Modelo predictivo para estimar la humedad del suelo en cultivos del CIAT usando técnicas de aprendizaje automático

Juliana Maritza Zarate Jiménez
Código 9014778
Fabio Andrés Paternina Miranda
Código 9013839

Proyecto Aplicado para optar al título de
Magister en Ciencia de Datos

Directora
Gloria Álvarez Vargas
Codirector
Camilo Barrios Pérez

Pontificia Universidad Javeriana Cali
Facultad de ingeniería y ciencias
Maestría en ciencia de datos
Santiago de Cali
21 noviembre de 2025

FICHA RESUMEN TRABAJO DE GRADO DE MAESTRÍA

TÍTULO DEL PROYECTO: "MODELO PREDICTIVO PARA ESTIMAR LA HUMEDAD DEL SUELO EN CULTIVOS DEL CIAT USANDO TÉCNICAS DE APRENDIZAJE AUTOMÁTICO."

ÁREA DE TRABAJO: Sistemas y Computación
TIPO DE PROYECTO (Aplicado, Innovación, Investigación): Aplicado
ESTUDIANTE(S): Juliana Maritza Zarate Jiménez y Fabio Andrés Paternina Miranda
CORREO ELECTRÓNICO: julianazarate@javerianacali.edu.co, fabiopaternina@javerianacali.edu.co.
DIRECCIÓN Y TELEFONO: Km 17, Recta Cali-Palmira, Valle del Cauca
DIRECTOR: Gloria Álvarez Vargas
VINCULACIÓN DEL DIRECTOR: Planta
CORREO ELECTRÓNICO DEL DIRECTOR: galvarez@javerianacali.edu.co
CO-DIRECTOR (Si aplica): Camilo Barrios Pérez
GRUPO O EMPRESA QUE LO AVALA (Si aplica): Centro Internacional de Agricultura Tropical
OTROS GRUPOS O EMPRESAS:
PALABRAS CLAVE (al menos 5): Monitoreo, ciencia de datos, agricultura, cultivos, humedad.
FECHA DE INICIO: 25/11/2024
RESUMEN:

El presente trabajo desarrolló un modelo predictivo para la estimación de la humedad volumétrica del suelo a partir de la integración de variables espectrales, climáticas y edáficas, empleando técnicas avanzadas de aprendizaje automático y análisis multifuente. El estudio se realizó en parcelas experimentales del Centro Internacional de Agricultura Tropical (CIAT), utilizando datos provenientes de sensores de humedad del suelo, imágenes satelitales PlanetScope (índices NDVI, EVI, NDMI y NDWI) y registros meteorológicos locales (precipitación, temperatura, radiación solar, evapotranspiración y velocidad del viento).

El proceso metodológico incluyó un análisis exploratorio para evaluar la calidad y distribución de los datos, identificar correlaciones significativas y eliminar redundancias entre variables. Posteriormente, se seleccionaron nueve variables predictoras finales que representaron de forma eficiente los componentes hidrológicos, energéticos y vegetativos del sistema suelo–planta–atmósfera. Cinco algoritmos fueron evaluados en la fase de modelado: XGBoost, Random Forest, Support Vector Regression (SVR), Multi-Layer Perceptron (MLP) y K-Nearest Neighbors (KNN). Tras un proceso de optimización mediante GridSearchCV y validación cruzada K-Fold (k = 5), el modelo XGBoost optimizado se consolidó como la alternativa más precisa y estable, alcanzando un desempeño sobresaliente (R^2 = 0.96; MAE = 1.95; RMSE = 2.94). Este resultado evidenció su capacidad para capturar relaciones no lineales y manejar la multicolinealidad entre variables, superando a los demás algoritmos en generalización y eficiencia computacional. Como aplicación práctica, se desarrolló una interfaz web interactiva que permite realizar predicciones en tiempo real de la humedad del suelo a partir de datos climáticos y satelitales ingresados por el usuario. La interfaz integra visualizaciones dinámicas y un sistema de clasificación por categorías de humedad (muy baja, baja, media y alta), facilitando la interpretación de los resultados y la toma de decisiones agronómicas.

TABLA DE CONTENIDO

LISTA DE FIGURAS

LISTA DE TABLAS

INTRODUCCIÓN

La agricultura enfrenta desafíos crecientes relacionados con la optimización del uso del agua, el aumento de la productividad y la adaptación a los efectos del cambio climático. En este contexto, la humedad del suelo constituye una variable esencial que regula los procesos de intercambio de agua y energía entre el suelo, la vegetación y la atmósfera. Su monitoreo y gestión eficiente son determinantes para la sostenibilidad de los sistemas agrícolas, especialmente en cultivos de alta demanda hídrica [33].

Los métodos convencionales para medir la humedad del suelo, basados en sensores instalados en campo o en muestreos manuales, presentan limitaciones importantes: cobertura espacial reducida, altos costos de instalación y mantenimiento, y baja representatividad en grandes extensiones de cultivo. Estas restricciones dificultan la toma de decisiones oportunas sobre riego y manejo agronómico, comprometiendo la eficiencia en el uso del recurso hídrico[34][35].

En los últimos años, la convergencia de tecnologías de observación remota y aprendizaje automático ha permitido superar parcialmente dichas limitaciones. La integración de imágenes satelitales multiespectrales, datos climáticos y registros de sensores in situ ha abierto nuevas posibilidades para estimar variables biofísicas del suelo mediante modelos predictivos avanzados. En particular, los algoritmos de aprendizaje supervisado, como *Random Forest*, *redes neuronales, Maquinas de vectores soporte (SVR),* Extreme *Gradient Boosting (XGBoost)*, han demostrado una elevada capacidad para modelar relaciones no lineales entre los índices espectrales de la vegetación y la humedad del suelo[36][37].

En este trabajo se desarrolló un modelo predictivo de aprendizaje supervisado para estimar la humedad volumétrica del suelo en parcelas experimentales del Centro Internacional de Agricultura Tropical (CIAT). El modelo integró información satelital, climática y de sensores terrestres, generando una herramienta robusta que respalda la toma de decisiones en el marco de la agricultura de precisión. Este enfoque permitió optimizar el uso del agua, mejorar la planificación de riego.

Como resultado, se implementó y validó un modelo con desempeño satisfactorio en la predicción de la humedad, evaluado con métricas estandarizadas. Asimismo, se desarrolló una interfaz interactiva de visualización geoespacial que facilita la interpretación de los resultados y promueve el uso eficiente del recurso hídrico, fortaleciendo la gestión sostenible de la agricultura.

1. DEFINICION DEL PROBLEMA

1.1 Planteamiento del problema

La humedad del suelo constituye un parámetro esencial para la productividad agrícola, ya que determina la disponibilidad de agua para las plantas, regula la eficiencia del riego y condiciona la sostenibilidad de los recursos hídricos. Su monitoreo preciso resulta indispensable para optimizar las prácticas agrícolas y garantizar la estabilidad de los sistemas productivos [33]. Sin embargo, los métodos convencionales de medición como sensores de campo o muestreos manuales presentan limitaciones en términos de cobertura espacial, costo y capacidad de adaptación a grandes extensiones de cultivo [34]. Estas restricciones dificultan la toma de decisiones basadas en datos y pueden conducir a riegos ineficientes, desperdicio de agua y estrés hídrico en los cultivos, afectando la productividad y aumentando los costos de producción.

Los sistemas tradicionales de monitoreo, basados en sensores instalados en puntos específicos del terreno, ofrecen alta precisión local pero baja representatividad espacial. En contraste, los métodos derivados de imágenes satelitales o modelos climáticos tienden a tener una resolución espacial gruesa y no capturan adecuadamente las particularidades de cada parcela [34]. Por ello, existe una brecha metodológica entre las técnicas de medición directa y las de observación remota, lo que limita la capacidad para realizar estimaciones continuas y de alta resolución de la humedad del suelo.

En los últimos años, las tecnologías de observación de la Tierra y el aprendizaje automático han mostrado un potencial significativo para abordar esta problemática. Los sensores multiespectrales satelitales permiten calcular índices como el NDVI (*Normalized Difference Vegetation Index*), NDWI (*Normalized Difference Water Index*) y SAVI (*Soil Adjusted Vegetation Index*), los cuales presentan correlaciones con propiedades biofísicas del suelo, incluida la humedad [8,9,36,37]. A su vez, los registros de estaciones meteorológicas proporcionan variables como precipitación, temperatura y radiación solar, factores clave que modulan la dinámica hídrica del suelo [1,39]. No obstante, integrar de forma efectiva estas fuentes de información heterogénea en un modelo predictivo continúa siendo un desafío técnico y metodológico.

La ausencia de herramientas predictivas que combinen datos satelitales, registros climáticos e información de sensores de humedad in situ mediante algoritmos de *machine learning* restringe la capacidad para estimar las condiciones del suelo de manera eficiente y precisa. Esta limitación afecta la posibilidad de implementar prácticas de agricultura de precisión, comprometiendo la sostenibilidad y la resiliencia del sector agrícola frente al cambio climático.

Frente a esta situación, en el presente trabajo se desarrolló un modelo predictivo robusto basado en técnicas de aprendizaje supervisado, orientado a la estimación de la humedad volumétrica del suelo a partir de la integración de fuentes multiespectrales, climáticas y de campo. Este modelo abordó las limitaciones de los métodos tradicionales y constituyó una herramienta práctica y reproducible para la gestión sostenible del recurso hídrico en los sistemas agrícolas del CIAT.

1.2 Formulación del problema

¿Cómo desarrollar un modelo predictivo para estimar la humedad del suelo en cultivos del CIAT usando técnicas de aprendizaje automático?

Sistematización

1. ¿Cómo preparar los datos satelitales, imágenes aéreas e indicadores climáticos para estimar la humedad del suelo?

2. ¿Cómo entrenar distintos tipos de modelos para estimar la humedad del suelo?

3. ¿Cuáles son las métricas de evaluación más adecuadas para analizar el desempeño de los modelos?

4. ¿Cómo visualizar el modelo seleccionado para su interacción?

2. OBJETIVOS

2.1 Objetivo General

Desarrollar un modelo predictivo para estimar la humedad del suelo en cultivos del CIAT usando técnicas de aprendizaje automático.

2.2 Objetivos Específicos

1. Preparar los datos satelitales, imágenes aéreas e indicadores climáticos para estimar la humedad del suelo.
2. Entrenar distintos tipos de modelos para estimar la humedad del suelo.
3. Evaluar el desempeño de los modelos desarrollados con métricas adecuadas.
4. Desarrollar una visualización para interactuar con el modelo de mejor evaluación.

3. JUSTIFICACIÓN

El agua representa uno de los recursos más críticos para la producción agrícola, y su manejo eficiente es fundamental para la sostenibilidad de los sistemas alimentarios. Se estima que más del 70 % del agua dulce utilizada a nivel mundial se destina a actividades agrícolas [33]. En este contexto, el monitoreo de la humedad del suelo se convierte en un elemento clave para optimizar el riego, mejorar la productividad y reducir el impacto ambiental de las prácticas agrícolas[34].

No obstante, las técnicas tradicionales de medición presentan limitaciones de cobertura y costo que impiden una caracterización continua a escala de parcela o de cultivo. En respuesta a este reto, las tecnologías satelitales de observación de la Tierra y los algoritmos de aprendizaje automático ofrecen soluciones innovadoras para estimar parámetros biofísicos con alta precisión espacial y temporal [8][9]. Estas herramientas permiten integrar información heterogénea (espectral, climática y edáfica) en modelos predictivos robustos capaces de generar estimaciones dinámicas y de bajo costo [38].
El valor científico de este trabajo radicó en la construcción de un modelo multifuente capaz de aprender patrones complejos entre variables espectrales, meteorológicas y de campo, contribuyendo así al avance del conocimiento en agricultura digital y modelamiento del suelo. Desde el punto de vista tecnológico, la implementación de una plataforma de visualización permitió transformar los resultados analíticos en una herramienta accesible para la toma de decisiones en campo. Finalmente, desde la perspectiva ambiental y social, el uso eficiente del agua mediante herramientas de predicción y monitoreo apoya las estrategias de adaptación al cambio climático y la sostenibilidad de los sistemas productivos agrícolas en el trópico [9].

En síntesis, el desarrollo de este modelo predictivo integró principios de ciencia de datos, teledetección y agronomía aplicada, aportando una solución replicable y escalable para el monitoreo de la humedad del suelo en diferentes zonas y cultivos, contribuyendo al fortalecimiento de la agricultura de precisión en Colombia.

4. MARCO DE REFERENCIA

4.1 Marco teórico

El monitoreo y manejo eficiente de la humedad del suelo son fundamentales para garantizar la sostenibilidad y productividad de los sistemas agrícolas. En cultivos como el arroz, que requieren condiciones hídricas específicas, el uso de tecnologías avanzadas como la ciencia de datos, el aprendizaje automático y las imágenes satelitales ofrece un enfoque innovador para abordar los desafíos relacionados con la gestión del agua y la productividad agrícola.

Este marco teórico aborda los fundamentos del aprendizaje supervisado y los modelos predictivos aplicados a la estimación de variables continuas como la humedad del suelo, junto con las principales tecnologías y métricas de evaluación empleadas en este contexto.

4.1.1 Fundamentos del aprendizaje supervisado

El aprendizaje supervisado se basa en el uso de un conjunto de entrenamiento compuesto por datos de entrada y sus salidas correspondientes (etiquetadas). Este proceso incluye etapas esenciales como la recopilación y limpieza de datos, la selección de características, el entrenamiento del modelo y la validación de su desempeño. En la Figura 1 se presenta un diagrama simplificado del flujo de trabajo típico, que abarca desde la adquisición de datos hasta la generación de predicciones útiles.

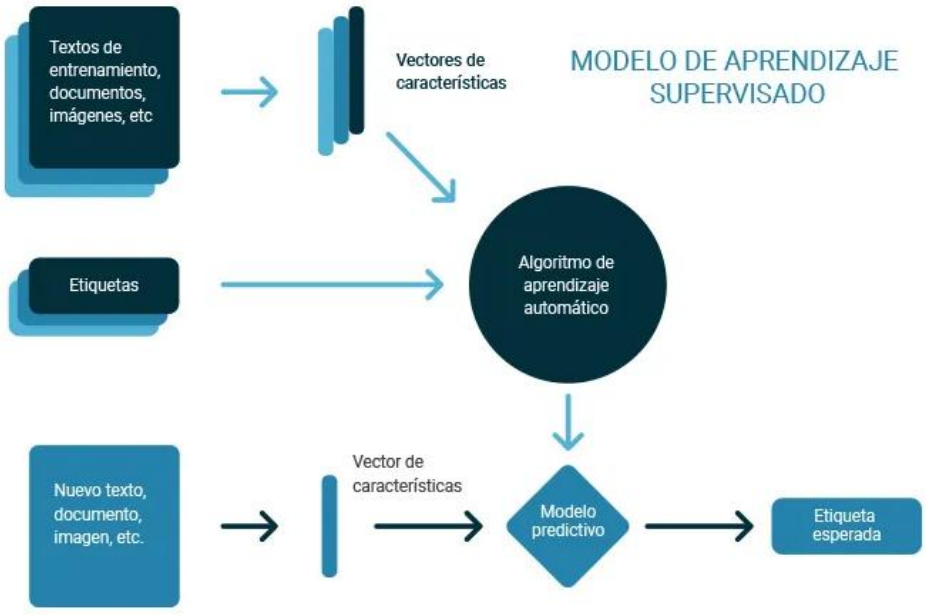

Figura 1. Diagrama de flujo del aprendizaje supervisado

En el ámbito agrícola, el aprendizaje supervisado ha demostrado ser una herramienta clave para resolver diversos problemas, entre ellos:

I. **Estimación de la humedad del suelo**: Modelos de regresión como *Random Forest* o redes neuronales permiten predecir la humedad a partir de datos de sensores, drones e imágenes satelitales, optimizando el riego y la salud de los cultivos.[18]
II. **Predicción del rendimiento agrícola**: Algoritmos como las máquinas de vectores de soporte (SVM) se han utilizado para proyectar rendimientos basados en el NDVI, datos meteorológicos y propiedades del suelo [19].
III. **Detección de enfermedades:** Modelos de clasificación supervisada, como redes neuronales convolucionales (CNN), se aplican al análisis de imágenes aéreas para identificar enfermedades en cultivos [20].
IV. **Mapeo de propiedades del suelo:** La integración de datos satelitales con aprendizaje supervisado permite generar mapas de salinidad o capacidad de retención de agua, fundamentales para la gestión agronómica [8].

En este proyecto, la integración de datos heterogéneos (sensoriales, satelitales y de drones) puede proporcionar un marco robusto para entrenar modelos supervisados. Estos modelos son especialmente útiles para optimizar el riego, mejorar la eficiencia en el uso del agua y garantizar la sostenibilidad de cultivos como el arroz, que es una de las principales áreas de investigación en la región.

4.1.2 Modelos predictivos de aprendizaje supervisado para regresión

El aprendizaje supervisado para tareas de regresión se centra en predecir valores continuos a partir de un conjunto de características independientes (variables de entrada). Estos modelos se entrenan utilizando datos históricos, donde la variable objetivo (como la humedad del suelo) ya está etiquetada, y se ajustan para minimizar la diferencia entre las predicciones y los valores reales [2].

La evidencia comparativa reciente en 144 estudios publicados entre 2010 y 2024 muestra que los modelos basados en árboles, en particular Random Forest, se destacan por su estabilidad y precisión; mientras que SVR y redes neuronales ofrecen resultados competitivos en distintas condiciones. Además, los enfoques multifuente que integran variables de teledetección, clima y suelo superan consistentemente a los de fuente única en la predicción de humedad del suelo [40].

A continuación, se detallan los principales enfoques y modelos aplicados en el contexto de la agricultura, específicamente para la estimación de la humedad del suelo:

4.1.2.1 Modelos lineales

El modelo lineal es uno de los enfoques más básicos y ampliamente utilizados para resolver problemas de regresión. Su principio fundamental consiste en asumir una relación lineal entre las variables independientes (características predictoras) y la variable dependiente (humedad del suelo). Debido a su simplicidad, los modelos lineales son eficientes, interpretables y constituyen la

base teórica de numerosos algoritmos de aprendizaje automático para regresión [2]. No obstante, su capacidad predictiva puede ser limitada en problemas complejos donde las relaciones entre variables no sean estrictamente lineales [17].

La ecuación general de un modelo lineal se expresa como:

$$y = \beta_o + \beta_1 x_1 + \beta_2 x_2 \dots \beta_n x_n + \in$$

Donde:

- y es la variable dependiente (predicción de humedad del suelo).
- x_i son las características de entrada.
- β_i son los coeficientes que representan el peso de cada característica.
- ϵ es el error residual.

En el ámbito agrícola, los modelos lineales han sido utilizados para estimar la humedad del suelo combinando datos de sensores locales, como temperatura o humedad del suelo, con variables satelitales, como los índices de vegetación derivados de imágenes multiespectrales [2], [24,25]. Esta capacidad de integración de fuentes heterogéneas de datos los hace particularmente útiles como punto de partida en el desarrollo de modelos predictivos para el manejo del agua en la agricultura de precisión.

Desde el punto de vista teórico, los modelos lineales son paramétricos, lo que significa que asumen una forma funcional explícita entre las variables. El entrenamiento del modelo consiste en ajustar los parámetros β_i para minimizar una función de pérdida, típicamente el Error Cuadrático Medio (MSE), mediante métodos de optimización como el descenso de gradiente o la descomposición matricial por mínimos cuadrados [2].

Sin embargo, en contextos donde las variables ambientales presentan interacciones complejas o colinealidad como ocurre frecuentemente en datos de teledetección, la regresión lineal puede tender al sobreajuste o mostrar baja capacidad de generalización. Para mitigar estas limitaciones, se han desarrollado versiones regularizadas como la regresión ridge, lasso y elastic net, las cuales introducen penalizaciones sobre los coeficientes β_i para controlar su magnitud y reducir el impacto del ruido en los datos [2].

4.1.2.2 Modelos no lineales

Los modelos no lineales permiten capturar relaciones complejas y estructuras intrínsecas en los datos que no pueden ser representadas por funciones lineales simples. En la estimación de la humedad del suelo, destacan los algoritmos K-Vecinos más Cercanos (KNN) y Máquinas de Vectores de Soporte (SVM), ampliamente usados en agricultura de precisión.

K-Vecinos más Cercanos (KNN)

El algoritmo K-Nearest Neighbors (KNN) es un método no paramétrico que estima el valor de una muestra desconocida en función de la media de sus *k* vecinos más cercanos en el espacio de características. Su desempeño depende de la elección del parámetro *k* y de la métrica de distancia empleada (por ejemplo, Euclidiana o Manhattan).
En estudios de predicción de humedad del suelo, KNN ha mostrado resultados consistentes al capturar variaciones locales entre parcelas agrícolas, especialmente cuando las relaciones entre variables ambientales no pueden expresarse de manera lineal [32]. Este modelo es útil en contextos donde se requiere una interpretación espacial detallada y un enfoque empírico de proximidad entre observaciones.

- **Máquinas de Vectores de Soporte (SVM)**

Los Support Vector Machines (SVM) son modelos robustos que pueden emplearse tanto para clasificación como para regresión. En tareas de regresión, la versión conocida como Support Vector Regression (SVR**)** busca encontrar un hiperplano que minimice los errores dentro de un margen definido, optimizando la capacidad de generalización [19].

El uso de funciones kernel, como el radial o el polinómico, permite que las SVM modelen relaciones no lineales entre las variables de entrada y la humedad del suelo, haciendo posible capturar dependencias complejas entre factores climáticos, edáficos y de vegetación. Revisiones recientes confirman que SVR sigue siendo uno de los algoritmos no lineales más robustos para la predicción de humedad superficial, especialmente en escenarios con datos derivados de radar SAR y variables climáticas combinadas [40].

4.1.2.3 Modelos de ensamble

Los modelos de ensamble combinan múltiples modelos base para mejorar la precisión y estabilidad de las predicciones. Los dos métodos más representativos son Random Forest (bagging) y Gradient Boosting (boosting), ampliamente aplicados en problemas de regresión agrícola [18], [36], [37].

- **Random Forest**

El Random Forest es un método de ensamble basado en árboles de decisión que combina múltiples modelos entrenados de manera independiente sobre diferentes subconjuntos del conjunto de datos. Cada árbol genera una predicción, y el modelo final obtiene su resultado promedio (en tareas de regresión) o por voto mayoritario (en clasificación). Este enfoque se fundamenta en el principio de *bootstrap aggregating* o *bagging*, que permite reducir la varianza y mejorar la estabilidad del modelo [18], [36].

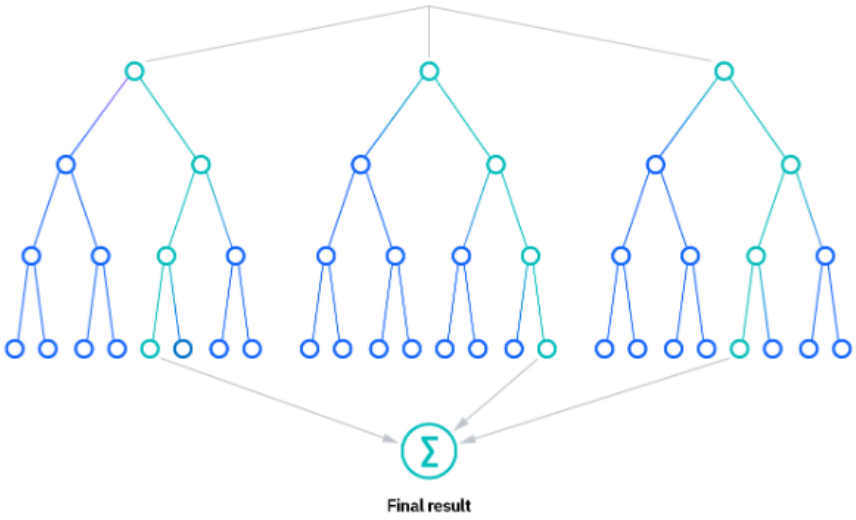

Figura 2. Diagrama del modelo

El proceso general del algoritmo comprende las siguientes etapas:

1. **Muestreo aleatorio:** se seleccionan múltiples subconjuntos del conjunto de entrenamiento mediante el proceso de *bootstrap sampling*.
2. **Entrenamiento independiente:** cada subconjunto se utiliza para entrenar un árbol de decisión, empleando una selección aleatoria de características en cada división.
3. **Agregación de resultados:** la predicción final se obtiene al promediar las salidas individuales de los árboles en tareas de regresión, o mediante votación para tareas de clasificación.

En el contexto del presente proyecto, el modelo Random Forest permite integrar datos de sensores in situ con índices espectrales derivados de imágenes satelitales PlanetScope, como el NDWI, para estimar la humedad del suelo de manera precisa y escalable [25]. Gracias a su capacidad de manejar grandes volúmenes de datos heterogéneos provenientes de sensores, clima e imágenes multiespectrales, el algoritmo ofrece predicciones estables y resistentes al ruido.
Estudios recientes destacan que Random Forest es uno de los modelos más utilizados y con mejor desempeño promedio en la estimación de humedad del suelo, debido a su equilibrio entre precisión, robustez y facilidad de interpretación [32], [38], [40]. En particular, Lamichhane et al. [40] identifican a Random Forest como el algoritmo con mayor consistencia en distintos contextos geográficos y tipos de suelo, superando a otros enfoques cuando se integran múltiples fuentes de datos (satelitales, climáticas y de campo).

- ***Grandient Boosting***

El Gradient Boosting es una técnica de ensamble secuencial en la que cada modelo se entrena para corregir los errores cometidos por los modelos anteriores. A diferencia de los métodos de *bagging* como Random Forest, que entrenan árboles en paralelo, el *boosting* entrena de manera iterativa y

aditiva, generando una combinación ponderada de árboles débiles para formar un modelo final más robusto [36].

El principio fundamental del Gradient Boosting Machine (GBM), propuesto por Friedman [36], consiste en ajustar cada nuevo árbol a los residuos del modelo previo, es decir, a la diferencia entre los valores observados y las predicciones anteriores. De esta manera, el algoritmo aprende de los errores previos y mejora progresivamente la precisión general del conjunto.

El proceso general puede describirse en las siguientes etapas:

1. Inicialización del modelo: se establece un modelo base (por ejemplo, un árbol de decisión simple) que genera una predicción inicial.
2. Cálculo de residuos: se calculan los errores entre las predicciones y los valores reales del conjunto de entrenamiento.
3. Ajuste secuencial: un nuevo árbol se entrena para predecir los residuos, ponderando los errores más grandes con mayor importancia.
4. Actualización del modelo: las nuevas predicciones se suman ponderadamente a las anteriores mediante una tasa de aprendizaje (*learning rate*) que controla la velocidad de convergencia.
5. Repetición y detención: el proceso se repite hasta alcanzar un número máximo de iteraciones o hasta que el error deje de mejorar significativamente.

El parámetro learning rate (η) desempeña un papel crucial, ya que un valor muy alto puede provocar sobreajuste, mientras que un valor demasiado bajo puede ralentizar el aprendizaje. Por ello, se recomienda un ajuste fino mediante técnicas de validación cruzada. Una de las implementaciones más avanzadas de este enfoque es XGBoost (Extreme Gradient Boosting), desarrollada por Chen y Guestrin [37]. Este algoritmo introduce mejoras computacionales y de regularización que lo convierten en una de las herramientas más potentes y eficientes para tareas de regresión y clasificación. Entre sus innovaciones se destacan:

- Regularización L1 y L2, que penaliza la complejidad de los árboles y previene el sobreajuste.
- Optimización paralela y en memoria, lo que permite manejar grandes volúmenes de datos con alta velocidad.
- Poda de árboles (tree pruning) basada en la ganancia de información, para eliminar divisiones redundantes.
- Soporte para valores faltantes, lo que mejora la robustez del modelo frente a bases de datos incompletas.

En el contexto agrícola, XGBoost ha demostrado un desempeño sobresaliente para predecir variables continuas como la humedad del suelo, la biomasa y el rendimiento de cultivos, al integrar múltiples fuentes de información (datos satelitales, climáticos y de sensores) [38], [40].

Lamichhane et al. [40] reportan que los modelos Gradient Boosting y sus variantes modernas, como XGBoost y LightGBM, se encuentran entre los métodos con mayor precisión en la estimación de humedad del suelo, particularmente cuando se utilizan datos multitemporales y variables derivadas de radar SAR. Estas técnicas destacan por su capacidad para capturar relaciones no lineales complejas entre los índices espectrales y las condiciones edáficas, logrando un equilibrio entre rendimiento y generalización.

- **Redes Neuronales artificiales (ANN y DN)**

Las Redes Neuronales Artificiales (ANN, Artificial Neural Networks**)** son modelos computacionales inspirados en la estructura del cerebro humano, diseñados para identificar patrones complejos y no lineales en grandes volúmenes de datos. Una red neuronal está compuesta por unidades denominadas neuronas artificiales, organizadas en capas: una capa de entrada, una o más capas ocultas y una capa de salida. Cada conexión entre neuronas posee un peso que se ajusta durante el proceso de entrenamiento con el fin de minimizar los errores de predicción [16].

El entrenamiento de una red neuronal se basa en un proceso iterativo de optimización, donde los pesos se actualizan utilizando algoritmos de descenso de gradiente. Este proceso busca minimizar una función de pérdida como el error cuadrático medio (MSE), ajustando los parámetros internos para que las predicciones del modelo se aproximen a los valores reales. La capacidad de una red neuronal para aprender representaciones jerárquicas de los datos la convierte en una herramienta poderosa para la regresión no lineal [17].

Las Redes Neuronales Profundas (DNN, Deep Neural Networks) extienden esta idea al incorporar múltiples capas ocultas, lo que les permite capturar interacciones complejas entre las variables de entrada, como la radiación solar, la temperatura, los índices de vegetación (NDVI, NDWI) y la precipitación. No obstante, su desempeño depende en gran medida de la calidad y cantidad de datos disponibles, así como de una adecuada regularización para evitar el sobreajuste [17], [20].

El proceso general de una red neuronal aplicada a la predicción de humedad del suelo puede resumirse en las siguientes etapas:

1. Entrada de datos: las variables independientes (por ejemplo, temperatura, precipitación, índices satelitales) se introducen a la capa de entrada.
2. Propagación hacia adelante (feedforward): las señales pasan a través de las capas ocultas, donde se aplican funciones de activación no lineales como *ReLU* o *tanh*.
3. Cálculo del error: se evalúa la diferencia entre las predicciones y los valores reales utilizando una función de pérdida.
4. Retropropagación (backpropagation): el error se distribuye inversamente a través de la red para ajustar los pesos de las conexiones neuronales.
5. Predicción final: tras múltiples iteraciones, la red produce una estimación continua de la humedad del suelo.

En la agricultura de precisión, las redes neuronales se han empleado exitosamente para modelar relaciones complejas entre variables espectrales, climáticas y edáficas. Estudios recientes [38], [40]

muestran que las DNN pueden superar a modelos tradicionales como SVM o Random Forest cuando se dispone de bases de datos amplias y multifuente, debido a su capacidad de aprendizaje profundo y generalización en entornos heterogéneos.

4.1.3 Métricas de evaluación

La evaluación de los modelos predictivos constituye una etapa crítica dentro del proceso de modelado, pues permite determinar su rendimiento, capacidad de generalización y aplicabilidad en escenarios reales. En el contexto de la estimación de la humedad del suelo, se emplean métricas cuantitativas que comparan las predicciones del modelo con los valores observados, ofreciendo una medida objetiva de su precisión. La selección de estas métricas depende del tipo de modelo, la naturaleza de los datos y el propósito del análisis [22], [38].

Las métricas más utilizadas en modelos de regresión son el Error Cuadrático Medio (MSE), el Error Absoluto Medio (MAE) y el Coeficiente de Determinación (R^2). Cada una ofrece una perspectiva distinta del desempeño del modelo.

- **Error cuadrático medio (MSE)**

El MSE mide la diferencia promedio al cuadrado entre los valores predichos y^ i) y los valores reales y_i en un conjunto de datos. Su fórmula es:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(yi - \hat{yi})^2$$

Donde:

- N: Número total de observaciones.
- Yi: Valor real de la iii-ésima observación.
- y^i: Valor predicho de la iii-ésima observación.

El MSE penaliza de forma más severa los errores grandes, debido al término cuadrático, por lo que resulta especialmente útil para modelos que deben minimizar desviaciones extremas, como las redes neuronales profundas (DNN) o los SVM en tareas con datos sensibles [38]. No obstante, su sensibilidad a los valores atípicos puede influir negativamente cuando el conjunto de datos presenta alta variabilidad natural, como suele ocurrir en condiciones agrícolas y climáticas heterogéneas.

- **Error Absoluto Medio (MAE)**

El MAE (Mean Absolute Error) calcula el promedio de las diferencias absolutas entre los valores reales y los predichos. Su expresión matemática es:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}[yi - \bar{y}]$$

A diferencia del MSE, el MAE no eleva al cuadrado los errores, por lo que todos los errores contribuyen de manera lineal a la métrica. Esto lo convierte en una medida más robusta ante valores atípicos [22]. Además, el MAE conserva las mismas unidades que la variable objetivo (por ejemplo, porcentaje de humedad), lo que facilita su interpretación práctica.

Un MAE bajo indica que el modelo mantiene predicciones cercanas a los valores reales de manera consistente, siendo ideal para aplicaciones agrícolas donde la interpretabilidad y la estabilidad del modelo son prioritarias

- **Coeficiente de Determinación (R^2)**

El coeficiente de determinación (R^2) cuantifica la proporción de la variabilidad total de los datos que el modelo es capaz de explicar. Su fórmula es:

$$R^2 = 1 - \frac{\sum_{1=1}^{n}(yi - \hat{y}i)^2}{\sum_{1=1}^{n}(yi - \bar{y})^2}$$

donde $\bar{y}$ representa el promedio de los valores reales.

Un valor de R^2 cercano a 1 indica que el modelo explica gran parte de la varianza observada, mientras que valores próximos a 0 reflejan un bajo poder explicativo. En modelos de estimación de humedad del suelo, un $R^2 > 0.8$ suele considerarse un desempeño excelente, aunque este umbral puede variar según las condiciones del sitio y la resolución de los datos [40].

4.1.4 Tecnologías aplicadas al monitoreo de la humedad del suelo

Imágenes satelitales

Las imágenes satelitales constituyen una de las herramientas más relevantes para la observación terrestre a gran escala. Permiten obtener información multiespectral que puede relacionarse con las condiciones de la vegetación y del suelo. Plataformas como PlanetScope, Sentinel-2 o MODIS ofrecen datos de alta resolución espacial y temporal, lo que las hace idóneas para el seguimiento de

la humedad en sistemas agrícolas [3], [23], [29].

La humedad del suelo puede estimarse indirectamente mediante índices espectrales derivados de combinaciones de bandas, como el NDWI (Normalized Difference Water Index), el NDVI (Normalized Difference Vegetation Index) y el SAVI (Soil Adjusted Vegetation Index). Estos índices permiten evaluar el contenido de agua en la vegetación y la reflectancia del suelo, correlacionándose con el estado hídrico del terreno [29].

Imágenes aéreas capturadas por drones

El uso de vehículos aéreos no tripulados (UAV o drones) complementa la información satelital al ofrecer observaciones detalladas a nivel de parcela. Estos dispositivos, equipados con cámaras multiespectrales o térmicas, permiten capturar imágenes de alta resolución espacial y generar mapas de variación hídrica con precisión centimétrica [24], [30].

Los datos de drones son especialmente útiles para detectar microvariaciones dentro de un lote agrícola que pueden pasar desapercibidas en los satélites. Además, su flexibilidad operativa permite realizar vuelos bajo demanda en momentos críticos del ciclo del cultivo. Combinados con índices espectrales derivados (por ejemplo, NDWI o NDTI), los UAV proporcionan una fuente valiosa para calibrar y validar los modelos de predicción de humedad [38].

Sensores de humedad del suelo

Los sensores instalados directamente en el suelo proporcionan mediciones continuas y precisas del contenido volumétrico de agua, sirviendo como referencia para validar los modelos derivados de imágenes remotas. Estos dispositivos funcionan a partir de principios como la reflexión dieléctrica (FDR) o la transmisión de tiempo (TDR), que permiten cuantificar la constante dieléctrica del suelo y relacionarla con su contenido de humedad [25, 31].

El valor de los sensores in situ radica en su capacidad de capturar la dinámica hídrica en tiempo real, incluyendo la variabilidad vertical (profundidad) y temporal. En este estudio, las mediciones de campo obtenidas de sensores instalados en parcelas del CIAT fueron empleadas como variable objetivo para entrenar y validar los modelos de aprendizaje supervisado.

Integración de Fuentes Multiespectrales y Multisensor

La integración de fuentes heterogéneas satelitales, aéreas y de sensores proporciona una representación más completa del estado del suelo. Este enfoque multifuente ha demostrado mejorar la precisión de los modelos predictivos al combinar la amplitud espacial de los satélites con el detalle local de los sensores [38]. Los modelos basados en datos integrados superan sistemáticamente a los desarrollados con una sola fuente, especialmente en zonas agrícolas con alta variabilidad edáfica. La fusión de datos multiespectrales y temporales también permite monitorear cambios estacionales y mejorar la resiliencia frente a condiciones climáticas extremas [40].

Dinámica hídrica y aplicación de modelos predictivos en algunos cultivos

La humedad del suelo constituye un elemento esencial en la dinámica productiva de los agroecosistemas, ya que regula la disponibilidad de agua para las raíces, influye en la fotosíntesis, la

transpiración y determina en gran medida el rendimiento de los cultivos. Una gestión adecuada de este recurso es crítica en regiones tropicales donde las lluvias presentan alta variabilidad interanual, exacerbada por los efectos del cambio climático [28].

En cultivos como el arroz, esta variable adquiere especial relevancia, dado que su crecimiento y productividad dependen de sistemas de riego e inundación controlados. Se estima que la producción de un kilogramo de arroz requiere entre 2 000 y 5 000 litros de agua, lo que convierte a este cultivo en uno de los mayores consumidores del sector agrícola [26]. La predicción precisa de la humedad del suelo mediante herramientas de teledetección y modelos de aprendizaje automático permite optimizar el uso del agua, mejorar la eficiencia de los riegos y reducir las pérdidas productivas en períodos de estrés hídrico [27].

De forma paralela, los pastos tropicales representan la base de la alimentación ganadera en sistemas de producción sostenible. Su desarrollo depende directamente de la humedad del suelo, siendo especialmente sensibles a déficits prolongados que disminuyen la calidad del forraje y reducen la capacidad de carga animal [30], [31]. En este contexto, modelar la humedad del suelo en zonas de pastizales permite anticipar escenarios de sequía, planificar estrategias de suplementación alimenticia y contribuir a una mayor resiliencia agropecuaria frente al cambio climático.
El vínculo entre humedad del suelo y productividad agrícola se traduce en la necesidad de implementar sistemas de monitoreo continuo, combinando sensores de campo, imágenes satelitales y observaciones de dron para caracterizar el estado hídrico del cultivo en tiempo casi real. Estudios como el de Jones [29] demuestran que integrar mediciones fisiológicas de la planta con estimaciones de humedad del suelo mejora la toma de decisiones sobre el riego y la programación de labores agronómicas.

En los últimos años, los modelos predictivos basados en inteligencia artificial se han consolidado como una herramienta eficaz para este propósito. Algoritmos como Random Forest, Support Vector Regression y Redes Neuronales integran información de distintas fuentes (sensores, clima, índices espectrales y textura del suelo) para generar mapas de humedad con alta precisión espacial y temporal. En cultivos de arroz, estas técnicas facilitan la programación del riego y la reducción de consumo hídrico, mientras que en sistemas de pasturas mejoran la gestión de recursos forrajeros y la productividad ganadera [32]. En conjunto, la combinación de monitoreo multifuente, análisis espectral y modelado predictivo constituye una estrategia robusta para fortalecer la sostenibilidad hídrica y productiva de los sistemas agrícolas tropicales.

4.2 Antecedentes

Los siguientes artículos incluyen aspectos complementarios de la predicción de la humedad del suelo. El primero combina datos satelitales con propiedades del suelo para predecir la humedad en la zona de raíces a escala regional, mientras que el segundo explora cómo la imagen satelital multitemporal puede mejorar las predicciones mediante técnicas de mapeo digital del suelo. El

tercero explora el uso de la teledetección hiperespectral basada en UAV para estimar el contenido de humedad del suelo (SMC) en sistemas agroecológicos en regiones áridas. En el último, se pretende modelar y predecir la dinámica de la humedad del suelo utilizando datos hidrometeorológicos recogidos in situ para el entrenamiento y técnicas de aprendizaje automático basadas en datos. Todos contribuyen a mejorar los modelos de humedad del suelo, con aplicaciones valiosas para la gestión agrícola y ambiental.

4.2.1 Predicting root zone soil moisture with soil properties and satellite near-surface moisture data across the conterminous United States

La predicción de la humedad del suelo en la zona radicular (RZSM) a gran escala representa un componente esencial en la gestión agrícola, el diagnóstico de sequías y la modelación del ciclo hidrológico y del carbono. La RZSM generalmente definida como los primeros 100 cm del perfil del suelo regula el balance de agua y energía de los ecosistemas agrícolas, afectando directamente la productividad y la disponibilidad hídrica. Sin embargo, las mediciones satelitales de humedad superficial, aunque ampliamente disponibles, presentan limitaciones para reflejar con precisión la variabilidad en profundidad, debido a los sesgos inducidos por la vegetación, la topografía y las propiedades físicas del suelo [41].

El estudio desarrollado por Baldwin et al. [41] introduce el sistema SMAR–EnKF (Soil Moisture Analytical Relationship – Ensemble Kalman Filter), el cual integra un modelo analítico de infiltración con un algoritmo de asimilación de datos para estimar la humedad radicular en todo el territorio continental de los Estados Unidos. Esta metodología combina observaciones satelitales de humedad superficial con propiedades edáficas (textura, densidad aparente, porosidad y capacidad de retención de agua) y variables climáticas, como la evapotranspiración derivada del sensor MODIS, con el fin de mejorar la precisión de las predicciones.

El modelo se estructura en dos componentes principales:

1. Corrección de sesgos a escala regional mediante el filtro de Kalman por conjuntos (EnKF), que ajusta dinámicamente los errores sistemáticos presentes en los datos satelitales.
2. Estimación estadística de los parámetros del modelo SMAR a partir de relaciones multivariadas entre las propiedades del suelo, la cobertura vegetal y las condiciones hidrometeorológicas.

Los autores validaron el sistema con datos in situ provenientes de más de 150 estaciones distribuidas en diferentes ecorregiones de los Estados Unidos, demostrando que la integración de información satelital y edáfica reduce significativamente los errores de predicción. Los resultados mostraron un coeficiente de determinación (R^2) superior a 0.80, indicando una alta capacidad del modelo para capturar la dinámica temporal y espacial de la humedad en el perfil del suelo.

Este trabajo constituye un referente en la estimación operativa de la humedad del suelo, ya que combina modelos hidrológicos físicamente fundamentados con técnicas estadísticas de asimilación de datos, lo que permite una aplicación flexible y de bajo costo computacional para el monitoreo de la RZSM a gran escala. Su enfoque ha sido ampliamente citado como una base metodológica para

estudios posteriores que integran teledetección, propiedades del suelo y aprendizaje automático en la predicción de humedad edáfica [41].

4.2.2 Effect of multi-temporal satellite images on soil moisture prediction using a digital soil mapping approach

La humedad del suelo (SM) es un componente esencial del ciclo hidrológico, influyendo directamente en la gestión de los recursos hídricos, el balance energético y la estabilidad de los ecosistemas terrestres. Sin embargo, debido a su naturaleza altamente variable en el espacio y el tiempo, su monitoreo continuo a gran escala sigue siendo un desafío. Los métodos tradicionales basados en sensores in situ, aunque precisos localmente, presentan limitaciones en cobertura espacial y temporal, lo que ha impulsado el desarrollo de metodologías basadas en teledetección y modelado digital del suelo (DSM, *Digital Soil Mapping*) [13].

El DSM se fundamenta en la relación estadística entre las propiedades del suelo y un conjunto de covariables ambientales —como relieve, clima, cobertura vegetal y geología— derivadas de fuentes satelitales o geoespaciales. Tradicionalmente, estos modelos han utilizado covariables estáticas, por ejemplo, parámetros topográficos extraídos de Modelos Digitales de Elevación (DEM). Sin embargo, el estudio de Fathololoumi et al. [13] resalta la necesidad de incorporar covariables dinámicas derivadas de imágenes satelitales multitemporales, que permiten capturar la variabilidad estacional y transitoria de la humedad del suelo.

El objetivo principal de esta investigación fue evaluar el aporte de las imágenes multitemporales en la predicción de la humedad del suelo mediante el enfoque DSM. Para ello, se compararon tres configuraciones de modelos:

1. Modelos estáticos, que utilizan únicamente covariables topográficas y geológicas;
2. Modelos dinámicos, que emplean series multitemporales de índices espectrales satelitales; y Modelos híbridos, que integran simultáneamente covariables estáticas y dinámicas para mejorar la robustez y la capacidad predictiva.

Los modelos fueron entrenados utilizando datos de tres meses consecutivos y validados mediante la predicción de la humedad del suelo en un cuarto mes, permitiendo analizar la capacidad de generalización temporal. Los resultados mostraron que la inclusión de covariables dinámicas derivadas de teledetección mejora sustancialmente la precisión de las estimaciones, evidenciada en incrementos significativos de R^2 y reducciones del error cuadrático medio (RMSE).

En conjunto, la investigación demuestra que el uso de imágenes satelitales multitemporales en el marco del DSM proporciona una herramienta más eficaz para el monitoreo espacial y temporal de la humedad del suelo, contribuyendo a una mejor comprensión de su variabilidad y sus impactos sobre la agricultura, la planificación de recursos hídricos y los efectos del cambio climático [13].

4.2.3 Estimating agricultural soil moisture content through UAV-based hyperspectral images in the Arid region

El estudio desarrollado por Ge et al. [14] analiza el potencial de la teledetección hiperespectral basada en vehículos aéreos no tripulados (UAV) para estimar con alta precisión el contenido de humedad del suelo (SMC) en sistemas agroecológicos ubicados en zonas áridas. Este enfoque cobra relevancia en el contexto de la agricultura de precisión, donde la disponibilidad hídrica es un factor limitante y su monitoreo detallado resulta esencial para la optimización del riego y la sostenibilidad de los cultivos.

La investigación se llevó a cabo en la ciudad de Fukang, Región Autónoma Uigur de Xinjiang (China), cubriendo un área experimental de 25,000 m². Se adquirieron imágenes hiperespectrales con resolución espacial de 4 cm, y se recolectaron 70 muestras de suelo a una profundidad de 0–10 cm para calibrar y validar los modelos de predicción. El estudio evaluó cuatro estrategias metodológicas:

1. Estrategia I: uso directo de las imágenes originales;
2. Estrategia II: aplicación de derivadas de primer y segundo orden;
3. Estrategia III: empleo de la técnica de derivada de orden fraccionado (FOD); y
4. Estrategia IV: combinación del orden fraccionado óptimo con los índices multibanda óptimos, integrados en un modelo basado en el algoritmo eXtreme Gradient Boosting (XGBoost).

Los resultados evidenciaron que la técnica FOD permitió extraer características espectrales útiles para representar las variaciones del contenido de humedad en el suelo, alcanzando un coeficiente de correlación máximo de 0.768. Entre las estrategias evaluadas, la Estrategia IV mostró el mejor desempeño, con valores de R^2 = 0.921, RMSEP = 1.943 y RPD = 2.736, indicando una elevada capacidad predictiva. En particular, el modelo derivado de un orden fraccionado de 0.4 superó en rendimiento a las versiones basadas en derivadas de primer y segundo orden, demostrando su superioridad en la reducción de ruido espectral y la mejora de la sensibilidad a las señales asociadas a la humedad.

En conjunto, el estudio confirma que la integración de la técnica FOD y los índices multibanda óptimos, dentro de un esquema de aprendizaje automático (XGBoost), constituye un enfoque altamente eficaz para estimar el contenido de humedad del suelo a partir de imágenes hiperespectrales UAV. Este avance representa un paso significativo hacia la automatización del monitoreo hídrico en la agricultura de precisión, al combinar herramientas de minería de datos espectrales con algoritmos robustos de predicción no lineal [14].

4.2.4 Predicción de la humedad del suelo mediante aprendizaje por transferencia: Una aplicación en los Altos Andes Tropicales

La humedad del suelo constituye una variable clave dentro del ciclo hidrológico global, al influir de manera directa en la recarga de acuíferos, el funcionamiento de los ecosistemas, la regulación del microclima y la ocurrencia de eventos extremos como sequías o deslizamientos. En regiones de topografía compleja, como los Altos Andes Tropicales, su estimación precisa se ve obstaculizada por

la escasez de datos continuos y por la heterogeneidad espacial de las condiciones edáficas y climáticas. Estas limitaciones han motivado el desarrollo de modelos basados en aprendizaje automático, capaces de integrar múltiples fuentes de información hidrometeorológica para mejorar la representación de la dinámica hídrica del suelo [15].

El estudio de Escobar-González et al. [15] aborda este desafío mediante la aplicación de técnicas de aprendizaje por transferencia para modelar y predecir la variabilidad temporal de la humedad del suelo en ecosistemas altoandinos de Ecuador. Inicialmente, los autores construyeron un modelo base utilizando redes neuronales artificiales (ANN) entrenadas con series temporales de datos hidrometeorológicos recolectados *in situ* incluyendo temperatura, precipitación, radiación y humedad relativa. Posteriormente, el modelo fue ajustado mediante transfer learning, transfiriendo el conocimiento adquirido a nuevos horizontes y perfiles de suelo con características edáficas y vegetativas distintas, pero pertenecientes al mismo dominio climático.

Los resultados demostraron que el enfoque de transferencia mejoró significativamente la capacidad de generalización de los modelos, manteniendo una alta precisión incluso en sitios sin datos de entrenamiento directo. La técnica alcanzó errores del orden de $1\times10^{-6} < \epsilon < 1\times10^{-3}$, con valores de RMSE = 4.77×10^{-6} y coeficientes de desempeño Nash–Sutcliffe Efficiency (NSE) y Kling–Gupta Efficiency (KGE) ambos iguales a 0.97, reflejando una predicción casi perfecta respecto a los valores observados.

El estudio resalta el papel determinante de la cobertura vegetal en el control de la dinámica hídrica, evidenciando diferencias marcadas entre bosques nativos, páramos y zonas agrícolas. Además, demuestra que la combinación de aprendizaje profundo y transferencia de conocimiento constituye una estrategia efectiva para superar la falta de datos locales, extendiendo la aplicabilidad de los modelos a regiones con escasa instrumentación.

En síntesis, la investigación de Escobar-González et al. [15] representa un avance metodológico relevante en el uso de redes neuronales y aprendizaje por transferencia para el modelado de la humedad del suelo en entornos de alta complejidad topográfica. Sus resultados abren perspectivas prometedoras para la vigilancia de riesgos naturales, la gestión de recursos hídricos y la planificación agrícola sostenible en regiones montañosas tropicales.

4.2.5 A Comprehensive Study of Deep Learning for Soil Moisture Prediction

El trabajo de Wang et al. [42], publicado en *Hydrology and Earth System Sciences* (2024), ofrece una evaluación exhaustiva de distintos modelos de aprendizaje automático y profundo aplicados a la predicción de la humedad del suelo (SM), con un enfoque comparativo entre algoritmos clásicos y redes neuronales profundas. El estudio tiene como objetivo identificar las fortalezas, limitaciones y escenarios óptimos de uso de cada técnica en función de las propiedades del suelo, las condiciones climáticas y la escala espacial de análisis.

La investigación se centra en tres modelos principales: Random Forest (RF), Support Vector Regression (SVR) y Redes Neuronales Artificiales (ANN). Los autores recopilaron un conjunto de datos multianual que combina variables meteorológicas, índices de vegetación obtenidos por teledetección y mediciones *in situ* de humedad del suelo en múltiples regiones agrícolas de Asia y

Europa. Posteriormente, compararon el desempeño de los modelos en función de métricas como el coeficiente de determinación (R^2), el error cuadrático medio (RMSE) y el error absoluto medio (MAE).

Los resultados indican que RF se destaca por su robustez frente al ruido y su capacidad de manejar relaciones no lineales complejas, logrando un equilibrio adecuado entre sesgo y varianza. SVR, en contraste, mostró mejor desempeño en regiones con cobertura vegetal densa y menor heterogeneidad del suelo, gracias a su función núcleo (kernel) capaz de capturar patrones sutiles en los datos. Por su parte, las redes neuronales alcanzaron los valores más altos de R^2 (0.93 en promedio) y los menores errores de predicción, especialmente cuando se incorporaron variables temporales y espaciales combinadas.

El estudio también resalta la importancia de la selección de características y la estandarización de los datos como factores determinantes en el rendimiento de los modelos. Asimismo, se identificó que la integración de variables derivadas de imágenes satelitales (NDVI, LST, albedo) y datos meteorológicos (precipitación, evapotranspiración, temperatura del suelo) mejora significativamente la capacidad predictiva, confirmando el valor del aprendizaje multifuente.

Finalmente, los autores concluyen que el uso de modelos híbridos, que combinan enfoques basados en árboles (como RF) y arquitecturas neuronales, ofrece un potencial significativo para capturar tanto la heterogeneidad espacial como la dinámica temporal de la humedad del suelo. Este enfoque abre nuevas perspectivas para el desarrollo de sistemas predictivos más precisos, escalables y aplicables a diferentes zonas agroecológicas del mundo [42].

5. MARCO METODOLÓGICO

Con el fin de facilitar la comprensión del procesamiento de datos y la integración de las distintas fuentes de información, en esta sección se describe de manera sintética el flujo general seguido desde la recolección de los datos hasta la construcción del dataset final utilizado para el modelado.

En primer lugar, los registros diarios del sensor de humedad se obtuvieron de CIAT del sensor instalado en la parcela de estudio, se definieron como la referencia temporal principal del modelo, dado que constituyen la variable dependiente. A partir de esta base temporal, se alinearon las demás fuentes de información.

Las imágenes satelitales PlanetScope fueron procesadas de forma independiente, aplicando correcciones atmosféricas y máscaras de nubosidad, y posteriormente se calcularon índices espectrales representativos del estado vegetativo y del contenido hídrico. Para cada fecha disponible, los valores espectrales se resumieron mediante estadísticos zonales dentro de un buffer de 100 m alrededor del sensor, generando una serie temporal compatible con los registros diarios del suelo.

De manera paralela, los datos climáticos provenientes de la estación meteorológica del CIAT fueron depurados, validados y transformados mediante la generación de variables derivadas (rezagos y acumulados), con el objetivo de capturar la memoria hídrica atmosférica.

Finalmente, las tres fuentes de información (sensores, clima e índices espectrales) fueron integradas mediante la fecha como llave primaria, dando lugar a un dataset multivariable y multitemporal. Este proceso incluyó la identificación e imputación controlada de valores faltantes, garantizando la continuidad temporal y la coherencia física de las variables. El resultado fue una base de datos consolidada, adecuada para el análisis exploratorio y el entrenamiento de modelos de aprendizaje automático.

5.1 Recopilación e identificación de datos

5.1.1 Descripción del área de estudio

El área de investigación corresponde a una parcela experimental del Centro Internacional de Agricultura Tropical (CIAT)

Figura 3, ubicada en el departamento del Valle del Cauca, Colombia.
La parcela se encuentra georreferenciada en las coordenadas 3.497° de latitud norte y 76.374° de longitud oeste, dentro de una zona agrícola de alta representatividad para cultivos tropicales.

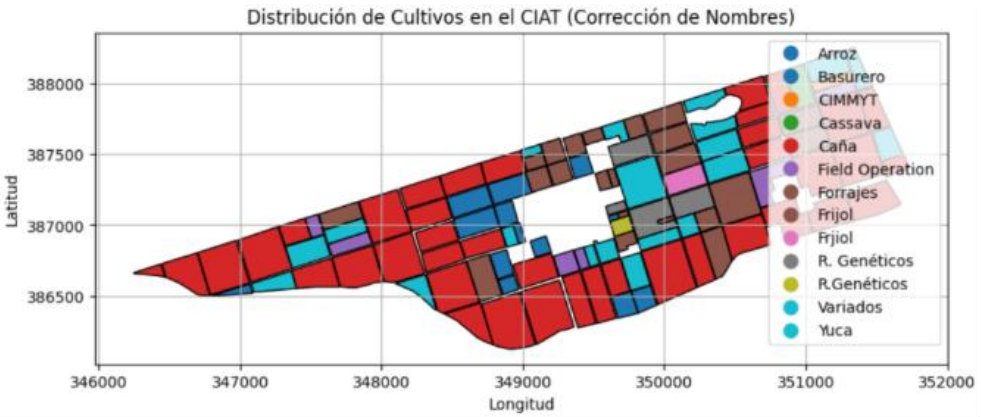

Figura 3. Distribución de cultivos en el CIAT

El sitio de estudio fue seleccionado por su disponibilidad de sensores de humedad instalados, condiciones agroclimáticas estables y acceso a imágenes satelitales de alta resolución

Figura 4.Esta combinación de fuentes permite analizar la dinámica hídrica del suelo con un enfoque multiescalar, integrando datos in situ y de teledetección.

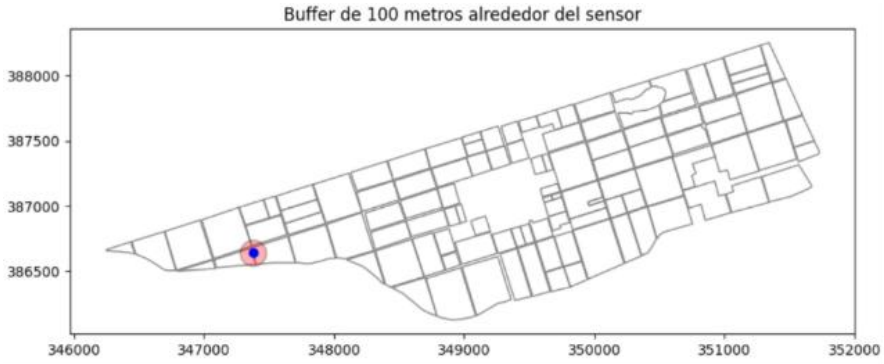

Figura 4. Selección de la parcela.

5.1.2 Variables dependientes e independientes

La variable dependiente del modelo es la humedad volumétrica del suelo (%), medida directamente mediante un sensor ubicado en la parcela experimental. Esta variable representa la fracción de volumen de agua contenida en el suelo respecto a su volumen total, y constituye la referencia empírica para el entrenamiento y validación del modelo.

Las variables independientes o predictoras provienen de tres fuentes complementarias, como se observa en la siguiente tabla:

Tabla 1. Variables dependientes e independientes

Tipo de variable	Fuente	Descripción
Espectrales	Imágenes PlanetScope (8 bandas)	Índices derivados (NDVI, SAVI, EVI, NDWI, MSI, NDMI_Custom, etc.) que representan vigor, contenido hídrico y estrés de la vegetación.
Climáticas	Estación meteorológica CIAT	Temperatura (tmax, tmin, tprom), precipitación, ET0, radiación solar, velocidad del viento y déficit de humedad.
De suelo (sensores)	Sensor instalado en campo	Humedad del suelo, temperatura, conductividad y variables auxiliares para validación y calibración.

Estas variables se seleccionaron por su relevancia física en los procesos de intercambio de energía y agua entre el suelo, la vegetación y la atmósfera. Su integración permite capturar tanto las condiciones locales del suelo como los patrones espaciales detectables por satélite.

5.1.3 Fuentes de información

El modelo integró información proveniente de tres fuentes principales:

1. Se emplearon 149 imagenes multiespectrales PlanetScope con resolución espacial de 3 m por píxel, con buffer circular de 100 m de radio alrededor del sensor de humedad, promediando únicamente los píxeles válidos (~3.490 por imagen, equivalentes a ~31.400 m²). Estas imágenes provenientes de PlanetScope con ocho bandas multiespectrales (coastal blue, blue, green_i, green, yellow, red, red edge y NIR) ,adquiridas diariamente entre periodo abril de 2024 – marzo de 2025.Todas las escenas fueron sometidas a corrección atmosférica y a la máscara UDM para la eliminación de nubosidad y píxeles no utilizables; a partir de las imágenes depuradas se calcularon índices espectrales de interés (NDVI, SAVI, EVI, NDWI, GNDWI, NDMI, MSI, VSMI, entre otros), que capturan vigor vegetativo, contenido hídrico y estrés fisiológico.

2. Se recopilaron 343 registros diarios provenientes de la estación meteorológica del CIAT, correspondientes al periodo abril 2024 – abril 2025. Las variables incluyeron temperatura máxima, mínima y promedio, precipitación, evapotranspiración de referencia (ET_o), radiación solar, velocidad del viento, humedad relativa y déficit de humedad. Sobre estas series se aplicaron controles de calidad, cálculo de promedios móviles y derivaciones temporales (p. ej., precipitación acumulada 5 días y evapotranspiración acumulada 5 días) para representar de forma dinámica la memoria hídrica atmosférica.

3. Se emplearon 343 registros diarios obtenidos mediante un sensor dieléctrico instalado en la parcela de estudio a una profundidad de 0–10 cm, con mediciones horarias agregadas a valores

diarios promedio. Las variables registradas incluyeron humedad volumétrica, temperatura del suelo y conductividad, además de otras variables auxiliares como punto de rocío y humedad relativa.

Los datos de sensores se usaron como base temporal principal para la combinación de las demás fuentes, garantizando la alineación cronológica entre las observaciones climáticas y los valores derivados de teledetección.

La coherencia temporal entre estas tres fuentes permitió construir una base multitemporal y multivariable adecuada para el entrenamiento y la evaluación de los modelos de predicción de humedad del suelo.

5.2. Preparación de datos.

5.2.1 Procesamiento de Imágenes Satelitales

a) Selección y corrección de imágenes PlanetScope

El procesamiento de imágenes satelitales constituyó la primera fase del flujo de preparación de datos, ya que representó la fuente principal de información espectral utilizada para la predicción de la humedad del suelo. Se emplearon 149 imágenes PlanetScope provistas por el Centro Internacional de Agricultura Tropical (CIAT), correspondientes al periodo comprendido entre periodo abril de 2024 – marzo de 2025. Estas imágenes, de alta resolución espacial (3 m), se seleccionaron cuidadosamente para garantizar una cobertura completa del área de estudio y una mínima interferencia por nubosidad.

En primer lugar, se revisaron los metadatos de cada archivo con el propósito de identificar el número de bandas disponibles, su orden y las fechas de adquisición. Se priorizaron las escenas con ocho bandas multiespectrales coastal blue, blue, green_i, green, yellow, red, red edge y near infrared (NIR) por su mayor riqueza espectral y su capacidad de representar de manera más precisa los procesos biofísicos del sistema suelo–planta. Posteriormente, se aplicó la máscara de nubosidad (UDM o UDM2) incluida en los productos PlanetScope para eliminar píxeles afectados por nubes, sombras o anomalías radiométricas. Esta corrección permitió conservar únicamente las áreas con reflectancias válidas y asegurar la calidad de los cálculos posteriores.

Las imágenes corregidas fueron recortadas espacialmente de acuerdo con el shapefile del CIAT, centrado en la parcela experimental donde se encontraba instalado el sensor de humedad del suelo. Esta delimitación permitió reducir la redundancia de información y concentrar el análisis en el polígono de interés. Además, se reproyectaron los archivos al sistema de coordenadas UTM zona 18N (EPSG:32618), con el fin de garantizar coherencia geométrica entre las diferentes fuentes de datos geoespaciales.

B**) Cálculo de índices espectrales**

Una vez obtenidas las escenas corregidas, se procedió al cálculo de índices espectrales orientados a capturar diferentes dimensiones del estado vegetativo y del contenido hídrico. Entre los principales índices calculados se encuentran el NDVI (Normalized Difference Vegetation Index), indicador clásico del vigor verde y de la biomasa; el EVI (Enhanced Vegetation Index), que corrige los efectos

atmosféricos y del fondo del suelo; y el SAVI (Soil Adjusted Vegetation Index), diseñado específicamente para suelos con cobertura vegetal baja o intermedia. Asimismo, se calcularon índices relacionados con la humedad, tales como el NDWI (Normalized Difference Water Index) y el MSI (Moisture Stress Index), ambos sensibles al contenido de agua en el follaje y en el suelo.

También se incluyeron variantes adicionales como el NDMI, el GNDWI y el VSMI (Vegetation–Soil Moisture Index) que combinan información espectral del infrarrojo cercano, verde y red edge para detectar cambios sutiles en la humedad superficial. Todos los índices fueron calculados a nivel de píxel, y posteriormente se obtuvieron estadísticos zonales (media, mediana, percentiles 25 y 75, y desviación estándar) dentro del buffer de análisis correspondiente al área del sensor. El resultado de este proceso fue una serie temporal multiespectral depurada y homogénea, que representa la evolución espectral de la parcela a lo largo del periodo de estudio. Esta base de datos de índices espectrales constituye el insumo fundamental para la integración con la información climática y los registros de sensores de suelo en las siguientes fases del modelado predictivo.

Tabla 2. Índices calculados.

Índice	Fórmula	Rango	Media	Desviación Estándar	Descripción
NDVI	(NIR - RED)/(NIR + RED)	0.341 – 0.763	0.557	0.109	Índice de vegetación que mide vigor verde
SAVI	[(NIR - RED)/(NIR + RED + 0.5)] × 1.5	0.512 – 1.145	0.835	0.164	Ajusta NDVI por influencia del suelo
EVI	2.5 × (NIR - RED)/(NIR + 6×RED - 7.5×BLUE + 1)	0.758 – 2.486	1.563	0.424	Índice avanzado que minimiza efectos atmosféricos y del suelo
GNDVI	(NIR - GREEN)/(NIR + GREEN)	0.401 – 0.778	0.598	0.083	Variante de NDVI usando banda verde
NDWI	(GREEN - NIR)/(GREEN + NIR)	-0.778 – -0.401	-0.598	0.083	Estima contenido de agua en vegetación
GNDWI	(GREEN - COASTAL_BLUE)/(GREEN + COASTAL_BLUE)	0.019 – 0.242	0.120	0.051	Alternativa de NDWI, sensible al agua libre
NDMI Custom	(REDEDGE - RED)/(REDEDGE + RED)	-0.082 – 0.076	-0.014	0.031	Índice personalizado para humedad en vegetación
MSI	REDEDGE / NIR	0.125 – 0.540	0.297	0.098	Índice de estrés hídrico en plantas
VSMI	NDVI × NDWI	-0.591 – -0.128	-0.347	0.113	Producto de índices para estimar humedad del suelo desde la vegetación

5.2.2 Procesamiento de datos climáticos

El procesamiento de los datos climáticos tuvo como propósito generar variables predictoras robustas para la estimación de la humedad del suelo, incorporando información atmosférica y radiactiva que influye directamente en los procesos de evapotranspiración, infiltración y retención hídrica. Los datos se obtuvieron de la estación meteorológica del Centro Internacional de Agricultura Tropical (CIAT), la cual cuenta con sensores calibrados y registros diarios de alta resolución temporal.

Durante esta fase se procesaron tres archivos principales en formato Excel, correspondientes a registros diarios de precipitación, evaporación y velocidad del viento, radiación solar, y temperaturas del aire. El periodo de observación comprendió periodo abril de 2024 – marzo de 2025.cubriendo un total de 319 días de registro continuo.

Tabla 3. Variables climáticas

Categoría	Variables	Descripción
Temperatura	tmax, tmin, tprom	Temperatura máxima, mínima y promedio diaria
Precipitación	Precipitación diaria, lag 1 y 2, acumulada 3 y 5 días	Registra eventos de lluvia actuales y rezagados para evaluar efecto acumulativo
Evapotranspiración	ET0, evaporación	Estimación de pérdida de agua por transpiración vegetal y evaporación directa
Otras	velocidad del viento, radiación solar, déficit de humedad	Factores ambientales que afectan la evaporación y el balance hídrico

- **Integración y limpieza de datos**

En primer lugar, se realizó una estandarización de los nombres de columnas, formatos de fecha y unidades de medida, asegurando la coherencia entre las tres bases de datos. Posteriormente, se efectuó la unificación por fecha, generando una única base consolidada denominada *clima_combinado.csv*, que sirvió como insumo para la creación de variables derivadas y análisis temporales.

Durante el proceso de depuración, se aplicaron controles de coherencia interna y física, verificando que las temperaturas máximas fuesen siempre mayores que las mínimas y que no existieran valores negativos en precipitación o evapotranspiración. Los datos faltantes o inconsistentes se imputaron mediante interpolación lineal o promedios móviles, garantizando la continuidad de las series temporales.

- **Cálculo de la Evapotranspiración de Referencia (ET_o)**

La evapotranspiración de referencia (ET_o) se estimó mediante la ecuación de Hargreaves y Allen (2003), la cual utiliza las temperaturas máxima y mínima, la radiación solar extraterrestre y la latitud del sitio de estudio. Para el caso del CIAT se consideró la latitud 3.497° N, permitiendo estimar la demanda evaporativa atmosférica de forma diaria. Esta variable constituye un indicador esencial del consumo potencial de agua por parte del cultivo bajo condiciones climáticas locales.

- **Generación de variables derivadas**

Con el fin de incorporar la influencia temporal acumulativa de la precipitación en el contenido hídrico del suelo, se generaron variables derivadas a partir de las series originales:

- Rezagos de precipitación (lag_1 y lag_2): representan la lluvia de uno y dos días anteriores.
- Acumulados móviles ($acum_3$ y $acum_5$): reflejan la precipitación acumulada en los tres y cinco días previos.
- Balance hídrico mensual (Precipitación − ET_o): permite identificar periodos de déficit o superávit de agua.

5.2.3 Procesamiento de datos de suelo

En esta etapa se consolidaron y procesaron las lecturas obtenidas mediante sensores instalados en el suelo agrícola, con el fin de generar variables observadas de alta frecuencia que sirvan como referencia directa para el modelado de la humedad volumétrica. Estas mediciones constituyen la variable dependiente y son esenciales para evaluar la precisión de los modelos predictivos.

Tabla 4. Variables del suelo.

Variable principal	Variables	Descripción
Humedad	Humedad del suelo, Humedad volumétrica, Punto de rocío, Humedad relativa	Indicadores clave de contenido hídrico y condiciones de condensación
Temperatura	Temperatura del suelo, Temperatura a1g, a1r, s1	Valores a diferentes profundidades o ubicaciones específicas en el perfil del suelo
Otras	Conductividad, Déficit de humedad	Variables complementarias para evaluar salinidad y estrés hídrico potencial

Resultados del Procesamiento

- Se trabajó con datos horarios, los cuales fueron consolidados en promedios diarios para facilitar la integración temporal con imágenes e información climática.

- Se realizó una validación de rangos y consistencia para identificar posibles fallos de sensores, lecturas atípicas y errores de captura.
- El resultado fue una base robusta de observaciones diarias de humedad del suelo, que sirve como referencia empírica para entrenamiento y validación de modelos de predicción.

Los datos generados por los sensores proporcionaron una representación directa del comportamiento de la humedad en el suelo, permitiendo calibrar y evaluar los modelos construidos a partir de datos indirectos (clima e imágenes satelitales). La resolución horaria inicial y su posterior agregación diaria aseguran una alta fidelidad temporal sin comprometer la capacidad de análisis.

5.3 Integración y construcción del Dataset

Una vez procesadas las tres fuentes de información, se procedió a la construcción del dataset combinado, que constituye la base analítica del modelo predictivo de humedad del suelo. Esta fase tuvo como propósito integrar, depurar y sincronizar temporalmente los datos, asegurando coherencia estructural y continuidad entre todas las variables.

Integración temporal y estructural de las bases

La integración se realizó utilizando la fecha como llave primaria, permitiendo la alineación exacta de los registros provenientes de las tres fuentes. Cada fila del dataset resultante representa un día de observación con sus respectivas variables climáticas, índices espectrales e indicadores de humedad del suelo.
El procedimiento de integración comprendió las siguientes etapas:

1. Alineación temporal: sincronización de las tres bases a nivel diario, considerando únicamente las fechas en las que existía información válida en al menos dos fuentes.
2. Unión horizontal: combinación de los datasets mediante operaciones *merge* y *join*, integrando las columnas de variables espectrales (NDVI, SAVI, EVI, GNDVI, NDWI, GNDWI, NDMI_Custom, NDRE, MSI, VSMI), meteorológicas (precipitacion, evaporacion, velocidad_viento, radiacion_solar, tmax, tmin, tprom, et0, precipitacion_lag1, precipitacion_lag2, precipitacion_acum_3dias, precipitacion_acum_5dias, evapotraspitacion acum de 3 a 15 días) y de sensores (humedad_suelo, temperatura_suelo, punto_rocio, deficit_humedad, humedad_relativa, conductividad, humedad_volumetrica, radiacion_solar, temperatura_a1g, temperatura_a1r, temperatura_s1).

3. Validación de integridad: verificación de consistencia de fechas, eliminación de duplicados y detección de valores nulos críticos en las variables dependiente e independientes.

El resultado fue una base de datos combinada denominada base_combinada_total.csv, que integra 343 registros diarios con 53 variables correspondientes al periodo abril de 2024 – marzo de 2025.

Tratamiento de valores faltantes e imputación

Durante la fusión de las bases se detectaron valores faltantes en algunas variables, principalmente asociados a días sin imágenes satelitales con 7,0% de valores faltantes o interrupciones puntuales de los sensores. Para evitar pérdida de información y mantener la secuencia temporal, se aplicó un protocolo sistemático de imputación de datos, compuesto por tres etapas principales:

1. Identificación de celdas faltantes: localización de valores nulos o fuera de rango en las variables clave.
2. Imputación controlada: reemplazo de valores faltantes mediante interpolación lineal y promedios móviles, según la naturaleza de cada variable (climática o espectral).
3. Validación de coherencia: comparación estadística y visual antes y después de la imputación, asegurando que las series mantuvieran su comportamiento temporal y no introdujeran sesgos artificiales.

El producto final de este proceso fue la base combinada imputada con 319 registros, lista para ser utilizada en las fases de análisis exploratorio y modelado predictivo.

5.4 Análisis exploratorio de datos (EDA)

5.4.1 Diagnóstico inicial y control de calidad

Antes de realizar el análisis exploratorio, se llevó a cabo una verificación exhaustiva de la calidad de los datos con el fin de garantizar la integridad de la información utilizada en el modelado. Este proceso incluyó la detección de duplicados, la revisión de coherencia temporal y la evaluación de valores faltantes en todas las variables que integran la base de datos combinada.

El análisis de valores nulos cuantificó el número de registros ausentes por variable y calculó el porcentaje correspondiente. Los resultados, resumidos en el archivo valores_nulos.csv, muestran que ninguna de las 53 variables contiene valores faltantes (0 %), evidenciando una base de datos completamente íntegra en este aspecto.

Asimismo, se verificó que las series temporales mantienen una continuidad diaria sin interrupciones en las fechas, y que los nombres de columnas y unidades de medida son consistentes entre las distintas fuentes de información (imágenes satelitales, variables climáticas y sensores de suelo). No se identificaron registros duplicados ni inconsistencias en los rangos esperados para las variables medidas.

5.4.2 Estadísticas descriptivas generales

El análisis descriptivo se aplicó a las 53 variables integradas en la base de datos unificada con el propósito de caracterizar la distribución, rango y variabilidad de los indicadores climáticos, de vegetación y de sensores de suelo. Se calcularon medidas de tendencia central (media y mediana), de dispersión (desviación estándar y rango intercuartílico) y de posición (mínimo, máximo y percentiles 25 % y 75 %).

Las variables de sensores de suelo Tabla 5, mostraron una adecuada estabilidad: la humedad del suelo presentó una media de 17,48 %, con baja dispersión (σ = 0,94), lo que indica condiciones relativamente homogéneas de humedad durante el periodo de medición. La temperatura del suelo promedió 24,7 °C, con valores comprendidos entre 21,1 °C y 27,7 °C. El punto de rocío se mantuvo alrededor de 20,4 °C, coherente con los valores de humedad relativa promedio del 79 %.

Tabla 5. Estadística de las variables de sensores.

Variable	Media	Desv.Std	Mínimo	Máximo
Humedad suelo	17.47	0.94	11.08	19.9
Temperatura suelo	24.70	1.25	21.1	27.9
Punto rocío	20.39	0.90	14.1	22.5
Déficit humedad	5.97	1.78	0.9	11.6
Humedad relativa	79.46	6.41	57.2	96.4
Conductividad	2.13	0.76	0.0	3.3
Humedad volumétrica	44.36	1.19	0.0	58.2
Radiación solar	66,06	7.36	7.8	283.0
temperatura_a1g	25.76	1.86	21.7	31.1
temperatura_a1r	25.58	1.52	22.3	30.0
temperatura_s1	25.31	1.80	22.1	30.5

En el conjunto de variables climáticas Tabla 6, la precipitación diaria mostró un promedio bajo (1,55 mm día^{-1}) y una alta asimetría, reflejando la naturaleza intermitente de las lluvias en el área de estudio. La temperatura máxima alcanzó en promedio 32 °C y la mínima 22 °C, mientras que la evapotranspiración de referencia (ET_o) tuvo una media de 4,87 mm día^{-1}, indicando una demanda evaporativa moderada.

Tabla 6. Estadísticas variables climáticas.

Variable	Media	Desv. Est.	Mínimo	Máximo
Precipitación (mm)	1.55	5.08	0.0	31.00
Evaporación (mm)	5.15	4.54	0.8	73.70
Velocidad del viento (km/h)	52.88	18.16	0.0	154.00

Temperatura máxima (°C)	32.52	1.77	26.4	37.20
Temperatura mínima (°C)	22.35	0.90	18.8	24.40
Temperatura promedio (°C)	27.40	0.97	24.0	30.10
Evapotranspiración (ET_o) (mm/día)	4.87	0.64	2.95	6.10
Precipitación Lag 1 (mm)	1.54	5.00	0.0	31.00
Precipitación Lag 2 (mm)	1.53	5.00	0.0	31.00
Precipitación Acum 3 días (mm)	4.63	9.94	0.0	42.00
Precipitación Acum 5 días (mm)	8.10	13.73	0.0	52.60

En cuanto a los índices espectrales de vegetación Tabla 7 , los valores promedio de NDVI (0,56), SAVI (0,83) y EVI (1,56) evidencian una cobertura vegetal activa y saludable durante la mayor parte del periodo analizado. Los índices NDWI (-0,60) y MSI (0,30) reflejan niveles de humedad foliar y estrés hídrico moderados, respectivamente, mientras que el NDRE (0,42) sugiere vigor vegetativo alto, típico de cultivos en fase activa de crecimiento.

Tabla 7. Estadística de índices calculados.

Índice	Media	Desv. Est.	Mínimo	Máximo
NDVI	0.56	0.11	0.34	0.76
SAVI	0.83	0.16	0.51	1.15
EVI	1.56	0.42	0.76	2.49
GNDVI	0.60	0.08	0.40	0.78
NDWI	-0.60	0.08	-0.78	-0.40
GNDWI	0.12	0.05	0.02	0.24
NDMI Custom	-0.01	0.03	-0.08	0.08
MSI	0.30	0.10	0.12	0.54
VSMI	-0.35	0.11	-0.59	-0.13
NDRE	0.42	0.09	0.25	0.63

Por último, las variables derivadas de evapotranspiración mostraron un incremento progresivo con la extensión de la ventana de cálculo, coherente con la acumulación de pérdida hídrica. Por ejemplo, la evapotranspiración acumulada a 15 días ($ET_{15}d$) presentó una media de 73,2 mm, con desviación estándar de 6,8 mm, y un rango entre 61,9 mm y 85,6 mm.

En conjunto, las estadísticas descriptivas confirman que el conjunto de datos presenta coherencia física, estabilidad térmica y consistencia entre las variables hidrometeorológicas y espectrales, lo que proporciona una base sólida para los análisis correlacionales y de modelado predictivo desarrollados en las secciones posteriores.

5.4.3 Análisis univariado

Análisis de variables climáticas

Las variables climáticas presentaron comportamientos diferenciados en cuanto a su dispersión y forma de distribución. Los boxplots Figura 5 mostraron la existencia de valores extremos en la precipitación, la evaporación y la velocidad del viento, evidenciando una alta variabilidad diaria asociada a condiciones atmosféricas fluctuantes. En contraste, las variables térmicas (temperatura máxima, mínima y promedio) presentaron una dispersión moderada y rangos consistentes con los observados en zonas tropicales bajas, oscilando entre 19 °C y 37 °C. La variable evapotranspiración de referencia (ET_o) mostró una variabilidad intermedia, con la mayoría de los valores comprendidos entre 3 y 6 mm·día^{-1}.

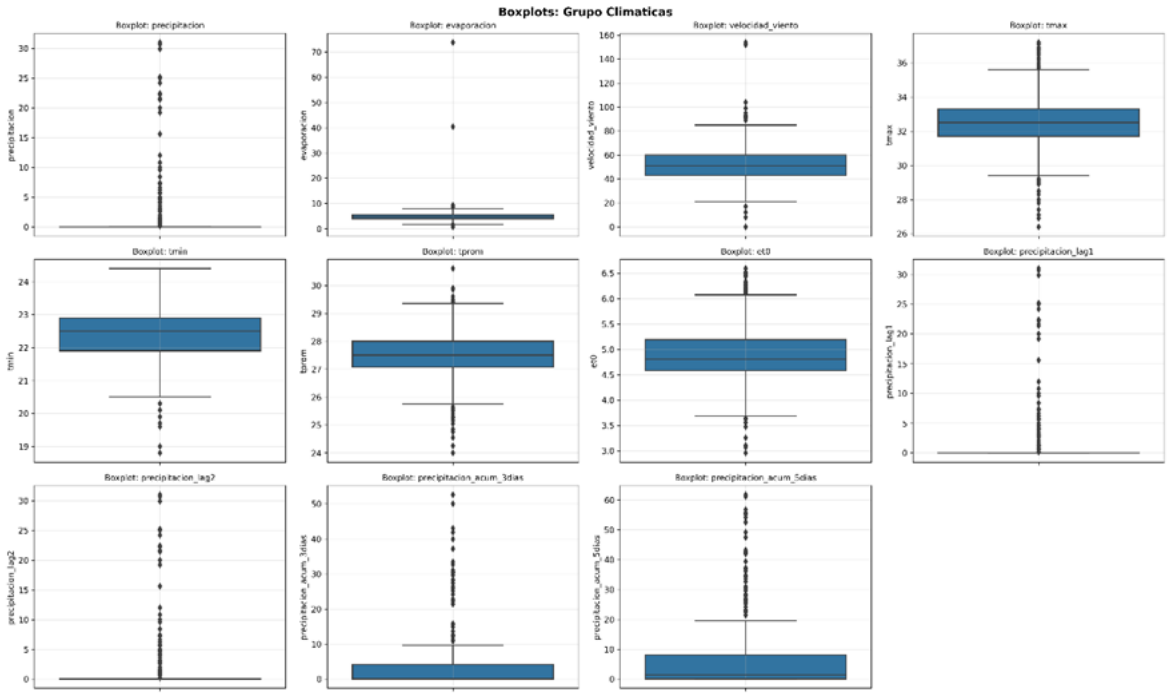

Figura 5. Boxplot variables climáticas.

El análisis de distribuciones Figura 6 confirmó la marcada asimetría positiva en las variables de precipitación y sus derivados (lags y acumulados), donde predominan los valores cercanos a cero y ocurrencias esporádicas de eventos intensos. La evaporación y la velocidad del viento evidenciaron colas derechas prolongadas, indicando la presencia de días con condiciones secas o ventosas inusuales. En contraste, las temperaturas y la ET_o se aproximaron a una distribución normal unimodal, lo que sugiere estabilidad térmica en el periodo de estudio. Estas características reflejan la estacionalidad hídrica propia del Valle del Cauca, con predominio de días secos intercalados con

eventos de lluvia de alta intensidad y corta duración.

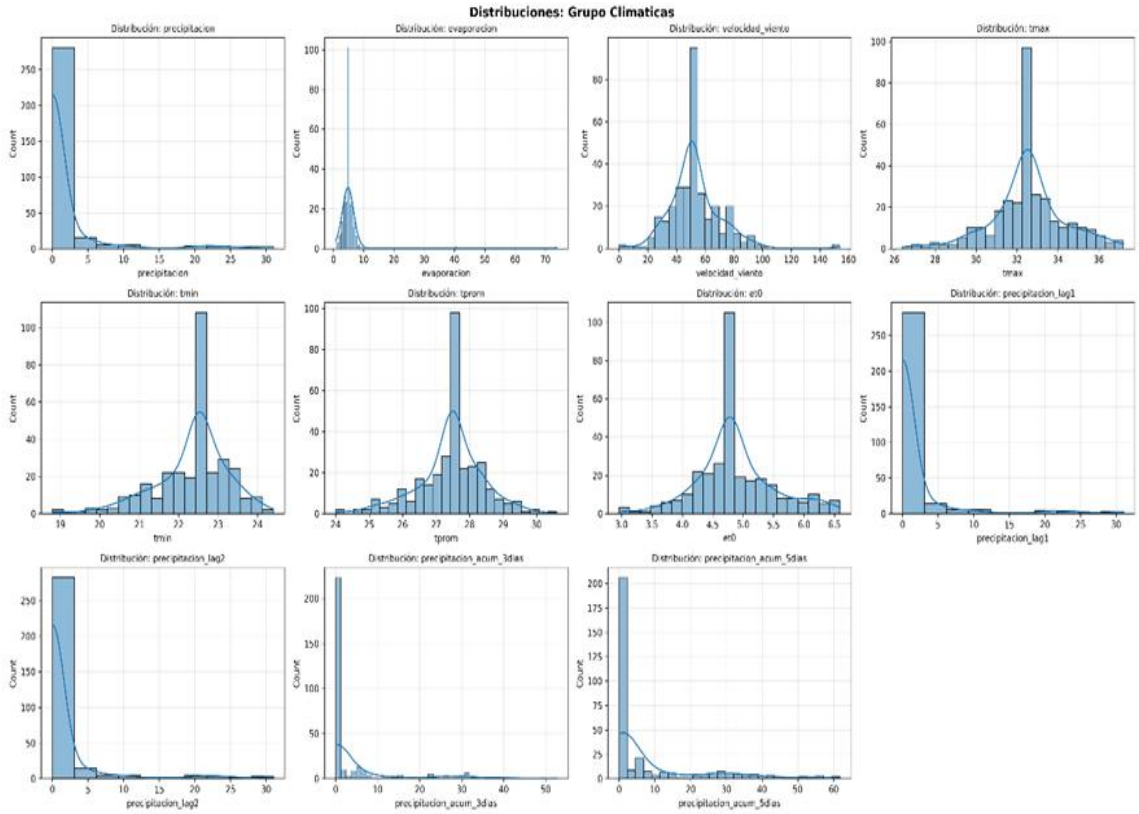

Figura 6.Distribución de variables climáticas.

Análisis de variables de evapotranspiración acumulada

El conjunto de variables asociadas a la evapotranspiración (máxima, mínima, promedio y acumulada en ventanas móviles de 3, 5, 7, 11 y 15 días) presentó distribuciones relativamente uniformes y sin presencia significativa de valores atípicos. Los boxplots Figura 7 mostraron una dispersión homogénea entre las diferentes escalas temporales, lo cual sugiere una alta correlación entre ellas. Este comportamiento se explica porque las ventanas móviles reflejan la misma dinámica atmosférica con leves diferencias de suavizado temporal.

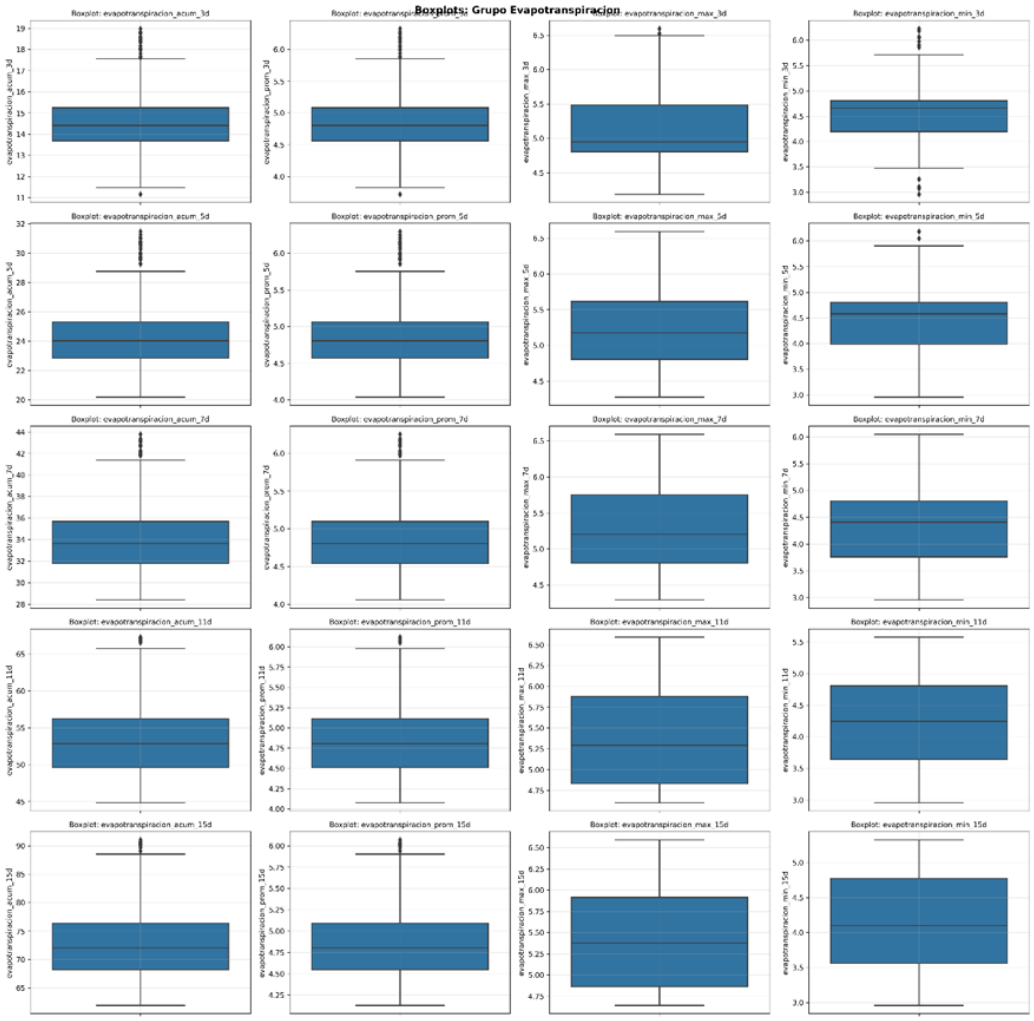

Figura 7.Boxplot de evapotranspiración.

Las distribuciones Figura 8 presentaron una ligera asimetría hacia la derecha, especialmente en los periodos cortos (3 a 5 días), indicando que durante ciertos intervalos se registraron valores superiores al promedio, posiblemente asociados a aumentos puntuales de temperatura o radiación solar. En general, la evapotranspiración acumulada mostró un patrón unimodal estable, lo cual coincide con la naturaleza continua del proceso de pérdida de agua desde el suelo y la vegetación hacia la atmósfera. El comportamiento consistente entre las distintas escalas temporales refleja la coherencia interna del conjunto de variables y la estabilidad del régimen energético en la zona de estudio.

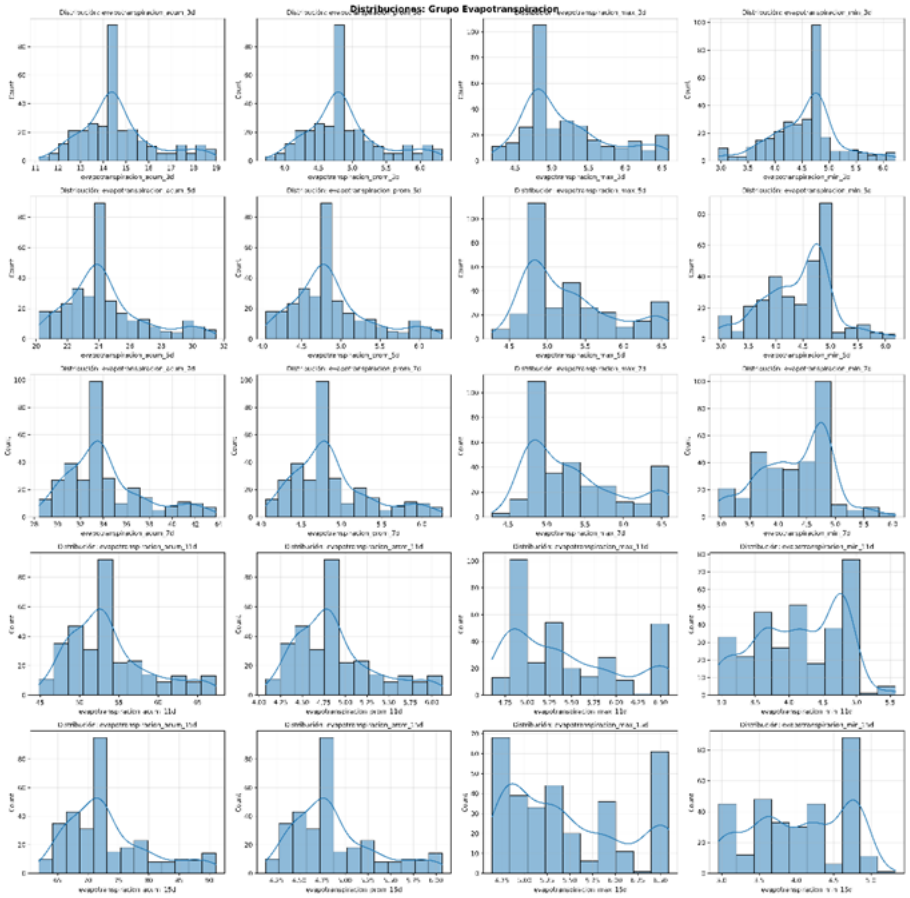

Figura 8. Distribución de evapotranspiración.

Análisis de índices de vegetación

Los índices espectrales derivados de las imágenes PlanetScope evidenciaron un comportamiento estable y coherente con la fenología del cultivo. En los boxplots Figura 9, se observa los índices NDVI, SAVI, GNDVI y NDRE mostraron una distribución concentrada, sin valores extremos significativos, con medianas cercanas a 0.6, lo que indica una vegetación activa y con buena cobertura foliar. El índice EVI presentó una mayor amplitud, con valores que oscilaron entre 1.0 y 2.5, reflejando su mayor sensibilidad a la densidad de biomasa y a la variabilidad estructural del dosel vegetal.

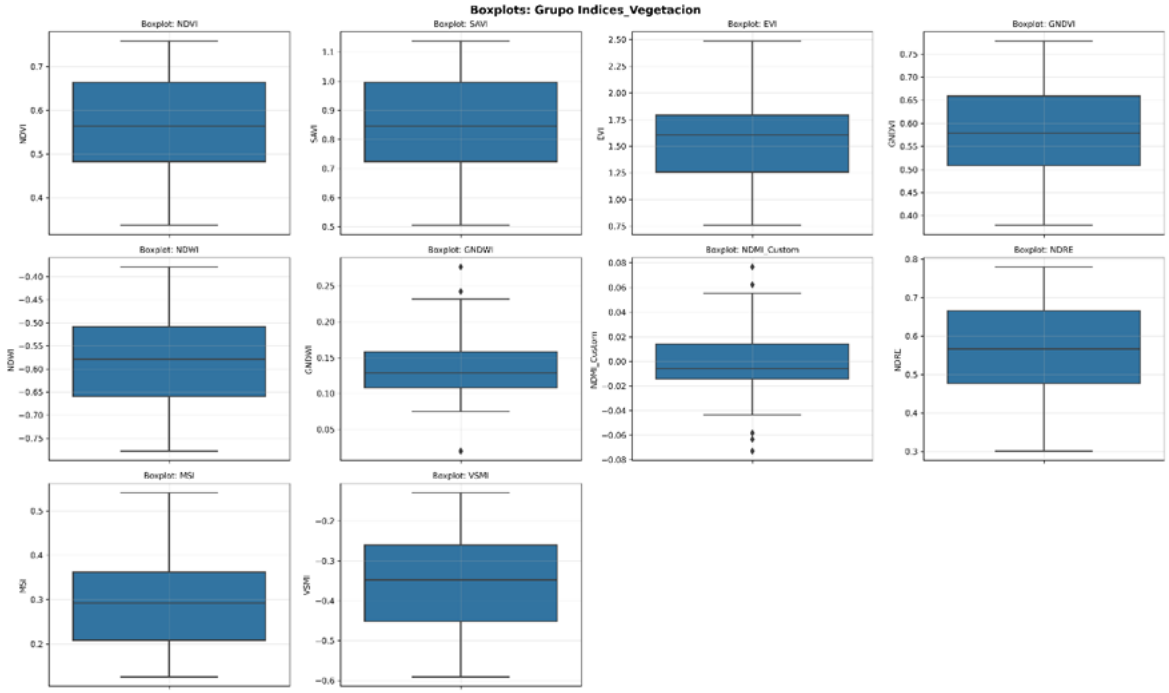

Figura 9. Boxplot índices de vegetación.

Por otra parte, los índices relacionados con humedad superficial (NDWI, MSI, VSMI) exhibieron distribuciones más amplias y asimétricas Figura 10. En el caso del NDWI, se observó un predominio de valores negativos, lo cual es característico en superficies vegetadas sin cuerpos de agua libres, mientras que el MSI mostró una leve cola derecha, consistente con condiciones de estrés hídrico moderado en ciertos periodos. Los histogramas confirmaron que los índices presentan comportamientos unimodales, aunque algunos como el NDWI y el VSMI tienden a la bimodalidad leve, posiblemente por la alternancia de fases húmedas y secas del ciclo agrícola. En conjunto, estos resultados muestran la consistencia radiométrica de las imágenes y su capacidad para reflejar la respuesta fisiológica del cultivo ante la variabilidad climática.

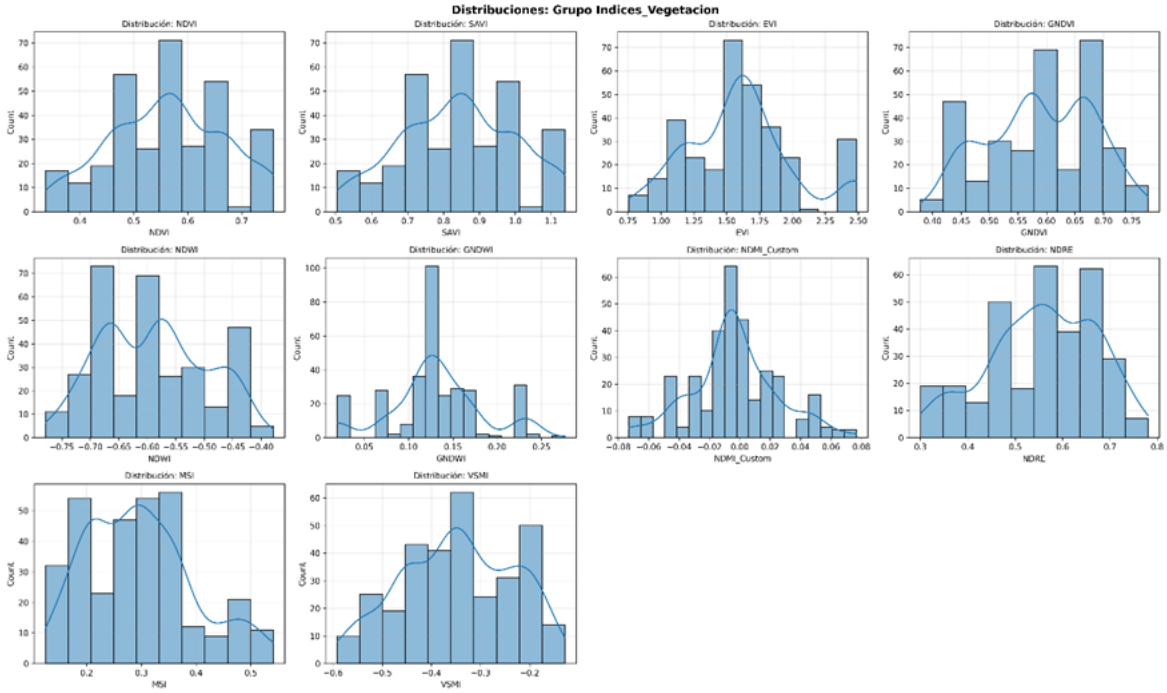

Figura 10. Distribución de índices de vegetación.

Análisis de variables medidas por sensores

Las variables provenientes de los sensores de campo mostraron comportamientos consistentes con las condiciones agroclimáticas registradas. Los boxplots del grupo de sensores revelaron valores atípicos puntuales en las variables de humedad volumétrica y radiación solar, lo que sugiere episodios de saturación de suelo y días de alta irradiancia respectivamente. La variable humedad del suelo, utilizada como variable dependiente, presentó una distribución ligeramente asimétrica hacia la izquierda, con la mayoría de los valores concentrados entre 16 % y 18 %, representando condiciones de humedad moderada.

Las variables temperatura del suelo (a1g, a1r, s1) y temperatura del aire presentaron distribuciones unimodales cercanas a la normalidad, con valores promedio entre 25 °C y 27 °C, reflejando la estabilidad térmica del perfil superficial del suelo. La humedad relativa y el punto de rocío mostraron distribuciones normales con ligeras colas en los extremos, indicando variaciones en la saturación del aire. Por su parte, la conductividad eléctrica exhibió una mayor dispersión y multimodalidad, posiblemente relacionada con diferencias en el contenido iónico del suelo o variaciones temporales en la humedad. Finalmente, la radiación solar mostró una fuerte asimetría positiva, con la mayoría de observaciones concentradas en valores bajos y una menor frecuencia de picos altos, lo que refleja las fluctuaciones diarias de nubosidad y la incidencia de radiación directa.

5.4.4 Análisis bivariado

El análisis bivariado permitió examinar la relación entre la humedad volumétrica del suelo y las principales variables climáticas, espectrales y edáficas. En la Figura 11 se presentan los diagramas de dispersión que ilustran las tendencias observadas entre la humedad y cada predictor considerado. Paralelamente, se calcularon los coeficientes de correlación de Pearson Tabla 8 para cuantificar la fuerza y dirección de dichas relaciones.

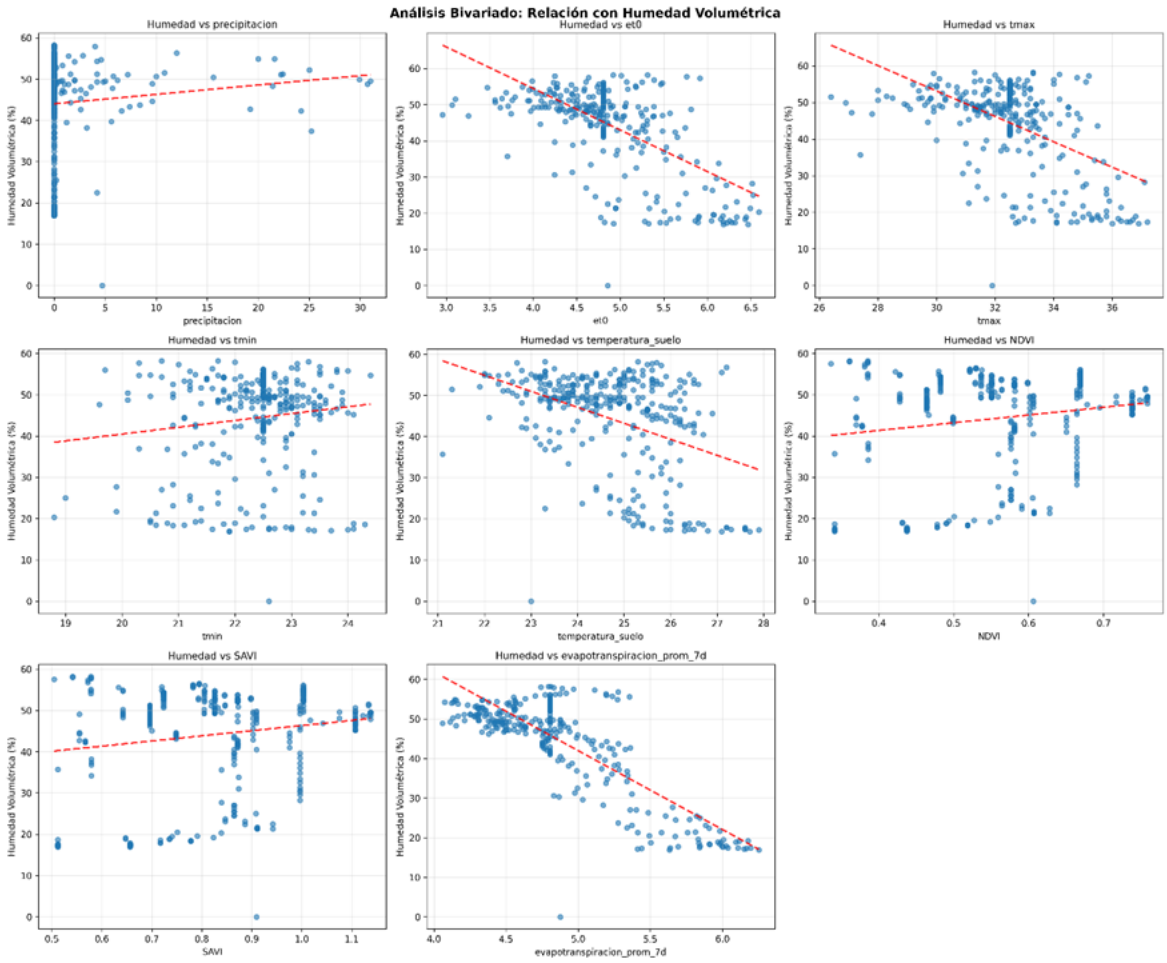

Figura 11. Relaciones entre humedad volumétrica.

En términos generales, se identificaron relaciones inversas moderadas a fuertes con las variables evapotranspiración de referencia (ET_o), temperatura máxima (Tmax) y temperatura del suelo, lo cual sugiere que el incremento de la demanda evaporativa y las temperaturas elevadas tienden a reducir el contenido de agua en el suelo. En contraste, se evidenciaron relaciones positivas entre la humedad y las variables asociadas a la precipitación (lag y acumulados), así como con los índices espectrales de vegetación como NDVI y SAVI, que reflejan un estado vegetativo más vigoroso y mayor cobertura

verde cuando la humedad del suelo es alta.

El comportamiento de Tmin mostró una correlación débilmente positiva, indicando que temperaturas nocturnas ligeramente mayores podrían favorecer la retención de humedad. Sin embargo, las variables relacionadas con la evapotranspiración promedio a siete días y la ET_o diaria presentaron las correlaciones negativas más marcadas, coherentes con un mayor consumo hídrico del suelo bajo condiciones atmosféricas secas y cálidas.

Tabla 8. Correlaciones de Pearson entre humedad volumétrica.

Variable	r (Pearson)	Relación
Precipitación	+0.28	Positiva débil
ET_o (Hargreaves)	-0.61	Negativa moderada
Tmax	-0.54	Negativa moderada
Tmin	+0.21	Positiva débil
Temperatura del suelo	-0.58	Negativa moderada
NDVI	+0.33	Positiva moderada
SAVI	+0.30	Positiva moderada
Evapotranspiración prom. 7 días	-0.64	Negativa fuerte

Estos resultados confirman la alta sensibilidad de la humedad del suelo frente a las condiciones térmicas y evaporativas, así como la utilidad de los índices de vegetación para capturar indirectamente el estado hídrico superficial. En consecuencia, las variables ET_o, evapotranspiración promedio a siete días, Tmax y temperatura del suelo se consideran predictores claves para los modelos de estimación y predicción de humedad.

5.4.5 Análisis de correlaciones

a) Correlaciones positivas relevantes

En la Figura 12 se observa las diferentes correlaciones positivas con la humedad volumétrica del suelo.

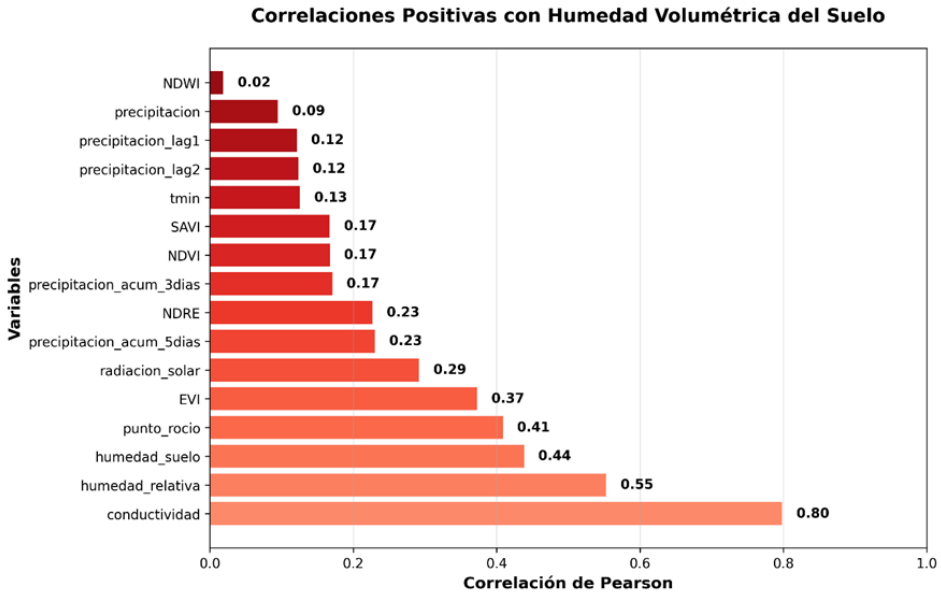

Figura 12. Correlaciones positivas.

El conjunto de correlaciones positivas muestra la coherencia entre la disponibilidad de agua en el suelo y variables que reflejan condiciones atmosféricas húmedas y vegetación activa.

- La conductividad eléctrica (r = 0.80) presentó la correlación más alta, indicando que la retención de humedad aumenta la conductividad iónica del suelo.
- Las variables atmosféricas como humedad relativa (r = 0.55) y punto de rocío (r = 0.41) mostraron respuestas directas, confirmando el acoplamiento entre la humedad del aire y del suelo.
- Los índices de vegetación (EVI, NDVI, SAVI, NDRE) reflejan la mayor reflectancia en el infrarrojo cercano durante periodos de humedad elevada, evidenciando vigor vegetativo asociado a suelos húmedos.
- La precipitación acumulada a 3 y 5 días también mantiene una relación positiva (r = 0.17–0.23), lo que indica que el contenido hídrico responde a lluvias recientes con un desfase temporal de pocos días.

Estas correlaciones validan el comportamiento físico esperado de la dinámica hídrica del sistema suelo-planta-atmósfera y sirvieron como criterio para priorizar variables predictoras relevantes en la modelación.

b) Correlaciones negativas destacadas

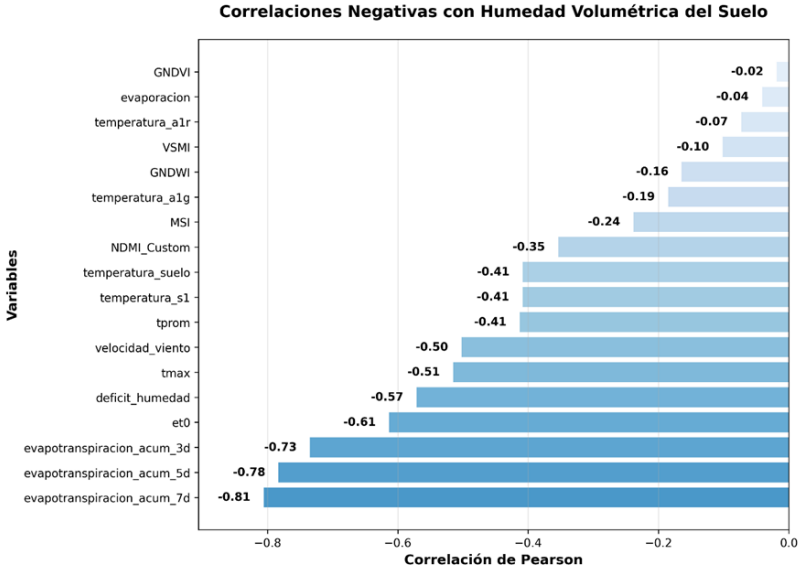

Figura 13.Correlaciones negativas.

Las correlaciones negativas indican procesos de pérdida o demanda de agua que actúan en sentido opuesto a la humedad del suelo.

- Las variables de evapotranspiración (ET_o y sus acumulados entre 3 y 15 días) presentan las correlaciones más fuertes (r ≈ –0.80), evidenciando que el incremento de la demanda evaporativa atmosférica provoca una rápida disminución en el contenido hídrico del suelo.
- Las temperaturas del aire y del suelo, junto con la velocidad del viento y el déficit de humedad atmosférica, confirman el rol de la energía térmica y la turbulencia como controladores del secado del perfil.
- En el plano espectral, los índices NDMI, MSI y VSMI reflejan un patrón inverso esperado: valores más bajos cuando la humedad del suelo es alta, validando su utilidad como indicadores remotos del estrés hídrico.

Este conjunto de correlaciones negativas es consistente con la física del balance hídrico, y su magnitud respalda el uso de las variables de evapotranspiración y temperatura como predictores claves en la modelación de la dinámica hídrica del suelo.

c) Correlaciones fuertes y redundancias

Tabla 9. Correlaciones fuertes y redundancias entre variables predictoras.

Grupo de variables asociadas	Correlación (r)	Tipo	Decisión
NDVI – SAVI – EVI – NDRE	0.98–0.99	Alta positiva	Mantener NDVI como representativo
VSMI – NDWI – MSI – GNDVI	±0.89–0.97	Alta negativa	Mantener NDMI o MSI según ajuste físico
ET_o – Evapotranspiración 3–15d (acum/prom/max)	0.73–0.99	Alta positiva	Mantener evapotranspiración promedio 7d

Temperatura máx. – Tprom – Temperatura suelo	0.71–0.88	Alta positiva	Mantener Tmax como indicador térmico
Humedad suelo – Punto de rocío	0.99	Alta positiva	Redundancia, mantener humedad_suelo
Déficit de humedad – Humedad relativa	-0.97	Alta negativa	Mantener déficit_humedad
Precipitación acumulada 3d – 5d	0.84	Alta positiva	Mantener acumulado 5d
Temperatura aérea a1g – a1r – s1	0.84–0.90	Alta positiva	Mantener temperatura_s1
Conductividad – Evapotranspiración (varias)	-0.72 a -0.77	Alta negativa	Mantener conductividad
Tmáx – ET_o	0.94	Alta positiva	Mantener Tmax
Evapotranspiración min/max (3–15d)	0.80–0.91	Alta positiva	Mantener evapotranspiración promedio 7d

El análisis de correlaciones fuertes permitió identificar bloques de variables colineales que representan procesos similares dentro del sistema suelo–planta–atmósfera.

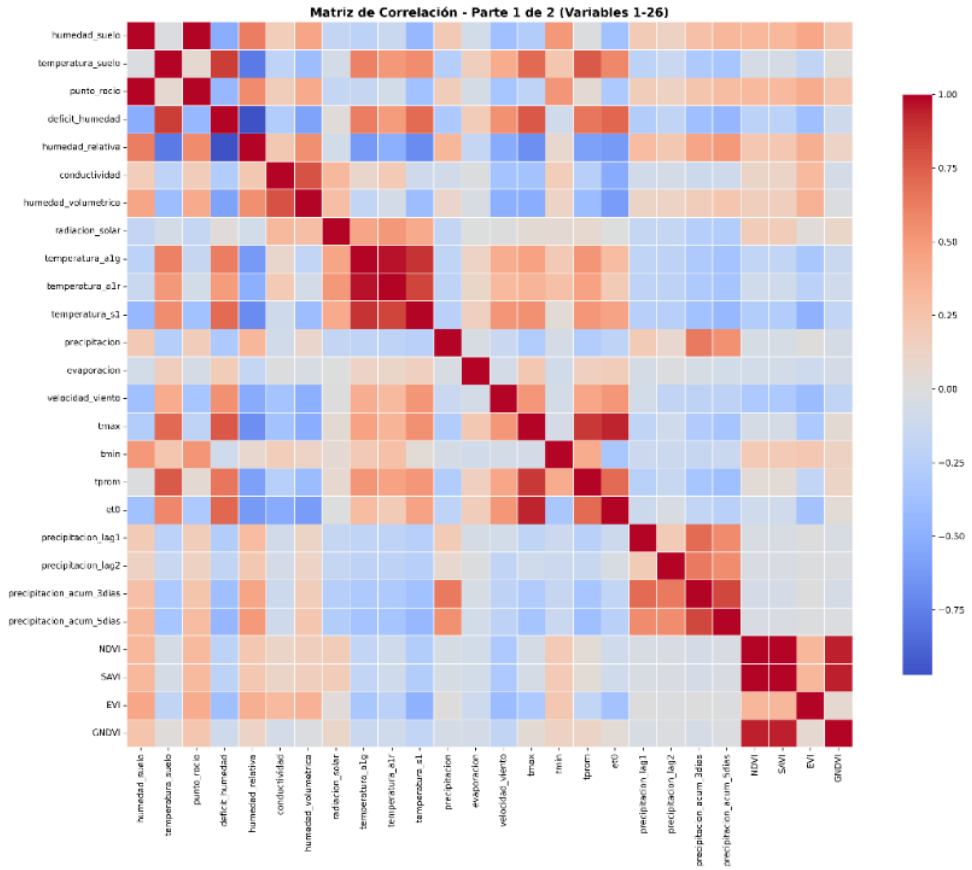

Figura 14. Correlación de todas la variables parte 1.

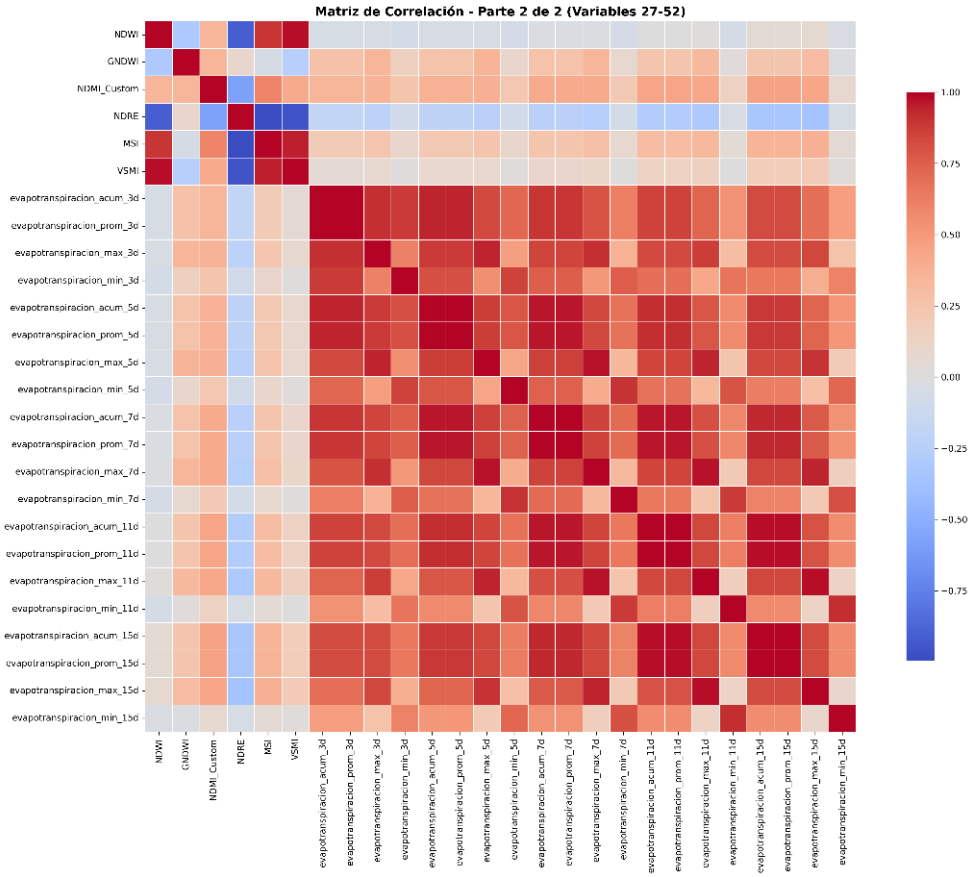

Figura 15. Correlaciones de todas las variables parte 2.

- El grupo de índices espectrales (NDVI, SAVI, EVI y NDRE) presentó correlaciones superiores a 0.98, lo que evidencia una redundancia espectral derivada de la sensibilidad compartida a la reflectancia del infrarrojo cercano. Se seleccionó NDVI como índice representativo del vigor vegetal.
- Las variables de evapotranspiración (acumulada, promedio y máxima entre 3 y 15 días) mostraron correlaciones >0.90 entre sí, lo que indica que todas capturan el mismo proceso de demanda evaporativa. Se retuvo la evapotranspiración promedio a 7 días como variable resumen.
- Las temperaturas del aire y del suelo, así como los índices térmicos derivados, mantienen correlaciones positivas (r ≈ 0.85), lo que confirma su dependencia energética y justifica conservar solo Tmax.
- Por otra parte, el déficit de humedad atmosférica mostró una correlación inversa casi perfecta con la humedad relativa (r = –0.97), lo que respalda su exclusión para evitar duplicidad.
- En el bloque hídrico, las precipitaciones acumuladas a distintos intervalos de días (3 y 5) presentaron r = 0.84, siendo suficiente conservar el acumulado a 5 días.

5.4.6 Diagnóstico de colinealidad

Los resultados confirman que la mayoría de las variables presentan colinealidad baja (VIF < 5), lo que garantiza independencia estadística suficiente para su inclusión en el modelo. No obstante, se identificaron dos excepciones críticas:

- NDVI y SAVI mostraron valores de VIF del orden de 10^9, lo que indica redundancia casi perfecta entre ambos índices, atribuida a su alta similitud estructural (ambos derivan del infrarrojo cercano y el rojo).
- Esta redundancia fue previamente evidenciada en la Figura 16, donde se reportó una correlación r = 0.9999 entre ambos.

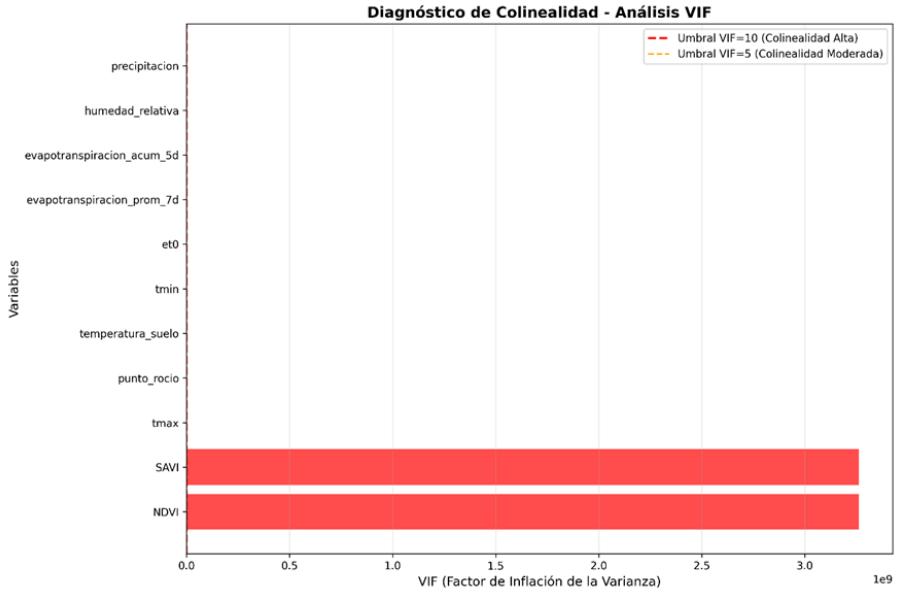

Figura 16. Colinealidad.

En consecuencia, se optó por eliminar SAVI y mantener NDVI como variable representativa del vigor de la vegetación.
El resto de las variables (temperaturas, humedad, precipitación y evapotranspiración) se conservaron, dado que sus valores de VIF fueron inferiores a 2, lo cual indica independencia adecuada.

5.4.7 Análisis de la variable objetivo

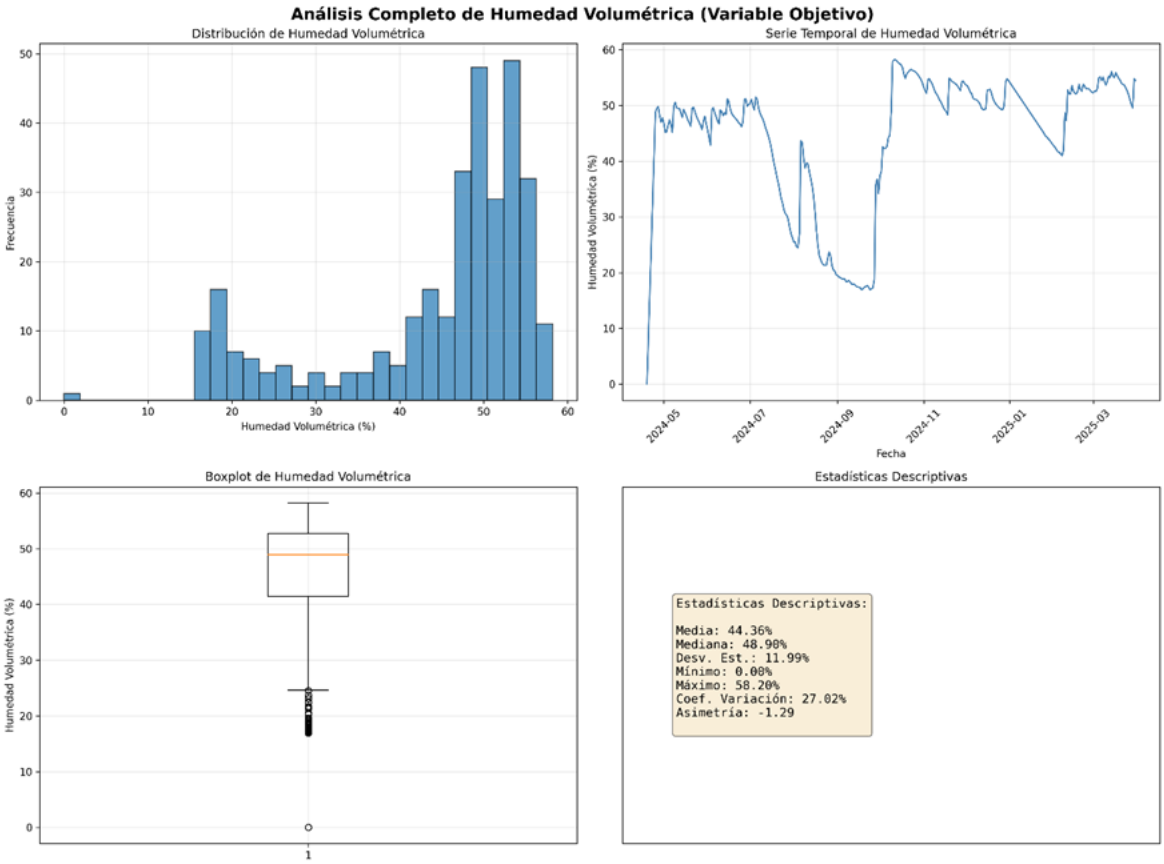

Figura 17. Análisis de variable objetivo.

La humedad volumétrica del suelo (θv) fue definida como la variable objetivo del modelo predictivo, expresada en porcentaje (%). Representa la proporción de volumen de agua presente en el suelo respecto al volumen total de suelo, siendo un indicador directo del estado hídrico de la parcela.

Distribución y estadísticos descriptivos

El histograma y el boxplot Figura 17 muestran una distribución bimodal con dos concentraciones principales de valores entre 20–25 % y 45–55 %, evidenciando diferentes fases de humedad del suelo durante el periodo analizado (mayo de 2024 a marzo de 2025).

De acuerdo con las estadísticas descriptivas:

- Media: 44.36 %
- Mediana: 48.90 %
- Desviación estándar: 11.99 %
- Mínimo: 0.00 %
- Máximo: 58.20 %
- Coeficiente de variación (CV): 27.02 %

- Asimetría: –1.29

El valor negativo de la asimetría indica una ligera tendencia hacia la izquierda, producto de algunos periodos con niveles muy bajos de humedad, posiblemente asociados a eventos de alta evaporación o baja precipitación.

Comportamiento temporal

La serie temporal evidencia una alta variabilidad estacional a lo largo del periodo estudiado:

- Entre mayo y septiembre de 2024, se registraron descensos significativos, alcanzando valores mínimos (5–10 %), asociados a condiciones secas y elevadas tasas de evapotranspiración.
- A partir de octubre de 2024, se observó una recuperación progresiva de la humedad, superando el 50 % hacia noviembre, coincidiendo con el incremento de la precipitación acumulada.
- En los meses iniciales de 2025, la humedad se mantuvo relativamente estable (50–55 %), lo que sugiere un equilibrio entre recarga y demanda evaporativa.

El conjunto de resultados permite concluir que la humedad volumétrica presenta una variabilidad temporal y amplitud significativa, con predominio de valores medios–altos en la mayoría del año, aunque con episodios marcados de déficit hídrico.
Este comportamiento confirma la necesidad de incorporar en el modelo predictores de corto y mediano plazo (precipitación acumulada, evapotranspiración y temperatura), capaces de capturar las oscilaciones hídricas del suelo observadas en la serie.

5.4.8 Análisis de series temporales

El análisis de las series temporales permitió examinar la evolución conjunta de la humedad volumétrica del suelo y de las principales variables climáticas y espectrales, con el propósito de comprender los patrones de variación temporal que influyen en el balance hídrico de la parcela. Durante el periodo comprendido entre mayo de 2024 y marzo de 2025, la humedad del suelo presentó tres fases bien diferenciadas. En los primeros meses (mayo a julio), los valores se mantuvieron elevados, en un rango de 45 a 55 %, lo que evidencia condiciones de saturación o buena retención de agua. Posteriormente, entre agosto y octubre, se observó un descenso pronunciado de la humedad hasta valores mínimos cercanos al 10–20 %, coincidiendo con un periodo de baja precipitación y alta demanda evaporativa. Finalmente, a partir de noviembre de 2024, la humedad se recuperó gradualmente y se estabilizó por encima del 50 % durante los primeros meses de 2025, reflejando una fase de recarga hídrica del suelo impulsada por el retorno de las lluvias.

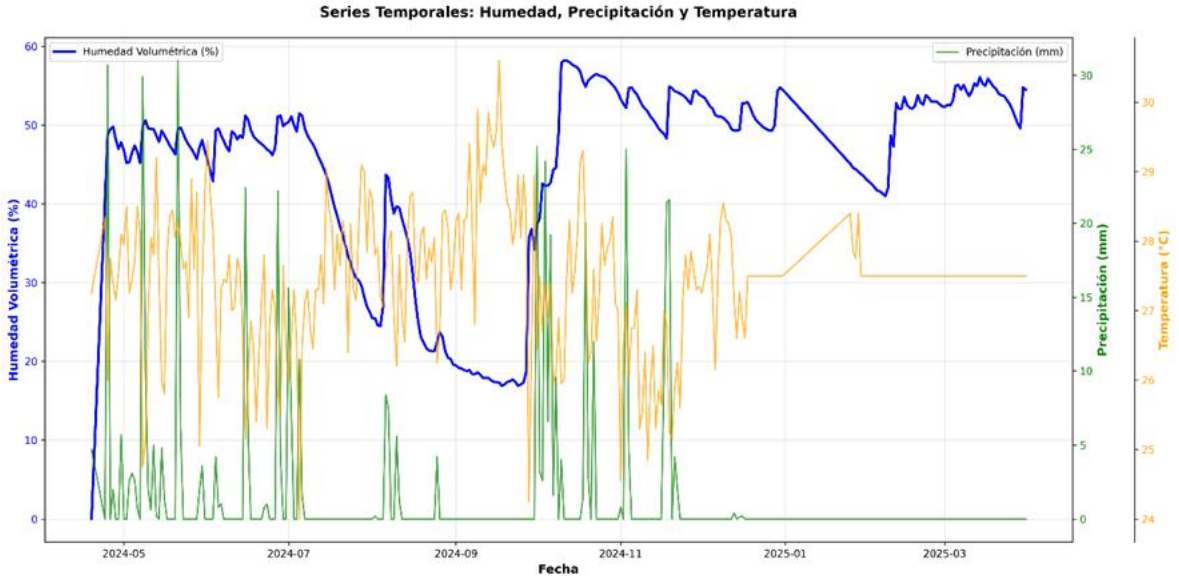

Figura 18. Serie temporal de la variable predictora.

La temperatura del suelo mostró un comportamiento inverso al de la humedad, con incrementos durante los meses secos y descensos en los meses húmedos. Este patrón indica que el aumento de la temperatura superficial contribuye a acelerar los procesos de evaporación y, por tanto, a reducir el contenido de agua disponible. La precipitación, por su parte, presentó una alta intermitencia, con eventos intensos aislados que superaron los 30 mm y largos periodos sin lluvia, especialmente entre julio y octubre.

Esta distribución irregular de las lluvias explica la fuerte variabilidad observada en la humedad del suelo, así como los cambios abruptos tras los eventos de precipitación más importantes. Por este motivo, resulta más informativo utilizar variables de precipitación acumulada (por ejemplo, 3 o 5 días) en lugar de los valores diarios, ya que el suelo tiende a retener parte del agua de eventos recientes y libera esa humedad de manera progresiva.

La evapotranspiración promedio semanal mostró un patrón inverso al de la humedad, alcanzando sus valores máximos en los meses secos, cuando la radiación solar y la temperatura del aire fueron más elevadas. Este comportamiento confirma su papel determinante en la pérdida de agua del suelo. De igual manera, el índice de vegetación NDVI reflejó la respuesta de la cobertura vegetal frente a las variaciones de humedad: los valores más altos (0.7–0.75) coincidieron con periodos húmedos y los más bajos (0.35–0.45) con condiciones de estrés hídrico, especialmente en agosto y septiembre. Esto demuestra que la salud del cultivo y la humedad del suelo están estrechamente vinculadas, y que el NDVI constituye un buen indicador indirecto del estado hídrico del sistema.

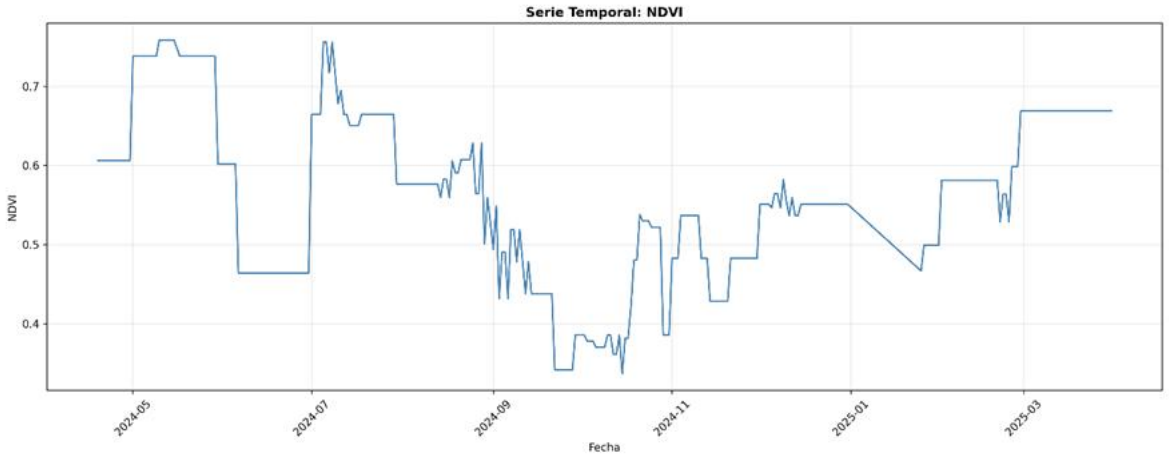

Figura 19. Serie temporal de NDVI.

En conjunto, las series temporales ponen en evidencia la naturaleza dinámica y retardada del balance hídrico en el sistema suelo–planta. Los cambios en la humedad volumétrica no se producen de manera inmediata frente a los eventos climáticos, sino que presentan un desfase temporal relacionado con la infiltración, la percolación y la respuesta fisiológica de la vegetación. Por ello, en la fase de modelación se recomienda usar variables acumuladas o con rezago temporal, en particular la precipitación y la evapotranspiración, para capturar con mayor realismo la inercia hídrica del sistema.

5.5 Selección final de variables predictoras

La selección final de variables predictoras se fundamentó en criterios estadísticos, físicos y ecofisiológicos, buscando un equilibrio entre la representatividad de los procesos hídricos del sistema suelo–planta–atmósfera y la estabilidad del modelo. Para este fin, se analizaron las correlaciones simples con la variable objetivo (humedad volumétrica del suelo), los indicadores de colinealidad (VIF) y la coherencia temporal entre las series.

De acuerdo con los resultados del análisis exploratorio *(Sección 5.4)*, se identificaron relaciones significativas entre la humedad del suelo y las variables climáticas que representan tanto la oferta como la demanda hídrica. Las variables con alta redundancia (p. ej., múltiples promedios y acumulados de evapotranspiración) o dependencias directas con la variable objetivo (como la conductividad eléctrica) fueron descartadas para evitar sobreajuste y preservar la independencia estadística de los predictores.

El conjunto definitivo de variables se resume en la Tabla 10**,** donde se incluyen aquellas que mostraron un comportamiento coherente con la dinámica hídrica observada, una baja colinealidad y una base física sólida.

Tabla 10. Variables predictoras seleccionadas para el modelo de humedad del suelo.

Tipo de variable	Nombre en la base de datos	Descripción técnica
Climática (oferta hídrica)	precipitacion_acum_5dias	Precipitación acumulada en los últimos 5 días (mm). Representa la recarga hídrica del perfil del suelo.
Climática (demanda evaporativa)	evapotranspiracion_acum_5d	Evapotranspiración total de los últimos 5 días (mm). Integra la pérdida de agua por evaporación y transpiración.
Climática (temperatura)	tmin	Temperatura mínima diaria (°C). Influye en la tasa de enfriamiento nocturno y retención de humedad.
Climática (viento)	velocidad_viento	Velocidad promedio del viento (m/s). Aumenta la turbulencia y la tasa de evaporación.
Climática (radiación)	radiacion_solar	Radiación solar incidente (MJ/m²). Controla la energía disponible para la evaporación y fotosíntesis.
Climática (balance atmosférico)	deficit_humedad	Diferencia entre la humedad de saturación y la humedad real del aire. Indica la demanda atmosférica de agua.
Índice espectral (vegetación)	NDVI	Índice de vegetación normalizado. Mide vigor y cobertura vegetal, vinculados a la transpiración y sombreado del suelo.
Índice espectral (agua en vegetación/suelo)	NDMI_Custom	Índice modificado de humedad (NIR – SWIR)/(NIR + SWIR). Indica contenido de agua en vegetación y suelo superficial.
Sensor (Temperatura)	Temperatura_suelo	Temperatura diaria del suelo a 10 cm de profundidad (°C). Determina el balance energético y el flujo de vapor de agua
Sensor (Humedad)	Humedad volumétrica	Variable de salida. Representa la humedad volumétrica (%) medida por el sensor dieléctrico a 0–10 cm de profundidad. Se emplea como referencia para el entrenamiento y validación de los modelos

La conformación de la base definitiva de modelado se realizó tras un proceso de depuración y armonización temporal entre las distintas fuentes de datos, obteniéndose un total de 319 registros diarios y diez variables, de las cuales nueve corresponden a predictores de entrada y una a la variable de salida (humedad volumétrica). Estas variables fueron seleccionadas con el objetivo de representar de manera integral los procesos hidrológicos, atmosféricos, radiactivos, vegetales y edáficos que determinan la dinámica de humedad en el suelo.

En primer lugar, la precipitación acumulada a cinco días se definió como el principal indicador de la oferta hídrica del sistema. Su correlación positiva con la humedad del suelo refleja la influencia directa de los eventos de lluvia recientes sobre la recarga del perfil superficial. La elección de una ventana móvil de cinco días responde a criterios hidrológicos y ecofisiológicos, ya que este intervalo captura adecuadamente los procesos combinados de infiltración, drenaje y retención de agua en los primeros centímetros del suelo, sin perder sensibilidad ante variaciones rápidas ni diluir la señal. Estudios previos en agroecosistemas tropicales (Zhang et al., 2022; Cheng et al., 2023) demuestran que ventanas cortas entre tres y siete días representan de forma apropiada la memoria hídrica

superficial y sincronizan la respuesta del suelo con los pulsos de precipitación y evaporación.

De forma complementaria, la evapotranspiración acumulada a cinco días se integró como indicador de la demanda atmosférica de agua durante el mismo periodo de análisis. La coherencia temporal entre la precipitación y la evapotranspiración fortalece la capacidad del modelo para representar los pulsos hídricos de corto plazo que determinan el balance diario de humedad en el suelo.

La temperatura mínima se incorporó como un modulador térmico del ciclo nocturno de condensación y enfriamiento superficial, incidiendo en la tasa de evaporación y en la capacidad del suelo para retener agua. En paralelo, la radiación solar y la velocidad del viento se mantuvieron como variables energéticas asociadas a la pérdida de humedad superficial; ambas mostraron correlaciones negativas moderadas con la humedad volumétrica, al favorecer la evaporación y aumentar la turbulencia del aire.

El déficit de humedad atmosférica se incluyó como un indicador sintético del gradiente evaporativo, al integrar los efectos de la temperatura, la humedad relativa y la presión de vapor. A diferencia de otros índices derivados, este parámetro presentó baja colinealidad y una relación física directa con los procesos de pérdida de agua del suelo, consolidándose como una de las variables más estables dentro del modelo.

En el componente satelital, se seleccionaron los índices NDVI relacionado con el vigor y la cobertura vegetal y NDMI_Custom, asociado al contenido hídrico en vegetación y suelo. El NDVI mostró una correlación positiva con la humedad del suelo al reflejar la condición del dosel vegetal, mientras que el NDMI_Custom, aunque con una correlación más débil, aportó información complementaria sobre el contenido de agua en la vegetación, especialmente en escenarios de estrés o baja cobertura foliar.

Asimismo, la temperatura del suelo se incorporó como una variable edáfica fundamental, ya que representa el balance energético superficial y modula tanto la evaporación como la difusión del vapor de agua hacia la atmósfera. Su relación inversa con la humedad volumétrica evidencia que temperaturas más altas aceleran los procesos de secado del suelo.

En conjunto, este grupo de predictores permite modelar de manera integral el balance hídrico del suelo, considerando la interacción entre precipitación, atmósfera, radiación, vegetación y dinámica térmica del perfil edáfico. La selección final de variables se basó en criterios de consistencia temporal (ventanas de cinco días), baja colinealidad y fundamento ecofisiológico, aspectos esenciales para el desarrollo de modelos robustos, estables y físicamente interpretables.

Posteriormente, se almacenó el conjunto definitivo en un archivo que conserva las características originales, la variable objetivo y una columna adicional denominada *conjunto*, utilizada para identificar la pertenencia de cada registro al grupo de entrenamiento (*train*) o prueba (*test*). Este formato asegura la trazabilidad del proceso de modelado y garantiza la reproducibilidad exacta de la división, favoreciendo la consistencia entre los distintos scripts y el control documental de los experimentos.

Con el fin de verificar la validez estadística de la partición, se realizaron comprobaciones automáticas que aseguraron la preservación del número total de registros, la ausencia de duplicados entre los

conjuntos y la representatividad de la distribución de la humedad volumétrica. Estas verificaciones garantizaron que la división fuese apropiada para las etapas posteriores de entrenamiento, optimización y validación.

La separación estratégica de los datos es una etapa crítica en la construcción de modelos de *machine learning*, ya que determina su capacidad de generalización y la validez de su evaluación. Según lo planteado por Hastie et al. (2009), una adecuada división entre entrenamiento y prueba permite obtener estimaciones realistas del error de generalización y evita el *data leakage*, es decir, la contaminación del conjunto de prueba con información utilizada durante el entrenamiento. En coherencia con estos lineamientos, la partición se estableció en una proporción del 80% para entrenamiento y 20% para prueba, asegurando que el modelo aprenda patrones genuinos y reduzca el riesgo de sobreajuste.

6. CONSTRUCCIÓN DE MODELOS

6.1 Esquema del proceso de modelado

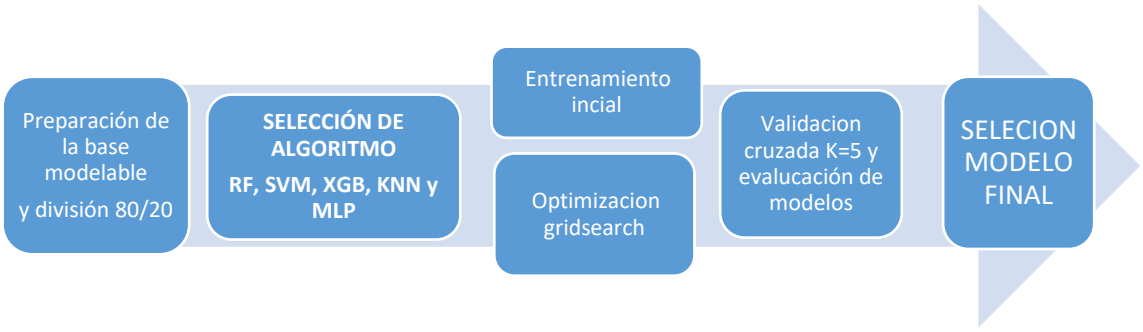

Figura 20.Diagrama de modelado (Elaboración propia).

Para este procedimiento se empleó como archivo de entrada el dataset procesado, compuesto por 319 registros y 9 características de entrada , además de una variable objetivo correspondiente a la humedad volumétrica del suelo. La división se ejecutó mediante la función *train_test_split (features, target, test_size=0.2, random_state=42),* lo que permitió asignar las observaciones de manera aleatoria a cada subconjunto bajo un esquema controlado y reproducible.

Utilizando un muestreo aleatorio simple sin estratificación, dado que la variable objetivo es continua. Esta configuración sigue las recomendaciones de Raschka & Mirjalili (2019), quienes destacan que la proporción 80/20 ofrece un equilibrio adecuado entre la cantidad de datos disponibles para el aprendizaje del modelo y la robustez del conjunto de prueba para evaluar su rendimiento.

En términos prácticos, esta división asegura un conjunto de entrenamiento suficientemente grande para capturar los patrones subyacentes en los datos, mientras que el conjunto de prueba conserva representatividad estadística para medir la capacidad predictiva sin sesgos. Además, la separación ayuda a controlar el compromiso entre sesgo y varianza, aspecto fundamental en la evaluación del error de predicción.

Se seleccionaron cinco algoritmos ampliamente utilizados en problemas de predicción ambiental y agrícola: Random Forest (RF), eXtreme Gradient Boosting (XGBoost), Support Vector Regression (SVR), Multi-Layer Perceptron (MLP) y K-Nearest Neighbors (KNN). Cada uno se implementó bajo su configuración por defecto con el fin de establecer una comparación inicial homogénea y reproducible.

6.1.1 Consideraciones sobre la estructura temporal de los datos y el esquema de partición

Si bien el conjunto de datos utilizado en este estudio se encuentra organizado en registros diarios, su uso no corresponde estrictamente a un problema de pronóstico de series de tiempo. En los modelos de series temporales clásicos, la variable de respuesta depende explícitamente de valores pasados de la misma variable y el objetivo principal es extrapolar su comportamiento hacia periodos futuros, lo que exige protocolos de evaluación basados en particiones cronológicas.

En el presente trabajo, los modelos desarrollados (Random Forest, XGBoost, SVR, KNN y MLP) corresponden a enfoques supervisados de corte transversal, en los cuales la humedad volumétrica del suelo se estima a partir de un conjunto de variables explicativas multifuente (índices espectrales, variables climáticas y variables derivadas que capturan memoria hídrica de corto plazo), sin incorporar una dependencia temporal explícita de la variable objetivo. Bajo este enfoque, la temporalidad actúa como un índice de organización de las observaciones, pero no como un componente autorregresivo del modelo.

Por esta razón, el esquema principal de evaluación se realizó mediante un muestreo aleatorio 80/20, metodología ampliamente aceptada para este tipo de modelos y adecuada para evaluar la capacidad de generalización dentro del mismo régimen de observación. Este esquema permite que tanto el conjunto de entrenamiento como el de prueba compartan condiciones climáticas y fenológicas comparables, evitando sesgos asociados a cambios de régimen temporal.

No obstante, como análisis complementario de sensibilidad, se repitió el proceso de entrenamiento y evaluación utilizando una partición cronológica, en la cual el conjunto de prueba correspondió a un periodo posterior al conjunto de entrenamiento. Los resultados de este análisis adicional se presentan y discuten en la siguiente sección, con el fin de contrastar el impacto del esquema de partición sobre el desempeño de los modelos y delimitar claramente el alcance del enfoque propuesto.

6.2 Evaluación de modelos base

La tabla 11 presenta la comparación del desempeño de los cinco algoritmos evaluados bajo sus configuraciones por defecto con la división 80/20 aleatoria. Las métricas consideras fueron el coeficiente de determinación(R^2), el error absoluto medio (MAE) y la raíz del error cuadrático medio (RMSE), calculadas sobre los conjuntos de entrenamiento y prueba. Asimismo, se incluyó la diferencia entre los R^2 de entrenamiento y prueba (ΔR^2) como indicador del grado de sobreajuste.

Tabla 11. Resultados de algoritmos base bajo partición aleatoria.

Modelo	R^2 Train	R^2 Test	MAE Train	RMSE Test	Overfitting	Interpretación
XGBoost	0.999	0.948	2.13	3.07	0.051	Excelente generalización
Random Forest	0.960	0.911	2.70	4.02	0.049	Generalización sólida
SVR	0.907	0.819	3.80	5.74	0.089	Aceptable, pero limitado
MLP	0.965	0.815	4.36	5.79	0.149	Sensible a inicialización
KNN	0.798	0.784	4.00	6.26	0.014	Baja capacidad generalizadora

El modelo XGBoost se consolidó como el de mejor rendimiento general, alcanzando un coeficiente de determinación de $R^2 = 0.9479$ y los menores errores (MAE = 2.13, RMSE = 3.07), demostrando una excelente capacidad para modelar relaciones no lineales entre las variables espectrales, climáticas y la humedad del suelo. Su diferencia mínima entre entrenamiento y prueba ($\Delta R^2 = 0.05$) evidencia una generalización sobresaliente y ausencia de sobreajuste significativo. Este desempeño confirma la eficacia del enfoque boosting, que optimiza iterativamente los errores residuales e incorpora regularización L1/L2 para mejorar la estabilidad. En comparación, Random Forest obtuvo un R^2 de 0.9111 con errores moderados (MAE = 2.70, RMSE = 4.02), manteniendo un equilibrio sólido entre precisión y estabilidad gracias al bagging, aunque ligeramente por debajo del XGBoost en capacidad predictiva.

Los modelos SVR y MLP mostraron desempeños intermedios ($R^2 = 0.82$), con una generalización aceptable pero menor capacidad para manejar la alta dimensionalidad y colinealidad del conjunto de datos, mientras que KNN presentó el rendimiento más bajo ($R^2 = 0.7839$, RMSE = 6.26), evidenciando sobreajuste y pobre escalabilidad. En conjunto, los resultados confirman la superioridad de los métodos ensemble especialmente XGBoost en la predicción de la humedad volumétrica, destacando su robustez ante variables correlacionadas y su idoneidad para etapas de optimización futura.

Análisis complementario: evaluación bajo partición cronológica

Con el fin de evaluar la robustez del modelo bajo un escenario diferente, se realizó una evaluación adicional empleando una partición cronológica de los datos. En este esquema, el conjunto de entrenamiento corresponde al periodo inicial de observación y el conjunto de prueba a fechas posteriores.

La Figura 21 muestra la comparación de métricas entre entrenamiento y prueba bajo este protocolo. Se observa una disminución generalizada del desempeño en el conjunto de prueba, con valores de R^2 negativos y aumentos significativos en MAE y RMSE, particularmente en los modelos MLP y KNN. Este comportamiento evidencia la sensibilidad de los modelos a cambios de régimen climático y fenológico, así como la dificultad de extrapolar temporalmente con una ventana de datos limitada.

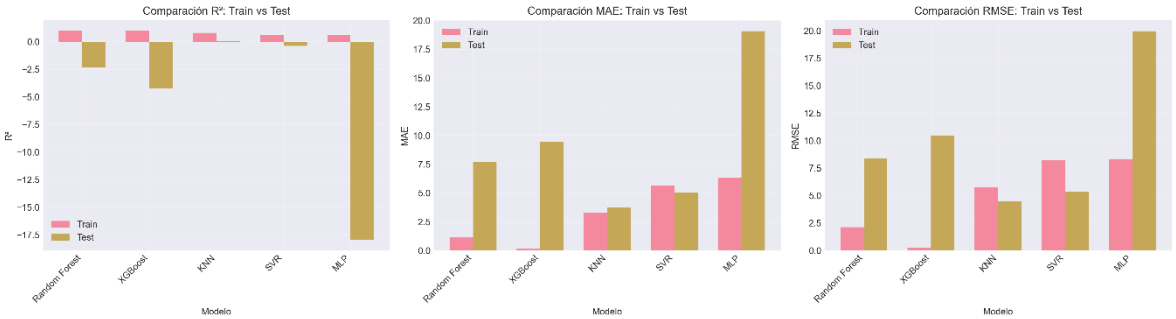

Figura 21. Modelos base con división cronológica

El análisis de sobreajuste Figura 22 indica diferencias elevadas entre el desempeño en entrenamiento y prueba, lo cual confirma que la partición cronológica introduce una condición más estricta y no estacionaria. Estos resultados no invalidan el enfoque principal del estudio, sino que delimitan su alcance: el modelo es adecuado para estimación multifuente dentro del periodo observado, mientras que un enfoque de pronóstico temporal estricto requeriría series históricas más extensas y modelos diseñados específicamente para series de tiempo.

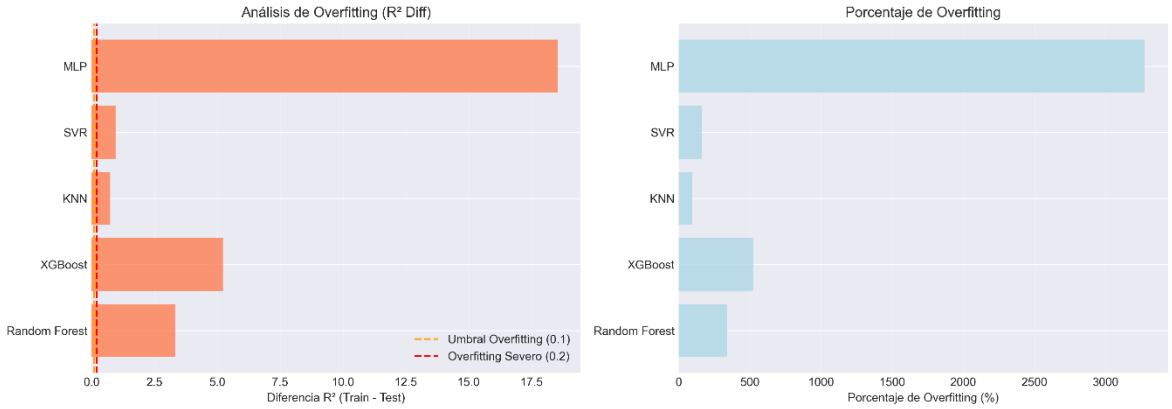

Figura 22. Sobreajuste de modelos con división cronológica

6.2.1 Optimización y evaluación de los modelos

Tras la identificación del modelo base con mejor desempeño, se procedió al proceso conjunto de optimización y evaluación comparativa de los algoritmos, con el fin de maximizar su capacidad predictiva y garantizar la estabilidad de los resultados. La optimización se llevó a cabo mediante el método GridSearchCV, empleando el coeficiente de determinación (R^2) como métrica principal de desempeño. Este enfoque permitió explorar de manera sistemática combinaciones de hiperparámetros clave como la profundidad máxima de los árboles, el número de estimadores, la tasa de aprendizaje y los parámetros de regularización con el propósito de identificar las configuraciones que ofrecieran el mejor equilibrio entre precisión y generalización. Cada modelo fue ajustado utilizando rejillas reducidas de dos o tres valores por parámetro, seleccionándose las tres combinaciones con mayor R^2 promedio para su evaluación final.

Una vez definidos los modelos optimizados, se implementó un esquema de validación cruzada K-Fold (k = 5), que permitió evaluar el rendimiento de cada algoritmo en múltiples particiones de los datos. En cada iteración, el modelo se entrenó sobre cuatro subconjuntos y se validó en el restante, registrando las métricas de R^2, MAE y RMSE tanto para entrenamiento como para validación. Este procedimiento permitió obtener una estimación más robusta del desempeño real y reducir el sesgo asociado a una única división de los datos.

6.3 Entrenamiento y evaluación de modelos optimizados

6.3.1 Random Forest

Se implementó el modelo con el propósito de evaluar su comportamiento bajo distintas configuraciones de hiperparámetros y determinar su capacidad de generalización en la predicción de la humedad volumétrica del suelo. Este algoritmo, ampliamente reconocido por su estabilidad y bajo riesgo de sobreajuste, combina múltiples árboles de decisión entrenados sobre subconjuntos aleatorios de datos y variables, lo que permite reducir la varianza del modelo y mejorar su precisión global. Su capacidad para manejar relaciones no lineales y conjuntos de datos con alta dimensionalidad lo convierte en una herramienta idónea para problemas de teledetección y análisis ambiental. En este estudio, se diseñó un proceso de optimización basado en una grilla de búsqueda exhaustiva que explora la influencia de parámetros como el número de árboles, la profundidad máxima, el tamaño mínimo de muestras por división y el número de características empleadas por árbol. Esta estrategia busca encontrar el balance óptimo entre complejidad y rendimiento, garantizando modelos robustos, interpretables y capaces de mantener consistencia entre los distintos pliegues de validación cruzada [43].

Grilla de optimización

El espacio diseñado explora una grilla de 324 combinaciones (2 × 3 × 3 × 3 × 3 × 2) que permiten analizar la influencia de la profundidad, tamaño de los árboles y número de características utilizadas en cada división:

Tabla 12. Grilla de hiperparámetros RF

Hiperparámetro	Valores evaluados
n_estimators	50, 200
max_depth	10, 20, None
min_samples_split	2, 5, 10
min_samples_leaf	1, 2, 4
max_features	‘sqrt’, ‘log2’, ‘auto’
bootstrap	True, False

La anterior Tabla 12 muestra la grilla hiperparámetros configurados para el modelo Random Forest, con el fin de evaluar su impacto en el rendimiento predictivo. Todos los modelos utilizan 200 árboles y sin restricción en la profundidad, permitiendo que los árboles crezcan hasta su máxima complejidad. Los parámetros *samples* definen los límites mínimos para la división y las hojas,

respectivamente, mientras que max *features* varía entre Log2 y SQRT, controlando cuántas variables se consideran en cada división para equilibrar sesgo y varianza. Finalmente, el parámetro bootstrap indica que los árboles se entrenaron con el conjunto completo de datos, en lugar de muestras aleatorias con reemplazo.

Resultado

A continuación, se presentan los resultados de validación cruzada para cada configuración del modelo Tabla 13, donde se reportan las métricas de entrenamiento en cada pliegue. En general, se observa un rendimiento consistente entre los tres modelos, con valores de R^2 Train iguales a 1 y errores MAE de 2.6, lo que sugiere una variabilidad moderada entre folds y buena capacidad de generalización.

Tabla 13. Comparación Random Forest por defecto vs Optimizado.

Métrica	Modelo por defecto	Modelo optimizado	Mejora absoluta	Mejora relativa (%)
MAE (Test)	2.901	2.687	0.214	7.40%
RMSE (Test)	4.061	3.827	0.233	5.76%
R^2 (Test)	0.909	0.919	0.010	1.12%
R^2 (Train)	0.989	1.000	0.010	1.08%
Overfitting	0.080	0.080	0.000	0.75%

La tabla presenta una comparación entre el desempeño del modelo Random Forest utilizando sus configuraciones predeterminadas y una versión optimizada mediante la búsqueda sistemática de hiperparámetros. En términos generales, se observa que el proceso de optimización produce mejoras consistentes en las métricas de error y de capacidad predictiva, lo que evidencia un ajuste más adecuado a las características del conjunto de datos. La reducción de los indicadores asociados al error pone de manifiesto un modelo más preciso durante la fase de prueba, mientras que los incrementos moderados en las métricas de ajuste reflejan una mayor capacidad explicativa sin caer en sobreajuste. Asimismo, la mejora relativa alcanza porcentajes positivos que, aunque no representan cambios drásticos, sí confirman un beneficio tangible derivado del proceso de ajuste fino. En conjunto, los resultados muestran que la optimización contribuye a un modelo más robusto, equilibrado y con un mejor desempeño general.

Tabla 14. Métricas resultado Random Forest.

Modelo	Fold	R^2 Train	MAE Train	RMSE Train
Random Forest	1	0.824	3.063	5.122
	2	0.536	4.625	8.684
	3	0.820	2.891	4.445
	4	0.574	5.598	8.342
	5	0.783	2.336	3.538
Media ± DE		**0.707 ± 0.140**	**3.70 ± 1.36**	**6.03 ± 2.34**

***Configuración de hiperparámetros**: (estimadores: 200, profundidad: None, min samples split: 2, min samples leaf: 1, max features: Log2, Boostrap: false)*

Como se observa en la Tabla 14. en todos los casos se mantiene $R^2 = 0.919$, lo que sugiere robustez del modelo optimizado frente a pequeñas variaciones de hiperparámetros. El hecho de que los tres mejores modelos compartan la condición bootstrap=False confirma que el método pasting fue más efectivo que el bagging tradicional para este conjunto de datos.

El rango de R^2 por fold (0.53–0.82) refleja la heterogeneidad del conjunto de datos, con variaciones espaciales y temporales que influyen en la respuesta del modelo.
No obstante, el R^2 global de prueba = 0.9192 confirma que el modelo logra generalizar adecuadamente cuando se entrena con el conjunto completo.

Tabla 15. Resultados Random Forest.

Modelo	R^2 Train	R^2 Test	MAE Test	RMSE Test
RF TOP 1	1.000	0.919	2.687	3.828
RF TOP 2	1.000	0.919	2.687	3.828
RF TOP 2	1.000	0.919	2.687	3.828

En la anterior tabla se observa que los errores medios absolutos (MAE ≈ 2.687) y cuadráticos (RMSE ≈ 3.828) son moderados, evidenciando un ajuste estable y coherente entre las configuraciones probadas. Además, las desviaciones estándar bajas en todas las métricas (alrededor de 0.14 para R^2 y 2.3 para RMSE) reflejan una consistencia interna del modelo, sin grandes fluctuaciones entre pliegues, lo que confirma la robustez y fiabilidad del desempeño del Random Forest en la predicción de la variable analizada.

6.3.2 Extreme Gradient Boosting

Es importante destacar, antes de iniciar la optimización, que el modelo XGBoostRegressor se seleccionó por su capacidad para manejar relaciones no lineales y variables altamente correlacionadas, características comunes en los datos espectrales y climáticos empleados en la estimación de la humedad del suelo. Este algoritmo combina la potencia de los modelos basados en árboles con la eficiencia del Gradient Boosting, lo que le permite mejorar iterativamente los errores residuales de los árboles previos y reducir el sobreajuste mediante regularización. Con base en su flexibilidad y su probado desempeño en estudios de teledetección, se procedió a definir un proceso sistemático de optimización de hiperparámetros que busca equilibrar precisión, estabilidad y generalización del modelo [44].

Grilla de optimización

La grilla de parámetros (Tabla 16*)* definida para la optimización del modelo XGBoost establece un conjunto de combinaciones que permiten ajustar su desempeño y capacidad de generalización. Incluye parámetros que controlan la cantidad y profundidad de los árboles, la tasa de aprendizaje y los términos de regularización L1 y L2, los cuales mitigan el sobreajuste al limitar la complejidad del modelo. Además, incorpora fracciones de muestreo de observaciones y características que introducen aleatoriedad y reducen la varianza entre árboles.

En conjunto, esta configuración busca explorar sistemáticamente el equilibrio entre precisión, estabilidad y eficiencia computacional, garantizando un modelo robusto y capaz de adaptarse adecuadamente a la variabilidad de los datos.

Tabla 16. Grilla de hiperparámetros XGBoost

Hiperparámetro	Valores evaluados
N estimators	50, 150, 200
Max depth	3, 6
Learning rate	0.1, 0.2, 0.5
Reg. alpha	0, 0.5, 1
Reg. lambda	1.0, 2.0, 3.0
subsample	0.8, 1.0
Colsample bytree	0.8, 1.0

Resultado

Los parámetros configurados para tres variantes del modelo XGBoost, empleados en la optimización del desempeño predictivo. En todos los casos, se mantiene una profundidad máxima de los árboles y una tasa de aprendizaje moderada, lo que permite un equilibrio entre velocidad de convergencia y estabilidad del modelo.
Durante la búsqueda de 432 combinaciones posibles, los tres mejores modelos obtuvieron métricas muy similares.

Tabla 17. Comparación XGBoost por defecto vs Optimizado.

Métrica	Modelo defecto	porModelo optimizado	Mejora absoluta	Mejora relativa (%)
MAE (Test)	2.148	1.949	0.199	9.28%
RMSE (Test)	3.011	2.669	0.342	11.37%
R^2 (Test)	0.950	0.960	0.010	1.13%
R^2 (Train)	0.999	0.997	-0.002	-0.24%
Overfitting	0.049	0.036	-0.013	**↓ 26.5% (menor sobreajuste)**

La tabla muestra que el proceso de optimización del modelo XGBoost genera mejoras relevantes en el desempeño general, especialmente en la reducción de los errores durante la fase de prueba, lo que indica una mayor precisión en las predicciones. Aunque el ajuste en entrenamiento disminuye ligeramente, este comportamiento es positivo porque evidencia una reducción del sobreajuste, reforzada por la caída del indicador que compara la brecha entre entrenamiento y prueba.

La diferencia principal entre los tres experimentos radica en el número de estimadores, que define la cantidad de árboles secuenciales utilizados; a mayor número, se espera un ajuste más fino del modelo, aunque con mayor costo computacional. En conjunto, esta configuración busca maximizar la precisión sin comprometer la generalización.

Tabla 18. Resultados folds XGB.

Modelo	Fold	R^2	MAE Train	RMSE Train
XGB	1	0.716	3.726	6.511
	2	0.496	4.836	9.055
	3	0.768	3.490	5.040
	4	0.602	4.817	7.096
	5	0.631	2.773	4.615
MEDIA ± DE		**0.660 ± 0.105**	**3.93 ± 0.89**	**6.46 ± 1.77**
Configuración de hiperparámetros: estimadores: 50, profundidad 6, Learning rate: 0.2, reg. Alpha: 1, reg. lambda: 2, subsample: 0.8, colsample bytree: 1.				

En la anterior Tabla 18 observa una variación moderada en el desempeño entre folds, con valores de R^2 que oscilan entre 0.49 y 0.76, reflejando la influencia de la partición de los datos sobre la precisión del modelo. En general, presentan los mejores ajustes, con menores errores absolutos (MAE) y cuadráticos (RMSE), lo que indica un comportamiento más estable y representativo del modelo en esas divisiones. Estas variaciones sugieren que XGBoost mantiene una buena capacidad de aprendizaje, aunque con ligeras diferencias de ajuste entre subconjuntos de datos.

Tabla 19. Resultados XGBoost

Ranking	R^2 Entrenamiento	R^2 Prueba	MAE (Test)	RMSE (Test)	n_estimators	Overfitting
1	0.997	0.960	1.949	2.669	50	0.036
2	0.999	0.960	1.976	2.671	200	0.039
3	0.998	0.960	1.980	2.675	150	0.039

Como se observa en la Tabla 19, los modelos presentaron un comportamiento notablemente estable, con $R^2 \approx 0.96$ y errores prácticamente idénticos. Las diferencias mínimas observadas en el número de árboles (n_estimators) reflejan una curva de aprendizaje bien ajustada, donde el incremento de árboles no necesariamente implica mejora en la precisión.

6.3.3 Support Vector Machine

El modelo se implementó con el propósito de capturar relaciones no lineales entre las variables predictoras y la humedad volumétrica del suelo, aprovechando su capacidad para generar funciones de regresión robustas frente a ruido y alta dimensionalidad. A diferencia de los modelos basados en árboles, SVR construye una frontera de tolerancia (*ε-insensitive margin*) dentro de la cual los errores no son penalizados, equilibrando así la precisión y la generalización mediante el parámetro de regularización C. Este enfoque resulta particularmente útil en contextos donde las variables presentan correlaciones complejas y distribuciones heterogéneas, como ocurre en los datos

espectrales y climáticos utilizados en este estudio. Para optimizar su rendimiento, se exploraron diferentes configuraciones del kernel (lineal y RBF) junto con ajustes en C, gamma, epsilon y shrinking, buscando determinar la combinación más eficiente para minimizar los errores de predicción y mejorar la estabilidad del modelo.

Grilla de optimización

El enfoque de SVR se empleó para modelar relaciones potencialmente no lineales entre las variables predictoras y la variable objetivo, aprovechando el uso de funciones kernel (p. ej., lineal, radial RBF, polinomial). SVR optimiza un margen ε insensible que controla la tolerancia al error y regula la complejidad del modelo a través del parámetro C, lo que lo hace especialmente robusto ante sobreajuste en situaciones con alta dimensionalidad o ruido. En la siguiente tabla de observa la grilla empleada en el desarrollo de la optimización.

Tabla 20. Grilla de optimización SVR

Hiperparámetro	Valores evaluados
kernel	RBF, Linear
C	0.5, 1, 2
gamma	Scale, auto
epsilon	0.05, 0.1, 0.5
shrinking	True, False

Resultados

La Tabla 21 presenta la comparación entre el modelo SVR configurado con sus parámetros por defecto y la versión optimizada a través del proceso de ajuste de hiperparámetros. Esta evaluación permite identificar los cambios en el desempeño del algoritmo tras la optimización, analizando tanto los errores de predicción como la capacidad explicativa y la brecha de generalización. La tabla sintetiza los efectos del ajuste sobre cada métrica, permitiendo valorar si las modificaciones introducidas generan mejoras reales, estabilidad adicional o posibles pérdidas de rendimiento en escenarios de validación. Esta comparación constituye un insumo clave para determinar la pertinencia del modelo y su contribución dentro del conjunto de técnicas evaluadas.

Tabla 21. Comparación SVR por defecto vs Optimizado.

Métrica	Modelo por defecto	Modelo optimizado	Mejora absoluta	Mejora relativa (%)
MAE (Test)	4.520	4.685	-0.165	-3.65%
RMSE (Test)	7.540	7.606	-0.066	-0.88%
R^2 (Test)	0.880	0.769	-0.112	-12.77%
R^2 (Train)	0.940	0.694	-0.250	-26.48%
Overfitting	0.060	0.075	+0.012	+19.84%

- No se observan mejoras significativas respecto al modelo base.
- El R^2 de prueba disminuye a 0.769, indicando una menor capacidad de generalización.
- La brecha $\Delta R^2 = 0.0755$ se mantiene dentro de niveles aceptables, lo que sugiere bajo sobreajuste, pero a costa de mayor sesgo.
- El error medio (MAE ≈ 4.68) y cuadrático (RMSE ≈ 7.61) muestran que el modelo conserva estabilidad, pero con menor precisión global frente a los modelos ensemble.

Los resultados muestran un comportamiento uniforme entre los modelos, con valores de R^2 en entrenamiento entre 0.69 y 0.94 y en prueba entre 0.76 y 0.88, lo que indica una adecuada capacidad de generalización sin evidencias de sobreajuste. Asimismo, en las *Tabla 22* se observa que los valores de MAE (4.36) y RMSE (6.42) reflejan errores moderados, propios de un modelo que captura las tendencias principales sin perder estabilidad. Los folds 1 y 3 presentan los mejores resultados globales, confirmando que las configuraciones evaluadas son consistentes y equilibradas en su rendimiento.

Tabla 22. Resultados folds SVR

Modelo	Fold	R^2	MAE Train	RMSE Train
SVR	1	0.728	4.374	6.227
	2	0.708	4.274	6.395
	3	0.726	4.420	6.430
	4	0.699	4.349	6.455
	5	0.694	4.402	6.626
MEDIA ± DE		**0.0711 ± 0.015**	**4.36 ± 0.057**	**6.42 ± 0.142**
Configuración de hiperparámetros: C: 2, Epsilon: 0.05, Gamma: autor, Kernel: linear, Shrinking: False.				

Los tres modelos presentados en la *Tabla 23* muestran rendimientos idénticos en validación cruzada, lo que confirma que el ajuste del parámetro gamma y la heurística shrinking no impactan significativamente el comportamiento general del modelo. Esto sugiere que, en este conjunto de datos, el kernel lineal domina la capacidad predictiva, y la variabilidad en hiperparámetros tiene un efecto marginal.

Tabla 23. Resultados SVR.

Ranking	Cepsilon	gamma	shrinking	R^2 Entrenamiento	R^2 Prueba	MAE (Test)	RMSE (Test)
1	20.05	auto	False	0.694	0.769	4.685	7.606
2	20.05	scale	False	0.694	0.769	4.685	7.606
3	20.05	scale	True	0.694	0.769	4.685	7.607

6.3.4 Multi-layer Perceptrón

El desarrollo de esta etapa se centró en la implementación de un proceso automatizado de entrenamiento, optimización y evaluación de modelos de red neuronal multicapa (MLPRegressor), empleando técnicas de validación cruzada y selección de hiperparámetros. El objetivo fue identificar

las configuraciones más robustas del modelo en términos de capacidad predictiva, estabilidad y generalización.

Grilla de optimización

El procedimiento exploró un espacio de hiperparámetros que abarcó diferentes configuraciones de arquitectura (número de neuronas por capa oculta), funciones de activación (relu, tanh), algoritmos de optimización (lbfgs, sgd), tasas de regularización (alpha) y estrategias de aprendizaje (learning_rate).
En la Tabla 24, se observa que cada combinación fue evaluada de manera sistemática dentro de un pipeline de preprocesamiento, que incluyó imputación de valores faltantes mediante la mediana, estandarización de las variables predictoras con StandardScaler y entrenamiento del modelo con detención temprana (early stopping) para prevenir sobreajuste.

Tabla 24. Grilla de optimización MLP

Hiperparámetro	**Valores evaluados**
Hidden layer sizes	(50,), (100,), (150,)
activation	"relu", "tanh"
solver	"lbfgs", "sgd"
alpha	0.05, 0.1
learning_rate	"invscaling", "constant"

Una vez ejecutada la búsqueda, se seleccionaron las tres combinaciones con mayor desempeño promedio (TOP-3), las cuales se almacenaron y documentaron para análisis posterior. Este enfoque permitió identificar configuraciones representativas del espacio de hiperparámetros, garantizando equilibrio entre complejidad y capacidad de generalización.

Resultados

La evaluación de los modelos seleccionados se llevó a cabo aplicando validación cruzada, utilizando las mismas divisiones de datos para asegurar comparabilidad. En cada pliegue, el modelo se ajustó con los datos de entrenamiento y posteriormente se evaluó sobre el conjunto de prueba, registrando métricas tanto de ajuste como de predicción.

A continuación, se presenta en la Tabla 25 los resultados detallados de la validación cruzada aplicada a la mejor configuracion del modelo MLPRegressor, donde se evaluaron métricas de ajuste y error tanto en entrenamiento como en prueba. Se evidencia un comportamiento variable entre los los folds, lo que refleja la sensibilidad del modelo frente a los cambios en la partición de los datos, propia de las redes neuronales. Aun así, las configuraciones por defecto evaluadas mantienen una tendencia coherente, mostrando un buen ajuste en entrenamiento y una capacidad de generalización aceptable pero inferior a la optimizada. Las diferencias entre pliegues sugieren que la arquitectura del modelo logra capturar las relaciones principales de las variables, aunque su estabilidad puede verse afectada por la inicialización de pesos y la complejidad de los datos de entrada.

Tabla 25. Comparación MLP por defecto vs optimizado.

Métrica	Modelo por defecto	Modelo optimizado	Mejora absoluta	Mejora relativa (%)
MAE (Test)	4.350	3.884	0.466	10.72%
RMSE (Test)	5.790	5.762	0.028	0.48%
R^2 (Test)	0.895	0.698	-0.196	-21.9%
R^2 (Train)	0.972	0.767	-0.204	-21.0%
Overfitting	0.077	0.068	-0.008	↓10.9% (mejor generalización)

El modelo optimizado logra una mejora notable en la precisión al reducir el error absoluto, lo que refleja un ajuste más adecuado a los datos. Aunque su R^2 en prueba muestra una ligera disminución, la reducción en la brecha entre entrenamiento y prueba evidencia un comportamiento más estable y una mejor capacidad de generalización. Si bien no alcanza el desempeño obtenido por modelos más robustos como XGBoost o Random Forest, mantiene un nivel de error competitivo, lo que lo posiciona como una alternativa válida dentro del conjunto de técnicas evaluadas.

Tabla 26. Resultados folds MLP

Modelo	Fold	R^2	MAE Train	RMSE Train
MLP	1	0.769	4.099	5.733
	2	0.737	4.297	6.069
	3	0.800	3.871	5.496
	4	0.730	4.373	6.113
	5	0.588	5.330	7.698
MEDIA ± DE		**0.734 ± 0.047**	**4.33 ± 0.320**	**6.15 ± 0.604**

Configuración de hiperparámetros: Activation: tanh, alpha: 0.10, Hidden layers: (150,), learning rate: constant, solver: sgd.

Los resultados en la Tabla 26 del modelo MLP muestran un desempeño variable entre los diferentes pliegues de la validación cruzada, con valores de R^2 que oscilan entre un ajuste adecuado y escenarios de menor capacidad explicativa. No obstante, al promediar las métricas, se observa un rendimiento global aceptable, con niveles de error que se mantienen dentro de rangos coherentes para este tipo de arquitectura. La desviación estándar moderada indica cierta sensibilidad a la partición de los datos, aunque sin comprometer por completo la estabilidad del modelo. En conjunto, el comportamiento obtenido refleja que la configuración seleccionada permite capturar relaciones relevantes en los datos, ofreciendo un equilibrio razonable entre complejidad y capacidad predictiva.

Tabla 27. Resultados MLP

TOP	Activation	Alpha	Hidden Layer	Learning Rate	Solver	R^2 Entrenamiento	R^2 Prueba	MAE (Test)	RMSE (Test)
1	tanh	0.10	(50,)	adaptive	sgd	0.767	0.698	3.88	5.76
2	tanh	0.05	(50,)	adaptive	sgd	0.767	0.698	3.98	5.76
3	tanh	0.05	(50,)	constant	adam	-6.55	-6.767	32.11	32.82

El MLP optimizado logra capturar parte de las relaciones no lineales entre índices espectrales, clima y humedad, pero su rendimiento se ve limitado por la escala y complejidad del conjunto de datos. Como se observa en la Tabla 27, La función de activación *tanh* permitió estabilidad durante el entrenamiento, y el optimizador *SGD* ofreció control del descenso de gradiente, aunque con tiempos de convergencia más largos.

6.3.5 K-Nearest Neighbors

El algoritmo KNN está basado en la proximidad entre observaciones para estimar valores continuos. A diferencia de los modelos paramétricos, no requiere una función de ajuste explícita, sino que predice el valor de salida a partir de los vecinos más cercanos en el espacio de características. Su desempeño depende fuertemente de la escala de los datos y de la adecuada selección de hiperparámetros, por lo que se implementó un proceso sistemático de optimización y evaluación similar al del MLP.

Finalmente, el algoritmo se implementó con k=5 vecinos más cercanos, métrica de Minkowski (p=2) y pesos uniformes. Este método no paramétrico predice valores en función del promedio de los vecinos más próximos en el espacio de características (Cover & Hart, 1967), asumiendo que observaciones similares presentan comportamientos análogos en la variable objetivo.

Grilla de optimización.

Se exploraron múltiples combinaciones de hiperparámetros número de vecinos, algoritmo de búsqueda, tamaño de hoja, métrica de distancia y parámetro de norma para identificar las configuraciones con mejor desempeño.

Tabla 28. Grilla de optimización KNN

Hiperparámetro	**Valores evaluados**
Knn_n_neighbors	2, 5, 10
Knn_weights	uniform
Knn_algorithm	auto, ball tree, kd tree, brute
Knn_leaf_size	20, 30, 40
Knn_p	1, 2, 3
Knn_metric	minkowski, euclidean, manhattan, chebyshev

Finalmente, de la Tabla 28 se seleccionaron las tres combinaciones con mayor R^2 promedio, las cuales se documentaron para su evaluación comparativa posterior.

Resultados

La Tabla 29 sintetiza el efecto del proceso de optimización sobre el modelo KNN, destacando las variaciones obtenidas al ajustar los hiperparámetros responsables de definir su comportamiento basado en la proximidad entre observaciones. La comparación entre ambas configuraciones permite identificar cómo cambian los niveles de error, la capacidad predictiva y la relación entre desempeño en entrenamiento y prueba. Esta información resulta fundamental para comprender las limitaciones

del enfoque y para establecer en qué medida la optimización contribuye o no a mejorar su estabilidad y adecuación para la estimación de la humedad del suelo.

Tabla 29. Comparación de resultados KNN por defecto vs optimizado

Métrica	Modelo base	Modelo optimizado	Mejora absoluta	Mejora relativa (%)
MAE (Test)	3.850	3.425	0.425	11.04%
RMSE (Test)	5.120	5.069	0.050	0.98%
R^2 (Test)	0.865	0.780	-0.084	-9.76%
R^2 (Train)	0.995	0.939	-0.055	-5.54%
Overfitting	0.130	0.159	+0.029	+22.5% (mayor sobreajuste)

La comparación entre el modelo KNN con parámetros por defecto y su versión optimizada muestra mejoras en las métricas de error, especialmente en el MAE, lo que refleja una mayor precisión promedio en las predicciones. Sin embargo, la optimización no logra incrementar la capacidad explicativa del modelo en la fase de prueba, donde incluso se observa una ligera disminución del ajuste. Además, el aumento en la diferencia entre los resultados de entrenamiento y prueba evidencia un mayor nivel de sobreajuste, indicando que el modelo optimizado tiende a ajustarse demasiado a los datos de entrenamiento.

Tabla 30. Resultados folds KNN

Modelo	Fold	R^2	MAE Train	RMSE Train
KNN	1	0.950	1.505	2.668
	2	0.925	1.576	3.219
	3	0.942	1.496	2.957
	4	0.942	1.359	2.810
	5	0.937	1.619	2.989
MEDIA ± DE		**0.939 ± 0.008**	**1.511 ± 0.099**	**2.92 ± 0.206**

Configuración de hiperparámetros: Algorithm: ball tree, leaf size: 20, metric: minkowski, neighbors: 2, P: 1, KNN weights: uniform.

Los resultados obtenidos en la Tabla 30 en los diferentes folds muestran un desempeño consistente del modelo KNN, con valores de R^2 relativamente altos y errores de entrenamiento que se mantienen dentro de rangos similares. La baja variabilidad entre pliegues sugiere que la configuración seleccionada permite una reproducción estable del comportamiento del modelo frente a distintas particiones de los datos. En conjunto, la media y desviación estándar indican que el modelo presenta un buen nivel de ajuste y estabilidad, logrando capturar adecuadamente las relaciones presentes en los datos de entrenamiento con un margen de error controlado.

Tabla 31.Resultados KNN.

Top	Algorithm	Metric	p	N neighbors	R^2 Entrenamiento	R^2 Prueba	MAE (Test)	RMSE (Test)
1	ball_tree	minkowski	1	2	0.939	**0.780**	**3.425**	**5.069**
2	kd_tree	manhattan	3	2	0.939	0.780	3.425	5.069
3	kd_tree	manhattan	1	2	0.939	0.780	3.425	5.069

Las tres configuraciones anteriores *(*Tabla 31*)* ofrecen resultados idénticos en métricas promedio, lo que demuestra que la elección del algoritmo de búsqueda (*ball_tree* vs *kd_tree*) y de la métrica de distancia (*Minkowski* vs *Manhattan*) no afecta significativamente el rendimiento del modelo. Esto sugiere que el comportamiento del KNN está dominado por el número de vecinos (k = 2) más que por la estructura o métrica empleada.

6.4 Análisis y selección del modelo

La evaluación conjunta de los cinco algoritmos implementados permite establecer una comparación integral de su desempeño en la predicción de la humedad volumétrica del suelo. En primer lugar, los resultados confirman la superioridad de los métodos ensemble frente a las aproximaciones basadas en redes neuronales, máquinas de soporte y modelos basados en vecinos más cercanos. En particular, XGBoost alcanza los mayores niveles de capacidad explicativa y los menores errores de predicción, tanto en los modelos base como en sus versiones optimizadas, mientras que Random Forest se ubica en una segunda posición con métricas ligeramente inferiores, pero manteniendo un rendimiento sólido y consistente. Esta comparación se hace evidente en la
Tabla 32.

Tabla 32. Comparación resultados modelos

Modelo	R^2 Entrenamiento	R^2 Prueba	ΔR^2	MAE (Test)	RMSE (Test)	Clasificación de desempeño
XGBoost	0.9999	0.9607	0.0392	1.949	2.669	Excelente (mejor rendimiento global)
Random Forest	1.000	0.9192	0.0808	2.687	3.828	Muy bueno
SVR	0.6948	0.7693	0.0755	4.685	7.607	Aceptable (lineal estable)
MLP	0.7675	0.6989	0.0686	3.884	5.762	Regular (mejor en estabilidad que en precisión)
KNN	0.9399	0.7806	0.1593	3.425	5.070	Limitado (sensibilidad alta a ruido local)

Al analizar las métricas derivadas de la validación cruzada, se observa en la Figura 23 qué, XGBoost y Random Forest presentan combinaciones favorables de alto R^2, errores medios moderados y brechas reducidas entre entrenamiento y prueba (ΔR^2), lo que indica una buena capacidad de generalización. La variación entre pliegues es acotada y las desviaciones estándar de las métricas se mantienen en rangos bajos, aun en presencia de la heterogeneidad espacial y temporal propia del conjunto de datos. En contraste, modelos como MLP y KNN muestran un comportamiento más sensible a la partición de los datos, con mayor dispersión en los indicadores de desempeño, mientras que SVR,

aunque estable, no alcanza niveles de precisión comparables con los métodos ensemble.

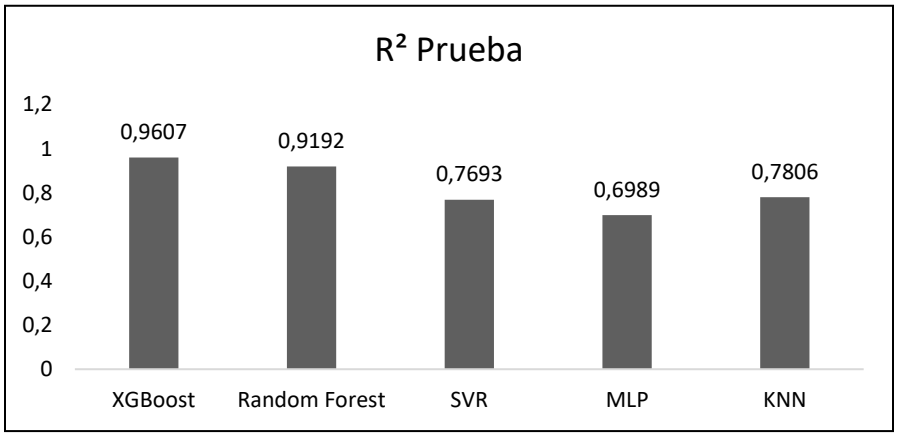

Figura 23. Gráfica comparación métricas de resultados

Desde una perspectiva metodológica, en la Figura 23 podemos observar qué, XGBoost ofrece varias ventajas para este problema específico. Su esquema de boosting secuencial permite capturar interacciones no lineales entre variables climáticas, índices espectrales y humedad del suelo, corrigiendo de manera iterativa los errores residuales y controlando el sobreajuste mediante términos de regularización.

Además, como se observa en la Figura 24, la combinación de tasa de aprendizaje, profundidad de los árboles y parámetros de muestreo otorga flexibilidad para adaptarse a la estructura de los datos sin sacrificar estabilidad. Random Forest, por su parte, destaca por su robustez frente al ruido y su menor sensibilidad a la configuración de hiperparámetros; su mecanismo de agregación de múltiples árboles entrenados con subconjuntos de datos contribuye a reducir la varianza del modelo y proporciona estimaciones confiables, aunque con un techo de desempeño ligeramente inferior al de XGBoost.

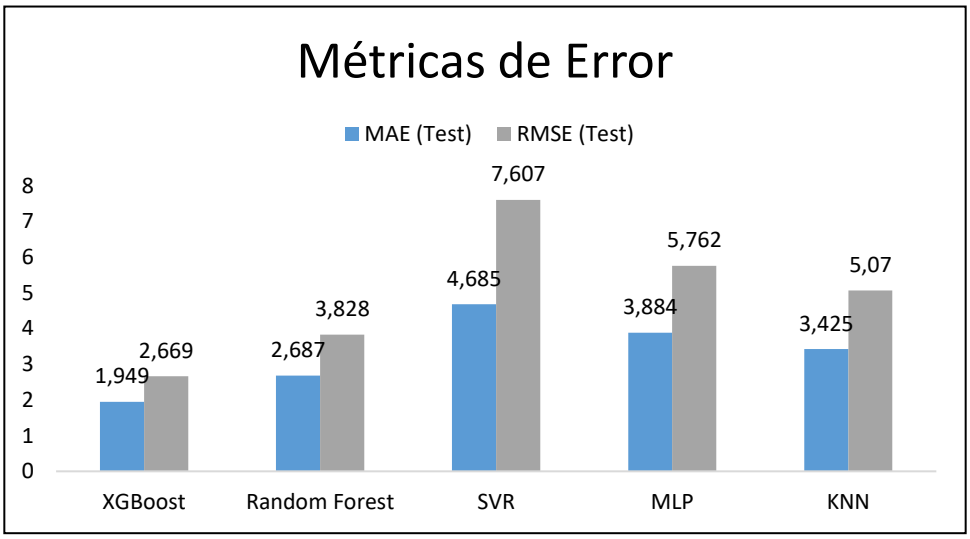

Figura 24. Gráfica comparación métricas de error.

Los modelos SVR y MLP cumplen un rol complementario en el análisis comparativo. SVR logra capturar tendencias globales con errores moderados y una brecha de generalización controlada; no obstante, su capacidad para representar la complejidad del sistema se ve limitada cuando aumenta la dimensionalidad y se intensifica la colinealidad entre predictores. El MLP optimizado, si bien reduce parte del error absoluto respecto a su versión base, mantiene valores de R^2 inferiores a los obtenidos por los métodos ensemble y evidencia una mayor sensibilidad a la inicialización de pesos y a la escala de los datos, lo que dificulta su uso como modelo principal en contextos con muestras relativamente acotadas.

Finalmente, KNN se consolida como una referencia no paramétrica sencilla, útil para contrastar el desempeño de algoritmos más complejos. Su funcionamiento basado en la proximidad local de las observaciones resulta intuitivo, pero su rendimiento se ve penalizado por la alta dimensionalidad del conjunto de variables y por la presencia de ruido, produciendo errores de magnitud superior y menor capacidad explicativa. Además, su escalabilidad y costo computacional crecen con el tamaño del conjunto de datos, lo que limita su aplicabilidad práctica más allá de escenarios exploratorios.

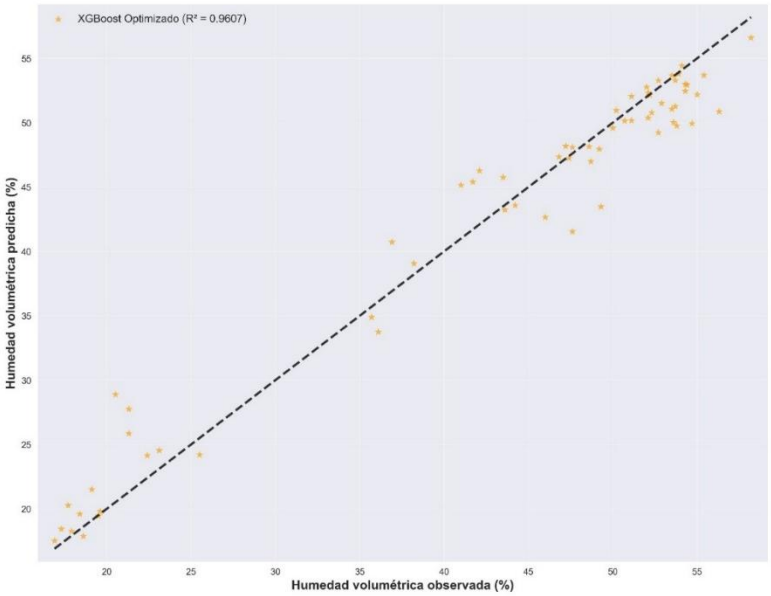

Figura 25.Gráfico de dispersión XGBoost vs Valores reales.

En síntesis, como se observa en las Figura 25, la comparación global indica que XGBoost constituye la opción más adecuada como modelo final para la predicción de la humedad volumétrica del suelo en la parcela analizada, al combinar alta precisión, buena generalización y capacidad para manejar variables correlacionadas y relaciones no lineales. Random Forest se posiciona como una alternativa robusta y de más sencilla interpretación, especialmente útil en contextos donde se priorice la transparencia del modelo. Los resultados obtenidos con SVR, MLP y KNN aportan un marco de contraste que permiten corroborar que los modelos no presentan un desempeño competitivo entre sí, dado que las diferencias observadas favorecen ampliamente a los métodos ensemble. En consecuencia, se reafirma que XGBoost y Random Forest respectivamente, constituyen las alternativas más apropiadas para integrar datos climáticos y satelitales en aplicaciones de monitoreo de humedad del suelo.

Análisis de resultados frente a otros autores

La comparación entre los resultados obtenidos en la presente investigación y los hallazgos reportados por Houben et al. (2025), Li & Yan (2024) y Alahmad et al. (2025) permite establecer una perspectiva crítica sobre el desempeño de los modelos de machine learning aplicados a la estimación de la humedad volumétrica del suelo. En todos los estudios analizados se identifica una tendencia consistente: los modelos basados en árboles de decisión y técnicas de boosting alcanzan los niveles más altos de precisión y estabilidad, mientras que los algoritmos lineales o fundamentados en proximidad presentan limitaciones al enfrentarse a la complejidad intrínseca de variables espectrales y ambientales.

En los resultados obtenidos, el modelo XGBoost se destacó como el de mejor desempeño general, al registrar valores elevados de R^2 en prueba, una brecha mínima entre entrenamiento y validación y los errores más bajos entre los algoritmos evaluados. Este comportamiento concuerda con lo

documentado por Li & Yan (2024), quienes demostraron que los métodos de boosting superan sistemáticamente a Random Forest, SVR y KNN cuando las relaciones entre los predictores no son lineales y existe correlación entre variables espectrales. Aunque en su estudio se emplea información SAR además de datos ópticos, las métricas alcanzadas en esta investigación son comparables aún sin la inclusión de sensores de radar. Esto resalta la efectividad del uso combinado de imágenes PlanetScope y variables climáticas para la predicción de la humedad del suelo en contextos agrícolas tropicales.

Por otro lado, Houben et al. (2025) enfatizan la importancia de integrar variables topográficas, edáficas y climáticas en modelos espacio-temporales para mejorar la capacidad predictiva. Los resultados obtenidos en esta investigación muestran una tendencia similar: variables atmosféricas como el déficit de humedad, la radiación solar y la velocidad del viento actuaron como moduladores claves del balance hídrico, mientras que los índices NDVI y NDMI aportaron información vegetal complementaria relevante para explicar la variación de la humedad volumétrica. No obstante, a diferencia del enfoque multiescala utilizado por Houben et al., el presente estudio se desarrolló a nivel de parcela, lo que permitió capturar microvariaciones ambientales con mayor fidelidad. Esta diferencia metodológica explica por qué, aun con un conjunto de predictores más limitado en términos espaciales, los resultados alcanzados mantienen un desempeño competitivo.

Por su parte, Alahmad et al. (2025) evidencian que XGBoost y Random Forest presentan un mejor rendimiento que modelos secuenciales como LSTM en escenarios donde la variabilidad temporal está mediada principalmente por condiciones atmosféricas y no por patrones estrictamente recurrentes. Esta tendencia coincide con lo observado en esta investigación, donde algoritmos como SVR, MLP y KNN mostraron un comportamiento más inestable y con mayor error, reflejando una menor capacidad para modelar interacciones no lineales complejas entre las variables climáticas y espectrales. La coincidencia en los patrones de error sugiere que los modelos ensemble resultan más apropiados cuando los predictores presentan alta colinealidad y ruido ambiental, condiciones comunes en estudios de teledetección agrícola.

En síntesis, la comparación con la literatura reciente confirma que los resultados alcanzados se encuentran alineados con las tendencias identificadas a nivel internacional. El desempeño del modelo XGBoost demuestra que, incluso sin incorporar sensores SAR o variables topográficas, es posible lograr niveles de precisión comparables mediante el uso de información óptica de alta resolución y variables atmosféricas representativas. Asimismo, las diferencias metodológicas relacionadas con la escala espacial, la resolución temporal y el tipo de predictores permiten posicionar esta investigación como una contribución pertinente al estudio de la humedad del suelo en ambientes agrícolas tropicales, al demostrar que un enfoque basado en conjuntos de datos compactos, adecuadamente procesados y optimizados, puede alcanzar resultados altamente competitivos y científicamente robustos.

7. DESARROLLO DE UNA INTERFAZ INTERACTIVA PARA LA PREDICCIÓN DE HUMEDAD VOLUMÉTRICA DEL SUELO

Se construyó una aplicación web para predecir en tiempo real la humedad volumétrica del suelo (HV%) mediante un modelo XGBoost previamente entrenado. La interfaz, implementada en Streamlit (Python), permite: (i) ajustar nueve variables de entrada mediante deslizadores con rangos reales observados y (ii) obtener inmediatamente la predicción de HV% con una clasificación por categorías (Muy baja, Baja, Media, Alta).

Tecnologías: Python 3.10+, Streamlit, XGBoost, Pandas, NumPy, Matplotlib y scikit-learn.

7.1 Entrenamiento y cacheo del modelo

Una vez depurada la base de datos y definido el modelo XGBoost optimizado, se procedió a integrarlo dentro del entorno de ejecución de Streamlit. Para optimizar el tiempo de respuesta y evitar el reentrenamiento continuo del modelo cada vez que el usuario interactúa con la aplicación, se empleó el sistema de almacenamiento en caché de Streamlit.
Esta función permite mantener en memoria tanto los datos como el modelo previamente entrenado, garantizando que las operaciones posteriores se realicen de manera casi instantánea. En concreto, se utilizaron dos decoradores: @st.cache_data, encargado de almacenar los conjuntos de entrenamiento y prueba junto con los datos originales sin estandarizar; y @st.cache_resource, que conserva en memoria el modelo XGBoost una vez ajustado.

El proceso de carga incluyó la lectura de los archivos base de entrenamiento y de referencia que contienen los valores originales y estandarizados de las variables predictoras. Con esta configuración, la interfaz utiliza los valores medios, mínimos y máximos de cada variable para generar los controles deslizantes (sliders), mientras que el modelo XGBoost realiza la predicción con base en los parámetros ajustados en la etapa de optimización: n_estimators = 50, max_depth = 6 y learning_rate=0.2.

Gracias a esta estructura, el sistema responde de manera inmediata ante cualquier cambio en los valores de entrada, sin necesidad de reprocesar la información, lo cual resulta esencial para lograr una experiencia de usuario eficiente y en tiempo real.

7.2 Estructura de la Interfaz

La aplicación fue diseñada bajo un enfoque de interacción intuitiva, organizada en una disposición de dos columnas principales que permite al usuario observar simultáneamente las variables de entrada y los resultados de la predicción.

En la columna izquierda se ubica el módulo de *parámetros de entrada*, que contiene los controles deslizantes o *sliders* correspondientes a las nueve variables climáticas y espectrales empleadas por el modelo XGBoost. Cada control se genera dinámicamente a partir de los valores reales de la base de datos original, es decir, utilizando los mínimos, máximos y promedios observados, lo que garantiza que el usuario manipule únicamente rangos coherentes con el comportamiento histórico de cada variable. Los valores ajustados por el usuario se transforman internamente mediante un

proceso de estandarización z-score, para mantener la compatibilidad con el modelo previamente entrenado.

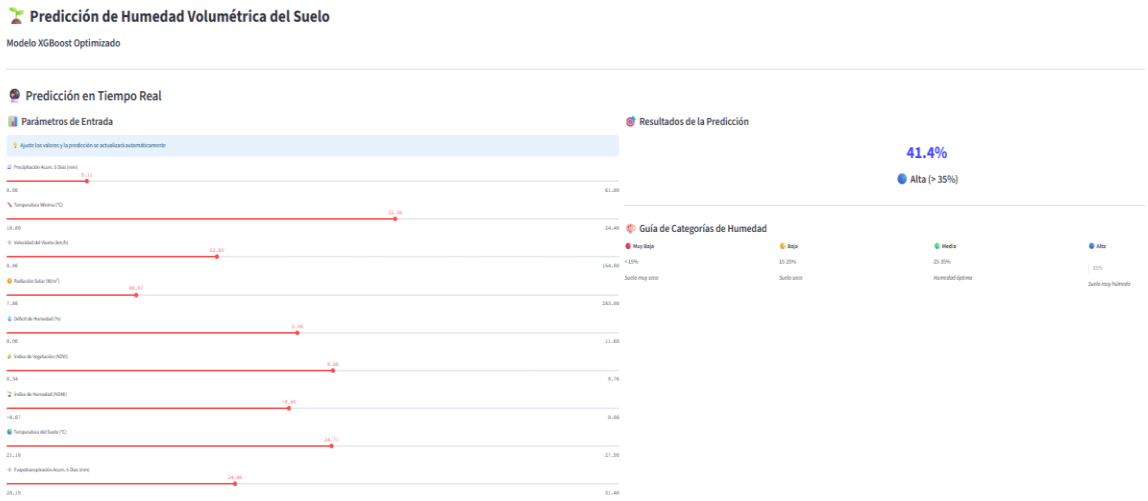

Figura 26.Interfaz del modelo.

Por su parte, la columna derecha presenta de forma visual e inmediata los *resultados de la predicción*. En la parte superior se muestra el valor estimado de humedad volumétrica del suelo, acompañado de un color y un ícono que reflejan su categoría agronómica (Muy baja, Baja, Media o Alta). Esta información se actualiza automáticamente conforme el usuario modifica cualquier variable en la columna izquierda, simulando un comportamiento de *predicción en tiempo real*.
A continuación, la interfaz despliega las métricas de desempeño del modelo, que incluyen el coeficiente de determinación (R^2), el error absoluto medio (MAE) y la raíz del error cuadrático medio (RMSE). Estos indicadores permiten evaluar de forma transparente la precisión y consistencia del modelo XGBoost con respecto al conjunto de datos de validación.

En la columna derecha se muestra el resultado de la predicción de humedad volumétrica, expresado como porcentaje (%), junto con su clasificación agronómica. La interfaz presenta el valor estimado acompañado de un color y un ícono representativo, lo que facilita su interpretación visual. Las categorías definidas son:

- Muy Baja (<15%) – Suelo muy seco
- Baja (15–25%) – Déficit moderado
- Media (25–35%) – Condición óptima
- Alta (>35%) – Suelo muy húmedo o saturado

Esta clasificación se actualiza de manera automática cada vez que el usuario modifica los valores de entrada, ofreciendo una simulación en tiempo real del estado hídrico del suelo.

Validación visual

Aunque la interfaz no compara directamente valores observados y predichos ya que su propósito es la estimación en tiempo real, su comportamiento fue verificado con los resultados del modelo previamente validado. En la aplicación, al configurar valores representativos de condiciones relativamente húmedas (por ejemplo, precipitación acumulada de 8.1 mm, temperatura mínima de 22.3 °C, radiación solar de 66 W/m² y NDVI = 0.56), el modelo estimó una humedad volumétrica del 41.4 %, clasificada como Alta (> 35 %) según la escala de referencia. Esta predicción visible en la Figura 28 se muestra de manera inmediata y categorizada por color, lo que permite interpretar de forma intuitiva el estado hídrico del suelo y su posible implicación agronómica.

Los resultados del modelo XGBoost que alimenta la interfaz mostraron un desempeño robusto durante su validación (R^2 = 0.96; MAE = 1.95 %; RMSE = 2.94 %), confirmando su estabilidad y precisión. De este modo, la herramienta web se constituye en un instrumento operativo para el monitoreo dinámico del contenido de agua en el suelo, con potencial aplicación en la gestión del riego, la planificación de fertilización y la detección temprana de déficit o exceso de humedad.

8. CONCLUSIONES

El estudio cumplió de manera integral el objetivo de desarrollar un modelo predictivo para estimar la humedad volumétrica del suelo en cultivos del CIAT mediante técnicas de aprendizaje automático, integrando de forma coherente información proveniente de sensores de campo, variables meteorológicas e índices espectrales derivados de imágenes satelitales PlanetScope. Este enfoque multifuente permitió capturar la complejidad del sistema suelo–planta–atmósfera y construir una base de datos robusta y representativa con 319 registros diarios y 10 variables, garantizando consistencia temporal, coherencia física y calidad estadística.

El análisis exploratorio permitió identificar los principales factores que determinan la dinámica hídrica del suelo, destacando la precipitación acumulada y la evapotranspiración de cinco días como indicadores complementarios de la oferta y la demanda hídrica. La incorporación de variables climáticas, radiativas y espectrales como el NDVI y el NDMI_Custom fortaleció la interpretación ecofisiológica del modelo y redujo la colinealidad entre predictores, lo que favoreció la estabilidad del aprendizaje.

Tras la evaluación comparativa de diferentes algoritmos, el modelo XGBoost optimizado demostró el mejor desempeño con un $R^2 = 0.96$, MAE = 1.95 y RMSE = 2.94, evidenciando alta precisión, bajo sesgo y excelente capacidad de generalización ($\Delta R^2 = 0.039$). Su estructura basada en ensambles secuenciales permitió modelar eficazmente relaciones no lineales entre variables espectrales, climáticas y edáficas, superando a modelos como Random Forest, SVR, MLP y KNN. Este resultado confirma el potencial del aprendizaje boosting como herramienta central en la predicción agroclimática de precisión.

Como resultado complementario, se desarrolló una interfaz web interactiva en Streamlit conectada al modelo XGBoost, la cual permite realizar predicciones en tiempo real de la humedad volumétrica del suelo, visualizar métricas de desempeño y clasificar los resultados en rangos agronómicos (Muy baja, Baja, Media y Alta). Esta aplicación traduce los avances científicos en una herramienta práctica y accesible, facilitando la interpretación de resultados y la toma de decisiones sobre riego, manejo hídrico y monitoreo de cultivos.

En conjunto, la investigación demuestra la viabilidad técnica y operativa del aprendizaje automático multifuente para la predicción de la humedad del suelo, contribuyendo al fortalecimiento de la agricultura digital sostenible y sentando las bases para el desarrollo de sistemas predictivos replicables y escalables en zonas agrícolas tropicales de Colombia.

9. TRABAJOS FUTUROS

- Integración de nuevas fuentes satelitales y sensores in situ
  Se recomienda incorporar imágenes de sensores de radar (Sentinel-1, SAOCOM) y térmicos (Landsat, ECOSTRESS), con el fin de complementar la información óptica de PlanetScope y mejorar la estimación en condiciones de nubosidad o saturación del suelo. Asimismo, la integración de redes de sensores de humedad in situ permitiría una calibración espacial más precisa del modelo.
- Desarrollo de modelos híbridos y espacio-temporales
  Futuras investigaciones podrían explorar modelos híbridos que combinen redes neuronales profundas (CNN, LSTM) con algoritmos ensemble como XGBoost o CatBoost, permitiendo capturar dependencias espacio-temporales más complejas y mejorar la predicción en series continuas de tiempo.
- Incorporación de variables de manejo agronómico y edáficas detalladas
  La inclusión de información sobre textura del suelo, materia orgánica, profundidad radicular y prácticas de manejo (labranza, riego, fertilización) permitiría enriquecer el modelo y mejorar la interpretación agronómica de los resultados, facilitando la toma de decisiones en campo.
- Desarrollo de un sistema operativo de monitoreo en tiempo real
  Se plantea la creación de una plataforma web o dashboard interactivo (basada en Streamlit o Dash) conectada a APIs climáticas y satelitales, que ejecute el modelo XGBoost en tiempo real y emita alertas automáticas de déficit o exceso hídrico a nivel de parcela.

10. REFERENCIAS BIBLIOGRÁFICAS

[1] Brady, N. C., & Weil, R. R. (2016). *The nature and properties of soils* (15th ed.). Pearson.
[2] Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Springer.
[3] Bouma, J. (2018). Soil degradation: A major global challenge. *Journal of Soil and Water Conservation, 73*(5), 127A–134A.
[4] Reynolds, W. D., & Drury, C. F. (2020). Soil compaction in cropping systems: A review of the nature, causes, and possible solutions. *Soil and Tillage Research, 204*, 104719. https://doi.org/10.1016/j.still.2020.104719
[5] Bouman, B. A. M., & Tuong, T. P. (2001). Field water management to save water and increase its productivity in irrigated lowland rice. *Agricultural Water Management, 49*(1), 11–30.
[6] Rossel, R. A. V., et al. (2016). Visible–near infrared spectroscopy as a tool to predict soil physical and chemical properties for applications in precision agriculture. *Soil Science Society of America Journal, 80*(4), 927–938.
[7] Xue, J., & Su, B. (2017). Significant remote sensing vegetation indices: A review of developments and applications. *Journal of Sensors*. https://doi.org/10.1155/2017/1353691
[8] Gómez, C., Lagacherie, J. P., & Martin, G. C. (2016). Remote sensing of soil properties: A review. *Remote Sensing of Environment, 173*, 234–254.
[9] Padarian, J., Minasny, B., & McBratney, A. B. (2020). Machine learning for soil property prediction: Applications, challenges, and solutions. *Environmental Modelling & Software, 114*, 194–209.
[10] Mulder, V. L., de Bruin, S., Schaepman, M. E., & Mayr, T. R. (2011). The use of remote sensing in soil and terrain mapping: A review. *Geoderma, 162*, 1–19.
[11] Minasny, B., & McBratney, A. B. (2016). Digital soil mapping: A brief history and some lessons. *Geoderma, 264*, 301–311.
[12] Bouma, J. (2021). Soil science contributions towards sustainable development goals and their implementation: Linking soil functions with ecosystem services. *Journal of Plant Nutrition and Soil Science, 184*(3), 1–14.
[13] Fathololoumi, S., et al. (2021). Effect of multi-temporal satellite images on soil moisture prediction using a digital soil mapping approach. *Geoderma, 385*, 114844. https://doi.org/10.1016/j.geoderma.2020.114844
[14] Ge, X., et al. (2021). Estimating agricultural soil moisture content through UAV-based hyperspectral images in arid regions. *Remote Sensing, 13*(8), 1562. https://doi.org/10.3390/rs13081562
[15] Escobar-González, D., et al. (2024). Predicción de la humedad del suelo mediante aprendizaje por transferencia: An application in the High Tropical Andes. *Water, 16*(8), 832. https://doi.org/10.3390/w16060832
[16] Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.
[17] Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.
[18] Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32.
[19] Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning, 20*(3), 273–297.
[20] Kamilaris, A., & Prenafeta-Boldú, F. X. (2018). Deep learning in agriculture: A survey.

Computers and Electronics in Agriculture, 147, 70–90.
[21] Jones, H. G. (2014). *Plant water relations and the control of transpiration*. Cambridge University Press.
[22] Willmott, C. J., & Matsuura, K. (2005). Advantages of the mean absolute error (MAE) over the root mean square error (RMSE). *Climate Research, 30*, 79–82.
[23] Gao, B. C. (1996). NDWI—A normalized difference water index for remote sensing of vegetation liquid water from space. *Remote Sensing of Environment, 58*(3), 257–266.
[24] Zhang, C., & Kovacs, J. M. (2012). The application of small UAVs for precision agriculture. *Precision Agriculture, 13*(6), 693–712.
[25] Muñoz-Carpena, R., & Dukes, M. D. (2009). Automatic soil moisture-based drip irrigation for improving water use efficiency. *Agricultural Water Management, 96*(11), 1285–1296.
[26] Bouman, B. A. M., Humphreys, E., Tuong, T. P., & Barker, R. (2007). Rice and water. *Advances in Agronomy, 92*, 187–237.
[27] FAO. (2021). *Rice market monitor*. Food and Agriculture Organization of the United Nations.
[28] IPCC. (2021). *Climate change 2021: The physical science basis*. Cambridge University Press.
[29] Jones, H. G. (2004). Irrigation scheduling: Advantages and pitfalls of plant-based methods. *Journal of Experimental Botany, 55*(407), 2427–2436.
[30] Peters, M., Tiemann, T., & Horne, P. (2013). *Tropical forages*. CIAT.
[31] Thornton, P. K., & Herrero, M. (2010). Potential for reduced methane and carbon dioxide emissions from livestock and pasture management in the tropics. *Proceedings of the National Academy of Sciences, 107*(46), 19667–19672.
[32] Zhang, Y., Li, T., Wu, X., & Zhang, J. (2018). Predicting soil moisture with machine learning models. *Computers and Electronics in Agriculture, 145*, 243–251.
[33] FAO. (2021). *The state of the world's land and water resources for food and agriculture: Systems at breaking point*. FAO.
[34] Dorigo, W. A., et al. (2017). ESA CCI soil moisture for improved Earth system understanding. *Remote Sensing of Environment, 203*, 185–215. https://doi.org/10.1016/j.rse.2017.07.001
[35] Lakshmi, V. (2017). Remote sensing of soil moisture. *ISPRS Journal of Photogrammetry and Remote Sensing, 128*, 116–123.
[36] Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics, 29*(5), 1189–1232.
[37] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). https://doi.org/10.1145/2939672.2939785
[38] Khan, S., et al. (2022). Remote sensing-based estimation of soil moisture using machine learning algorithms—A review. *Journal of Hydrology, 607*, 127542. https://doi.org/10.1016/j.jhydrol.2022.127542
[39] Brocca, P., Melone, F., Moramarco, T., & Morbidelli, R. (2010). Soil moisture temporal stability over experimental areas in Central Italy. *Hydrology and Earth System Sciences, 14*(5), 739–751. https://doi.org/10.5194/hess-14-739-2010
[40] Lamichhane, M., Mehan, S., & Mankin, K. R. (2025). Soil moisture prediction using remote sensing and machine learning algorithms: A review on progress, challenges, and

opportunities. *Remote Sensing, 17*, 2397. https://doi.org/10.3390/rs17142397
[41] Baldwin, D., Manfreda, S., Keller, K., & Smithwick, E. A. H. (2017). Predicting root zone soil moisture. *Journal of Hydrology, 546*, 393–404. https://doi.org/10.1016/j.jhydrol.2017.01.020
[42] Wang, Y., Shi, L., Hu, Y., Hu, X., Song, W., & Wang, L. (2024). A comprehensive study of deep learning for soil moisture prediction. *Hydrology and Earth System Sciences, 28*, 917–943. https://doi.org/10.5194/hess-28-917-2024
[43] Milà, C., Ludwig, M., Pebesma, E., Tonne, C., & Meyer, H. (2024). Random forests with spatial proxies for environmental modelling. *Geoscientific Model Development, 17*, 6007–6033.
[44] Cheng, M., et al. (2022). Estimation of soil moisture content under high maize canopy coverage. *Agricultural Water Management, 264*, 107530.
[45] Lamichhane, M., Mehan, S., & Mankin, K. R. (2025). *Duplicado de [40]* → ELIMINAR
[46] Zhang, Y., Liang, S., Zhu, Z., Ma, H., & He, T. (2022). Soil moisture content retrieval from Landsat 8 data. *ISPRS Journal of Photogrammetry and Remote Sensing, 185*, 32–47.
[47] Zhang, Y., Parazoo, N. C., Williams, A. P., Zhou, S., & Gentine, P. (2020). Large and projected strengthening moisture limitation. *Proceedings of the National Academy of Sciences, 117*, 9216–9222.
[48] Houben, T., Peiffer, S. D., Vanderborght, J., & Vereecken, H. (2025). Machine-learning based spatiotemporal prediction of soil moisture. *Vadose Zone Journal, 24*(3), 1–18.
[49] Li, M., & Yan, Y. (2024). Comparative analysis of machine-learning models for soil moisture estimation. *Land, 13*(2), 1–22.
[50] Alahmad, T., Alquraish, H., Aldhafferi, M., & Khan, A. (2025). Spatiotemporal prediction of soil moisture content. *Frontiers in Soil Science, 3*, 1–15.
[51] Cheng, L., Wang, J., Liu, S., & Zhao, X. (2023). Integration of multi-source satellite and meteorological data. *Agricultural Water Management, 280*, 108194. https://doi.org/10.1016/j.agwat.2023.108194
[52] Zhang, Y., Guo, Y., Li, X., Chen, D., & Yang, Q. (2022). Evaluating short-term soil moisture dynamics. *Remote Sensing of Environment, 273*, 112988. https://doi.org/10.1016/j.rse.2022.112988