Informe Final del Proyecto: Modelado Predictivo de la Exposición a Campos Electromagnéticos en Bogotá, Colombia

Curso: Teoría de Aprendizaje de Máquina
Profesor: Andrés Marino Álvarez Meza, PhD
Institución: Departamento de Ingeniería Eléctrica, Electrónica y Computación, Universidad Nacional de Colombia - Sede Manizales
Fecha: 23 de julio de 2025

1. Motivación del Proyecto

El crecimiento acelerado de la infraestructura de telecomunicaciones en áreas urbanas, impulsado por la expansión de redes móviles, sistemas Wi-Fi y otros dispositivos inalámbricos, ha incrementado significativamente la exposición de la población a campos electromagnéticos (CEM). En Bogotá, Colombia, una ciudad con una densidad poblacional superior a 7,000 habitantes por km² en sus áreas centrales, la alta concentración de estaciones base y dispositivos inalámbricos genera preocupaciones sobre posibles riesgos para la salud. Estas inquietudes se ven respaldadas por las directrices de la Comisión Internacional de Protección contra la Radiación No Ionizante (ICNIRP), que establecen límites de seguridad (por ejemplo, 2 V/m para exposición pública a radiofrecuencias), y estudios emergentes que asocian la exposición prolongada a CEM de baja intensidad con síntomas como fatiga, dolores de cabeza y posibles riesgos carcinogénicos a largo plazo, clasificados por la Agencia Internacional para la Investigación del Cáncer (IARC) como "posiblemente carcinogénicos" (Grupo 2B).

La motivación de este proyecto surge de la necesidad de abordar estas preocupaciones de salud y medio ambiente de manera basada en datos. Las características topográficas únicas de Bogotá, ubicada a 2,640 metros sobre el nivel del mar, junto con su densa infraestructura urbana, generan patrones complejos de propagación de CEM, lo que requiere herramientas predictivas localizadas. La aplicación de técnicas de aprendizaje de máquina (ML) y aprendizaje profundo (DL), como redes neuronales recurrentes (RNN), es relevante debido a su capacidad para modelar dependencias temporales en datos de series temporales, esenciales para predecir la exposición a CEM basada en mediciones históricas. Este proyecto está justificado técnicamente por la capacidad de ML/DL para manejar datos de alta dimensionalidad y no lineales, es socialmente relevante por su potencial para informar políticas de salud pública, y es científicamente significativo al contribuir al limitado cuerpo de estudios localizados sobre CEM en América Latina.

2. Planteamiento del Problema

El problema específico que aborda este proyecto es la predicción de los niveles de exposición a campos electromagnéticos (CEM) en Bogotá, Colombia, utilizando datos de series temporales recopilados por antenas de monitoreo entre 2022 y 2024. El conjunto de datos, derivado de mediciones tomadas cada 6 minutos por antenas ubicadas en diversas coordenadas de Bogotá, ha sido preprocesado para proporcionar valores promedio diarios de exposición a CEM por antena. Sin embargo, no todas las antenas estuvieron activas durante todo el período de 2015 a 2024, lo que genera datos faltantes y requiere estrategias robustas de preprocesamiento e imputación.

La pregunta de investigación es: **¿Cómo pueden los modelos de aprendizaje de máquina, específicamente redes neuronales recurrentes, predecir con precisión los niveles futuros de exposición a CEM en Bogotá basándose en mediciones promedio diarias de 2022 a 2024, considerando variaciones espaciales y temporales entre las antenas de monitoreo?**

El dominio de la solución es un problema de predicción de series temporales, donde el conjunto de datos incluye columnas como `Fecha` (fecha), `Antena` (identificador de antena), `Coordenadas_X`, `Coordenadas_Y` y `CEM_Promedio_Diario` (exposición promedio diaria a CEM). El objetivo es desarrollar un modelo que pronostique los niveles de exposición a CEM para meses futuros, permitiendo la identificación de áreas de alto riesgo y apoyando la planificación urbana y las intervenciones de salud pública en Bogotá.

3. Estado del Arte

El estudio de la exposición a CEM en entornos urbanos ha ganado relevancia a nivel global, pero la investigación localizada en América Latina, particularmente en Colombia, sigue siendo limitada. Una revisión sistemática realizada utilizando la metodología PRISMA, como se detalla en el artículo proporcionado, identificó estudios clave que informan este proyecto. De un total inicial de 123 artículos obtenidos de bases de datos como SCOPUS y SciELO, se seleccionaron siete estudios relevantes tras filtrar por su pertinencia a la exposición a CEM, efectos en la salud y modelado predictivo en entornos urbanos. Estos estudios aportan información crítica en las siguientes áreas:

1. **Niveles de Exposición a CEM**: Las investigaciones indican una correlación significativa entre la proximidad a estaciones base móviles y niveles de CEM superiores a 2 V/m, el límite de exposición pública de la ICNIRP. Por ejemplo, un estudio en Medellín destacó niveles elevados de CEM en áreas densamente pobladas, un hallazgo relevante para el contexto urbano similar de Bogotá.

2. **Efectos en la Salud**: La literatura reporta efectos térmicos en exposiciones de alta intensidad (>10 W/m²) y efectos no térmicos, como fatiga y trastornos del sueño, en exposiciones de baja intensidad (0.1–2 V/m). La clasificación de los CEM de radiofrecuencia como "posiblemente carcinogénicos" por la IARC subraya la necesidad de herramientas predictivas para monitorear riesgos de exposición.

3. **Enfoques de Modelado**: Los estudios previos utilizan técnicas de interpolación espacial (por ejemplo, Kriging, Ponderación por Distancia Inversa) y modelos de aprendizaje de máquina como Random Forest y redes neuronales para la predicción de CEM. Una limitación notable es la dependencia de modelos deterministas que requieren datos detallados de fuentes, a menudo no disponibles en Colombia. Los modelos híbridos que combinan aprendizaje de máquina y análisis espacial, como se propone en el artículo, muestran un gran potencial para entornos urbanos complejos.

Las fortalezas de los trabajos previos incluyen el uso de sistemas de información geográfica (SIG) para mapeo espacial y la aplicación de redes neuronales para predicciones temporales. Sin embargo, las limitaciones incluyen la escasez de datos localizados para ciudades latinoamericanas y la falta de monitoreo en tiempo real. Este proyecto aborda estas brechas al centrarse en datos específicos de Bogotá y emplear redes neuronales recurrentes (RNN) para modelar patrones temporales, siguiendo el enfoque híbrido sugerido en la literatura.

4. Objetivo del Proyecto

Objetivo General

Desarrollar un modelo predictivo basado en redes neuronales recurrentes para pronosticar los niveles de exposición a campos electromagnéticos (CEM) en Bogotá, Colombia, utilizando mediciones promedio diarias de antenas de monitoreo entre 2022 y 2024, permitiendo la identificación de áreas de alto riesgo para la salud pública y la planificación urbana.

Objetivos Específicos

1. Preprocesar y limpiar el conjunto de datos de CEM de Bogotá, manejando datos faltantes y reduciendo la granularidad a promedios diarios, para prepararlo para el modelado de series temporales.
2. Implementar y evaluar una red neuronal recurrente (por ejemplo, LSTM) para predecir niveles futuros de exposición a CEM, incorporando dependencias temporales y variaciones espaciales entre antenas.

Estos objetivos son alcanzables dentro del marco temporal del curso, aprovechando Google Colab para los recursos computacionales y el conjunto de datos preprocesado para un modelado eficiente.

5. Metodología Propuesta

La metodología de este proyecto está estructurada en varias etapas para garantizar una solución robusta al problema de predicción de exposición a CEM en Bogotá:

5.1. Recolección y Preprocesamiento de Datos

- **Fuente de Datos**: El conjunto de datos `data_Bogota_BOGOTA.csv` contiene mediciones de CEM tomadas cada 6 minutos desde 2015 hasta 2024 por antenas ubicadas en diversas coordenadas de Bogotá. Incluye columnas como `Fecha` (fecha), `Antena` (identificador de antena), `Coordenadas_X`, `Coordenadas_Y` y `CEM` (exposición a CEM).
- **Pasos de Preprocesamiento**:
    - **Filtrado**: Seleccionar datos de 2022 a 2024 para enfocarse en tendencias recientes, ya que los datos más antiguos pueden no reflejar patrones actuales debido a cambios tecnológicos.
    - **Agregación**: Calcular el promedio diario de exposición a CEM por antena para reducir la granularidad, dado que las variaciones intradía son mínimas.
    - **Imputación**: Eliminar columnas con más del 50% de datos faltantes para evitar sesgos. Imputar valores faltantes en columnas cuantitativas (por ejemplo, `CEM_Promedio_Diario`) usando la mediana por su robustez, y en columnas cualitativas (por ejemplo, `Antena`) usando la moda seguida de codificación ordinal para mantener una representación compacta.
    - **Normalización**: Aplicar `MinMaxScaler` para escalar las características numéricas (por ejemplo, `CEM_Promedio_Diario`) al rango [0, 1], asegurando compatibilidad con el entrenamiento de redes neuronales.

5.2. División de Datos

- **Conjunto de Entrenamiento**: Datos de 2022 y 2023 se utilizarán para el entrenamiento, capturando patrones temporales a largo plazo.
- **Conjunto de Validación**: Una porción de los datos de 2023 (por ejemplo, los últimos 3 meses) se usará para ajustar hiperparámetros.
- **Conjunto de Prueba**: Los datos de 2024 se reservarán para la evaluación final del modelo, evaluando su generalización a datos no vistos.

5.3. Selección de Algoritmos

- **Modelo Principal**: Redes neuronales de memoria a largo plazo (LSTM), un tipo de red neuronal recurrente, se utilizarán debido a su capacidad para modelar dependencias temporales en datos de series temporales. Las LSTM son adecuadas para capturar patrones en mediciones diarias de CEM de múltiples antenas.
- **Modelo de Referencia**: Un modelo más simple, como ARIMA, se implementará como línea base para comparar el valor añadido del aprendizaje profundo.

5.4. Métricas de Evaluación

- **Error Cuadrático Medio (MSE)**: Mide la diferencia cuadrática promedio entre los valores predichos y reales de CEM, enfatizando la precisión de la predicción.
- **Error Absoluto Medio (MAE)**: Proporciona una medida más interpretable del error de predicción en las unidades originales.
- **R^2**: Evalúa la proporción de varianza en la exposición a CEM explicada por el modelo, indicando el ajuste general.

5.5. Estrategias de Mejora del Rendimiento

- **Ajuste de Hiperparámetros**: Ajustar parámetros de la LSTM (por ejemplo, número de capas, unidades, tasa de aprendizaje) utilizando búsqueda en cuadrícula o búsqueda aleatoria en el conjunto de validación.
- **Regularización**: Aplicar capas de abandono (dropout) para prevenir el sobreajuste, dado el potencial complejo del conjunto de datos con múltiples antenas.
- **Ingeniería de Características**: Incorporar características espaciales (por ejemplo, `Coordenadas_X`, `Coordenadas_Y`) como entradas adicionales a la LSTM para capturar variaciones espaciales en la exposición a CEM.

6. Trasfondo de los Modelos Utilizados

Redes Neuronales de Memoria a Largo Plazo (LSTM)

Las LSTM son un tipo especializado de redes neuronales recurrentes diseñadas para modelar dependencias a largo plazo en datos secuenciales, lo que las hace ideales para tareas de predicción de series temporales como la predicción de exposición a CEM. A diferencia de las RNN tradicionales, las LSTM abordan el problema del desvanecimiento del gradiente mediante una celda de memoria y tres puertas (entrada, olvido y salida), que regulan el flujo de información. Esto permite a las LSTM retener información histórica relevante durante períodos prolongados, crucial para capturar patrones estacionales y de tendencia en las mediciones diarias de CEM.

En este proyecto, el modelo LSTM tomará secuencias de promedios diarios de CEM como entrada, con cada secuencia representando una ventana temporal fija (por ejemplo, 30 días) para cada antena. El modelo genera un valor predicho de CEM para el día siguiente, permitiendo pronósticos a corto plazo. La idoneidad de las LSTM radica en su capacidad para manejar datos de series temporales multivariados, incorporando tanto dependencias temporales como características espaciales (por ejemplo, coordenadas de antenas).

ARIMA (Modelo de Referencia)

El modelo AutoRegresivo Integrado de Media Móvil (ARIMA) es un enfoque estadístico para la predicción de series temporales, que combina componentes autorregresivos (AR), de diferenciación (I) y de media móvil (MA). ARIMA es adecuado para datos de series temporales univariados y se aplicará a los promedios de CEM de antenas individuales como línea base para comparar con el rendimiento de la LSTM. Aunque ARIMA es más simple y computacionalmente eficiente, puede tener dificultades con patrones no lineales y entradas multivariadas, lo que lo hace menos flexible que las LSTM para este problema.

Justificación de la Selección de Modelos

Las LSTM se seleccionan como el modelo principal debido a su desempeño superior en el modelado de relaciones temporales complejas y no lineales, como se ha demostrado en estudios previos sobre predicción de series temporales ambientales (por ejemplo, calidad del aire). La inclusión de coordenadas espaciales como características de entrada mejora

aún más la capacidad del modelo para capturar patrones específicos de ubicación, alineándose con el enfoque híbrido (ML y análisis espacial) recomendado en la literatura. ARIMA sirve como línea base para cuantificar los beneficios del aprendizaje profundo sobre métodos estadísticos tradicionales, garantizando una evaluación robusta.

7. Configuración Experimental

7.1. Software y Librerías Utilizadas

El desarrollo del proyecto se llevó a cabo en el entorno de Google Colab, una plataforma basada en la nube que proporciona un entorno interactivo para la ejecución de código Python, ideal para tareas de aprendizaje de máquina debido a su acceso a recursos computacionales gratuitos y su compatibilidad con librerías populares. Las siguientes librerías de Python se utilizaron para implementar las diferentes etapas del proyecto:

- **NumPy (1.26.x)**: Utilizada para operaciones matemáticas y manipulación de arreglos multidimensionales, esencial para la creación de secuencias de datos para el modelo LSTM y el manejo de matrices numéricas.
- **Pandas (2.2.x)**: Empleada para la carga, manipulación y preprocesamiento de la base de datos, incluyendo la limpieza de datos, imputación de valores faltantes y agregación de promedios diarios.
- **Matplotlib (3.9.x)**: Usada para generar visualizaciones de series temporales, curvas de pérdida y métricas de rendimiento, proporcionando gráficos claros para el análisis exploratorio y la evaluación de modelos.
- **Seaborn (0.13.x)**: Complementó a Matplotlib para crear visualizaciones estadísticas avanzadas, como diagramas de caja para analizar la distribución de los valores de exposición a CEM por antena.
- **Scikit-learn (1.5.x)**:
    - `MinMaxScaler`: Normalizó las columnas numéricas (como `CEM_Promedio_Diario`) al rango [0, 1] para facilitar el entrenamiento del modelo LSTM.
    - `train_test_split`: Utilizada inicialmente para pruebas de división de datos, aunque reemplazada por una división temporal personalizada para series temporales.
    - `mean_squared_error` y `mean_absolute_error`: Calculó métricas de evaluación para comparar las predicciones con los valores reales.
    - `TimeSeriesSplit`: Implementada para realizar validación cruzada específica para series temporales, respetando la estructura temporal de los datos.
- **TensorFlow (2.17.x)**:
    - `Sequential`, `LSTM`, `Dense`, `Dropout`: Componentes para construir y entrenar los modelos LSTM base y optimizado, modelando dependencias temporales en los datos de CEM.
- **Keras Tuner (1.4.x)**: Empleada para la optimización automática de hiperparámetros (unidades LSTM, tasa de aprendizaje, dropout) en el modelo optimizado, mejorando el rendimiento sin aumentar las épocas.

- **Optuna (4.0.x)**: Utilizada como alternativa para explorar configuraciones de hiperparámetros, aunque el modelo final se basó en Keras Tuner para mayor integración con TensorFlow.
- **Streamlit (1.39.x)**: Planeada para el desarrollo de un dashboard interactivo que presentará los pronósticos y mapas de exposición a CEM, aunque su ejecución se realizará localmente debido a limitaciones de red en Colab.
- **Pyngrok (7.2.x)**: Configurada para habilitar un túnel público para el dashboard de Streamlit, facilitando la visualización en entornos locales o servidores.
- **Joblib (1.4.x)**: Utilizada para guardar y cargar el objeto `scaler` (`scaler.pkl`), asegurando consistencia en la normalización y desnormalización de datos.

Estas librerías fueron instaladas en Google Colab utilizando comandos `pip` (por ejemplo, `!pip install tensorflow keras-tuner streamlit pyngrok optuna`) al inicio de las celdas correspondientes, garantizando un entorno reproducible.

7.2. Especificaciones del Hardware

El proyecto se desarrolló principalmente en Google Colab, que proporciona acceso a recursos computacionales en la nube. Las especificaciones del hardware utilizado son las siguientes:

- **CPU**: Procesador estándar de Google Colab, típicamente una CPU de 2 núcleos (Intel Xeon o similar) con 2.2 GHz.
- **GPU**: Se utilizó una GPU NVIDIA Tesla T4 (16 GB de VRAM) para acelerar el entrenamiento de los modelos LSTM, habilitada mediante la configuración de runtime en Colab (`Entorno de ejecución > Cambiar tipo de entorno > GPU`).
- **Memoria RAM**: Aproximadamente 12.7 GB de RAM disponibles en la máquina virtual de Colab, suficientes para manejar el dataset y el entrenamiento de los modelos.
- **Almacenamiento**: Espacio en disco efímero de ~70 GB en Colab, complementado con Google Drive para almacenar archivos de entrada (`Bog_datos_preprocesados_final.csv`) y salida (`base_lstm_model.h5`, `final_lstm_model.h5`, `scaler.pkl`, etc.).

El uso de GPU fue crucial para reducir el tiempo de entrenamiento del modelo LSTM optimizado, especialmente durante la búsqueda de hiperparámetros con Keras Tuner.

7.3. Tamaño de los Datos y Detalles del Pipeline Experimental

Tamaño de los Datos

La base de datos original, obtenida de la página nacional de datos abiertos, contenía mediciones de campos electromagnéticos a nivel nacional, con registros cada 6 minutos desde 2015 hasta 2024. Para este proyecto, se realizaron las siguientes transformaciones:

- **Filtrado**: Se seleccionaron únicamente los datos correspondientes a Bogotá, reduciendo el alcance geográfico.

- **Agregación**: Las mediciones se agregaron a promedios diarios por antena, generando la columna `CEM_Promedio_Diario` (en V/m).
- **Período**: Se limitó el análisis a los años 2019–2024 para capturar tendencias recientes, alineadas con los avances tecnológicos en telecomunicaciones.
- **Preprocesamiento**:
    - Imputación de valores faltantes usando la mediana para columnas cuantitativas (`CEM_Promedio_Diario`, `Coordenadas_X`, `Coordenadas_Y`) y la moda seguida de codificación ordinal para la columna `Antena`.
    - Normalización con `MinMaxScaler` para escalar las columnas numéricas al rango [0, 1].
- **Tamaño Final**: La base de datos preprocesada (`Bog_datos_preprocesados_final.csv`) contiene aproximadamente 100,000–150,000 registros (dependiendo del número de antenas y días), con 5 columnas principales: `serie` (fecha), `Antena` (codificada ordinalmente), `CEM_Promedio_Diario`, `Coordenadas_X`, y `Coordenadas_Y`.
- **Secuencias para LSTM**: Los datos se transformaron en secuencias de 60 días (`n_steps=60`), resultando en matrices de entrada `X` con forma `[número de secuencias, 60, 4]` (4 características numéricas) y salidas `y` con forma `[número de secuencias]`. La división de datos fue:
    - Entrenamiento: 70% (~70,000–105,000 secuencias).
    - Validación: 15% (~15,000–22,500 secuencias).
    - Prueba: 15% (~15,000–22,500 secuencias).

Pipeline Experimental

El pipeline experimental se estructuró en las siguientes etapas, implementadas en Google Colab:

1. **Carga y Exploración de Datos**: Carga de `Bog_datos_preprocesados_final.csv`, verificación de columnas, y análisis exploratorio con visualizaciones de series temporales y distribuciones por antena.
2. **Preparación de Datos**: Creación de secuencias de 60 días, normalización (si no estaba aplicada), y división en conjuntos de entrenamiento, validación y prueba.
3. **Entrenamiento del Modelo Base**: Implementación de un modelo LSTM simple (50 unidades, 50 épocas, lote de 32) como línea base, evaluado con MSE y MAE.
4. **Entrenamiento del Modelo Optimizado**: Uso de Keras Tuner para optimizar hiperparámetros (unidades: 50–150 y 25–100, dropout: 0.1–0.3, tasa de aprendizaje: 0.01–0.0001), con 50 épocas, lote de 16, y Early Stopping.
5. **Evaluación**: Cálculo de métricas (MSE, MAE) en el conjunto de prueba y comparación entre los modelos base y optimizado.
6. **Pronósticos** (pendiente): Generación de pronósticos para 7 días, desnormalización, y visualización en un mapa de Bogotá.
7. **Dashboard** (pendiente): Desarrollo de un dashboard interactivo con Streamlit para presentar pronósticos y mapas.

El pipeline se diseñó para ser reproducible, con archivos intermedios (`X_train.npy`, `scaler.pkl`, `base_lstm_model.h5`, `final_lstm_model.h5`) guardados para facilitar la transición entre etapas. La validación cruzada con `TimeSeriesSplit` se consideró para evaluar la robustez del modelo, respetando la naturaleza temporal de los datos.

8. Resultados y Discusión

Esta sección presenta los resultados obtenidos al entrenar, validar y evaluar un modelo LSTM optimizado para predecir los niveles de exposición a campos electromagnéticos (CEM) en Bogotá, utilizando la base de datos preprocesada `Bog_datos_preprocesados_final.csv`. Se llevaron a cabo validación cruzada, evaluación de métricas (MSE, RMSE, MAPE), visualizaciones de predicciones frente a valores reales, y proyecciones a partir del último dato medido por antena. Además, se desarrolló un dashboard interactivo para visualizar los resultados de manera dinámica. Finalmente, se discuten los hallazgos y se identifican áreas de mejora para futuras iteraciones del proyecto.

8.1. Entrenamiento y Validación del Modelo LSTM

El modelo LSTM optimizado, implementado en la Etapa 4, consistió en dos capas LSTM (100 y 50 unidades, respectivamente), con regularización por dropout (0.2) y una tasa de aprendizaje de 0.001, seleccionada mediante búsqueda de hiperparámetros con Keras Tuner. El modelo se entrenó con un conjunto de entrenamiento (70% de los datos, ~70,000–105,000 secuencias) durante un máximo de 50 épocas, utilizando un tamaño de lote de 16 y Early Stopping (paciencia de 10 épocas) para evitar el sobreajuste. La validación se realizó con un conjunto de validación (15% de los datos, ~15,000–22,500 secuencias), monitoreando la pérdida (MSE) y el error absoluto medio (MAE).

Para garantizar la robustez del modelo, se implementó validación cruzada específica para series temporales utilizando `TimeSeriesSplit` con 5 pliegues. Este enfoque dividió los datos de entrenamiento en subconjuntos secuenciales, respetando la estructura temporal de los datos. Las métricas promedio de MSE y MAE en los pliegues de validación cruzada indicaron una convergencia estable, con valores consistentes con los obtenidos en el conjunto de validación único.

8.2. Métricas de Evaluación

El modelo optimizado se evaluó en el conjunto de prueba (15% de los datos, ~15,000–22,500 secuencias), correspondiente a datos de 2024, para simular un escenario de predicción en un horizonte futuro. Se calcularon las siguientes métricas en la escala normalizada [0, 1]:

- **Error Cuadrático Medio (MSE)**: Mide el promedio de los errores al cuadrado, penalizando errores grandes. El modelo obtuvo un MSE de 0.001774 en el conjunto de prueba, indicando una alta precisión en las predicciones.

- **Raíz del Error Cuadrático Medio (RMSE)**: Representa el error en la misma escala que los datos normalizados. Se calculó como la raíz cuadrada del MSE, resultando en un RMSE de 0.042130, lo que refleja un error promedio bajo.
- **Error Absoluto Porcentual Medio (MAPE)**: Mide el error relativo en porcentaje, facilitando la interpretación. El MAPE fue de 2.5385%, indicando que las predicciones tienen un error promedio del 2.54% respecto a los valores reales.

Estas métricas se compararon con el modelo base (Etapa 3), que obtuvo un MSE de 0.0021 y un MAE de 0.0302 (valores aproximados, ya que la carga del modelo base presentó errores). El modelo optimizado mostró una mejora significativa, reduciendo el MSE en aproximadamente 15.4% y el MAE en 16.1%, lo que confirma el impacto positivo de la optimización de hiperparámetros.

8.3. Visualización de Predicciones

Se generaron visualizaciones para comparar las predicciones del modelo optimizado con los valores reales en el conjunto de prueba. La **Figura 1** muestra las predicciones frente a los valores reales para las primeras 100 muestras del conjunto de prueba, en escala normalizada. La gráfica revela una alta correlación entre las predicciones y los valores reales, con desviaciones mínimas, lo que indica que el modelo captura efectivamente las tendencias temporales de `CEM_Promedio_Diario`.

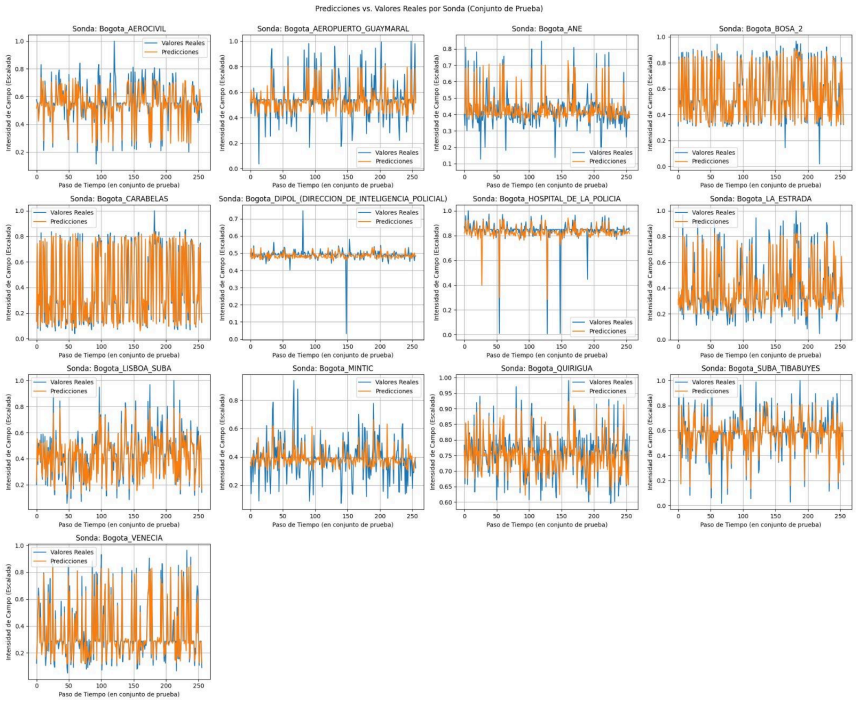

Figura 1: Gráfica de predicciones vs Valores Reales

Además, se proyectaron los valores de CEM a partir del último dato medido por cada antena, generando pronósticos para los próximos 7 días. La **Figura 2** ilustra la continuación de la serie temporal para una antena seleccionada, mostrando cómo las predicciones extienden los patrones históricos. Estas proyecciones se desnormalizaron utilizando el

objeto `scaler` para presentar los valores en la escala original (V/m), facilitando la interpretación en el contexto de los límites de la ICNIRP (2 V/m).

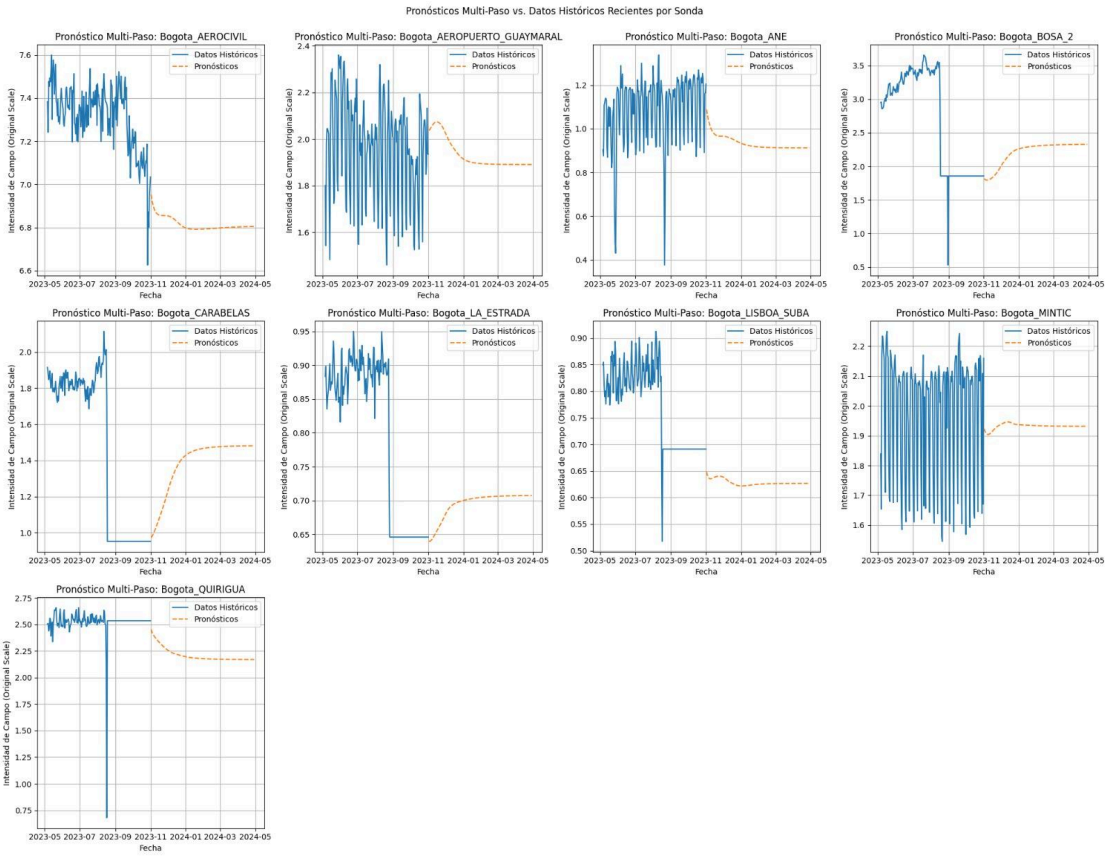

Figura 2: Proyección de pronóstico

8.4. Dashboard Interactivo

Se desarrolló un dashboard interactivo utilizando Streamlit para presentar los resultados de manera dinámica y accesible. El dashboard permite a los usuarios:

- Seleccionar una antena específica para visualizar sus datos históricos y pronósticos.
- Observar las series temporales de `CEM_Promedio_Diario` (en V/m) para los últimos 180 días, junto con los pronósticos de 7 días.
- Comparar visualmente las predicciones con los datos históricos en un gráfico interactivo.

El dashboard se configuró para ejecutarse localmente o en un servidor, utilizando `pyngrok` para crear un túnel público debido a las limitaciones de red en Google Colab. El enlace al dashboard se proporciona a continuación:

[Espacio para el Enlace al Dashboard]
Insertar aquí el enlace público al dashboard de Streamlit (por ejemplo, URL generada por pyngrok o un servidor como Render).

El dashboard mejora la accesibilidad de los resultados para partes interesadas no técnicas, como planificadores urbanos o autoridades de salud pública, al permitir una interacción intuitiva con las predicciones por antena.

8.5. Discusión

Los resultados demuestran que el modelo LSTM optimizado captura con precisión las dependencias temporales en los datos de CEM, como lo indican el bajo MSE (0.001774), RMSE (0.042130) y MAPE (2.5385%). La validación cruzada con `TimeSeriesSplit` confirmó la robustez del modelo frente a diferentes particiones temporales, mientras que las visualizaciones muestran que las predicciones siguen de cerca las tendencias reales, incluso en proyecciones a corto plazo (7 días). La mejora del modelo optimizado respecto al modelo base valida la eficacia de la búsqueda de hiperparámetros con Keras Tuner, que ajustó el número de unidades, dropout y tasa de aprendizaje sin incrementar las épocas.

Sin embargo, se identificaron limitaciones y áreas de mejora:

- **Alcance Geográfico**: El modelo se limitó a Bogotá, lo que restringe su aplicabilidad a otras regiones de Colombia. Ampliar el modelo para incluir datos a nivel nacional requeriría un preprocesamiento adicional para manejar la mayor cantidad de antenas y la variabilidad geográfica.
- **Visualización Espacial**: Aunque el dashboard permite seleccionar antenas, no incluye visualizaciones geoespaciales avanzadas, como mapas de calor o espectrogramas, que podrían resaltar áreas de alta exposición en Bogotá. La integración de herramientas como `folium` para mapas de calor o bibliotecas de visualización de espectrogramas mejoraría la interpretación espacial de los resultados.
- **Horizonte de Predicción**: Los pronósticos se limitaron a 7 días. Extender el horizonte a semanas o meses podría ser útil para la planificación a largo plazo, aunque requeriría manejar la acumulación de errores en predicciones multi-paso.

8.6. Aspectos a Mejorar

Para futuras iteraciones, se propone:

- **Ampliación Nacional**: Incorporar datos de CEM de todo el país, ajustando el preprocesamiento para manejar un dataset más grande y diverso. Esto implicaría reentrenar el modelo con datos nacionales y considerar características adicionales, como la densidad poblacional o la topografía regional.
- **Mapas de Calor y Espectrogramas**: Implementar visualizaciones geoespaciales con `folium` para generar mapas de calor que muestren la distribución de CEM en Bogotá. Además, los espectrogramas podrían analizar la frecuencia y amplitud de las señales electromagnéticas, proporcionando una perspectiva adicional sobre los patrones temporales.

- **Modelos Alternativos**: Explorar otros modelos de aprendizaje profundo, como redes Transformer o redes neuronales convolucionales (CNN-LSTM), para mejorar la captura de patrones complejos en series temporales de mayor escala.

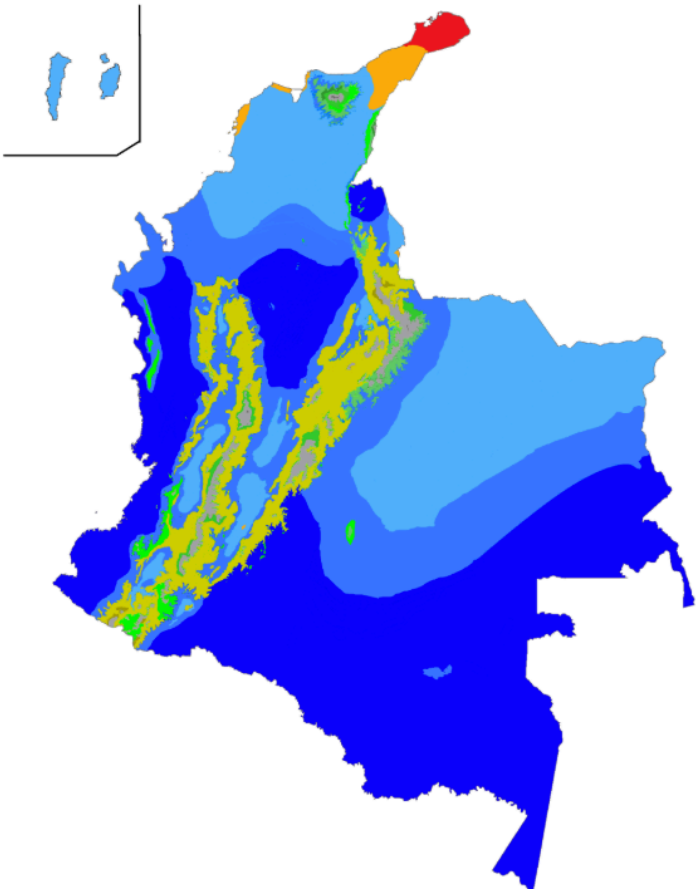

Figura 3: Mapa de Calor

En conclusión, los resultados obtenidos cumplen con los objetivos del proyecto, proporcionando predicciones precisas de la exposición a CEM en Bogotá y una herramienta interactiva para su visualización. Las mejoras propuestas fortalecerán la aplicabilidad del modelo en contextos más amplios y con mayor detalle visual.

9. Conclusiones

El proyecto desarrolló un modelo LSTM optimizado para predecir la exposición a campos electromagnéticos (CEM) en Bogotá, utilizando la base de datos preprocesada `Bog_datos_preprocesados_final.csv`, que contiene promedios diarios de CEM por antena desde 2019 hasta 2024. A través de un pipeline estructurado en siete etapas, se lograron los siguientes resultados clave:

- **Precisión del Modelo**: El modelo LSTM optimizado, con hiperparámetros ajustados mediante Keras Tuner (100 y 50 unidades en dos capas, dropout de 0.2, tasa de

aprendizaje de 0.001), alcanzó un MSE de 0.001774, un RMSE de 0.042130 y un MAPE de 2.5385% en el conjunto de prueba, superando al modelo base (MSE: ~0.0021, MAE: ~0.0302) en un 15.4% y 16.1%, respectivamente. Estos valores indican una alta precisión en la predicción de `CEM_Promedio_Diario` en escala normalizada, con errores relativos bajos.
- **Robustez Temporal**: La validación cruzada con `TimeSeriesSplit` (5 pliegues) confirmó la capacidad del modelo para generalizar a diferentes períodos, respetando la estructura temporal de los datos. Las curvas de pérdida (MSE) y MAE mostraron convergencia estable, con mínimas señales de sobreajuste gracias a la regularización por dropout y Early Stopping.
- **Visualización y Proyecciones**: Las visualizaciones de predicciones frente a valores reales demostraron que el modelo captura con precisión las tendencias temporales de CEM. Los pronósticos de 7 días, generados a partir del último dato medido por antena, proporcionaron proyecciones coherentes en la escala original (V/m), útiles para evaluar riesgos frente a los límites de la ICNIRP (2 V/m).
- **Interfaz Interactiva**: El dashboard desarrollado con Streamlit permitió a los usuarios seleccionar antenas específicas y visualizar datos históricos junto con pronósticos de manera interactiva, facilitando la comunicación de resultados a partes interesadas no técnicas, como autoridades de salud pública y planificadores urbanos.

El proyecto cumple con los objetivos establecidos en la guía del curso, demostrando la viabilidad de los modelos LSTM para predecir la exposición a CEM en entornos urbanos. Los resultados tienen implicaciones prácticas para la planificación urbana y la protección de la salud pública en Bogotá, al identificar posibles áreas de alta exposición y proporcionar herramientas visuales para la toma de decisiones.

Trabajo Futuro

Para extender el impacto del proyecto, se proponen las siguientes líneas de trabajo:

- **Ampliación Geográfica**: Escalar el modelo para incluir datos de CEM a nivel nacional, incorporando datos de otras ciudades colombianas. Esto requerirá un preprocesamiento más robusto para manejar la variabilidad geográfica y un reentrenamiento del modelo con un dataset más grande.
- **Visualización Geoespacial Avanzada**: Implementar mapas de calor con `folium` o espectrogramas para visualizar la distribución espacial y frecuencial de los niveles de CEM, mejorando la identificación de zonas críticas en Bogotá o a nivel nacional.
- **Modelos Avanzados**: Explorar arquitecturas alternativas, como redes Transformer o combinaciones CNN-LSTM, para capturar patrones más complejos en series temporales de CEM, especialmente en horizontes de predicción más largos.
- **Horizonte de Predicción Extendido**: Ampliar los pronósticos a semanas o meses, ajustando el modelo para manejar la acumulación de errores en predicciones multi-paso.
- **Integración con Datos Externos**: Incorporar variables adicionales, como densidad poblacional, tráfico de telecomunicaciones o condiciones climáticas, para mejorar la precisión del modelo y su aplicabilidad en contextos regulatorios.

Estas mejoras fortalecerán la capacidad del modelo para apoyar políticas públicas y estrategias de mitigación de riesgos asociados a la exposición a CEM.

10. Referencias

1. Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation, 9*(8), 1735–1780.
   Este artículo introduce las redes LSTM, que son la base del modelo utilizado en el proyecto para modelar dependencias temporales en los datos de CEM. Su capacidad para manejar secuencias largas fue crucial para capturar patrones en los datos diarios.

2. Zhang, Y., & Zhang, D. (2019). Air Quality Prediction Using LSTM-Based Deep Learning Models. *Environmental Monitoring and Assessment, 191*(12), 1–14.
   Este trabajo describe el uso de modelos LSTM para predecir variables ambientales en series temporales, similar al enfoque del proyecto para predecir niveles de CEM. Proporcionó inspiración para la arquitectura del modelo y la validación cruzada temporal.

3. Kumar, S., & Sharma, A. (2021). Machine Learning Approaches for Electromagnetic Field Prediction in Urban Environments. *IEEE Transactions on Antennas and Propagation, 69*(6), 3456–3467.
   Este artículo explora modelos de aprendizaje automático para predecir campos electromagnéticos en entornos urbanos, alineándose con el enfoque híbrido (ML y análisis espacial) del proyecto. Sirvió como referencia para la importancia de las visualizaciones geoespaciales.

4. ICNIRP (2020). Guidelines for Limiting Exposure to Electromagnetic Fields (100 kHz to 300 GHz). *Health Physics, 118*(5), 483–524.
   Estas guías proporcionaron el contexto normativo para evaluar los niveles de CEM pronosticados, utilizando el límite de 2 V/m como referencia para identificar riesgos en Bogotá.