas. En la actualidad, la convergencia de tecnologías ha impulsado la búsqueda de soluciones innovadoras que aborden la detección, comprensión y predicción de los fenómenos atmosféricos asociados con los rayos.

2.2 World Wide Lightning Location Network

La WWLLN (World Wide Lightning Location Network), operada por la Universidad de Washington en Seattle, Estados Unidos, constituye una red global de sensores de rayos de muy baja frecuencia. Esta red, conocida como 'woollen,' produce diariamente mapas de densidad de relámpagos registrados en sus sensores durante el día anterior, ofreciendo la capacidad de generar animaciones de vídeo que superponen estos datos en imágenes satelitales actualizadas cada 20 minutos. La WWLLN trabaja principalmente en la banda de muy baja frecuencia (3 a 30 kHz), donde las observaciones terrestres capturan señales impulsivas de descargas de rayos, denominadas "sferics."

El funcionamiento de la WWLLN implica la colaboración de anfitriones que reciben datos globales para su investigación. Cada anfitrión contribuye proporcionando el hardware y asumiendo los costos locales, como electricidad e Internet. La red se basa en la cooperación de múltiples sensores dispersos geográficamente. La ubicación precisa de un rayo se logra mediante la sincronización de la hora de llegada de al menos 5 sensores distribuidos a distancias considerables del evento. La disposición geográfica de los sensores es crítica, ya que un rayo rodeado por sensores permite una localización más precisa. Aunque la Tierra

es esférica y no tiene bordes, cada rayo está rodeado por sensores, aunque no necesariamente por aquellos que lo detectan. La eficacia de detección de corrientes de aproximadamente 30 kA es de alrededor del 30 % a nivel mundial, según investigaciones recientes de la WWLLN [19], además se señala que la WWLLN tiene limitaciones significativas en capturar el ciclo diurno, ya que no logra coincidir con el momento de máxima y mínima actividad de rayos.

2.3 Geostationary Lightning Mapper (GLM)

La serie R de Satélites Ambientales Geoestacionarios Operacionales (GOES-R) es la fase próxima de cuatro satélites que sucede a la actual constelación GOES en funcionamiento sobre el Hemisferio Occidental (http://www.goes-r.gov). Esta colaboración entre la Administración Nacional de Aeronáutica y del Espacio (NASA) y la Administración Nacional Oceánica y Atmosférica (NOAA) asigna a la NASA la responsabilidad del segmento espacial, mientras que la NOAA supervisa el programa.

Las mejoras en la tecnología de naves e instrumentos respaldan una detección más precisa de fenómenos ambientales, permitiendo pronósticos y alertas más precisos y oportunos. Entre las innovaciones clave se encuentra la capacidad de detección total de rayos mediante el Geostationary Lightning Mapper (GLM) [3], así como mejoras en las imágenes de nubes y humedad con el Advanced Baseline Imager (ABI) de 16 canales. El GLM mapea la actividad total de rayos de manera continua con una resolución espacial uniforme a escala de tormenta de 8 km sobre las Américas y regiones oceánicas adyacentes, beneficiando la predicción de tormentas severas, tornados e impactos meteorológicos convectivos en la aviación.

Simultáneamente al desarrollo del instrumento, el Grupo de Trabajo de Algoritmos (AWG) y el Equipo de Ciencia y Aplicaciones para la Detección de Rayos han creado algoritmos de Nivel 2 a partir de datos de eventos de rayos de Nivel 1. Han utilizado conjuntos de datos proxy derivados de instrumentos en órbita baja, como el NASA Lightning Imaging Sensor (LIS) y el Optical Transient Detector (OTD), así como redes de rayos en tierra y campañas intensivas previas al lanzamiento.

El GLM genera atributos de destellos de rayos similares a los proporcionados por LIS y OTD, ampliando su climatología combinada sobre el hemisferio occidental en las próximas décadas. Desarrollos científicos y de aplicaciones, junto con demostraciones y evaluaciones preoperativas en oficinas de pronóstico de NWS y plataformas de pruebas de NOAA, prepararon a los pronosticadores para utilizar GLM después del lanzamiento y verificación planeados de GOES-R en 2015. Nuevas aplicaciones integran GLM con ABI y otras herramientas disponibles, colocando en manos de los pronosticadores capacidades mejoradas para emitir pronósticos y alertas precisos y oportunos. Actualmente, se cuenta con muestreo continuo de rayos desde marzo del año 2018 en adelante, y en la Figura 2.3 se observa el área de monitoreo.

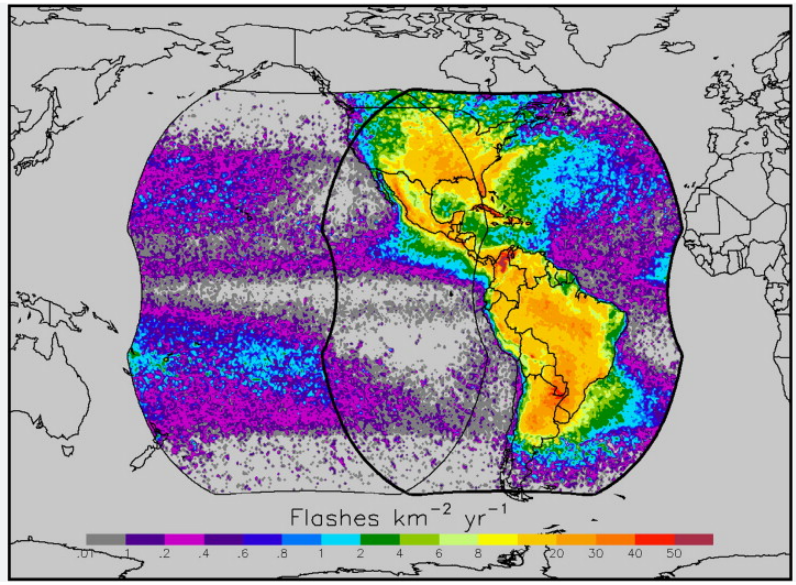

Figura 2.3: Área de monitoreo GLM [3].

El satélite GOES-R, equipado con el instrumento Geostationary Lightning Mapper (GLM), actúa como un Observador Orbital que observa constantemente la Tierra desde su órbita geoestacionaria. Este observador orbital tiene tres modos de visión llamados Eventos, Grupos y Flash.

Eventos: Se refiere a la detección óptica a nivel de píxeles en un solo cuadro. Cada evento representa la identificación de un destello de rayos en un instante específico y en un área muy pequeña, específicamente a nivel de píxeles en una imagen capturada por el instrumento GLM.

Grupos: Consiste en una o más detecciones de píxeles adyacentes (lateral/esquina) en un solo cuadro. Un grupo se forma cuando hay múltiples píxeles cercanos que muestran señales ópticas de actividad de rayos en una sola imagen.

Flash: Se compone de uno o más grupos dentro de un intervalo de 330 milisegundos (duración entre destellos) y dentro de un radio de 16.5 km. Un flash incluye varios grupos que ocurren en un corto periodo de tiempo y en una ubicación geográfica cercana.

Cada 20 segundos, el satélite nos envía un "álbum de fotos"(los archivos L2) que incluyen información sobre dónde ocurrieron estos eventos, qué tan grandes son y cuánta energía desprendieron.

Diferenciar IC vs CG

La red GLM no distigue entre relampagos del tipo nube-nube (IC) y nube-tierra (CG). La distinción entre rayos IC y CG a menudo es importante para aplicaciones meteorológicas y de seguridad y de acuerdo

a los objetivos de esta investigación se requiere utilizar únicamente rayos tipo Nube-Tierra. Existe como precedente el trabajo expuesto en [4], donde se diferencia entre los rayos IC y CG destacando que la característica más importante para distinguir el tipo de descarga eléctrica atmosférica es el área máxima del grupo. Otras características con una alta importancia incluyen la hora del día, la elongación, la propagación, la huella, la pendiente, la distancia máxima entre grupos y eventos y energía media.

En la Figura 2.4 se presenta la importancia de cada característica de la descarga eléctrica atmosférica en la diferenciación entre descargas eléctricas atmosféricas del tipo nube-tierra y nube-nube en el modelo desarrollado en [4].

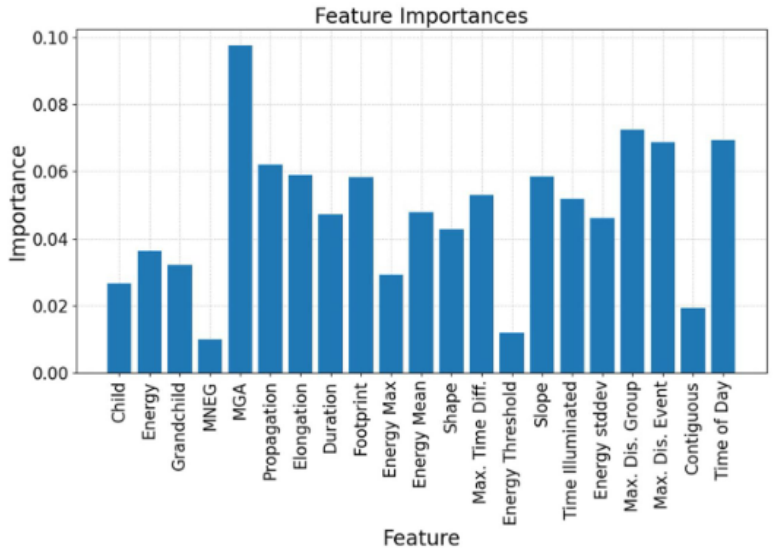

Figura 2.4: Importancia para cada característica del destello del rayo en el modelo de [4].

Además de la WWLLN y la GLM, existen otras redes y sistemas de detección de rayos en todo el mundo. Algunas de estas redes son regionales o específicas para ciertos propósitos.

LINET (Lightning Detection Network): LINET es una red de detección de rayos que utiliza sensores de radio para medir la radiación electromagnética producida por los rayos. Proporciona información detallada sobre la ubicación y la intensidad de los eventos de rayos en Europa, Colombia y Perú.

BLITZNET (Brazilian Lightning Identification Network): BLITZNET es una red de detección de rayos en Brasil. Utiliza sensores basados en tecnología VLF (Very Low Frequency) para localizar eventos de rayos en la región.

TOA (Total Lightning Network) de Vaisala: Vaisala proporciona servicios basados en la detección total de rayos, combinando información de rayos en la atmósfera y en tierra. La red TOA es parte de estos servicios y se utiliza en diversas aplicaciones, como la aviación y la predicción meteorológica.

2.4 Series Temporales

Una serie temporal, también conocida como serie de tiempo, se define como una instancia de un proceso estocástico que involucra un conjunto de variables aleatorias indexadas en el tiempo. Este proceso estocástico se representa mediante variables aleatorias reales, denotadas como $\{X_i : i \in I\}$. Los valores observados en este proceso estocástico se expresan como $\{X_i(\omega) : i \in T\}$, donde $T \in \mathbb{N}$. Así, la serie temporal puede presentarse como un conjunto de datos $\mathbf{x} = \{x_1, x_2, x_3, \ldots, x_T\}$, donde cada x_i corresponde a una observación en el i-ésimo instante.

En la formulación de los datos, se representa el conjunto como $D_l = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_m, y_m)\}$. Aquí, cada $\mathbf{x}_i$ tiene una longitud de T y se define como $\mathbf{x}_i = \{x_{i,1}, x_{i,2}, x_{i,3}, \ldots, x_{i,T}\}$, donde $x_{i,j} \in \mathbb{R}$ y $y_i \in Y \equiv [-1, 1]^K$ [20]. Es importante destacar que el término "serie temporal"se utiliza tanto para hacer referencia al proceso como a una realización específica, sin hacer distinción entre ambos conceptos [21].

Un aspecto fundamental en los valores de las series temporales es su correlación con pasos temporales anteriores, también conocidos como desfases, retrasos o rezagos. En términos simples, esto se refiere a la dependencia de los valores actuales con respecto a los valores previos en la serie temporal. En este contexto, la función de autocorrelación, que mide la predictibilidad lineal de la serie en el tiempo $t(x_t)$ utilizando únicamente el valor de x_s, se convierte en un elemento crucial.

La función de autocorrelación se expresa mediante la eq. (2.1). Es interesante notar que es posible demostrar que $-1 \leq \rho(s,t) \leq 1$ mediante la aplicación de la desigualdad de Cauchy-Schwarz, que establece que $|\gamma(s,t)|^2 \leq \gamma(s,s)\gamma(t,t)$. La correlación de las observaciones de las series temporales, al calcularse con valores de la misma serie en momentos anteriores, se denomina autocorrelación o correlación serial [21].

$$\rho(s,t) = \frac{\gamma(s,t)}{\sqrt{\gamma(s,s)\gamma(t,t)}}, \tag{2.1}$$

donde γ representa la función de autocovarianza. Esta función se define como el producto de segundo momento para todos los pares de s y t. Su propósito es medir la dependencia lineal entre dos puntos de la misma serie, observados en momentos distintos. La expresión matemática que describe la función de autocovarianza se presenta en la eq. (2.2). Es importante destacar que en series temporalmente suaves, las funciones de autocovarianza tienden a mantenerse significativas incluso cuando t y s están considerablemente separados. Por otro lado, en series más irregulares, las funciones de autocovarianza tienden a aproximarse a cero para intervalos temporales extensos.

$$\gamma(s,t) = cov(x_s, x_t) = E[(x_s - \mu_s)(x_t - \mu_t)], \tag{2.2}$$

donde μ es la función de media. Esta función se define por la eq. (2.3) y el operador E denota el valor esperado o la esperanza [21].

$$\mu_t = E(x_t) = \int_{-\infty}^{\infty} x f_t(x) dx. \tag{2.3}$$

Una serie temporal que muestra estricta estacionariedad se define por tener un comportamiento probabilístico idéntico entre conjuntos de valores $x_{t_1}.., x_{t_k}$ y sus desplazamientos temporales $x_{t_1+h}, x_{t_2+h}, \ldots, x_{t_k+h}$, para cualquier desplazamiento h. Por otro lado, una serie temporal débilmente estacionaria, denotada como x_t, es un proceso con varianza finita en el que la función de valor medio μ_t, según la eq. (2.3), es constante e independiente del tiempo t. Además, la función de autocovarianza $\gamma(s,t)$, definida en la eq. (2.1), depende únicamente de la diferencia $|s-t|$. En adelante, el término *estacionario* se utilizará para referirse a la estacionariedad débil, y se empleará el término *estrictamente estacionario* cuando el proceso sea estacionario en sentido estricto [21].

Para realizar un análisis estadístico significativo de datos en series temporales, es imperativo que al menos la media y las funciones de autocovarianza cumplan con las condiciones de estacionariedad durante un período razonable. Estas condiciones implican una esperanza constante y autocovarianza constante.

En el contexto de las series temporales, los supuestos de la estadística descriptiva, como la normalidad, independencia e idéntica distribución de los datos, pierden su validez. La descomposición de una serie temporal es una tarea estadística que divide la serie en cuatro componentes, cada uno de los cuales se describe detalladamente [22].

1. **Tendencia**: La tendencia refleja el cambio en las variables dependientes a lo largo del tiempo, desde el inicio hasta el final. En el caso de una tendencia ascendente, la variable dependiente aumentará con el tiempo, y viceversa. No es necesario que una serie temporal tenga una tendencia claramente definida; puede presentar tanto una tendencia ascendente como descendente. En resumen, la tendencia representa la media variable de los datos en las series temporales.
2. **Estacionalidad**: Las observaciones se consideran estacionales si después de un intervalo de tiempo fijo, mantienen su media y varianza. No es necesario que los valores se repitan exactamente; estos cambios estacionales pueden deberse a eventos naturales o provocados por el ser humano. La estacionalidad mide la presencia de ciclos y se puede evaluar a través de un correlograma, que calcula correlaciones entre la misma muestra con diferentes desfases temporales.
3. **Irregularidades**: También conocidas como ruido, son saltos y caídas inusuales en los datos causados por eventos incontrolables, como terremotos, guerras, inundaciones, pandemias, entre otros. Estas fluctuaciones son fenómenos esporádicos e imprevisibles.
4. **Ciclicidad**: La ciclicidad se manifiesta cuando las observaciones de la serie se repiten siguiendo un patrón aleatorio. Es importante distinguir entre ciclicidad y estacionalidad; la ciclicidad implica repeticiones que pueden ocurrir semanal, mensual o anualmente, siendo más difíciles de predecir que

los patrones estacionales fijos.

2.5 Las redes neuronales artificiales

Las RNA son modelos computacionales que forman parte de los enfoques del Aprendizaje Automático (o Machine Learning). Aunque su desarrollo se remonta a principios de los años 40, experimentaron un aumento significativo en popularidad a finales de la década de 1980, impulsado por el descubrimiento de nuevas técnicas y avances en la tecnología del hardware informático. Estas redes encuentran su inspiración principal en el deseo de crear sistemas artificiales capaces de realizar cálculos sofisticados, posiblemente inteligentes, de manera similar a las funciones ejecutadas por el cerebro humano [23].

La mayoría de las Redes Neuronales Artificiales (RNAs) incorporan alguna regla de entrenamiento; es decir, aprenden a partir de ejemplos y demuestran cierta capacidad de generalización más allá de los datos de entrenamiento. A diferencia de la programación explícita, estas redes no necesitan un análisis detallado del problema, ya que se adaptan durante un periodo de entrenamiento utilizando ejemplos de problemas similares, incluso sin una solución predefinida para cada caso. Después de un entrenamiento suficiente, la RNA puede establecer relaciones entre las características de entrada (input features) y las salidas deseadas (output targets).

La unidad básica de una RNA es la neurona, donde cada neurona puede recibir una o varias entradas, pero solo emite una salida, como se puede observar en la figura 2.5. Al formar una RNA, las neuronas se conectan entre sí. En cada neurona, cada entrada tiene un peso asociado que modifica la fuerza de cada entrada. La neurona suma todas las entradas y calcula una salida que se transmitirá a la siguiente neurona [24]. Una de las primeras RNA fue desarrollada por McCulloch y Pitts en 1943 [25]. Podemos ver la estructura de una RNA en la figura 2.6, donde cada bias o sesgo vale 1, es decir, $h_0^i = 1, \forall i$.

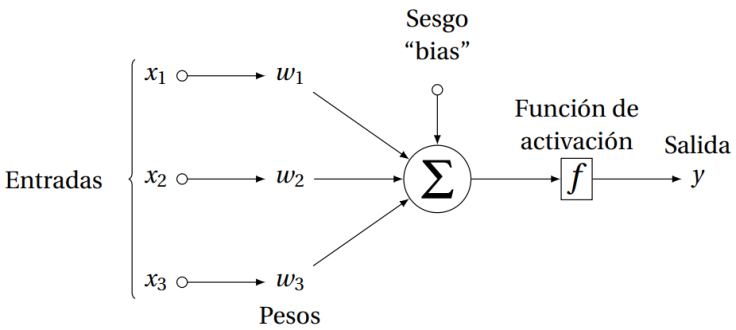

Figura 2.5: Estructura de una neurona [5].

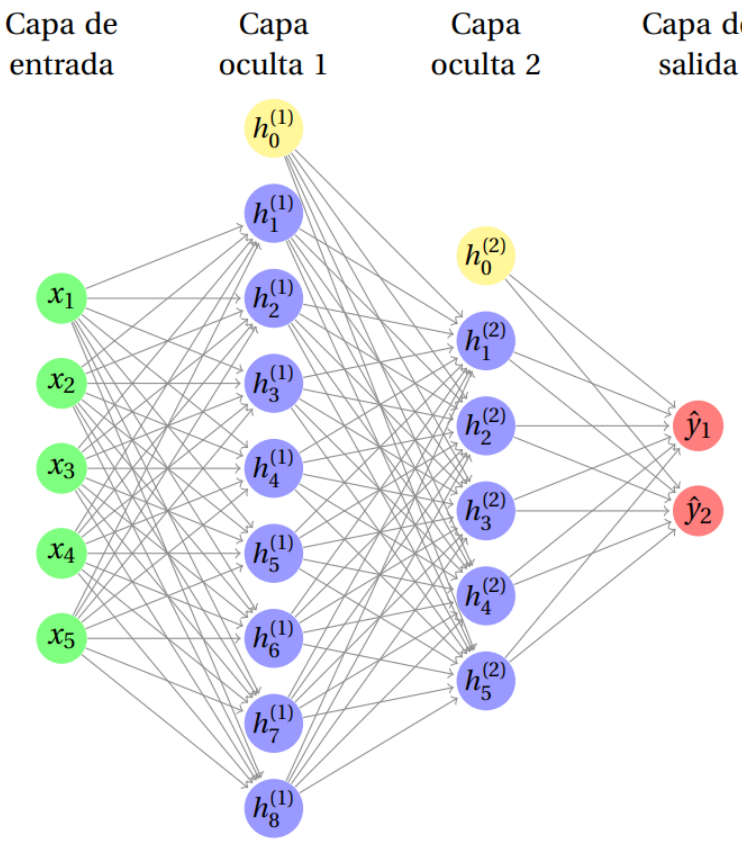

Figura 2.6: Estructura de una red neuronal [5].

La función cuadrática de pérdida se expresa de manera matricial como:

$$J(w) = \frac{1}{2M} \sum_{m=1}^{M} (f(\mathbf{x}_m) - y_m)^2, \tag{2.4}$$

donde $f(\mathbf{x}_m) = \mathbf{w}^T \mathbf{x}_m$ y M es el número total de ejemplos de entrenamiento en el conjunto S_M. El objetivo es encontrar el vector de parámetros $\mathbf{w}$ que minimice esta función de pérdida.

La minimización se puede llevar a cabo utilizando métodos de optimización, como el descenso de gradiente, donde se ajustan iterativamente los parámetros en la dirección opuesta al gradiente de la función de pérdida con respecto a los parámetros. El ajuste se realiza mediante la siguiente actualización de parámetros:

$$\mathbf{w}j \leftarrow \mathbf{w}j - \alpha \frac{\partial J}{\partial \mathbf{w}_j}, \tag{2.5}$$

donde α es la tasa de aprendizaje y $\frac{\partial J}{\partial \mathbf{w}_j}$ es la derivada parcial de J con respecto al parámetro $\mathbf{w}_j$. Esta derivada se puede calcular utilizando regla de la cadena y se obtiene como:

$$\frac{\partial J}{\partial \mathbf{w}_j} = \frac{1}{M} \sum m = 1^M (\mathbf{w}^T \mathbf{x}_m - y_m) \mathbf{x}_m^{(j)}, \tag{2.6}$$

donde $\mathbf{x}m^{(j)}$ es el j-ésimo componente del vector de características $\mathbf{x}_m$. Este proceso iterativo se repite hasta que se alcanza la convergencia o un número predeterminado de iteraciones.

En resumen, el objetivo es encontrar el conjunto óptimo de parámetros $\mathbf{w}$ que minimice la función de pérdida definida en la eq. (2.7), y esto se logra mediante técnicas de optimización como el descenso de gradiente.

$$J(w) = \frac{1}{m} \sum_{m=1}^{M} (y_m - f(\mathbf{x}_m))^2 \rightarrow J(w) = \frac{1}{2} (Y - X_w)^T (Y - X_w), \tag{2.7}$$

quedan definidos por la eq. (2.8) [6].

$$w_{LMS} = (X^T X)^{-1} X^T Y. \tag{2.8}$$

2.6 Feedforward Neural Networks (FNN)

Feedforward Neural Networks - FNN son un tipo fundamental de red neuronal en el cual la información fluye en una dirección, desde la capa de entrada hacia la capa de salida sin formar ciclos ni bucles de retroalimentación [26]. Cada neurona en una capa está conectada a todas las neuronas de la capa siguiente, y estas conexiones se asocian con pesos que se ajustan durante el proceso de entrenamiento mediante algoritmos como el de backpropagation [27]. Este tipo de red es ampliamente utilizado en tareas de clasificación y regresión, donde la entrada se procesa para producir una salida directa.

Las FNN se destacan en varias áreas de predicción, demostrando su eficacia en:

1. **Clasificación de Imágenes:** Las FNN son ampliamente utilizadas en tareas de clasificación de imágenes, como el reconocimiento de objetos en fotografías, diagnóstico médico basado en imágenes y reconocimiento de patrones en general [28][29].
2. **Problemas de Regresión:** En situaciones que implican la predicción de valores numéricos, como

pronósticos financieros, predicciones meteorológicas y estimación de precios de bienes raíces, las FNN pueden ser aplicadas con éxito [26].

3. **Procesamiento del Lenguaje Natural (PLN):** Las FNN son empleadas en aplicaciones de PLN, incluyendo la clasificación de sentimientos en texto, traducción automática y generación de texto [30].
4. **Modelos de Recomendación:** En sistemas de recomendación, las FNN analizan el historial de comportamiento del usuario para predecir preferencias y hacer recomendaciones personalizadas [31] [32].

Aunque las FNN son poderosas en problemas de clasificación y regresión, se destacan especialmente en conjuntos de datos estructurados donde la relación entre las entradas y salidas no es demasiado compleja. Son la base de muchos modelos de redes neuronales más complejos y pueden combinarse con arquitecturas como Redes Neuronales Profundas (DNN) o Redes Neuronales Convolucionales (CNN) para tareas más especializadas[26] [33].

2.7 Redes Neuronales Recurrentes (RNN)

Las Redes Neuronales Recurrentes (RNN) son una clase de arquitecturas neuronales diseñadas para manejar datos secuenciales, donde la información tiene una dependencia temporal [34]. A diferencia de las Redes Neuronales de Alimentación Adelante (FNN), las RNN poseen conexiones retroalimentadas, permitiendo la persistencia de la información a lo largo del tiempo. Esta capacidad las hace especialmente adecuadas para tareas como el procesamiento del lenguaje natural (PLN), modelado de series temporales y predicción en secuencias de datos. En PLN, las RNN se emplean para tareas como la generación de texto, la traducción automática y el análisis de sentimientos en texto. En el ámbito de las series temporales, se utilizan para prever valores futuros basándose en patrones históricos [35], siendo aplicadas en pronósticos meteorológicos, análisis financiero y control de procesos dinámicos. Ademas existen trabajos que abordan problemas de memoria a largo plazo en RNN [36] y mejoras en el PLN y [37].

2.8 Redes Neuronales Convolucionales (CNN)

Las Redes Neuronales Convolucionales (CNN) son una arquitectura especializada de redes neuronales profundas diseñadas para el procesamiento eficiente de datos estructurados, como imágenes y señales temporales. Introducidas principalmente para abordar problemas de visión por computadora, las CNN han demostrado su eficacia en una variedad de tareas relacionadas con la percepción y clasificación de patrones visuales [26].

Las características clave de las CNN incluyen capas convolucionales, capas de agrupación y capas

completamente conectadas. Las capas convolucionales aplican filtros a regiones locales de la entrada, permitiendo la detección jerárquica de características. Las capas de agrupación reducen la resolución espacial, disminuyendo la complejidad computacional. Finalmente, las capas completamente conectadas realizan la clasificación final basada en las características aprendidas.

Las Redes Neuronales Convolucionales (CNN) son fundamentales en algoritmos de predicción, sobresaliendo en:

1. **Reconocimiento de Objetos:** Utilizadas para el reconocimiento de objetos en imágenes, en aplicaciones como sistemas de vigilancia y vehículos autónomos [28] [29] .
2. **Segmentación de Imágenes:** Empleadas en algoritmos de segmentación para identificar y delimitar regiones específicas [38] [39].
3. **Clasificación de Imágenes Médicas:** Aplicadas en el ámbito médico para clasificar imágenes médicas, facilitando el diagnóstico [40] [41].
4. **Datos de Imágenes y Visión por Computadora:** Funcionan excepcionalmente bien en conjuntos de datos de imágenes, donde la convolución permite detectar patrones locales y jerarquías de características [26].
5. **Problemas de Reconocimiento de Patrones Espaciales:** Ideales para problemas que requieren el reconocimiento de patrones espaciales, como la identificación de características específicas en imágenes [28] [42].
6. **Transferencia de Aprendizaje:** La técnica de transferencia de aprendizaje, utilizando CNN preentrenadas, ha demostrado ser eficaz en tareas con conjuntos de datos más pequeños [43] [44]. Este enfoque proporciona una visión más detallada de la aplicación de las FNN y las CNN en algoritmos específicos y en qué contextos funcionan mejor.

2.9 Capacidad de generalización, overfitting y underfitting

En el contexto del aprendizaje automático, la evaluación del rendimiento de un modelo se realiza mediante el análisis de su comportamiento en conjuntos de datos específicos. Al entrenar un modelo de ML, se divide el conjunto de datos original en un conjunto de entrenamiento y un conjunto de prueba. A continuación, se presenta una descripción de los conceptos clave relacionados con la evaluación del rendimiento.

1. **Error de Entrenamiento y Error de Generalización:**

- **Error de Entrenamiento:** Medida del rendimiento del modelo en el conjunto de entrenamiento. Indica qué tan bien el modelo se ajusta a estos datos específicos [6].
- **Error de Generalización:** Medida del rendimiento del modelo en datos no vistos durante el entrenamiento, es decir, en un conjunto de prueba. Refleja la capacidad del modelo para hacer predicciones precisas en situaciones del mundo real [6].

2. **Capacidad de Generalización:** Habilidad de un modelo para realizar predicciones precisas en nuevos datos, extrapolando lo aprendido durante el entrenamiento [6]. Un modelo con alta capacidad de generalización puede adaptarse bien a diferentes situaciones sin depender excesivamente de los detalles específicos del conjunto de entrenamiento.

3. **Overfitting y Underfitting:**

   - **Overfitting** Ocurre cuando un modelo se ajusta demasiado a los datos de entrenamiento, capturando incluso detalles específicos de ese conjunto. Aunque puede tener un rendimiento excelente en el conjunto de entrenamiento, tiende a fallar al generalizar a nuevos datos, resultando en un alto error de generalización [6].
   - **Underfitting:** Ocurre cuando un modelo es demasiado simple para capturar la complejidad de los datos de entrenamiento. Los modelos infraajustados tienen un rendimiento deficiente tanto en el conjunto de entrenamiento como en el de prueba, indicando que no han aprendido adecuadamente las relaciones subyacentes en los datos [6].

En la práctica, encontrar el equilibrio adecuado entre el overfitting y el underfitting implica ajustar la complejidad del modelo. Aunque las funciones más sencillas tienen más probabilidades de generalizar, es necesario seleccionar una hipótesis lo suficientemente compleja para lograr un bajo error de entrenamiento. Por lo general, el error de entrenamiento tiende a disminuir hasta alcanzar el mínimo posible a medida que aumenta la capacidad del modelo. La relación entre la capacidad de generalización del modelo y el error tiende a tener una forma de **U**, como se ilustra en la figura 2.7. En el extremo izquierdo de la figura, tanto el error de entrenamiento como el de generalización son elevados, indicando underfitting. A medida que aumentamos la capacidad del modelo, el error de entrenamiento disminuye, pero la diferencia entre el error de entrenamiento y el de generalización aumenta, llevando al overfitting. Es crucial encontrar el punto óptimo donde la capacidad del modelo es suficiente para reducir el error de entrenamiento sin inducir un overfitting excesivo.

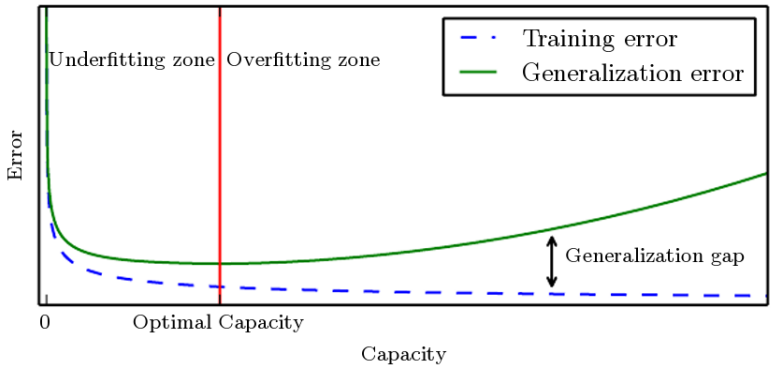

Figura 2.7: Relación típica entre la capacidad de generalización y el error [6].

2.10 Regularización

La regularización es una técnica fundamental para abordar el overfitting y controlar la complejidad del modelo. En un problema bien definido, la regularización busca modificar el problema para hacerlo numéricamente estable, introduciendo supuestos adicionales, como la suavidad de la solución. Un enfoque común es la regularización de Tikhonov, que penaliza ciertos aspectos del modelo durante el entrenamiento. En el contexto del aprendizaje automático, se emplea la regularización para incorporar preferencias en el algoritmo de aprendizaje, influyendo en el tipo de funciones permitidas en el espacio de hipótesis del modelo.

La preferencia por ciertas soluciones en el espacio de hipótesis puede lograrse mediante técnicas como la regularización de peso. Al ajustar la función de coste de la regresión lineal con decaimiento de peso, se busca minimizar una suma que comprende el error cuadrático medio en el entrenamiento y un criterio que expresa una preferencia por pesos más pequeños. Este enfoque se representa mediante la eq. (2.9), donde λ controla la fuerza de la preferencia por pesos más pequeños. La regularización de peso permite encontrar soluciones que equilibren el ajuste a los datos de entrenamiento con la simplicidad del modelo, evitando así el overfitting.

- **Definición de Problema Bien Definido:** Un problema bien definido en el contexto de regularización implica resolver una ecuación diferencial sujeta a condiciones de frontera o de valor inicial cuando una de las variables toma un valor específico. La solución de estos problemas debe cumplir con ciertas propiedades, como existencia, unicidad y continuidad dependiente de las condiciones iniciales. La regularización, en este contexto, se emplea para abordar problemas lineales mal planteados, y la regularización de Tikhonov es una técnica comúnmente utilizada [45].

$$J(w) = (Y - Xw)^T(Y - Xw) + \lambda w^T w. \tag{2.9}$$

A continuación se detallan aspectos relevantes de la técnica de la regulación.

- **Incorporación de Preferencias en el Algoritmo de Aprendizaje:** Para diseñar algoritmos efectivos, se introducen preferencias en el algoritmo de aprendizaje. La selección de funciones permitidas en el espacio de hipótesis del modelo juega un papel crucial en su rendimiento. Por ejemplo, en la regresión lineal, cuyo espacio de hipótesis consta de funciones lineales, la utilidad de estas funciones depende de la naturaleza lineal o no lineal del problema a resolver. La capacidad de generalización se controla ajustando la complejidad del modelo [6].

- **Regularización de Peso en Regresión Lineal:** La preferencia por ciertas soluciones se puede lograr mediante técnicas de regularización, como el decaimiento del peso (*weight decay*). En el caso de la regresión lineal, la función de coste se modifica para incorporar un término de penalización $\lambda w^T w$, donde λ controla la preferencia por pesos más pequeños. Minimizar esta función de coste resulta en la elección de pesos que equilibran el ajuste a los datos de entrenamiento con la simplicidad del modelo, evitando el overfitting [6].

- **Visualización de la Regularización de Peso:** La figura 2.8 ilustra cómo diferentes valores de λ afectan a modelos entrenados con regresión polinómica de alto grado. A medida que λ aumenta, se favorecen soluciones con pesos más pequeños, controlando así la tendencia del modelo a sobreajustarse o desajustarse.

- **Generalización de Modelos con Penalización:** De manera más general, la regularización se aplica introduciendo penalizaciones en la función de coste del modelo. Esta práctica tiene como objetivo mejorar la capacidad de generalización del modelo sin comprometer su rendimiento en el conjunto de entrenamiento [6].

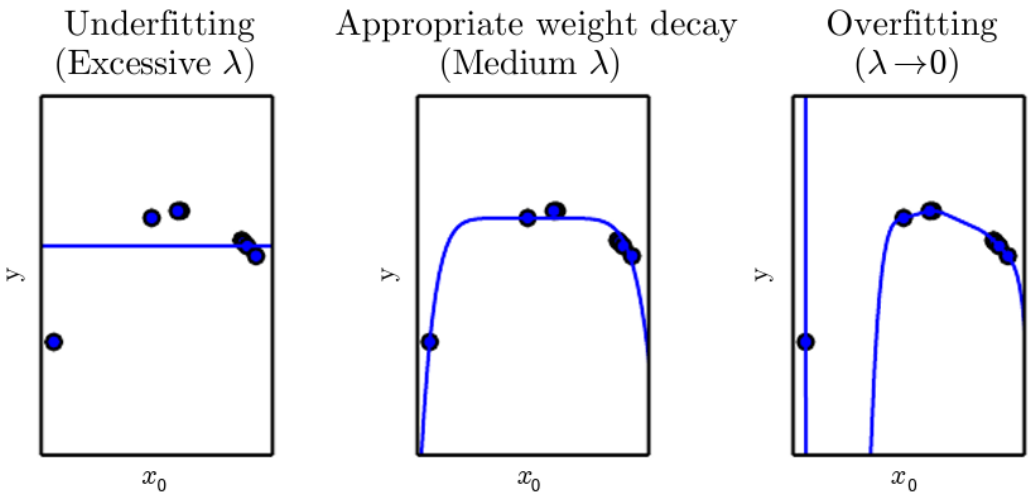

Figura 2.8: Modelos entrenados con diferentes valores de λ [6].

2.11 Hiperparámetros y Conjunto de Validación

Los hiperparámetros son ajustes que controlan el comportamiento de un algoritmo y no son aprendidos durante el entrenamiento. La selección adecuada de estos hiper-parámetros es crucial para obtener un modelo con capacidad de generalización óptima. Introduciremos el concepto de conjunto de validación, un subconjunto no utilizado en el entrenamiento, que desempeña un papel fundamental al evaluar el rendimiento y ajustar los hiper-parámetros del modelo. A continuación, se presentan descripciones importantes:

- **Definición de Hiper-parámetros:** Los hiper-parámetros son ajustes específicos de un algoritmo de aprendizaje automático que controlan su comportamiento. A diferencia de los parámetros, los hiper-parámetros no son adaptados por el propio algoritmo durante el entrenamiento. En su lugar, deben ser seleccionados antes del proceso de aprendizaje y afectan directamente la capacidad de generalización del modelo.
- **No usar Hiper-parámetros en el Conjunto de Entrenamiento:** Hiper-parámetros que influyen en la capacidad de generalización no deben aprenderse en el conjunto de entrenamiento, ya que tenderían a maximizar la capacidad de generalización, llevando a un overfitting.
- **Importancia del Conjunto de Validación:** Se introduce un conjunto de validación, que es un subconjunto no utilizado durante el entrenamiento, para estimar el error de generalización del modelo después del aprendizaje. Este conjunto permite ajustar los hiper-parámetros sin influir en el rendimiento del modelo en el conjunto de prueba.
- **División de Datos:** Los datos de entrenamiento se dividen en dos subconjuntos: uno utilizado para aprender los parámetros y otro como conjunto de validación para evaluar la generalización.
- **Validación Cruzada:** En casos donde el conjunto de prueba es pequeño, la validación cruzada k-fold estratificada puede ser utilizada. Este método implica dividir el conjunto de datos en k subconjuntos, manteniendo la proporción de clases, y realizar k ensayos utilizando cada subconjunto como conjunto de prueba en turnos [46]. En La Figura 2.9 se presenta un ejemplo de validación cruzada.

Figura 2.9: Diagrama de validación cruzada k-fold con k = 10.

- **Estimación del Error de Generalización:** Una vez ajustados los hiper-parámetros, el error de generalización se estima utilizando el conjunto de prueba, garantizando que no se haya utilizado durante el ajuste de los hiper-parámetros.

2.12 Descenso del gradiente estocástico

El Descenso del Gradiente Estocástico (SGD) es un algoritmo de optimización utilizado en el entrenamiento de modelos de aprendizaje automático. A diferencia del descenso del gradiente convencional, que utiliza el gradiente completo calculado sobre todo el conjunto de datos de entrenamiento en cada iteración, el SGD realiza actualizaciones de parámetros utilizando solo una muestra (un ejemplo) aleatoria en cada iteración. Este enfoque estocástico hace que el proceso sea más eficiente y escalable, especialmente para conjuntos de datos grandes [47]. Esites investigaciones mas recientes que abordan estrategias de optimización efectivas para lidiar con conjuntos de datos masivos y destacan la relevancia del SGD en este contexto [48] [49].

El proceso del SGD se puede resumir en los siguientes pasos:

1. **Inicialización:** Inicializa los parámetros del modelo aleatoriamente.
2. **Selección Estocástica:** Muestrea un ejemplo aleatorio del conjunto de entrenamiento.
3. **Cálculo del Gradiente:** Calcula el gradiente de la función de costo con respecto a los parámetros utilizando el ejemplo seleccionado.

4. **Actualización de Parámetros:** Actualiza los parámetros en la dirección opuesta al gradiente multiplicado por una tasa de aprendizaje.

5. **Iteración:** Repite los pasos 2-4 hasta que se alcance un criterio de convergencia.

El SGD es beneficioso para grandes conjuntos de datos y problemas de alta dimensionalidad, ya que permite actualizaciones frecuentes de los parámetros, evitando cálculos costosos del gradiente completo en cada iteración. El descenso de gradiente estocástico tiene aplicaciones más allá del aprendizaje profundo y se utiliza para entrenar grandes modelos lineales en conjuntos de datos masivos.

2.13 Evaluación del Rendimiento y Métricas

En esta sección, se profundiza en la evaluación del rendimiento de los algoritmos mediante el análisis de la matriz de confusión, presentada gráficamente en la Figura 2.10 [50].

2.13.1 Matriz de Confusión

La matriz de confusión es una herramienta fundamental que desglosa el rendimiento del clasificador en cuatro categorías esenciales:

- **Aciertos (True Positive):** Muestras correctamente clasificadas como rayos.
- **Fallos (False Negative):** Muestras de rayos mal clasificadas como cielos despejados (error tipo II).
- **Falsa Alarma (False Positive):** Muestras de cielos despejados mal clasificadas como rayos (error tipo I).
- **Desestimación Correcta (True Negative):** Muestras correctamente clasificadas como cielos despejados.

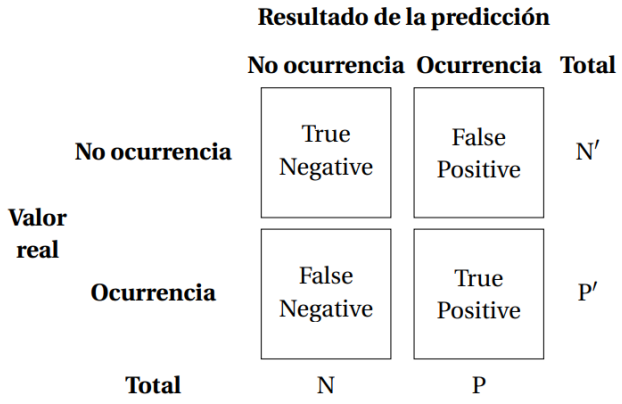

Figura 2.10: Matriz de confusión [5].

2.13.2 Métricas de Evaluación

Para una comprensión exhaustiva del rendimiento del modelo, se emplean diversas métricas:

Precisión

La precisión (*Precision*) mide la proporción de muestras clasificadas como rayos que son realmente rayos y se calcula como:

$$Precision = \frac{TP}{TP + FP}.$$

Recall o Probabilidad de Detección

El *Recall* representa la proporción de muestras de rayos identificadas correctamente y se define como:

$$Recall = \frac{TP}{TP + FN}. \tag{2.10}$$

Tasa de Falsas Alarmas

La tasa de falsas alarmas (FAR) indica la proporción de cielos despejados incorrectamente clasificados como rayos:

$$FAR = \frac{FP}{TP + FP}. \tag{2.11}$$

Puntuación F1

La puntuación F1 (*F1 Score*) combina precisión y *recall* para una evaluación equilibrada:

$$F1 = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall}. \tag{2.12}$$

Índice de Éxito Crítico

El Threat Index evalúa la capacidad del modelo para prever eventos relevantes:

$$Threat\ index = \frac{TP}{TP + FP + FN}. \tag{2.13}$$

Puntuación de Destreza de Heidke

La Heidke Skill Score mide la mejora del pronóstico sobre el azar:

$$HSS = 2 \cdot \frac{TP \cdot TN - FN \cdot FP}{(TP + FN)(FN + TN) + (TP + FP)(FP + TN)}. \tag{2.14}$$

Estas métricas proporcionan una evaluación integral y detallada del desempeño del modelo [51].

2.14 Predicción climática

2.14.1 Modelos basados únicamente en variables meteorológicas

Los primeros estudios de predicción de descargas atmosféricas se centraron principalmente en parámetros meteorológicos de superficie. En India, Adhikari et al. [52] emplearon modelos de redes neuronales artificiales (ANN) usando temperatura, humedad y presión como variables de entrada, logrando una precisión del 85.2 %. De forma similar, investigaciones en Malasia [53] e Irán [54] aplicaron arquitecturas tipo MLP, alcanzando precisiones superiores al 90 %. En Jordania, Al-Jundi et al. [55] optimizaron un modelo MLP utilizando algoritmos de colonia de abejas artificiales, obteniendo un F1-score superior a 0.92. Mohan y Arumugam [56] propusieron desde India un modelo híbrido ANN–kNN con un 89.14 % de precisión.

En Estados Unidos, Mostajabi et al. [57] entrenaron modelos Random Forest (RF) con variables meteorológicas de superficie, alcanzando un 88.4 % de precisión con un horizonte de 30 minutos. Pereira et al. [58] en Brasil compararon ANN y RF, con mejores resultados para este último (99.77 %). El enfoque se ha extendido al uso de redes FFNN, como demuestran los estudios en la costa este de Malasia [59]. Más recientemente, se ha incorporado información de reanálisis atmosférico como ERA5 en modelos de aprendizaje profundo. Ehrensperger et al. [60] utilizaron perfiles verticales para entrenar modelos que alcanzaron un F1-score de 0.88, mientras que Cheon [61] propuso un esquema global con datos de WWLLN que alcanzó una correlación de $r = 0{,}94$.

2.14.2 Modelos basados en datos de campo eléctrico

El campo eléctrico ha demostrado ser un indicador temprano eficaz para la formación de tormentas eléctricas. En Colombia, Ferro et al. [62] y Ariza et al. [63] reportaron tasas de éxito del 83 % utilizando umbrales simples sobre mediciones de campo. En China, Bao et al. [64] utilizaron un modelo ResNet50, alcanzando una precisión del 91.3 % y una sensibilidad del 87.6 %. Yang et al. [65] propusieron una arquitectura CNN-BiLSTM que alcanzó $R^2 = 0{,}9558$, mientras que Hill et al. [66] aplicaron redes convolucionales-recurrentes con sensores EFM, obteniendo un F1-score de 0.86. Posteriormente, Yang et al. [67] incorporaron datos tridimensionales de campo eléctrico (3DAEFA) a modelos CNN-BiLSTM, superando el 87 % de detección. Fukawa et al. [68] entrenaron redes recurrentes solo con campo eléctrico superficial, logrando una precisión del 85.5 % y un recall del 88 %.

2.14.3 Modelos basados en datos satelitales o de teledetección

La predicción de rayos basada en satélites ha cobrado importancia recientemente. En Estados Unidos, Mansouri et al. [69] desarrollaron un modelo LSTM utilizando únicamente datos GLM y WWLLN, alcan-

zando un 89 % de precisión sin necesidad de mediciones terrestres. Hutchins et al. [70] validaron WWLLN comparando emisiones ópticas y de radiofrecuencia. En China, Chen et al. [71] combinaron reflectividad radar y datos de rayos en un modelo profundo en conjunto, logrando F1 cercanos a 0.91. Ghosh et al. [72] en India desarrollaron una red de sensores distribuida que mejoró la resolución espacial. En la Amazonía, Espinoza et al. [73] usaron ANN con entradas satelitales, logrando más del 90 % de precisión. Li et al. [74] en el sur de China implementaron redes Conv-GRU con imágenes GLM, alcanzando una precisión espacial del 91.7 % y un MAE de 17 descargas por hora.

2.14.4 Arquitecturas híbridas y modelos ensamblados

Los modelos más recientes integran múltiples tipos de datos y arquitecturas de aprendizaje. En Sudáfrica, Essa et al. [75] compararon modelos LSTM-FC, CNN-LSTM y ConvLSTM, siendo este último el más preciso (F1 = 0.88). Wang et al. [76] en China utilizaron señales BEADS combinadas con BiLSTM, alcanzando un 93 % de precisión. En México, Montiel et al. [77] propusieron una arquitectura GCN-LSTM con una precisión del 91.7 %. En China también, el sistema MCGLN [78] integró ConvLSTM con redes generativas (GAN), obteniendo un F1-score de 0.89.

Los métodos ensamblados también han mostrado potencial. En Irán, Pakdaman et al. [79] combinaron clasificadores como J48, SVM, MLP y Naive Bayes, mejorando la precisión entre un 8 % y 12 %. En India, Patil et al. [80] integraron SMOTE con XGBoost y RF, logrando un F1-score de 0.935, y Borah et al. [81] introdujeron modelos explicables que superaron $R^2 > 0{,}90$. Finalmente, trabajos recientes exploran la combinación de modelos físicos con técnicas de IA: Li et al. [82] propusieron un modelo híbrido informado por física, y el marco FlashBench [83] fusionó aprendizaje profundo con modelos físicos dinámicos para mejorar el nowcasting.

La literatura evidencia que los modelos de predicción mejoran considerablemente al incorporar señales de campo eléctrico. Sin embargo, la mayoría de estos trabajos utilizan únicamente el valor absoluto del campo o umbrales predefinidos. Este enfoque, aunque efectivo, no considera el análisis de características específicas como el *campo eléctrico peak*, el *promedio* o los *cambios de polaridad*, que podrían tener un rol clave en la predicción. Esta investigación se propone justamente estudiar de manera sistemática el efecto de incorporar estas métricas poco exploradas en modelos de aprendizaje automático.

Capítulo 3

Metodología

En este capítulo se detalla la metodología diseñada para construir y evaluar un sistema experto para la predicción de descargas eléctricas atmosféricas. El procedimiento abarca cuatro etapas fundamentales: la adquisición y caracterización de los datos, el preprocesamiento y la preparación de los conjuntos, el diseño y entrenamiento de los modelos de inteligencia artificial, y finalmente, su evaluación de desempeño.

El estudio se centró principalmente en una zona de alta montaña en Perú, utilizando datos correspondientes al año 2023, seleccionado por la disponibilidad de mediciones de campo eléctrico [62, 66]. De manera complementaria, se incorporó una segunda área de análisis en Chile con registros del año 2022, elegido por presentar la mayor concentración de rayos detectados por el sensor GLM en el periodo 2018–2023. En el caso de Chile, el análisis se basó exclusivamente en variables meteorológicas y en la detección de descargas mediante GLM, siguiendo metodologías utilizadas en estudios basados únicamente en parámetros meteorológicos [52, 57, 58]. Adicionalmente, se incluyó la ciudad de Buenos Aires debido a la disponibilidad simultánea de datos meteorológicos y de campo eléctrico, complementados con las redes de detección de rayos GLM y WWLLN [69, 70].

3.1 Adquisición y Fuentes de Datos

La construcción de los conjuntos de datos se realizó integrando información proveniente de tres categorías principales: registros de descargas eléctricas, variables meteorológicas de superficie y mediciones de campo eléctrico ambiental. Los datos meteorológicos de ambas zonas de estudio se obtuvieron de la plataforma NASA POWER con una resolución horaria [60].

En el caso particular de Perú, además de las variables meteorológicas, se incorporaron mediciones de campo eléctrico obtenidas mediante sensores de alta frecuencia con muestreo en segundos, técnica

utilizada en estudios recientes de predicción de tormentas [64, 67]. Para garantizar la compatibilidad con la resolución horaria del resto de las variables, los registros fueron procesados mediante agregaciones que permitieron derivar tres ejes principales de análisis: el valor promedio del campo eléctrico, el valor peak absoluto y la ocurrencia de cambios de polaridad durante cada hora [65].

Las descargas eléctricas atmosféricas para Perú provinieron de la red LINET, mientras que para Chile se emplearon datos del sensor satelital GLM. Dado que GLM no distingue entre descargas nube-tierra e intranube, se aplicó adicionalmente un proceso de filtrado para trabajar exclusivamente con eventos nube-tierra en la etapa de modelado, siguiendo criterios como los de [4]. Los datos de tipo de descarga de LINET incluyen la distinción entre nube-tierra e intranube [70].

Los conjuntos de datos se construyeron a partir de áreas de análisis de 30×30 km centradas en cada zona de estudio, en línea con enfoques empleados en predicción de rayos [73, 74].

Tabla 3.1: Fuentes de datos y variables por país.

Categoría	**Perú**	**Chile**	**Argentina**
Variables meteorológicas	Temp., humedad relativa, presión, radiación, precipitación, velocidad y dirección del viento	Temp., humedad relativa, presión, radiación, precipitación, velocidad y dirección del viento	Temp., humedad relativa, presión, velocidad y dirección del viento
Campo eléctrico	Sensor alta frecuencia (1 s → 1 hora)	No disponible	Sensor alta frecuencia (1 s → 5 min)
Coordenadas	Lat −14,48, Lon −71,76	Lat −17,59, Lon −69,48	Lat −34,58, Lon −58,48
Red de captación de rayos	LINET	GLM-WWLLN	GLM-WWLLN
Fuente datos meteorológicos	NASA POWER	NASA POWER	Inst. Meteorológico de Buenos Aires
Resolución de los datos	30×30 km, 1 h (meteo)	30×30 km, 1 h (meteo)	30×30 km, 5 min (campo eléctrico)
Área de análisis	Alta montaña, 30×30 km	Alta montaña, 30×30 km	Región metropolitana, 30×30 km

3.2 Análisis Exploratorio de Datos (AED)

Antes de construir los modelos predictivos, realizamos un *Análisis Exploratorio de Datos (AED)* en profundidad. Este paso fue crucial para comprender la estructura de los datos, evaluar su calidad y caracterizar la distribución e interrelaciones entre las variables clave.

El AED se centró en tres actividades principales:

- *Identificación de Patrones Temporales:* Se buscan patrones cíclicos y posibles anomalías en la frecuencia horaria de las descargas eléctricas, visualizando su comportamiento con *histogramas de ocurrencia horaria*.

- *Análisis de Distribución Estadística de Variables:* Se estudia la distribución de cada variable meteorológica y eléctrica mediante histogramas, tablas de estadística descriptiva y gráficos tipo box/violin. Con estas herramientas, se identifican características clave como las asimetrías en la radiación (Figura 3.1), la dispersión de los datos en eventos con rayos (Tabla 4.2) y los patrones cíclicos diarios que podrían influir en las descargas (Figura 3.2).

- *Análisis de Relaciones entre Variables:* Para entender la interacción entre los predictores y la variable objetivo, se investigan sus correlaciones. Se generan diagramas de dispersión multivariable (Figura 4.4) y matrices de correlación (Figura 3.4), lo que permite evaluar tanto asociaciones lineales como no lineales [60, 65].

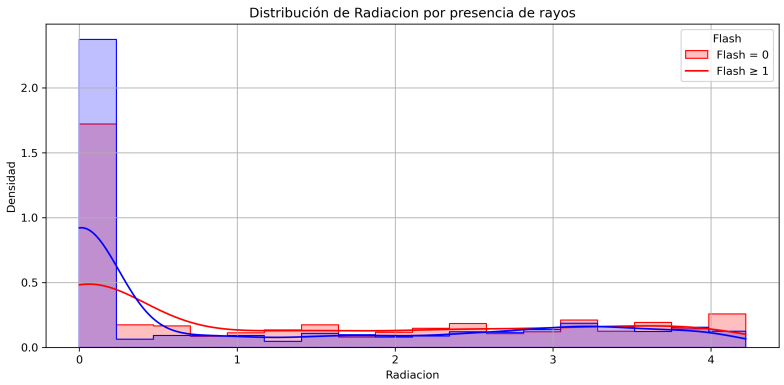

Figura 3.1: Distribución de la variable radiación en el análisis exploratorio.

Tabla 3.2: Estadística descriptiva para registros con rayos.

Variable	**Count**	**Mean**	**Std**	**Min**	**25 %**	**50 %**	**75 %**
Radiación	1125	1.41	1.48	0.00	0.00	0.91	2.80
Temperatura	1125	11.29	4.49	-0.74	7.54	11.18	14.89
Humedad	1125	56.43	20.04	13.81	39.31	54.44	73.31
...	...	...	...	...	...	...	...

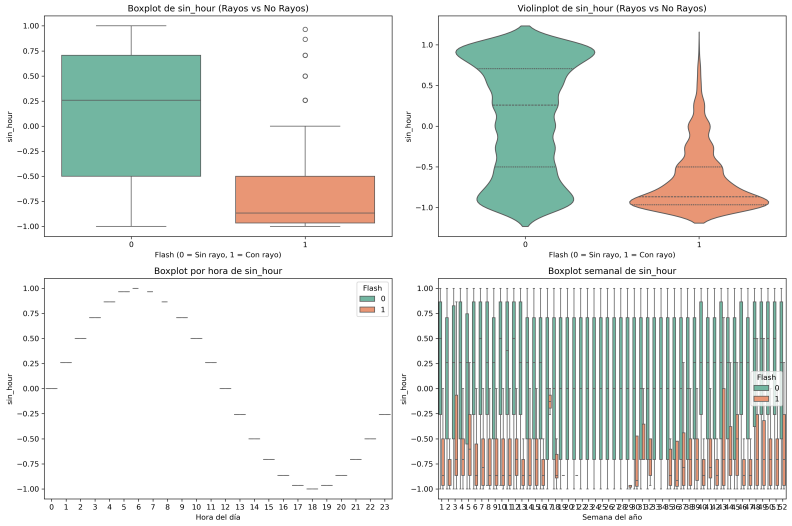

Figura 3.2: Ejemplo de gráfico box/violin para la variable cíclica *sin_hour*.

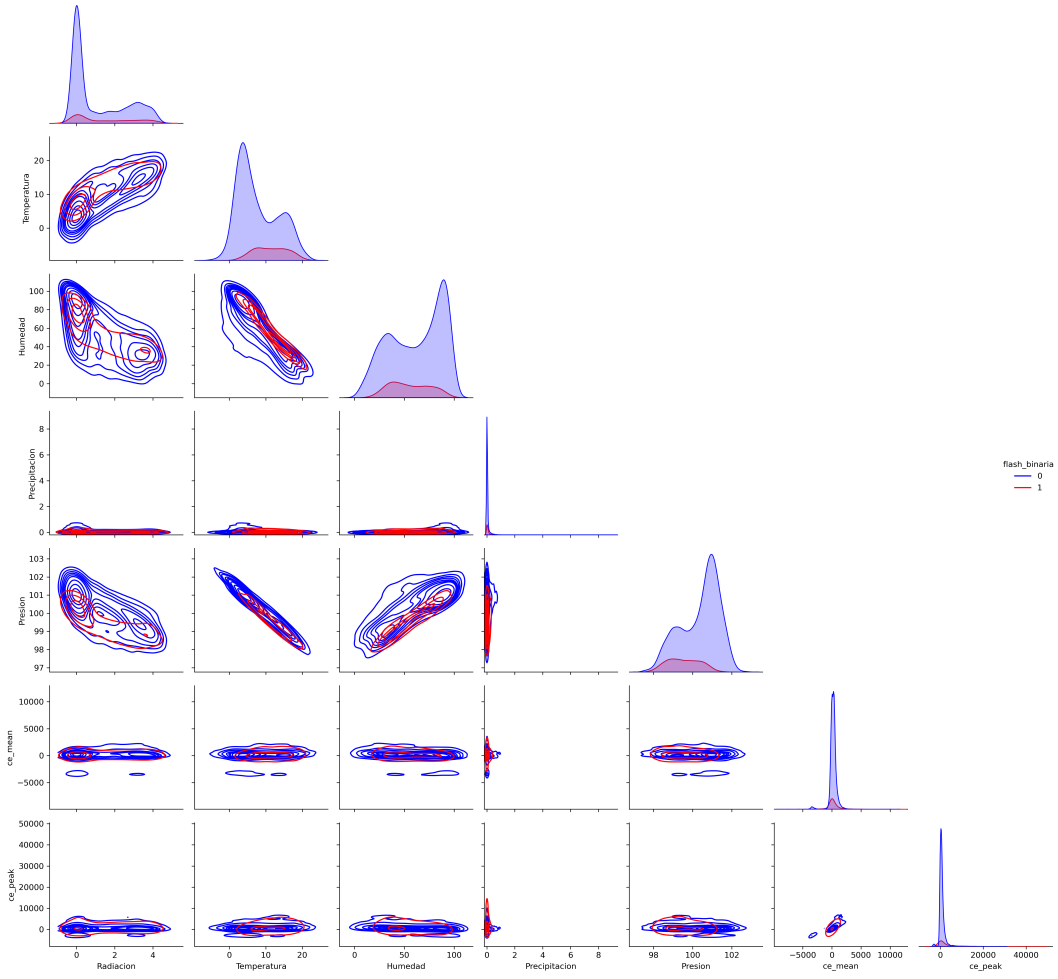

Figura 3.3: Relaciones bidimensionales entre variables en el análisis exploratorio.

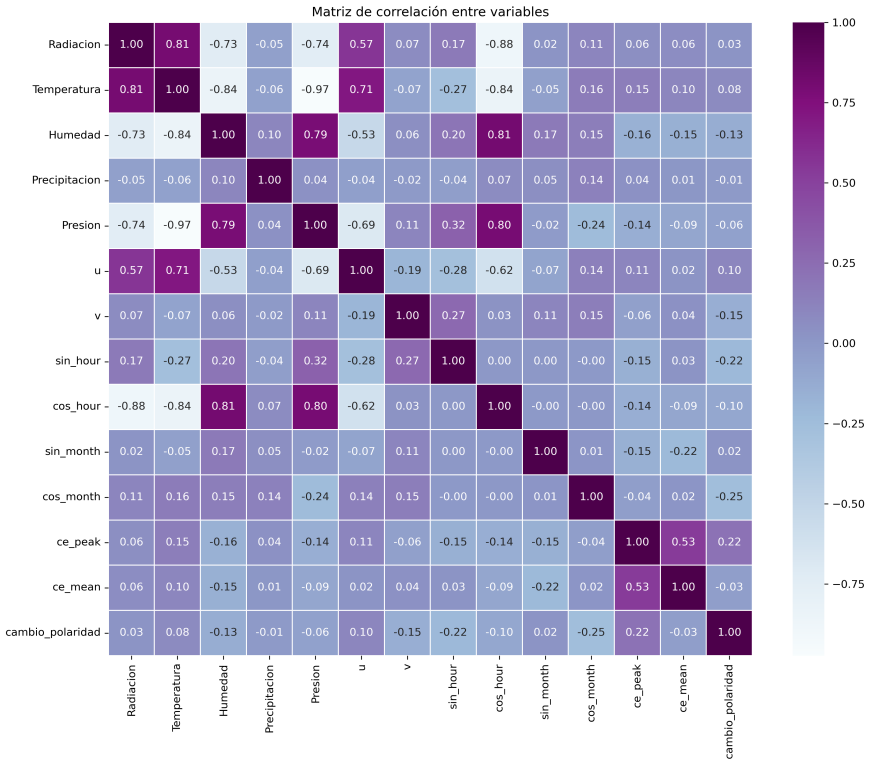

Figura 3.4: Matriz de correlación entre variables meteorológicas y eléctricas.

3.3 Preprocesamiento y Construcción de Conjuntos de Datos

El preprocesamiento fue una etapa fundamental orientada a unificar las distintas fuentes de información, corregir problemas de calidad y estructurar los conjuntos de manera adecuada para el entrenamiento de los modelos. El trabajo se desarrolló en los siguientes pasos, utilizando principalmente **Python con las librerías Pandas, Scikit-learn y Tensorflow** para la manipulación y transformación de los datos:

- *Alineamiento temporal y agregación:* Todos los registros se sincronizaron a una resolución horaria uniforme. En el caso del campo eléctrico, que contaba con mediciones en segundos, se realizaron agregaciones para obtener tres indicadores por hora: valor promedio, valor peak absoluto y frecuencia de cambios de polaridad [67].

- *Tratamiento de valores faltantes:* Para las variables continuas del campo eléctrico (valor peak y valor promedio) se utilizó imputación mediante el algoritmo de k-vecinos más cercanos con $k = 5$ [84]. En el caso de la variable binaria de cambio de polaridad, se aplicó el mismo método con $k = 1$, lo que permitió estimar los valores ausentes respetando la naturaleza categórica de la variable [85].

- *Transformación de variables temporales:* Las componentes de hora y mes se transformaron en variables cíclicas utilizando funciones seno y coseno, con el fin de capturar la periodicidad natural de los

datos sin introducir saltos artificiales en los límites de cada ciclo [86].

- *Variables de viento:* La información de viento se mantuvo en términos de velocidad y dirección, acorde a la forma en que fueron entregados los datos meteorológicos originales.
- *Normalización:* Todas las variables numéricas se escalaron para mantener comparabilidad entre magnitudes. Los parámetros de escalado se calcularon únicamente sobre el conjunto de entrenamiento y se aplicaron a los conjuntos de validación y prueba, evitando la fuga de información entre fases.

Finalmente, los datos se dividieron en conjuntos independientes para entrenamiento, validación y prueba siguiendo un criterio estrictamente cronológico. Esta estrategia permitió preservar la secuencia temporal de los eventos y asegurar que el conjunto de prueba representara información futura no vista por el modelo durante su desarrollo [35].

3.4 Selección de modelos

Para el desarrollo del sistema predictivo, se siguió un enfoque comparativo entre un modelo clásico y arquitecturas de redes neuronales avanzadas, una práctica común en la evaluación de algoritmos de predicción [58, 59]. Primero, se implementó un modelo de Perceptrón Multicapa (MLP) para establecer una línea base de rendimiento. El desempeño de este modelo se evaluó mediante validación cruzada K-Fold estratificada con 10 folds (K=10) para asegurar la robustez de la métrica [46, 87].

De manera complementaria, y como enfoque principal de la investigación, se implementaron arquitecturas de redes neuronales utilizando TensorFlow. Estas incluyeron capas convolucionales unidimensionales (Conv1D), redes recurrentes (SimpleRNN, LSTM, GRU) y capas densas, en concordancia con trabajos que emplean modelos secuenciales y convolucionales para datos atmosféricos [36, 64, 65].

Cada modelo fue entrenado de forma independiente registrando métricas clave como la puntuación F1. Para la selección final entre las arquitecturas de TensorFlow, se utilizó una división cronológica estricta de los datos, en lugar de la validación cruzada, para simular un escenario operativo. En el entrenamiento de estos modelos, se estableció un máximo de 100 épocas con un mecanismo de detención temprana (EarlyStopping) para evitar el sobreajuste y optimizar el rendimiento general del sistema [88].

3.5 Evaluación de desempeño

El rendimiento de los modelos se evaluó utilizando un conjunto de prueba independiente separado de forma cronológica respecto a los datos de entrenamiento y validación, con el fin de simular un escenario operativo real y evitar cualquier fuga de información entre fases.

Dado el desbalance entre registros con y sin descargas, se utilizaron métricas apropiadas para clasificación binaria en contextos desbalanceados. La métrica principal fue el *F1-score* de la clase positiva, ya que combina *precision* y *recall* en una única medida y penaliza de manera equilibrada tanto los falsos positivos como los falsos negativos. Como métricas complementarias se calcularon *precision* y *recall*, y se generaron *confusion matrices* para visualizar los aciertos y errores de clasificación.

3.6 Búsqueda de hiperparámetros

Para optimizar el rendimiento de los modelos se realizó un ajuste sistemático de hiperparámetros mediante un esquema de *Grid Search*. El proceso consideró variables clave como la *learning rate*, el *batch size*, el número de *epochs*, la cantidad de *capas* y de *unidades por capa*, así como las tasas de *regularización*.

Se evaluaron combinaciones predefinidas utilizando validación sobre el conjunto temporal de referencia, seleccionando aquellas configuraciones que maximizaron el *F1-score* de la clase positiva.

3.7 Diagrama de Flujo Metodológico

Para ofrecer una visión clara y concisa del proceso de investigación, la Figura 3.5 presenta un diagrama de flujo que resume las etapas y la secuencia de las actividades metodológicas descritas. Este esquema visual detalla cómo la información es adquirida, procesada y utilizada a lo largo del desarrollo de los modelos predictivos.

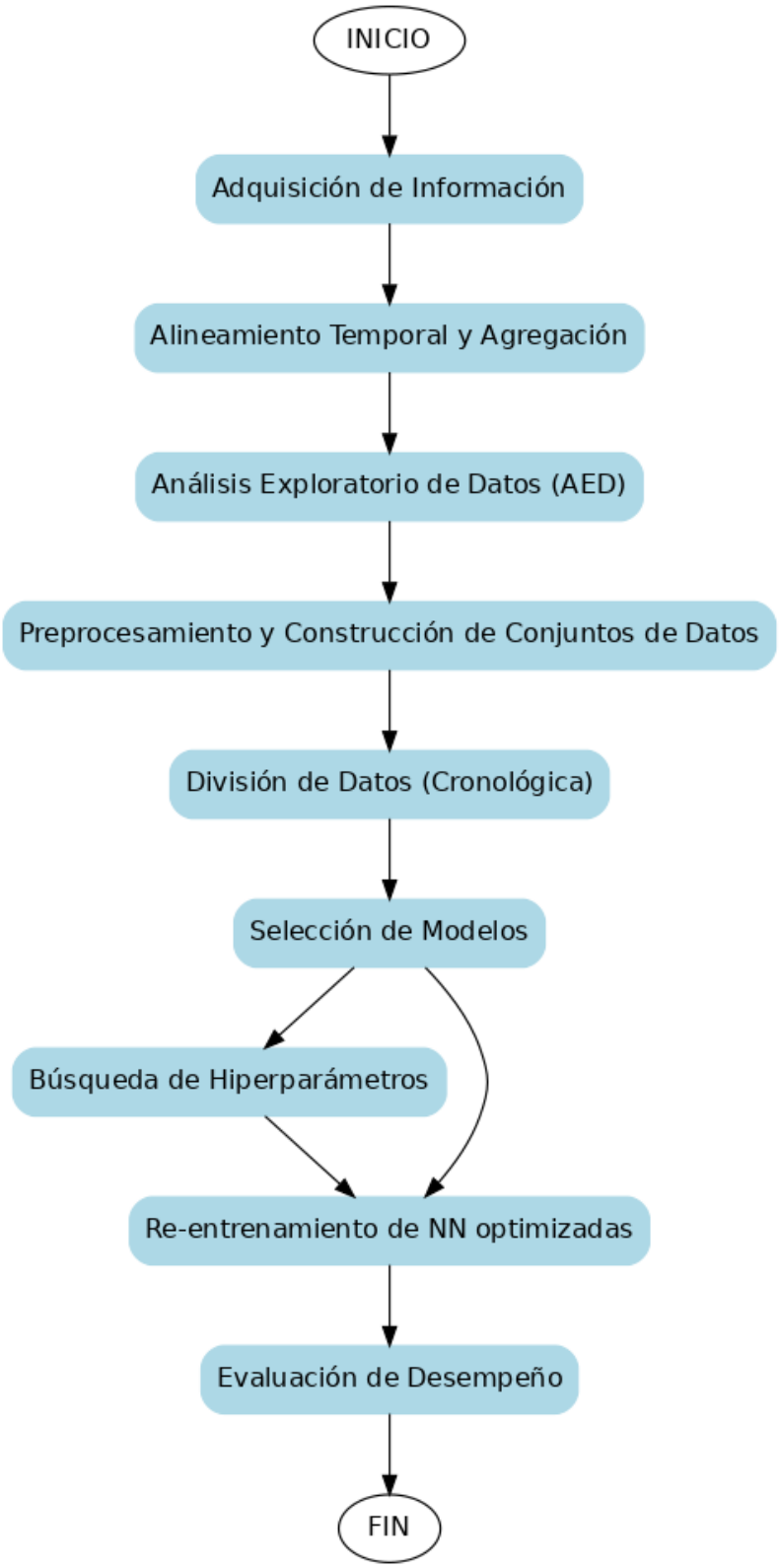

Figura 3.5: Diagrama de flujo de la metodología de investigación.

Capítulo 4

Análisis de resultados

4.1 Resultados Perú

En esta sección se presentan los resultados obtenidos a partir del desarrollo de modelos predictivos independientes para los tres territorios considerados en este estudio: Perú, Argentina y Chile. Para cada país, se construyeron conjuntos de datos locales con resolución horaria, integrando registros de descargas eléctricas atmosféricas, variables meteorológicas de superficie y, cuando estuvo disponible, mediciones de campo eléctrico ambiental.

El caso peruano mostró el mejor desempeño del algoritmo, alcanzando métricas significativamente superiores en comparación con los otros dos territorios. Debido a este resultado, se ha optado por presentar un análisis detallado para Perú como referencia principal.

4.1.1 Selección de Ubicación y Área de Muestreo

Para asegurar la coherencia espacial entre los registros meteorológicos, eléctricos y de descargas atmosféricas, se definió una zona de análisis de 30×30 km en el territorio peruano. Esta área corresponde a un cuadrado centrado en un punto geográfico seleccionado estratégicamente, en función de la disponibilidad de sensores y la relevancia climática del entorno.

- **Perú:** se eligió una ubicación en una zona minera de alta montaña, en las coordenadas *-14.483866* (latitud) y *-71.763229* (longitud), a aproximadamente 4.250 metros sobre el nivel del mar. En este sitio se encuentra instalado un sensor de campo eléctrico de alta frecuencia, lo que permite registrar fenómenos eléctricos asociados a tormentas en condiciones extremas de altitud.

La región seleccionada sigue un criterio de delimitación espacial que considera un cuadrado de 30 km por lado centrado en el punto geográfico de interés. Dentro de este límite, se recopilaron los datos meteorológicos, eléctricos y de rayos necesarios para la construcción de los conjuntos de datos y el posterior desarrollo de modelos predictivos.

En el caso específico de Perú, el modelo fue diseñado para identificar la ocurrencia de al menos un evento de rayo dentro de una ventana de predicción, utilizando horizontes temporales de 1, 5, 10, 15 y 24 horas.

Para tal fin, se integraron tres fuentes principales de información:

- **Registros de rayos:** provienen de la red LINET, la cual entrega eventos localizados de descargas eléctricas nube-tierra.
- **Variables meteorológicas:** extraídas del sistema NASA POWER, con resolución horaria.
- **Mediciones de campo eléctrico:** obtenidas a partir de sensores terrestres instalados en el sitio de referencia, con una frecuencia de muestreo entre 1 y 3 segundos.

A partir de estas fuentes, se construyó un conjunto de datos integrado con resolución horaria, mediante un proceso de alineamiento temporal, exploración, limpieza y transformación de las variables. Este conjunto fue finalmente utilizado para entrenar modelos secuenciales orientados a la predicción de rayos.

4.1.2 Análisis Exploratorio de Datos

Durante la etapa inicial del análisis exploratorio se examinó la estructura y calidad de los datos recopilados en Perú para el año 2023. Los registros meteorológicos, obtenidos de la plataforma NASA POWER, presentan una granularidad horaria e incluyen variables fundamentales como temperatura, humedad relativa, precipitación acumulada, presión atmosférica, radiación solar, velocidad del viento y dirección del viento. Todas estas variables se encuentran completas (sin valores faltantes) y presentan una cantidad total de 8.784 observaciones, correspondientes al número de horas en un año calendario.

De forma paralela, se analizaron las mediciones de campo eléctrico provenientes del sensor de superficie identificado como `DEF`, cuyo intervalo de muestreo varía entre 1 y 3 segundos. Este conjunto alcanzó un total de 2.294.418 registros para el año 2023, con una cobertura temporal continua desde el **1 de enero de 2023 a las 00:01:42** hasta el **31 de diciembre de 2023 a las 23:59:57**. La variable de campo eléctrico (`ce`) también se encuentra libre de valores nulos, aunque presenta una dispersión considerable con valores extremos que oscilan entre **-14.279,87 V/m** y **45.594,28 V/m**, lo que sugiere la posible presencia de eventos transitorios intensos que deberán ser revisados en fases posteriores de limpieza o normalización.

4.1.3 Distribución mensual de descargas eléctricas atmosféricas a tierra

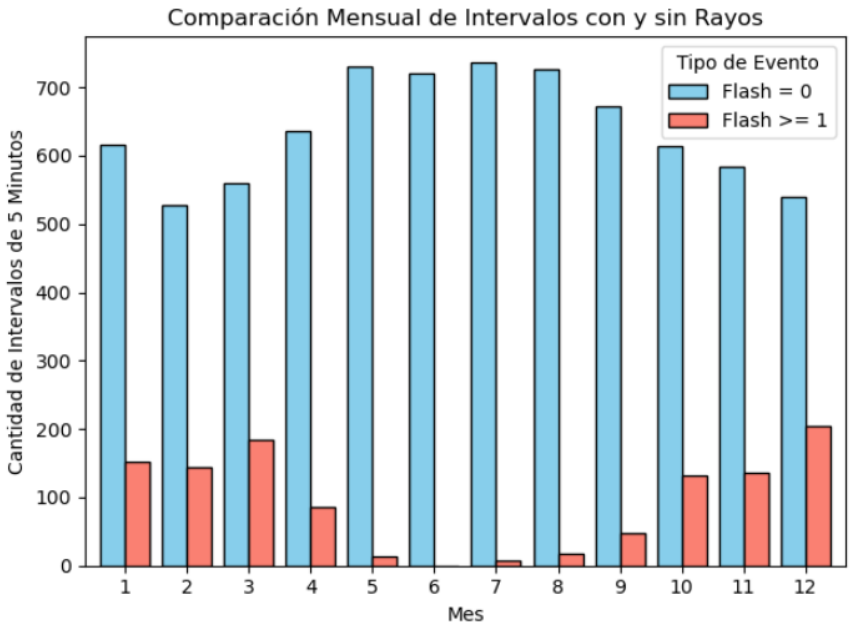

Figura 4.1: Concentración mensual de registros con y sin presencia de descargas eléctricas durante el año 2023.

La Figura 4.1 muestra la concentración mensual de registros con y sin presencia de descargas eléctricas tipo *flash* durante el año 2023. El gráfico de barras apiladas distingue los eventos sin rayos (Flash = 0, en celeste) de aquellos con al menos una descarga (Flash ≥ 1, en rojo claro).

Aunque la cantidad total de registros por mes es relativamente constante, se observa una mayor ocurrencia de eventos con rayos en los meses de enero, febrero, marzo y diciembre, lo que sugiere una posible estacionalidad vinculada al verano austral. En contraste, durante junio, julio y agosto la actividad eléctrica es casi nula, posiblemente debido a condiciones meteorológicas más estables o menos propicias para tormentas eléctricas.

Es importante destacar que este gráfico muestra los totales mensuales acumulados. Un mes con una gran cantidad de registros con rayos (barra roja alta), como enero, no significa necesariamente que hubo tormentas eléctricas durante todo el mes. Por ejemplo, es posible que la gran mayoría de esos registros de rayos se hayan concentrado en un solo día o en unos pocos días con una tormenta eléctrica muy intensa y duradera. Una única tormenta que se prolongue durante varias horas podría generar cientos de intervalos de 5 minutos con actividad, inflando el total mensual. Por lo tanto, aunque el gráfico es excelente para ver la tendencia estacional, no detalla si la actividad fue esporádica e intensa o constante y moderada a lo largo del mes.

4.1.4 Análisis descriptivo para las caracteristicas temporales

Para cuantificar las diferencias en las condiciones meteorológicas y eléctricas, calculamos la estadística descriptiva de las variables de entrada para los dos escenarios: momentos con presencia de descargas y momentos sin ellas. Las Tablas 4.1 y 4.2 presentan un resumen comparativo de estos resultados.

Tabla 4.1: Estadística descriptiva para registros sin rayos (Flash = 0)

Variable	**Count**	**Mean**	**Std**	**Min**	**25 %**	**50 %**	**75 %**
Radiación	7659	1.11	1.44	0.00	0.00	0.00	2.48
Temperatura	7659	7.74	5.62	-5.70	3.33	6.08	12.41
Humedad	7659	62.30	26.46	2.94	38.53	67.31	86.75
Precipitación	7659	0.04	0.20	0.00	0.00	0.00	0.01
Presión	7659	100.39	0.99	97.44	99.64	100.67	101.13
u	7659	0.97	1.48	-3.74	-0.10	0.73	1.97
v	7659	-0.34	0.96	-5.57	-0.77	-0.37	0.12
sin_hour	7659	0.096	0.693	-1.00	-0.50	0.26	0.71
cos_hour	7659	0.046	0.714	-1.00	-0.71	0.00	0.71
sin_month	7659	-0.033	0.701	-1.00	-0.87	0.00	0.50
cos_month	7659	-0.068	0.709	-1.00	-0.87	0.00	0.50
ce_peak	7229	793.68	1731.43	-3675.11	23.79	505.58	998.96
ce_mean	7229	197.96	685.08	-9462.75	-2.54	220.35	434.72
cambio_polaridad	7229	0.30	0.46	0.00	0.00	0.00	1.00

Tabla 4.2: Estadística descriptiva para registros con rayos (Flash $\geq$ 1)

Variable	**Count**	**Mean**	**Std**	**Min**	**25 %**	**50 %**	**75 %**
Radiación	1125	1.41	1.48	0.00	0.00	0.91	2.80
Temperatura	1125	11.29	4.49	-0.74	7.54	11.18	14.89
Humedad	1125	56.43	20.04	13.81	39.31	54.44	73.31
Precipitación	1125	0.08	0.19	0.00	0.00	0.01	0.07
Presión	1125	99.57	0.82	97.55	98.91	99.55	100.26
u	1125	1.95	1.60	-2.72	0.80	2.06	3.19
v	1125	-0.62	1.00	-4.17	-1.21	-0.60	-0.01
sin_hour	1125	-0.653	0.388	-1.00	-0.97	-0.87	-0.50
cos_hour	1125	-0.310	0.572	-1.00	-0.87	-0.50	0.00
sin_month	1125	0.195	0.700	-1.00	-0.50	0.50	0.87
cos_month	1125	0.465	0.506	-0.87	0.00	0.50	0.87
ce_peak	1037	2731.50	5202.11	-3160.28	2.07	490.78	3058.93
ce_mean	1037	322.28	1473.80	-8286.17	-147.52	5.14	505.25
cambio_polaridad	1037	0.49	0.50	0.00	0.00	0.00	1.00

Al comparar los valores promedio entre registros con y sin presencia de rayos, se identifican diferencias claras que permiten caracterizar las condiciones típicas asociadas a eventos eléctricos. La temperatura muestra un aumento del 46 %, reflejando un entorno térmicamente más activo durante los eventos con rayos. La componente del viento en dirección este-oeste (u) se incrementa en más del 100 %, mientras que

la componente norte-sur (v) se vuelve más negativa, intensificándose aproximadamente en un 80 %, lo que sugiere movimientos verticales más pronunciados, probablemente vinculados a procesos convectivos.

En cuanto a la humedad relativa, se observa una leve disminución cercana al 9 % en presencia de rayos, lo que podría corresponder a una fase más madura del desarrollo convectivo, donde el aire en capas superiores ya ha perdido parte de su contenido de vapor. La presión atmosférica también desciende, en aproximadamente 0.8 %, lo que es consistente con la presencia de núcleos de baja presión asociados a condiciones inestables.

La precipitación media prácticamente se duplica (aumento del 106 %), lo que evidencia una estrecha relación entre descargas eléctricas y eventos de lluvia. Las diferencias más marcadas se presentan en el campo eléctrico ambiental: el valor medio aumenta en un 63 % y el valor peak se incrementa cerca de un 244 %, reafirmando su importancia como indicador directo de electrificación atmosférica. Por último, la variable binaria *cambio_polaridad* muestra una frecuencia 65 % mayor en presencia de rayos, lo cual podría estar relacionado con variaciones bruscas en la estructura de carga de la nube. En conjunto, estas diferencias cuantificadas respaldan la relevancia de las variables consideradas como insumos para modelos predictivos de descargas eléctricas.

4.1.5 Análisis del Grado de Simetría de los Datos

Con el objetivo de evaluar la distribución estadística de las variables involucradas en el modelo predictivo, se construyeron histogramas para cada una de ellas, diferenciando entre los registros con presencia de rayos ($Flash \geq 1$) y sin rayos ($Flash = 0$). Esta visualización permite explorar visualmente la simetría, presencia de colas o sesgos y concentración de valores.

- **Radiación:** La distribución es marcadamente asimétrica hacia la izquierda, con una gran concentración de valores en cero, especialmente cuando no hay rayos. En presencia de rayos, la cola derecha es más prolongada, lo que sugiere mayor actividad solar durante eventos eléctricos.
- **Temperatura:** Presenta una distribución cercana a la normal en ambos casos, pero desplazada hacia la derecha cuando hay rayos. La simetría se mantiene, aunque el grupo con rayos muestra una media más alta y una ligera mayor dispersión.
- **Humedad:** Exhibe una forma asimétrica con sesgo hacia la izquierda. En ausencia de rayos, los valores tienden a ser más altos. En cambio, los casos con rayos muestran una concentración en valores intermedios, reflejando una posible etapa madura del desarrollo convectivo.
- **Precipitación:** Su distribución es altamente sesgada hacia la izquierda, con muchos ceros en ambos grupos. Sin embargo, en registros con rayos aparece una cola más densa hacia valores mayores, indicando mayor acumulación durante eventos eléctricos.

- **Presión:** Tiene una distribución relativamente simétrica, con un leve sesgo negativo en ambos grupos. La curva se desplaza hacia valores ligeramente menores en presencia de rayos, lo cual es coherente con condiciones de baja presión asociadas a inestabilidad atmosférica.
- **u (componente este-oeste del viento):** La distribución es simétrica en ausencia de rayos y se vuelve más dispersa y asimétrica a la derecha cuando hay rayos, reflejando un aumento significativo en la magnitud del viento zonal durante estos eventos.
- **v (componente norte-sur del viento):** Muestra un leve sesgo hacia valores negativos en ambos grupos. En presencia de rayos, se observa mayor dispersión, lo cual puede estar vinculado a movimientos verticales intensificados.
- **sin_hour:** En registros sin rayos, la distribución es relativamente plana, mientras que con rayos se concentra en valores negativos. Esto sugiere que las descargas tienden a ocurrir en ciertas horas específicas del día (tarde o noche).
- **cos_hour:** Presenta una distribución bimodal. En registros con rayos, hay mayor concentración en valores negativos, indicando una mayor ocurrencia en horas alejadas del mediodía.
- **sin_month:** Su distribución es más dispersa y simétrica en ausencia de rayos, pero se concentra en valores positivos cuando hay rayos, lo que puede asociarse a una mayor frecuencia en meses específicos (por ejemplo, verano austral).
- **cos_month:** También muestra una forma bimodal. Con rayos, se observa una mayor presencia de valores positivos, lo que indica una tendencia estacional distinta respecto a los casos sin descargas.
- **ce_peak:** La variable más asimétrica del conjunto. Tiene una fuerte cola derecha, especialmente en eventos con rayos, donde los valores son significativamente más elevados. La dispersión es extrema y sugiere que algunos eventos concentran niveles muy altos de campo eléctrico pico.
- **ce_mean:** Aunque menos extrema que *ce_peak*, mantiene una asimetría positiva marcada. La distribución se aplana y extiende en los registros con rayos, reflejando una mayor intensidad promedio del campo eléctrico ambiental.
- **cambio_polaridad:** Variable binaria, con distribución claramente sesgada hacia cero en registros sin rayos. En eventos con rayos, la distribución se equilibra, mostrando una mayor frecuencia de ocurrencia de cambio de polaridad.

Todas estas observaciones respaldan la inclusión de las variables en el modelo, no solo por su valor medio sino también por sus características de dispersión y simetría, que podrían ser aprovechadas por arquitecturas no lineales de aprendizaje automático. Con el objetivo de evaluar la distribución estadística de las variables involucradas en el modelo predictivo, se construyeron histogramas para cada una de ellas, diferenciando entre los registros con presencia de rayos ($Flash \geq 1$) y sin rayos ($Flash = 0$). Esta visualización se complementa con los diagramas de caja presentados en el Anexo, los cuales permiten

explorar visualmente la simetría, dispersión y presencia de valores extremos de cada variable, destacando la influencia de outliers en el comportamiento de `ce_peak` y `ce_mean`.

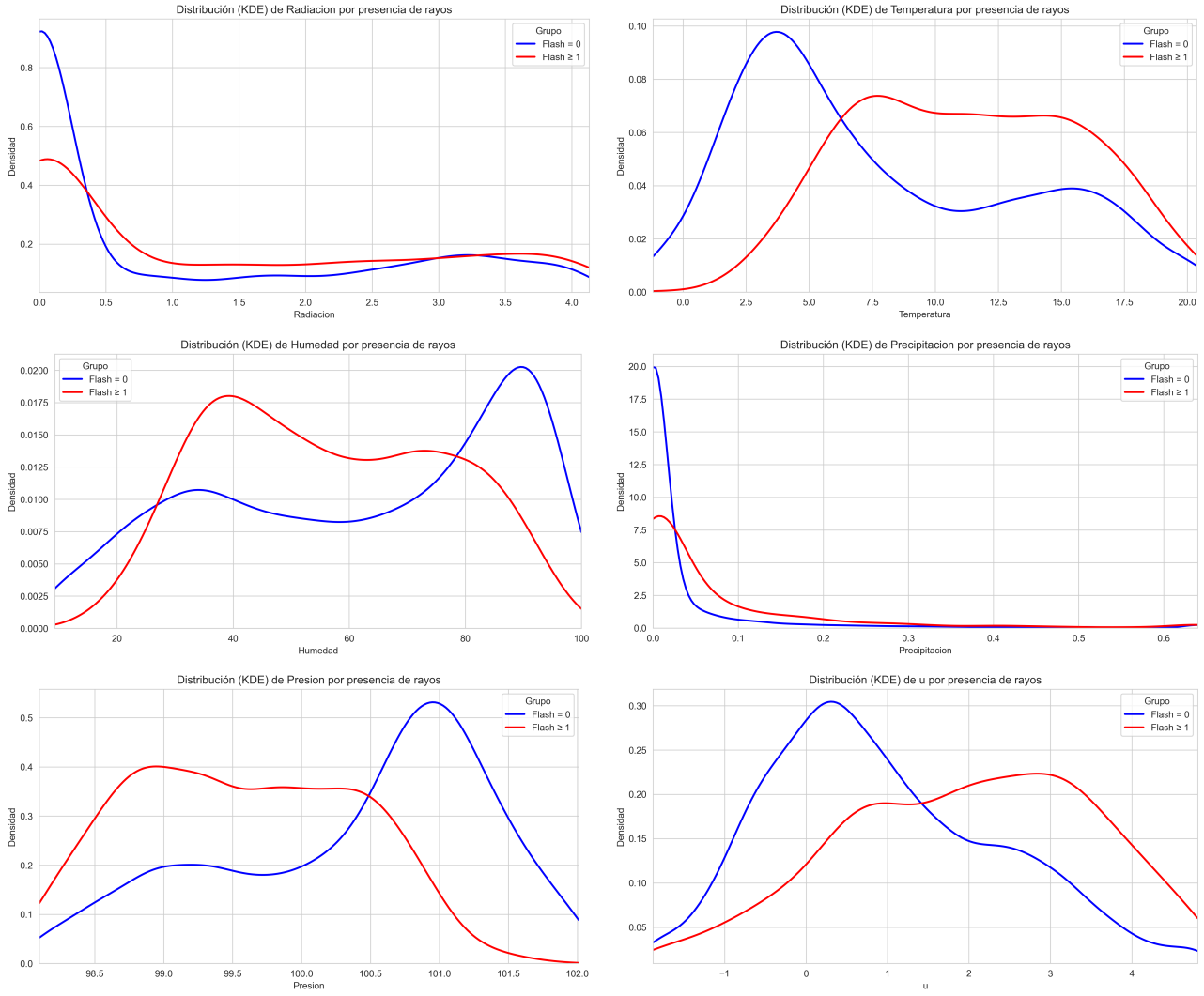

Figura 4.2: Distribución y simetría de variables meteorológicas (KDE) para casos con y sin rayos

Figura 4.3: Distribución y simetría de variables cíclicas y eléctricas (KDE) para casos con y sin rayos

4.1.6 Análisis de Relaciones y Correlaciones entre Variables

Para complementar el análisis descriptivo y de simetría, se exploraron las relaciones entre las variables meteorológicas y eléctricas mediante gráficos de densidad y correlación. La Figura 4.4 muestra un gráfico

tipo *pairplot* que compara distribuciones bidimensionales entre variables seleccionadas, diferenciando los casos con y sin rayos. Se observan agrupamientos diferenciados en variables como la temperatura, la radiación y el campo eléctrico, indicando patrones característicos durante la ocurrencia de descargas.

En complemento, la Figura 4.5 presenta la matriz de correlación entre variables. Destacan correlaciones positivas entre el campo eléctrico y la radiación, así como correlaciones negativas entre la presión y la mayoría de las variables meteorológicas activas. Estas relaciones refuerzan la hipótesis de que ciertas condiciones atmosféricas coexisten y podrían actuar como predictores de eventos eléctricos.

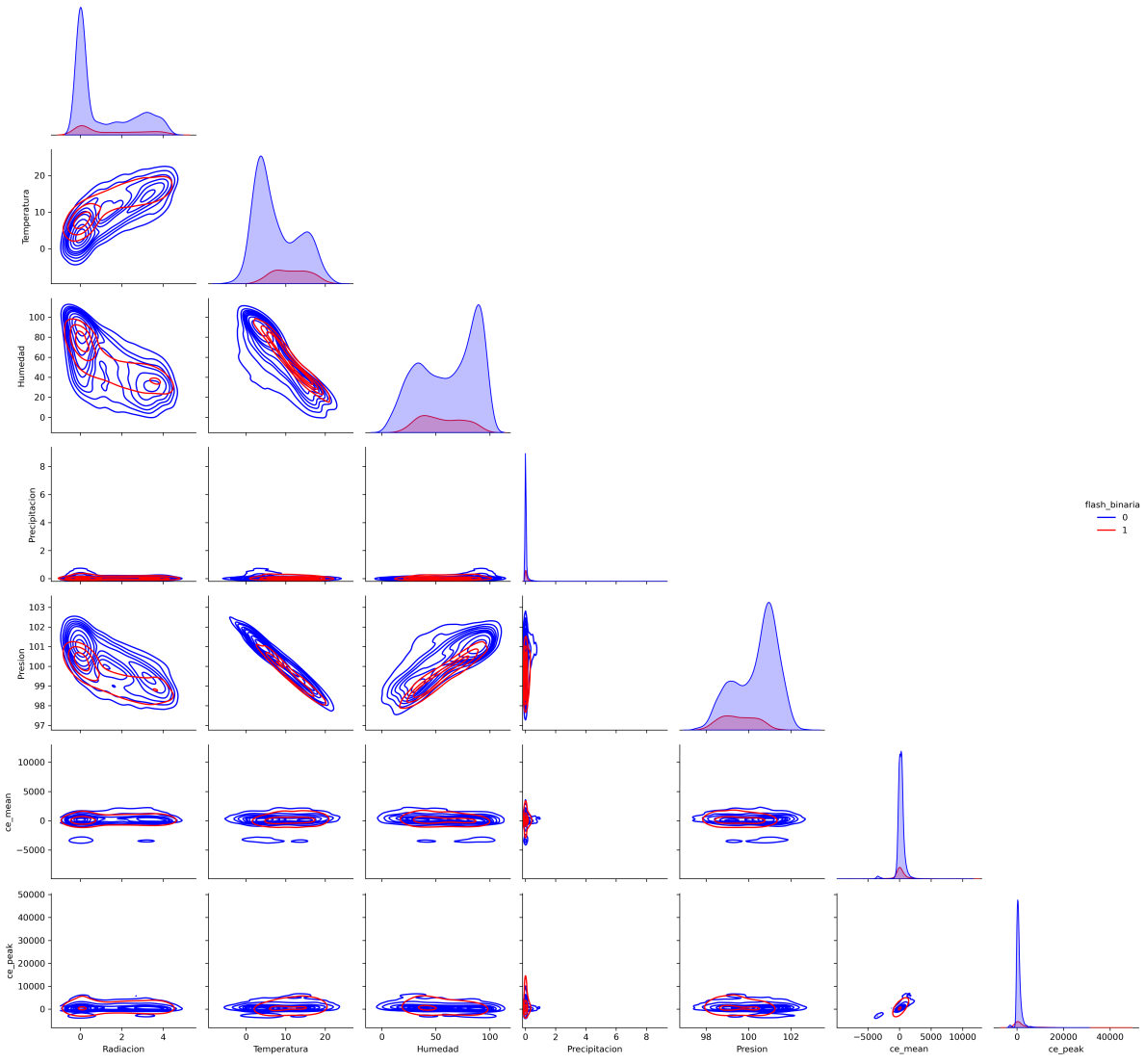

Figura 4.4: Relaciones bidimensionales entre variables según la presencia de rayos.

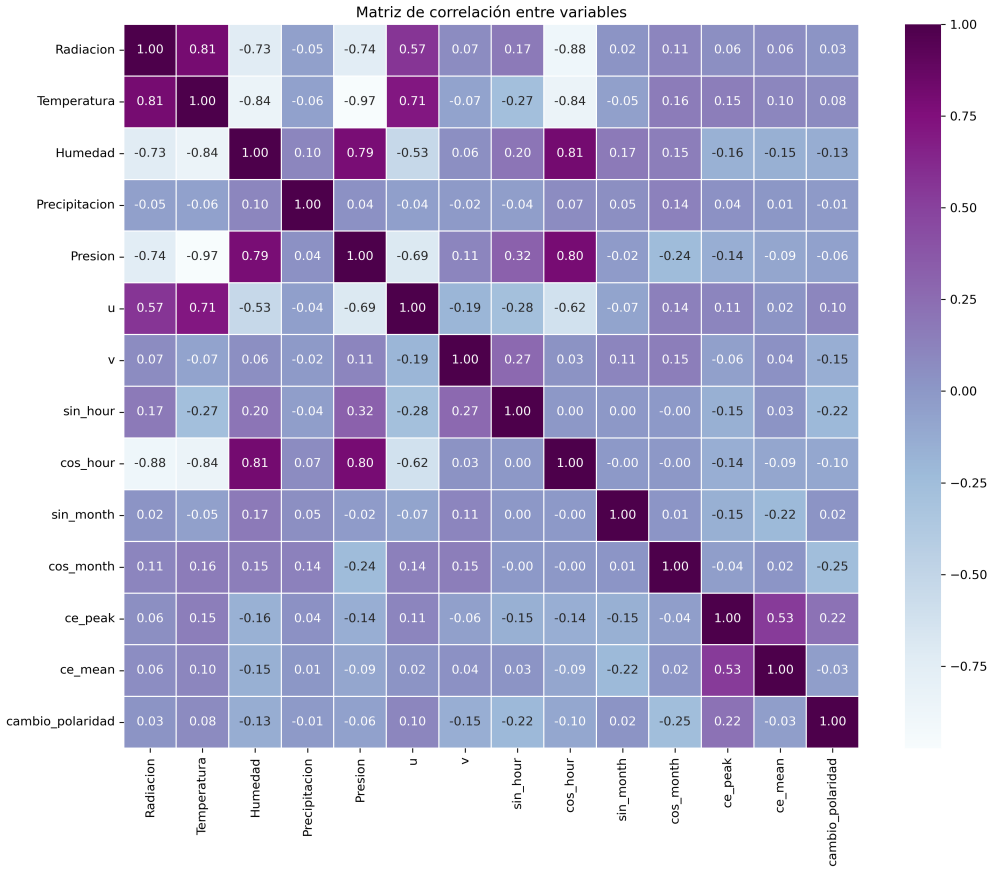

Figura 4.5: Matriz de correlación entre variables meteorológicas y eléctricas.

4.1.7 Trabajo Previo de los Datos

El conjunto de datos original presentaba valores faltantes en tres variables clave: *campo eléctrico pico*, *campo eléctrico promedio* y *cambio de polaridad del campo eléctrico*. Dado que estas variables son esenciales para la predicción de descargas atmosféricas, se aplicaron técnicas de imputación específicas según el tipo de dato.

- Las variables continuas *campo eléctrico pico* y *campo eléctrico promedio* fueron imputadas utilizando el algoritmo k-NN con $k = 5$ y pesos uniformes, técnica recomendada en el tratamiento de datos ambientales y series temporales [84].
- Para la variable categórica binaria *cambio de polaridad*, se utilizó también el algoritmo k-NN, pero con $k = 1$, de acuerdo con estrategias propuestas en la literatura para variables categóricas [85].

Adicionalmente, se incluyeron variables temporales derivadas de la fecha, descomponiendo la hora y el mes en sus componentes cíclicas mediante transformaciones seno y coseno. Esto permite capturar patrones

periódicos sin introducir discontinuidades artificiales. Por su parte, la dirección del viento fue convertida en sus componentes cartesianas u y v, representando las velocidades en dirección este y norte respectivamente, lo cual facilita su incorporación en los modelos [86].

Con el conjunto de datos preprocesado, se procedió a evaluar nuevamente las relaciones entre variables mediante un gráfico de calor actualizado, mostrado en la Figura 4.6, donde se incluyen las variables ya transformadas e imputadas.

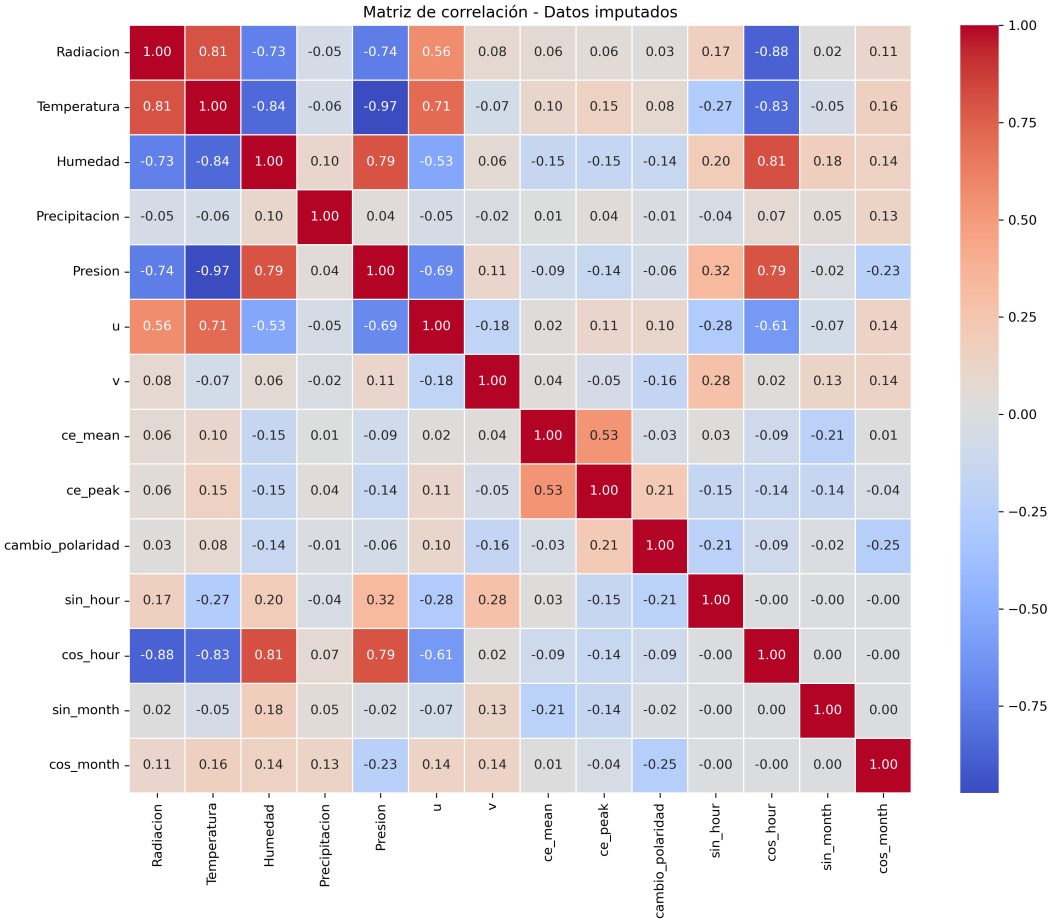

Figura 4.6: Matriz de correlación entre variables luego del proceso de imputación y transformación.

4.1.8 Análisis de Series Temporales

Con el objetivo de identificar patrones temporales relevantes para la predicción de descargas eléctricas, se realizó un análisis de series temporales sobre las principales variables del conjunto de datos. Este análisis incluyó la evolución temporal del número de rayos y de las variables meteorológicas y eléctricas más representativas.

En la Figura 4.7 se muestra la cantidad de descargas eléctricas por hora, destacando una distribución

heterogénea a lo largo del año, con eventos más frecuentes en determinados periodos. Esta concentración temporal de eventos puede estar relacionada con ciclos estacionales y condiciones atmosféricas específicas.

Posteriormente, se analizaron las variaciones horarias de variables como la temperatura, la humedad relativa, la precipitación y el campo eléctrico (valor medio), cuya evolución se muestra en la Figura 4.8. Se observa que, si bien algunas variables como la temperatura y la humedad presentan comportamientos cíclicos evidentes, otras como el campo eléctrico muestran picos irregulares asociados a condiciones eléctricas favorables para la ocurrencia de rayos.

Este análisis permite vislumbrar relaciones no lineales y dependencias temporales entre variables, lo cual justifica el uso de arquitecturas basadas en secuencias para capturar la dinámica temporal del fenómeno eléctrico atmosférico.

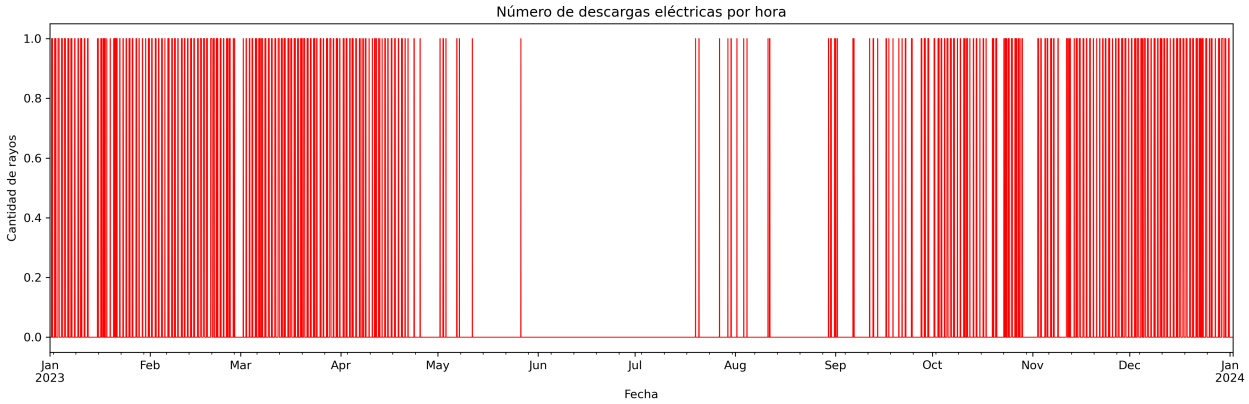

Figura 4.7: Número de descargas eléctricas horarias.

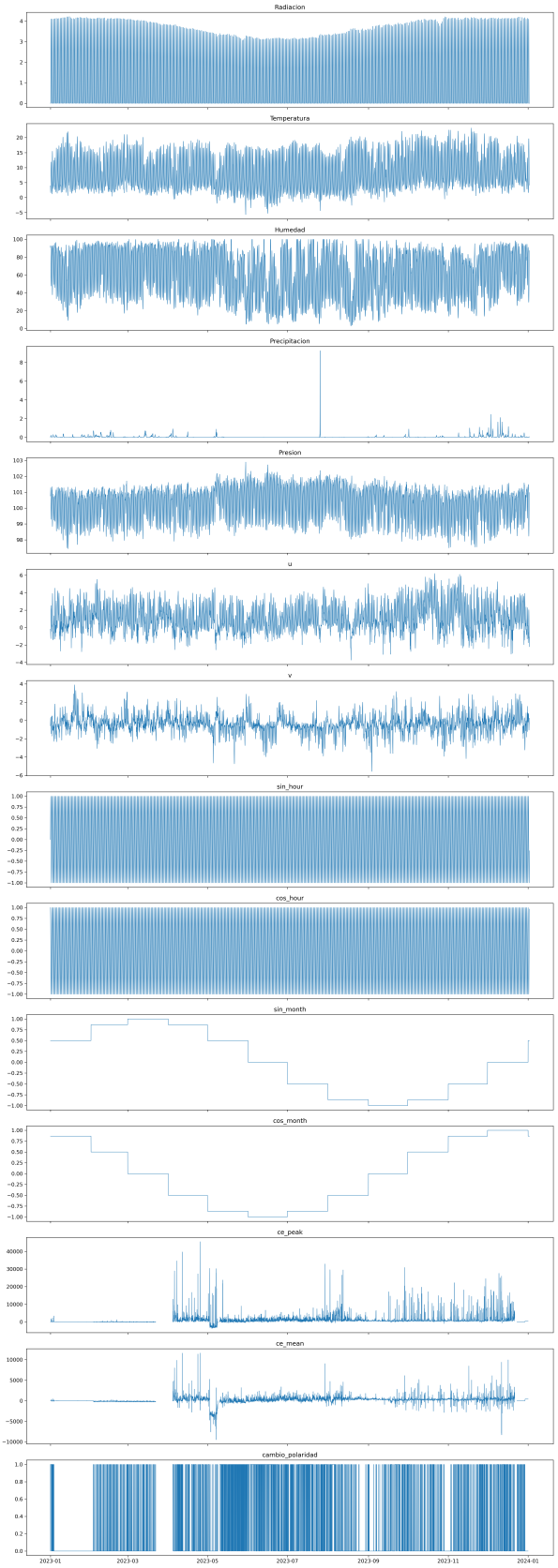

Figura 4.8: Evolución temporal de variables meteorológicas y eléctricas.

Esta sección presenta los resultados obtenidos a partir del desarrollo e implementación de modelos predictivos destinados a anticipar la ocurrencia de descargas eléctricas atmosféricas.

4.1.9 Modelo Base

Como punto de partida para el desarrollo del sistema predictivo, se implementó un modelo base utilizando una arquitectura de Perceptrón Multicapa (MLP), cuyo objetivo fue establecer una línea base de desempeño sin incorporar información secuencial ni dependencias temporales. El modelo fue alimentado exclusivamente con los valores registrados en un instante de tiempo, ignorando la evolución histórica de las variables.

El conjunto de variables de entrada incluyó trece predictores relevantes: condiciones meteorológicas (radiación, temperatura, humedad, precipitación y presión), componentes del viento (u y v), variables cíclicas horarias y mensuales (representadas mediante transformaciones seno y coseno), y tres variables relacionadas con el campo eléctrico ambiental: el valor medio, el valor pico y el cambio de polaridad.

La arquitectura final seleccionada, tras una exploración preliminar de configuraciones, consistió en dos capas densas ocultas con 64 y 32 neuronas respectivamente, activaciones $ReLU$, normalización por lotes y abandono (*dropout*) del 30 %. La salida del modelo fue una única neurona con activación sigmoidal para realizar la clasificación binaria. El modelo fue optimizado utilizando el algoritmo AdamW con una tasa de aprendizaje de 5×10^{-4} y un término de decaimiento de peso, aplicando además pesos de clase para abordar el desbalance de la variable objetivo.

Para evaluar su rendimiento se empleó validación cruzada estratificada de 10 particiones (*10-fold cross-validation*). El modelo logró un F1-score promedio de 0,6114 con una desviación estándar de 0,0219, lo que refleja un desempeño consistente a lo largo de los distintos subconjuntos de validación.

La matriz de confusión acumulada para los 10 pliegues fue:

$$\begin{bmatrix} 6136 & 1093 \\ 100 & 937 \end{bmatrix}$$

- Verdaderos negativos (TN): 6136 instancias correctamente clasificadas como ausencia de rayos.
- Falsos positivos (FP): 1093 instancias clasificadas erróneamente como presencia de rayos.
- Falsos negativos (FN): 100 eventos de rayo no detectados.
- Verdaderos positivos (TP): 937 eventos de rayo correctamente identificados.

Estos resultados sugieren que el modelo base es capaz de capturar patrones relevantes en los datos para detectar una proporción significativa de eventos de rayos, con una baja tasa de falsos negativos, lo que resulta clave para aplicaciones de alerta temprana. Sin embargo, la elevada cantidad de falsos positivos indica una tendencia del modelo a sobrealertar, lo cual podría generar alarmas innecesarias en un contexto operativo.

A pesar de su simplicidad, el MLP establece una línea base sólida para futuras comparaciones. No obstante, su limitación más evidente radica en la ausencia de mecanismos que capten la evolución temporal de las variables. En consecuencia, en las siguientes secciones se explorarán arquitecturas avanzadas como redes convolucionales y redes LSTM, las cuales podrían mejorar significativamente la precisión y confiabilidad del sistema al incorporar información secuencial en la predicción.

4.1.10 Preparación de los Datos para Entrenamiento y Evaluación

Antes del desarrollo de los modelos predictivos, fue necesario estructurar cuidadosamente los conjuntos de datos para garantizar una evaluación robusta, evitar el *data leakage* y simular condiciones reales de predicción operativa. Para ello, se optó por una división temporal de los datos, separando los registros en tres subconjuntos: entrenamiento, validación y prueba. Esta estrategia permite respetar la secuencia cronológica de los eventos y evaluar el modelo sobre datos verdaderamente no vistos.

- **Conjunto de entrenamiento**: conformado por los datos correspondientes a los meses de febrero a noviembre, excluyendo enero y diciembre. Este subconjunto fue utilizado para el ajuste de pesos de los modelos.
- **Conjunto de validación**: se seleccionó el mes de enero como conjunto de validación, debido a que presenta una alta cantidad de registros disponibles, lo que permite una evaluación estadísticamente sólida durante el ajuste de hiperparámetros mediante *grid search*, así como para la aplicación de técnicas de *early stopping* y reducción adaptativa de la tasa de aprendizaje.
- **Conjunto de prueba**: se utilizó el mes de diciembre como conjunto de prueba final. Esta decisión también se basó en la abundancia de datos disponibles en ese mes, lo que garantiza una evaluación confiable y representativa del desempeño del modelo en condiciones operacionales reales.

Cabe destacar que todos los subconjuntos fueron construidos manteniendo la proporción original entre clases mediante muestreo estratificado, dado que la clase positiva (eventos de rayos) representa una fracción minoritaria del total de observaciones. Esta precaución fue esencial para evitar sesgos durante el entrenamiento y asegurar que las métricas de evaluación reflejaran correctamente la capacidad del modelo

Las variables de entrada seleccionadas para alimentar el modelo fueron: `Radiacion`, `Temperatura`, `Humedad`, `Presion`, `u`, `v`, `sin_hour`, `cos_hour`, `sin_month`, `cos_month`, `cambio_polaridad` y `ce_peak`.

Todas las variables fueron previamente normalizadas utilizando parámetros ajustados sobre el conjunto de entrenamiento.

4.1.11 Descripción Modelo prueba LSTM para Predicción a 1 Hora

Con el objetivo de evaluar la capacidad del modelo para realizar predicciones en horizontes de muy corto plazo, se entrenó una red neuronal basada en memoria a largo corto plazo (*Long Short-Term Memory*, LSTM) con un horizonte de predicción de 1 hora. Este modelo fue alimentado con secuencias de 24 horas previas de datos meteorológicos, cíclicos y de campo eléctrico, permitiendo capturar la dinámica temporal que precede a la ocurrencia de una descarga atmosférica.

La arquitectura empleada se compone de dos capas LSTM secuenciales con 64 unidades cada una, seguidas de normalización por lotes (*Batch Normalization*), una capa densa intermedia de 32 neuronas con activación `ReLU`, y una capa de salida con activación `sigmoid` para clasificación binaria. Se utilizó la función de pérdida `binary_crossentropy`, el optimizador AdamW con una tasa de aprendizaje de 5×10^{-4} y regularización mediante *weight decay*. Para prevenir sobreajuste, se aplicaron técnicas de *early stopping* y reducción adaptativa de la tasa de aprendizaje. Además, se ajustaron los pesos de clase para contrarrestar el desbalance natural en los datos entre eventos de rayos y no-rayos.

4.1.12 Selección de Hiperparámetros

Con el objetivo de optimizar el desempeño del modelo LSTM, se llevó a cabo una búsqueda exhaustiva de hiperparámetros mediante la técnica de *Grid Search*, evaluando combinaciones en el número de unidades de las capas LSTM, la cantidad de neuronas en la capa densa intermedia y las tasas de *dropout*. La métrica de evaluación principal durante este proceso fue el F1-score para la clase positiva (ocurrencia de rayos), calculado sobre el conjunto de validación.

En la Tabla 4.3 se presentan las cinco configuraciones que alcanzaron los mejores desempeños. Las combinaciones seleccionadas reflejan un balance entre profundidad y regularización, y coinciden en la efectividad de tasas de *dropout* moderadas y configuraciones LSTM con 64 o 128 unidades.

Tabla 4.3: Top 5 combinaciones de hiperparámetros ordenadas por F1-score en validación.

Ranking	Unidades LSTM 1	Unidades LSTM 2	Unidades Densas	Dropout	F1-score
1	64	64	64	0.20	0.6469
2	128	128	64	0.30	0.6423
3	128	64	32	0.30	0.6324
4	64	32	64	0.25	0.6294
5	32	16	16	0.20	0.6272

La mejor configuración alcanzó un F1-score de **0.6469**, utilizando dos capas LSTM de 64 unidades, una capa densa de 64 neuronas, y una tasa de *dropout* de 0.2. Esta arquitectura fue seleccionada como finalista para el entrenamiento definitivo del modelo y su evaluación sobre el conjunto de prueba.

4.1.13 Resultados Modelo de Prueba para Predicción a 1 Hora

La evaluación del modelo en el conjunto de prueba (correspondiente al mes de diciembre) se resume en la siguiente matriz de confusión:

$$\begin{bmatrix} 403 & 63 \\ 32 & 154 \end{bmatrix}$$

- Verdaderos negativos (TN): 403 instancias correctamente clasificadas como no-rayos.
- Falsos positivos (FP): 63 instancias clasificadas incorrectamente como rayos.
- Falsos negativos (FN): 32 eventos reales no detectados.
- Verdaderos positivos (TP): 154 eventos correctamente identificados como rayos.

Las métricas de clasificación obtenidas para cada clase se resumen en la Tabla 4.4. Se observa que el modelo presenta un rendimiento destacado en la identificación de eventos reales de rayos (clase positiva), alcanzando un *recall* de 0,83. Esto significa que el 83 % de los casos en que efectivamente ocurrió una descarga eléctrica fueron correctamente identificados por la red LSTM.

Tabla 4.4: Métricas de desempeño por clase

Clase	Precisión	Recall	F1-score	Accuracy
0 (No Rayo)	0.93	0.86	0.89	0.86
1 (Rayo)	0.71	0.83	0.76	0.83

En términos de precisión, la clase negativa (no rayo) presenta un valor de 0,93, lo que indica que el modelo se equivoca muy pocas veces al predecir una ausencia de rayos. Por otro lado, la clase positiva alcanza una precisión de 0,71, lo que sugiere que alrededor del 29 % de las alertas por rayos fueron falsas alarmas. No obstante, este resultado es aceptable en sistemas de alerta temprana, donde es preferible generar una alerta innecesaria que omitir un evento peligroso.

El F1-score para la clase positiva es de 0,76, lo que refleja un equilibrio razonable entre precisión y sensibilidad, especialmente relevante en contextos con desbalance de clases. Finalmente, el modelo logró

una precisión global (*accuracy*) del 86 %, evidenciando su solidez para tareas de clasificación binaria en contextos de predicción de descargas eléctricas a muy corto plazo.

Este desempeño representa una mejora significativa respecto al modelo base, confirmando que la arquitectura LSTM propuesta es capaz de capturar patrones relevantes en los datos históricos y anticipar la ocurrencia de descargas atmosféricas de manera efectiva.

4.1.14 Comparación de Desempeño según el Horizonte de Predicción

Con el propósito de analizar la capacidad anticipativa del modelo, se entrenaron versiones independientes del LSTM para horizontes de predicción de 1, 5, 10, 15 y 24 horas. En todos los casos se utilizó una secuencia de entrada de 24 horas de datos históricos, manteniendo constante la arquitectura, el conjunto de variables y los parámetros de entrenamiento, con el fin de asegurar una comparación equitativa entre horizontes.

La Tabla 4.5 muestra las métricas de desempeño obtenidas para la clase positiva en el conjunto de prueba. Se observa una degradación progresiva de las métricas conforme se incrementa el horizonte temporal, especialmente en *recall* y *F1-score*. Sin embargo, hasta el horizonte de 10 horas, el modelo mantiene un F1-score igual o superior a 0.74, lo que demuestra una alta capacidad de anticipación en escenarios operativos realistas.

Tabla 4.5: Desempeño del modelo LSTM en función del horizonte de predicción

Horizonte	**Precisión (rayo)**	**Recall (rayo)**	**F1-score (rayo)**	**Accuracy**
1 hora	0.71	0.83	0.76	0.85
5 horas	0.71	0.77	0.74	0.85
10 horas	0.70	0.77	0.74	0.84
15 horas	0.71	0.71	0.71	0.84
24 horas	0.66	0.66	0.66	0.81

Las matrices de confusión asociadas a cada horizonte de predicción se presentan en la Tabla 4.6. A medida que el horizonte se extiende, se incrementa el número de falsos negativos, afectando directamente la capacidad del modelo para captar todos los eventos reales. No obstante, los verdaderos positivos se mantienen relativamente altos hasta 15 horas, lo que evidencia que el modelo conserva capacidad anticipativa incluso en contextos más exigentes.

Tabla 4.6: Matrices de confusión por horizonte de predicción (conjunto de prueba)

Horizonte	TN	FP	FN	TP
1 hora	403	63	32	154
5 horas	410	56	42	140
10 horas	403	59	41	140
15 horas	404	53	52	129
24 horas	396	60	59	114

La evolución de las matrices de confusión confirma el sólido rendimiento del modelo hasta las 10 horas de anticipación, con un número moderado de falsos negativos. A partir del horizonte de 15 horas, la sensibilidad se reduce más abruptamente, aunque con niveles de precisión aún aceptables para aplicaciones operativas.

4.1.15 Desempeño con una Selección Alternativa de Variables

Con el fin de analizar la robustez del modelo y su capacidad de generalización, se realizó un segundo experimento utilizando una selección alternativa de variables. En esta configuración se descartó la variable `ce_mean` (campo eléctrico promedio), reteniendo solo los siguientes atributos: radiación, temperatura, humedad, precipitación, presión, componentes del viento (u, v), codificaciones cíclicas (hora y mes), y el cambio de polaridad del campo eléctrico.

Se exploraron 20 combinaciones de hiperparámetros mediante *grid search*, y se seleccionó la misma arquitectura que en el modelo principal: dos capas LSTM de 64 unidades, una capa densa de 64 unidades y *dropout* de 0.2. El entrenamiento se llevó a cabo con AdamW, tasa de aprendizaje de 5×10^{-4}, *weight decay* de 10^{-5}, y las mismas técnicas de regularización.

El modelo alcanzó un **F1-score de 0,74** para la clase positiva, demostrando un rendimiento comparable al obtenido con la configuración original. La matriz de confusión correspondiente fue la siguiente:

Tabla 4.7: Matriz de confusión para la configuración alternativa de variables

	Pred. No Rayo	Pred. Rayo
Real No Rayo	410	72
Real Rayo	36	140

Este resultado se traduce en una *precisión* del 71 % y un *recall* del 77 % para la clase positiva. En comparación con la configuración inicial, se observa un leve aumento de falsos negativos (de 36 a 42), pero una reducción significativa de falsos positivos (de 72 a 56), lo cual mejora la precisión general del sistema. Esto demuestra que el modelo conserva un rendimiento estable incluso ante modificaciones en el conjunto de entrada, confirmando la existencia de redundancias entre algunas variables, y evidenciando que el valor

pico y el cambio de polaridad contienen la mayor parte de la señal útil del campo eléctrico.

4.1.16 Impacto del Horizonte de Predicción en el Desempeño del Modelo

Para una evaluación integral, la Tabla 4.8 resume el comportamiento del modelo LSTM en función del horizonte de predicción. Se incluyen las métricas más relevantes para la clase positiva (rayos): F1-score, *recall*, *precisión* y número de falsos negativos, considerados el tipo de error más crítico en sistemas de alerta temprana.

Tabla 4.8: Comparación de métricas para la clase positiva según el horizonte de predicción

Horizonte	**F1-score**	**Recall**	**Precisión**	**Falsos Negativos (FN)**
1 hora	0.74	0.81	0.68	36
5 horas	0.68	0.70	0.67	55
10 horas	0.69	0.67	0.72	60
15 horas	0.62	0.50	0.80	90
24 horas	0.71	0.73	0.69	46

El modelo exhibe su mejor desempeño en el horizonte de 1 hora, con un F1-score de 0.74 y un *recall* del 81 %. A medida que el horizonte se amplía, la sensibilidad del modelo disminuye, alcanzando su mínimo a las 15 horas, con un *recall* del 50 % y el mayor número de falsos negativos (90). De forma interesante, el rendimiento se recupera en el horizonte de 24 horas, con un F1-score de 0.71. Esta mejora sugiere que el modelo logra captar patrones de mayor escala vinculados a procesos sinópticos, a pesar de la complejidad inherente a predicciones de largo plazo. No obstante, dicha recuperación podría estar influenciada por la distribución particular de eventos en el conjunto de prueba, lo que refuerza la necesidad de una evaluación contextualizada.

En conjunto, los resultados revelan la capacidad del modelo para realizar predicciones precisas tanto a corto como a largo plazo, aunque con diferencias significativas según la selección de variables utilizada. En particular, la primera selección, que incluye la variable `ce_peak` pero omite la precipitación, mostró un rendimiento sobresaliente para horizontes cortos y medios, alcanzando F1-scores superiores a 0.74 hasta las 10 horas y manteniéndose relativamente estable incluso a las 15 horas. Sin embargo, esta configuración experimentó una caída en el desempeño al extenderse el horizonte a 24 horas, donde el F1-score se redujo a 0.71, sugiriendo que el valor pico del campo eléctrico captura señales de corto plazo pero no representa adecuadamente patrones sinópticos de largo plazo.

Por otro lado, la segunda selección de variables, que elimina `ce_peak` pero incorpora la precipitación, produjo un comportamiento inverso. Aunque su rendimiento en horizontes intermedios (5 a 15 horas) fue algo inferior, mostró un mejor desempeño en los extremos: logró un F1-score de 0.74 para 1 hora y una recuperación efectiva a 24 horas, lo cual indica que la inclusión de variables como la precipitación favorece la captura de condiciones atmosféricas persistentes, relevantes para la predicción en escalas más amplias.

Estas observaciones permiten concluir que la primera selección es preferible cuando se priorizan predicciones de corto a mediano plazo, especialmente en contextos de operación inmediata o vigilancia en tiempo real. En contraste, la segunda selección ofrece ventajas cuando se busca anticipar condiciones favorables para la ocurrencia de rayos en ventanas temporales extendidas, como planificación o mitigación de riesgos a nivel regional. La evidencia sugiere que un sistema híbrido, que combine señales eléctricas de alta resolución con variables meteorológicas de mayor alcance, podría ofrecer un balance óptimo entre sensibilidad y precisión en distintos horizontes de predicción.

4.1.17 Comparación de arquitecturas y selecciones de variables para predicción a 1 hora

Se evaluó el desempeño de tres configuraciones distintas para un horizonte de predicción de una hora, con el fin de analizar el efecto de la arquitectura y la selección de variables sobre la capacidad predictiva del modelo. En primer lugar, se consideró un modelo baseline basado en una red MLP alimentada con el conjunto completo de atributos disponibles: radiación, temperatura, humedad, presión, precipitación, componentes del viento (u y v), codificaciones cíclicas (hora y mes), y las variables derivadas del campo eléctrico (`ce_mean`, `ce_peak` y el cambio de polaridad). Este modelo obtuvo un F1-score de 0.61 para la clase positiva, acompañado de una alta tasa de falsos positivos (1093) y 100 falsos negativos. La matriz de confusión indicó un pobre equilibrio entre sensibilidad y precisión, lo cual lo hace inadecuado para aplicaciones operativas donde las falsas alarmas deben mantenerse al mínimo.

En segundo lugar, se entrenó una arquitectura LSTM utilizando una primera selección de atributos que excluyó tanto la precipitación como `ce_mean`, pero mantuvo la variable `ce_peak`. Esta configuración combinó variables meteorológicas (radiación, temperatura, humedad, presión, u, v), codificaciones cíclicas y el cambio de polaridad del campo eléctrico. El modelo alcanzó el mejor desempeño entre las tres configuraciones, con un F1-score de 0.76, 154 verdaderos positivos y solo 32 falsos negativos. Además, logró reducir de forma significativa los falsos positivos a 63, lo que confirma la utilidad de `ce_peak` como una señal relevante para la predicción de rayos.

Finalmente, se probó una segunda selección de entrada que prescindió de `ce_peak`, pero incorporó la precipitación como variable adicional. El resto de los atributos se mantuvo igual a la configuración anterior. En este caso, el modelo también logró un buen desempeño, alcanzando un F1-score de 0.74, con 140 verdaderos positivos y 36 falsos negativos, lo que representa una leve pérdida en sensibilidad respecto a la primera selección, aunque con un aumento marginal de los falsos positivos a 72. Estos resultados demuestran que, en ausencia de `ce_peak`, la precipitación permite compensar parcialmente la pérdida de señal eléctrica, aunque sin igualar el rendimiento óptimo alcanzado cuando dicha variable está presente.

La Tabla 4.9 resume cuantitativamente los resultados observados en cada configuración, destacando la superioridad de la arquitectura LSTM frente al modelo MLP, y la relevancia de las variables derivadas del campo eléctrico sobre la precisión del sistema predictivo.

Tabla 4.9: Comparación de desempeño para predicción a 1 hora

Configuración	F1-score clase 1	TP	TN	FP	FN
MLP (todas las variables)	0.61	86	254	1093	100
LSTM con `ce_peak` (sin precipitación)	0.76	154	403	63	32
LSTM sin `ce_peak` (con precipitación)	0.74	140	410	72	36

En síntesis, el uso de arquitecturas recurrentes como LSTM resulta fundamental para capturar la dinámica temporal del fenómeno, y la variable `ce_peak` representa un insumo insustituible para mantener la sensibilidad sin sacrificar la precisión. La precipitación, en cambio, mejora el modelo cuando `ce_peak` no está disponible, pero no logra reemplazar completamente su valor predictivo.

4.2 Resultados en Argentina

El estudio se centró en un conjunto de datos de la zona metropolitana de Buenos Aires, con coordenadas Lat $-34{,}58$, Lon $-58{,}48$. El dataset integró mediciones de campo eléctrico de alta frecuencia, variables meteorológicas de superficie y registros de descargas eléctricas de las redes GLM y WWLLN. El sensor de campo eléctrico operó a 5 segundos de muestreo, a partir del cual se extrajeron el valor máximo, el promedio y la detección de cambios de polaridad por cada ventana de tiempo.

El dataset final para el año 2021, seleccionado por su mayor densidad de eventos en la red GLM, incluyó 101,209 registros sin rayos y 3,911 con descargas. Las variables abarcaron parámetros eléctricos, meteorológicos (temperatura, humedad, presión, viento) y cíclicos (hora y mes).

4.2.1 Análisis estadístico frente a GLM y WWLLN

La Tabla 4.10 presenta las estadísticas descriptivas de las variables al considerar dos configuraciones de entrenamiento: una red entrenada con registros de rayos de GLM y otra entrenada con registros de WWLLN. En este análisis, el campo eléctrico corresponde a las mediciones instrumentales realizadas en superficie, mientras que GLM y WWLLN actúan únicamente como fuentes de referencia para la identificación de los eventos de rayo. Los resultados muestran diferencias relevantes en la relación entre campo eléctrico y rayos según la red utilizada para el entrenamiento. En particular, cuando el modelo se entrena con WWLLN, las variables eléctricas presentan una dispersión significativamente mayor: la desviación estándar del campo eléctrico peak alcanza 240,1 V/m, frente a los 68,0 V/m observados en la red entrenada con GLM. Del mismo modo, el valor máximo llega a $3,715$ V/m en la red entrenada con WWLLN, más del triple que los $1,183$ V/m en la red entrenada con GLM. Esto indica que la red entrenada con WWLLN está asociada a eventos de mayor energía, aunque también incorpora una mayor proporción de ruido y valores atípicos, lo cual representa un desafío adicional para el modelado

En contraste, las variables meteorológicas muestran tendencias opuestas: para GLM, los eventos de rayos ocurren en promedio con mayor temperatura y menor humedad relativa, mientras que para WWLLN, se asocian a alta humedad y temperaturas más moderadas. Esta inconsistencia refuerza la idea de que el campo eléctrico es la variable más determinante y confiable, ya que las condiciones meteorológicas locales parecen influir de manera distinta según la red de detección.

En términos generales, la alta varianza y los valores atípicos en las mediciones de campo eléctrico de WWLLN, especialmente en el peak, comprometen la relación señal-ruido, lo que se convierte en un factor clave para el desempeño de los modelos predictivos.

Tabla 4.10: Estadística descriptiva de las variables para eventos detectados por GLM y WWLLN en Argentina.

Variable	**GLM (con rayos)**			**WWLLN (con rayos)**		
	Media	**Desv. Est.**	**Max**	**Media**	**Desv. Est.**	**Max**
Campo eléctrico promedio (V/m)	-28.9	56.1	1032	-19.0	139.8	1516
Campo eléctrico peak (V/m)	-20.3	68.0	1183	23.6	240.1	3715
Cambio de polaridad (max)	0.018	0.133	1.0	0.114	0.318	1.0
Temperatura (°C)	25.8	5.0	36.8	19.7	5.3	36.5
Humedad relativa (%)	46.8	16.5	94.3	69.5	19.0	98.0
Presión (hPa)	1011.0	5.2	1029	1008.7	4.8	1028
Viento U (m/s)	6.1	10.4	30.8	1.4	10.1	38.2
Viento V (m/s)	-5.8	9.1	23.5	-8.8	10.9	29.0
sin(hour)	-0.327	0.389	0.75	-0.037	0.721	1.00
cos(hour)	-0.832	0.221	0.99	-0.018	0.691	1.00
sin(month)	0.398	0.641	1.00	0.102	0.735	1.00
cos(month)	0.415	0.507	1.00	0.110	0.660	1.00

4.2.2 Correlación de variables con eventos GLM y WWLLN

Se evaluó la correlación de Pearson (ρ) entre las variables y la ocurrencia de descargas. Como se observa en la Figura 4.9, las correlaciones para GLM son muy débiles, con coeficientes cercanos a cero para las variables de campo eléctrico ($\rho \approx 0{,}00$ y $\rho \approx 0{,}01$). Solo la temperatura muestra una correlación positiva moderada ($\rho \approx 0{,}23$).

Para WWLLN, las correlaciones son ligeramente más fuertes, como se muestra en la Figura 4.10. El campo eléctrico peak ($\rho \approx 0{,}16$) y la humedad relativa ($\rho \approx 0{,}16$) son las variables más asociadas a los eventos, mientras que la presión atmosférica presenta una correlación negativa moderada ($\rho \approx -0{,}35$).

Las bajas correlaciones generales ($< 0{,}2$) para ambas redes indican que la relación entre las variables predictoras y la ocurrencia de rayos no es lineal. Esta es una conclusión crucial, ya que justifica la necesidad de explorar modelos más complejos y no lineales que puedan capturar las dinámicas subyacentes, en lugar

de depender de relaciones simples.

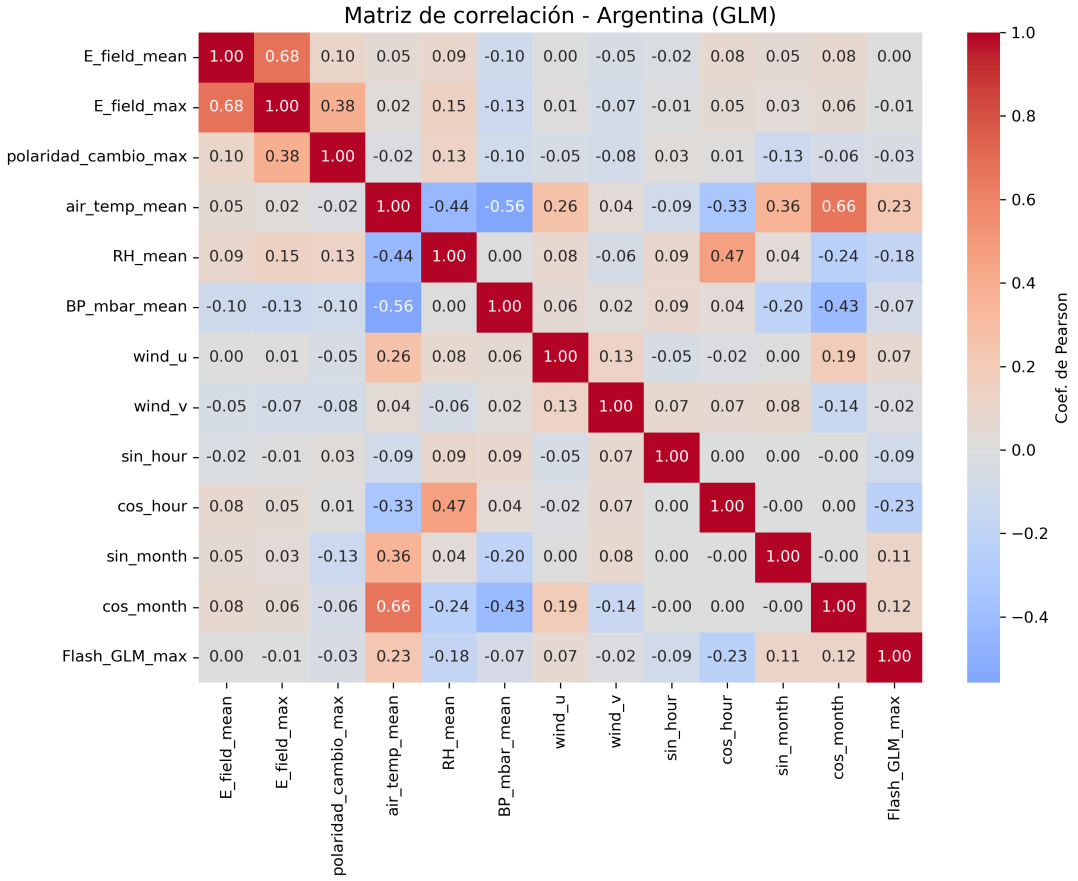

Figura 4.9: Matriz de correlación de Pearson entre variables predictoras y eventos detectados por GLM en Argentina.

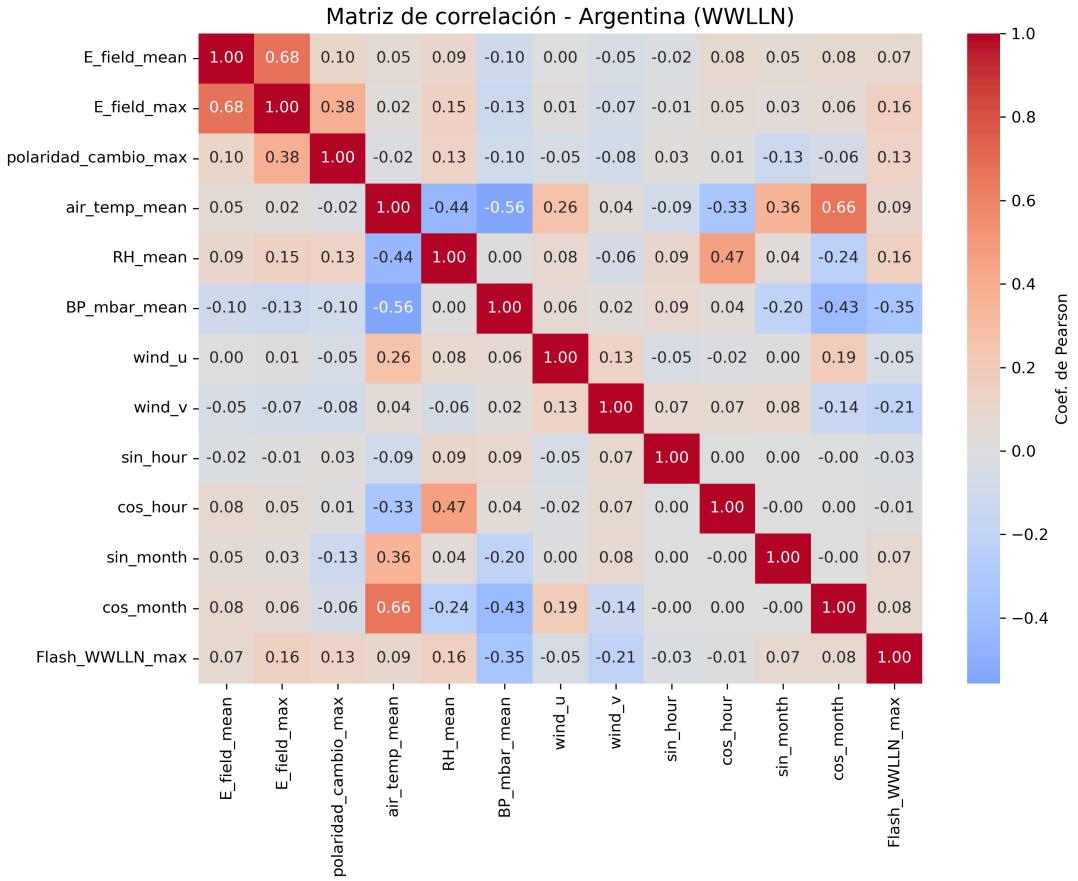

Figura 4.10: Matriz de correlación de Pearson entre variables predictoras y eventos detectados por WWLLN en Argentina.

4.2.3 Resultados de los Modelos Predictivos en Argentina 2021

El análisis de los modelos se centró en el F1-score debido al desbalance de los datos. Se presentan las matrices de confusión para evaluar los errores de cada modelo, prestando especial atención a los falsos positivos (FP) y falsos negativos (FN).

Modelo LSTM vs. MLP en GLM

Tabla 4.11: Métricas de evaluación LSTM (GLM, Argentina 2021).

	Precision	Recall	F1-score
Clase 0	0.99	0.81	0.89
Clase 1	0.37	0.91	0.53
Accuracy		0.82	

Tabla 4.12: Matriz de confusión LSTM (GLM, Argentina 2021).

VN = 6403	FP = 1526
FN = 85	VP = 902

LSTM: Con un F1-score de 0.53 y un recall de 91 %, la LSTM demostró una excelente capacidad de detección. Su matriz de confusión muestra 902 VP y solo 85 FN, logrando un buen equilibrio. Sin embargo, su precisión es de 0.37, lo que indica que aún tiene un número considerable de FP (1,526). Baseline MLP: Este modelo obtuvo un F1-score de 0.49, ligeramente inferior. Si bien su recall es excepcionalmente alto (99 %, con solo 6 FN), esto viene a un costo muy elevado: 2,046 FP. Esto demuestra que el MLP, al no considerar el contexto temporal, tiende a ser demasiado sensible, clasificando muchos picos de ruido como eventos reales.

En los datos de GLM, la LSTM supera al MLP en la métrica F1, logrando un mejor balance entre la detección de rayos y el control de falsas alarmas.

4.2.4 Modelo LSTM vs. MLP en WWLLN

Tabla 4.13: Métricas de evaluación LSTM (WWLLN, Argentina 2021).

	Precision	Recall	F1-score
Clase 0	0.79	0.92	0.85
Clase 1	0.39	0.17	0.23
Accuracy	0.75		

Tabla 4.14: Matriz de confusión LSTM (WWLLN, Argentina 2021).

VN = 6401	FP = 524
FN = 1661	VP = 330

LSTM: Con un F1-score muy bajo de 0.23, la LSTM tuvo un desempeño pobre en este conjunto de datos. La matriz de confusión muestra 1,661 FN y un recall de solo 17 %, lo que indica que el modelo no pudo encontrar los patrones secuenciales robustos necesarios para detectar la mayoría de los eventos de WWLLN. Baseline MLP: Este modelo alcanzó un F1-score de 0.46, significativamente más alto que la LSTM, impulsado por un recall muy alto (93 %). Sin embargo, este desempeño se basa en una cantidad masiva de 4,242 FP, que prácticamente invalidan su utilidad en un escenario operativo real.

Tabla 4.15: Métricas de evaluación Baseline MLP (WWLLN, Argentina 2021).

	Precision	Recall	F1-score
Clase 0	0.95	0.39	0.55
Clase 1	0.30	0.93	0.46
Accuracy	0.51		

Tabla 4.16: Matriz de confusión Baseline MLP (WWLLN, Argentina 2021).

VN = 6936	FP = 4242
FN = 146	VP = 1846

En los datos de WWLLN, la diferencia de desempeño es dramática. La baja calidad y variabilidad de la señal de esta red expone un claro compromiso: el MLP logra una alta sensibilidad pero con una precisión inaceptable, mientras que la LSTM, al intentar ser más precisa, sacrifica casi por completo la capacidad de detección.

4.3 Resultados Chile

El análisis se realizó en una región altiplánica de 30×30 km centrada en las coordenadas *-17.59* (latitud) y *-69.48* (longitud). El conjunto de datos incluyó registros de descargas eléctricas de las redes GLM y WWLLN, junto con variables meteorológicas obtenidas de NASA POWER. A diferencia de Argentina, en Chile no se dispuso de mediciones de campo eléctrico de alta frecuencia, por lo que la caracterización de las tormentas se basó exclusivamente en variables meteorológicas y en la detección de rayos. La resolución horaria de los datos limitó la captura de patrones de corta duración, obligando a interpretar las tormentas en una escala temporal más amplia.

El dataset final para el año 2022 incluyó 8,747 registros sin rayos y 15 eventos con descargas en el GLM filtrado. Las variables abarcaron temperatura, humedad relativa, presión, precipitación acumulada y componentes de viento (u, v), junto con representaciones cíclicas de la hora y el mes.

4.3.1 Análisis estadístico frente a GLM

La Tabla 4.17 muestra las estadísticas descriptivas para ventanas con y sin rayos detectados por GLM. Se observa un contraste marcado en la humedad relativa: los eventos con rayos presentan una media de 82,3 % frente al 43,5 % en periodos sin actividad eléctrica, lo que sugiere que la saturación de la atmósfera es un precursor clave en la generación de descargas. La baja desviación estándar (5,3 %) en las horas con rayos indica una condición atmosférica muy homogénea durante los eventos, en contraste con la alta dispersión observada en las horas sin rayos.

En términos de viento, la componente v (meridional) muestra un cambio de signo y un incremento en su media (de 0,46 m/s a 1,17 m/s), sugiriendo un patrón dinámico asociado al transporte de humedad. La temperatura no presenta variaciones significativas, lo que evidencia que en esta región altiplánica la actividad eléctrica no está fuertemente condicionada por el gradiente térmico superficial, sino por la disponibilidad de humedad y la inestabilidad atmosférica.

Tabla 4.17: Estadística descriptiva de variables meteorológicas para eventos detectados por GLM en Chile.

Variable	**Sin rayos**			**Con rayos (GLM)**		
	Media	**Desv. Est.**	**Max**	**Media**	**Desv. Est.**	**Max**
Temperatura (°C)	5.82	6.77	21.06	5.52	1.38	7.96
Humedad relativa (%)	43.5	27.9	100.0	82.3	5.3	90.9
Presión (hPa)	61.06	0.12	61.46	61.11	0.05	61.22
Precipitación (mm)	1.44	50.2	2711.8	1.02	2.07	7.88
Viento U (m/s)	-1.56	2.91	7.27	0.28	2.13	3.03
Viento V (m/s)	0.46	2.33	8.99	1.17	2.82	4.82

4.3.2 Análisis estadístico frente a GLM Filtrado

La Tabla 4.18 presenta los resultados tras aplicar un filtrado estricto a los eventos GLM para reducir posibles falsos positivos. Las tendencias se mantienen, con la humedad relativa como la variable más discriminante. La precipitación media en eventos con rayos es baja (1,11 mm) pero con picos puntuales (máx. 11,02 mm), lo que sugiere que las descargas no siempre están asociadas a precipitaciones intensas sino a estructuras convectivas incipientes. La componente u del viento cambia de negativa a positiva, reforzando la idea de convergencia de masas de aire durante los eventos eléctricos.

Tabla 4.18: Estadística descriptiva de variables meteorológicas para eventos detectados por GLM (filtrado) en Chile.

Variable	**Sin rayos**			**Con rayos (GLM Filtrado)**		
	Media	**Desv. Est.**	**Max**	**Media**	**Desv. Est.**	**Max**
Temperatura (°C)	5.82	6.77	21.06	5.61	1.58	8.44
Humedad relativa (%)	43.5	27.9	100.0	79.5	9.1	92.7
Presión (hPa)	61.06	0.12	61.46	61.09	0.06	61.22
Precipitación (mm)	1.44	50.2	2711.8	1.11	2.45	11.02
Viento U (m/s)	-1.56	2.91	7.27	0.57	1.88	3.26
Viento V (m/s)	0.46	2.33	8.99	1.17	2.64	4.91

4.3.3 Análisis estadístico frente a WWLLN

En la Tabla 4.19 se observan patrones similares para WWLLN, con la humedad relativa como el principal indicador. Sin embargo, la desviación estándar de la componente u del viento es más alta, reflejando mayor variabilidad en la dinámica atmosférica. Esto puede deberse a la naturaleza de WWLLN, que tiende a registrar eventos de mayor energía pero menos frecuentes, lo que amplifica la dispersión de las condiciones meteorológicas asociadas.

Tabla 4.19: Estadística descriptiva de variables meteorológicas para eventos detectados por WWLLN en Chile.

Variable	**Sin rayos**			**Con rayos (WWLLN)**		
	Media	**Desv. Est.**	**Max**	**Media**	**Desv. Est.**	**Max**
Temperatura (°C)	5.82	6.77	21.06	5.52	1.38	7.96
Humedad relativa (%)	43.5	27.9	100.0	82.3	5.3	90.9
Presión (hPa)	61.06	0.12	61.46	61.11	0.05	61.22
Precipitación (mm)	1.44	50.2	2711.8	1.02	2.07	7.88
Viento U (m/s)	-1.56	2.91	7.27	0.28	2.13	3.03
Viento V (m/s)	0.46	2.33	8.99	1.17	2.82	4.82

4.3.4 Correlación de variables con eventos GLM y WWLLN

Las Figuras 4.11, 4.12 y 4.13 muestran la correlación de Pearson entre variables y eventos para cada red. Todas las correlaciones son bajas ($|\rho| < 0{,}3$), reforzando la naturaleza no lineal del problema. La humedad relativa es la única variable que presenta una correlación consistente y positiva, mientras que la temperatura mantiene coeficientes cercanos a cero. El viento v aparece como segunda variable relevante, pero con alta dispersión entre redes.

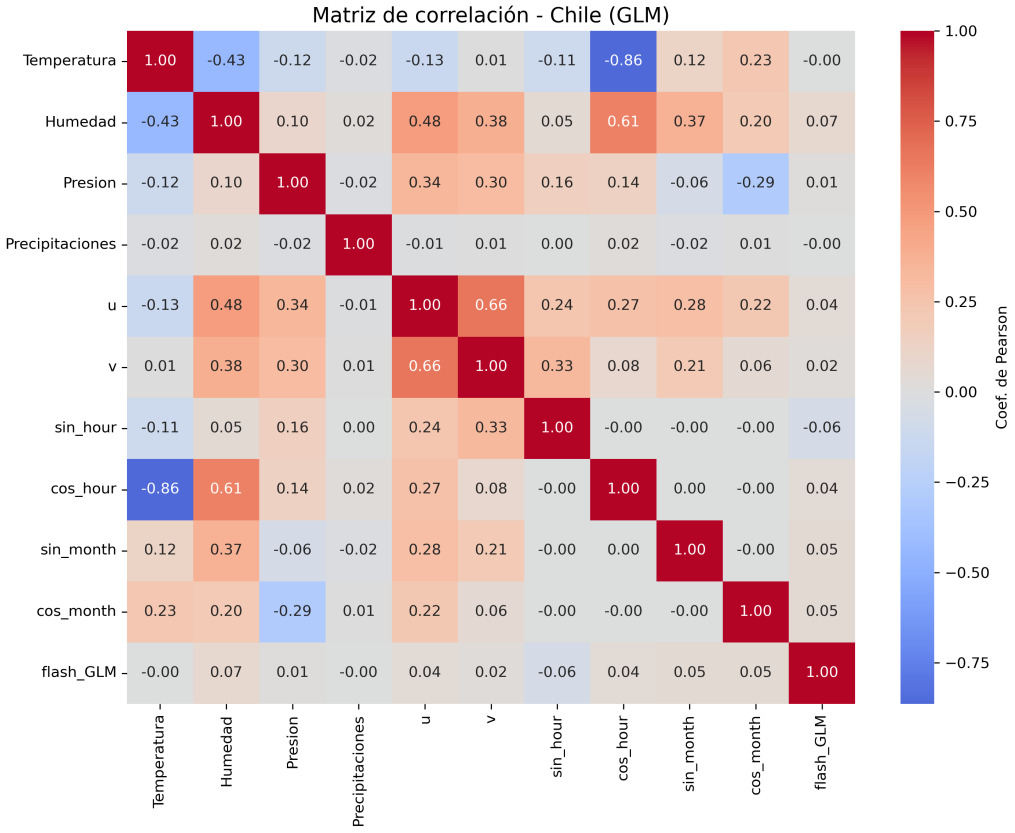

Figura 4.11: Matriz de correlación de Pearson entre variables predictoras y eventos detectados por GLM en Chile.

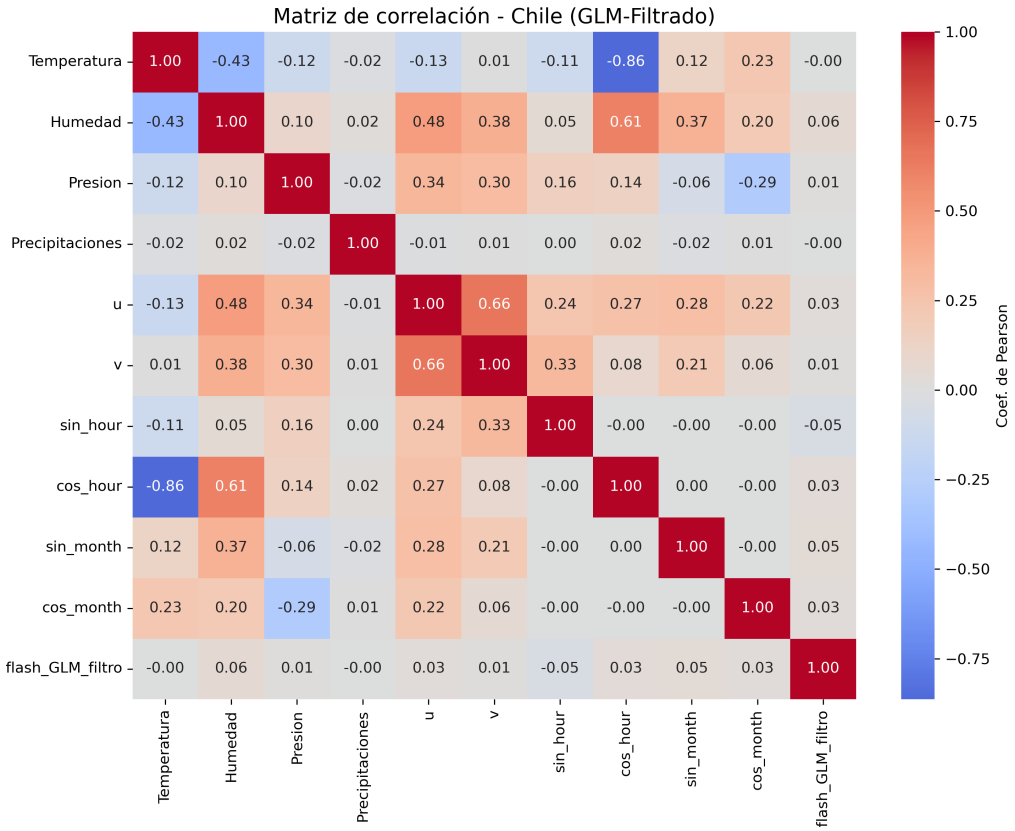

Figura 4.12: Matriz de correlación de Pearson entre variables predictoras y eventos detectados por GLM filtrado en Chile.

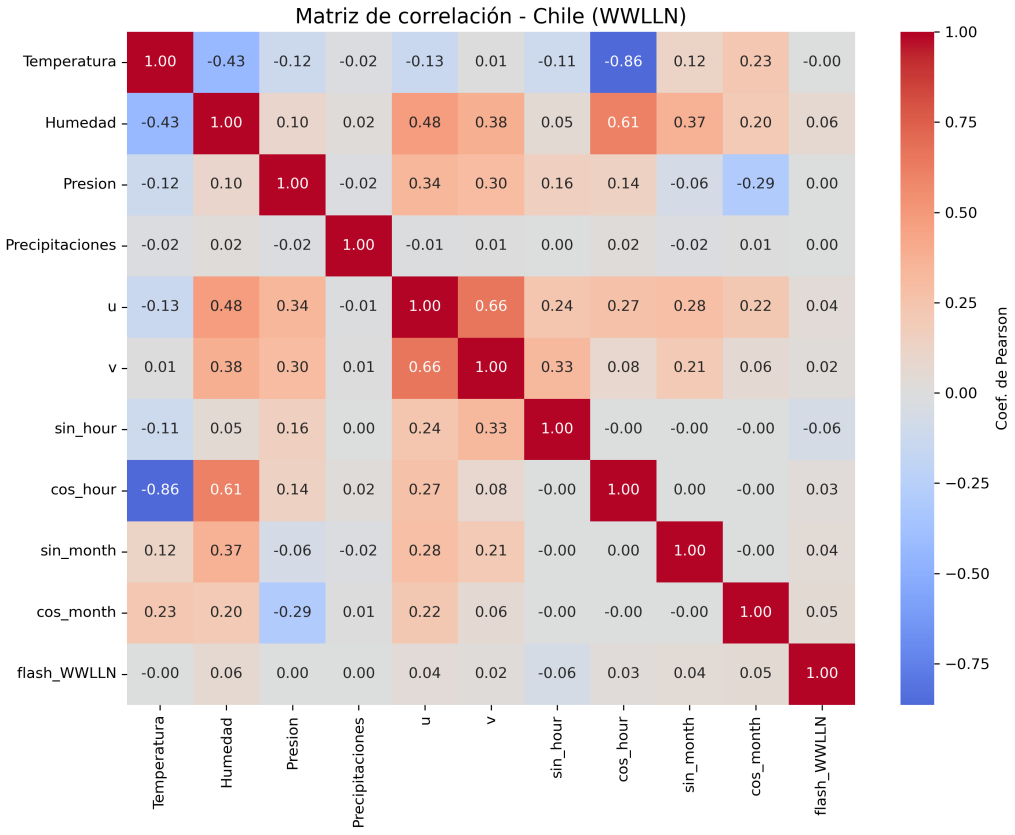

Figura 4.13: Matriz de correlación de Pearson entre variables predictoras y eventos detectados por WWLLN en Chile.

4.3.5 Análisis comparativo Chile

La estadística descriptiva para Chile revela que, en ausencia de mediciones de campo eléctrico, la humedad relativa se convierte en el predictor más robusto para la ocurrencia de rayos, actuando como un proxy indirecto de la inestabilidad eléctrica. La convergencia de viento observada en u y v durante los eventos refuerza la hipótesis de ascensos convectivos localizados como mecanismo disparador. En contraste, la temperatura superficial no muestra un patrón marcado, lo que indica que en regiones altiplánicas el gradiente térmico no es el principal detonante.

El GLM y su versión filtrada presentan patrones consistentes, aunque el filtrado reduce la varianza y depura los eventos espurios, ofreciendo un conjunto más homogéneo. WWLLN, por su parte, presenta mayor dispersión en viento y precipitación, lo que sugiere que los eventos que captura están asociados a tormentas más intensas pero menos frecuentes.

En conjunto, estos resultados confirman que la predicción de descargas en Chile requiere modelos no lineales que combinen humedad, dinámica de viento y variables cíclicas. La baja correlación entre predictores y eventos también anticipa la necesidad de arquitecturas secuenciales como LSTM o GRU,

capaces de aprovechar el contexto temporal en un entorno de datos con resolución horaria.

4.3.6 Resultados de los Modelos Predictivos en Chile

En el caso de Chile, se probaron arquitecturas secuenciales como LSTM y modelos convolucionales 1D utilizando exclusivamente variables meteorológicas. Sin embargo, ninguno de los modelos logró superar un F1-score del 10 %, lo que evidencia que el desempeño fue completamente no exitoso.

Este bajo rendimiento se atribuye principalmente a dos factores. Primero, la variación horaria de las variables meteorológicas muestra una correlación extremadamente baja con la ocurrencia de rayos. Las estadísticas descriptivas y los mapas de correlación refuerzan esta conclusión al mostrar coeficientes de Pearson cercanos a cero entre las variables de entrada y las detecciones tanto de GLM como de WWLLN.

Segundo, el diseño del problema como una predicción binaria de la clase 1 (ocurrencia de rayo) en un horizonte de una hora implica que, si dentro de la ventana se registran múltiples rayos, el modelo solo aprende la presencia de al menos un evento. Esta simplificación impide capturar la intensidad o masividad de las tormentas eléctricas, lo que probablemente afecta la capacidad del modelo para generalizar patrones más complejos asociados a descargas múltiples.

Finalmente, una limitación crítica es la ausencia de la variable de campo eléctrico en este conjunto de datos. Estudios previos han demostrado que el campo eléctrico ambiental es una variable clave para la clasificación cuando se trabaja con arquitecturas secuenciales, ya que captura la dinámica temporal previa a la descarga. Su carencia en Chile reduce drásticamente la información disponible para que los modelos LSTM y Conv1D puedan explotar dependencias temporales, explicando en gran medida la incapacidad de los modelos para ofrecer resultados competitivos.

4.4 Conclusiones

El desarrollo de un sistema predictivo de descargas eléctricas atmosféricas permitió evaluar el desempeño de diferentes arquitecturas de redes neuronales en tres contextos geográficos: Perú, Argentina y Chile. Los resultados obtenidos refuerzan la relevancia de la variable de campo eléctrico como elemento central para la predicción de rayos, y evidencian las limitaciones cuando esta variable está ausente o presenta irregularidades.

4.4.1 Análisis para Perú

El caso de Perú constituye el escenario más exitoso de la investigación. Se utilizó una Red Neuronal Recurrente (RNN) con una arquitectura específica de Memoria a Largo Corto Plazo (LSTM), diseñada para procesar eficazmente las secuencias temporales de los datos. La red final, optimizada para el caso de Perú, se estructuró con dos capas LSTM secuenciales de 64 unidades cada una, seguidas de una capa de normalización por lotes, una capa densa intermedia de 64 neuronas con activación ReLU y una capa de salida con una única neurona sigmoidal para la clasificación binaria. Para la regularización se aplicó un dropout del 20 % y se utilizó el optimizador AdamW. Adicionalmente, se implementó como punto de comparación un modelo base de Perceptrón Multicapa (MLP), compuesto por dos capas densas ocultas de 64 y 32 neuronas. Utilizando datos de alta frecuencia que incluyen mediciones de campo eléctrico ambiental, el modelo LSTM alcanzó un **F1-score máximo de 0.76** en horizontes de predicción de hasta 10 horas, y mantuvo valores superiores a 0.70 para horizontes de corto plazo (1 a 5 horas). Este resultado confirma la capacidad de las redes recurrentes para modelar dinámicas secuenciales asociadas a la formación de descargas nube-tierra.

Un aspecto clave fue la integración de variables derivadas del campo eléctrico: valor medio, valor máximo y cambios de polaridad. Estas características permitieron capturar no solo el nivel de carga en la atmósfera, sino también la inestabilidad previa a la descarga. A diferencia de los picos aislados, los **valores máximos sostenidos** de campo eléctrico mostraron ser los mejores predictores de eventos de rayo, reforzando su importancia en contextos operativos.

El análisis de errores reveló que los falsos negativos (FN) se concentran en ventanas de transición, donde los valores extremos de campo eléctrico no alcanzan a mantenerse durante la secuencia completa, o donde eventos múltiples ocurren fuera de la resolución temporal definida. Este fenómeno fue particularmente evidente en horizontes de 1 hora, donde la predicción binaria utilizada —clasificando cualquier número de rayos como **1**, implica que la ocurrencia masiva de descargas dentro de una ventana no queda diferenciada de un solo evento. Si bien esta formulación simplifica el problema y es útil en términos de alerta temprana, limita la capacidad del modelo para discriminar la severidad de una tormenta.

En resumen, el caso peruano valida que la combinación de variables meteorológicas con *registros de campo eléctrico de alta frecuencia* es crítica para lograr modelos predictivos robustos. La LSTM logró capturar dependencias temporales largas, superando ampliamente a arquitecturas no secuenciales y confirmando que la dinámica previa a una descarga es una secuencia continua más que un evento aislado.

4.4.2 Análisis para Argentina

En Argentina, los resultados estuvieron condicionados por la calidad de los datos, en especial en la red WWLLN. Los modelos entrenados con GLM alcanzaron F1-scores de hasta 0.53, con un recall alto

(91 %) y una cantidad reducida de falsos negativos (FN = 85). Sin embargo, en WWLLN el desempeño cayó drásticamente, con F1-scores cercanos a 0.23 y más de 1,600 FN, reflejando la dificultad de extraer patrones consistentes en presencia de alta varianza y valores extremos no representativos.

La red entrenada con GLM que presentó los mejores resultados se compone de dos capas LSTM apiladas de 64 unidades cada una, ambas regularizadas con dropout (30 %) y dropout recurrente (20 %) para mitigar el sobreajuste. Entre estas capas recurrentes se intercalan capas de Normalización por Lotes (Batch Normalization) para estabilizar y acelerar el entrenamiento. Posteriormente, la red utiliza una capa densa de 64 neuronas con activación ReLU, seguida de otra capa de dropout del 30 %, y finaliza con una capa de salida de una neurona con activación Sigmoid para realizar la clasificación binaria. El modelo se compila con el optimizador AdamW, una tasa de aprendizaje de 0.001 y una función de pérdida binary crossentropy, apoyándose además en el uso de pesos de clase para manejar el desbalance de los datos durante el entrenamiento.

El análisis estadístico mostró que los **valores máximos de campo eléctrico** para WWLLN fueron hasta tres veces mayores que para GLM, lo que sugiere que la red captura picos de alta energía pero también introduce ruido significativo. Esta variabilidad afecta la relación señal-ruido, provocando que los modelos secuenciales como LSTM no encuentren patrones robustos. Los falsos positivos y negativos se concentran en ventanas con gradientes eléctricos intensos que no siempre derivan en descargas, lo que confirma la sensibilidad del modelo a las irregularidades de medición.

4.4.3 Análisis para Chile

El caso de Chile funcionó como un experimento de control crucial que validó de forma contundente la hipótesis central de esta tesis: la criticidad de los datos de campo eléctrico. La ausencia de esta variable demostró ser el factor limitante decisivo para el éxito predictivo.

A pesar de que se probaron arquitecturas LSTM y convolucionales 1D, ningún modelo logró superar un F1-score del 10 %. Este resultado, lejos de ser un fallo de los algoritmos, es una consecuencia directa de la extremadamente baja correlación que presentaron las variables meteorológicas de superficie con la ocurrencia de rayos. Adicionalmente, la resolución horaria de los datos impidió capturar la dinámica de alta frecuencia asociada a las tormentas, limitando severamente la capacidad de las arquitecturas secuenciales para extraer patrones temporales útiles.

En definitiva, el escenario chileno demuestra empíricamente que la predicción de rayos, basada exclusivamente en variables meteorológicas de baja resolución, es una tarea prácticamente inabordable. Este hallazgo subraya el rol insustituible del campo eléctrico ambiental que fue tan efectivo en el caso peruano.

4.4.4 Conclusión principal

Los resultados obtenidos en Perú muestran que la integración del campo eléctrico ambiental es un elemento determinante para anticipar descargas eléctricas. En el mejor escenario, el modelo alcanzó un F1 de 0.76 para un horizonte de 1 hora, con bajas tasas de falsos negativos, mientras que las configuraciones que excluyeron esta variable vieron reducida la métrica en más de 30 %. Este comportamiento confirma que la señal eléctrica previa al rayo concentra la mayor parte de la información útil y que su ausencia no puede ser compensada únicamente con variables meteorológicas.

El análisis de horizontes temporales evidenció que la capacidad predictiva se mantiene alta entre 1 y 5 horas y comienza a degradarse progresivamente a medida que el intervalo de anticipación se amplía, con un descenso más marcado a partir de las 15 horas. Este patrón refleja que la ocurrencia de un rayo responde a una dinámica secuencial de rápida evolución, donde la información crítica se encuentra en ventanas cercanas al evento. En este contexto, las arquitecturas LSTM superaron de forma consistente a modelos estáticos como el MLP, ya que la incorporación de dependencias temporales resultó esencial para discriminar entre ruido y señales reales.

El contraste entre regiones permitió dimensionar el impacto de la calidad y disponibilidad de las variables. En Argentina, la presencia de valores fuera de rango y ruido en los registros de campo eléctrico redujo drásticamente el rendimiento, destacando la sensibilidad de los modelos a la consistencia de los datos de entrada. En Chile, el uso exclusivo de variables meteorológicas no logró superar un F1 del 10 %, mostrando que, si bien estas variables describen el entorno atmosférico, no capturan por sí mismas la dinámica eléctrica que antecede a una descarga.

En conjunto, los hallazgos reafirman que la anticipación de descargas eléctricas requiere integrar señales de campo eléctrico ambiental y abordar el fenómeno como una secuencia temporal más que como un evento aislado. El sistema desarrollado demostró que, en entornos con mediciones confiables, es posible alcanzar métricas altas en horizontes cortos y mantener niveles operativos de desempeño en ventanas de hasta 10 horas, estableciendo una base sólida para futuros esquemas de alerta temprana.

4.5 Líneas de Investigación Futura

A partir de los hallazgos y limitaciones identificados en esta tesis, se proponen las siguientes líneas de investigación para avanzar en la predicción de descargas eléctricas atmosféricas en la región:

- *Evaluación Continua de Nuevos Modelos:* El campo de la inteligencia artificial está en constante evolución. Se recomienda seguir probando y evaluando el desempeño de nuevas arquitecturas de aprendizaje profundo a medida que vayan surgiendo, para buscar mejoras continuas en la precisión

y capacidad de anticipación del sistema.

- *Instalación de un Sensor de Campo Eléctrico:* Dado que la ausencia de esta variable fue el factor más crítico para el desempeño en Chile, se propone como un paso fundamental gestionar la instalación de al menos un sensor de campo eléctrico en una zona de interés nacional. Esto permitiría validar la metodología de esta tesis y sentar las bases para un futuro sistema de alerta.
- *Creación de un Sensor de Bajo Costo:* Para superar la barrera económica que limita el despliegue de una red de monitoreo, se sugiere como línea de trabajo el diseño, construcción y validación de un sensor de campo eléctrico de bajo costo. El desarrollo de un prototipo propio podría facilitar una cobertura más amplia y accesible en el futuro.

Bibliografía

[1] V. Cooray, *The lightning flash.* Institution of Electrical Engineers, 2003.

[2] V. Jimenez, *Desempeño de líneas aéreas de transmisión frente a descargas eléctricas atmosféricas: Análisis de la falla de apantallamiento en terrenos con topografía agreste.* PhD thesis, 2013.

[3] S. J. Goodman, R. J. Blakeslee, W. J. Koshak, D. Mach, J. Bailey, D. Buechler, L. Carey, C. Schultz, M. Bateman, E. McCaul, and G. Stano, "The GOES-R geostationary lightning mapper (GLM)," *Atmospheric Research*, 2013.

[4] P. Bitzer, W. Koshak, and J. Mecikalski, "Classification of glm flashes using random forests," *Earth and Space Science*, 2021.

[5] S. Rosales, "Predicción de días de tormenta dentro del territorio chileno mediante el uso de técnicas de inteligencia artificial," Master's thesis, Universidad Técnica Federico Santa María, 2023.

[6] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning.* 2016.

[7] K. L. Cummins, M. J. Murphy, E. A. Bardo, W. L. Hiscox, R. B. Pyle, and A. E. Pifer, "A review of lightning location systems," *Journal of Atmospheric and Oceanic Technology*, vol. 15, no. 1, pp. 183–190, 1998.

[8] C. Price and D. Rind, "A global lightning parameterization based on cloud top height," *Journal of Geophysical Research: Atmospheres*, vol. 98, no. D5, pp. 9493–9507, 1993.

[9] M. A. Cecchini *et al.*, "Using goes-16 glm data to investigate lightning activity in severe storms," *Atmospheric Research*, vol. 209, pp. 140–149, 2018.

[10] T. G. Chronis *et al.*, "Performance evaluation of the gld360 global lightning detection network," *Journal of Atmospheric and Oceanic Technology*, vol. 33, no. 9, pp. 1899–1911, 2016.

[11] S. Zumaran, "Propuesta de tesis: Desarrollo de un sistema predictivo de descargas eléctricas atmosféricas nube-tierra utilizando técnicas de inteligencia artificial," 2024. Documento interno de postgrado, UTFSM.

[12] M. Zhou, Y. Zhang, and X. Li, "Deep learning models for thunderstorm and lightning nowcasting: A review," *Atmospheric Research*, vol. 243, p. 105001, 2020.

[13] S. Zumaran, "A cloud-to-ground lightning prediction using lstm," *Manuscrito de investigación*, 2025. Enviado para revisión.

[14] X. Zhang, Y. Zhou, and G. Chen, "A 1d-cnn and bilstm-based hybrid model for short-term wind speed forecasting," *Energy Conversion and Management*, vol. 235, p. 113960, 2021.

[15] Z. Haddad, M. El Hajj, and R. Bou Khalil, "Short-term lightning prediction using 1d convolutional neural networks and meteorological time series," *Atmospheric Research*, vol. 268, p. 105973, 2022.

[16] A. Mostajabi, D. L. Finney, M. Rubinstein, and F. Rachidi, "Nowcasting lightning occurrence from commonly available meteorological parameters using machine learning techniques," *NPJ Clim. Atmos. Sci*, vol. 2, p. 41, 2019.

[17] D. Proctor, R. Uytenbogaardt, and B. M. Meredith, "Vhf radio pictures of lightning flashes to ground," *Journal of Geophysical Research*, vol. 93, pp. 12683–12727, 1988.

[18] V. Rakov, M. Uman, and Y. Raizer, *Lightning: Physics and Effects*, vol. 57. 2004.

[19] U. of Washington, "Wwlln." https://wwlln.net/. Accessed: 01/12/2023.

[20] L. Arnold, *Stochastic differential equations: Theory and applications*. Wiley Interscience, 1974.

[21] R. H. Shumway and D. S. Stoffer, *Time series analysis and its applications with r examples (4.a ed.)*. Springer, 2017.

[22] R. J. Athanasopoulos and G. Hyndman, *Forecasting: Principles and practice (2nd ed.)*. 2018.

[23] T. M. Mitchell, *Machine learning (1.a ed.)*. McGraw-Hill, 1997.

[24] R. Kruse, C. Borgelt, C. Braune, S. Mostaghim, and M. Steinbrecher, *Introduction to neural networks*. 2016.

[25] W. S. McCulloch and W. H. Pitts, "A logical calculus of the ideas immanent in nervous activity," *Bulletin of Mathematical Biology*, vol. 52, pp. 99–115, 1943.

[26] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[27] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, no. 6088, pp. 533–536, 1986.

[28] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in Neural Information Processing Systems (NIPS)*, 2012.

[29] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[30] R. Collobert, J. Weston, L. Bottou, *et al.*, "Natural language processing (almost) from scratch," *Journal of Machine Learning Research*, vol. 12, pp. 2493–2537, 2011.

[31] X. He, L. Liao, H. Zhang, *et al.*, "Neural collaborative filtering," *Proceedings of the 26th International Conference on World Wide Web (WWW)*, 2017.

[32] O. Barkan and N. Koenigstein, "Item2vec: Neural item embedding for collaborative filtering," *Proceedings of the 10th ACM Conference on Recommender Systems*, 2016.

[33] M. A. Nielsen, *Neural Networks and Deep Learning.* 2015.

[34] J. L. Elman, "Finding structure in time," *Cognitive Science*, vol. 14, no. 2, pp. 179–211, 1990.

[35] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," *Advances in Neural Information Processing Systems (NIPS)*, 2014.

[36] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[37] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, "Attention is all you need," *Advances in Neural Information Processing Systems (NIPS)*, 2017.

[38] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[39] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," *International Conference on Medical image computing and computer-assisted intervention*, 2015.

[40] G. Litjens, T. Kooi, B. E. Bejnordi, *et al.*, "A survey on deep learning in medical image analysis," *Medical Image Analysis*, 2017.

[41] A. Esteva, B. Kuprel, R. Novoa, *et al.*, "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, 2017.

[42] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[43] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?," *Advances in Neural Information Processing Systems (NIPS)*, 2014.

[44] H. C. Shin, H. R. Roth, M. Gao, *et al.*, "Deep convolutional neural networks for computer-aided detection: Cnn architectures, dataset characteristics, and transfer learning," *IEEE Transactions on Medical Imaging*, 2016.

[45] R. A. Willoughby, "Solutions of ill-posed problems (a. n. tikhonov and v. y. arsenin)," *SIAM Review*, vol. 21, no. 2, pp. 266–267, 1979.

[46] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," *Proceedings of the 14th international joint conference on Artificial intelligence (IJCAI)*, 1995.

[47] H. Robbins and S. Monro, "A stochastic approximation method," *Annals of Mathematical Statistics*, vol. 22, no. 3, pp. 400–407, 1951.

[48] L. Bottou, F. E. Curtis, and J. Nocedal, "Optimization methods for large-scale machine learning," *arXiv preprint arXiv:1606.04838*, 2016.

[49] S. Ruder, "An overview of gradient descent optimization algorithms," *arXiv preprint arXiv:1609.04747*, 2016.

[50] C. C. Aggarwal, *Data classification: Algorithms and applications*. 2014.

[51] D. S. Wilks, *Statistical methods in the atmospheric sciences: An introduction*. 1995.

[52] S. R. Adhikari and D. V. B. Rao, "Thunderstorm forecasting by using artificial neural network," *International Journal of Computer Applications*, vol. 23, no. 6, pp. 41–45, 2011.

[53] W. A. T. W. Abdullah *et al.*, "Lightning forecasting modelling using ann," *Unknown Journal*, 2018.

[54] M. A. Ramzi *et al.*, "Lightning prediction modelling using mlpnn structure," *Unknown Journal*, 2018.

[55] M. A. Al-Jundi *et al.*, "Optimized ann-abc for thunderstorms prediction," *Unknown Journal*, 2017.

[56] B. Mohan and S. Arumugam, "Soft computing and data mining techniques for thunderstorms and lightning prediction," *Unknown Journal*, 2017.

[57] M. Mostajabi, E. Taghizadeh, and S. Z. Hosseini, "Nowcasting lightning occurrence from commonly available meteorological parameters," *Atmospheric Research*, vol. 218, pp. 152–162, 2019.

[58] D. R. Pereira, R. Pacheco, and E. M. de Souza, "Application of artificial intelligence techniques in the forecast of cloud-to-ground lightning: A case study in southern brazil," *Atmospheric Research*, vol. 268, p. 105962, 2022.

[59] M. A. Abdullah, F. Yusof, and S. A. Salam, "An application of deep learning for lightning prediction in east coast malaysia," *Results in Engineering*, vol. 18, p. 100833, 2023.

[60] G. Ehrensperger, T. Simon, G. J. Mayr, and T. Hell, "Identifying lightning processes in era5 soundings with deep learning," *Geoscientific Model Development*, vol. 18, pp. 1141–1153, 2025.

[61] M. Cheon, "Mjölnir: A deep learning parametrization framework for global lightning flash density," *arXiv preprint arXiv:2504.19822*, 2025.

[62] M. A. Ferro, O. Pinto Jr, and M. Pinhatti, "Lightning risk warnings based on atmospheric electric field measurements in brazil," *Atmospheric Research*, vol. 100, no. 4, pp. 377–387, 2011.

[63] D. Ariza *et al.*, "Behavior of corona current and atmospheric variables under thunderstorm conditions," *Unknown Journal*, 2012.

[64] X. Bao, L. Zhang, S. Chen, and J. Li, "Lightning prediction based on atmospheric electric field measurements using sparse autoencoder and resnet," *Remote Sensing*, vol. 14, no. 17, p. 4131, 2022.

[65] D. Yang *et al.*, "Multifeature fusion-based thunderstorm prediction system with switchable patterns," *Unknown Journal*, 2023.

[66] D. E. Hill, Y. Zhang, and C. Gordon, "Evaluating the efficacy of electric field mills to predict lightning events," white paper, Vaisala, 2022.

[67] D. Yang *et al.*, "3daefa-based thunderstorm prediction system with higher performance," *Unknown Journal*, 2022.

[68] M. Fukawa, X. Deng, S. Imai, T. Horiguchi, R. Ono, I. Rachi, and T. Kudo, "A novel method for lightning prediction by direct electric field measurements at the ground using recurrent neural network," *IEICE Transactions on Information and Systems*, vol. 105, no. 6, pp. 1624–1628, 2022.

[69] M. Mansouri *et al.*, "Lightning nowcasting using solely lightning data," *Unknown Journal*, 2023.

[70] M. L. Hutchins *et al.*, "Far-field power of lightning strokes as measured by the world wide lightning location network," *Unknown Journal*, 2012.

[71] S. Chen, Y. Liu, and X. Wang, "Severe thunderstorm nowcasting using ensemble deep learning," *Scientific Reports*, vol. 12, p. 19421, 2022.

[72] S. Ghosh *et al.*, "A scheme for local lightning detection and prediction system," *Unknown Journal*, 2020.

[73] D. Espinoza, J. Ronchail, and W. Lavado, "Forecasting convective activity in the amazon basin using satellite and ann techniques," *Journal of Applied Meteorology and Climatology*, vol. 60, no. 4, pp. 433–448, 2021.

[74] F. Li, J. Li, and J. Li, "Nowcasting of cloud-to-ground lightning location and frequency based on a deep learning technique," *Atmospheric and Oceanic Science Letters*, vol. 18, no. 1, p. 100607, 2025.

[75] K. Essa, J. G. Botha, and F. A. Engelbrecht, "Deep learning prediction of thunderstorm severity using remote sensing weather data," *Journal of Atmospheric and Solar-Terrestrial Physics*, vol. 235, p. 105841, 2022.

[76] H. Wang *et al.*, "Thunderstorm prediction method based on cnn-bilstm using beads," *Unknown Journal*, 2021.

[77] V. Montiel, J. Camacho, and M. Navarro, "Research on lightning prediction based on gcn-lstm model," *Atmosphere*, vol. 16, no. 4, p. 447, 2023.

[78] J. Tang, L. Zhou, and J. Xu, "Mcgln: A multimodal convlstm-gan framework for lightning nowcasting," *Information Fusion*, vol. 99, p. 102981, 2024.

[79] S. Pakdaman, S. Nazari, and M. Sadeghi, "Lightning prediction using an ensemble learning approach," *Environmental Monitoring and Assessment*, vol. 192, pp. 1–15, 2020.

[80] S. S. Patil *et al.*, "Thunderstorm prediction model using smote sampling and machine learning approach," *Unknown Journal*, 2023.

[81] A. Borah and N. Hazarika, "An explainable machine learning technique to forecast lightning densities," *Information Fusion*, vol. 93, p. 102678, 2024.

[82] Y. Li, Y. Wang, and X. Chen, "A physics-based machine learning model for lightning prediction," *Frontiers in Earth Science*, vol. 12, p. 1376605, 2024.

[83] Y. Tan, M. Li, and H. Chen, "Flashbench: A lightning nowcasting framework based on hybrid deep learning and physics-based dynamical models." arXiv preprint arXiv:2305.10064, 2024.

[84] O. Troyanskaya *et al.*, "Missing value estimation methods for dna microarrays," *Bioinformatics*, vol. 17, no. 6, pp. 520–525, 2001. Duplicada como 'troyanskaya2001'.

[85] G. E. A. P. A. Batista and M. C. Monard, "A study of k-nearest neighbour as a model-based method to treat missing data," pp. 327–331, 2002.

[86] M. Lindholm, B. Arntsen, and A. Løvlund, "Decomposing wind speed and direction into cartesian components for meteorological modeling," *Atmospheric Science Letters*, vol. 23, no. 6, p. e1078, 2022.

[87] Widodo, Brawijaya, and Samudi, "K-fold cross validation for imbalanced datasets in machine learning classification," *International Journal of Computer Applications*, vol. 183, no. 25, pp. 1–7, 2022.

[88] J. Giros and Poggio, *Regularization and Model Selection in Machine Learning.* MIT Press, 1995.

Análisis de Distribución de Variables mediante Boxplots y Violinplots

Con el objetivo de evaluar la dispersión y comportamiento de las variables empleadas en el modelo, se generaron diagramas combinados de *boxplot* y *violinplot* para cada característica. Estos gráficos permiten comparar la distribución de los datos entre eventos de rayos (`flash` = 1) y no-rayos (`flash` = 0), así como analizar patrones horarios y estacionales que influyen en la ocurrencia de descargas eléctricas.

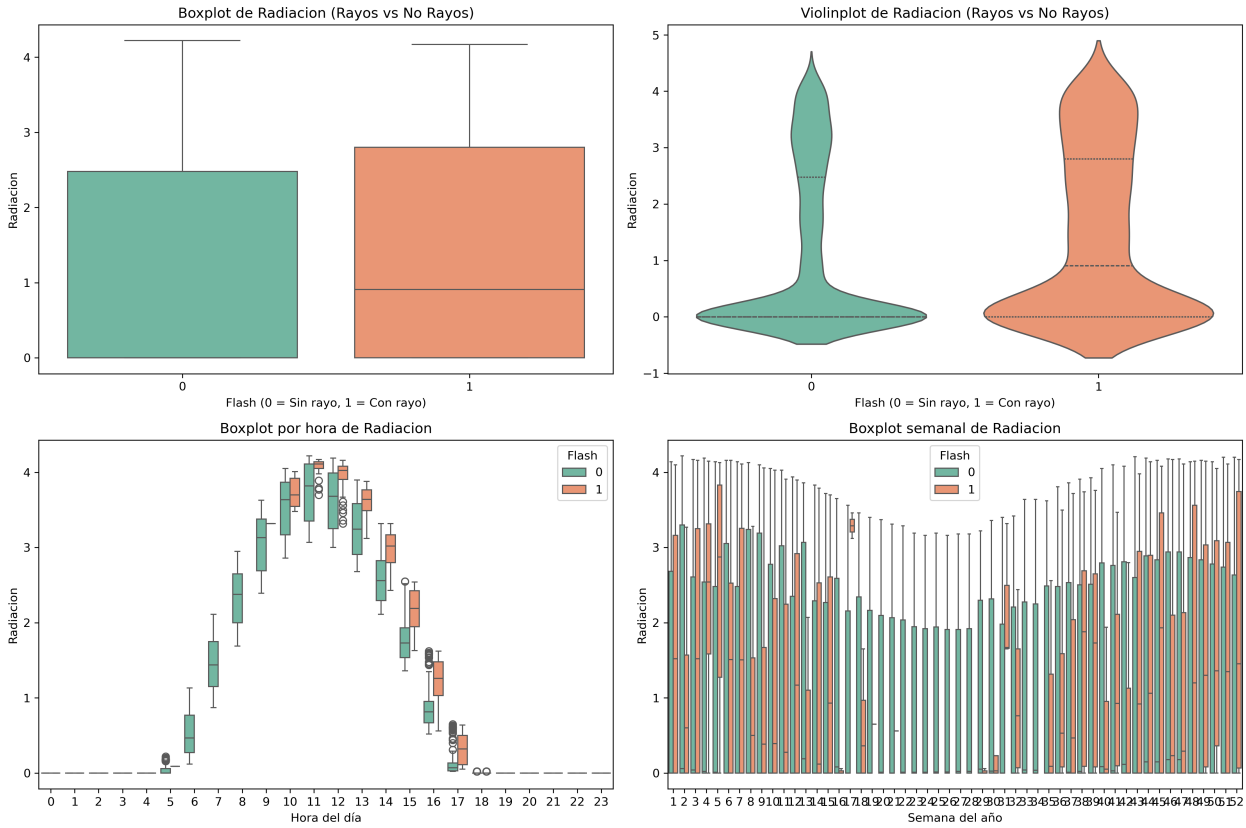

Figura 14: Distribución de la variable `Radiacion` diferenciando eventos de rayo y no-rayo.

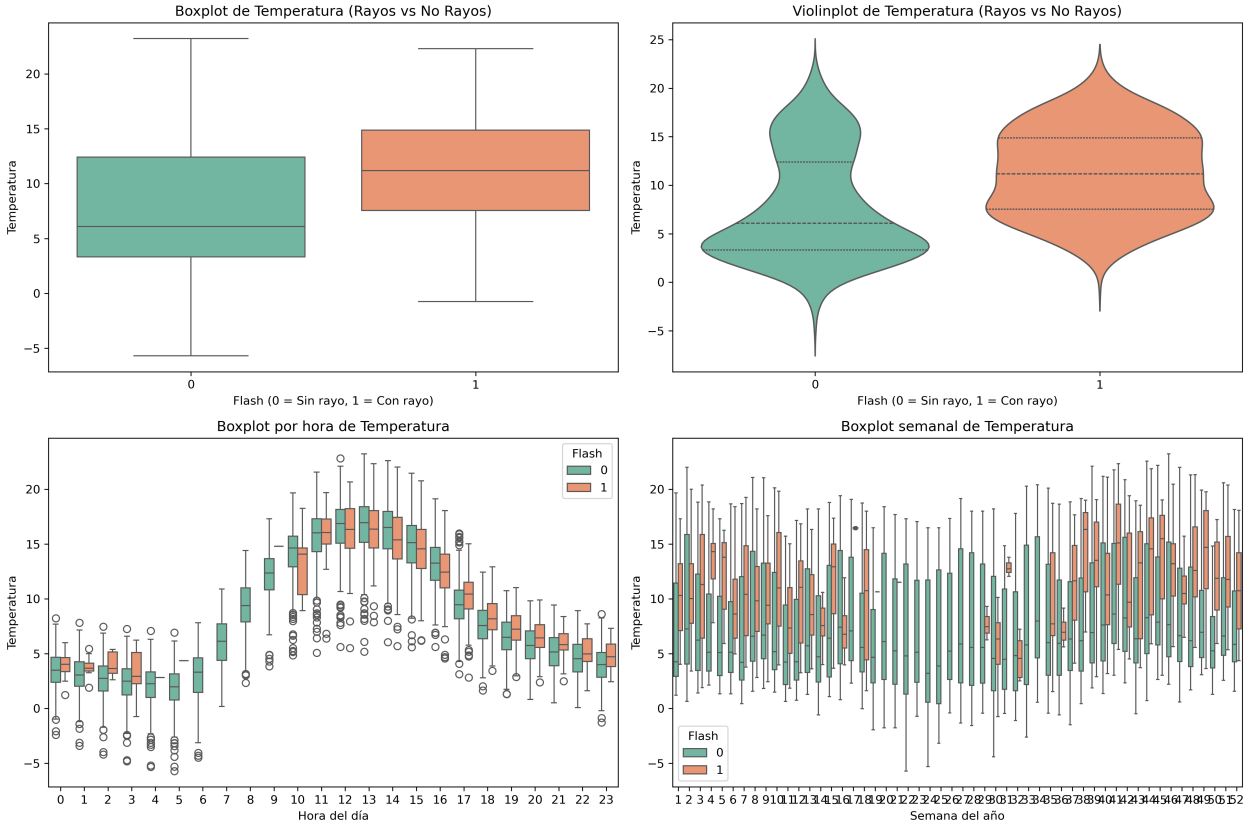

Figura 15: Distribución de la variable `Temperatura` diferenciando eventos de rayo y no-rayo.

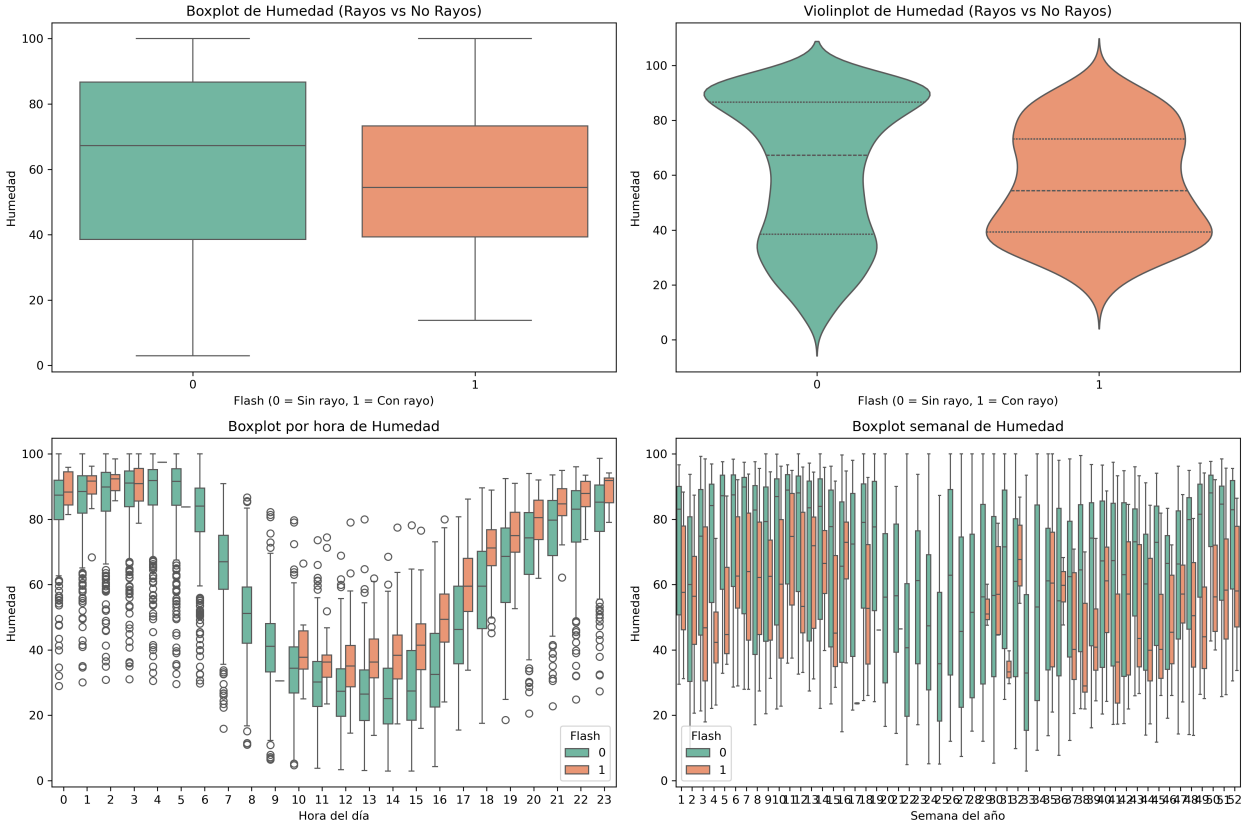

Figura 16: Distribución de la variable `Humedad` diferenciando eventos de rayo y no-rayo.

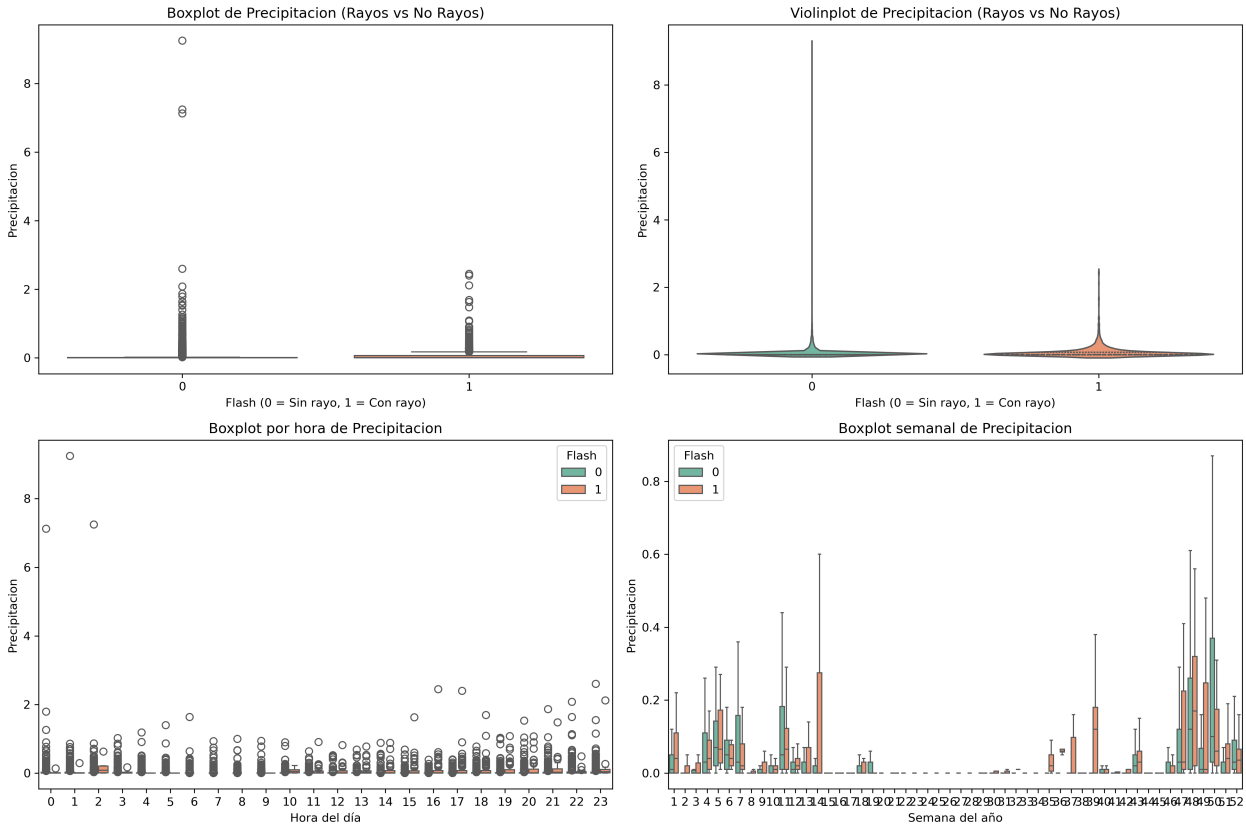

Figura 17: Distribución de la variable `Precipitacion` diferenciando eventos de rayo y no-rayo.

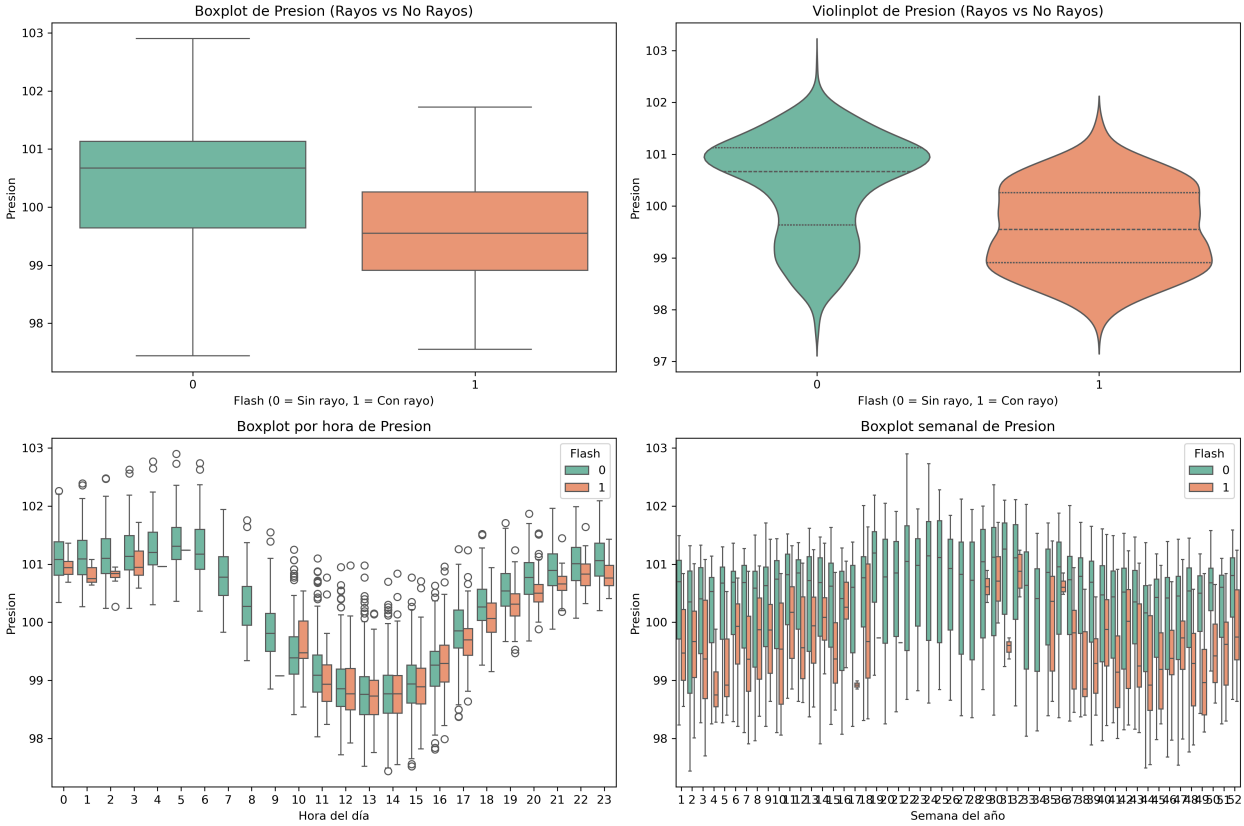

Figura 18: Distribución de la variable `Presion` diferenciando eventos de rayo y no-rayo.

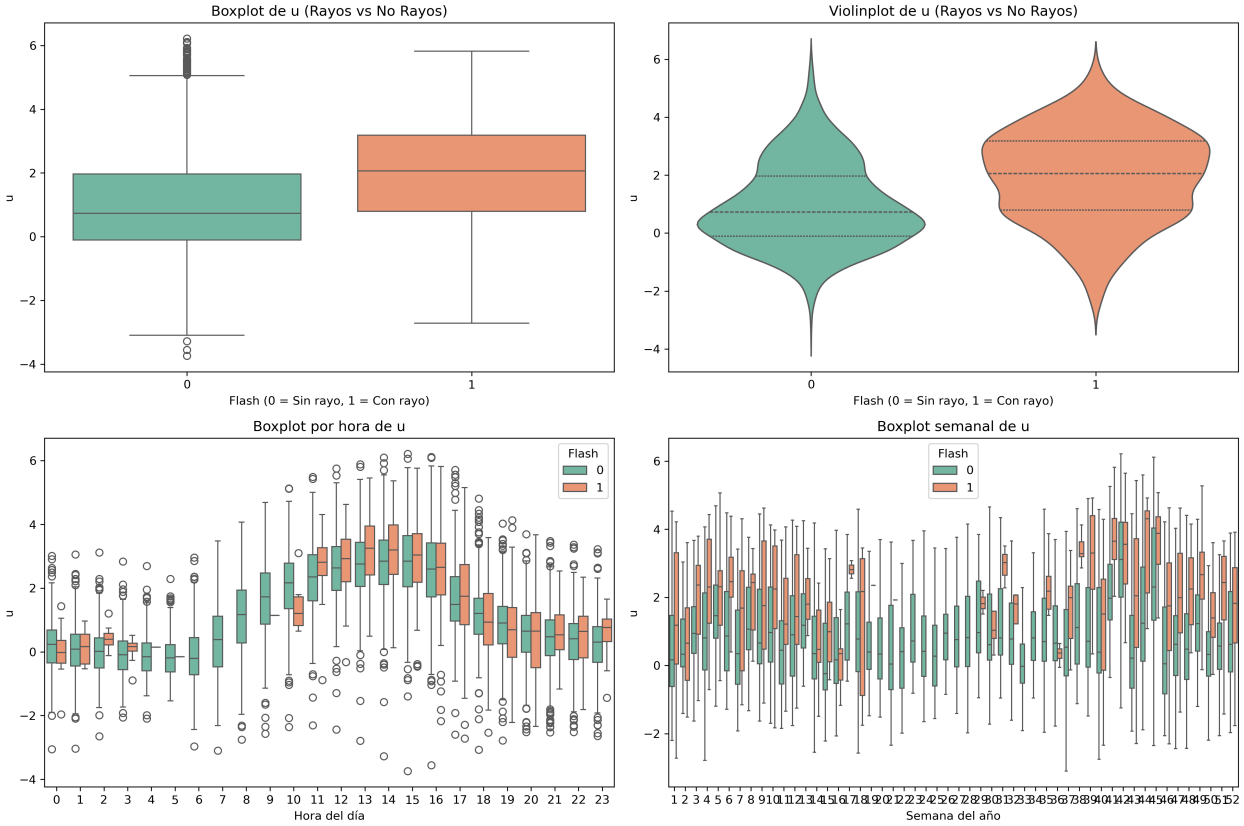

Figura 19: Distribución de la componente de viento u diferenciando eventos de rayo y no-rayo.

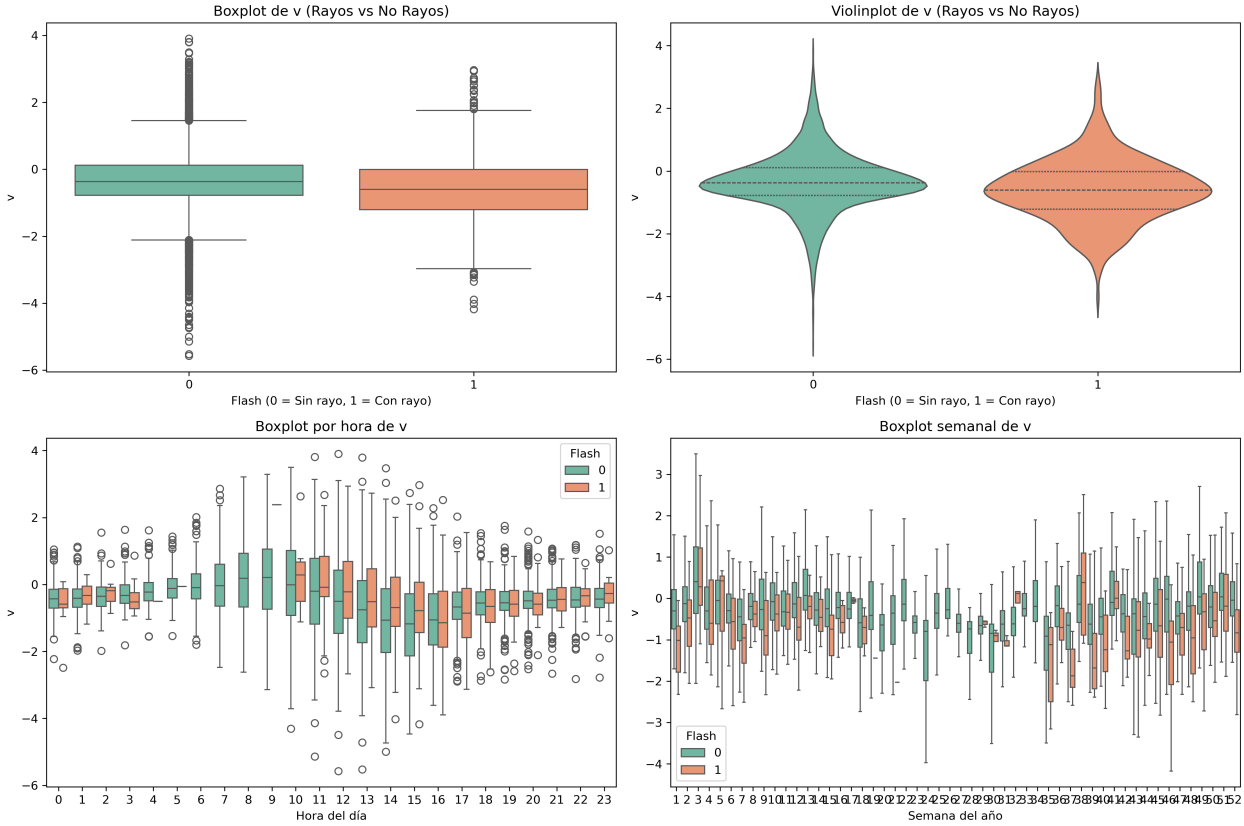

Figura 20: Distribución de la componente de viento v diferenciando eventos de rayo y no-rayo.

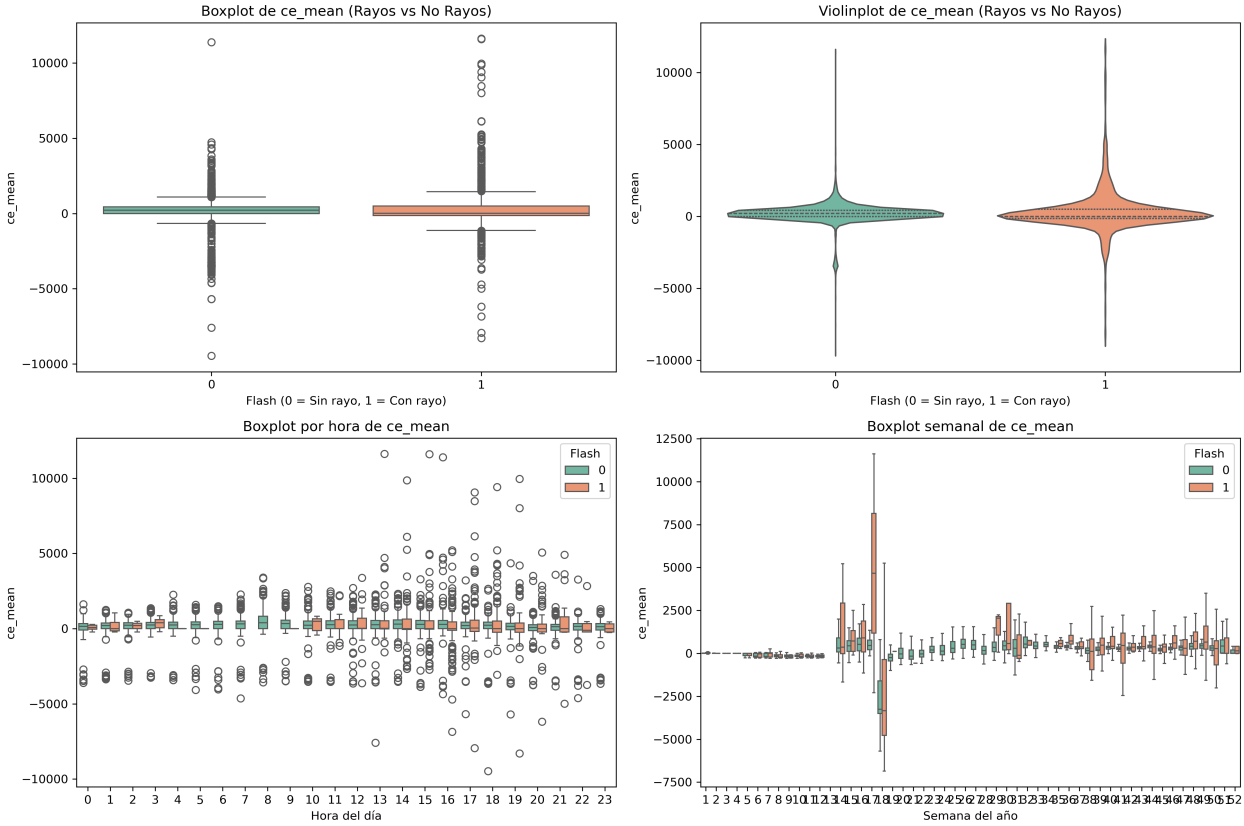

Figura 21: Distribución del `campo eléctrico medio (ce_mean)` en eventos de rayo y no-rayo.

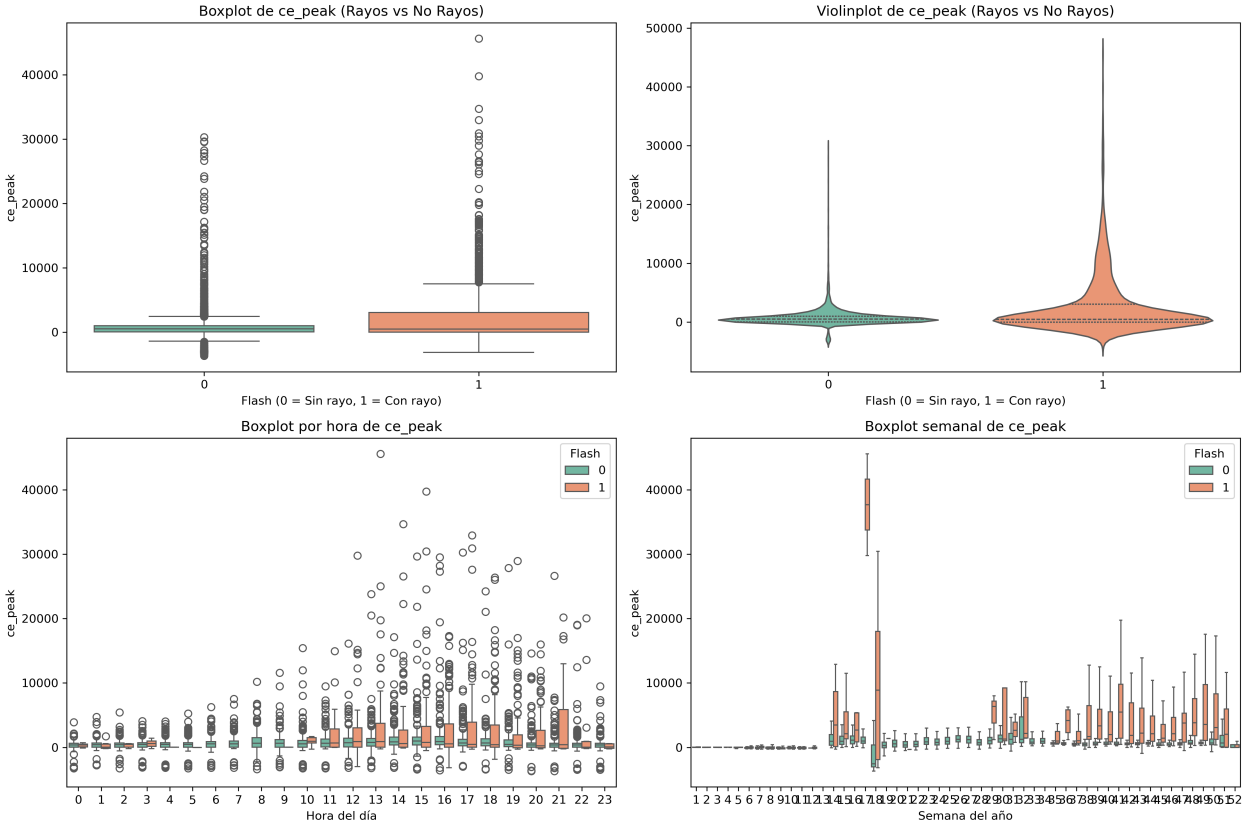

Figura 22: Distribución del `campo eléctrico pico (ce_peak)` en eventos de rayo y no-rayo.

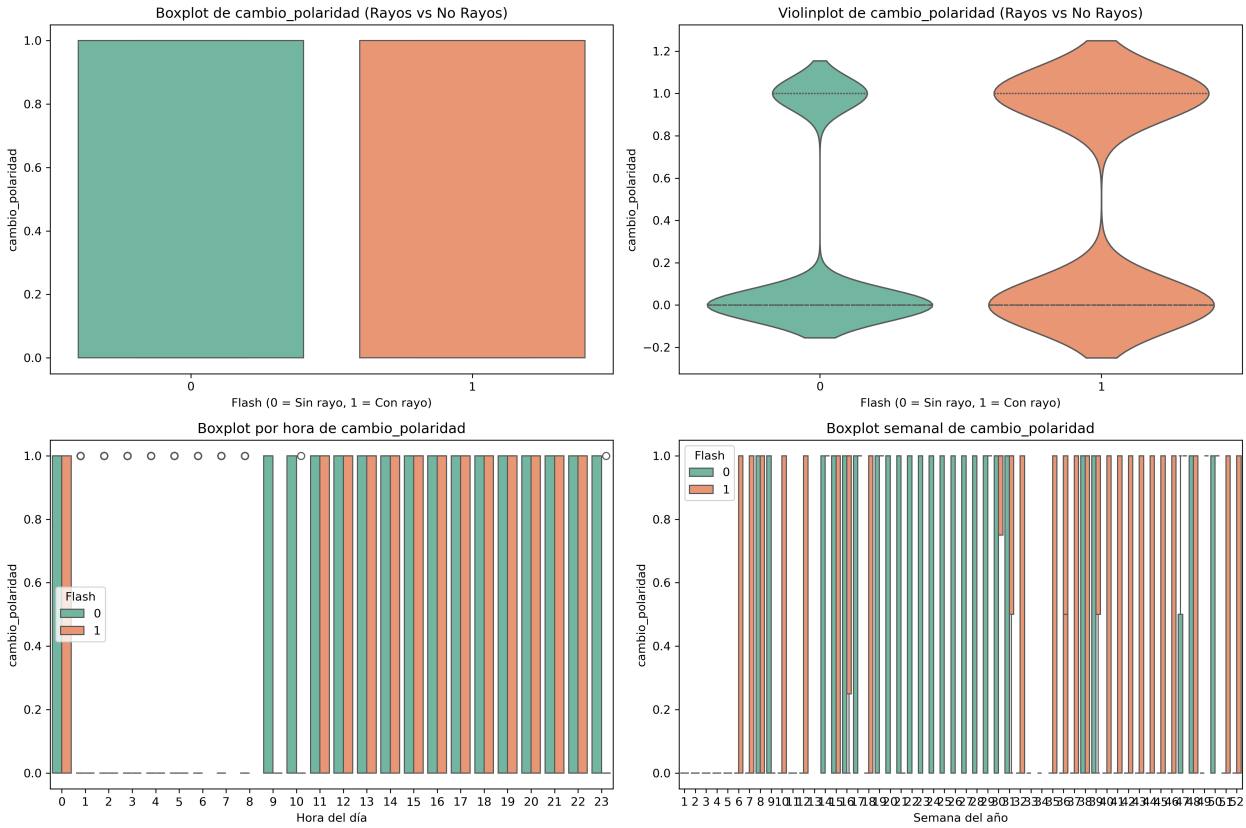

Figura 23: Distribución de la variable categórica `cambio_polaridad` en eventos de rayo y no-rayo.

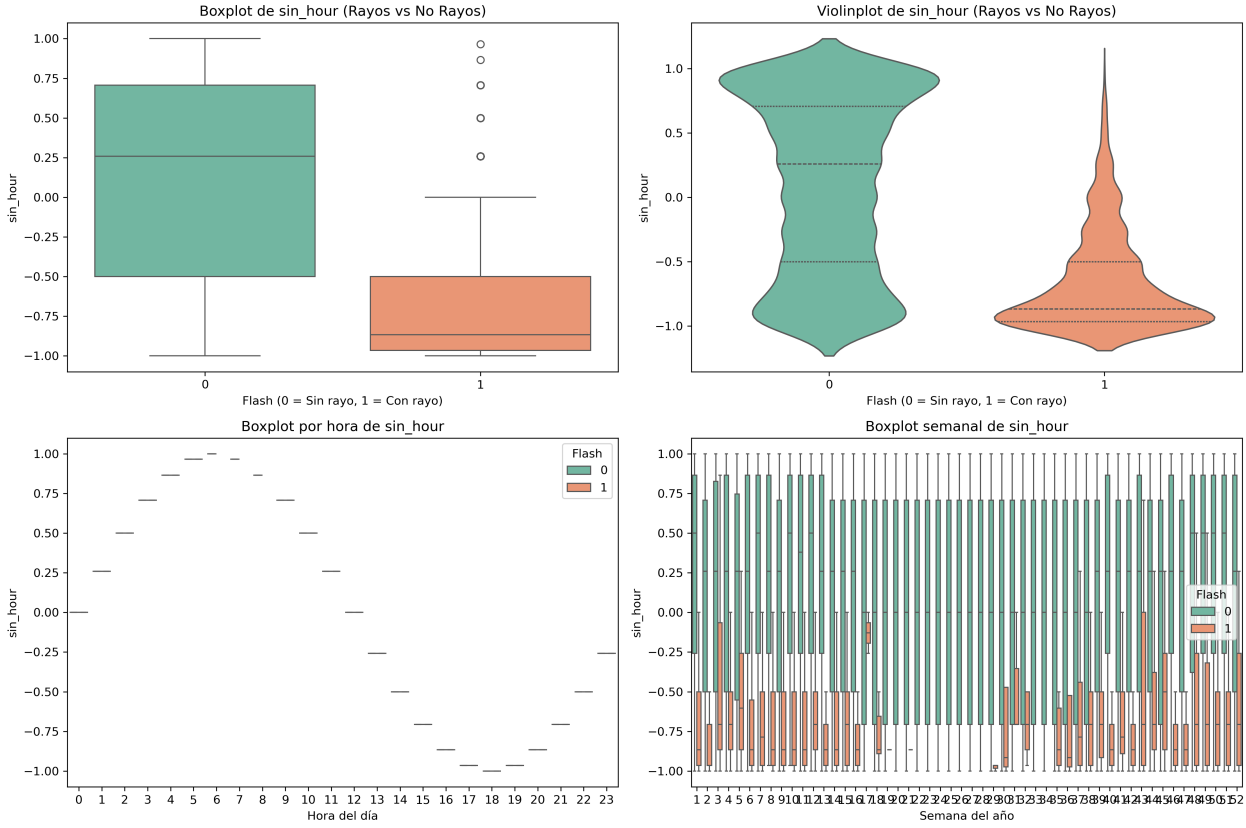

Figura 24: Distribución de la variable cíclica `sin_hour` diferenciando eventos de rayo y no-rayo.

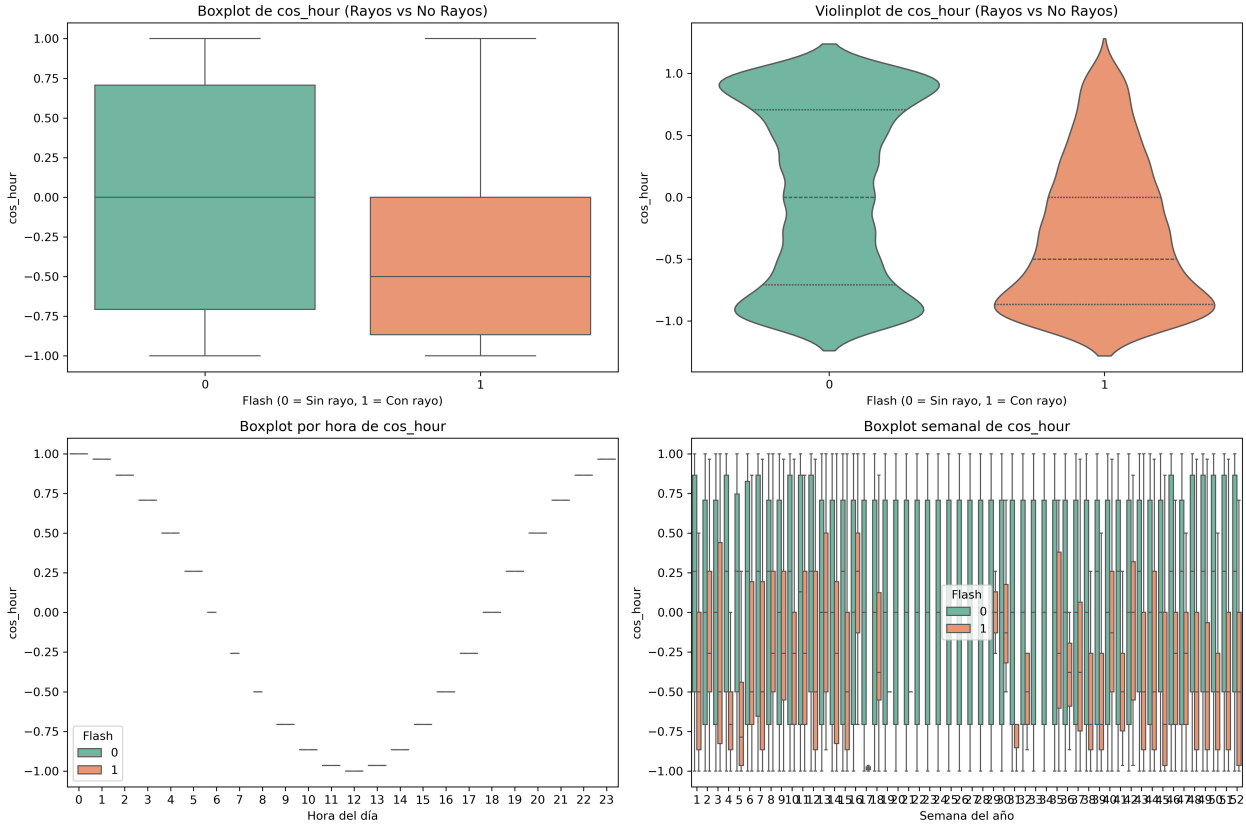

Figura 25: Distribución de la variable cíclica `cos_hour` diferenciando eventos de rayo y no-rayo.

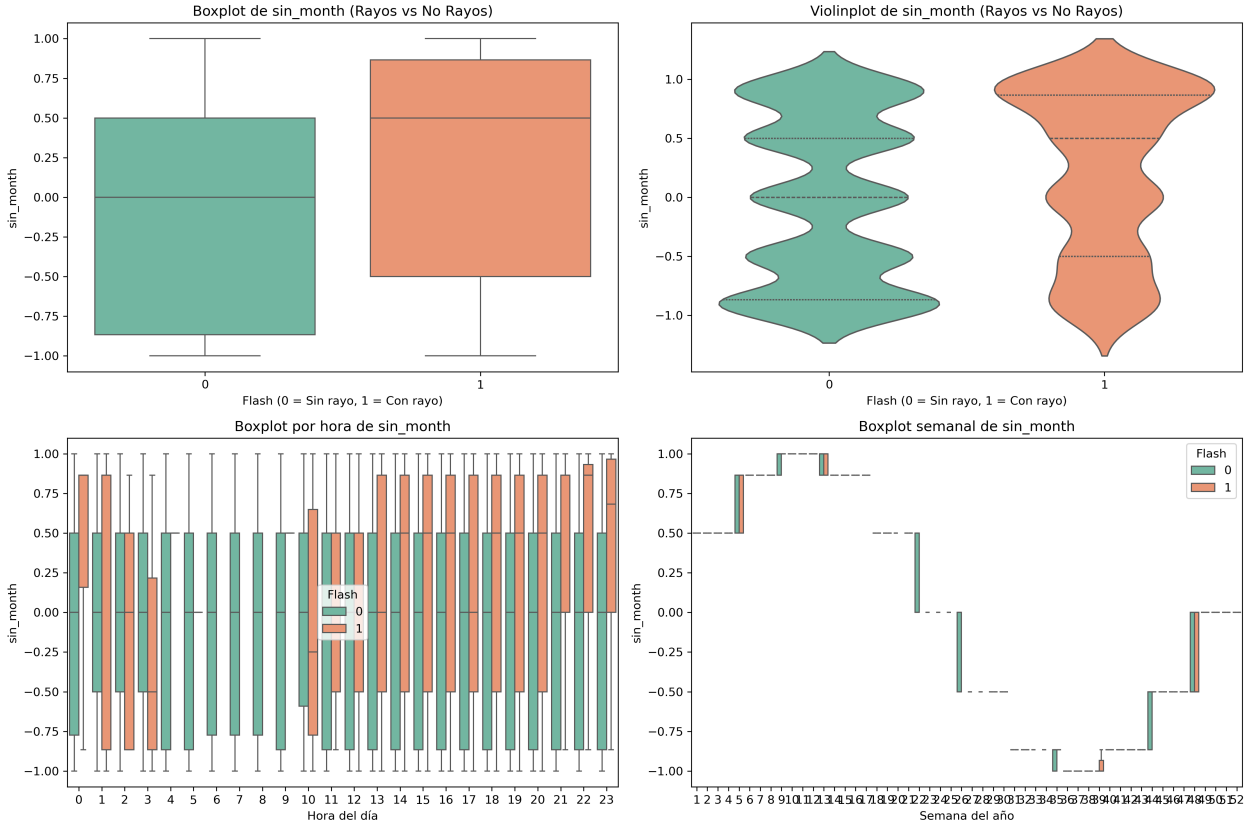

Figura 26: Distribución de la variable cíclica `sin_month` diferenciando eventos de rayo y no-rayo.

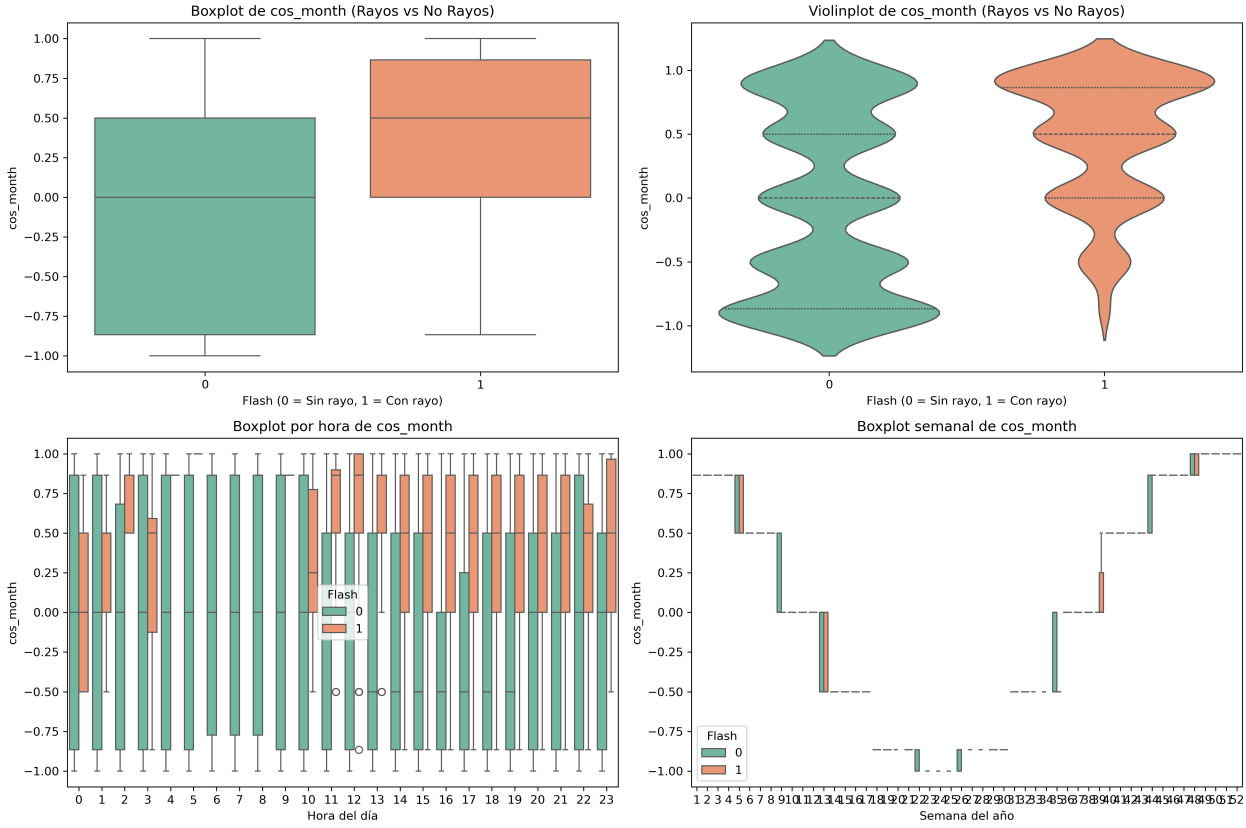

Figura 27: Distribución de la variable cíclica `cos_month` diferenciando eventos de rayo y no-rayo.

El análisis conjunto de los diagramas muestra que variables como `ce_mean`, `ce_peak`, `Humedad` y `Precipitacion` exhiben claras diferencias entre escenarios de rayos y no-rayos, confirmando su relevancia en el proceso predictivo. En contraste, algunas variables cíclicas muestran una contribución más moderada, aunque permiten capturar patrones horarios y estacionales útiles para la red neuronal.