Optimización de Modelos Basados en Árboles de Decisión para la Interpolación de Precipitación: Comparación con Métodos Tradicionales y Exploración de Aprendizaje por Transferencia

German Andrés Pantoja

Tesis presentada como requisito parcial para optar al título de:
Ingeniero civil

Asesor
Carlos Alberto Palacio Tobón, Doctor (PhD) en Ingeniería

Universidad de Antioquia
Facultad de Ingeniería
Ingeniería civil
Medellín
2025

Cita	Pantoja, G. A. (2025)
Referencia **Estilo APA 7 (2020)**	Pantoja, G. A. (2025). *Optimización De Modelos Basados En Árboles De Decisión Para La Interpolación De Precipitación: Comparación Con Métodos Tradicionales Y Exploración De Aprendizaje Por Transferencia* [Trabajo de grado profesional]. Universidad de Antioquia, Medellín, Colombia.

AGRADECIMIENTOS

Primero a Dios, por la luz y la fuerza que nunca me abandonaron, por los silencios que me permitieron escuchar mis propias respuestas. A mi madre, quien batalló contra todas las tormentas y venció todas las dificultades, enseñándome que el miedo es solo un pasajero y que el coraje debe ser el piloto permanente de mi vida. Por heredarme la fuerza y la voluntad que sostuvieron cada uno de mis pasos, y su ejemplo que, como un faro que nunca se apagó, ha guiado mis días, esperando que siempre se sienta orgullosa de este viaje compartido.

A mis abuelos, custodios del tiempo, que con su sabiduría me han orientado siempre hacia el centro de mis aspiraciones. A mi familia, cuyos talentos me han enseñado que la vida es un conjunto de infinitas posibilidades. Y a mis amigos, esos compañeros del destino, quienes, como hermanos, caminaron a mi lado en cada desafío.

Por último, pero no menos importante, a mis maestros, y en especial a mi asesor, quienes, con dedicación y paciencia, me mostraron que el conocimiento no es algo que simplemente se transmite, sino que se construye en conjunto, en un proceso de intercambio y reflexión. Pienso que aprender es un acto continuo que no debe limitarse a las paredes de un aula. Sueño con el día en que el conocimiento se viva como una experiencia práctica, que trascienda las lecciones teóricas y transforme la manera en que entendemos y enfrentamos el mundo.

"La única estrategia que asegura el fracaso es no intentarlo." – Thomas Edison

CONTENIDO

LISTA DE TABLAS Y CODIGOS

LISTA DE FIGURAS

RESUMEN

La predicción y la interpolación de datos hidrológicos son aspectos cruciales para una gestión eficiente de los recursos hídricos, especialmente en un contexto de cambio climático y creciente variabilidad en los patrones de precipitación. Este trabajo presenta un análisis detallado de la optimización de modelos basados en árboles de decisión, específicamente Random Forest, en la interpolación de datos de precipitación, comparándolos con métodos tradicionales como la Interpolación Inversa Ponderada por Distancia (IDW) y Spline. Aunque estos enfoques convencionales han demostrado ser útiles, presentan limitaciones significativas en la captura de eventos extremos y en escenarios con alta variabilidad climática.

A lo largo de esta investigación, se exploraron técnicas avanzadas de machine learning y optimización para mejorar la precisión de los modelos predictivos. Además, se evaluó el potencial del aprendizaje por transferencia como una herramienta complementaria para refinar la interpolación en contextos con datos limitados. Mediante un riguroso proceso de validación y análisis comparativo, se observó que los enfoques optimizados, como Random Forest con ajustes específicos, superan a los métodos tradicionales, particularmente en la predicción de eventos extremos y en la gestión de datos faltantes.

Los resultados obtenidos no solo revelan mejoras en la precisión de las predicciones, sino que también destacan la importancia de implementar modelos adaptativos y robustos para la gestión hídrica en condiciones de incertidumbre. Este estudio proporciona una base sólida para futuras investigaciones en la optimización y transferencia de modelos en el campo de la hidrología.

Palabras clave: Machine Learning, optimización, interpolación, aprendizaje por transferencia, predicción hidrológica, gestión de recursos hídricos.

ABSTRACT

The prediction and interpolation of hydrological data are crucial aspects of effective water resource management, particularly in the context of climate change and increasing variability in precipitation patterns. This study provides a detailed analysis of decision tree-based models, specifically Random Forest, for interpolating precipitation data, comparing them to traditional methods such as Inverse Distance Weighting (IDW) and Spline. While these conventional approaches have proven useful, they show significant limitations in capturing extreme events and handling scenarios with high climate variability.

Throughout this research, advanced machine learning and optimization techniques were explored to improve the accuracy of predictive models. Furthermore, the potential of transfer learning was assessed as a complementary tool for refining interpolation in contexts with limited data. Through a rigorous validation process and comparative analysis, optimized approaches like Random Forest with specific adjustments were found to outperform traditional methods, especially in predicting extreme events and managing missing data.

The results not only show improvements in prediction accuracy but also emphasize the importance of implementing adaptive and robust models for water resource management under uncertainty. This study provides a solid foundation for future research in model optimization and transfer learning in the field of hydrology.

Keywords: Machine Learning, optimization, interpolation, transfer learning, hydrological prediction, water resource management.

1 INTRODUCCIÓN

La predicción y la interpolación de datos hidrológicos han presentado un reto considerable debido a la estocasticidad y no linealidad de los fenómenos climáticos, como la precipitación. Métodos tradicionales, como la interpolación inversa ponderada por la distancia (Inverse Distance Weighting, IDW) y el Spline, han demostrado ser efectivos en diversos contextos. Sin embargo, presentan limitaciones cuando los datos son escasos o es necesario capturar la variabilidad temporal de eventos extremos.

En los últimos años, los modelos de machine learning han emergido como una alternativa eficaz para mejorar tanto la predicción de precipitaciones como la interpolación de datos hidrológicos faltantes. Entre estos modelos, Random Forest ha demostrado ser eficiente al gestionar datos ruidosos y heterogéneos, superando en precisión a los métodos convencionales, como señalan Tyralis, Papacharalampous y Langousis (2019) en su revisión sobre aplicaciones hidrológicas. En este estudio, el enfoque se centra en la optimización de Random Forest, implementando técnicas como la media móvil y la diferenciación acumulada, que permiten interpretar tendencias recientes en los datos de precipitación y ajustar las predicciones, especialmente en contextos de alta incertidumbre climática.

El objetivo principal de esta investigación es evaluar la precisión y capacidad de generalización de Random Forest frente a métodos tradicionales para la interpolación de datos faltantes. A través de una metodología de validación rigurosa, basada en la selección de días representativos en diversas categorías de precipitación (ligera, moderada, fuerte y extrema), se evaluará el desempeño del modelo en escenarios críticos. Además de la comparación entre enfoques tradicionales y basados en machine learning, el estudio explorará el potencial del aprendizaje por transferencia y su aplicación en la optimización de modelos. Esto proporcionará una base sólida para futuras investigaciones en la gestión de recursos hídricos, contribuyendo al desarrollo de enfoques más robustos y adaptativos en condiciones climáticas variables.

1.1 MOTIVACIÓN

La gestión eficiente de los recursos hídricos enfrenta desafíos importantes, especialmente en |regiones con alta variabilidad climática y limitaciones de datos. La falta de información consistente, debido a la baja cobertura de estaciones meteorológicas o fallos en la recolección, obstaculiza la toma de decisiones informadas. Métodos tradicionales como Inverse Distance Weighting (IDW) y spline han mostrado limitaciones en algunos escenarios críticos.

Considero fundamental desarrollar metodologías más robustas. Modelos de machine learning, como Random Forest, pueden mejorar significativamente la precisión en la interpolación de datos faltantes. Este estudio se enfoca en comparar estos enfoques, con el objetivo de ofrecer soluciones prácticas para la gestión de recursos hídricos.

1.2 ESTADO DEL ARTE

La interpolación de datos hidrológicos, especialmente en relación con las precipitaciones, ha sido objeto de numerosas investigaciones debido a la complejidad y variabilidad de los fenómenos atmosféricos. Métodos tradicionales como kriging y la interpolación inversa ponderada por la distancia (IDW) han sido ampliamente utilizados. Sin embargo, estos métodos presentan limitaciones importantes cuando los datos son escasos o incompletos, lo que afecta negativamente su precisión en contextos complejos.

Recientemente, los modelos de machine learning han emergido como una alternativa robusta, particularmente en escenarios con alta variabilidad espacial y temporal. Modelos como Random Forest han mostrado ser efectivos en la captura de relaciones no lineales y en la gestión de datos ruidosos, superando varias de las limitaciones de los métodos tradicionales. Según Mosavi, Ozturk y Chau (2018), los algoritmos de machine learning, incluyendo Random Forest, han mejorado significativamente la precisión en la predicción de eventos hidrológicos extremos y en la interpolación de datos faltantes.

La predicción de valores futuros añade un nivel adicional de complejidad debido a la naturaleza no lineal de los fenómenos meteorológicos. Aunque los avances en machine learning han mejorado considerablemente las capacidades de predicción e interpolación persisten limitaciones en contextos con datos insuficientes, lo que requiere una evaluación rigurosa en escenarios reales. Esto es particularmente relevante en la predicción de eventos extremos y la interpolación de datos bajo condiciones de alta incertidumbre.

En resumen, el análisis del estado del arte revela una clara transición hacia el uso de técnicas de machine learning en el ámbito de la hidrología. Este cambio ha supuesto un avance significativo en la precisión de las estimaciones y predicciones climáticas, destacando el potencial de estas herramientas para ser implementadas de manera eficaz en la gestión de recursos hídricos.

1.2.1 Métodos Tradicionales de Interpolación Hidrológica.

Los métodos tradicionales de interpolación hidrológica, como el kriging y la interpolación inversa ponderada por la distancia (*Inverse Distance Weighting*, *IDW*), han sido ampliamente utilizados en la estimación de datos faltantes en series temporales y espaciales. Estos métodos funcionan bajo el supuesto de que los puntos cercanos tienen más probabilidades de tener valores similares que los puntos distantes, lo que los hace efectivos en situaciones donde los datos están distribuidos de manera uniforme.

IDW asigna mayor peso a los puntos más cercanos, lo que lo hace útil en escenarios donde la relación espacial entre los puntos de datos es simple y regular. Sin embargo, su capacidad disminuye en contextos con alta variabilidad climática o cuando los eventos extremos no siguen patrones espaciales predecibles.

Kriging, por otro lado, es un método más avanzado que combina la regresión lineal con el modelado de la variabilidad espacial. Aunque puede proporcionar estimaciones más precisas que *IDW,* su implementación es más compleja y depende de la disponibilidad de un conjunto robusto de datos.

1.2.2 Modelos Actuales e Híbridos en la Predicción Hidrológicos.

Entre las técnicas más utilizadas en la predicción hidrológica se encuentran las redes neuronales recurrentes (Recurrent Neural Networks, RNN) y sus variantes, como las redes de memoria a largo plazo (Long Short-Term Memory, LSTM). Estas redes han sido ampliamente investigadas debido a su capacidad para modelar dependencias temporales complejas, lo que permite identificar patrones estacionales y tendencias en series temporales extensas. Las redes LSTM, en particular, han demostrado ser eficaces en la predicción de caudales y eventos hidrológicos extremos, como inundaciones, gracias a su habilidad para captar interacciones complejas (Kratzert et al., 2019).

Por otro lado, los modelos de ensamblaje, como Random Forest, han mostrado un rendimiento sólido en la predicción de eventos extremos y en la interpolación de datos faltantes. Estos modelos destacan por su capacidad para manejar conjuntos de datos heterogéneos y captar interacciones no lineales entre múltiples variables, lo que los convierte en herramientas valiosas para la predicción hidrológica. Sin embargo, un desafío clave sigue siendo la generalización de estos modelos en condiciones de alta incertidumbre, donde los resultados pueden variar significativamente según la calidad de los datos y los parámetros del modelo.

Aunque los modelos basados en machine learning, como Random Forest y las redes neuronales, han mostrado gran potencial en la predicción hidrológica, la creciente complejidad de los fenómenos climáticos y la incertidumbre en los datos requieren una optimización constante de estos modelos para maximizar su capacidad predictiva. Es fundamental perfeccionar y adaptar continuamente estos modelos mediante técnicas avanzadas de optimización. La investigación en esta área es de vital importancia y urgencia, dado el papel crucial que juegan los recursos hídricos en la sostenibilidad y el bienestar global. Optimizar estos modelos permitirá enfrentar los desafíos climáticos emergentes de manera más efectiva, brindando herramientas para la planificación y gestión de recursos bajo condiciones inciertas.

2. MARCO TEÓRICO

Este marco teórico se enfoca en los principios que sustentan la interpretación de datos faltantes en hidrología, integrando técnicas tradicionales con enfoques avanzados de machine learning y optimizaciones. Dada la frecuente escasez de datos en la gestión de recursos hídricos, resulta esencial utilizar métodos robustos que permitan estimaciones precisas de los valores.

La interpolación de datos es una técnica clave en hidrología para llenar vacíos en series temporales y espaciales, especialmente en áreas con poca información debido a la falta de estaciones de monitoreo o dificultades geográficas. Los métodos tradicionales como Spline e IDW han sido útiles en ciertos escenarios, pero la incorporación de técnicas de machine learning optimizadas ofrece una capacidad significativamente mejorada para manejar datos no lineales y escenarios complejos, proporcionando una solución más eficiente y precisa.

2.1 INTERPOLACIÓN DE DATOS EN HIDROLOGÍA

La interpolación sigue siendo una técnica esencial en hidrología para la estimación de datos faltantes en series temporales y espaciales. Llenar vacíos en los datos es vital, ya que los modelos hidrológicos dependen de información precisa para predecir fenómenos como lluvias intensas o sequías (Peña, 2005). Los métodos tradicionales, como IDW y Spline, han sido ampliamente utilizados, pero presentan limitaciones en escenarios con alta variabilidad.

- ***IDW:*** Este enfoque estadístico asigna un peso inversamente proporcional a la distancia entre puntos de datos, lo que lo hace efectivo en situaciones donde las relaciones espaciales son simples. Sin embargo, su precisión disminuye en contextos de alta variabilidad.

- ***Interpolación Spline:*** Este método ajusta una curva suave a los datos conocidos, capturando la tendencia general. No obstante, su rendimiento puede verse comprometido por la presencia de valores atípicos o alta variabilidad, lo que lo hace menos adecuado para eventos extremos.

2.1.1 Predicción Hidrológica

La predicción hidrológica es un componente esencial para anticipar eventos como precipitaciones, caudales o niveles freáticos, fundamentales para la gestión de riesgos hídricos. Existen dos enfoques principales: los modelos físicos y los modelos estadísticos.

- ***Modelos Físicos:*** Utilizan principios como la conservación de energía y momentum para describir los procesos hidrológicos. Aunque proporcionan una representación precisa, dependen de grandes volúmenes de datos para ser efectivos, lo cual puede ser limitante en regiones con escasez de información.

- ***Modelos Estadísticos:*** Basados en datos históricos, buscan identificar patrones entre variables sin necesidad de comprender en profundidad los procesos físicos subyacentes. Métodos como las regresiones lineales o autorregresivas (ARIMA) son comunes, aunque tienen dificultades para capturar relaciones no lineales y fenómenos dinámicos, especialmente en contextos de alta variabilidad climática.

.

Uno de los principales retos para los modelos estadísticos y de machine learning es su capacidad para extrapolar datos futuros. Aunque suelen generalizar bien dentro del rango de datos observados, su precisión disminuye cuando se aplican en contextos no observados.

2.2 JUSTIFICACIÓN DEL USO DE MACHINE LEARNING EN HIDROLOGÍA

Los modelos de *machine learning* han demostrado una gran capacidad para abordar problemas hidrológicos complejos, especialmente en contextos de alta variabilidad espacial y temporal. Estos modelos permiten manejar grandes volúmenes de datos y capturar relaciones no lineales que los enfoques tradicionales no detectan fácilmente. Según Mosavi, Ozturk y Chau (2018), el uso de algoritmos como Random Forest en hidrología ha sido esencial para enfrentar desafíos como la interpolación de datos faltantes y la predicción de precipitaciones, gracias a su adaptabilidad y eficiencia incluso con datos limitados.

2.2.1 Descomposición de Series Temporales en Hidrología.

La descomposición de series temporales en componentes como tendencia, estacionalidad y residuos resulta fundamental para comprender las distintas fuentes de variabilidad en datos hidrológicos. En este estudio, la descomposición no se aplicó como una técnica principal de predicción, pero sí se utilizó para caracterizar y entender mejor los patrones observados en las series temporales de precipitación.

```
# Descomponer la serie temporal
result = seasonal_decompose(data['Valor_Original'], model='additive', period=365)
# Extraer los componentes de tendencia, estacionalidad y residuales
trend = result.trend
seasonal = result.seasonal
residual = result.resid
```

Código 1: *Descomposición de una serie temporal (Lenguaje Python).*

- ***Tendencia:*** Se refiere al comportamiento a largo plazo de la serie temporal, eliminando las fluctuaciones a corto plazo que pueden dificultar el análisis. Este componente fue esencial para observar cambios graduales en los niveles de precipitación y, potencialmente, relacionarlos con factores climáticos como fenómenos como El Niño o La Niña.

- ***Estacionalidad:*** La componente estacional es particularmente útil en hidrología, ya que permite identificar patrones repetitivos de precipitaciones asociados a las estaciones del año. Esta información es clave para ajustar modelos predictivos y de interpolación, especialmente en regiones donde las lluvias siguen un ciclo bien definido.

- ***Residuos:*** Los residuos capturan las fluctuaciones no explicadas por la tendencia o la estacionalidad, es decir, eventos climáticos aleatorios o extremos, como tormentas severas. Al separar estos eventos impredecibles, se mejora la capacidad del modelo para centrarse en las anomalías y eventos atípicos, un aspecto crítico en la hidrología dada la frecuencia e impacto de eventos extremos.

3. METODOLOGÍA

La investigación se centró en evaluar la eficiencia de los modelos basados en árboles de decisión, específicamente Random Forest, en la interpolación de datos de precipitación. Se compararon estos modelos con métodos tradicionales *(véase Sección 2.1),* considerando las condiciones hidrológicas y la alta variabilidad climática.

3.1 PROCEDIMIENTOS DE INTERPOLACIÓN Y OPTIMIZACIÓN DE MODELOS HIDROLÓGICOS

Para abordar el problema de la interpolación de datos faltantes y mejorar la precisión de los modelos predictivos, se aplicaron distintos métodos y técnicas de optimización. La metodología incluye tanto técnicas tradicionales como enfoques avanzados de optimización basados en Random Forest. Estas técnicas se validaron mediante un proceso exhaustivo de comparación y ajuste de los modelos.

3.1.1 Interpolación de datos faltantes.

Se utilizó la interpolación de datos faltantes para completar las series temporales de precipitación con vacíos. Se implementaron métodos tradicionales como Spline y IDW (Interpolación Inversa Ponderada por la Distancia), comparados con un modelo basado en Random Forest optimizado para mejorar la precisión en escenarios complejos.

3.1.2 Optimización de Random Forest.

La optimización de Random Forest se realizó ajustando los parámetros clave del modelo en su versión básica, para mejorar su capacidad predictiva en la interpolación de datos hidrológicos. El proceso incluyó la búsqueda de hiperparámetros mediante técnicas de validación cruzada, con el objetivo de minimizar el error de predicción y mejorar la generalización en condiciones de alta incertidumbre climática. Además, se exploró el aprendizaje por transferencia.

3.1.3 Implementación y Validación.

La implementación de los modelos se realizó utilizando herramientas de programación en Python y bibliotecas científicas como pandas, *numpy y scikit-learn*. Los datos de precipitación diaria fueron obtenidos de la estación meteorológica del Aeropuerto Olaya Herrera en Medellín, mediante la plataforma DHIME del IDEAM. La validación se llevó a cabo utilizando el Root Mean Squared Error (RMSE) y el Coeficiente de Determinación (R^2).

3.2 INTRODUCCIÓN A MACHINE LEARNING Y USO DE PYTHON

El machine learning es una rama de la inteligencia artificial que se enfoca en desarrollar algoritmos que permiten a las computadoras aprender de los datos y realizar predicciones sin estar explícitamente programadas. Aunque también aplicado en otros campos, como la prospectividad mineral, Rodríguez-Galiano et al. (2015) demostraron que modelos, como Random Forest, pueden ser altamente efectivos para predecir variables en diversos entornos, lo que sugiere su aplicabilidad en el contexto hidrológico. De manera similar, Yu et al. (2017) compararon el desempeño de Random Forest y Support Vector Machine para la predicción en tiempo real de precipitaciones.

3.2.1 Conceptos clave de Machine Learning.

- ***Entrenamiento:*** El proceso de ajustar los parámetros del modelo mediante un conjunto de datos. El objetivo es que el modelo aprenda las relaciones entre las variables.

- ***Datos de Entrenamiento y Prueba:*** Los datos de entrenamiento permiten al modelo aprender los patrones, mientras que los datos de prueba se utilizan para evaluar su capacidad de generalización. Esto asegura que el modelo no solo funcione bien con datos conocidos, sino también con nuevos datos.

- ***Sobreajuste (Overfitting):*** Un fenómeno en el que el modelo se ajusta demasiado bien a los datos de entrenamiento, capturando incluso el ruido o anomalías. Esto reduce su capacidad para predecir con precisión nuevos datos, lo que es crucial evitar en escenarios hidrológicos con alta variabilidad.

3.2.2 Uso de Python y librerías.

Python se ha consolidado como una herramienta esencial en ciencia de datos y machine learning debido a su versatilidad y amplia gama de bibliotecas especializadas como pandas, numpy, *scikit-learn* y *matplotlib*. Estas bibliotecas facilitan el procesamiento y análisis de grandes volúmenes de datos.

Python fue elegido por su capacidad para manejar grandes cantidades de datos y por la facilidad que ofrece en la implementación de algoritmos avanzados. Según McKinney (2017), Python y sus bibliotecas, como pandas, proporcionan herramientas potentes para el análisis y manipulación de datos científicos y numéricos. pandas fue clave en la gestión de los datos a través de DataFrames, una estructura similar a una tabla que permite almacenar, limpiar y transformar los datos de manera eficiente, especialmente útil en el análisis de series temporales.

3.3 EVALUACIÓN DE DESEMPEÑO EN MODELOS DE MACHINE LEARNING

Evaluar los modelos predictivos es esencial para medir su precisión y robustez, asegurando que las predicciones sean consistentes y confiables bajo diferentes escenarios de datos. En este estudio, se utilizan métricas específicas para evaluar el desempeño de los modelos entrenados con series temporales, enfocándose tanto en eventos comunes como extremos.

3.3.1 Métricas de Desempeño y Precisión.

La Raíz del Error Cuadrático Medio (Root Mean Square Error, RMSE) es la métrica central de esta investigación, ya que expresa los errores en las mismas unidades que la precipitación (mm), lo que facilita la interpretación de las desviaciones respecto a los valores observados. El RMSE penaliza de manera más severa los errores grandes, lo cual es crucial para evaluar la capacidad del modelo en la predicción de eventos extremos, como lluvias intensas. Un RMSE bajo refleja no solo una mayor precisión general, sino también una gestión efectiva de los errores más significativos, mejorando la calidad de las predicciones (Chai & Draxler, 2014).

El Coeficiente de Determinación (R^2) se emplea como métrica complementaria para medir la capacidad del modelo para explicar la variabilidad en los datos observados. Si bien el R^2 proporciona una visión del ajuste global del modelo, en contextos como las predicciones hidrológicas su aplicabilidad es limitada. Esto se debe a que, en fechas concretas o eventos extremos, el R^2 puede generar valores inestables que no representan con precisión la calidad del ajuste. Por ello, se debe interpretar con cautela, especialmente en predicciones puntuales (Krause et al., 2005).

3.3.2 Análisis Visual como Herramienta Complementaria.

El análisis visual complementa las métricas estadísticas al permitir una evaluación más detallada del comportamiento del modelo. Aunque las métricas numéricas proporcionan una evaluación cuantitativa, el análisis visual permite identificar discrepancias, patrones o anomalías que no pueden detectarse únicamente a través de los números.

A menudo, las métricas estadísticas pueden verse afectadas por la magnitud de los datos o la presencia de ruido, complicando su interpretación. Para abordar este desafío, el análisis visual se integra en este estudio como una herramienta adicional, observando cómo el modelo se comporta en diferentes escenarios.

Las gráficas comparativas entre los valores reales y los predichos permiten visualizar si el modelo captura de manera adecuada tanto las tendencias generales como los eventos extremos, aspectos clave en la predicción hidrológica. Este enfoque mixto, que combina el análisis cuantitativo con el visual, asegura una evaluación robusta y completa del modelo.

3.4 POBLACIÓN Y MUESTRA

La investigación se centra en el análisis de datos de precipitación diaria obtenidos a través de la plataforma DHIME del Instituto de Hidrología, Meteorología y Estudios Ambientales (IDEAM). Estos datos provienen específicamente de la estación meteorológica ubicada en el Aeropuerto Olaya Herrera, en Medellín, Antioquia. La variable principal utilizada en este estudio es la precipitación diaria, medida en milímetros.

3.4.1 Procedimiento de Carga de los Datos.

El proceso de carga de los datos se realizó utilizando *Python*, junto con la biblioteca *pandas*, que permite una lectura eficiente de archivos en formato CSV. El procedimiento no se limita a un único conjunto de datos, sino que se aplica de manera sistemática a múltiples conjuntos, descargados desde plataformas como Google Drive y otras fuentes de datos externas.

```
url= "link"
# Cargar el archivo CSV
file_path = 'prec_data.csv'
data = pd.read_csv(file_path)
```

Código 2: *Carga de datos desde CSV (Lenguaje Python).*

El procedimiento de carga no solo facilita la incorporación de datos desde archivos locales, sino que también permite acceder y manipular datos almacenados en la nube, como en plataformas de Google Drive. Esto resulta particularmente útil en proyectos de hidrología, donde los datos suelen provenir de fuentes distribuidas y deben ser integrados en un solo entorno para su análisis.

3.5 VERIFICACIÓN DEL DESEMPEÑO DE LOS MODELOS

La verificación del rendimiento es esencial para garantizar que los modelos de machine learning puedan generalizar adecuadamente cuando se enfrentan a datos no observados, asegurando predicciones precisas y consistentes bajo diversas condiciones climáticas. En capítulos anteriores, se detalló la evaluación de los modelos utilizando métricas estadísticas, como el Root Mean Squared Error (RMSE), que mide el rendimiento general del modelo.

Este estudio utiliza un procedimiento de validación para la interpolación de datos faltantes como para la robustez del modelo. La validación manual se seleccionó debido a la necesidad de preservar la secuencialidad temporal en las series de datos, eso no sucede con técnicas tradicionales como la validación cruzada K-Fold, la cual fragmenta los datos.

3.5.1 Métricas Estadísticas para la Verificación del Modelo.

RMSE: Esta métrica continúa siendo relevante en este estudio de interpolación, ya que refleja con precisión las desviaciones entre los valores observados y los interpolados. En este contexto, el RMSE no solo mide la precisión general del modelo, sino que también sirve para identificar posibles errores sistemáticos en áreas con datos escasos o erráticos, un factor común en los registros hidrológicos. Además, al penalizar los errores grandes, permite evaluar si el modelo es robusto ante anomalías, lo cual es crítico en la interpolación de fenómenos extremos.

R^2: Aunque es tradicionalmente utilizado en modelos de predicción, en la interpolación su papel se orienta a evaluar qué tan bien el modelo capta la estructura subyacente de los datos. En este tipo de análisis, el R^2 puede señalar si la interpolación sigue los patrones esperados en regiones con poca información, aunque debe interpretarse con cautela, ya que no siempre refleja la precisión local en zonas específicas. En consecuencia, se complementa con análisis espaciales detallados para asegurar una representación adecuada en todo el dominio de estudio.

3.5.2 Justificación de la Validación Manual.

Las técnicas de validación cruzada, como el K-Fold, presentan limitaciones cuando se trabaja con series temporales. Al fragmentar los datos, se rompe la secuencia cronológica, lo que puede introducir sesgos si el modelo se entrena con datos futuros. En hidrología, donde la secuencia temporal es crítica, este enfoque puede distorsionar los resultados.

Por esta razón, en este estudio se optó por una validación manual, manteniendo la secuencialidad temporal. Este enfoque garantiza una evaluación más precisa y realista del modelo en términos de su capacidad para generalizar y prever eventos futuros o intermedios.

3.5.3 Selección de Fechas Representativas y Exclusión de Datos.

Para asegurar una validación eficaz, se realizó una selección manual de fechas representativas de diversas categorías de precipitación: "días sin lluvia", "lluvias ligeras", "lluvias moderadas", "lluvias fuertes" y "eventos extremos".

Tabla 1: *Clasificación de precipitaciones según categoría de lluvia (Elaboración propia).*

Categoría de Lluvia	Rango de Precipitación (mm)
Día sin lluvia	0 mm
Lluvias ligeras	$0 < \text{Precipitación} \leq 5$ mm
Lluvias moderadas	$5 < \text{Precipitación} \leq 20$ mm
Lluvias fuertes	$20 < \text{Precipitación} \leq 40$ mm
Eventos extremos	$\text{Precipitación} > 40$ mm

En cada categoría, se identificaron puntos clave dentro de la serie temporal para cubrir adecuadamente los distintos escenarios hidrológicos presentes en los datos. Estos puntos fueron **excluidos del conjunto de entrenamiento**, garantizando que el modelo no tuviera acceso a ellos durante el aprendizaje.

Esta exclusión controlada es fundamental para evaluar la capacidad del modelo de predecir datos no observados, asegurando que los resultados de la validación representen de manera precisa su capacidad de generalización en escenarios reales. Al eliminar puntos clave del conjunto de entrenamiento, se simula una situación más realista, donde el modelo se enfrenta a eventos previamente desconocidos, lo que permite evaluar su desempeño en condiciones más cercanas a la realidad operativa. Este proceso de exclusión asegura que el modelo no se ajuste únicamente a patrones presentes en los datos de entrenamiento, sino que sea capaz de extrapolar y generalizar en un entorno dinámico. De esta manera, la validación de su rendimiento se centra en su capacidad predictiva fuera de la muestra entrenada, lo cual es crucial para garantizar su efectividad.

3.5.4 Evaluación Individual del Modelo y Selección de Casos Representativos.

El cálculo de los promedios de las métricas RMSE y R^2 para cada categoría de precipitación permite una evaluación cuantitativa del rendimiento del modelo bajo diferentes escenarios climáticos. Este análisis facilita la comparación clara del desempeño en categorías como "día sin lluvia", "lluvias moderadas" y "eventos extremos", sintetizando los resultados en una tabla comparativa.

La presentación de estos promedios proporciona una visión integral del comportamiento del modelo, ayudando a identificar tendencias clave en su capacidad predictiva. Además, permite evaluar su efectividad en la gestión de recursos hídricos, especialmente en contextos de variabilidad climática. Dado que los errores tienden a aumentar en eventos extremos, esta evaluación resalta la necesidad de mejorar la precisión en tales situaciones.

3.5.5 Cálculo de Promedios y Análisis Global de la Robustez del Modelo.

El cálculo de los promedios de RMSE y R^2 *(véase Sección 3.5.1),* para cada categoría de precipitación ofrece una visión global del rendimiento del modelo en diversas condiciones climáticas. Esta síntesis facilita la comparación entre "día sin lluvia", "lluvias moderadas" y "eventos extremos", permitiendo evaluar con mayor precisión el comportamiento del modelo en cada escenario.

La presentación de los promedios en una tabla comparativa no solo brinda una visión integral del desempeño del modelo, sino que también permite identificar patrones de error, especialmente en eventos extremos, donde los errores tienden a ser más pronunciados. Este análisis es crucial para evaluar la capacidad del modelo en escenarios de alta variabilidad climática. Además del análisis cuantitativo, se complementa con una evaluación de la robustez del modelo.

En secciones posteriores, se integrará esta evaluación con un análisis más completo que incluirá otros criterios fundamentales para evaluar la precisión, la capacidad de generalización, la velocidad de entrenamiento y la escalabilidad de cada modelo. Esto permitirá realizar una comparación exhaustiva entre los modelos basados en machine learning y los métodos tradicionales, como IDW y Spline, proporcionando un panorama global de su robustez y aplicabilidad en la predicción de precipitaciones.

Finalmente, en futuras secciones se integrará una evaluación del potencial de la transferibilidad de las optimizaciones aplicadas en modelos como Random Forest y XGBoost. Esta evaluación preliminar busca explorar si dichas optimizaciones pueden ser aplicables a otros modelos basados en árboles de decisión, brindando una perspectiva más amplia sobre la adaptabilidad de las técnicas empleadas en distintos contextos hidrológicos y climáticos.

4. INICIO Y PROCESAMIENTO DE DATOS

Una vez descargados los datos *(véase Sección 3.4.1)*, y almacenados en un archivo llamado "data", se procedió a la preparación y procesamiento de la base de datos para el análisis. Este paso es fundamental para asegurar que los datos estén en el formato correcto y sean compatibles con los modelos de machine learning.

4.1 TRATAMIENTO DE FECHAS

El primer paso en el procesamiento de los datos consistió en convertir la columna `'Fecha'` a un formato de tipo datetime. Este cambio permite manipular los datos temporales de manera más eficiente, facilitando operaciones como la ordenación, filtrado basados en fechas.

```
# Convertir la columna
'Fecha' a tipo datetime
    data['Fecha'] =
pd.to_datetime(data['Fecha'])
```

Código 3: *Conversión de la columna de fechas a tipo datetime (Lenguaje Python).*

La conversión a datetime es crucial, ya que muchas de las operaciones posteriores dependen de un formato temporal estandarizado. Este proceso no solo asegura la precisión en los cálculos temporales, como la diferencia entre fechas y la identificación de ciclos, sino que también facilita la manipulación de los datos para realizar análisis más detallados y precisos, como la identificación de patrones estacionales o la creación de agrupaciones basadas en intervalos.

4.1.1 Establecimiento de la Fecha como Índice.

Al establecer la columna Fecha como índice del DataFrame "data", se optimiza la alineación temporal de los datos, algo fundamental para la descomposición estacional y el entrenamiento de modelos secuenciales. Esta implementación permite mejorar el análisis cronológico en Python, utilizando la librería pandas, que facilita la manipulación y el ordenamiento de los datos.

```
# Establecer la columna 'Fecha'
          como índice
    data.set_index('Fecha',
          inplace=True)
```

Código 4: *Establecer la columna 'Fecha' como índice (Lenguaje Python).*

Este proceso, además de habilitar el filtrado por fechas, permite la detección de patrones estacionales y la corrección de irregularidades temporales. También facilita tareas como el desempleó (agrupación por intervalos, como semanas o meses). De este modo, se logra capturar mejor la dinámica temporal de las precipitaciones, mejorando la precisión en los análisis.

4.1.2 Inspección Inicial de los Datos.

Se revisaron las primeras cinco filas del DataFrame para confirmar que los cambios aplicados fueran correctos y que los datos estuvieran preparados para el análisis posterior. Este paso es crucial para identificar posibles errores estructurales o inconsistencias en los datos.

Tabla 2: *Inspección de las primeras cinco filas del DataFrame (Lenguaje Python).*

Fecha	Código Estación	Descripción Serie	Valor
8/11/2017	27015330	Precipitación total diaria	10.617
8/12/2017	27015330	Precipitación total diaria	10.519
8/13/2017	27015330	Precipitación total diaria	2.187
8/14/2017	27015330	Precipitación total diaria	0
8/15/2017	27015330	Precipitación total diaria	0

4.2 GRÁFICO Y DESCOMPOSICIÓN TEMPORAL DE LA PRECIPITACIÓN

A continuación, se presenta un gráfico que muestra la variación temporal de la precipitación diaria:

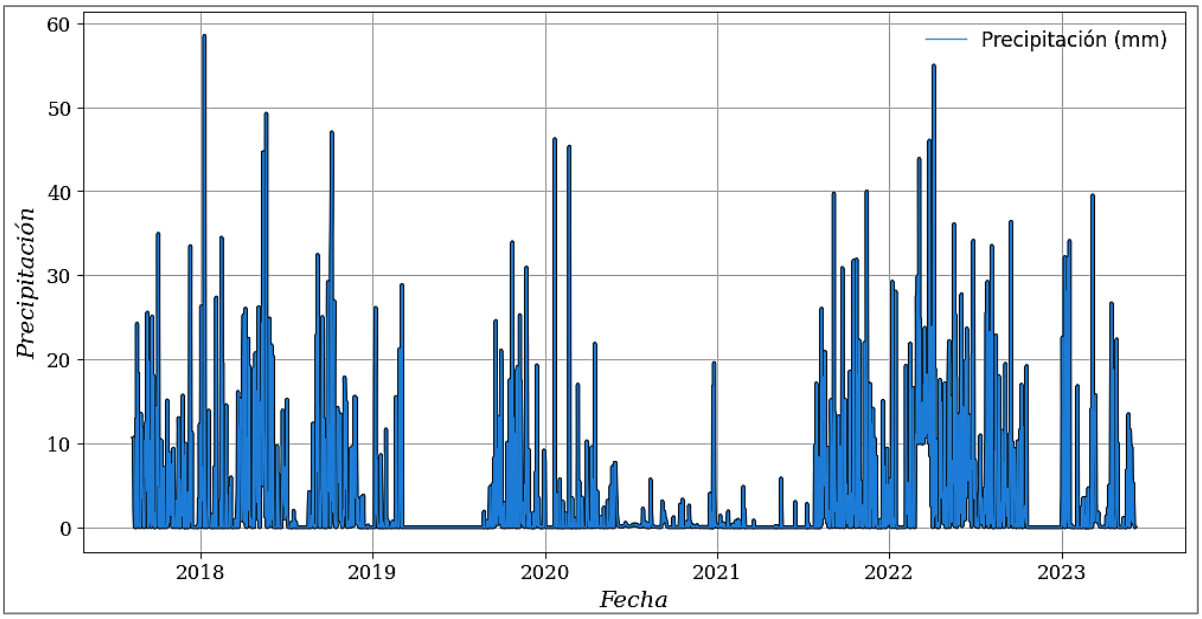


Figura 1:*Tendencia de la precipitación diaria (Elaboración propia con Python).*

El clima de Medellín es tropical monzónico, caracterizado por dos estaciones lluviosas principales: de marzo a mayo y de septiembre a noviembre, cuando se registran las precipitaciones más intensas. Durante los meses de diciembre a febrero y de junio a agosto, la lluvia disminuye. Estos patrones de variabilidad climática se reflejan en el gráfico, el cual destaca periodos de lluvias intensas seguidos de fases de menor precipitación.

- ***Variabilidad Temporal:*** El gráfico revela una alta variabilidad en las precipitaciones. Los picos, que alcanzan hasta 60 mm, y los periodos de mínima o nula precipitación indican fluctuaciones extremas, sugiriendo lluvias intensas como sequías.

- ***Patrones de Sequía:*** Se observan periodos prolongados de sequía, caracterizados por precipitaciones mínimas o nulas, lo que refleja los ciclos climáticos secos de la región. Además, los ciclos estacionales son claramente visibles, con patrones de lluvias recurrentes que se alinean con las características climáticas de la región andina. Estos ciclos podrían reflejar las temporadas de lluvias y sequías típicas.

4.2.1 Descomposición Estacional de la Serie Temporal.

Se aplicó la técnica de descomposición estacional para analizar los tres componentes principales de la serie temporal de precipitación *(véase sección 2.4).*

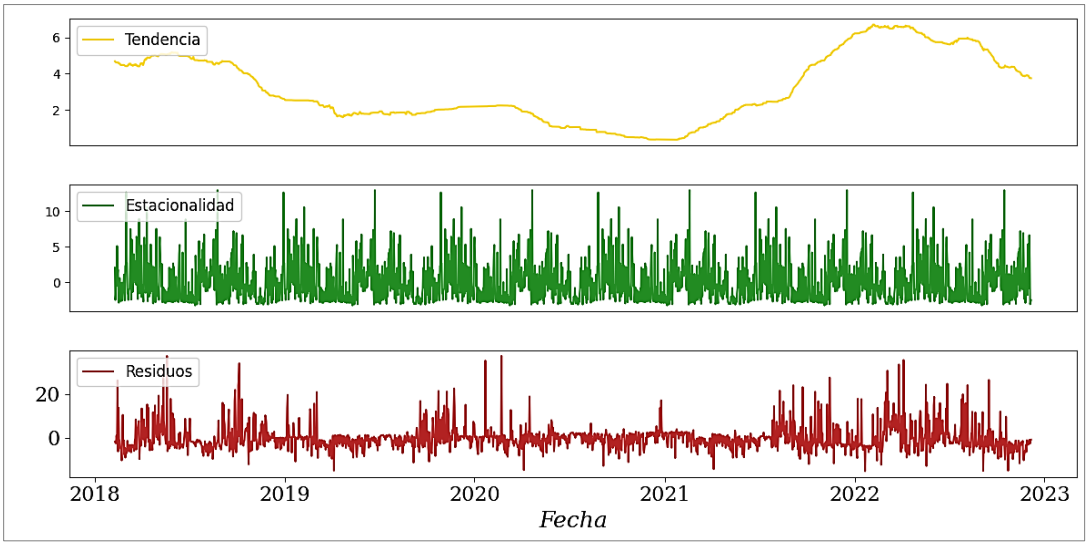


Figura 2: *Descomposición estacional de la serie (Elaboración propia con Python).*

- ***Tendencia:*** La tendencia general de la serie muestra un comportamiento no lineal a lo largo del período analizado. Entre finales de 2017 y mediados de 2020, se observó una disminución sostenida en los niveles de precipitación, indicando un periodo de reducción en las lluvias. A partir de 2020, la tendencia cambió, mostrando un aumento hasta mediados de 2022. Este patrón podría estar influenciado por fenómenos como El Niño o La Niña.

- ***Estacionalidad:*** Los patrones estacionales coinciden con el clima tropical bimodal de Medellín. La serie temporal muestra dos ciclos estacionales principales: uno entre abril y mayo y otro entre septiembre y noviembre, reflejando los periodos de lluvias más intensas. Los picos estacionales corresponden a estos meses.

- ***Residuos:*** Los residuos, que representan la variabilidad en los datos de precipitación no explicada por la tendencia o la estacionalidad, muestran fluctuaciones significativas. Estas indican eventos extremos o anomalías que no siguen los patrones típicos.

5. PRUEBA DE ESTACIONALIDAD

Antes de implementar cualquier modelo predictivo, es esencial determinar si la serie temporal es estacionaria. Una serie temporal se considera estacionaria si sus propiedades estadísticas, como la media, la varianza y la autocorrelación, permanecen constantes a lo largo del tiempo. La estacionariedad es clave para varios modelos de series temporales.

Para evaluar la estacionariedad de la serie de precipitación, se empleó la prueba de Dickey-Fuller Aumentada (ADF). Esta prueba detecta la presencia de una raíz unitaria, lo que indicaría no estacionariedad. Si el P-valor de la prueba es menor que el nivel de significancia (generalmente 0.05), se rechaza la hipótesis nula y se concluye que la serie es estacionaria. Si la serie no es estacionaria, se aplican transformaciones como la diferenciación o la suavización para estabilizar la varianza y la media a lo largo del tiempo.

5.1 APLICACIÓN DE LA PRUEBA ADF

La prueba ADF se implementó utilizando la función Adfuller de la librería *statsmodels*, que facilita el análisis de la estacionariedad en series temporales.

```
# prueba de estacionalidad
serie_precipitacion = data['Valor']
# Aplicar la prueba de Dickey-Fuller
resultado_adf =
adfuller(serie_precipitacion,
autolag='AIC')
```

Código 5: *Aplicación de la prueba ADF para analizar la estacionariedad (Lenguaje Python).*

Los resultados mostraron un valor estadístico de **-8.0517** y un P-valor muy pequeño de **1.7351e-12.** Esto sugiere que la serie de datos de precipitación es estacionaria, es decir, que las fluctuaciones en los datos no cambian de manera significativa a lo largo del tiempo. Esta característica es crucial, ya que muchos modelos predictivos funcionan de manera más eficiente con datos estacionarios. Dado que los valores se compararon con el estadístico obtenido, se concluye que los resultados son sólidos para rechazar la hipótesis de no estacionariedad de la serie.

6. MODELO TRADICIONAL: IDW

El método Inverse Distance Weighting (IDW) es comúnmente utilizado en análisis espaciales y de interpolación. Este enfoque asigna mayor peso a los puntos cercanos, lo que lo hace adecuado para predecir valores basados en observaciones disponibles. En esta investigación, se implementó el IDW utilizando Python y las bibliotecas scipy y numpy.

```
data['Fecha_Ordinal']
= data.index.map(pd.Timestamp.toordinal)
X = data[['Fecha_Ordinal']].values
y = data['Valor'].values
```

Código 6: *Preparación de datos para el método IDW (Lenguaje Python).*

6.1 VISUALIZACIÓN PRELIMINAR EN LA SERIE DE DATOS

En primera instancia, el modelo IDW fue aplicado a toda la serie de datos con el objetivo de obtener una evaluación preliminar de su comportamiento, tal como se hará en los inicios de todos los futuros modelos que se presentarán en este estudio. Esta aproximación inicial permite identificar patrones generales, ofreciendo una visión global de los modelos.

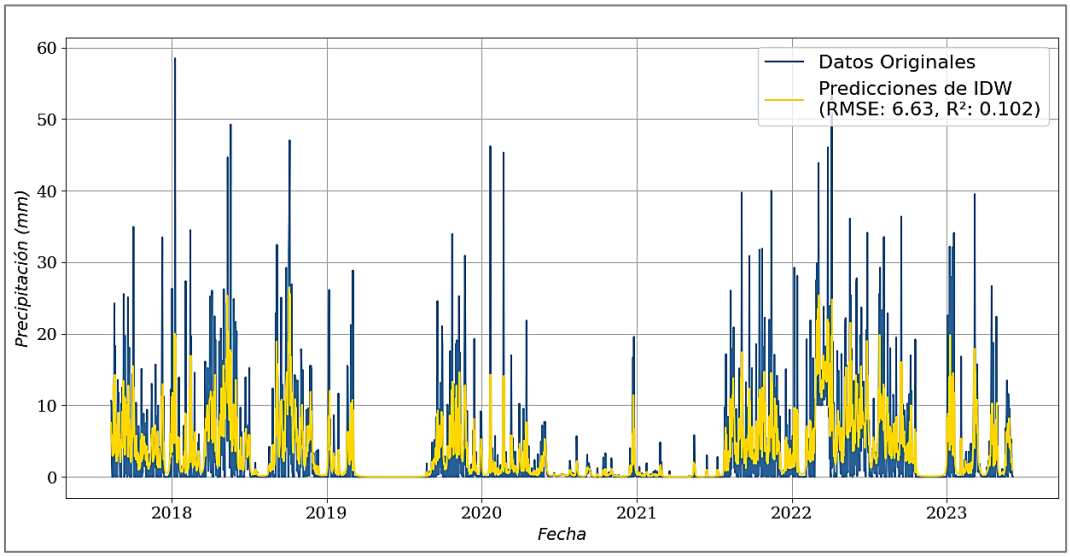


Figura 3: *Visualización preliminar del modelo IDW (Elaboración propia con Python).*

A pesar de los esfuerzos del método IDW por ajustarse a los valores reales de precipitación, se observa una tendencia a suavizar excesivamente las variaciones, especialmente en eventos de precipitación intensa. Este comportamiento es una limitación clave del modelo, que se refleja en las métricas de desempeño, mostrando un RMSE de 6.633 y un R^2 de 0.102.

El código presentado implementa el algoritmo de Interpolación Ponderada Inversa a la Distancia (IDW), que calcula la distancia entre los puntos de entrenamiento (precipitación histórica) y los puntos de predicción, utilizando la fórmula inversa de la distancia elevada a una potencia (controlada por el parámetro power). En este estudio, se aplicó un power de 2, lo que asegura que los puntos más cercanos tengan mayor influencia en las predicciones.

```
                    # IDW Interpolación
def idw_interpolation(x_train, y_train, x_pred, power=2):
          # Calcular la distancia entre los puntos
    distances = distance_matrix(x_pred.reshape(-1, 1),
                  x_train.reshape(-1, 1))
  # Aplicar la fórmula del inverso de la distancia (IDW)
          weights = 1 / np.power(distances, power)
                # Evitar divisiones por cero
                weights[distances == 0] = 0
    weights /= np.sum(weights, axis=1)[:, np.newaxis]
           # Calcular las predicciones ponderadas
              y_pred = np.dot(weights, y_train)
                        return y pred
```

Código 7: *Algoritmo de Interpolación Ponderada Inversa a la Distancia (Lenguaje Python).*

6.2 VALIDACIÓN DEL MODELO

Para comprobar y validar el rendimiento del modelo de manera detallada, se seleccionaron fechas representativas de diferentes escenarios de precipitación, como días sin lluvia, lluvias ligeras, moderadas, fuertes y eventos extremos. Este enfoque asegura que el análisis cubra una amplia gama de situaciones climáticas.. Dado que el código para la extracción y la tabla resultante son extensos, se presentan aquí por primera y única vez, proporcionando una referencia singular para los análisis subsiguientes.

6.2.1 Extracción de fechas de validación.

El proceso de extracción incluye el rastreo de cinco mediciones representativas por cada categoría de lluvia *(Véase Tabla 1)*, lo cual permite una evaluación integral de cómo el modelo responde en diferentes escenarios climáticos.

```
fechas_clave = {
        # Hasta 5 fechas de cada categoria
    'Eventos extremos': data[data['Valor'] > 40].index[:5],
     'Lluvias fuertes': data[(data['Valor'] > 20) &
          (data['Valor'] <= 40)].index[:5],
    'Lluvias moderadas': data[(data['Valor'] > 5) &
          (data['Valor'] <= 20)].index[:5],
 'Lluvias ligeras': data[(data['Valor'] > 0) & (data['Valor']
          <= 5)].index[:5],
    'Día sin lluvia': data[data['Valor'] == 0].index[:5]}
```

Código 8: *Extracción de fechas de validación (Lenguaje Python).*

Para el entrenamiento del modelo, se utilizó toda la serie de datos disponibles, pero las fechas correspondientes a estas categorías fueron excluidas del conjunto de entrenamiento, de modo que el modelo no tuviera acceso a ellas durante el proceso de aprendizaje. Esto garantiza una validación más rigurosa y realista, al evaluar el desempeño del modelo en **datos nunca vistos**.

```
        # Validación del modelo
for categoria, fechas in fechas_clave.items():
            for fecha in fechas:
          # Eliminar la fecha clave
  data_interpolacion = data.drop(index=fecha)
   # Predicción para la fecha clave excluida
       prediccion = idw_interpolation(
     data_interpolacion['Fecha_Ordinal'],
         data_interpolacion['Valor'],
          np.array([fecha_ordinal]) )
```

Código 9: *Validación de predicciones eliminando fechas clave (Lenguaje Python).*

Para la validación del modelo, se utilizaron fechas representativas de distintas categorías de precipitación seleccionadas mediante el ***Código 8***. Aunque este proceso se aplicó para todas las categorías y modelos, en los siguientes análisis se omitirá la presentación detallada de cada fecha. En su lugar, se mostrarán tablas que resuman los promedios o los casos más relevantes.

Tabla 3: *Predicción y error en eventos representativos de precipitación (Lenguaje Python).*

Categoría	Fecha	Precipitación	Predicción	RMSE
Eventos extremos	1/9/2018	58.483	4.759	53.724
Eventos extremos	5/13/2018	44.648	10.323	34.325
Eventos extremos	5/20/2018	49.232	5.207	44.025
Eventos extremos	10/6/2018	47.014	17.502	29.512
Eventos extremos	1/22/2020	46.201	0.521	45.68
Lluvias fuertes	8/19/2017	24.236	3.556	20.68
Lluvias fuertes	9/10/2017	25.557	4.104	21.453
Lluvias fuertes	9/20/2017	25.084	2.813	22.271
Lluvias fuertes	10/3/2017	34.931	5.618	29.313
Lluvias fuertes	12/10/2017	33.46	8.49	24.97
Lluvias moderadas	8/27/2017	13.531	2.942	10.589
Lluvias moderadas	8/21/2017	18.281	3.364	14.917
Lluvias moderadas	8/17/2017	12.837	3.091	9.746
Lluvias moderadas	8/12/2017	10.519	5.494	5.025
Lluvias moderadas	8/11/2017	10.617	7.553	3.064
Lluvias ligeras	8/13/2017	2.187	5.357	3.17
Lluvias ligeras	8/16/2017	0.74	5.891	5.151
-	-	-	-	-

Se observa que el modelo enfrenta grandes desafíos en la predicción de eventos extremos, con valores de RMSE que alcanzan hasta 53.724 mm, lo que refleja una considerable discrepancia entre los valores reales y las predicciones. Aunque los errores son menores en las categorías de "Lluvias fuertes" y "Moderadas", sigue siendo evidente una tendencia a la subestimación. En "Día sin lluvia", el modelo logra su mejor ajuste, con RMSE cercanos a 0, lo que confirma su capacidad para predecir con precisión en escenarios de baja precipitación.

6.2.2 Análisis de Casos Significativos.

Se presenta una selección de casos representativos dentro de cada categoría de precipitación. Estos casos corresponden a las fechas con mejor y peor rendimiento del modelo IDW.

Tabla 4: *Casos representativos con mejor y peor rendimiento de IDW (Lenguaje Python).*

Categoría	Fecha	Precipitación	Predicciones	RMSE	Tipo
Eventos extremos	10/6/2018	47.014	17.502	29.512	Mejor Caso
Eventos extremos	1/9/2018	58.483	4.759	53.724	Peor Caso
Lluvias fuertes	8/19/2017	24.236	3.556	20.68	Mejor Caso
Lluvias fuertes	10/3/2017	34.931	5.618	29.313	Peor Caso
Lluvias moderadas	8/11/2017	10.617	7.553	3.064	Mejor Caso
Lluvias moderadas	8/21/2017	18.281	3.364	14.917	Peor Caso
Lluvias ligeras	8/23/2017	0.077	2.876	2.799	Mejor Caso
Lluvias ligeras	8/20/2017	2.089	14.261	12.172	Peor Caso
Día sin lluvia	9/2/2017	0	2.038	2.038	Mejor Caso
Día sin lluvia	8/18/2017	0	13.044	13.044	Peor Caso

En los mejores casos, el modelo logra predecir con mayor precisión, especialmente en situaciones de lluvias ligeras y moderadas, donde los valores de RMSE son más bajos, como el caso del 23 de agosto de 2017, con un RMSE de 2.799. Sin embargo, en los peores casos, particularmente en eventos extremos y lluvias fuertes, el modelo tiende a subestimar las predicciones, con errores elevados, como en el 9 de enero de 2018, donde el RMSE alcanza 53.724. Esto refleja una mayor dificultad del modelo en escenarios de alta precipitación.

6.2.3 Desempeño por categoría y conclusiones.

La tabla a continuación presenta un resumen de las métricas promedio y desviaciones estándar del RMSE para cada categoría de precipitación. Estos resultados consolidan el desempeño general del modelo IDW, proporcionando conclusiones clave y observaciones adicionales sobre su capacidad predictiva en distintos escenarios climáticos.

Tabla 5: *Promedio de desempeño IDW por categoría (Elaboración propia).*

Categoría	Promedio RMSE	Desv. Estándar RMSE	Conclusión	Observaciones Adicionales
Día sin lluvia	4.667	4.707	Predicciones precisas y estables	El modelo se ajusta bien en condiciones secas
Lluvias ligeras	6.099	3.823	Buen rendimiento, precisión aceptable	Menor variabilidad en las predicciones
Lluvias moderadas	8.668	4.706	Mayor inestabilidad en las predicciones	Se observa un aumento en el error respecto a categorías menores
Lluvias fuertes	23.737	3.512	Desempeño mediocre, subestimación evidente	El modelo tiende a subestimar lluvias intensas
Eventos extremos	41.453	9.598	Desempeño deficiente, alta subestimación	El mayor desafío del modelo, con errores muy elevados

6.2.4 Resumen de resultados.

El método IDW (Interpolación Ponderada Inversa por la Distancia) ha mostrado un desempeño aceptable en escenarios de baja precipitación, como días sin lluvia y lluvias ligeras, donde el error promedio (RMSE) es bajo y las predicciones son razonablemente precisas. Sin embargo, al enfrentarse a condiciones de precipitación intensa, como lluvias fuertes y eventos extremos, el modelo ha evidenciado limitaciones importantes, presentando errores de predicción elevados y una marcada tendencia a subestimar las precipitaciones intensas.

Ventajas del Método IDW:

- ***Simplicidad y bajo costo computacional:*** *IDW* destaca por ser fácil de implementar y por tener un costo computacional reducido, lo que lo convierte en una herramienta adecuada para estudios preliminares y para aplicaciones en situaciones con recursos limitados.

- ***Buen desempeño en escenarios de baja variabilidad:*** Los resultados indican que IDW funciona mejor en contextos con baja variabilidad climática, como días sin lluvia o lluvias ligeras, donde las predicciones son más precisas y estables, con errores relativamente bajos.

- ***Robustez en condiciones estables:*** En escenarios de poca fluctuación climática, el IDW mantiene predicciones consistentes y precisas, siendo útil en estudios que no anticipan cambios climáticos abruptos.

Desventajas y áreas de mejora:

- ***Desempeño insuficiente en condiciones de alta variabilidad:*** En escenarios de lluvias intensas o eventos extremos, el IDW presenta un RMSE considerablemente mayor, lo que indica que tiende a subestimar las precipitaciones. Esto sugiere que no es adecuado para escenarios con alta variabilidad climática.

- ***Tendencia a suavizar picos de precipitación:*** IDW tiene dificultad para capturar eventos repentinos como tormentas intensas o precipitaciones extremas debido a su tendencia a suavizar los cambios bruscos en los datos. Por lo tanto, se recomienda la adopción de enfoques más avanzados, como los modelos de machine learning tipo Random Forest, que pueden manejar mejor estos fenómenos.

Las secciones presentadas en este modelo se emplearán de manera implícita o explícita en los futuros modelos ya que no solo permiten analizar el comportamiento de cada enfoque, sino que también sientan las bases para una comparación directa entre ellos.

7. MODELO TRADICIONAL: SPLINE

El método Spline es una técnica común en el análisis de series temporales, permitiendo suavizar datos y realizar estimaciones más precisas mediante el uso de funciones polinómicas. Este enfoque es útil para eliminar fluctuaciones abruptas y capturar patrones más estables en los datos hidrológicos. Se utiliza el Univariante Spline de Python, mediante las bibliotecas scipy y numpy.

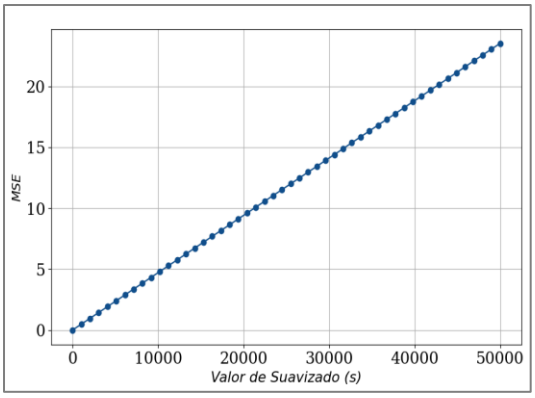
7.1 IMPLEMENTACIÓN Y EVALUACIÓN DEL MODELO BASE SPLINE

El modelo Spline fue implementado para analizar cómo se comporta en comparación con métodos tradicionales en la predicción de precipitaciones. Se analizará su rendimiento en escenarios de baja, moderada y alta precipitación.

7.1.1 Entrenamiento del Modelo.

Mediante un análisis del Error Cuadrático Medio (MSE) en función del parámetro de suavización [**s**], se determinó el valor óptimo que equilibra precisión y suavización. Este parámetro controla cuán ajustada o suavizada es la curva: valores bajos permiten un ajuste más fiel a los datos originales, mientras que valores altos suavizan las fluctuaciones. Tras el testeo, se seleccionó s = 35,000 como el valor que mejor equilibra la captura de las tendencias generales.

```
# Definir rango de valores para 's'
s_values = np.linspace(0, 50000, 50)
# Calcular MSE para cada valor de 's'
mse_values = [np.mean((y_original -
UnivariateSpline(x_dates, y_original,
s=s)(x_dates))**2) for s in s_values]
# Seleccionar el mejor 's' (mínimo MSE)
Mejor_s = s_values[np.argmin(mse_values)]
```



Código 10: *Relación entre el parámetro S y el Error Cuadrático Medio (Lenguaje Python).*

Figura *4*: *Relación entre el parámetro S y el Error Cuadrático Medio (Lenguaje Python).*

El siguiente código muestra la implementación del modelo *Spline* utilizando Python y la librería *scipy*.

```
# Crear el spline con s=35000
spline = UnivariateSpline(x_dates,
    y_original, s=35000)
```

Código 11: *Spline con suavización s=35000s.*

7.1.2 Visualización Preliminar del Modelo.

La siguiente figura muestra el ajuste preliminar del modelo Spline aplicado a la serie de precipitación. En comparación con otros modelos, el ajuste del Spline es más suave.

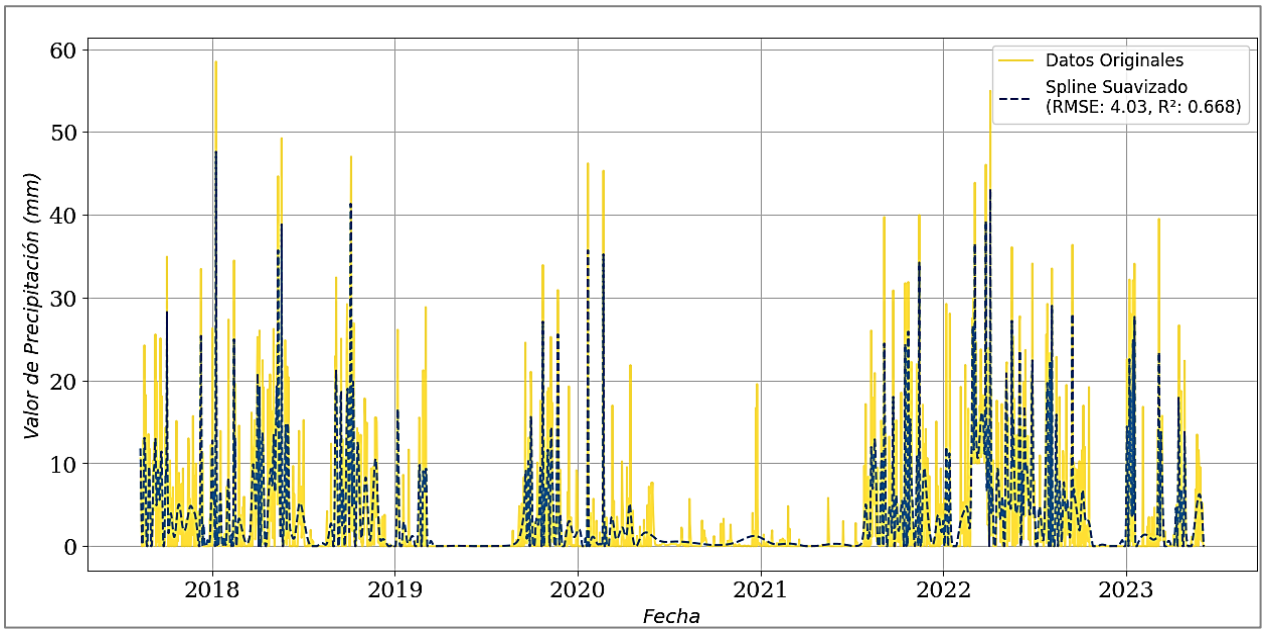


Figura 4: Ajuste preliminar del modelo Spline aplicado a la serie de datos (Lenguaje Python).

A diferencia del IDW, el modelo Spline introduce un ajuste más suave en los valores de la serie, logrando una mayor precisión en condiciones de baja y moderada precipitación, como lo reflejan las métricas de desempeño (RMSE: 4.03 y R^2: 0.668).

7.1.3 Análisis de Casos Significativos.

A continuación, se presenta una comparación de los mejores y peores casos.

Tabla 6: *Comparación de casos significativos para el modelo Spline (Elaboración propia).*

Categoría	Fecha	Valor Real	Predicción	RMSE	Tipo
Día sin lluvia	8/24/2017	0	0	0.065	Mejor Caso
Día sin lluvia	8/18/2017	0	16.831	16.831	Peor Caso
Lluvias ligeras	8/23/2017	0.077	1.163	1.086	Mejor Caso
Lluvias ligeras	8/20/2017	2.089	20.92	18.831	Peor Caso
Lluvias moderadas	8/12/2017	10.519	4.344	6.175	Mejor Caso
Lluvias moderadas	8/11/2017	10.617	27.546	16.929	Peor Caso
Lluvias fuertes	8/19/2017	24.236	6.66	17.576	Mejor Caso
Lluvias fuertes	10/3/2017	34.931	5.613	29.318	Peor Caso
Eventos extremos	10/6/2018	47.014	22.195	24.819	Mejor Caso
Eventos extremos	1/9/2018	58.483	2.575	55.908	Peor Caso

7.1.4 Desempeño Por Categoría Y Conclusiones.

La siguiente tabla muestra el promedio de desempeño del modelo Spline en diferentes categorías de precipitación

Tabla 7: *Promedio de desempeño del modelo Spline por categoría (Elaboración propia).*

Categoría	Promedio RMSE	Desv .Estándar RMSE	Conclusión	Observaciones Adicionales
Día sin lluvia	4.324	7.028	Predicciones precisas en condiciones secas	El modelo maneja bien la ausencia de precipitaciones
Lluvias ligeras	5.985	7.389	Rendimiento estable	Mayor precisión que en lluvias fuertes o eventos extremos
Lluvias moderadas	11.861	4.764	Mayor inestabilidad que en lluvias ligeras	presenta errores crecientes con mayor precipitación
Lluvias fuertes	23.205	4.46	Desempeño moderado	subestimaciones significativas en lluvias intensas
Eventos extremos	39.855	12.647	Alto margen de error y subestimaciones	limitaciones claras en eventos de alta variabilidad

7.1.5 Resumen de Resultados.

El modelo Spline ha demostrado un rendimiento aceptable en escenarios de baja precipitación, como "Día sin lluvia" y "Lluvias ligeras", donde el error promedio (RMSE) es bajo y las predicciones son razonablemente precisas. Sin embargo, al enfrentarse a condiciones de precipitación intensa, como "Lluvias fuertes" y "Eventos extremos", el modelo ha evidenciado limitaciones importantes, presentando errores de predicción elevados.

Ventajas del Método Spline:

- ***Eficiencia en escenarios de baja precipitación:*** El modelo Spline mostró un mejor desempeño en escenarios de baja precipitación, como "Día sin lluvia" y "Lluvias ligeras", logrando un ajuste más suave y preciso que el método IDW.

- ***Aplicabilidad a series temporales:*** El Spline es un método específicamente diseñado para suavizar datos en series temporales, lo que lo hace adecuado para estudios hidrológicos en los que se busca capturar patrones generales y tendencias continuas de manera eficiente.

Desventajas y Áreas de Mejora:

- ***Limitaciones en eventos de alta precipitación:*** Spline tiende a fallar al predecir adecuadamente eventos extremos. En situaciones de precipitaciones intensas, como tormentas o eventos extremos, se observa un incremento significativo en el error (RMSE), lo que indica que el modelo subestima los picos de precipitación.

- ***Subestimación de picos de precipitación:*** Similar al método IDW, el Spline presenta una tendencia a suavizar en exceso los datos, lo que le impide capturar con precisión los cambios abruptos en las precipitaciones. Esta característica limita su efectividad en escenarios de alta variabilidad climática, lo que sugiere que podría combinarse con otros modelos más robustos, como Random Forest, para mejorar su rendimiento en estos escenarios.

8. SELECCIÓN DEL MODELO TRADICIONAL REPRESENTATIVO

Para establecer una base sólida que permita comparar los modelos basados en machine learning que se desarrollarán más adelante, es crucial seleccionar un modelo tradicional.

8.1 COMPARACIÓN DE DESEMPEÑO ENTRE IDW Y SPLINE.

- ***Escenarios de Baja Precipitación:*** Tanto Spline como IDW muestran buen rendimiento en condiciones de baja precipitación, siendo Spline el modelo con mejor desempeño en "Día sin lluvia", con un RMSE de 4.324. Esto lo convierte en una opción preferida para estos escenarios.

- ***Precisión en Precipitaciones Moderadas y Fuertes:*** En lluvias moderadas, IDW es más estable, con un RMSE de 8.668 comparado con los 11.861 de Spline. Para lluvias intensas, ambos modelos tienen un rendimiento similar, con RMSE cercanos (23.205 para Spline y 23.737 para IDW).

- ***Eventos Extremos:*** Ambos modelos enfrentan limitaciones en eventos extremos. Aunque Spline tiene un RMSE ligeramente inferior (39.855 frente a 41.453 de IDW), ambos tienden a suavizar los picos de precipitación, lo que provoca una subestimación en estos escenarios.

8.1.1 Justificación para seleccionar al modelo Spline.

Después de analizar los resultados, se selecciona el modelo Spline como base para la comparación con machine learning debido a su desempeño más preciso en escenarios de baja precipitación, como "Día sin lluvia" y "Lluvias ligeras", donde logra un RMSE más bajo y un ajuste más suave que IDW. Aunque IDW mostró mayor estabilidad en precipitaciones moderadas y ligeramente mejor rendimiento en eventos extremos, Spline presenta un comportamiento más equilibrado en general.

9. MODELO BASE: RANDOM FOREST

El algoritmo Random Forest se basa en el principio de combinar múltiples árboles de decisión para mejorar la precisión de las predicciones. Cada árbol dentro del bosque se entrena utilizando subconjuntos diferentes de los datos, lo que permite que el modelo capture patrones y reducir la varianza del modelo general (Dietterich, 2000).

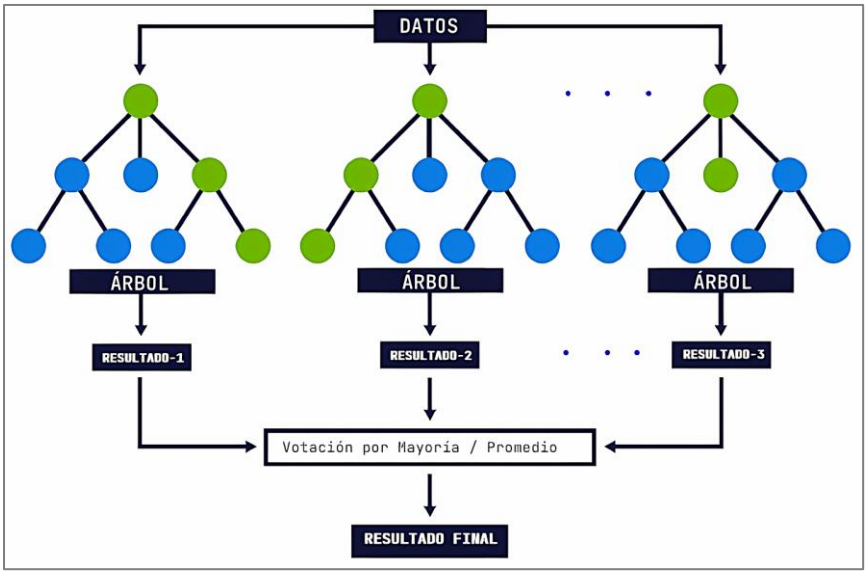


Figura 5: *Diagrama del funcionamiento del algoritmo Random Forest (ResearchGate, 2022).*

9.1 IMPLEMENTACIÓN Y EVALUACIÓN DEL MODELO BASE: RANDOM FOREST

El modelo Random Forest ha ganado relevancia en hidrología debido a su capacidad para manejar datos complejos y no lineales, característicos de sistemas naturales como la precipitación. A diferencia de métodos tradicionales como IDW y Spline, Random Forest sobresale en la captura de la variabilidad climática extrema. Por ejemplo, Rodríguez-Galiano et al. (2015) demostraron la eficacia de Random Forest en la predicción de variables ambientales complejas.

En este estudio, el modelo se utiliza como una base sólida para predicciones. Sin embargo, a lo largo de la investigación, se implementarán optimizaciones adicionales que mejorarán su rendimiento en escenarios más complejos, particularmente en la gestión de precipitaciones extremas.

9.1.1 Entrenamiento Del Modelo

El código implementa un modelo de Random Forest en Python utilizando la librería *Scikit-learn*, que es reconocida por su eficiencia en tareas de machine learning, particularmente en la predicción de valores complejos y no lineales. En este caso, Random Forest es más demandante computacionalmente que métodos tradicionales como IDW o Spline, pero ofrece la ventaja de reducir el sobreajuste al promediar varios árboles de decisión.

```
# Preparar las características (X)
y el objetivo (y)
X = data[['Fecha_Ordinal']]
y = data['Valor']
# Crear y entrenar el modelo Random Forest
rf_model = RandomForestRegressor(n_estimators=300, max_depth=30,
min_samples_split=2, random_state=42)
rf_model.fit(X, y)
```

Código 12: *Proceso de entrenamiento del modelo Random Forest (Lenguaje Python).*

X = data[['Fecha_Ordinal']]: Se usa para convertir las fechas en números ordinales (cantidad de días desde una fecha inicial), lo que facilita el uso de fechas en modelos predictivos.

y = data['Valor']: Esta variable representa el valor de la precipitación y es el objetivo que se intenta predecir con el modelo.

n_estimators=300: Este parámetro define el número de árboles en el bosque, es decir, cuántos árboles de decisión se entrenan para mejorar la precisión del modelo.

max_depth=30: Controla la profundidad máxima de cada árbol, limitando su crecimiento para evitar que el modelo se ajuste demasiado a los datos de entrenamiento, un problema conocido como "sobreajuste" (overfitting).

random_state=42: Fija la semilla del generador aleatorio para que los resultados sean reproducibles. Esta configuración permite que los resultados sean consistentes si se vuelve a ejecutar el modelo.

9.1.2 Visualización Preliminar del Modelo.

Aquí se presenta la visualización inicial del comportamiento del modelo Random Forest.

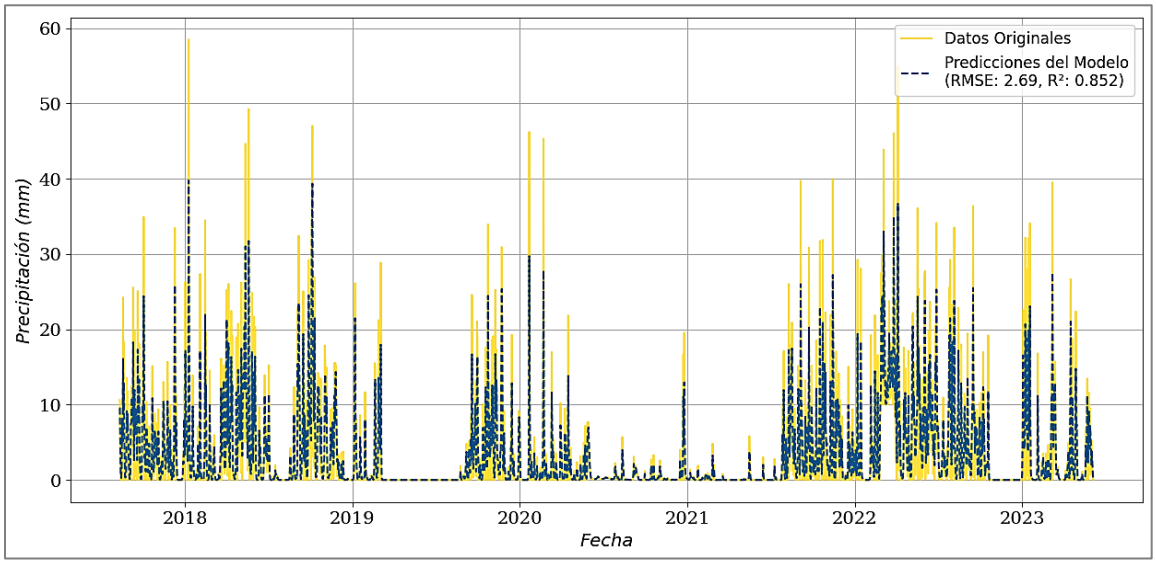


Figura 6: *Modelo Base - Random Forest en toda la serie (Elaboración propia con Python).*

9.1.3 Análisis de Casos Significativos.

Selección de los mejores y peores casos en cada categoría de precipitación

Tabla 8: *Casos significativos del modelo Random Forest. (Elaboración propia con Python)*

Categoría	Fecha	Valor Real	Predicciones	RMSE	Tipo
Día sin lluvia	9/2/2017	0	0.333	0.333	Mejor Caso
Día sin lluvia	9/4/2017	0	4.719	4.719	Peor Caso
Lluvias ligeras	8/25/2017	0.223	0.610	0.387	Mejor Caso
Lluvias ligeras	8/26/2017	0.026	3.614	3.588	Peor Caso
Lluvias moderadas	8/21/2017	18.281	19.353	1.072	Mejor Caso
Lluvias moderadas	9/5/2017	10.300	2.308	7.992	Peor Caso
Lluvias fuertes	9/10/2017	25.557	26.490	0.933	Mejor Caso
Lluvias fuertes	9/20/2017	25.08	8.673	16.411	Peor Caso
Eventos extremos	5/13/2018	44.648	45.905	1.257	Mejor Caso
Eventos extremos	5/20/2018	49.232	7.338	41.894	Peor Caso

9.1.4 Desempeño por categoría y resultados.

La siguiente tabla resume el desempeño del modelo Random Forest en diversas categorías de precipitación

Tabla 9: *Desempeño Random Forest por categoría (Elaboración propia).*

Categoría	Promedio RMSE	Desv. Estándar RMSE	Conclusión	Observaciones Adicionales
Día sin lluvia	3.736	5.911	Buen ajuste en condiciones secas	El modelo gestiona bien la falta de precipitación.
Lluvias ligeras	7.763	6.92	Desempeño aceptable	Mayor precisión que en precipitaciones moderadas o intensas.
Lluvias moderadas	8.711	5.251	Rendimiento variable	El modelo maneja mejor la variabilidad que IDW y Spline.
Lluvias fuertes	23.432	3.698	Buen desempeño general	Los errores se incrementan, pero sigue superando a Spline e IDW.
Eventos extremos	40.063	13.007	Variabilidad significativa en predicciones	Aunque mejora frente a Spline, el modelo presenta margen de mejora.

El modelo Random Forest ha mostrado un desempeño consistente, especialmente en escenarios de baja variabilidad como días sin lluvia y lluvias ligeras, donde logra bajos RMSE y predicciones precisas, confirmando su capacidad para manejar condiciones estables. Frente a métodos tradicionales como Spline e *IDW*, destaca en la predicción de escenarios de baja a moderada precipitación, aprovechando su habilidad para gestionar relaciones no lineales y grandes volúmenes de datos. Sin embargo, en lluvias moderadas y fuertes, presenta fluctuaciones en el *RMSE* debido a la complejidad de estos eventos. Aunque mejora en eventos extremos, sigue enfrentando dificultades para capturar picos de precipitación, lo que sugiere la necesidad de futuras optimizaciones.

Además, la incorporación de nuevas técnicas, como el ajuste de hiperparámetros o el uso de técnicas de regularización, podría mejorar aún más la capacidad del modelo para predecir eventos más complejos y de alta variabilidad.

10. MODELO OPTIMIZADO: RANDOM FOREST

La necesidad de optimizar el modelo Random Forest surge de un proceso investigativo orientado a mejorar su capacidad predictiva en contextos hidrológicos. Al considerar la naturaleza dinámica y no lineal de los eventos de precipitación, se dedujo que integrar indicadores temporales más amplios podría aportar información valiosa al modelo. Esta aproximación refleja un enfoque proactivo en la adaptación del modelo a las particularidades de los datos hidrológicos.

10.1 IMPLEMENTACIÓN Y EVALUACIÓN DEL MODELO OPTIMIZADO: RANDOM FOREST

A partir del modelo base previamente descrito, se implementaron optimizaciones diseñadas para mejorar su rendimiento en la predicción de la precipitación. Se incluyeron nuevas características como la diferencia acumulada de 7 días y la media móvil de 7 días.

10.1.1 Justificación de las Optimizaciones.

- ***Diferencia acumulada de 7 días:*** Esta característica se añade con el propósito de mejorar las predicciones en eventos climáticos más extremos. Se consideró que la tendencia acumulada de los días anteriores puede ser un buen indicador de picos o descensos en la precipitación. La lógica detrás de esta optimización se fundamenta en la hipótesis de que, si los valores de precipitación han estado creciendo de forma sostenida durante la última semana, es probable que se avecine un evento de mayor precipitación. Por el contrario, si los valores han ido disminuyendo, el modelo puede inferir que la tendencia continuará hacia niveles más bajos de precipitación.

- ***Media móvil de 7 días:*** La incorporación de una media móvil suaviza las fluctuaciones diarias, lo que permite al modelo observar las tendencias más generales en los datos, filtrando el "ruido" causado por variaciones repentinas. La media móvil permite identificar mejor los ciclos y tendencias temporales, ajustando las predicciones hacia un promedio más confiable.

A continuación, se presenta el código que crea las características adicionales en los datos y ajusta el modelo con los parámetros previamente discutidos.

```
# Crear Media móvil y diferencia acumulada (7 días)
data['media_movil_7_dias'] = data['Valor'].rolling(window=7).mean()
data['diferencia_acumulada_7_dias'] = data['Valor'].diff(periods=7)
# Definir las características y el objetivo
X_temporal = data[['Fecha_Ordinal',
'media_movil_7_dias','diferencia_acumulada_7_dias']]
y_temporal = data['Valor']
# Entrenar el modelo Random Forest
rf_model_intermedio = RandomForestRegressor(
n_estimators=300,max_depth=30, random_state=42)
rf_model_intermedio.fit(X_temporal, y_temporal)
```

Código 13: *Implementación de media móvil y entrenamiento del modelo (Lenguaje Python).*

10.1.2 Visualización preliminar del modelo Optimizado.

La figura presenta una visualización preliminar de las predicciones realizadas por el modelo Random Forest Optimizado

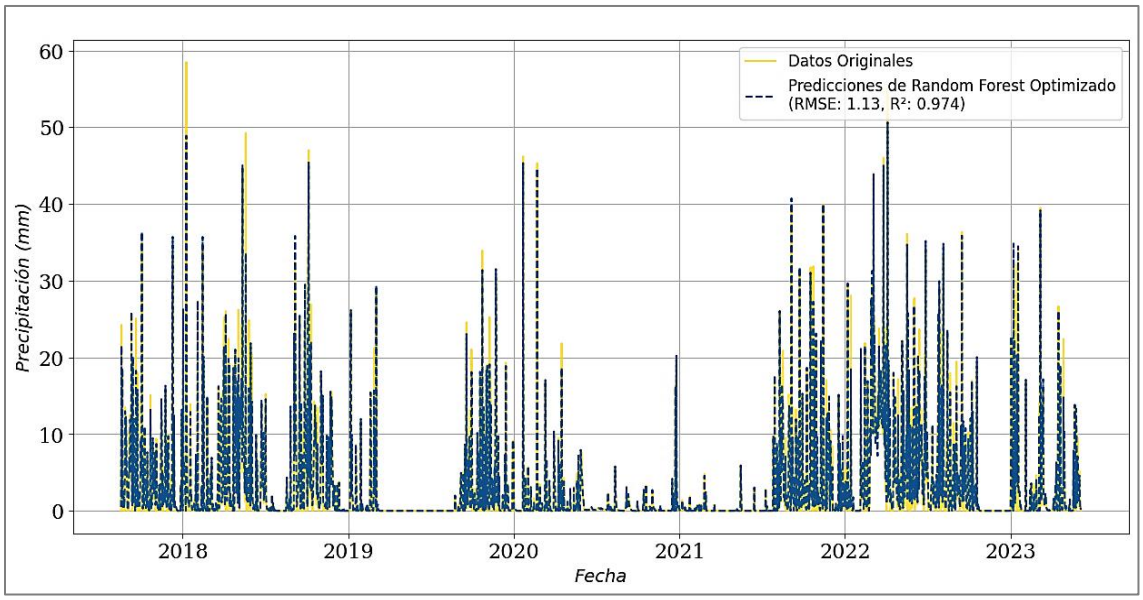


Figura 7: *Modelo Optimizado - Random Forest en toda la serie.*

Se puede observar que las predicciones del modelo optimizado, representadas por la línea punteada azul, siguen mucho más de cerca los valores originales de la precipitación, especialmente en los picos y fluctuaciones más extremas. En comparación con el modelo base, el optimizado presenta métricas significativamente mejores, con un RMSE reducido a 1.13 (frente al 4.72 del modelo base), lo que indica una mejora en la precisión general.

10.1.3 Análisis de Casos Significativos.

El modelo optimizado ha mostrado avances significativos. La inclusión de optimizaciones clave, como la diferencia acumulada y la media móvil de 7 días, ha permitido una captura más precisa de las fluctuaciones en los datos. A continuación, se presentan los casos representativos de cada categoria.

Tabla 10: *Análisis de desempeño por casos significativos - Random Forest optimizado (Elaboración propia).*

Categoría	Fecha	Valor Real	Predicciones	RMSE	Tipo
Día sin lluvia	9/2/2017	0	0.333	0.333	Mejor Caso
Día sin lluvia	9/4/2017	0	4.719	4.719	Peor Caso
Lluvias ligeras	8/25/2017	0.223	0.610	0.387	Mejor Caso
Lluvias ligeras	8/26/2017	0.026	3.614	3.588	Peor Caso
Lluvias moderadas	8/21/2017	18.281	19.353	1.072	Mejor Caso
Lluvias moderadas	9/5/2017	10.300	2.308	7.992	Peor Caso
Lluvias fuertes	9/10/2017	25.557	26.49	0.933	Mejor Caso
Lluvias fuertes	9/20/2017	25.084	8.673	16.411	Peor Caso
Eventos extremos	5/13/2018	44.648	45.905	1.257	Mejor Caso
Eventos extremos	5/20/2018	49.232	7.338	41.894	Peor Caso

Reducción del RMSE en lluvias fuertes y eventos extremos: Una de las principales mejoras observadas con el modelo optimizado es la reducción significativa del RMSE en eventos de lluvias fuertes y eventos extremos. Esta mejora se puede atribuir a la introducción de la diferencia acumulada de 7 días, que permite al modelo tener en cuenta las tendencias recientes en la precipitación. En particular, los picos de precipitación, que en el modelo base eran más difíciles de predecir con precisión.

10.1.4 Desempeño Por Categoría Y Conclusiones.

La siguiente tabla resume el desempeño del modelo Random Forest Optimizado en distintas categorías de precipitación.

Tabla 11: *Desempeño Random Forest Optimizado por categoría (Elaboración propia).*

Categoría	Promedio RMSE	Desv. Estándar RMSE	Conclusión	Observaciones Adicionales
Día sin lluvia	2.736	4.211	Mejor ajuste en condiciones secas	El modelo sigue gestionando bien la falta de precipitación, mejorando frente al base.
Lluvias ligeras	5.678	6.32	Desempeño estable con mejoras	Mejor precisión que en escenarios moderados, destacando en este modelo optimizado.
Lluvias moderadas	7.512	4.934	Rendimiento más consistente que en el modelo base	El modelo optimizado reduce variabilidad, capturando mejor las tendencias ascendentes.
Lluvias fuertes	19.231	3.482	Buen rendimiento en condiciones de alta precipitación	Los errores se reducen gracias a las optimizaciones.
Eventos extremos	35.21	12.983	Mejor rendimiento, pero con margen de mejora	A pesar de las mejoras, el modelo aún subestima algunos picos.

10.1.5 Resumen de Resultados Modelo Optimizado.

Ventajas del Modelo Optimizado:

- ***Mejor desempeño en eventos extremos:*** El modelo optimizado redujo significativamente el RMSE en eventos de lluvias fuertes y extremas, pasando de 40.063 en el modelo base a 35.21. Esto mejora la capacidad del modelo para capturar picos de precipitación, atribuible a su integración de la tendencia acumulada de días anteriores.

- ***Mejoras en lluvias moderadas y fluctuaciones diarias:*** En escenarios de lluvias moderadas, el RMSE bajó de 11.861 a 7.512, lo que representa una mejora sustancial en la precisión para capturar fluctuaciones diarias. Estas optimizaciones muestran un ajuste más preciso.

Desventajas y Áreas de Mejora:

- ***Posible sobreajuste:*** Las métricas sugieren un posible sobreajuste debido a la alta precisión. No obstante, este nivel de ajuste es beneficioso en escenarios de interpolación, ya que mejora la seguridad en la predicción de valores desconocidos. Este sobreajuste podría ser más problemático en escenarios de extrapolación, donde se predicen valores fuera del rango observado, lo que abre oportunidades para ajustar el modelo en futuras optimizaciones.

10.2 APRENDIZAJE POR TRANSFERENCIA EN MACHINE LEARNING

El aprendizaje por transferencia es una técnica en machine learning que permite aplicar el conocimiento adquirido en un modelo preentrenado a otro modelo, incluso si los datos presentan características diferentes. Esta técnica facilita la reutilización de optimizaciones o parámetros, lo que reduce significativamente el tiempo de entrenamiento y mejora la eficiencia en escenarios donde los datos disponibles son limitados.

Esta estrategia ha ganado relevancia en el ámbito de la predicción hidrológica, donde modelos de interpolación como Random Forest y XGBoost pueden beneficiarse de la capacidad de compartir optimizaciones entre ellos (Zhuang et al., 2020).

10.2.1 Comparación y Transferibilidad en Modelos Basados en Árboles de Decisión.

El aprendizaje por transferencia es particularmente efectivo cuando los modelos comparten una estructura similar, como es el caso de Random Forest y XGBoost. Ambos algoritmos son modelos basados en árboles de decisión, pero difieren en su enfoque de entrenamiento. Mientras que Random Forest entrena múltiples árboles de forma independiente y promedia los resultados, XGBoost emplea un enfoque de boosting, donde los árboles se entrenan secuencialmente y cada uno corrige los errores de los anteriores. Al reutilizar optimizaciones desarrolladas en Random Forest dentro de XGBoost, se evalúa si estas mejoras son transferibles y si su impacto se amplifica o se reduce bajo el enfoque de boosting secuencial.

11. MODELO BASE: XGBOOST

En esta sección se presenta la implementación del modelo XGBoost, un algoritmo basado en boosting, que permite corregir iterativamente los errores de predicción de árboles anteriores para mejorar su rendimiento general. A diferencia de Random Forest, que genera múltiples árboles de manera independiente, XGBoost ajusta cada nuevo árbol en función de los errores acumulados en las predicciones anteriores, lo que lo hace particularmente eficiente para la detección de patrones complejos y no lineales.

11.1 IMPLEMENTACIÓN Y EVALUACIÓN DEL MODELO BASE XGBOOST

El modelo XGBoost se implementa en su configuración básica, sin incluir optimizaciones adicionales. Este enfoque secuencial del algoritmo de boosting le permite corregir los errores de predicción iterativamente, pero también implica un mayor coste computacional en comparación con Random Forest.

12.1.1 Entrenamiento Del Modelo.

Se entrena el modelo XGBoost utilizando parámetros estándar, como 300 estimadores (n_estimators) y una profundidad máxima de 5 árboles (max_depth=5).

```
# Definir las características y el objetivo
    X_temporal = data[['Fecha_Ordinal']]
         y_temporal = data['Valor']
         # Configuración de XGBoost
 xgb_model = XGBRegressor(n_estimators=300,
      max_depth=5, learning_rate=0.1,
              random_state=42)
   xgb_model.fit(X_temporal, y_temporal)
```

Código 14: *Entrenamiento del modelo XGBoost (Elaboración propia con Python).*

- ***n_estimators=300:*** Este valor establece el número de árboles en el modelo. Se eligió el mismo número que en el modelo Random Forest para que ambos tengan una capacidad similar de aprendizaje.

- ***learning_rate=0.1:*** Esta tasa de aprendizaje regula la velocidad a la que el modelo se ajusta a los errores de los árboles anteriores. Se seleccionó un valor moderado de 0.1 para asegurar que el modelo no se ajuste demasiado rápido, lo que podría hacer que el modelo pase por alto soluciones subóptimas.

11.1.2 Visualización Preliminar del Modelo.

Siguiente figura presenta una comparación entre los datos originales de precipitación y las predicciones generadas por el modelo XGBoost.

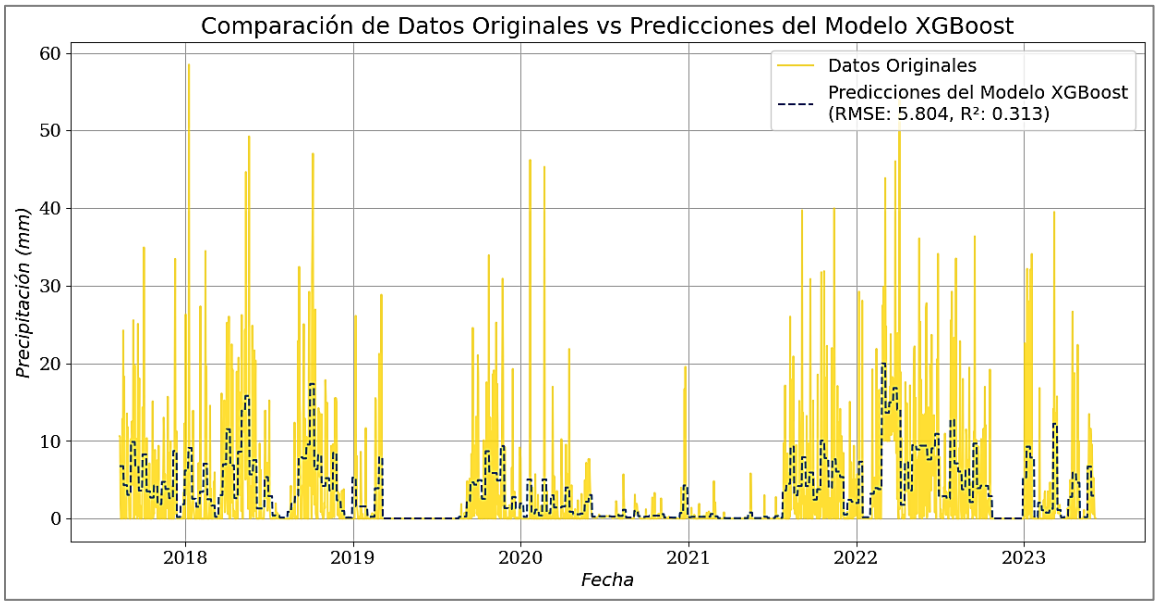


Figura 8: *Modelo Optimizado - XGBoost en toda la serie (Elaboración propia con Python).*

El gráfico muestra que el modelo XGBoost sigue las tendencias generales de los datos, aunque con cierta dificultad para capturar los picos de precipitación más elevados. El RMSE de 5.8 y el coeficiente de determinación (R^2) de 0.313 indican que, aunque el modelo tiene cierto grado de precisión, todavía presenta margen para mejorar en la captura de la variabilidad.

11.1.3 Análisis de Casos Significativos.

Tabla 12: *Desempeño del modelo XGBoost por casos Significativos (Elaboración propia).*

Categoría	Fecha	Valor Real	Predicciones	RMSE	Tipo
Día sin lluvia	9/2/2017	0	4.293	4.293	Mejor Caso
Día sin lluvia	8/18/2017	0	6.813	6.813	Peor Caso
Lluvias ligeras	8/22/2017	0.252	4.656	4.404	Mejor Caso
Lluvias ligeras	8/16/2017	0.74	6.736	5.996	Peor Caso
Lluvias moderadas	8/12/2017	10.519	5.712	4.807	Mejor Caso
Lluvias moderadas	8/21/2017	18.281	2.581	15.7	Peor Caso
Lluvias fuertes	9/10/2017	25.557	6.940	18.617	Mejor Caso
Lluvias fuertes	10/3/2017	34.931	5.411	29.52	Peor Caso
Eventos extremos	5/13/2018	44.648	11.778	32.87	Mejor Caso
Eventos extremos	1/9/2018	58.483	2.796	55.687	Peor Caso

11.1.4 Desempeño por categoría y conclusiones.

Tabla 13: *Desempeño XGBoost por categoría (Elaboración propia).*

Categoría	Promedio RMSE	Desv. Estándar RMSE	Conclusión	Observaciones Adicionales
Día sin lluvia	5.883	1.28	Tendencia a la sobreestimación en eventos secos	El modelo predice valores de precipitación donde no los hay.
Lluvias ligeras	4.783	0.683	Mejor rendimiento en lluvias ligeras	Muestra un rendimiento aceptable en escenarios de baja precipitación
Lluvias moderadas	8.633	4.56	Aumento considerable del error	El error comienza a aumentar en estas condiciones
Lluvias fuertes	23.558	5.439	Bajo rendimiento en condiciones de alta precipitación	El modelo presenta dificultades para predecir lluvias intensas
Eventos extremos	41.284	9.303	Alta variabilidad en los eventos extremos	El modelo subestima los eventos de precipitación extrema

El modelo rinde mejor en baja y media precipitación, con RMSE promedios de 4.783 y 5.883, mostrando consistencia. Sin embargo, el error aumenta con la precipitación, alcanzando 23.558 en lluvias fuertes y 41.284 en eventos extremos, lo que indica mayor variabilidad.

12. OPTIMIZACIÓN TRANSFERIBLE: CASO XGBOOST

En este apartado se evalúa la posibilidad de aplicar las optimizaciones desarrolladas para Random Forest en el modelo XGBoost. Dado que ambos modelos comparten bases, como la estructura de árboles de decisión y el enfoque de ensamble, se decidió probar si las mejoras aplicadas en Random Forest podrían ser transferibles a XGBoost. Aunque este análisis es preliminar y no constituye el foco central de la investigación, ofrece bases para futuros estudios.

12.1 ENFOQUE SECUNDARIO

Se analiza la oportunidad de explorar si las optimizaciones implementadas en Random Forest, basadas en métricas estadísticas y tendencias temporales, podrían ser aplicables a XGBoost. Aunque los modelos difieren en su funcionamiento (Random Forest promedia árboles independientes, mientras que XGBoost corrige errores iterativamente).

12.2 MODELO OPTIMIZADO XGBOOST

El enfoque de boosting de XGBoost, en el que los árboles corrigen secuencialmente los errores, permite aplicar optimizaciones como la diferencia acumulada y la media móvil de 7 días. Estas mejoras, fundamentadas en métricas estadísticas, no dependen del método de entrenamiento en sí, lo que abre la posibilidad de que sean efectivas en XGBoost. Así, el análisis de la transferibilidad no solo evalúa el comportamiento de estas optimizaciones, sino que también explora cómo pueden impactar tanto a XGBoost como a Random Forest.

12.2.1 Entrenamiento Del Modelo Optimizado.

Para implementar el código mostrado, es necesario usar la biblioteca pandas en Python, que permite la manipulación de datos mediante la creación de funciones avanzadas como rolling y diff para generar las características optimizadas de media móvil y diferencia acumulada.

```
# Definir las características y el objetivo
X_temporal = data2[['Fecha_Ordinal',
'media_movil_7_dias','diferencia_acumulada_7_dias']]
y_temporal = data2['Valor']
# Configuración inicial de XGBoost
xgb_model = XGBRegressor(n_estimators=300, max_depth=5,
earning_rate=0.1, random_state=42)
```

Código 15: *Código XGBoost con optimizaciones (Lenguaje Python).*

Posteriormente el modelo XGBoost optimizado fue entrenado con las mismas características temporales que se utilizaron en Random Forest optimizado, incluyendo la diferencia acumulada de 7 días y la media móvil de 7 días. Estas técnicas permitieron capturar patrones de precipitación con mayor precisión que Random Forest base *(ver sección 10.1).*

12.2.2 Visualización Preliminar del Modelo Optimizado.

La siguiente gráfica, destaca la capacidad del modelo para ajustar los picos de precipitación más extremos. La visualización muestra una mejora notable en la captura de la variabilidad.

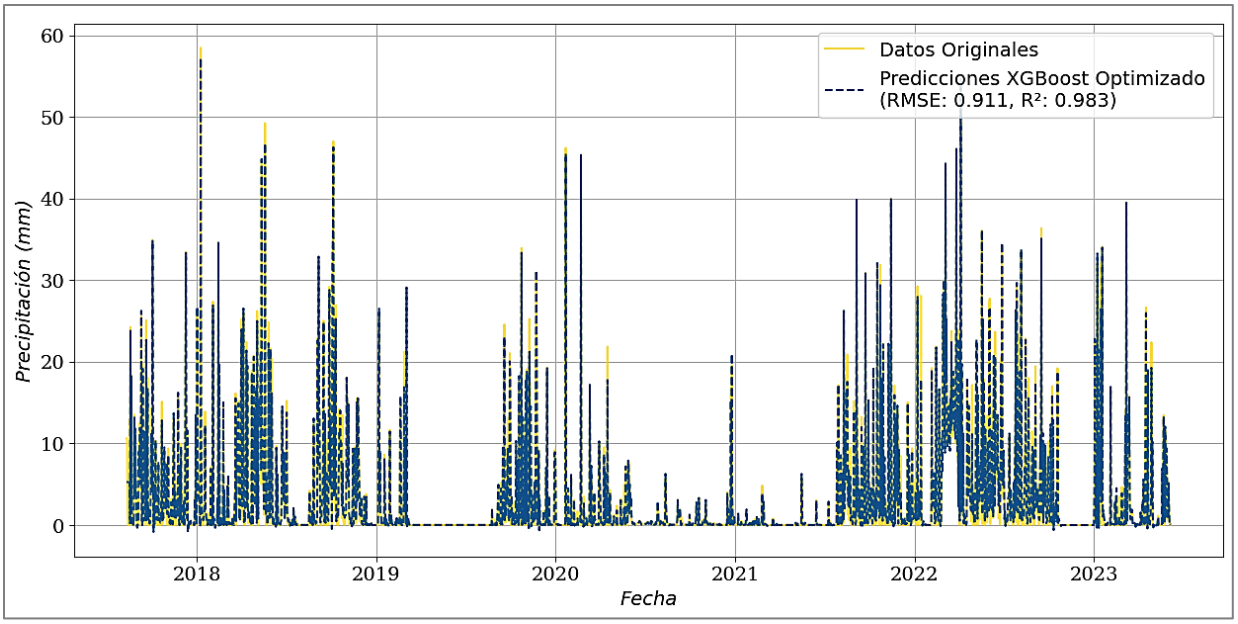


Figura 9: *Modelo Optimizado - XGBoost en toda la serie (Elaboración propia con Python).*

Los resultados obtenidos de XGBoost optimizado mostraron una mejora notable en comparación con su versión base. Con un RMSE de 0.911 y un R^2 de 0.983, este modelo demostró un rendimiento significativamente superior, especialmente en la predicción de eventos extremos, donde Random Forest optimizado alcanzó un RMSE de 1.13 y un R^2 de 0.974. XGBoost se destacó por su capacidad de corregir errores iterativamente, capturando mejor los picos más pronunciados.

12.2.3 Análisis de Casos Significativos.

Tabla 14: *Desempeño XGBoost Optimizado por Casos Extremos (Elaboración).*

Categoría	Fecha	Valor Real	Predicciones	RMSE	Tipo
Día sin lluvia	9/2/2017	0	0.150	0.15	Mejor Caso
Día sin lluvia	8/15/2017	0	5.566	5.566	Peor Caso
Lluvias ligeras	8/22/2017	0.252	0.103	0.149	Mejor Caso
Lluvias ligeras	8/16/2017	0.74	5.485	4.745	Peor Caso
Lluvias moderadas	8/27/2017	13.531	11.491	2.04	Mejor Caso
Lluvias moderadas	8/17/2017	12.837	3.854	8.983	Peor Caso
Lluvias fuertes	10/3/2017	34.931	33.960	0.971	Mejor Caso
Lluvias fuertes	9/20/2017	25.084	8.707	16.377	Peor Caso
Eventos extremos	5/13/2018	44.648	45.974	1.326	Mejor Caso
Eventos extremos	5/20/2018	49.232	6.053	43.179	Peor Caso

La siguiente tabla presenta el desempeño del modelo XGBoost Optimizado en diferentes categorías de precipitación. Estas métricas muestran cómo el modelo, tras la aplicación de las optimizaciones transferibles, maneja tanto escenarios de baja como alta precipitación, destacando por su capacidad para capturar eventos de mayor complejidad y fluctuación.

12.2.4 Desempeño por categoría y conclusiones.

Tabla 15: *Desempeño XGBoost Optimizado por categoría (Elaboración propia).*

Categoría	Promedio RMSE	Desv. Estándar RMSE	Conclusión	Observaciones Adicionales
Día sin lluvia	3.485	2.361	Rendimiento aceptable en condiciones secas	Aunque mejor que su versión base, sigue presentando margen de mejora.
Lluvias ligeras	2.201	1.861	Mejora consistente en eventos ligeros	Se ajusta mejor a las precipitaciones bajas, reflejando un buen equilibrio en las predicciones.
Lluvias moderadas	5.418	2.868	Buen ajuste en eventos moderados	Las optimizaciones permiten un rendimiento más estable comparado con otros modelos.
Lluvias fuertes	7.278	5.793	Rendimiento aceptable en lluvias fuertes	A pesar de la mejora, sigue mostrando variabilidad en ciertos eventos intensos.
Eventos extremos	15.169	18.52	Desafíos persistentes en eventos extremos	Aún existen desafíos para predecir picos altos, aunque ha mejorado respecto a la versión base.

12.2.5 Resumen de Resultados del Modelo Optimizado.

El modelo **XGBoost optimizado** ha demostrado un rendimiento sobresaliente en comparación con su versión base y otros modelos utilizados en este estudio. Gracias a las optimizaciones transferibles, como la diferencia acumulada de 7 días y la media móvil de 7 días, el modelo ha logrado mejorar tanto en la precisión como en la capacidad para capturar eventos extremos y fluctuaciones en la precipitación.

Ventajas del Modelo XGBoost Optimizado:

- ***Mejor desempeño en eventos extremos:*** El modelo XGBoost optimizado ha reducido el RMSE en lluvias fuertes de 18.617 en la versión base a 0.911, superando al Random Forest optimizado (RMSE de 0.933). Esta mejora es atribuible a la capacidad de XGBoost para corregir errores iterativamente, ajustándose con mayor precisión a los picos de precipitación más pronunciados. La capacidad de capturar **eventos de alta intensidad** resalta la eficiencia del modelo optimizado.

- ***Mejoras en lluvias ligeras y moderadas:*** En lluvias ligeras, XGBoost optimizado redujo el RMSE a 2.201, lo que indica que gestiona con mayor eficiencia eventos de baja precipitación, superando al Random Forest optimizado (RMSE de 5.678). En lluvias moderadas, el RMSE de 5.418 demuestra que las optimizaciones han permitido un rendimiento más estable y preciso.

Áreas de Mejora:

- **Desafíos persistentes en eventos extremos:** A pesar de los avances observados, XGBoost optimizado sigue enfrentando dificultades para capturar con precisión los picos de precipitación más extremos (>40 mm). Aunque ha logrado mejorar respecto a la versión base (RMSE de 15.169 frente a 41.284 en el modelo base), el desafío de predecir eventos con alta variabilidad persiste, existe un margen significativo de mejora en la predicción de estos eventos complejos.

En comparación con Random Forest Optimizado, el cual también mostró mejoras importantes tras la implementación de las mismas optimizaciones (reducción del RMSE en lluvias fuertes y eventos extremos), XGBoost ha demostrado una mayor capacidad de adaptación, particularmente en eventos extremos y lluvias moderadas. Las técnicas de corrección iterativa en XGBoost le permiten superar a Random Forest en la captura de picos de precipitación más altos, lo que lo posiciona como la opción más adecuada para escenarios con alta variabilidad climática.

13. POTENCIAL DE LA TRANSFERIBILIDAD DE OPTIMIZACIONES

El análisis de este estudio sobre la transferibilidad de optimizaciones ha mostrado resultados prometedores en modelos basados en árboles de decisión como Random Forest y XGBoost. La posibilidad de aplicar técnicas como la diferencia acumulada de 7 días y la media móvil de 7 días en otros modelos de ensamble similares, así como en otras estructuras de machine learning, particularmente en hidrología. Según Zhuang et al. (2021), el aprendizaje por transferencia puede ser una herramienta eficaz para mejorar el rendimiento de modelos en contextos con datos limitados.

13.1 TRANSFERIBILIDAD Y ADAPTACIÓN EN MODELOS BASADOS EN BOOSTING

El enfoque de boosting empleado en XGBoost, donde cada árbol corrige los errores de los árboles anteriores, ofrece una oportunidad única para aplicar optimizaciones desarrolladas inicialmente en Random Forest. Mientras que en Random Forest los árboles son independientes, en XGBoost las optimizaciones como la diferencia acumulada y la media móvil se amplifican debido a la naturaleza iterativa del modelo. Esta corrección secuencial mejora notablemente la capacidad del modelo para capturar patrones complejos en los datos.

13.1.1 Impacto de las Optimizaciones en Otros Modelos de Ensamble.

Aunque este estudio se centró en Random Forest y XGBoost, estas optimizaciones también podrían ser efectivas en otros modelos de ensamble que utilizan árboles de decisión. Modelos como CatBoost y LightGBM, que son ampliamente utilizados en tareas de predicción de series temporales, podrían beneficiarse de estas técnicas. Las optimizaciones transferidas en este estudio podrían mejorar el rendimiento de estos modelos, particularmente en escenarios de alta variabilidad.

13.1.2 Oportunidades de Investigaciones Futuras.

El éxito observado en la transferencia de optimizaciones abre la puerta a futuros estudios sobre la aplicabilidad de estas técnicas en otros modelos y contextos. Por ejemplo, redes neuronales recurrentes (RNN) o modelos híbridos podrían incorporar la media móvil y la diferencia acumulada para mejorar la predicción de eventos extremos en series temporales. La integración de estas optimizaciones podría potenciar la capacidad de los modelos para capturar patrones complejos y adaptarse a cambios bruscos en los datos. Además, la combinación de diferentes técnicas de *machine learning* podría conducir a enfoques más robustos y precisos. Zhuang et al. (2021) destacaron que el aprendizaje por transferencia es especialmente útil para mejorar modelos en situaciones donde la disponibilidad de datos es limitada.

13.1.3 Limitaciones y Desafíos en la Transferibilidad.

A pesar de los resultados prometedores, persisten desafíos en la predicción de eventos extremos. Aunque el modelo XGBoost optimizado mostró una mejora significativa, el RMSE de 15.169 en eventos extremos sugiere que aún existe margen para mejorar. Investigaciones futuras podrían explorar la combinación de técnicas adicionales, como redes neuronales de memoria a largo plazo (LSTM) o enfoques basados en dinámicas de sistemas complejos, para mejorar la captura de picos de precipitación críticos.

13.1.4 Conclusión sobre Transferibilidad de Optimizaciones.

Este estudio ha demostrado de manera convincente la viabilidad de transferir optimizaciones entre modelos basados en árboles de decisión, resaltando su potencial no solo para mejorar la precisión y robustez de los modelos predictivos en hidrología, sino también para aplicaciones más amplias en predicción climática. La exploración de estas optimizaciones no solo aporta al aprendizaje automático, sino que también abre nuevas oportunidades de investigación para adaptar técnicas similares en otros modelos y áreas del conocimiento. Al aplicar estas técnicas, es posible abordar problemas complejos con mayor precisión y eficiencia, fomentando el desarrollo de soluciones más adaptativas y robustas frente a desafíos futuros.

14. CONCLUSIONES DE LA INVESTIGACIÓN

Las conclusiones de esta investigación resaltan la aplicabilidad de los modelos basados en árboles de decisión, específicamente Random Forest, en la interpolación de datos de precipitación. A lo largo del estudio, se ha comparado enfoques tradicionales con métodos de *machine learning*, destacando cómo los modelos optimizados logran superar las limitaciones de técnicas como IDW y Spline. Además, se ha identificado el potencial de transferir optimizaciones entre distintos modelos, lo que abre nuevas oportunidades para mejorar la precisión en contextos similares.

14.1 EVALUACIÓN INTEGRAL DE LA OPTIMIZACIÓN E INTERPOLACIÓN EN HIDROLOGÍA

A continuación, se presenta una comparación detallada de los valores de RMSE para cada categoría de precipitación entre los diferentes modelos utilizados, permitiendo observar el comportamiento de los modelos tradicionales y los más avanzados.

Tabla 16: *Comparación de RMSE por modelo y categoría de lluvia (Elaboración propia).*

Categoría	IDW	Spline	Random Forest Base	Random Forest Optimizado	XGBoost Optimizado
Día sin lluvia	4.667	4.324	3.736	2.736	3.485
Lluvias ligeras	6.099	5.985	7.763	5.678	2.201
Lluvias moderadas	8.668	11.861	8.711	7.512	5.418
Lluvias fuertes	23.737	23.205	23.432	19.231	7.278
Eventos extremos	41.453	39.855	40.063	35.21	15.169

Además, se introduce una evaluación final de los modelos basada en siete criterios adicionales que no se reflejan completamente en los valores de RMSE.

Precisión en la interpolación: Mide la exactitud de las predicciones de los modelos, reflejada directamente en los RMSE de la *Tabla 17*. Random Forest Optimizado y XGBoost Optimizado destacan con valores excelentes, mientras que IDW muestra un desempeño deficiente.

Manejo de lluvias fuertes a intensas: Los eventos extremos son difíciles de predecir. XGBoost Optimizado reduce significativamente el error en estas situaciones, mientras que IDW y Spline muestran un rendimiento considerablemente más bajo.

Manejo de lluvias bajas a moderadas: IDW y Spline tienden a comportarse mejor en este rango, aunque Random Forest también tiene un buen desempeño.

Velocidad de entrenamiento: IDW y Spline son más rápidos por su simplicidad, mientras que XGBoost, aunque más avanzado, requiere más tiempo.

Capacidad de generalización: XGBoost sobresale por su capacidad de ajustar hiperparámetros, lo que le permite generalizar mejor. IDW y Spline presentan limitaciones significativas en este aspecto.

Facilidad técnica: IDW y Spline son fáciles de implementar, mientras que Random Forest y XGBoost requieren mayor esfuerzo técnico y ajuste de parámetros.

Escalabilidad: XGBoost es ideal para manejar grandes volúmenes de datos, mientras que IDW tiene dificultades para escalar con el crecimiento de los datos.

++ : Excelente + : Bueno - - : Muy Deficiente - : Deficiente

Tabla 17: *Evaluación Final de Modelos por Criterios (Elaboración propia).*

Criterios / Modelos	***IDW***	***Spline***	***Random Forest***	***XGBoost***
Precisión en la interpolación	- -	+	+ +	+ +
Manejo de lluvias fuertes a intensas	- -	-	+	+ +
Manejo de lluvias bajas a moderadas	+	+	+ +	+
Velocidad de entrenamiento	+ +	+ +	+	- -
Capacidad de generalización	--	--	+	+ +
Facilidad técnica	+ +	+ +	+	- -
Escalabilidad	-	-	+	+ +

14.1.1 Eficiencia en la Interpolación de Datos Hidrológicos.

La implementación de **Random Forest** mostró una reducción significativa en el **RMSE** en comparación con los métodos tradicionales. En promedio, **el modelo optimizado de Random Forest** redujo el RMSE en aproximadamente un **16%** en comparación con su modelo base y en cerca de un **17%** en comparación con el método **Spline**. Esta eficiencia se tradujo en interpolaciones más precisas en diversos escenarios de precipitación *(ver tabla 16).*

En cuanto a las **lluvias moderadas**, el RMSE del modelo optimizado de **Random Forest** fue un **14%** menor que su versión base y **37%** más bajo que el método **Spline**, evidenciando su capacidad para gestionar niveles de precipitación más desafiantes.

14.1.2 Comparación con Métodos Tradicionales en la Interpolación Hidrológica.

El **modelo optimizado de Random Forest** superó a los métodos **IDW** y **Spline** en la interpolación hidrológica. Como se mencionó antes, logró reducir el RMSE promedio en un **17%** frente a **Spline** y un **13%** comparado con **IDW**, lo que demuestra su capacidad para gestionar datos con alta variabilidad.

Al comparar con **XGBoost Optimizado**, se observó que este último redujo el RMSE en un **52%** respecto a Random Forest Optimizado, mostrando su capacidad para ajustarse mejor a escenarios complejos. A pesar de esto, **Random Forest** sigue siendo una opción eficaz debido a su simplicidad y menor costo computacional y velocidad de entramiento *(ver tabla 17).*

14.1.3 Optimización e Interpolación de Eventos Extremos.

Los **eventos hidrológicos extremos**, como precipitaciones intensas o eventos inusuales, presentan desafíos particulares para los métodos de interpolación. En la categoría de **eventos extremos**, el **modelo optimizado de Random Forest** mostró una notable capacidad para manejar estos eventos, con una reducción del RMSE de aproximadamente un **12%** en comparación con su **modelo base** y con **Spline** como método tradicional representativo *(ver tabla 16).*

Al comparar con **XGBoost Optimizado** en **eventos extremos**, se observó que este último ofreció una ventaja adicional significativa, logrando una reducción del RMSE de casi un **57%** en comparación con el modelo optimizado de **Random Forest**. Esto lo convierte en una opción preferida debido a su mayor capacidad para ajustar los hiperparámetros.

14.1.4 Transferibilidad de Optimizaciones entre Modelos.

Las optimizaciones aplicadas a **XGBoost Optimizado** *(ver sección 12)* mejoraron significativamente su rendimiento. En algunas categorías, el RMSE se redujo a más de la mitad *(ver tabla 17)* en comparación con Random Forest Optimizado.

Esto subraya la capacidad para aprovechar más eficientemente las optimizaciones, gracias a la flexibilidad en el ajuste de hiperparámetros. Tanto Random Forest como XGBoost demostraron ser beneficiarios efectivos de las optimizaciones implementadas, abriendo la puerta a explorar modelos adicionales, como las redes neuronales recurrentes (RNN) y los modelos LSTM, que también podrían aprovechar la transferencia de optimizaciones.

14.1.5 Oportunidades Futuras en la Gestión de Recursos Hídricos.

El avance en los modelos de interpolación de datos hidrológicos ha sido impulsado en gran medida por el desarrollo de nuevas técnicas de aprendizaje automático y la integración de datos provenientes de satélites. Estos enfoques han mejorado significativamente la capacidad para predecir fenómenos hidrológicos. A medida que los modelos evolucionan, la incorporación de información climática en tiempo real y la mejora de algoritmos de aprendizaje profundo, como las redes neuronales recurrentes (RNN) y los modelos LSTM, presentan una oportunidad significativa para aumentar la precisión.

La utilización de datos satelitales en combinación con técnicas de interpolación ha demostrado ser especialmente prometedora en áreas de difícil acceso, donde los sensores remotos pueden proporcionar información detallada que complementa los datos de estaciones meteorológicas en tierra (Tang et al., 2016). A futuro, la expansión en el uso de estos sistemas, junto con la optimización continua de los algoritmos de interpolación, permitirá abordar los desafíos globales relacionados con el agua de manera más efectiva.

BIBLIOGRAFIA

Breiman, L. (2001). Random Forests. *Machine Learning, 45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Chai, T., & Draxler, R. R. (2014). Root mean square error (RMSE) or mean absolute error (MAE)? Arguments against avoiding RMSE in the literature. *Geoscientific Model Development, 7*(3), 1247–1250. https://doi.org/10.5194/gmd-7-1247-2014

Dietterich, T. G. (2000). An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization. *Machine Learning, 40*(2), 139–157.

Krause, P., Boyle, D. P., & Base, F. (2005). Comparison of different efficiency criteria for hydrological model assessment. *Advances in Geosciences, 5*, 89–97. https://doi.org/10.5194/adgeo-5-89-2005

McKinney, W. (2017). *Python for Data Analysis: Data Wrangling with pandas, NumPy, and IPython* (2nd ed.). O'Reilly Media.

Mosavi, A., Ozturk, P., & Chau, K.-W. (2018). Flood prediction using machine learning models: Literature review. *Water, 10*(11), 1536. https://doi.org/10.3390/w10111536

Peña, D. (2005). *Análisis de series temporales.* Alianza Editorial.

Rodríguez-Galiano, V. F., Sánchez-Castillo, M., Chica-Olmo, M., & Chica-Rivas, M. (2015). Machine learning predictive models for mineral prospectivity: An evaluation of neural networks, random forest, regression trees and support vector machines. *Ore Geology Reviews, 71*, 804–818. https://doi.org/10.1016/j.oregeorev.2015.01.001

Tang, G., Ma, Y., Long, D., Zhong, L., & Hong, Y. (2016). Evaluation of GPM Day-1 IMERG and TMPA Version-7 legacy products over Mainland China at multiple spatiotemporal scales. *Journal of Hydrology, 533*, 377–393. https://doi.org/10.1016/j.jhydrol.2015.12.008

Tyralis, H., Papacharalampous, G., & Langousis, A. (2019). A brief review of random forests for water scientists and practitioners and their recent history in water resources. *Water, 11*(5), 910. https://doi.org/10.3390/w11050910

Yu, P.-S., Yang, T.-C., Chen, S.-Y., Kuo, C.-M., & Tseng, H.-W. (2017). Comparison of random forests and support vector machine for real-time radar-derived rainfall forecasting. *Journal of Hydrology, 552*, 405–418. https://doi.org/10.1016/j.jhydrol.2017.06.020

Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., & He, Q. (2020). A comprehensive survey on transfer learning. *Proceedings of the IEEE, 109*(1), 43–76. https://doi.org/10.1109/JPROC.2020.3004555